Distributional Regression

Estimation of Conditional Distributions with Likelihood Ratio Order Constraint

Victor Ryan

Delft University of Technology

Distributional Regression

Estimation of Conditional Distributions with Likelihood Ratio Order Constraint

by

Victor Ryan

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on August 29 at 10:30 AM.

Student number:	4650840	
Project duration:	January 8, 2024 – August 29, 2024	
Thesis committee:	Prof. dr. ir. G. (Geurt) Jongbloed,	TU Delft, supervisor
	Dr. Ir. G.F. (Tina) Nane	Graduation Committee

Preface

Since high school, I have always enjoyed geography. In particular, I found it intriguing to learn how rain is formed, how earthquakes and tsunamis occur, and more. This thesis journey started with my search for a supervisor with a research interest in weather or climate sciences. Fortunately, I found a kind and supportive supervisor named Prof. Dr. Ir. G. (Geurt) Jongbloed. He introduced me to several papers and authors working in these fields. After reading them, I discovered the concept of ”distributional regression”. This type of regression focuses on predicting the probability of an event occurring, given an observed value. That is, one is interested in estimating the conditional distribution functions. The paper that I encountered applied this method in the context of weather prediction. Since regression is also one of my research interests I decided to delve into distributional regression.

I have been working on this thesis for eight months and have thoroughly enjoyed my research journey. It has been a great learning experience, allowing me to explore new concepts such as scoring rules and stochastic order. These new concepts brought new challenges, and redoing the proofs and finding examples helped me understand them better. Besides the theoretical part, the programming aspect of this project has also been difficult. This is due to the large code base of such a project is large, much larger than any modelling project that I had taken during my bachelor's studies (except my bachelor thesis). Organising and commenting on my code was crucial in helping review and debug my code. Comments also made the code more readable.

Although there were challenges that I overcame by myself, I also received significant help from those around me. First of all, I am grateful to my supervisor for guiding me into the topics and for his support during these eight months. He recommended several books related to the thesis and ensured that I was prepared for my thesis defence. I also thank him for connecting me with others in the scientific community who share my research interests. Further, I am grateful to Dr. A. F. F. (Alexis) Derumigny for letting me focus on working on the thesis while we also have a project that we almost finished. He taught me how to organize my code for larger projects and explained the process of publishing a research paper.

I am deeply thankful to my mother Fena Indra Suryani, my father Djoni Arryadi Karta and my sister Shirley Ramadhani for their support and for encouraging me to pursue what I love: studying and advancing my education. The mental support from my beloved girlfriend, Dwi Juliana Sari, has been invaluable. Her willingness to listen to my hardships kept me motivated to continue working on this project. Lastly, I would like to thank my close friends Sandra Fonhof and Kevin Luo for boosting my confidence to pursue an academic career. My years in TU Delft have also been enjoyable, thanks to Zuhair, Daniël Cohen and Huan.

I hope the reader finds the research in this thesis intriguing and is inspired to tackle the problem addressed in this thesis with their own methods.

Victor Ryan
Delft, 20 August 2024

Abstract

This thesis aims to estimate conditional distribution functions subject to the likelihood ratio order constraint. We use the modified gradient projection method to ensure that in each iteration, the point stays feasible while improving the objective function. Regarding the objective function, we use the continuous ranked probability score (CRPS), a loss function used for forecast evaluation. Given its strict propriety, we use it for estimation procedures. Our numerical experiments indicate that the estimated conditional distribution functions perform reasonably well as the number of iteration increases. However, the algorithm's long running time makes it impractical for use in practice. Furthermore, due to the reparametrization of the estimand, the objective function loses its convexity property while the feasible set is convex. This causes the algorithm to potentially return a local minimum, rather than a global minimum.

Contents

1 Introduction

Regression is a procedure which estimates the relationship between a variable Y that depends on a vector of covariates $\mathbf{X} = (X_1, \dots, X_d)^\intercal$. The relationship between Y and $\mathbf{X}$ is described through an unknown function f and a random error ε, that is,

$$Y = f(\mathbf{X}) + \varepsilon.$$

In linear regression, for instance, the relationship of Y with $\mathbf{X}$ is described by the conditional average $\mathbb{E}[Y|\mathbf{X}]$. This conditional expectation is then assumed to be linear in the parameter. Suppose for instance, the conditional expectation is modelled in the following manner: $\mathbb{E}[Y|\mathbf{X}] = \beta_0 + \sum_{k=1}^{d} \beta_k X_k$, then the goal in linear regression is to estimate the unknown coefficients β_i, for all $0 \leq i \leq d$. The rigidness of this relationship may be relaxed by assuming that $\mathbb{E}[Y|\mathbf{X}]$ belongs to a class of functions parameterized in a non-linear way. All in all, the relationship between Y and $\mathbf{X}$ is usually summarized by one quantity, which is the expected value.

There is another type of regression, in which the whole conditional distribution $Y|\mathbf{X}$ is of interest instead of only the conditional expectation. Such regression is called *distributional regression* (Fahrmeir et al., 2013). Similar to usual regression, there are parametric and non-parametric distributional regression. In the parametric case, the conditional distribution of $Y|\mathbf{X}$ is assumed to be a distribution P, which belongs to known class of distributions $\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$, where Θ denotes the parameter space with finite dimension. The goal is to estimate the unknown parameter $\boldsymbol{\theta}$. For non-parametric approaches, the parameter is the distribution function of $Y|\mathbf{X}$, which is only assumed to belong to a much bigger class of distributions that cannot be smoothly parameterized with a finite dimensional parameter set. Because the conditional distribution function is of interest, this type of regression is used when the forecasts are probabilistic.

An example of an application of distributional regression is in the meteorological field. The physics of the atmosphere may be described using partial differential equations, which are then solved numerically (Kalnay, 2002, p. 136). Due to the chaotic nature of these partial differential equations, the model predicts long-term weather inaccurately. Therefore, an ensemble of multiple runs of numerical weather prediction (NWP) models is used, to account for the uncertainty (Gneiting and Raftery, 2005). These runs differ in the initial conditions and/or the parameterization of the model. The model outputs are then statistically post-processed, to quantify the uncertainty of the model outputs (Gneiting and Katzfuss, 2014; Schefzik et al., 2013). One method of post-processing the outputs is to do distributional regression. The goal of this approach is to find the conditional distribution function of the weather quantity Y (e.g. temperature, wind or precipitation), given the ensemble member forecasts $\mathbf{X}$ (Gneiting et al., 2005).

Following from the example above, if multiple runs of NWP indicate a larger value of the weather quantity (e.g. high temperature, heavy precipitation), then regressing of these outputs on the weather quantity should be large as well intuitively (Henzi et al., 2021). This yields that the probability of observing large value of the weather quantity is also large. If we were to do the usual regression, then we expect that the conditional expectation increases with the covariate values. Such a regression is called the *isotonic regression*. The usual linear regression, in which the relationship is increasing, is an example of such a regression. In Henzi et al. (2021), the concept of isotonicity is extended for distributional regression. The isotonic relationship for this regression refers to the order constraint on the conditional distribution function of $Y|X = x$. To be more precise, the isotonic distributional regression preserves the order $F(y|x_1) \geq F(y|x_2)$ for any $y \in \mathbb{R}$, if $x_1, x_2 \in \mathbb{R}$ such that $x_1 \leq x_2$, where we denote $F(\cdot|x)$ as the conditional distribution function of $Y|X = x$. If the conditional distribution function satisfies this relationship, then we denote it as $[Y|X = x_1] \leq_{\mathrm{st}} [Y|X = x_2]$, which is an example of a stochastic order.

In fact, new inference techniques have been developed that impose different types of stochastic order constraints. An example of a stochastic order is the likelihood ratio order, which is a stronger order than we discussed previously. This order stipulates that if X and Y have density function f_X and f_Y respectively, then X is stochastically smaller than Y in likelihood ratio order if and only if the density ratio $f_Y(t)/f_X(t)$ increases in t. Imposing this order in a regression context means that the ratio of the conditional densities $f(y|x')/f(y|x)$ increases in y for $x \leq x'$. Recently, Mösching and Dümbgen (2024) show how to estimate the conditional distribution functions with this stochastic order as a constraint. Their technique estimates $F(y|x)$ non-parametrically by constructing empirical likelihood, which is maximized with the likelihood ratio order constraint.

Inspired by the work of Henzi et al. (2021) and Mösching and Dümbgen (2024), we attempt to estimate conditional distribution functions while adhering to the likelihood ratio order constraint. In this thesis, however, we use expected loss (risk) as the criterion function. We specifically choose the continuous ranked probability score (CRPS) for the loss function, which is an example of a *scoring rule*. A scoring rule takes two arguments: the probability distribution function used for the forecasting and the realization. It then assigns these two arguments to a numerical value (Gneiting and Katzfuss, 2014).

This numerical value may be used for evaluation of the forecast and estimation procedure (Gneiting and Raftery, 2007). In particular, we use a strictly proper scoring rule for the latter, such as CRPS, which will be explained later in the thesis. The use of strictly proper scoring rule ensures that the risk of probabilistic forecasting using a distribution function is minimum if and only if the said distribution function is the true distribution of the realization. Henzi et al. (2021) demonstrate using such a scoring rule (we will see that CRPS is strictly proper), to solve the isotonic distributional regression problem.

Therefore in this thesis, we aim to estimate the conditional distribution function similarly as in Mösching and Dümbgen (2024), but instead, we minimize the expected CRPS loss with the likelihood ratio order constraint. To solve the minimization problem numerically, we use the gradient projection method.

Before we delve deep into the main body of the thesis, we outline how this thesis is organized. Chapter two through chapter four review the concepts that we have mentioned in this introduction. We start in Chapter 2 which discusses the different types of stochastic orders. Here we will see that the likelihood ratio order is the strongest stochastic order. In Chapter 3, we discuss scoring rules and give examples of (strictly) proper scoring rules. It turns out that there is a connection between scoring rules and information theory. The algorithm that is proposed by Mösching and Dümbgen (2024) is described in Chapter 4. We include the information on their algorithm in order to understand their method. We also apply the likelihood approach when the conditional distribution functions are chosen to belong to a class of normal distribution. In Chapter 5, we formulate the empirical risk with CRPS as the loss function and the likelihood ratio order constraint. It turns out that this feasible set is non-convex. Although, we transform it so that it becomes a convex set, the objective function then becomes a non-convex function. Lastly, Chapter 6 describes the algorithm that we use to solve the optimization problem. We also show how the number of constraints can be lowered substantially, and present the results of a small simulation study. In these results, we visually compare the performance of the algorithm developed in this thesis with the one proposed by Mösching and Dümbgen (2024). Their algorithm are implemented in the programming language `R` (R Core Team, 2024), with a package name `LRDistReg` (see Mösching and Dümbgen (2022)).

2 Stochastic Orders

In the introduction, we have briefly introduced several terminologies such as stochastic order and scoring rules. In this chapter, we begin by defining the stochastic orders more formally. There are four orders that we will discuss: usual stochastic order, likelihood ratio order and the monotone hazard rate order. We will show examples of random variables that follow these orders. We also show how these orders is related. The definitions and theorems of the usual stochastic order, the hazard rate order and the likelihood ratio order are taken from Shaked and Shanthikumar (2007). The proofs of the theorems here are not new, but we add details on them for the clarity because the proofs/details are often omitted in these sources. Lastly, the definition of the monotone hazard ratio order is taken from Wu and Westling (2023).

2.1. The usual stochastic order

We start with the definition of the usual stochastic order.

Definition 2.1.1 (The usual stochastic order)**.** *Let X and Y be real-valued random variables. We say that X is smaller than Y in the usual stochastic order (denoted by $X \leq_{\mathrm{st}} Y$) if for any $x \in \mathbb{R}$,*

$$\mathbb{P}(X \leq x) \geq \mathbb{P}(Y \leq x). \tag{2.1}$$

Intuitively, we say $X \leq_{\mathrm{st}} Y$ if and only if the probability of X taking a smaller value is higher than Y taking that same value. From (2.1), we also have that $X \leq_{\mathrm{st}} Y$ if and only if for any $x \in \mathbb{R}$,

$$\mathbb{P}(X > x) \leq \mathbb{P}(Y > x).$$

One way to characterize the usual stochastic ordering is to use the *coupling* method. The following is a definition of coupling from van der Hofstad (2016).

Definition 2.1.2 (Coupling)**.** *Let X and Y be random variables on probability spaces $(\Omega_X, \mathcal{F}_X, \mathbb{P}_X)$ and $(\Omega_Y, \mathcal{F}_Y, \mathbb{P}_Y)$ respectively. The random variable $\left(\hat{X}, \hat{Y}\right)$ is a coupling of X and Y if there exists a new probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that the marginal distributions have the following properties:*

$$\mathbb{P}\left(\hat{X} \leq x\right) = \mathbb{P}_X(X \leq x) \quad \textit{and} \quad \mathbb{P}\left(\hat{Y} \leq y\right) = \mathbb{P}_Y(Y \leq y).$$

That is, the marginal distribution of $\hat{X}$ is the distribution X, similarly the marginal distribution of $\hat{Y}$ is the distribution of Y.

The following theorem from van der Hofstad (2016) states that there exists a coupling X and Y, say $\left(\hat{X}, \hat{Y}\right)$, such that $\hat{X} \leq \hat{Y}$ almost surely if and only if $X \leq_{\mathrm{st}} Y$.

Theorem 2.1.3. *Let X and Y be random variables. There exists coupling $\left(\hat{X}, \hat{Y}\right)$ of X and Y such that $\mathbb{P}\left(\hat{X} \leq \hat{Y}\right) = 1$ if and only if $X \leq_{\mathrm{st}} Y$.*

Proof. The proof of the coupling implies the usual stochastic order is shown by using the marginal property of the coupling. Indeed, suppose that X and Y are coupled by $(\hat{X}, \hat{Y})$ such that $\mathbb{P}(\hat{X} \leq \hat{Y}) = 1$. For any $x \in \mathbb{R}$, we have

$$\begin{aligned}\mathbb{P}(Y \leq x) = \mathbb{P}(\hat{Y} \leq x) &= \mathbb{P}(\hat{Y} \leq x | \hat{X} \leq \hat{Y}) \mathbb{P}(\hat{X} \leq \hat{Y}) + \mathbb{P}(\hat{Y} \leq x | \hat{X} > \hat{Y}) \mathbb{P}(\hat{X} > \hat{Y}) \\ &= \mathbb{P}(\hat{Y} \leq x | \hat{X} \leq \hat{Y}) \\ &= \mathbb{P}(\hat{X} \leq \hat{Y} \leq x) \\ &\leq \mathbb{P}(\hat{X} \leq x) = \mathbb{P}(X \leq x).\end{aligned}$$

The inequality is obtained by using $\mathbb{P}(A \cap B) \leq \min\{\mathbb{P}(A), \mathbb{P}(B)\}$, which is true since $\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B) \leq \mathbb{P}(B)$ and $\mathbb{P}(A \cap B) = \mathbb{P}(B|A)\mathbb{P}(A) \leq \mathbb{P}(A)$.

For the converse, we assume that $X \leq_{\text{st}} Y$. Let F^{-1} denote the generalized inverse function of the distribution function F, i.e. for $\alpha \in [0, 1]$,

$$F^{-1}(\alpha) = \inf\{x \in \mathbb{R} : F(x) \geq \alpha\}.$$

Then it is known that $F^{-1}(U)$ has distribution function F if U is uniform on $(0, 1)$. Indeed, for any $x \in \mathbb{R}$,

$$\mathbb{P}(F^{-1}(U) \leq x) = \mathbb{P}(U \leq F(x)) = F(x). \tag{2.2}$$

Now, let F_X and F_Y be the distribution function of X and Y respectively. Define $\hat{X} := F_X^{-1}(U)$ and $\hat{Y} := F_Y^{-1}(U)$. Then $(\hat{X}, \hat{Y})$ have the same marginal distribution as X and Y by (2.2). By the assumption $F_X(x) \geq F_Y(x)$ for any $x \in \mathbb{R}$, which follows $F_X^{-1}(\alpha) \leq F_Y^{-1}(\alpha)$ for any $\alpha \in [0, 1]$. Hence,

$$\mathbb{P}(\hat{X} \leq \hat{Y}) = \mathbb{P}(F_X^{-1}(U) \leq F_Y^{-1}(U)) = 1.$$

□

We give examples of two random variables that satisfy the usual stochastic order.

Example 2.1.4. Let X and Y follow the Bernoulli distribution with parameters p_1 and p_2 respectively, such that $0 \leq p_1 < p_2 \leq 1$. To show that $X \leq_{\text{st}} Y$, we couple the two random variables by using the uniform distribution. Let U be uniformly distribution on $(0, 1)$. Let $\hat{X} = \mathbb{1}(U \in (0, p_1])$ and $\hat{Y} = \mathbb{1}(U \in (0, p_2])$, the marginal distributions of $\hat{X}$ and $\hat{Y}$ correspond to the distribution of X and Y. Then,

$$\mathbb{P}(\hat{X} \leq \hat{Y}) = \mathbb{P}(\hat{X} = 0, \hat{Y} = 0) + \mathbb{P}(\hat{X} = 0, \hat{Y} = 1) + \mathbb{P}(\hat{X} = 1, \hat{Y} = 1).$$

These terms will sum up to one because

$$\begin{aligned}\mathbb{P}(\hat{X} = 0, \hat{Y} = 0) &= \mathbb{P}(U \in (p_2, 1]) = 1 - p_2 \\ \mathbb{P}(\hat{X} = 0, \hat{Y} = 1) &= \mathbb{P}(U \in (p_1, p_2]) = p_2 - p_1 \\ \mathbb{P}(\hat{X} = 1, \hat{Y} = 1) &= \mathbb{P}(U \in (0, p_1]) = p_1.\end{aligned}$$

In fact, the probability of the event $\{\hat{X} = 1, \hat{Y} = 0\}$ is zero since if $\hat{X} = 1$, then $U \in (0, p_1] \backslash (p_1, p_2]$ and so it is necessary that $\hat{Y} = 1$ as well. We conclude that,

$$\mathbb{P}(\hat{X} \leq \hat{Y}) = 1,$$

and therefore $X \leq_{\text{st}} Y$ by Theorem 2.1.3. △

Example 2.1.5. Let X and Y follow the geometric distributions with parameter p_1 and p_2 respectively, such that $0 < p_1 < p_2 \leq 1$. We interpret this distribution as the number of trials needed to get one success. The claim is that $Y \leq_{\text{st}} X$.

One may prove this claim directly by using the Definition 2.1.1. Indeed, for any $k \in \mathbb{N}$, by using the geometric series,

$$\mathbb{P}(X > k) = 1 - \sum_{j=1}^{k} p_1(1 - p_1)^{j-1} = (1 - p_1)^k \leq (1 - p_2)^k = \mathbb{P}(Y > k).$$

The inequality is true since $p \mapsto (1-p)^k$ is a decreasing function in $p \in (0,1)$. Hence, $Y \leq_{\mathrm{st}} X$.

A different approach is to use the coupling argument. The idea in this case is to consider two infinite sequence of i.i.d. Bernoulli random variables, in which each element of the sequence are coupled according to Example 2.1.4. The coupling of X and Y is constructed by finding which element of the sequence is a success.

We interpret 'success' as an event that takes a value one. Let $(Y_i)_{i\in\mathbb{N}}$ and $(Z_i)_{i\in\mathbb{N}}$ be i.i.d. Bernoulli sequence such that Y_i and Z_i follow the Bernoulli distributions with parameters p_1 and p_2 respectively, for any $i \in \mathbb{N}$. Let $\left(\hat{Y}_i, \hat{Z}_i\right)$ be a coupling from Example 2.1.4 for any $i \in \mathbb{N}$. The coupling of X and Y can be constructed by defining the following random variables

$$\hat{X}_1 := \min\left\{j : \hat{Y}_j = 1\right\} \quad \text{and} \quad \hat{X}_2 := \min\left\{k : \hat{Z}_k = 1\right\}.$$

Note that $\hat{X}_1$ and $\hat{X}_2$ follow the geometric distribution with parameter p_1 and p_2 respectively, since the tuple $\left(\hat{Y}_i, \hat{Z}_i\right)$ is independent for each $i \in \mathbb{N}$. Further $\mathbb{P}\left(\hat{X}_2 \leq \hat{X}_1\right) = 1$ because the $\hat{Z}_k$ will be equal to 1 earlier than $\hat{Y}_j$ since the event $\{\hat{Z}_k = 1\}$ has higher probability than the event $\{\hat{Y}_k = 1\}$. △

2.2. The hazard rate order

The term hazard rate may be encountered in survival analysis. It is a statistical analysis where the outcome of interest is the time at which a certain event happens, e.g. the mortality time of a patient, the lifespan of an electrical device. The hazard rate is the rate of a subject not surviving for an additional infinitesimally change in time $t \geq 0$, given that it has survived longer than t.

Let us define the hazard rate more formally, which is taken from Karim and Islam (2019). Let T be a non-negative random variable and $t \geq 0$, then the *hazard rate* is the following function:

$$h(t) := \lim_{\delta t \to 0^+} \frac{\mathbb{P}(t < T \leq t + \delta t | T \geq t)}{\delta t}. \tag{2.3}$$

In survival analysis, one usually considers non-negative random variables. This is not required for defining the hazard rate order.

Definition 2.2.1 (The hazard rate order). *Let X and Y be real-valued random variables, with $h_X(t)$ and $h_Y(t)$ being the associated hazard rates. We say that X is smaller than Y in the hazard rate order (denoted by $X \leq_{\mathrm{hr}} Y$) if for any $t \in \mathbb{R}$,*

$$h_X(t) \geq h_Y(t). \tag{2.4}$$

To gain an intuition of the hazard rate order, we first rewrite (2.3) in terms of the probability of surviving. Suppose T has a density f_T, then (2.3) may be written in terms of f_T and $\mathbb{P}(T \geq t)$. The quantity $S_T(t) := \mathbb{P}(T \geq t)$ is called the *survival function* in survival analysis. This function is non-increasing, since $S_T(t) = 1 - \mathbb{P}(T \leq t)$. From (2.3), we obtain that

$$h(t) = \frac{1}{\mathbb{P}(T \geq t)} \lim_{\delta t \to 0^+} \frac{\mathbb{P}(t < T \leq t + \delta t)}{\delta t} = \frac{f_T(t)}{S_T(t)} \geq 0. \tag{2.5}$$

The hazard rate is non-negative because density functions are non-negative and the range of the survival function is the interval $[0,1]$. If f_T is the derivative of the distribution function F_T w.r.t. t, then from (2.5),

$$h(t) = \frac{1}{S_T(t)} \frac{d}{dt} F_T(t) = \frac{1}{S_T(t)} \frac{d}{dt}[1 - S_T(t)] = -\frac{1}{S_T(t)} \frac{d}{dt} S_T(t) = -\frac{d}{dt} \log S_T(t).$$

Therefore, the hazard rate is the instantaneous change of $-\log S(t)$. A small value of hazard rate at a given point corresponds to a small change in $-\log S(t)$ for an infinitesimal increase in t. This corresponds to a slow decrease in the probability of survival and therefore a higher lifespan. On the other hand, a higher hazard rate implies a shorter lifespan.

Using (2.5), we can show that $X \leq_{\mathrm{hr}} Y$ implies $X \leq_{\mathrm{st}} Y$. This means that the hazard rate order is a stronger order than the usual stochastic order.

Theorem 2.2.2. *Let X and Y be two non-negative random variables with density function f_X and f_Y respectively. If $X \leq_{\mathrm{hr}} Y$, then $X \leq_{\mathrm{st}} Y$.*

Proof. If $X \leq_{\mathrm{hr}} Y$, then (2.4) is equivalent to

$$\frac{f_X(t)}{\mathbb{P}(X \geq t)} \geq \frac{f_Y(t)}{\mathbb{P}(Y \geq t)}.$$

The values on the numerator and denominator on both sides of the inequalities are always non-negative. Therefore, it is necessary that $f_X(t) \geq f_Y(t)$ and $\mathbb{P}(X \geq t) \leq \mathbb{P}(Y \geq t)$ for the inequality to be true. In any case, $\mathbb{P}(X \geq t) \leq \mathbb{P}(Y \geq t)$, which is equivalent to $\mathbb{P}(X < t) \geq \mathbb{P}(Y < t)$. Hence, $X \leq_{\mathrm{st}} Y$. □

Example 2.2.3. Let X and Y be exponentially distributed with parameters $\lambda_1, \lambda_2 > 0$ respectively, such that $\lambda_1 < \lambda_2$. Then,

$$S_X(t) = 1 - \int_0^t \lambda_1 e^{-\lambda_1 x}\, dx = e^{-\lambda_1 t} \qquad \text{and} \qquad S_Y(t) = 1 - \int_0^t \lambda_2 e^{-\lambda_2 x}\, dx = e^{-\lambda_2 t}.$$

Therefore, the corresponding hazard rates are

$$h_X(t) = \frac{\lambda_1 e^{-\lambda_1 t}}{e^{-\lambda_1 t}} = \lambda_1 \qquad \text{and} \qquad h_Y(t) = \frac{\lambda_2 e^{-\lambda_2 t}}{e^{-\lambda_2 t}} = \lambda_2.$$

Since $\lambda_2 > \lambda_1$, we conclude that $Y \leq_{\mathrm{hr}} X$. △

2.3. The likelihood ratio order

This particular stochastic order is the strongest order, in the sense that the monotone likelihood ratio order implies the other two previously mentioned orders. Let us define the likelihood ratio order.

Definition 2.3.1 (The likelihood ratio order)**.** *Let X and Y be real-valued random variables with density function f_X and f_Y respectively. Let $\mathrm{supp}(X)$ denote the support of X, i.e. $\mathrm{supp}(X) = \mathrm{cl}\{x \in \mathbb{R} : f_X(x) > 0\}$, where $\mathrm{cl}(A)$ means the closure of a set A. Then X is smaller than Y in the likelihood ratio order (denoted as $X \leq_{\mathrm{lr}} Y$) if*

$$\frac{f_Y(t)}{f_X(t)} \quad \textit{is increasing in } t \in \mathrm{supp}(X) \cup \mathrm{supp}(Y).$$

In other words, we have $X \leq_{\mathrm{lr}} Y$ if and only if for any $x, y \in \mathrm{supp}(X) \cup \mathrm{supp}(Y)$ and $x \leq y$, then

$$\frac{f_Y(y)}{f_X(y)} \geq \frac{f_Y(x)}{f_X(x)},$$

or equivalently,

$$f_X(x) f_Y(y) \geq f_X(y) f_Y(x) \quad \text{for any } x \leq y.$$

As a remark, this definition assumes that the probability measure is absolutely continuous w.r.t. the Lebesgue measure. Dümbgen and Mösching (2023) generalize it for arbitrary probability measures on $\mathbb{R}$ that are absolutely continuous w.r.t. a general measure.

The following theorem shows that the likelihood ratio order is a sufficient condition for random variables to follow the hazard rate order. This yields that the likelihood ratio order is the strongest stochastic order among the usual and hazard rate orders.

Theorem 2.3.2. *Let X and Y be real-valued random variables with density function f_X and f_Y respectively. If $X \leq_{\mathrm{lr}} Y$, then $X \leq_{\mathrm{hr}} Y$ and hence $X \leq_{\mathrm{st}} Y$.*

Proof. Let F_X and F_Y denote the distribution function of X and Y respectively. Assume $X \leq_{\mathrm{lr}} Y$, then for any $x \leq y$,

$$f_X(x) f_Y(y) \geq f_X(y) f_Y(x). \tag{2.6}$$

Integrating both sides of inequality in (2.6) w.r.t. x on interval $(-\infty, y]$ yields

$$\int_{-\infty}^{y} f_X(x) f_Y(y)\, dx \geq \int_{-\infty}^{y} f_X(y) f_Y(x)\, dx \quad \Leftrightarrow \quad f_Y(y) F_X(y) \geq f_X(y) F_Y(y). \tag{2.7}$$

We again integrate both sides of inequality in (2.6), but w.r.t. y on interval $[x,\infty)$ yields

$$\int_x^\infty f_X(x)f_Y(y)\,dy \geq \int_x^\infty f_X(y)f_Y(x)\,dy \quad \Leftrightarrow \quad f_X(x)(1-F_Y(x) \geq f_Y(x)(1-F_X(x)). \tag{2.8}$$

From (2.7) and (2.8), we obtain the following two inequalities:

$$\frac{f_Y(x)}{f_X(x)} \geq \frac{F_Y(x)}{F_X(x)} \quad \text{and} \quad \frac{f_Y(x)}{f_X(x)} \leq \frac{1-F_Y(x)}{1-F_X(x)}. \tag{2.9}$$

From (2.9), we obtain that

$$\frac{f_X(x)}{1-F_x(x)} = h_X(x) \geq h_Y(x) = \frac{f_Y(y)}{1-F_Y(y)},$$

and so $X \leq_{\mathrm{hr}} Y$. Further, we combine both inequalities in (2.9) to get

$$\frac{1-F_Y(x)}{1-F_X(x)} \geq \frac{F_Y(x)}{F_X(x)} \quad \Leftrightarrow \quad \frac{F_X(x)}{1-F_X(x)} \geq \frac{F_Y(x)}{1-F_Y(x)}.$$

Hence, $\mathbb{P}(X \leq x) \geq \mathbb{P}(Y \leq x)$, which means that $X \leq_{\mathrm{st}} Y$. As a remark, the proof of $X \leq_{\mathrm{hr}} Y$ implies $X \leq_{\mathrm{st}} Y$ is also in Theorem 2.2.2. Furthermore, the proof is for X and Y being continuous random variables. The proof for discrete random variables is similar, but we replace the integrals with sums. □

We demonstrate the likelihood ratio order by giving the following examples and an example in which two random variables follow the usual stochastic order, but not the likelihood ratio order.

Example 2.3.3. Let $X \sim \mathcal{N}(\mu_1,\sigma^2)$ and $Y \sim \mathcal{N}(\mu_2,\sigma^2)$ such that $\mu_1 \leq \mu_2$. Let f_X and f_Y be the densities of X and Y respectively. The ratio of the densities is

$$\frac{f_Y(t)}{f_X(t)} = \exp\left(\frac{(t-\mu_1)^2}{2\sigma^2} - \frac{(t-\mu_2)^2}{2\sigma^2}\right) = \exp\left(\frac{2t(\mu_2-\mu_1)+\mu_1^2-\mu_2^2}{2\sigma^2}\right). \tag{2.10}$$

The ratio in (2.10) is increasing in t, for any $t \in \mathbb{R}$. Indeed, because $\mu_1 \leq \mu_2$, we have $\mu_2 - \mu_1 \geq 0$. Therefore, the mapping $t \mapsto 2t(\mu_2-\mu_1)$ is increasing in t. Hence, $X \leq_{\mathrm{lr}} Y$.

However, the likelihood ratio order does not hold if we vary the variance when both random variables have the same fixed mean. Consider two centered normal distributions with different variances, i.e. $X \sim \mathcal{N}(0,\sigma_1^2)$ and $Y \sim \mathcal{N}(0,\sigma_2^2)$, such that $0 < \sigma_1 \leq \sigma_2$. Then for any $t \in \mathbb{R}$, we have

$$\frac{f_Y(t)}{f_X(t)} = \frac{\sigma_1}{\sigma_2}\exp\left(\frac{t^2(\sigma_2^2-\sigma_1^2)}{2\sigma_1^2\sigma_2^2}\right),$$

which is not increasing due to the term t^2. The ratio tends to infinity as $t \to \pm\infty$, and it has a minimum at $t=0$. △

Example 2.3.4. Let $X, Y \in \{1,2,3\}$ and we define their probability mass functions as follows

$$\mathbb{P}(X=x) = \begin{cases} 0.2 & \text{if } x=1, \\ 0.7 & \text{if } x=2, \\ 0.1 & \text{if } x=3. \end{cases} \quad \text{and} \quad \mathbb{P}(Y=x) = \begin{cases} 0.1 & \text{if } x=1, \\ 0.8 & \text{if } x=2, \\ 0.1 & \text{if } x=3. \end{cases}$$

We have that

$$\begin{gathered} \mathbb{P}(X \leq 1) = 0.2 \geq 0.1 = \mathbb{P}(Y \leq 1) \\ \mathbb{P}(X \leq 2) = 0.2+0.7 = 0.9 \geq 0.9 = 0.1+0.8 = \mathbb{P}(Y \leq 2) \\ \mathbb{P}(X \leq 3) = 1 \geq 1 = \mathbb{P}(Y \leq 3), \end{gathered}$$

which means that $X \leq_{\mathrm{st}} Y$. However,

$$\frac{\mathbb{P}(Y=1)}{\mathbb{P}(X=1)} = \frac{0.1}{0.2} = \frac{1}{2}, \qquad \frac{\mathbb{P}(Y=2)}{\mathbb{P}(X=2)} = \frac{0.8}{0.7} > 1, \qquad \frac{\mathbb{P}(Y=3)}{\mathbb{P}(X=3)} = \frac{0.1}{0.1} = 1.$$

The ratio $\mathbb{P}(Y=x)/\mathbb{P}(X=x)$ is not increasing in x, and therefore X is not smaller than Y in the likelihood ratio order. Note that we use the definition of the likelihood ratio order for discrete random variables. In this setting, the density function in the definition is replaced by the probability mass function. △

Example 2.3.5. We consider a distribution that is given by Henzi et al. (2021). Let Y and X be real-valued random variables. Suppose $X \sim \mathrm{Unif}(0,10)$ and

$$Y|X = x \sim \mathrm{Gamma}(k(x) = \sqrt{x}, \theta(x) = \min\{\max\{x,1\},6\}),$$

where $k(x) > 0$ for any $x > 0$ is the shape parameter and $\theta(x) > 0$ is the scale parameter. Then the density of $Y|X = x$ is

$$f_{Y|X}(y|x) = \frac{1}{\Gamma(k(x))\theta(x)^{k(x)}} y^{k(x)-1} e^{-y/\theta(x)} \mathbb{1}\{y > 0\}.$$

Assume $0 < x \leq x' < 10$ and let $(k(x), \theta(x))$ and $(k(x'), \theta(x'))$ be the shape and scale parameter of $Y|X = x$ and $Y|X = x'$ respectively. Then the ratio of densities is

$$\frac{f_{Y|X}(y|x')}{f_{Y|X}(y|x)} = \frac{\Gamma(k(x))}{\Gamma(k(x'))} \frac{\theta(x)^{k(x)}}{\theta(x')^{k(x')}} y^{k(x')-k(x)} e^{y/\theta(x)-y/\theta(x')} = C y^{k(x')-k(x)} e^{y/\theta(x)-y/\theta(x')}, \tag{2.11}$$

where

$$C := \frac{\Gamma(k(x))}{\Gamma(k(x'))} \frac{\theta(x)^{k(x)}}{\theta(x')^{k(x')}}$$

is a positive constant for a fixed x, x'.

Now, we will show that (2.11) is increasing in y for any $y \geq 0$. The mapping $y \mapsto y^{k(x')-k(x)}$ is increasing in y because $x \mapsto \sqrt{x}$ is an increasing function and so $k(x') - k(x) = \sqrt{x'} - \sqrt{x} \geq 0$. Further, the mapping $x \mapsto \min\{\max\{x,1\},6\}$ is increasing as well. The scale parameter then equal to either 1 or 6 for $x \in (0,1]$ and $x \in [6,10)$ respectively. For $x \in (1,6)$, the scale parameter increases linearly. Therefore, if $x \leq x'$, then $\theta(x) \leq \theta(x')$ and so $y \mapsto e^{y/\theta(x)-y/\theta(x')}$ is increasing in y because the exponent is non-negative. The ratio in (2.11) is increasing in y for any $y \geq 0$, and hence $[Y|X = x] \leq_{\mathrm{lr}} [Y|X = x']$ if $x \leq x'$. It also implies that $[Y|X = x] \leq_{\mathrm{st}} [Y|X = x']$. As a remark, $[Y|X = x] \leq_{\mathrm{st}} [Y|X = x']$ is still true for $X \sim \mathrm{Unif}(0,c)$ for any $c > 0$, and

$$Y|X = x \sim \mathrm{Gamma}(k(x) = \sqrt{x}, \theta(x) = \min\{\max\{x,a\},b\}), \qquad \text{for any } 0 < a < b < c.$$

△

Example 2.3.6. We adjust the following result from Lemma 3 in Mösching and Dümbgen (2024), which is valid for probability measures with dominating measures other than Lebesgue measure. Let X and Y be random variables which has density functions f_X and f_Y respectively, and assume $X \leq_{\mathrm{lr}} Y$. Let Z_λ be another random variable with density function $f_\lambda(z) := (1-\lambda)f_X(z) + \lambda f_Y(z)$, where $0 < \lambda < 1$. The claim is that $Z_\lambda \leq_{\mathrm{lr}} Z_\mu$ for any $0 < \lambda < \mu < 1$.

Indeed, let $0 < \lambda < \mu < 1$ and take any $z_1 \leq z_2$. Then

$$\begin{aligned} f_\lambda(z_1)f_\mu(z_2) - f_\lambda(z_2)f_\mu(z_1) &= [(1-\lambda)\mu - \lambda(1-\mu)]f_X(z_1)f_Y(z_2) \\ &\quad + [(1-\lambda)(1-\mu) - (1-\lambda)(1-\mu)]f_X(z_1)f_X(z_2) \\ &\quad + [\lambda(1-\mu) - (1-\lambda)\mu)]f_X(z_2)f_Y(z_1) + [\lambda\mu - \lambda\mu]f_Y(z_1)f_Y(z_2) \\ &= (\mu-\lambda)[f_X(z_1)f_Y(z_2) - f_Y(z_1)f_Y(z_2)] \geq 0, \end{aligned}$$

by the assumption that $X \leq_{\mathrm{lr}} Y$. Hence, $Z_\lambda \leq_{\mathrm{lr}} Z_\mu$ for any $0 < \lambda < \mu < 1$. △

2.4. The monotone hazard ratio order

Following our discussion about survival analysis, one might be interested in estimating the hazard ratio. This quantity gives the relative hazard of an experiment group with a control group. For instance, what the instantaneous risk of death of heavy-smoker patients relative to non-smoker patients is, at a time t, given the survival time t. A model that is often used for obtaining the hazard ratio is Cox's proportional hazard regression model, which was proposed by Cox (1972).

The Cox's proportional hazard regression model assumes that the hazard rate for an individual with covariates $\mathbf{X} \in \mathbb{R}^d$ is

$$h(t|\mathbf{X}) = h_0(t)\exp(\boldsymbol{\beta}^\top \mathbf{X}),$$

where h_0 represents the baseline hazard when the covariates $\mathbf{X} = \mathbf{0}$ and $\boldsymbol{\beta} \in \mathbb{R}^d$. Suppose $\mathbf{X} = X_1 \in \mathbb{R}$, then

$$\frac{h(t|X_1 + 1)}{h(t|X_1)} = \exp(\beta),$$

which yields that the hazard rate given an increase of X_1 of one unit, is proportional to the hazard rate given X_1. If $X_1 \in \{0, 1\}$, then, the ratio of hazard rate is $\exp(\beta)$. We interpret it as the mortality risk of patients that belong to group $X_1 = 1$ relative to patients from group $X_1 = 0$. In both of these cases, the hazard ratio is independent of time.

The assumption that the hazard ratio is constant with time, may be violated in practice. Fortunately, Wu and Westling (2023) developed a non-parametric estimator for estimating the hazard ratio, in the case that it exhibits a monotone behavior. In constructing the estimator, they define a new type of stochastic order called the monotone hazard ratio order.

Definition 2.4.1 (The monotone hazard ratio order)**.** *Let X and Y be real-valued random variables, with $h_X(t)$ and $h_Y(t)$ being the associated hazard rates. We say that X is smaller than Y in the hazard rate order (denoted by $X \leq_{\mathrm{mhr}} Y$) if*

$$\frac{h_Y(t)}{h_X(t)} \quad \textit{is non-decreasing in } t \in \mathrm{supp}(X) \cup \mathrm{supp}(Y). \tag{2.12}$$

As stated by Wu and Westling (2023), if two random variables are ordered in the monotone hazard ratio sense, then in general they are not ordered in the usual stochastic order, likelihood ratio, or hazard rate. The monotone hazard ratio order is also not implied by any of these three orders. This is illustrated by the following example taken from Wu and Westling (2023). We refer to the aforementioned article for more examples.

Example 2.4.2. Suppose first that we have two geometric random variables X_1 and X_2 with parameters $p_1, p_2 \in (0, 1]$ respectively. We still interpret this distribution as in Example 2.1.5. Then for $i \in \{1, 2\}$, the cumulative distribution is

$$F_{X_i}(t) = 1 - (1 - p_i)^t, \quad t \in \mathbb{N}.$$

By (2.5) the hazard rate for X_i's are

$$\begin{aligned} h_{X_i}(t) &= \frac{\mathbb{P}(X_i = t)}{\mathbb{P}(X_i \geq t)} \\ &= \frac{\mathbb{P}(X_i = t)}{\mathbb{P}(X_i = t) + \mathbb{P}(X_i > t)} \\ &= \frac{1}{1 + \mathbb{P}(X_i > t)/\mathbb{P}(X_i = t)} \end{aligned}$$

We have that

$$\frac{\mathbb{P}(X_i > t)}{\mathbb{P}(X_i = t)} = \frac{(1 - p_i)^t}{p_i(1 - p_i)^{t-1}} = \frac{1 - p_i}{p_i}$$

Therefore, the hazard rate for X_i is $h_{X_i}(t) = p_i$, and so the hazard ratio is

$$\frac{h_{X_2}(t)}{h_{X_1}(t)} = \frac{p_2}{p_1}\frac{1 - p_1}{1 - p_2},$$

which is non-decreasing in $t \in \mathbb{N}$. However, as we have seen in Example 2.1.5, we have to impose an order on the p_i's. If $p_1 > p_2$, then $X_1 \leq_{\mathrm{mhr}} X_2$ but $X_1 \leq_{\mathrm{st}} X_2$ is false. Therefore, $X_1 \leq_{\mathrm{lr}} X_2$ and $X_1 \leq_{\mathrm{hr}} X_2$ are also false. △

3

Scoring Rules

In this chapter, we define more formally what a scoring rule is and its interpretation. We also give examples of scoring rules and study their properties. Finally, there is a relation between the scoring rules and information theory, which we will discuss as well.

Suppose we forecast an event, in which its uncertainty is quantified by a probability distribution. The forecasts are based on a given data set. The question is, under the assumption that new events have the same distribution as those observed in the given data set, how well do the forecasts predict these new events? To assess the quality of the probabilistic forecasts, one might consider using scoring rules. In this section, we review well-known scoring rules and their properties. We mainly use Gneiting and Raftery (2007) for the basic information about the scoring rules. In this section, we also give several examples of scoring rules.

Let us first define scoring rules formally.

Definition 3.0.1 (Scoring rules)**.** *Let $\mathcal{P}$ be a convex class of probability measures on a measurable space $(\Omega, \mathcal{F})$. A function $S : \mathcal{P} \times \Omega \to \overline{\mathbb{R}}$ is called a scoring rule if for any $\mathbb{P} \in \mathcal{P}$, the function $S(\mathbb{P}, \cdot)$ is measurable, and the integral of $S(\mathbb{P}, \cdot)$ w.r.t. $\mathbb{Q}$ exists for any $\mathbb{Q} \in \mathcal{P}$.*

Before we interpret the scoring rule, there is a technicality that we need to address. Gneiting and Raftery (2007) requires S to be $\mathcal{P}$-quasi integrable, which means that for any $\mathbb{P} \in \mathcal{P}$, a real-valued measurable function f on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is $\mathbb{P}$-quasi integrable. Let $f = f^+ - f^-$, where $f^+ = \max\{f, 0\}$ and $f^- = \min\{-f, 0\}$. For f to be $\mathbb{P}$-quasi integrable, means that at least one of f^+ and f^- has a real integral w.r.t. $\mathbb{P}$. So then $\int f \, d\mathbb{P} \in \overline{\mathbb{R}}$. One may also say that the integral f w.r.t measure $\mathbb{P}$ exists (Bauer, 2001, p. 65).

We may interpret scoring rules in two ways: as a reward or a loss system. If the scoring rules are interpreted as a reward system, then the scoring rules are said to be *positively oriented*. If we view the scoring rules as a loss system, then the scoring rules are *negatively oriented*. In positive orientation, a forecaster that predicts the observed event well receives a high reward. Alternatively, scoring rules that are seen as a loss system give less loss on well-predicted events. In this section, we use the positive orientation, unless mentioned otherwise.

3.1. Proper and strictly proper scoring rules

A natural question that comes up is how can we construct forecasts such that the reward is high. Further, the scoring rule varies depending on the probability measure and the data. To this end, we will need a scoring rule such that the averaged reward is maximized when we predict events using the 'true' distribution of the data set. We refer to such scoring rule as *proper*. If the scoring rule has a unique distribution that maximizes the reward output, then the scoring rule is *strictly proper*.

Definition 3.1.1 (Proper and strictly proper scoring rules)**.** *Let $\mathcal{P}$ be a convex class of probability measures on a measurable space $(\Omega, \mathcal{F})$. Let $S : \mathcal{P} \times \Omega \to \overline{\mathbb{R}}$ be a scoring rule. Let $\mathbb{E}_{\mathbb{Q}}$ denote an expectation under the probability measure $\mathbb{Q} \in \mathcal{P}$. The scoring rule S is proper relative to $\mathcal{P}$, if for any $\mathbb{P}, \mathbb{Q} \in \mathcal{P}$,*

$$\mathbb{E}_{\mathbb{Q}}[S(\mathbb{Q}, \cdot)] = \int S(\mathbb{Q}, \omega) \, d\mathbb{Q}(\omega) \geq \int S(\mathbb{P}, \omega) \, d\mathbb{Q}(\omega) = \mathbb{E}_{\mathbb{Q}}[S(\mathbb{P}, \cdot)]. \tag{3.1}$$

We refer to S being strictly proper if the equality in (3.1) *holds if and only if* $\mathbb{P} = \mathbb{Q}$.

From the Definition 3.1.1, we observe the following. Suppose a given data set follows the distribution of $\mathbb{Q}$, but the forecaster uses $\mathbb{P}$ to predict events in the data. Then on average, the reward of predicting events using the true distribution should give the maximal reward, if the reward system (i.e. the scoring rule) is proper. If the distribution $\mathbb{Q}$ is the only distribution that gives maximal reward, then the scoring rule is strictly proper. We further observe that if S interpreted as a loss function, then predicting events using $\mathbb{Q}$ should give the least loss. In that case, (3.1) therefore becomes

$$\mathbb{E}_{\mathbb{Q}}[S(\mathbb{Q},\cdot)] \leq \mathbb{E}_{\mathbb{Q}}[S(\mathbb{P},\cdot)]. \tag{3.2}$$

We now give examples of scoring rules and show that they are (strictly) proper.

Example 3.1.2 (The Brier score)**.** The Brier score or the quadratic score is presented in Brier (1950) for weather forecast verification. For this score, the convex class of probability measures $\mathcal{P}$ is

$$\mathcal{P}_r := \left\{ \mathbf{p} = (p_\omega)_{\omega\in\Omega} \in \mathbb{R}^r : \forall \omega \in \Omega, p_\omega \geq 0, \sum_{\omega\in\Omega} p_\omega = 1 \right\}, \tag{3.3}$$

which is defined on a measurable space $(\Omega, \mathcal{F})$ such that $\Omega := \{1, 2, \ldots, r\}$. Then the Brier score is defined as follows:

$$S_{\mathrm{Brier}}(\mathbf{p}, j) = -\sum_{\omega=1}^{r} (p_\omega - \mathbb{1}(j = \omega))^2, \tag{3.4}$$

where $\mathbb{1}(j = \omega) = 1$ if the event ω occurs, otherwise its value is 0.

The interpretation of this score is quite straightforward. Suppose $\mathbf{p}$ is the probabilistic forecast and j is the event that we observe. Suppose the forecaster has 100% confidence that event ω' will occur, i.e. $p_{\omega'} = 1$ and $p_\omega = 0$ for any $\omega \neq \omega' \in \Omega$. If we observe event $j = \omega'$, then from (3.4), the Brier score takes a value of zero. On the other hand, if we observe event j that is not ω', then the score is negative. In this particular example, the score is equal to -2. The highest reward for 'perfect' forecasting is then zero.

To show that S_{Brier} is strictly proper, we compute $\mathbb{E}_{\mathbf{q}}[S_{\mathrm{Brier}}(\mathbf{q},\cdot)] - \mathbb{E}_{\mathbf{q}}[S_{\mathrm{Brier}}(\mathbf{p},\cdot)]$. For any fixed $j \in \Omega$, we have

$$\begin{aligned} S_{\mathrm{Brier}}(\mathbf{p}, j) &= -\sum_{\omega\in\Omega} p_\omega^2 - 2\mathbb{1}(j = \omega)p_\omega + \mathbb{1}(j = \omega)^2 \\ &= 2p_j - 1 - \sum_{\omega\in\Omega} p_\omega^2, \end{aligned}$$

and so,

$$S_{\mathrm{Brier}}(\mathbf{q}, j) - S_{\mathrm{Brier}}(\mathbf{p}, j) = 2(q_j - p_j) - \sum_{\omega\in\Omega} (q_\omega^2 - p_\omega^2).$$

Therefore for any $\mathbf{q}, \mathbf{p} \in \mathcal{P}$, we get

$$\begin{aligned} \mathbb{E}_{\mathbf{q}}[S_{\mathrm{Brier}}(\mathbf{q},\cdot)] - \mathbb{E}_{\mathbf{q}}[S_{\mathrm{Brier}}(\mathbf{p},\cdot)] &= \sum_{j\in\Omega} S_{\mathrm{Brier}}(\mathbf{q}, j)q_j - \sum_{j\in\Omega} S_{\mathrm{Brier}}(\mathbf{p}, j)q_j \\ &= \sum_{j\in\Omega} [S_{\mathrm{Brier}}(\mathbf{q}, j) - S_{\mathrm{Brier}}(\mathbf{p}, j)]q_j \\ &= \sum_{j\in\Omega} \left(2(q_j - p_j) - \sum_{\omega\in\Omega} (q_\omega^2 - p_\omega^2) \right) q_j \\ &= \sum_{j\in\Omega} 2q_j(q_j - p_j) - \sum_{\omega\in\Omega} (q_\omega^2 - p_\omega^2) \sum_{j\in\Omega} q_j \end{aligned}$$

$$
\begin{aligned}
&= \sum_{j\in\Omega} 2(q_j^2 - q_j p_j) - \sum_{j\in\Omega}(q_j^2 - p_j^2)\\
&= \sum_{j\in\Omega} q_j^2 - 2q_j p_j + p_j^2\\
&= \sum_{j\in\Omega}(q_j - p_j)^2 \geq 0,
\end{aligned}
$$

the squared Euclidean distance in $\mathbb{R}^r$ of $\mathbf{p}$ and $\mathbf{q}$. The Brier score satisfies the condition in (3.1), and therefore $S_{\text{Brier}}(\mathbf{p},\cdot)$ is a proper scoring rule. By the definition of metric distance, equality is attained if and only if $\mathbf{q} = \mathbf{p}$. △

Example 3.1.3 (The pseudo-spherical score)**.** While we find the pseudo-spherical score in Gneiting and Raftery (2007), the spherical score was already mentioned in Winkler and Murphy (1968). The measurable space $(\Omega, \mathcal{F})$ and the convex class of probability measures are the same as in Example 3.1.2. The pseudo-spherical score is defined as follows: for any $\alpha > 1$ and $\mathbf{p} \in \mathcal{P}_r$,

$$S_{\text{Spherical}}(\mathbf{p}, j) = \frac{p_j^{\alpha-1}}{\left(\sum_{k\in\Omega} p_k^\alpha\right)^{(\alpha-1)/\alpha}}.$$

According to Selten (1998), if $\alpha = 2$, then $S_{\text{Spherical}}$ is called spherical because the vector $\mathbf{p}$ is mapped to a vector on a unit sphere. Indeed, let $\alpha = 2$ and let

$$\mathbf{S}_{\text{Spherical}}(\mathbf{p}) := (S_{\text{Spherical}}(\mathbf{p}, 1), \dots, S_{\text{Spherical}}(\mathbf{p}, r))^\intercal.$$

Then the Euclidean norm of $\mathbf{S}_{\text{Spherical}}(\mathbf{p})$ is

$$||\mathbf{S}_{\text{Spherical}}(\mathbf{p})|| = \sqrt{\sum_{j\in\Omega}\left(\frac{p_j}{\left(\sum_{k\in\Omega} p_k^2\right)^{1/2}}\right)^2} = \sqrt{\frac{\sum_{j\in\Omega} p_j^2}{\sum_{k\in\Omega} p_k^2}} = 1.$$

For a general α, the norm of the vector $\mathbf{S}_{\text{Spherical}}(\mathbf{p})$ is generally not equal to one. The name pseudo-spherical scores most likely refer to the generalized version of the spherical score when $\alpha = 2$.

To show that the pseudo-spherical scoring rule is proper, we use Hölder's inequality. For any $\mathbf{p}, \mathbf{q} \in \mathcal{P}_r$ and $\alpha > 1$, we have

$$
\begin{aligned}
\mathbb{E}_{\mathbf{q}}[S_{\text{Spherical}}(\mathbf{p},\cdot)] &= \sum_{j\in\Omega} \frac{p_j^{\alpha-1}}{\left(\sum_{k\in\Omega} p_k^\alpha\right)^{(\alpha-1)/\alpha}} q_j\\
&= \left(\sum_{k\in\Omega} p_k^\alpha\right)^{-(\alpha-1)/\alpha}\left(\sum_{j\in\Omega} p_j^{\alpha-1} q_j\right) \qquad (3.5)
\end{aligned}
$$

and

$$\mathbb{E}_{\mathbf{q}}[S_{\text{Spherical}}(\mathbf{q},\cdot)] = \sum_{j\in\Omega} \frac{q_j^\alpha}{\left(\sum_{k\in\Omega} q_k^\alpha\right)^{(\alpha-1)/\alpha}} = \left(\sum_{j\in\Omega} q_j^\alpha\right)^{1/\alpha}. \qquad (3.6)$$

By using Hölder's inequality for $p = \alpha/(1-\alpha)$ and $q = \alpha$ on the second term of (3.5) and combine with (3.6) yield

$$
\begin{aligned}
\sum_{j\in\Omega} p_j^{\alpha-1} q_j &\leq \left(\sum_{j\in\Omega} p_j^\alpha\right)^{(\alpha-1)/\alpha}\left(\sum_{j\in\Omega} q_j^\alpha\right)^{1/\alpha}\\
&= \left(\sum_{j\in\Omega} p_j^\alpha\right)^{(\alpha-1)/\alpha} \mathbb{E}_{\mathbf{q}}[S_{\text{Spherical}}(\mathbf{q},\cdot)]. \qquad (3.7)
\end{aligned}
$$

Combining (3.5), (3.6) and (3.7) yield,

$$\mathbb{E}_{\mathbf{q}}[S_{\text{Spherical}}(\mathbf{p},\cdot)] \leq \mathbb{E}_{\mathbf{q}}[S_{\text{Spherical}}(\mathbf{q},\cdot)].$$

△

Example 3.1.4 (The logarithmic score)**.** Let $(\Omega, \mathcal{F})$ and $\mathcal{P}_r$ be measurable space and a convex class of probability measures as defined in Example 3.1.2. Then for $\mathbf{p} \in \mathcal{P}_r$ and $j \in \Omega$, the logarithm scoring rule is

$$S_{\log}(\mathbf{p}, j) = \log p_j. \tag{3.8}$$

This score is strictly proper since $\mathbb{E}_{\mathbf{q}}[S_{\log}(\mathbf{q},\cdot)] - \mathbb{E}_{\mathbf{q}}[S_{\log}(\mathbf{p},\cdot)]$ is related to the Kullback-Leibler divergence. Indeed,

$$\mathbb{E}_{\mathbf{q}}[S_{\log}(\mathbf{q},\cdot)] - \mathbb{E}_{\mathbf{q}}[S_{\log}(\mathbf{p},\cdot)] = \sum_{j\in\Omega}[\log(q_j) - \log(p_j)]q_j = \sum_{j\in\Omega}\log\left(\frac{q_j}{p_j}\right)q_j = -\sum_{j\in\Omega}\log\left(\frac{p_j}{q_j}\right)q_j. \tag{3.9}$$

By using the inequality $-\log(x) \geq -x + 1$ for any $x \in \mathbb{R}$, on (3.9), we have

$$\mathbb{E}_{\mathbf{q}}[S_{\log}(\mathbf{q},\cdot)] - \mathbb{E}_{\mathbf{q}}[S_{\log}(\mathbf{p},\cdot)] \geq \sum_{j\in\Omega}\left(-\frac{p_j}{q_j} + 1\right)q_j = \sum_{j\in\Omega} -p_j + q_j = 0. \tag{3.10}$$

The inequality in (3.10) is an equality if and only if $\mathbf{p} = \mathbf{q}$, which follows immediately from (3.9). Note also that if $S^*_{\log}(\mathbf{p}, j) := -S_{\log}(\mathbf{p}, j)$, then $S^*_{\log}(\mathbf{p}, j)$ is the negative orientation version of the logarithmic score and

$$\mathbb{E}_{\mathbf{q}}[S^*_{\log}(\mathbf{q}, j)] = -\sum_{j\in\Omega} q_j \log q_j,$$

which is the Shannon entropy. △

3.2. The (continuous) ranked probability score

In Section 3.1, we have seen several examples of (strictly) proper scoring rules. These scoring rules are defined on a finite sample space. Such scoring rules use the probability mass function and the observed event to give the forecaster a reward or a loss, depending on the orientation of the scoring rule. In this section, we discuss the development of a scoring rule that uses the distribution function directly instead of the probability mass function. This scoring rule is called the continuous ranked probability score (CRPS), which was introduced by Matheson and Winkler (1976). Before the CRPS, there was also a scoring rule called the ranked probability score (RPS), which was introduced by Epstein (1969). In this section, we will discuss both the RPS and the CRPS. Furthermore, even though Matheson and Winkler (1976) already stated how the RPS and the CRPS are related, we will prove the truthfulness of the statement mathematically as well.

3.2.1. The ranked probability score

The ranked probability score is constructed with the goal that it is sensitive to how 'far' lies a forecast from the observed event. Consider the following examples of forecasts for a hypothetical location, in which the wind speed is rarely very high and is often low. Suppose then the recorded wind speed can be partitioned and ranked from very high, high, moderate to low-speed. Now, a forecaster A and B forecast these events with the following probability respectively: $(0.1, 0.1, 0.2, 0.6)$ and $(0.2, 0.1, 0.1, 0.6)$. If a low wind speed is observed, then all scoring rules that we have discussed in the previous section will give the same score. However, forecaster B assigns a higher probability to the event as 'very high', and so one might argue that forecaster A has a better quality forecast. For this reason, a scoring rule that can distinguish such forecasts was developed by Epstein (1969).

The derivation of the ranked probability score uses the utility framework. Suppose the sample space $\Omega = \{1, 2\}$, where 1 is interpreted as bad weather that requires full protection and 2 is good weather which requires no protection. One can then compute the expected utility, by designing a utility matrix and a decision rule of when action to protect is taken. Instead of two possible outcomes, Epstein (1969) apply the utility framework on $\Omega = \{1, \ldots, r\}$ for $r \geq 2$. Suppose we again interpret the sample space as the outcome of the weather and we rank the elements of Ω from worst to the best weather, one then can compute the expected utility in this scenario. Unfortunately, the resulting expected utility will depend on the outcome of the weather and the ”goodness” of the prediction. We describe the computation and explanation of the utility in more detail in Appendix A. That is, if we predict the best weather perfectly, then its expected utility is higher than when the worst weather is perfectly predicted. The ranked probability score is defined by combining the resulting expected utility when we rank the weather from best to worst, and vice versa. Now, we state the definition of the ranked probability score.

Definition 3.2.1 (The ranked probability score)**.** *Let $(\Omega, \mathcal{F})$ be the probability space, where $\Omega = \{1, \dots, r\}$, such that the elements are ranked from ”good” to ”bad” or vice versa* [1]*. Let p_ω denote the probability of event ω occurs. Let $\mathcal{P}_r$ be a class of convex probability measures on $(\Omega, \mathcal{F})$, defined as in* (3.3)*. The ranked probability score (RPS) is a scoring rule defined as follows*

$$S_{\mathrm{RPS}}(\mathbf{p}, j) = \frac{3}{2} - \frac{1}{2(r-1)} \sum_{i=1}^{r-1} \left[\left(\sum_{k=1}^{i} p_k \right)^2 + \left(\sum_{k=i+1}^{r} p_k \right)^2 \right] - \frac{1}{r-1} \sum_{i=1}^{r} |i-j| p_i. \tag{3.11}$$

The ranked probability score gives the maximum reward for ’perfectly’ predicting an event j and it is independent of how the forecaster ranks the sample space. Indeed, let $\mathbf{p} = (p_1, \dots, p_r)$ such that $p_j = 1$ and $p_{j'} = 0$ for any $1 \le j \ne j' \le r$. Then for any $1 \le i \le r-1$, the summand of the second term in (3.11) is equal to 1. This is because if $1 \le j \le i$, then $\sum_{k=1}^{i} p_k = 1$ and $\sum_{k=i+1}^{r} p_k = 0$, and vice versa for $i \le j \le r$. Hence,

$$S_{\mathrm{RPS}}(\mathbf{p}, j) = \frac{3}{2} - \frac{1}{2(r-1)} \sum_{i=1}^{r-1} 1 = \frac{3}{2} - \frac{1}{2} = 1.$$

Now suppose we predict with 100% confidence that weather j' occurs, but the observed outcome is $j \ne j'$. In case we set $\mathbf{p} = \mathbf{p}_1 = (1, 0, \dots, 0)$ and $j > 1$, then

$$S_{\mathrm{RPS}}(\mathbf{p}_1, j) = \frac{3}{2} - \frac{1}{2(r-1)} \sum_{i=1}^{r-1} 1 - \frac{1}{r-1}(j-1) = \frac{3}{2} - \frac{1}{2} - \frac{j-1}{r-1} = \frac{r-j}{r-1}.$$

If we let $\mathbf{p}_2 = (0, 0, \dots, 1)$ but $j < r$, then

$$S_{\mathrm{RPS}}(\mathbf{p}_2, j) = \frac{3}{2} - \frac{1}{2(r-1)} \sum_{i=1}^{r-1} 1 - \frac{1}{r-1}(r-j) = \frac{3}{2} - \frac{1}{2} - \frac{j-1}{r-1} = \frac{j-1}{r-1}.$$

The score $S_{\mathrm{RPS}}(\mathbf{p}_1, j)$ is decreasing linearly with j, as opposed to linearly increasing $S_{\mathrm{RPS}}(\mathbf{p}_2, j)$ in j. These scores are intercepting at $j = (r-1)/2$, which is the weather in the middle range.

The following result states that the ranked probability score is strictly proper, which is already proven in Murphy (1966) and von Holstein (1970).

Theorem 3.2.2. *The ranked probability score is strictly proper relative to $\mathcal{P}_r$.*

Proof. Let X be a random variable that takes values in $\{1, \dots, r\}$ and following a distribution with probability mass vector $\mathbf{q} = (q_1, \dots, q_r) \in \mathcal{P}_r$. Before we compute $\mathbb{E}_{\mathbf{q}} \left[S_{\mathrm{RPS}}(\mathbf{p}, X) \right]$, let us rewrite the second term in $S_{\mathrm{RPS}}(\mathbf{p}, j)$. We prove

$$\sum_{i=1}^{r-1} \left[\left(\sum_{k=1}^{i} p_k \right)^2 + \left(\sum_{k=i+1}^{r} p_k \right)^2 \right] = \sum_{k=1}^{r} \sum_{l=1}^{r} [(r-1) - |k-l|] p_k p_l. \tag{3.12}$$

[1] The term ”good” and ”bad” depends on the context of the problem. If we consider temperature, we can discretize and assign an interpretation for each interval. The sample space Ω in this case would be sorted from hot to cold or vice versa

To this end put each summand in a $(r-1)\times 2$ array, where the first column represents $\left(\sum_{k=1}^{i} p_k\right)^2$ and the second column is $\left(\sum_{k=i+1}^{r} p_k\right)^2$:

$$\begin{array}{lr}
p_1^2 & (p_2+p_3+\cdots+p_{r-1}+p_r)^2 \\
(p_1+p_2)^2 & (p_3+\cdots+p_{r-1}+p_r)^2 \\
\vdots & \vdots \\
(p_1+p_2+p_3+\cdots+p_{r-2})^2 & (p_{r-1}+p_r)^2 \\
(p_1+p_2+p_3+\cdots+p_{r-2}+p_{r-1})^2 & p_r^2
\end{array}$$

If we expand the polynomials, we notice that $p_1^2, p_2^2, \dots, p_r^2$ appears in each $r-1$ rows . As for the cross terms, we let $1 \le n \ne m \le r$. If $n < m$, then $p_n p_m$ does not appear in the rows which the entries are $\left(\sum_{k=1}^{n} p_k\right)^2$ and $\left(\sum_{k=m'}^{r} p_k\right)^2$, where $n < m' \le m$. There are $n-m$ such rows. Similarly, if $n > m$, then $p_n p_m$ will not appear in $m-n$ rows. Hence, for any $1 \le n \ne m \le r$, there are $[(r-1)-|n-m|]$ terms of $p_n p_m$. Hence (3.12) is true.

Now, we let $\mathbf{p} = (p_1, \dots, p_r) \in \mathcal{P}_r$, then

$$\begin{aligned}
\mathbb{E}_{\mathbf{q}}\left[S_{\mathrm{RPS}}(\mathbf{p}, X)\right] &= \sum_{j=1}^{r} S_{\mathrm{RPS}}(\mathbf{p}, j) q_j \\
&= \sum_{j=1}^{r} \left\{ \frac{3}{2} - \frac{1}{2(r-1)} \sum_{i=1}^{r-1} \left[\left(\sum_{k=1}^{i} p_k \right)^2 + \left(\sum_{k=i+1}^{r} p_k \right)^2 \right] - \frac{1}{r-1} \sum_{i=1}^{r} |i-j| p_i \right\} q_j \\
&= \frac{3}{2} - \frac{1}{2(r-1)} \sum_{i=1}^{r-1} \left[\left(\sum_{k=1}^{i} p_k \right)^2 + \left(\sum_{k=i+1}^{r} p_k \right)^2 \right] - \frac{1}{r-1} \sum_{j=1}^{r} q_j \left\{ \sum_{i=1}^{r} |i-j| p_i \right\} \\
&= \frac{3}{2} - \frac{1}{2(r-1)} \sum_{k=1}^{r} \sum_{l=1}^{r} [(r-1) - |k-l|] p_k p_l - \frac{1}{r-1} \sum_{j=1}^{r} q_j \left\{ \sum_{i=1}^{r} |i-j| p_i \right\}.
\end{aligned}$$

We would like to find $\mathbf{p}$ such that $\mathbf{p}$ maximizes $\mathbb{E}_{\mathbf{q}}\left[S_{\mathrm{RPS}}(\mathbf{p}, X)\right]$, subject to $\sum_{j=1}^{r} p_j = 1$. To this end, we use the Lagrange multiplier, and therefore we solve

$$\breve{\mathbf{p}} = \operatorname*{arg\,max}_{\mathbf{p} \in \mathcal{P}_r} \left\{ \mathbb{E}_{\mathbf{q}}\left[S_{\mathrm{RPS}}(\mathbf{p}, X)\right] + \lambda \left(\sum_{j=1}^{r} p_j - 1 \right) \right\} =: \operatorname*{arg\,max}_{\mathbf{p} \in \mathcal{P}_r} f(\mathbf{p}, \lambda).$$

Then, if we differentiate f w.r.t. some entry in p_m in $\mathbf{p}$, we obtain

$$\begin{aligned}
\frac{\partial f}{\partial p_m}(\mathbf{p}, \lambda) &= -\frac{1}{r-1} \sum_{i=1}^{r} [(r-1) - |m-i|] p_i - \frac{1}{r-1} \sum_{j=1}^{r} |m-j| q_j + \lambda, \\
\frac{\partial^2 f}{\partial p_m^2}(\mathbf{p}, \lambda) &= -1.
\end{aligned}$$

The second derivative test then guarantees that the maximum exists. To find the stationary point, we use the constraint that the p_i's sum up to 1. So we then have

$$\begin{aligned}
\frac{\partial f}{\partial p_m}(\mathbf{p}, \lambda) &= -1 + \frac{1}{r-1} \sum_{i=1}^{r} |m-i| p_i - \frac{1}{r-1} \sum_{j=1}^{r} |m-j| q_j + \lambda \\
&= \lambda - 1 + \frac{1}{r-1} \sum_{j=1}^{r} |m-j| (p_j - q_j).
\end{aligned}$$

Which means

$$\frac{\partial f}{\partial p_m}(\mathbf{p},\lambda) = 0 \Longrightarrow \frac{1}{r-1}\sum_{j=1}^{r} |m-j|(p_j - q_j) = (r-1)(1-\lambda).$$

The trick that is used in Murphy (1966), is to subtract the above equation with another adjacent equation to eliminate $(r-1)(1-\lambda)$. So then for $1 \leq m \leq r-1$, we have

$$\begin{aligned}
&\sum_{j=1}^{r} |m-j|(p_j-q_j) - \sum_{j=1}^{r} |m+1-j|(p_j-q_j) = 0 \\
\Leftrightarrow\ &\sum_{j=m+1}^{r} (p_j-q_j) - \sum_{j=1}^{m}(p_j-q_j) = 0 \\
\Leftrightarrow\ &\sum_{j=1}^{m}(p_j-q_j) = \sum_{j=1}^{r}(p_j-q_j) - \sum_{j=1}^{m}(p_j-q_j) \\
\Leftrightarrow\ &2\sum_{j=1}^{m}(p_j-q_j) = 0, \qquad (3.13)
\end{aligned}$$

where we again use the constraint. Equation (3.13) is true if and only if $p_j = q_j$ for any $1 \leq j \leq m$ and $1 \leq m \leq r-1$. Hence, $\breve{\mathbf{p}} = \mathbf{q}$. This maximizer is unique and so the scoring rule is indeed strictly proper. □

Lastly, we would like to show another way to formulate the RPS. This will be needed when we link the RPS with the CRPS. It turns out that the RPS is an affine transformation of a certain function that depends on the forecast probabilities.

Lemma 3.2.3. *Let* $\mathbf{p} = (p_1, \dots, p_r) \in \mathcal{P}_r$ *and* $j \in \Omega = \{1, \dots, r\}$. *Let*

$$G_i = \sum_{k=1}^{i} p_i \qquad \textit{and therefore} \qquad 1 - G_i = \sum_{k=i+1}^{r} p_i. \qquad (3.14)$$

Then,

$$S_{\mathrm{RPS}}(\mathbf{p}, j) = 1 + \frac{1}{r-1}\left(-\sum_{i=1}^{r-1} G_i^2 + 2\sum_{i=j}^{r-1} G_i - (r-j)\right).$$

Proof. From the definition the RPS, and using (3.14) yields

$$\begin{aligned}
S_{\mathrm{RPS}}(\mathbf{p}, j) &= \frac{3}{2} - \frac{1}{2(r-1)}\sum_{i=1}^{r-1}\left[G_i^2 + (1-G_i)^2\right] - \frac{1}{r-1}\sum_{i=1}^{r}|i-j|p_i \\
&= \frac{3}{2} - \frac{1}{2(r-1)}\sum_{i=1}^{r-1}\left[2G_i^2 + 1 - 2G_i\right] - \frac{1}{r-1}\sum_{i=1}^{r}|i-j|p_i \\
&= \frac{3}{2} - \frac{1}{r-1}\sum_{i=1}^{r-1}\left[G_i^2 - G_i\right] - \frac{1}{2} - \frac{1}{r-1}\sum_{i=1}^{r}|i-j|p_i \\
&= 1 - \frac{1}{r-1}\sum_{i=1}^{r-1}\left[G_i^2 - G_i\right] - \frac{1}{r-1}\sum_{i=1}^{r}|i-j|p_i. \qquad (3.15)
\end{aligned}$$

Now, note that

$$\sum_{i=1}^{r}|i-j|p_i = \sum_{i=1}^{j-1}(j-i)p_i + \sum_{i=j+1}^{r}(i-j)p_i$$

$$
\begin{aligned}
&= \underbrace{(p_1 + p_1 + \cdots + p_1)}_{(j-1)\text{ terms}} + \underbrace{(p_2 + \cdots + p_2)}_{(j-2)\text{ terms}} + \cdots + \underbrace{(p_{j-1} + p_{j-1})}_{2\text{ terms}} + \underbrace{p_{j-1}}_{1\text{ term}} \\
&\quad + \underbrace{p_{j+1}}_{1\text{ term}} + \underbrace{(p_{j+2} + p_{j+2})}_{2\text{ terms}} + \cdots + \underbrace{(p_r + p_r + \cdots + p_r)}_{(r-j)\text{ terms}}
\end{aligned}
$$

$$
\begin{aligned}
&= \underbrace{p_1 + (p_1 + p_2) + \cdots (p_1 + \cdots + p_{j-1})}_{(j-1)\text{ terms}} \\
&\quad + \underbrace{(p_{j+1} + p_{j+2} + \cdots p_r) + (p_{j+2} + \cdots + p_r) + \cdots + p_r}_{(r-j)\text{ terms}}
\end{aligned}
\tag{3.16}
$$

$$
\begin{aligned}
&= \sum_{i=1}^{j-1} \sum_{k=1}^{i} kp_k + \sum_{i=j}^{r-1} \sum_{k=i+1}^{r} p_k \\
&= \sum_{i=1}^{j-1} G_i + \sum_{i=j}^{r-1} (1 - G_i) \\
&= \sum_{i=1}^{j-1} G_i + (r - j) - \sum_{i=j}^{r-1} G_i.
\end{aligned}
\tag{3.17}
$$

We combine (3.15) and (3.17) to get

$$
\begin{aligned}
S_{\text{RPS}}(\mathbf{p}, j) &= 1 - \frac{1}{r-1} \sum_{i=1}^{r-1} \left[G_i^2 - G_i \right] - \frac{1}{r-1} \left[\sum_{i=1}^{j-1} G_i - \sum_{i=j}^{r-1} G_i + (r - j) \right] \\
&= 1 + \frac{1}{r-1} \left(- \sum_{i=1}^{r-1} G_i^2 + \sum_{i=1}^{r-1} G_i - \sum_{i=1}^{j-1} G_i + \sum_{i=j}^{r-1} G_i - (r - j) \right) \\
&= 1 + \frac{1}{r-1} \left(- \sum_{i=1}^{r-1} G_i^2 + 2 \sum_{i=j}^{r-1} G_i - (r - j) \right)
\end{aligned}
$$

□

3.2.2. The continuous ranked probability score

In the previous section, we showed examples of (strictly) proper scoring rules defined on finite sample space. The convex class of probability measures in these examples consists of probability mass functions. We will discuss a scoring rule that takes directly the distribution function as input, together with an observed real-valued number. It turns out that this scoring rule is a generalization of the Brier score, and this scoring rule is referred to as the continuous ranked probability score (CRPS).

The connection between the CRPS and the Brier score becomes clear if we discuss how the CRPS is constructed. The CRPS is introduced by Matheson and Winkler (1976), and the motivation behind the score is to assess the probabilistic forecasts using the continuous probability distributions directly. Consider first the Brier score in Example 3.1.2, with the class $\mathcal{P}_2$ and the sample space $\{0, 1\}$. Then, for $p \in \mathcal{P}_2$, the Brier score may also be written as

$$
S_{\text{Brier}}(p, j) = -p^2 \mathbb{1}(j \neq \omega) - (1 - p)^2 \mathbb{1}(j = \omega).
$$

If the event ω occurs and the forecaster predicts that event with probability mass p, then the reward for the forecaster is $-(1 - p)^2$, the reward is $-p^2$ otherwise. Now, when the sample space is $\mathbb{R}$, then we usually compute the probability using the distribution function F. Matheson and Winkler (1976) generalizes the Brier score by replacing p with F. Further, instead of checking whether a value is observed, we check if the observed value is at least a fixed real number. That is for $x, t \in \mathbb{R}$, we define the score as follows

$$
S(F, x) = -F(t)^2 \mathbb{1}(t < x) - (1 - F(t))^2 \mathbb{1}(t \geq x).
\tag{3.18}
$$

The score in (3.18) depends on the threshold t. To remove such dependencies, all possible threshold value are integrated, i.e. we integrate (3.18) w.r.t. t on $\mathbb{R}$. This yields

$$\begin{aligned} S^*(F,x) &= -\int_{\mathbb{R}} F^2(t)\mathbb{1}(t < x)\,dt - \int_{\mathbb{R}} (1 - F(t))^2\mathbb{1}(t \geq x)\,dt \\ &= -\int_{\mathbb{R}} F^2(t)(1 - \mathbb{1}(t \geq x))\,dt - \int_{\mathbb{R}} (1 - F(t))^2\mathbb{1}(t \geq x)\,dt \\ &= -\int_{\mathbb{R}} F^2(t) - F^2(t)\mathbb{1}(t \geq x) + \mathbb{1}(t \geq x) - 2F(t)\mathbb{1}(t \geq x) + F^2(t)\mathbb{1}(t \geq x)\,dt \\ &= -\int_{\mathbb{R}} F^2(t) - 2F(t)\mathbb{1}(t \geq x) + \mathbb{1}(t \geq x)\,dt \\ &= -\int_{\mathbb{R}} [F(t) - \mathbb{1}(t \geq x)]^2\,dt. \end{aligned} \tag{3.19}$$

The score in (3.19) is called the *continuous ranked probability score* (CRPS). We formally define the score in the following.

Definition 3.2.4. *Let $\mathcal{B}(\mathbb{R})$ be the Borel set of $\mathbb{R}$ and $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is the measurable space. Let $\mathcal{P}$ be the convex class of probability measures $\mathbb{P}$ such that they are defined on $\mathcal{B}(\mathbb{R})$. Let F be the cumulative distribution function, which identifies the probability measure $\mathbb{P} \in \mathcal{P}$. The CRPS is defined as*

$$\mathrm{CRPS}(F,x) = -\int_{\mathbb{R}} [F(t) - \mathbb{1}(t \geq x)]^2\,dt. \tag{3.20}$$

There is an equivalent formulation of the CRPS, which is used to test the hypothesis that two distributions are equal. We will use this formulation to show the CRPS is strictly proper. To achieve this, we need the following lemma from Baringhaus and Franz (2004, Lemma 2.1).

Lemma 3.2.5. *Let X and Y be two independent real-valued random variables, such that their first moments are finite. Let F and G be the distribution function of X and Y respectively. Then*

$$\mathbb{E}[|X - Y|] = \int_{\mathbb{R}} F(t)(1 - G(t))\,dt + \int_{\mathbb{R}} (1 - F(t))G(t)\,dt.$$

Proof. Note that for any $x, y \in \mathbb{R}$, then

$$|x - y| = \int_{\mathbb{R}} \mathbb{1}(x \leq t < y) + \mathbb{1}(y \leq t < x)\,dt. \tag{3.21}$$

Indeed, let λ be the Lebesgue measure. If $x < y$, then $\mathbb{1}(y \leq t < x) = 0$, and so

$$|x - y| = \int_{\mathbb{R}} \mathbb{1}(x \leq t < y)\,dt = \lambda([x, y)) = x - y.$$

If $y > x$, then $\mathbb{1}(x \leq t < y) = 0$, which means that

$$|x - y| = \int_{\mathbb{R}} \mathbb{1}(y \leq t < x)\,dt = \lambda([y, x)) = y - x.$$

In case that $x = y$, then $\mathbb{1}(x \leq t < y) = \mathbb{1}(y \leq t < x) = 0$. So the integral representation in (3.21) is true. Therefore,

$$\mathbb{E}[|X - Y|] = \mathbb{E}\left[\int_{\mathbb{R}} \mathbb{1}(X \leq t < Y) + \mathbb{1}(Y \leq t < X)\,dt\right].$$

Since the indicator function is a measurable function, by Fubini-Tonelli's theorem, we have

$$\begin{aligned} \mathbb{E}[|X - Y|] &= \mathbb{E}\left[\int_{\mathbb{R}} \mathbb{1}(X \leq t < Y) + \mathbb{1}(Y \leq t < X)\,dt\right] \\ &= \int_{\mathbb{R}} \mathbb{E}[\mathbb{1}(X \leq t < Y)] + \mathbb{E}[\mathbb{1}(Y \leq t < X)]\,dt \end{aligned}$$

$$\begin{aligned}
&= \int_{\mathbb{R}} \mathbb{P}(X \leq t < Y) + \mathbb{P}(Y \leq t < X)\, dt \\
&= \int_{\mathbb{R}} \mathbb{P}(X \leq t)\mathbb{P}(Y > t) + \mathbb{P}(Y \leq t)\mathbb{P}(X > t)\, dt \\
&= \int_{\mathbb{R}} F(t)(1 - G(t))\, dt + \int_{\mathbb{R}} (1 - F(t))G(t)\, dt.
\end{aligned}$$

□

The following statement shows that a new formulation of the CRPS is equivalent to the one in Definition 3.2.4. It also shows that the score has the same unit as the observations. This result is stated in Gneiting and Raftery (2007) and we check that it is indeed correct.

Lemma 3.2.6. *Let $\mathcal{P}$ be a convex class of probability measures on a measurable space $(\Omega, \mathcal{F})$. Let* $\mathrm{CRPS} : \mathcal{P} \times \Omega \to \overline{\mathbb{R}}$ *be a scoring rule as defined in* (3.20)*. Let X be a random variable with distribution function F, which uniquely identifies the probability measure $\mathbb{P} \in \mathcal{P}$. Suppose X' is an independent copy of X. Then, for any $x \in \Omega$,*

$$\mathrm{CRPS}(F, x) = \frac{1}{2}\mathbb{E}[|X - X'|] - \mathbb{E}[|X - x|].$$

Proof. By Lemma 3.2.5, we have

$$\mathbb{E}[|X - X'|] = 2\int_{\mathbb{R}} F(t)(1 - F(t))\, dt = 2\int_{\mathbb{R}} F(t) - F^2(t)\, dt,$$

and by (3.21) and Fubini-Tonelli's theorem,

$$\begin{aligned}
\mathbb{E}[|X - x|] &= \int_{\mathbb{R}} F(t)\mathbb{1}(t < x)\, dt + \int_{\mathbb{R}} (1 - F(t))\mathbb{1}(t \geq x)\, dt \\
&= \int_{\mathbb{R}} F(t)(1 - \mathbb{1}(t \geq x))\, dt + \int_{\mathbb{R}} (1 - F(t))\mathbb{1}(t \geq x)\, dt.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\frac{1}{2}\mathbb{E}[|X - X'|] - \mathbb{E}[|X - x|] &= \int_{\mathbb{R}} F(t) - F^2(t) - F(t)(1 - \mathbb{1}(t \geq x)) - (1 - F(t))\mathbb{1}(t \geq x)\, dt \\
&= \int_{\mathbb{R}} -F^2(t) + 2F(t)\mathbb{1}(t \geq x) - \mathbb{1}(t \geq x)\, dt \\
&= -\int_{\mathbb{R}} (F(t) - \mathbb{1}(t \geq x))^2\, dt,
\end{aligned}$$

which is the definition of CRPS. □

We use Lemma 3.2.5 and Lemma 3.2.6 to show that CRPS is strictly proper. Although we use the lemmas here, one can also show this property by using the definition of CRPS directly.

Theorem 3.2.7. *The CRPS is a proper scoring rule relative to $\mathcal{P}$, the convex class of Borel probability measures. It is strictly proper relative to a class of Borel probability measures with finite first moment.*

Proof. Let $\mathbb{P}, \mathbb{Q} \in \mathcal{P}$, where $\mathbb{P}$ and $\mathbb{Q}$ are identified by the distribution function F and G respectively. Let X, X', Y, Y' be mutually independent random variables. Let F (resp. G) be the distribution function of X, X' (resp. Y, Y'). For any $x \in \mathbb{R}$,

$$\begin{aligned}
\mathrm{CRPS}(G, x) - \mathrm{CRPS}(F, x) &= \frac{1}{2}\mathbb{E}[|Y - Y'|] - \frac{1}{2}\mathbb{E}[|X - X'|] - \mathbb{E}[|Y - x|] + \mathbb{E}[|X - x|] \\
&= \int_{\mathbb{R}} \Big\{ G(t) - G^2(t) - F(t) + F^2(t) - G(t)(1 - \mathbb{1}(t \geq x)) \\
&\qquad - (1 - G(t))\mathbb{1}(t \geq x) + F(t)(1 - \mathbb{1}(t \geq x)) + (1 - F(t))\mathbb{1}(t \geq x) \Big\}\, dt
\end{aligned}$$

$$= \int_{\mathbb{R}} F^2(t) - G^2(t) + 2(G(t) - F(t))\mathbb{1}(t \geq x)\, dt.$$

By using Fubini-Tonelli, we obtain

$$\begin{aligned}
\mathbb{E}_G[\mathrm{CRPS}(G,x)] - \mathbb{E}_G[\mathrm{CRPS}(F,x)] &= \int_{\mathbb{R}} \left\{ \int_{\mathbb{R}} F^2(t) - G^2(t) + 2(G(t) - F(t))\mathbb{1}(t \geq x)\, dt \right\} dG(x) \\
&= \int_{\mathbb{R}} \left\{ \int_{\mathbb{R}} F^2(t) - G^2(t) + 2(G(t) - F(t))\mathbb{1}(t \geq x)\, dG(x) \right\} dt \\
&= \int_{\mathbb{R}} F^2(t) - G^2(t) + 2(G(t) - F(t))\left\{ \int_{\mathbb{R}} \mathbb{1}(t \geq x) dG(x) \right\} dt \\
&= \int_{\mathbb{R}} F^2(t) - G^2(t) + 2(G(t) - F(t))G(t)\, dt \\
&= \int_{\mathbb{R}} (G(t) - F(t))^2\, dt \geq 0.
\end{aligned}$$

We can use the Fubini-Tonelli theorem because the integrand is a measurable function. Hence, the CRPS scoring rule is indeed a proper scoring rule. Suppose F and G have finite first moment. If $F = G$, then it is clear that

$$\mathbb{E}_G[\mathrm{CRPS}(G,x)] - \mathbb{E}_G[\mathrm{CRPS}(F,x)] = \int_{\mathbb{R}} (F(t) - F(t))^2\, dt = 0.$$

If $F \neq G$, then

$$\mathbb{E}_G[\mathrm{CRPS}(G,x)] - \mathbb{E}_G[\mathrm{CRPS}(F,x)] = \int_{\mathbb{R}} (F(t) - G(t))^2\, dt > 0.$$

□

We have confirmed that CRPS is strictly proper, yet we need to explain the relationship between this score and the ranked probability score (see Section 3.2.1). According to Matheson and Winkler (1976), these two are equivalent in a sense that both are affine transformations of a certain function. To see this, Matheson and Winkler (1976) starts by generalizing CRPS by assuming that the integrated variable t is random and it follows a distribution $R(t)$. Therefore, we get a new scoring rule, which we denote as S^{**}. From the computations to get (3.19) yield

$$S^{**}(G,x) = -\int_{-\infty}^{x} G^2(t)\, dR(t) - \int_{x}^{\infty} (1 - G(t))^2\, dR(t). \tag{3.22}$$

If $R(t)$ is a step function, i.e. for $t_1 < \cdots < t_r$,

$$R(u) = \sum_{k=1}^{r-1} r_k \mathbb{1}(t_k \leq t < t_{k+1}),$$

then if $x = t_j$, and $G(r_k) := G_k$, we have

$$S^{**}(G,t_j) = -\sum_{k=1}^{j-1} G_k^2 r_k - \sum_{k=j}^{r-1} (1 - G_k)^2 r_k. \tag{3.23}$$

In case that $r_k = 1/r$ for any $k = 1, \dots, r-1$, we have

$$\begin{aligned}
S^{**}(G,t_j) &= -\frac{1}{r}\left(\sum_{k=1}^{j-1} G_k^2 + \sum_{k=j}^{r-1} (1 - G_k)^2 \right) \\
&= -\frac{1}{r}\left(\sum_{k=1}^{j-1} G_k^2 + \sum_{k=j}^{r-1} (1 - 2G_k + G_k^2) \right)
\end{aligned}$$

$$= -\frac{1}{r}\left(\sum_{k=1}^{j-1} G_k^2 + (r-j) - 2\sum_{k=j}^{r-1} G_k + \sum_{k=j}^{r-1} G_k^2\right). \tag{3.24}$$

Since

$$\sum_{k=j}^{r-1} G_k^2 = \sum_{k=1}^{r-1} G_k^2 - \sum_{k=1}^{j-1} G_k^2, \tag{3.25}$$

both (3.24) and (3.25) yield

$$\begin{aligned} S^{**}(G, t_j) &= -\frac{1}{r}\left(\sum_{k=1}^{j-1} G_k^2 + (r-j) - 2\sum_{k=j}^{r-1} G_k + \sum_{k=1}^{r-1} G_k^2 - \sum_{k=1}^{j-1} G_k^2\right) \\ &= \frac{1}{r}\left(-\sum_{k=1}^{r-1} G_k^2 + 2\sum_{k=j}^{r-1} G_k - (r-j)\right). \end{aligned} \tag{3.26}$$

We compare the result from Lemma 3.2.3 with (3.26). While the RPS is an affine transformation of $-\sum_{k=1}^{r-1} G_k^2 + 2\sum_{k=j}^{r-1} G_k - (r-j)$, the CRPS is a linear transformation of it. In any case, both the of these score are affine transformations of $-\sum_{k=1}^{r-1} G_k^2 + 2\sum_{k=j}^{r-1} G_k - (r-j)$.

3.3. The connection of scoring rules with information theory

So far, we use scoring rules to measure the quality of the forecasts. However, if S is a negatively oriented and a proper scoring rule, then we can also view scoring rules as a loss function. The expectation $\mathbb{E}_{\mathbb{Q}}[S(\mathbb{P}, \cdot)]$ is then the expected loss of using $\mathbb{P}$ to predict events, while the true distribution of the events is $\mathbb{Q}$. By (3.2), we then want the expected loss to be minimized, where $\mathbb{Q}$ is a minimizer of the expected loss. This minimizer is unique if S is strictly proper. If we choose a particular S, we can in fact obtain entropy and Kullback-Leibler divergence back. In this section, we discuss the connection the terminologies used in the information theory and review a generalization of these notions.

Although we have mentioned the entropy and the Kullback-Leibler divergence in the previous section, let us recall and define them more properly in this section.

Definition 3.3.1 (The Shannon entropy). *Let $(\Omega, \mathcal{F}, \mathbb{Q})$ be a probability space, such that Ω is finite. Let X be a random variable with probability mass function $q_x := \mathbb{Q}(X = x)$, where $x \in \Omega$. The entropy or the Shannon entropy is defined as*

$$H(\mathbb{Q}) := \mathbb{E}_q[-\log q(X)] = -\sum_{x\in\Omega} q_x \log(q_x). \tag{3.27}$$

The Shannon entropy indicates the average information that is contained in X (Yeung, 2002). Equivalently, it measures the uncertainty of the outcome of X. The larger the entropy of a random variable, the more uncertain one can predict an outcome of X.

If one wish to measure how ”far” the two measures p and q are apart, we can use a well-known measure called the Kullback-Leibler divergence. This measure is non-negative and equal to zero if and only if $p = q$. This measure is not a distance measure in a metric sense because it is not symmetric. Below is a definition of the Kullback-Leibler divergence.

Definition 3.3.2 (The Kullback-Leibler divergence). *Let Ω be a finite sample space. Let P and Q be probability distribution on Ω, with probability mass functions p_x and q_x respectively. The Kullback-Leibler divergence is defined as follows:*

$$D_{\mathrm{KL}}(P||Q) = \sum_{x\in\Omega} p_x \log\frac{p_x}{q_x} = -\sum_{x\in\Omega} p_x \log\frac{q_x}{p_x}.$$

Now let us rewrite the Shannon entropy. Let Ω be a finite sample space and $\mathcal{P}$ be a family distribution over Ω. Suppose a random variable X follows an unknown distribution $P \in \mathcal{P}$. Let $\mathcal{G}$ be a family of probability mass function of q. The entropy can also be written as

$$H(P) = \inf_{q \in \mathcal{G}} \mathbb{E}_P[-\log q(X)], \tag{3.28}$$

which is a consequence of the fact that $D_{\mathrm{KL}}(P||Q) \geq 0$ and equality is obtained if and only if $P = Q$. This statement is called the information inequality in Theorem 2.6.3 from Cover and Thomas (2005). To see that (3.28) is true, take any $q \in \mathcal{G}$. Then,

$$\begin{aligned} \mathbb{E}_P[-\log q(X)] &= -\sum_{x \in \Omega} p_x \log q_x \\ &= -\sum_{x \in \Omega} p_x \log\left(\frac{q_x}{p_x} p_x\right) \\ &= -\sum_{x \in \Omega} \left[p_x \log\left(\frac{q_x}{p_x}\right) + p_x \log p_x\right] \\ &= D_{\mathrm{KL}}(P||Q) + H(P) \geq H(P), \end{aligned}$$

where we use that $D_{\mathrm{KL}}(P||Q) \geq 0$ and $H(P) \geq 0$ since the probability mass function is always non-negative and $p_x \in [0, 1]$ for any $x \in \Omega$. Then the equality is obtained if and only if $p = q$.

In (3.28), we observe that the entropy is the smallest possible ”value” under the logarithmic loss. Grünwald and Dawid (2004) generalizes this definition by considering an arbitrary loss function $L : \mathcal{P} \times \Omega \to \mathbb{R}$. The generalized entropy function is then the following:

Definition 3.3.3. *Let $(\Omega, \mathcal{F})$ be a measurable space and $\mathcal{P}$ be a class of probability measures on $(\Omega, \mathcal{F})$. The generalized entropy function associated with the loss function $L : \mathcal{P} \times \Omega \to \overline{\mathbb{R}}$ is defined by*

$$H(P) := \inf_{Q \in \mathcal{P}} \mathbb{E}_P[L(Q, \cdot)]. \tag{3.29}$$

Now, we relate (3.29) with the scoring rule. If the scoring rule S is interpreted as a loss function (i.e. the scoring rule is negatively oriented), then the definition remains the same. If the scoring rule is seen as a reward system, then (3.29) becomes

$$H(P) := \sup_{Q \in \mathcal{P}} \mathbb{E}_P[S(Q, \cdot)], \tag{3.30}$$

where S is a proper scoring rule. Indeed, consider the logarithmic scoring rule which we orient positively. Then,

$$\mathbb{E}_P[\log q(X)] = H(P) - D_{KL}(P||Q) \leq H(P),$$

because $H(P) \geq 0$ and $-D_{KL}(P||Q) \leq 0$. To ensure that there exists $Q \in \mathcal{P}$ such that $\mathbb{E}_P[S(Q, \cdot)]$ is maximum, we need S to be proper.

Grünwald and Dawid (2004) also defines the divergence between two distributions.

Definition 3.3.4. *For any P, Q in class $\mathcal{P}$ and proper scoring rule S, the divergence function of P and Q is*

$$d(P, Q) := H(Q) - \mathbb{E}_Q[S(P, \cdot)] = \mathbb{E}_Q[S(Q, \cdot)] - \mathbb{E}_Q[S(P, \cdot)].$$

The divergence function $d(P, Q)$ is non-negative because the distribution Q maximizes the expected reward, and it is positive if and only if S is strictly proper. Additionally, $d(P, Q)$ is not necessarily equal to $d(Q, P)$. Indeed, if S is the logarithmic scoring rule, then the divergence function is the Kullback-Leibler divergence, which is known to be asymmetric (Sason, 2022). As a remark, if S is negatively oriented, the divergence function is

$$d(P, Q) = \mathbb{E}_Q[S(P, \cdot)] - H(Q),$$

which is still non-negative.

Let us compute the generalized entropy function and the divergence function for the Brier, pseudo-spherical, logarithmic score and CRPS.

Example 3.3.5. For the Brier score,

$$\begin{aligned} H(\mathbf{q}) = \mathbb{E}_{\mathbf{q}}[S_{\text{Brier}}(\mathbf{q},\cdot)] &= \sum_{j\in\Omega} S_{\text{Brier}}(\mathbf{q},j)q_j \\ &= \sum_{j\in\Omega}(2q_j - 1 - \sum_{\omega\in\Omega} q_\omega^2)q_j \\ &= 2\sum_{j\in\Omega} q_j^2 - 1 - \sum_{\omega\in\Omega} q_\omega^2 \\ &= \sum_{j\in\Omega} q_j^2 - 1. \end{aligned}$$

The divergence function is already computed, which is

$$d(\mathbf{q},\mathbf{p}) = \sum_{j\in\Omega}(q_j - p_j)^2,$$

the squared Euclidian norm (see Example 3.1.2). △

Example 3.3.6. We use the results from Example 3.1.3 to compute $H(\mathbf{q})$ and $d(\mathbf{p},\mathbf{q})$ of the pseudo-spherical score. The generalized entropy function is

$$H(\mathbf{q}) = \left(\sum_{j\in\Omega} q_j^\alpha\right)^{1/\alpha},$$

and the divergence function is

$$d(\mathbf{p},\mathbf{q}) = \left(\sum_{j\in\Omega} q_j^\alpha\right)^{1/\alpha} - \left(\sum_{j\in\Omega} p_j^{\alpha-1} q_j\right)\left(\sum_{k\in\Omega} p_k^\alpha\right)^{-(\alpha-1)/\alpha}.$$

△

Example 3.3.7. For the positively oriented logarithmic score,

$$H(\mathbf{q}) = \sum_{j\in\Omega} q_j \log(q_j),$$

which is the negative (Shannon) entropy. The divergence function is the Kullback-Leibler divergence as we have shown previously in Example 3.1.4:

$$d(\mathbf{p},\mathbf{q}) = \sum_{j\in\Omega} \log\left(\frac{q_j}{p_j}\right) q_j = -\sum_{j\in\Omega} \log\left(\frac{p_j}{q_j}\right) q_j.$$

△

Example 3.3.8. For the CRPS, the generalized entropy function is

$$\begin{aligned} H(G) &= \int_{\mathbb{R}}\int_{\mathbb{R}} (G(t) - \mathbb{1}(t \geq x))^2\, dt\, dG(x) \\ &= \int_{\mathbb{R}}\int_{\mathbb{R}} G^2(t) - 2G(t)\mathbb{1}(t \geq x) + \mathbb{1}(t \geq x)\, dt\, dG(x) \\ &= \int_{\mathbb{R}}\int_{\mathbb{R}} G^2(t) - 2G(t)\mathbb{1}(t \geq x) + \mathbb{1}(t \geq x)\, dG(x)\, dt \\ &= \int_{\mathbb{R}} G^2(t) - 2G^2(t) + G(t)\, dt \end{aligned}$$

$$= \int_{\mathbb{R}} G(t)(1 - G(t))\, dt.$$

The divergence function associated with CRPS is

$$d(F, G) = \int_{\mathbb{R}} (F(t) - G(t))^2\, dt.$$

△

4

Distributional Regression with Likelihood Ratio Order Constraint

Recall from Chapter 1, that the goal of distributional regression is to estimate the cumulative distribution function of the response variable given the covariates. We will restrict ourselves to univariate regression. In usual linear regression, we model the relationship between the response and the covariate by its conditional expectation. Then, we assume the expected response variable, given the covariate, is linear in the parameters. In isotonic regression, we only suppose that the relationship between the response and the covariates is monotonic. In other words, for any $x_1 < x_2$ we have $\mathbb{E}[Y|X = x_i] \leq \mathbb{E}[Y|X = x_2]$.

In isotonic distributional regression, we impose a certain order on the conditional distribution function. Let $y \mapsto F(\cdot|x)$ denote the conditional distribution of $Y|X = x$. Suppose we want that for any given covariates $x_1 \leq x_2$, we have $F(y|x_1) \geq F(y|x_2)$ for all y. Then we have seen in Chapter 2 that we impose the usual stochastic order: $[Y|X = x_1] \leq_{\mathrm{st}} [Y|X = x_2]$. Distributional regression under this constraint has been studied by Henzi et al. (2021). We also have seen that the likelihood ratio order is the strongest order compared with the hazard ratio and the usual stochastic order. Therefore, it is natural to develop a new estimation technique with likelihood ratio order as the constraint. Fortunately, Mösching and Dümbgen (2024) have developed such a technique, which we describe later in this chapter.

The structure of this chapter goes as follows. We start by discussing distributional regression for the parametric case. In Section 4.1, we model the conditional distribution using the normal distribution. It turns out that imposing likelihood ratio order when we choose this particular model, yields the isotonic regression. We then shift focus to a non-parametric approach to estimating the conditional distributions in Section 4.2. Here, we explain the estimator that Mösching and Dümbgen (2024) proposed. The method utilizes empirical likelihood and maximizes the likelihood function to obtain the estimators.

Before we proceed with the chapter, let us describe the setting and define some notations that we will use throughout this chapter. Let $\mathcal{D}_n := (X_i, Y_i)_{i=1}^n = (x_i, y_i)_{i=1}^n$ denotes a data set with n observations. Both $(Y_i)_{i=1}^n$ and $\mathbf{X} := (X_i)_{i=1}^n$ are real-valued random variables. Suppose each (X_i, Y_i) are identically distributed from P and $(Y_i)_{i=1}^n$ is independent given $\mathbf{X}$. We denote $F_{Y|x}(y|x) := \mathbb{P}(Y \leq y|X = x)$ to be the conditional distribution of $Y|X = x$, and we assume that it has a density function $f(y|x)$. We use $\mathcal{D}_n$ to estimate the family of conditional distributions $F_{Y|x}(y|x)$, such that

$$x_1 < x_2 \text{ and } y_1 < y_2 \Longrightarrow f(y_2|x_1)f(y_1|x_2) \leq f(y_1|x_1)f(y_2|x_2), \tag{4.1}$$

that is, $[Y|X = x_1] \leq_{\mathrm{lr}} [Y|X = x_2]$ (see Section 2.3 for a definition of this order). The resulting estimator will be denoted by $\widehat{F}_{Y|X}$.

4.1. The parametric estimation case: normal distribution

We have shown in Example 2.3.3 several instances in which random variables that follow normal distributions respect the likelihood ratio order. In this regression setting, we will first assume that the mean and the variance parameter depend on $\mathbf{X}$. More formally, suppose that a given data set $\mathcal{D}_n$ follows from an unknown distribution P, but we know that $Y|X_i \sim \mathcal{N}(\mu(X_i), \sigma(X_i)^2)$ for any $1 \leq i \leq n$, such that $x \mapsto \mu(x)$ and $x \mapsto \sigma(x)$ is an arbitrary functions of x. Let $(x_i, y_i)_{i=1}^n$ be the realization of $\mathcal{D}_n$. The goal is to estimate $\mu(x_i)$ and, if possible, as well as $\sigma(x_i)$ for each $1 \leq i \leq n$ under the likelihood ratio order constraint by using $\mathcal{D}_n$. It turns out that when the conditional distribution is modelled using the normal distribution under the likelihood order constraint, optimizing the likelihood of the model is equivalent to isotonic regression (see Appendix C).

We first find under what condition $[Y|X_s = x_s] \leq_{\mathrm{lr}} [Y|X_t = x_t]$, where $x_s \leq x_t$. Let $f_{Y|X_i}(y|x_i)$ be the density function of $\mathcal{N}(\mu(x_i), \sigma(x_i)^2)$. Then, for any $y \in \mathbb{R}$,

$$\frac{f_{Y|X_t}(y|x_t)}{f_{Y|X_s}(y|x_s)} = \frac{\sigma(x_s)}{\sigma(x_t)} \exp\left(\frac{(y-\mu(x_s))^2}{2\sigma(x_s)^2} - \frac{(y-\mu(x_t))^2}{2\sigma(x_t)^2}\right)$$
$$= \frac{\sigma(x_s)}{\sigma(x_t)} \exp\left(\frac{\sigma(x_t)^2(y-\mu(x_s))^2 - \sigma(x_s)^2(y-\mu(x_t))^2}{2\sigma(x_s)^2\sigma(x_t)^2}\right)$$
$$= \frac{\sigma(x_s)}{\sigma(x_t)} \times$$
$$\exp\left(\frac{y^2(\sigma(x_t)^2 - \sigma(x_s)^2) + 2y(\mu(x_t)\sigma(x_s)^2 - \mu(x_s)\sigma(x_t)^2) + \mu(x_s)^2\sigma(x_t)^2 - \mu(x_t)^2\sigma(x_s)^2}{2\sigma(x_s)^2\sigma(x_t)^2}\right).$$

From these results, we are unable to conclude that the density ratio is increasing y due to an unknown behavior of $y \mapsto y^2(\sigma(x_t)^2 - \sigma(x_s)^2)$. To remove this term, we can let $\sigma(x_t) = \sigma(x_s) =: \sigma$, i.e. the distribution of $Y|X_i$ has equal variance. Then, we are left with the following result

$$\frac{f_{Y|X_t}(y|x_t)}{f_{Y|X_s}(y|x_s)} = \exp\left(\frac{2y(\mu(x_t) - \mu(x_s)) + \mu(x_s)^2 - \mu(x_t)^2}{2\sigma^2}\right). \tag{4.2}$$

In this case, the ratio is increasing in y if and only if $\mu(x_s) \leq \mu(x_t)$ for any $x_s \leq x_t$. It means that $x \mapsto \mu(x)$ is an isotonic function of $x \in \mathbb{R}$. Indeed, if $x \mapsto \mu(x)$ is isotonic, then $x_s \leq x_t$ implies $\mu(x_s) \leq \mu(x_s)$, which yields $[Y|X_s = x_s] \sim \mathcal{N}(\mu(x_s), \sigma^2)$ is smaller than $[Y|X_t = x_t] \sim \mathcal{N}(\mu(x_t), \sigma^2)$ in the likelihood ratio.

With this reason, we assume that $[Y|X_i = x_i] \sim \mathcal{N}(\mu(x_i), \sigma^2)$ for all $1 \leq i \leq n$. Suppose for now that $\mu(x_i)$ and σ^2 for any $1 \leq i \leq n$ are unknown. We compute the likelihood $L(\mu(x), \sigma^2|\mathcal{D}_n)$ for each $1 \leq i \leq n$. We have

$$L(\mu(x), \sigma^2|\mathcal{D}_n) := L(\mu(x), \sigma^2|\mathcal{D}_n) = \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \mu(x_i))^2\right)$$
$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \mu(x_i))^2\right).$$

So the log-likelihood is

$$\log L(\mu(\mathbf{x}), \sigma^2|\mathcal{D}_n) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \mu(x_i))^2. \tag{4.3}$$

Let $\widehat{\mu(x)}$ be the estimated $\mu(x)$ and $\widehat{\sigma^2}$ is the estimated σ^2. Then we should solve the following optimization problem without likelihood ratio order constraint:

$$\left(\widehat{\mu(x)}, \widehat{\sigma^2}\right) := \underset{\mu(x),\sigma^2}{\arg\max}\left\{-\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \mu(x_i))^2\right\}.$$

To add the likelihood ratio order constraint, let us first sort the realizations $\mathcal{D}_n$ since the order matters:

$$\{x_1, \dots, x_n\} = \{x_{(1)}, \dots, x_{(\ell)}\} \quad \text{and} \quad \{y_1, \dots, y_n\} = \{y_{(1)}, \dots, y_{(m)}\},$$

where $x_{(1)} < \dots < x_{(\ell)}$ and $y_{(1)} < \dots < y_{(m)}$ for some $1 \leq \ell, m \leq n$. Then the log-likelihood in (4.3) becomes

$$\log L(\mu(x), \sigma^2 | \mathcal{D}_n) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{j=1}^{\ell}\sum_{k=1}^{m} w_{jk}(y_{(k)} - \mu(x_{(j)}))^2.$$

As we have explained previously, the likelihood ratio is increasing if $x \mapsto \mu(x)$ is an isotonic function. Therefore, we add the following constraint on $\mu(x)$:

$$\mu(x_{(j)}) \leq \mu(x_{(j+1)}) \qquad \forall\ 1 \leq j < \ell.$$

So, we state the optimization problem as the following:

$$\begin{aligned}\left(\widehat{\mu(x)}, \widehat{\sigma^2}\right) &:= \underset{\mu(x),\sigma^2}{\arg\max}\left\{-\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{j=1}^{\ell}\sum_{k=1}^{m} w_{jk}(y_{(k)} - \mu(x_{(j)}))^2\right\} \qquad (4.4)\\ \text{s.t.}\quad & \mu(x_{(j)}) \leq \mu(x_{(j+1)}) \qquad \forall\ 1 \leq j < \ell.\end{aligned}$$

Note that if we choose a specific model for μ, such that it is an isotonic function of x, then we can also remove this constraint. For instance, let $\mu(x) = \alpha + \beta x$ where α and $\beta \geq 0$. Then, the solution of the optimization problem is the same as the solution that is obtained from the ordinary least squares method. See Lemma E.1 in Appendix E for more detailed computations.

Another remark is that σ^2 is independent of the constraint. Therefore, we can maximize the log-likelihood w.r.t. σ^2. We have

$$\begin{aligned}&\frac{\partial L}{\partial\sigma^2}(\mu(x), \sigma^2|\mathcal{D}_n) = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}\sum_{j=1}^{\ell}\sum_{k=1}^{m} w_{jk}(y_{(k)} - \mu(x_{(j)}))^2 = 0\\ \Leftrightarrow\quad & \widehat{\sigma^2} = \frac{1}{n}\sum_{j=1}^{\ell}\sum_{k=1}^{m} w_{jk}(y_{(k)} - \mu(x_{(j)}))^2.\end{aligned}$$

If we substitute σ^2 with $\widehat{\sigma^2}$, the optimization problem in (4.4) becomes

$$\begin{aligned}\widehat{\mu(x)} &:= \underset{\mu(x)}{\arg\max}\left\{-\frac{n}{2}\log\left(\sum_{j=1}^{\ell}\sum_{k=1}^{m} w_{jk}(y_{(k)} - \mu(x_{(j)}))^2\right)\right\}\\ \text{s.t.}\quad & \mu(x_{(j)}) \leq \mu(x_{(j+1)}) \qquad \forall\ 1 \leq j < \ell.\end{aligned}$$

Equivalently, if we replace the log-likelihood with the negative log-likelihood, we want to solve

$$\widehat{\mu(x)} := \underset{\mu(x)}{\arg\min}\left\{\sum_{j=1}^{\ell}\sum_{k=1}^{m} w_{jk}(y_{(k)} - \mu(x_{(j)}))^2\right\} \qquad (4.5)$$

$$\text{s.t.}\quad \mu(x_{(j)}) \leq \mu(x_{(j+1)}) \qquad \forall\ 1 \leq j < \ell. \qquad (4.6)$$

As it turns out, the minimization problem in (4.5) is an isotonic regression (see Appendix C). To see this, let

$$w_{j+} := \sum_{k=1}^{m} w_{jk} \qquad \text{and} \qquad \overline{y}_j := \frac{\sum_{k=1}^{m} w_{jk}y_{(k)}}{w_{j+}}.$$

Then,

$$\sum_{j=1}^{\ell}\sum_{k=1}^{m} w_{jk}(y_{(k)} - \mu(x_{(j)}))^2 = \sum_{j=1}^{\ell}\sum_{k=1}^{m} w_{jk}(y_{(k)} - \overline{y}_j + \overline{y}_j - \mu(x_{(j)}))^2$$

$$= \sum_{j=1}^{\ell} \sum_{k=1}^{m} w_{jk} (y_{(k)} - \overline{y}_j)^2 + 2 \sum_{j=1}^{\ell} \sum_{k=1}^{m} w_{jk} (y_{(k)} - \overline{y}_j)(\overline{y}_j - \mu(x_{(j)}))$$
$$+ \sum_{j=1}^{\ell} w_{j+} (\overline{y}_j - \mu(x_{(j)}))^2$$

The first sum is independent of $\mu(x)$, and the second term is equal to zero (see Lemma E.2 in Appendix E). Hence, the minimization problem in (4.5) with the constraint (4.6) becomes

$$\widehat{\mu(x)} := \underset{\mu(x)}{\arg\min} \left\{ \sum_{j=1}^{\ell} w_{j+} (\overline{y}_j - \mu(x_{(j)}))^2 \right\}$$
$$\text{s.t.} \quad \mu(x_{(j)}) \leq \mu(x_{(j+1)}) \qquad \forall\, 1 \leq j < \ell.$$

Therefore, in this particular case, distributional regression under likelihood ratio order constraint is equivalent to isotonic regression of $\overline{y}_j$ with weights w_{j+}.

4.2. The non-parametric case

Now that we have treated the parametric case, we will discuss an estimator in which the model belongs to a much bigger class of distributions. In particular, Mösching and Dümbgen (2024) choose to model the conditional distributions that belong to some discrete distributions with unknown probabilities. These weights are then estimated subject to the likelihood ratio order constraint. In this section, we will explain how Mösching and Dümbgen (2024) estimate these unknown weights. The idea is to construct an empirical likelihood. In Section 4.2.1, we review what empirical likelihood is and apply it to the distributional regression problem. The empirical likelihood is rewritten in such a way that the maximization problem can be solved numerically by using a well-known method from isotonic regression. These will be explained in more detail in Section 4.2.2 and Section 4.2.3.

To apply this method, we sort the response and the covariate values in $\mathcal{D}_n$ from the smallest to the largest value. We also remove ties in the value, and so we use the following observations for the estimation:

$$\{X_1, \ldots, X_n\} = \{x_1, \ldots, x_\ell\} \quad \text{and} \quad \{Y_1, \ldots, Y_n\} = \{y_1, \ldots, y_m\},$$

with $x_1 < \cdots < x_\ell$ and $y_1 < \cdots < y_m$. The following constant defines how many observations have the same values as (x_j, y_k), where $1 \leq j \leq \ell$ and $1 \leq k \leq m$. We define

$$w_{jk} := \#\{i : (X_i, Y_i) = (x_j, y_k)\}.$$

4.2.1. The construction of the likelihood: empirical likelihood

To estimate the conditional distributions $F_{Y|x_j}$ for each $1 \leq j \leq \ell$, Mösching and Dümbgen (2024) used the empirical likelihood approach. Let us briefly review what this approach is, and apply it for the estimation problem.

The empirical approach uses the empirical likelihood of a cumulative distribution function, which is then maximized to get a non-parametric estimate of the distribution. The following is the definition of the empirical likelihood of a distribution from Owen (2001, p. 6)

Definition 4.2.1 (The empirical likelihood). *Let $(X_i)_{i=1}^n$ be any i.i.d. real-valued random variables. The empirical likelihood is*

$$L(F) := \prod_{i=1}^{n} [F(X_i) - F(X_i-)],$$

where $F(x) = \mathbb{P}(X \leq x)$ and $F(x-) = \mathbb{P}(X < x)$.

Note that if the distribution function is continuous, then the empirical likelihood is zero. If $(X_i)_{i=1}^n$ is a sequence of discrete random variables and i.i.d. of F, then $F(x) - F(x-) = \mathbb{P}(X_i = x)$. To have a non-zero likelihood, we must assign non-zero probabilities to each observation.

When we maximize the empirical likelihood unconstrained, we will get the cumulative empirical distribution function (ECDF). This result is also stated and proven in Theorem 2.1 in Owen (2001, p. 7). However, we decide to show it in a different way. Therefore, the ECDF is a non-parametric maximum likelihood estimator of a distribution. We state this result and reprove it in the following theorem.

Theorem 4.2.2. *Let $(X_i)_{i=1}^n$ be any i.i.d. real-valued random variables. Let F_n be the ECDF of $(X_i)_{i=1}^n$, i.e.*

$$F_n(x) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i \leq x).$$

Let F be any cumulative distribution function. Then F_n maximizes the empirical likelihood $L(F)$.

Proof. Let $\{X_1, \dots, X_n\} = \{x_1, \dots, x_\ell\}$ such that $x_1 < x_2 < \cdots < x_\ell$ for some $1 \leq \ell \leq n$. W.l.o.g. we assume that $p_j \neq 0$. Let $n_j = \#\{i : X_i = x_j\} > 0$, so that $\sum_{j=1}^{\ell} n_j = n$ and $p_j := F(X_j) - F(X_j-)$. Then,

$$f(\mathbf{p}) := \log L(F) = \sum_{j=1}^{\ell} n_j \log p_j. \tag{4.7}$$

We would like to maximize (4.7) subject to $\sum_{j=1}^{\ell} p_j = 1$. So then (4.7) is simply a function of $\mathbf{p} := (p_1, \dots, p_\ell)$, which is defined on a set

$$S := \left\{ \mathbf{p} \in (0,\infty)^\ell : \sum_{j=1}^{\ell} p_j = 1 \right\}.$$

We use the Lagrange multiplier to solve this optimization problem. Let

$$\mathcal{L}(\mathbf{p}, \lambda) := \log L(F) + \lambda \left(1 - \sum_{j=1}^{\ell} p_j \right).$$

Then, for any $1 \leq j \leq \ell$,

$$\frac{\partial \mathcal{L}}{\partial p_j}(\mathbf{p}, \lambda) = \frac{n_j}{p_j} - \lambda = 0 \qquad \Longrightarrow \qquad p_j = \frac{n_j}{\lambda}.$$

From the constraint, we obtain that

$$\sum_{j=1}^{\ell} p_j = \frac{1}{\lambda}\sum_{j=1}^{\ell} n_j = 1 \qquad \Longrightarrow \qquad \lambda = \sum_{j=1}^{\ell} n_j = n.$$

Therefore,

$$p_j = \frac{n_j}{n}.$$

Let $\widehat{\mathbf{p}} := n^{-1}(n_1, \dots, n_\ell)$, then $\widehat{\mathbf{p}}$ maximizes $f(\mathbf{p})$. This is because the function $f(\mathbf{p})$ is strictly concave in $\mathbf{p}$, hence $\widehat{\mathbf{p}}$ is the unique maximizer of $f(\mathbf{p})$. In the case that $\ell = n$, we have that $n_j = 1$, so that $p_j = 1/n$ for any $1 \leq j \leq \ell$. These are the probabilities for the ECDF. □

We now go back to our distributional regression problem. We would like to estimate $F_{Y|x}$, for any $x \in \mathcal{X}$ using $\mathcal{D}_n$. To use the empirical likelihood approach, we first assume that $F_{Y|x}$ has a support $\{y_1, \dots, y_m\}$ for each observed $x_j \in \mathcal{X}$, $1 \leq j \leq \ell$. Then the ECDF for each $x_j \in \mathcal{X}$ is

$$F_{Y|x_j}(y|x_j) := \sum_{k=1}^{m} q_{jk} \mathbb{1}(y_k \leq y),$$

where q_{jk} can be understood as the probability of observing y_k given the value x_j, which is unknown. The empirical likelihood is then

$$\begin{aligned} L(F_{Y|x_1}, \dots, F_{Y|x_\ell}) &= \prod_{j=1}^{n} \prod_{k=1}^{n} \left[F_{Y|x_j}(Y_k|x_j) - F_{Y|x_j}(Y_i - |x_j) \right] \\ &= \prod_{j=1}^{\ell} \prod_{k=1}^{m} \left[F_{Y|x_j}(y_k|x_j) - F_{Y|x_j}(y_k - |x_j) \right]^{w_{jk}} \\ &= \prod_{j=1}^{\ell} \prod_{k=1}^{m} q_{jk}^{w_{jk}}, \end{aligned}$$

and so its empirical log-likelihood is

$$\Lambda(\mathbf{q}) := \log L(F_{Y|x_1}, \dots, F_{Y|x_\ell}) = \sum_{j=1}^{\ell} \sum_{k=1}^{m} w_{jk} \log q_{jk}, \tag{4.8}$$

where $\mathbf{q} \in [0,1]^{\ell \times m}$. If we add the likelihood ratio order constraint (4.1), then the optimization problem that we want to solve is

$$\begin{aligned} \widehat{\mathbf{q}} &:= \underset{\mathbf{q} \in [0,1]^{\ell \times m}}{\arg\max} \ \Lambda(\mathbf{q}) \\ \text{s.t.} \quad & \sum_{k=1}^{m} q_{jk} = 1, \qquad \forall 1 \le j \le \ell, & (4.9) \\ & q_{j_1 k_2} q_{j_2 k_1} \le q_{j_1 k_1} q_{j_2 k_2}, \qquad \forall 1 \le j_1 < j_2 \le \ell, 1 \le k_1 < k_2 \le m. & (4.10) \end{aligned}$$

The constraint of the form in (4.10) will appear often throughout this chapter, and so we refer to this form of constraint as the "likelihood ratio order constraint". This optimization problem has ℓ constraints to ensure that weights q_{jk} are summed up to one for each j. There are further $\binom{\ell}{2}\binom{m}{2}$ inequality constraints for the likelihood ratio order constraints. It turns out that we can reduce the number of constraints, by relating this optimization problem with estimating the joint distribution (X, Y) under the likelihood order constraint.

4.2.2. The empirical likelihood of the joint distribution

As mentioned at the end of the previous section, we can reduce the number of constraints. Instead of estimating the conditional distribution for each covariate value, one can estimate the joint distribution (X, Y) using the empirical approach as well. Mösching and Dümbgen (2024) showed that these two estimation problems are equivalent.

The ECDF of the observed values $(x_j, y_k)_{j,k}$ is

$$F_{X,Y}(x, y) = \sum_{j=1}^{\ell} \sum_{k=1}^{m} h_{jk} \mathbb{1}(x_j \le x, y_k \le y),$$

where $h_{jk} \ge 0$ and $h_{++} := \sum_{j=1}^{\ell} \sum_{k=1}^{m} h_{jk} = 1$. When we construct the empirical likelihood, we obtain the same likelihood as in (4.8). However, the number of constraints that we use for this optimization problem is $\binom{\ell}{2}\binom{m}{2} + 1$, instead of $\binom{\ell}{2}\binom{m}{2} + \ell$. The optimization problem that we then want to solve is

$$\widehat{\mathbf{h}} := \underset{\mathbf{h} \in [0,1]^{\ell \times m}}{\arg\max} \sum_{j=1}^{\ell} \sum_{k=1}^{m} w_{jk} \log h_{jk} = \underset{\mathbf{h} \in [0,1]^{\ell \times m}}{\arg\max} \ \Lambda(\mathbf{h}) \tag{4.11}$$

$$\text{s.t.} \quad h_{++} = 1, \tag{4.12}$$

$$h_{j_1k_2}h_{j_2k_1} \leq h_{j_1k_1}h_{j_2k_2}, \qquad \forall 1 \leq j_1 < j_2 \leq \ell, 1 \leq k_1 < k_2 \leq m. \tag{4.13}$$

As a remark, since h_{jk} may interpreted as a probability mass function, then the additional constraint (4.13) implies that we estimate an unknown TP$_2$ distribution. The term TP$_2$ stands for 'total positive of order 2'. To see more details about TP$_2$ distribution, we refer the reader to Appendix B.

Note that we can remove the constraint (4.12) by using the Lagrange multiplier. Let

$$\mathcal{L}(\mathbf{h}, \lambda) = \Lambda(\mathbf{h}) + \lambda(1 - h_{++}).$$

Then for any $1 \leq j \leq \ell$ and $1 \leq k \leq m$,

$$\frac{\partial \mathcal{L}}{\partial h_{jk}} = \frac{w_{jk}}{h_{jk}} - \lambda \quad \text{and} \quad \frac{\partial \mathcal{L}}{\partial \lambda} = 1 - h_{++}.$$

Set both equations to zero and we solve for λ. We have

$$h_{jk} = \frac{w_{jk}}{\lambda} \quad \Longrightarrow \quad \sum_{j=1}^{\ell}\sum_{k=1}^{m} h_{jk} = \frac{\sum_{j=1}^{\ell}\sum_{k=1}^{m} w_{jk}}{\lambda} \quad \Longrightarrow \quad 1 = \frac{n}{\lambda} \quad \Longrightarrow \quad \lambda = n.$$

Let $\Lambda^*(\mathbf{h}) = \mathcal{L}(\mathbf{h}, n)$, then we need to show that the following optimization problem

$$\begin{aligned} \widehat{\mathbf{h}} = \underset{\mathbf{h} \in [0,\infty)^{\ell \times m}}{\arg\max}\, \Lambda^*(\mathbf{h}) &= \underset{\mathbf{h} \in [0,\infty)^{\ell \times m}}{\arg\max}\, \Lambda(\mathbf{h}) + n(1 - h_{++}) \\ \text{s.t.} \quad h_{j_1k_2}h_{j_2k_1} &\leq h_{j_1k_1}h_{j_2k_2}, \qquad \forall 1 \leq j_1 < j_2 \leq \ell, 1 \leq k_1 < k_2 \leq m, \end{aligned} \tag{4.14}$$

also solves (4.11) with constraints (4.12) and (4.13). Indeed, take any $\mathbf{h} \in [0, 1]^{\ell \times m}$ such that $h_{++} = 1$ and $\Lambda^*(\mathbf{h}) > -\infty$. Let $\widetilde{\mathbf{h}} := (h_{jk}/h_{++})_{j,k}$, then it is clear that $\widetilde{\mathbf{h}}$ satisfies (4.13) if and only if $\mathbf{h}$ also satisfies this constraint and $\widetilde{\mathbf{h}}_{++} = 1$. Therefore $\Lambda^*(\widetilde{\mathbf{h}}) = \Lambda(\widetilde{\mathbf{h}})$. Further,

$$\begin{aligned} \Lambda^*(\mathbf{h}) &= \sum_{j=1}^{\ell}\sum_{k=1}^{m} w_{jk} \log h_{jk} + n(1 - h_{++}) \\ &= \sum_{j=1}^{\ell}\sum_{k=1}^{m} w_{jk} \log\left(\frac{h_{jk}}{h_{++}} h_{++}\right) + n(1 - h_{++}) \\ &= \Lambda(\widetilde{\mathbf{h}}) + n(\log h_{++} - h_{++} + 1) \\ &\leq \Lambda(\widetilde{\mathbf{h}}) = \Lambda^*(\widetilde{\mathbf{h}}), \end{aligned}$$

where the inequality is obtained by using $\log x \leq x - 1$ for any $x \in [0, \infty)$. The inequality $\Lambda^*(\mathbf{h}) \leq \Lambda^*(\widetilde{\mathbf{h}})$ becomes equality if and only if $h_{++} = 1$, meaning that $\mathbf{h} = \widetilde{\mathbf{h}}$. Hence, a maximizer of Λ^* also maximizes Λ, because it satisfies the likelihood ratio order and the entries have to sum up to 1.

Lastly, the maximizer of $\Lambda^*(\mathbf{h})$ also maximizes $\Lambda(\mathbf{q})$, and vice versa. For an arbitrary $\mathbf{h} \in [0, \infty)^{\ell \times m}$ such that $\Lambda(\mathbf{h}) > -\infty$, write

$$h_{jk} = p_j q_{jk}, \quad \text{with } p_j := \sum_{k=1}^{m} h_{jk} = h_{j+} \text{ and } q_{jk} := \frac{h_{jk}}{h_{j+}}.$$

Note $\mathbf{q} = [q_{jk}]_{j,k} \in [0, \infty)^{\ell \times m}$, and $\mathbf{h}$ satisfies the likelihood ratio order constraint if and only if $\mathbf{q}$ does. Then,

$$\Lambda(\mathbf{h}) = \sum_{j=1}^{\ell}\sum_{k=1}^{m} w_{jk} \log h_{jk} = \sum_{j=1}^{\ell}\sum_{k=1}^{m} w_{jk}(\log p_j + \log q_{jk}) = \sum_{j=1}^{\ell} w_{j+} \log p_j + \Lambda(\mathbf{q}),$$

also,

$$h_{++} = \sum_{j=1}^{\ell} h_{j+} = \sum_{j=1}^{\ell} p_j \quad \text{and} \quad n = \sum_{j=1}^{\ell}\sum_{k=1}^{m} w_{jk} = \sum_{j=1}^{\ell} w_{j+}.$$

Hence, we rewrite (4.14) into the following expression

$$\begin{aligned}\Lambda^*(\mathbf{h}) &= \Lambda(\mathbf{h}) + n(1 - h_{++}) \\ &= \Lambda(\mathbf{q}) + \sum_{j=1}^{\ell} w_{j+} \log p_j + \sum_{j=1}^{\ell} w_{j+} \left(1 - \sum_{j=1}^{\ell} p_j\right) \\ &= \Lambda(\mathbf{q}) + \sum_{j=1}^{\ell} (w_{j+} \log p_j - np_j + w_{j+}).\end{aligned}$$

Let $\mathbf{p} = (p_1, \ldots, p_\ell)$ and $\mathrm{M}(\mathbf{p}) := \sum_{j=1}^{\ell} (w_{j+} \log p_j - np_j + w_{j+})$, then $\mathrm{M}(\mathbf{p})$ is uniquely maximized at $(w_{j+}/n)_{j=1}^{l}$. This easily verified by differentiating $\mathrm{M}(\mathbf{p})$ w.r.t. $\mathbf{p}$ and solve $\mathrm{M}(\mathbf{p})/d\mathbf{p} = 0$. Note that $(w_{j+}/n)_{j=1}^{l}$ sums up to 1. The uniqueness and existence of the maximizer is guaranteed because $\mathrm{M}(\mathbf{p})$ is convex. Now the equivalence of the two estimation problems can be explained by using the maximizer $\widehat{h}_{j+} = \widehat{p}_j = w_{j+}/n$.

- Assume $\widehat{\mathbf{h}}$ is a maximizer of $\Lambda^*(\mathbf{h})$ under the likelihood order constraint. Then $\widehat{q}_{jk} := \widehat{h}_{jk}/\widehat{h}_{j+}$ is also a maximizer of $\Lambda(\mathbf{q})$ that satisfies the constrains (4.9) and (4.10).
- If $\widehat{\mathbf{q}}$ maximizes $\Lambda(\mathbf{q})$ that satisfies the constraints (4.9) and (4.10), then $\widehat{h}_{jk} = \widehat{p}_j \widehat{q}_{jk}$ maximizes $\Lambda^*(\mathbf{h})$ under the likelihood ratio order constraint.

The equivalence of the estimation problem implies that we only need to find $\mathbf{h}$ that maximizes $\Lambda^*(\mathbf{h})$. We next describe the procedure of finding $\mathbf{h}$ that maximizes $\Lambda^*(\mathbf{h})$, under the likelihood ratio constraint in (4.13).

4.2.3. Estimation procedure

By estimating the joint distribution of (X, Y), we have seen in the previous section that the number of constraints is less than in the original problem. The optimization problem is further relaxed through the use a Lagrange multiplier. Now, we discuss the estimation procedure proposed by Mösching and Dümbgen (2024). Before we proceed, we need to discuss the reduction of the number of parameters and how the number of constraints can be further minimized drastically. Then, after rewriting and reparameterizing the objective again, the optimization problem is solved by using an iterative algorithm. The algorithm computes a new proposal and performs a linear search so that the objective function decreases. However, we will not discuss in detail how the linear search is done.

The reduction of the dimension and the number of constraints

The parameter space $\mathbf{h}$ is in $[0, \infty)^{\ell \times m}$. This space can be reduced to a smaller space, consequently reducing the number of constraints.

Define a set $\mathcal{P}$, which consists of pair (j, k) such that $m_j \le k \le M_j$, where

$$m_j := \min\{k : w_{j'k} > 0 \text{ for some } j' \ge j\} \quad \text{and} \quad M_j := \max\{k : w_{j'k} > 0 \text{ for some } j' \le j\},$$

or equivalently there exists ℓ_k and L_k such that $\ell_k \le \ell \le L_k$, where

$$\ell_k := \min\{\ell : w_{jk'} > 0 \text{ for some } k' \ge k\} \quad \text{and} \quad L_k := \max\{\ell : w_{jk'} > 0 \text{ for some } k' \le k\}.$$

An illustration of $\mathcal{P}$ is in Figure 4.1. We observe that $(j, k) \notin \mathcal{P}$ if at (j, k) there are no observed points in one of the shaded areas. As a remark, the set $\mathcal{P}$ may also be defined as follows:

$$\mathcal{P} := \{(j, k) : w_{j_1, k_2}, w_{j_2, k_1} > 0 \text{ for some } 1 \le j_1 \le j \le j_2 \text{ and } 1 \le k_1 \le k \le k_2 \le m\}. \tag{4.15}$$

The indices of the observed points are certainly in $\mathcal{P}$. In the example in Figure 4.1, the location of a point P, which is $(5, 4)$ is in $\mathcal{P}$ because we can choose $j_1 = 4, k_2 = 4$ and $j_2 = 5, k_1 = 3$. An equivalent representation of $\mathcal{P}$ is

$$\mathcal{P} = \{(j, m_j) : 1 \le j \le \ell\} \cup \bigcup_{k=2}^{m} \{(j, k) : \ell_k \le j \le L_{k-1}\}. \tag{4.16}$$

Let $A := \{(j, m_j) : 1 \leq j \leq \ell\}$. The set A represent the "lowest" points on j. In the example of Figure 4.1, the elements $(4,3)$ and $(5,3)$ are examples of points in A. The points on the left of $(6,4)$, e.g. the points $(4,4)$ and $(5,4)$ are in the set $\cup_{k=2}^{m}\{(j,k) : \ell_k \leq j \leq L_{k-1}\}$. So the latter set is a collection of points on the left of (j, L_k), up until the point (j, ℓ_k), for each $k = 2, \ldots, m$. Another equivalent representation of $\mathcal{P}$ is

$$\mathcal{P} = A \cup \{(j,k) : 1 \leq j \leq \ell, m_j < k \leq M_j\}.$$

The points in the set $\{(j,k) : 1 \leq j \leq \ell, m_j < k \leq M_j\}$ essentially represents the points above m_j up until M_j for each $1 \leq j \leq \ell$.

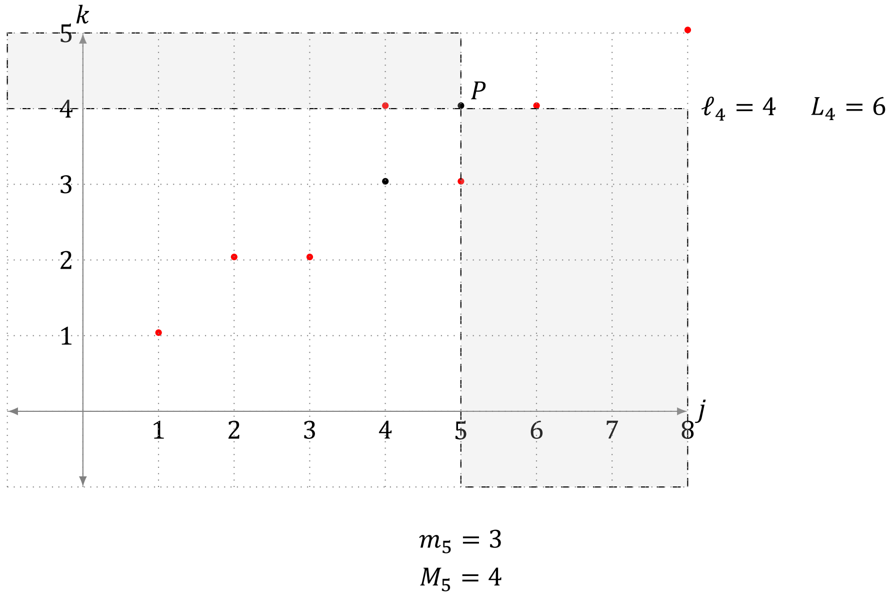

Figure 4.1: In this example, we have $\ell = 7$ and $m = 5$. The red dots are observations such that $w_{jk} > 0$ and the black dots are the unobserved points. These dots represent $\mathcal{P}$. The shaded area in the bottom right can be used to find m_j or L_k. The shaded area in the top left is used to find M_j or ℓ_k. The point P is at $(5,4)$ and it belongs to $\mathcal{P}$ because we can find m_5 and M_5 such that $m_5 \leq 4 \leq M_5$.

The points whose indices are outside of $\mathcal{P}$ does not worsen Λ^* and the number of constraints is drastically reduced. These properties of $\mathcal{P}$ are stated in Lemma 1 in Mösching and Dümbgen (2024). We state the results of this lemma below without proof.

1. We have $h_{jk} > 0$ for any $(j,k) \in \mathcal{P}$, if $\mathbf{h}$ satisfies the likelihood ratio order constraint and $\Lambda^*(\mathbf{h}) > -\infty$.

2. Let $\tilde{\mathbf{h}} := (\mathbb{1}((j,k) \in \mathcal{P})h_{jk})_{j,k}$, then $\tilde{\mathbf{h}}$ also satisfies the likelihood ratio order and $\Lambda^*(\tilde{\mathbf{h}}) \geq \Lambda^*(\mathbf{h})$ with equality if and only if $\mathbf{h} = \tilde{\mathbf{h}}$.

3. We can replace the likelihood ratio order constraint by

$$h_{j-1,k}h_{j,k-1} \leq h_{j-1,k-1}h_{j,k}, \quad 1 < j \leq \ell.1 < k \leq m, \tag{4.17}$$

if $\mathbf{h} \in [0,\infty)^{\ell \times m}$ such that $\{(j,k) : h_{jk} > 0\} = \mathcal{P}$. Consequently, the number of constraints reduces to $(\ell-1)(m-1)$ inequalities, which is much smaller than $\binom{\ell}{2}\binom{m}{2}$ for large ℓ and m.

From property (1) and (2), we set $h_{jk} := 0$ for any $(j,k) \notin \mathcal{P}$, and $h_{jk} > 0$ for any $(j,k) \in \mathcal{P}$. Instead of the parameter space being in $[0,\infty)^{\ell \times m}$, we now focus on $\mathbf{h} \in (0,\infty)^{\mathcal{P}}$. Further for any $1 < j \leq \ell$ and $1 < k \leq m$, it is sufficient that $(j-1,k), (j,k-1) \in \mathcal{P}$ so that $(j,k), (j-1,k-1) \in \mathcal{P}$. Indeed, if $(j-1,k), (j,k-1) \in \mathcal{P}$, then we choose $j_1 = j-1, j_2 = j$ and $k_1 = k-1, k_2 = k$ to prove that $(j,k), (j-1,k-1) \in \mathcal{P}$. Hence (4.17) becomes

$$h_{j-1,k}h_{j,k-1} \leq h_{j-1,k-1}h_{j,k} \qquad \text{if } (j-1,k), (j,k-1) \in \mathcal{P}. \tag{4.18}$$

Reparametrization of the parameter and reformulate the constraints

The purpose of reparametrization of the parameter is to transform the non-linear constraints into linear constraints. The feasible set also becomes a convex set. Furthermore, it will turn out that one can use a method from isotonic regression such as PAVA (see Appendix C), to find a search direction for the next iteration. For convenience purposes, we consider the negative log-likelihood and the objective is now to minimize it subject to the constraints.

First, we transform each entry of the matrix $\widetilde{\mathbf{h}} = (\mathbb{1}\{(j,k) \in \mathcal{P}) h_{jk}\}_{j,k}$ by using logarithm. Let $\boldsymbol{\theta} := (\log h_{jk})_{(j,k)\in\mathcal{P}} = (\theta_{jk})_{(j,k)\in\mathcal{P}}$, then the objective function $\Lambda^*(\mathbf{h})$ in (4.14) becomes

$$\sum_{(j,k)\in\mathcal{P}} (-w_{jk}\theta_{jk} - n(1 - \exp(\theta_{jk}))).$$

So, the goal is to minimize

$$f(\boldsymbol{\theta}) := \sum_{(j,k)\in\mathcal{P}} (-w_{jk}\theta_{jk} + n\exp(\theta_{jk})), \tag{4.19}$$

subject to

$$\theta_{jk} - \theta_{j,k-1} + \theta_{j-1,k-1} - \theta_{j-1,k} \geq 0, \quad 1 < j \leq \ell, 1 < k \leq m,$$

which is the constraint in (4.17).

Now, consider a transformation $T(\boldsymbol{\theta})$ of $\boldsymbol{\theta}$ for all $1 \leq j \leq \ell$, defined as follows:

$$T_{jk}(\boldsymbol{\theta}) := \widetilde{\theta}_{jk} := \begin{cases} \theta_{j,m_j} & \text{if } k = m_j, \\ \theta_{jk} - \theta_{j,k-1} & \text{if } m_j < k \leq M_j \end{cases}$$

Note that $\theta_{jk} = \sum_{k'=m_j}^{k} \widetilde{\theta}_{jk'}$, for all $1 \leq j \leq \ell$ and $m_j \leq k \leq M_j$. Then, substituting θ_{jk} with $\sum_{k'=m_j}^{k} \widetilde{\theta}_{jk'}$ yields

$$\begin{aligned}\widetilde{f}(\widetilde{\boldsymbol{\theta}}) &= \sum_{(j,k)\in\mathcal{P}} \left(-w_{jk} \left(\sum_{k'=m_j}^{k} \widetilde{\theta}_{jk'} \right) + n\exp\left(\sum_{k'=m_j}^{k} \widetilde{\theta}_{jk'} \right) \right) \\ &= \sum_{j=1}^{\ell} \sum_{k=m_j}^{M_j} \left(-w_{jk} \left(\sum_{k'=m_j}^{k} \widetilde{\theta}_{jk'} \right) + n\exp\left(\sum_{k'=m_j}^{k} \widetilde{\theta}_{jk'} \right) \right).\end{aligned}$$

For computing the first term of the summand, let us fix $1 \leq j \leq \ell$, we then have

$$\begin{aligned}\sum_{k=m_j}^{M_j} -w_{jk} \left(\sum_{k'=m_j}^{k} \widetilde{\theta}_{jk'} \right) &= -(w_{j,m_j} + \cdots + w_{j,M_j})\widetilde{\theta}_{j,m_j} - (w_{j,m_j+1} + \cdots + w_{j,M_j})\widetilde{\theta}_{j,m_j+1} - \cdots - \\ &\quad - w_{j,M_j}\widetilde{\theta}_{j,M_j} \\ &= \sum_{k=m_j}^{M_j} \left(\sum_{k'=k}^{M_j} w_{jk'} \right) \widetilde{\theta}_{jk}.\end{aligned}$$

Let $\underline{w}_{jk} = \sum_{k'=k}^{M_j} w_{jk'}$, then the objective function becomes

$$\widetilde{f}(\widetilde{\boldsymbol{\theta}}) = \sum_{j=1}^{\ell} \sum_{k=m_j}^{M_j} \left(-\underline{w}_{jk}\widetilde{\theta}_{jk} + n\exp\left(\sum_{k'=m_j}^{k} \widetilde{\theta}_{jk'} \right) \right).$$

As for the constraints, we take the log on both sides of the equation in (4.18). This yields

$$\theta_{j-1,k-1} + \theta_{jk} - \theta_{j-1,k} - \theta_{j,k-1} \geq 0, \qquad \text{if } (j-1,k), (j,k-1) \in \mathcal{P}. \tag{4.20}$$

Equivalently, we have

$$\widetilde{\theta}_{jk} \geq \widetilde{\theta}_{j-1,k}, \qquad \text{if } (j-1,k),(j,k-1) \in \mathcal{P}. \tag{4.21}$$

In (4.21), we observe that $\widetilde{\theta}_{jk}$ is increasing in index j. By a definition of $\mathcal{P}$ in (4.16), the constraint in (4.21) becomes

$$\left(\widetilde{\theta}_{jk}\right)_{j=\ell_k}^{L_{k-1}-\ell_k+1} \qquad \text{is an increasing sequence, for } 1 \leq k \leq m, L_{k-1} - \ell_k + 1 \geq 2.$$

Obtaining the search direction

Let us summarize what has been done and state some results from Mösching and Dümbgen (2024) without proof. First of all, the function $f(\boldsymbol{\theta})$ in (4.19) is strictly convex. Let Θ be the set of $\boldsymbol{\theta}$ such that (4.20) is satisfied, then the set Θ is a convex set. Theorem 1 in Mösching and Dümbgen (2024) shows that $f(\boldsymbol{\theta})$ has a minimum in the set Θ.

Next, we reparameterize $\boldsymbol{\theta}$ by $\widetilde{\boldsymbol{\theta}}$. We then obtain a function $\widetilde{f}(\widetilde{\boldsymbol{\theta}})$, with a new constraint (4.21). Let $\widetilde{\Theta}$ be a set of $\widetilde{\boldsymbol{\theta}}$ such that it satisfies (4.21). The function $\widetilde{f}$ and the feasible set $\widetilde{\Theta}$ are still strictly convex and convex respectively. To solve this minimization problem, the function $\widetilde{f}$ will be approximated by a quadratic function. The minimizer of the quadratic function yields the search direction. The full description of how to obtain the new point for the next iteration will not be explained here. We refer the reader to Section 3.5 in Mösching and Dümbgen (2024) for the explanation.

Let us apply second-order Taylor expansion of $\widetilde{f}$ at $\widetilde{\boldsymbol{\theta}}$. First of all, we apply derivatives and the inner product only for indices in $\mathcal{P}$, i.e.

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{(j,k)\in\mathcal{P}} x_{jk} y_{jk} \quad \text{and} \quad \nabla\widetilde{f}(\tilde{\mathbf{x}}) = \left(\frac{\partial \widetilde{f}(\tilde{x})}{\partial \tilde{x}_{jk}} \right)_{(j,k)\in\mathcal{P}}.$$

We have

$$\widetilde{f}(\tilde{\mathbf{x}}) \approx \widetilde{f}(\widetilde{\boldsymbol{\theta}}) + \left\langle \nabla\widetilde{f}(\widetilde{\boldsymbol{\theta}}), (\tilde{\mathbf{x}} - \widetilde{\boldsymbol{\theta}}) \right\rangle + \frac{1}{2} \sum_{(j,k)\in\mathcal{P}} \frac{\partial^2 \widetilde{f}}{\partial \tilde{x}_{jk}^2}(\widetilde{\boldsymbol{\theta}})(\tilde{x}_{jk} - \widetilde{\theta}_{jk})^2. \tag{4.22}$$

For the computation of $\nabla\widetilde{f}(\tilde{\mathbf{x}})$, we refer the reader to Lemma E.4 in Appendix E. Recall that $\theta_{jk} = \sum_{k'=m_j}^{k} \widetilde{\theta}_{jk'}$ by the definition of $\widetilde{\boldsymbol{\theta}}$. Therefore, by Lemma E.4, we get for any $(j,k) \in \mathcal{P}$,

$$\nabla\widetilde{f}(\widetilde{\boldsymbol{\theta}}) = \left(-\underline{w}_{jk} + n \sum_{k'=k}^{M_j} \exp(\theta_{jk'}) \right)_{(j,k)\in\mathcal{P}} \quad \text{and} \quad \frac{\partial^2 \widetilde{f}}{\partial \tilde{x}_{jk}^2}(\widetilde{\boldsymbol{\theta}}) = n \sum_{k'=k}^{M_j} \exp(\theta_{jk'}). \tag{4.23}$$

Note that the cross partial derivatives of $\widetilde{f}$ is equal to zero. Indeed, it is clear from $\nabla\widetilde{f}(\tilde{\mathbf{x}})$ that for any $1 \leq j \leq \ell$, $k = m_j$ and $j' \neq j$, then

$$\frac{\partial \widetilde{f}}{\partial \tilde{x}_{jk} \partial \tilde{x}_{j'k}} = \frac{\partial \widetilde{f}}{\partial \tilde{x}_{j'k} \partial \tilde{x}_{jk}} = 0.$$

In case that $2 \leq k \leq m$ and $\ell_k \leq j \leq L_{k-1}$, if we choose $k \neq k^*$, then we consider $\ell_{k^*} \leq j' \leq L_{k^*-1}$ which is a different coordinate of $\tilde{\mathbf{x}}$ that is absent in,

$$\left(\nabla\widetilde{f}(\widetilde{\boldsymbol{\theta}})\right)_{(j,k)} = -\underline{w}_{jk} + n \sum_{k'=k}^{M_j} \exp(\theta_{jk'}).$$

The second-order Taylor expansion of $\widetilde{f}$ at $\widetilde{\boldsymbol{\theta}}$ is then applied by combining (4.22) and (4.23). This yields

$$\widetilde{f}(\tilde{\mathbf{x}}) \approx \widetilde{f}(\widetilde{\boldsymbol{\theta}}) + \frac{1}{2} \sum_{(j,k)\in\mathcal{P}} \frac{\partial^2 \widetilde{f}}{\partial \tilde{x}_{jk}^2}(\widetilde{\boldsymbol{\theta}}) \left(2 \left(\frac{\partial^2 \widetilde{f}}{\partial \tilde{x}_{jk}^2}(\widetilde{\boldsymbol{\theta}}) \right)^{-1} \frac{\partial \widetilde{f}}{\partial \tilde{x}_{jk}}(\widetilde{\boldsymbol{\theta}})(\tilde{x}_{jk} - \widetilde{\theta}_{jk}) + \tilde{x}_{jk}^2 - 2\tilde{x}_{jk}\widetilde{\theta}_{jk} + \widetilde{\theta}_{jk}^2 \right)$$

$$= \widetilde{f}(\widetilde{\boldsymbol{\theta}}) + \frac{1}{2} \sum_{(j,k)\in\mathcal{P}} \frac{\partial^2 \widetilde{f}}{\partial \tilde{x}_{jk}^2}(\widetilde{\boldsymbol{\theta}}) \left[\left(\tilde{x}_{jk} - \widetilde{\theta}_{jk} + \left(\frac{\partial^2 \widetilde{f}}{\partial \tilde{x}_{jk}^2}(\widetilde{\boldsymbol{\theta}}) \right)^{-1} \frac{\partial \widetilde{f}}{\partial \tilde{x}_{jk}}(\widetilde{\boldsymbol{\theta}}) \right)^2 - \left(\left(\frac{\partial^2 \widetilde{f}}{\partial \tilde{x}_{jk}^2}(\widetilde{\boldsymbol{\theta}}) \right)^{-1} \frac{\partial \widetilde{f}}{\partial \tilde{x}_{jk}}(\widetilde{\boldsymbol{\theta}}) \right)^2 \right].$$

Let

$$\operatorname{const}(\boldsymbol{\theta}) := \widetilde{f}(\widetilde{\boldsymbol{\theta}}) - \frac{1}{2} \sum_{(j,k)\in\mathcal{P}} \frac{\partial^2 \widetilde{f}}{\partial \tilde{x}_{jk}^2}(\widetilde{\boldsymbol{\theta}}) \left(\left(\frac{\partial^2 \widetilde{f}}{\partial \tilde{x}_{jk}^2}(\widetilde{\boldsymbol{\theta}}) \right)^{-1} \frac{\partial \widetilde{f}}{\partial \tilde{x}_{jk}}(\widetilde{\boldsymbol{\theta}}) \right)^2,$$

then (4.22) becomes

$$\begin{aligned} \widetilde{f}(\tilde{\mathbf{x}}) &\approx \operatorname{const}(\boldsymbol{\theta}) + \frac{1}{2} \sum_{(j,k)\in\mathcal{P}} \frac{\partial^2 \widetilde{f}}{\partial \tilde{x}_{jk}^2}(\widetilde{\boldsymbol{\theta}}) \left(\tilde{x}_{jk} - \left(\widetilde{\theta}_{jk} - \left(\frac{\partial^2 \widetilde{f}}{\partial \tilde{x}_{jk}^2}(\widetilde{\boldsymbol{\theta}}) \right)^{-1} \frac{\partial \widetilde{f}}{\partial \tilde{x}_{jk}}(\widetilde{\boldsymbol{\theta}}) \right) \right)^2 \\ &= \operatorname{const}(\boldsymbol{\theta}) + \frac{1}{2} \sum_{j=1}^{\ell} \frac{\partial^2 \widetilde{f}}{\partial \tilde{x}_{j,m_j}^2}(\widetilde{\boldsymbol{\theta}}) \left(\tilde{x}_{j,m_j} - \left(\widetilde{\theta}_{j,m_j} - \left(\frac{\partial^2 \widetilde{f}}{\partial \tilde{x}_{j,m_j}^2}(\widetilde{\boldsymbol{\theta}}) \right)^{-1} \frac{\partial \widetilde{f}}{\partial \tilde{x}_{j,m_j}}(\widetilde{\boldsymbol{\theta}}) \right) \right)^2 \\ &\quad + \frac{1}{2} \sum_{k=2}^{m} \sum_{j=\ell_k}^{L_{k-1}} \frac{\partial^2 \widetilde{f}}{\partial \tilde{x}_{jk}^2}(\widetilde{\boldsymbol{\theta}}) \left(\tilde{x}_{jk} - \left(\widetilde{\theta}_{jk} - \left(\frac{\partial^2 \widetilde{f}}{\partial \tilde{x}_{jk}^2}(\widetilde{\boldsymbol{\theta}}) \right)^{-1} \frac{\partial \widetilde{f}}{\partial \tilde{x}_{jk}}(\widetilde{\boldsymbol{\theta}}) \right) \right)^2. \end{aligned}$$

Fix k such that $1 < k \leq m$. We note the minimization of the quadratic approximation of the objective function with the feasible set $\widetilde{\Theta}$ yields the isotonic regression of $(x_{jk})_{j=\ell_k}^{L_{k-1}}$ with weights $\frac{\partial^2 \widetilde{f}}{\partial \tilde{x}_{jk}^2}(\widetilde{\boldsymbol{\theta}})$. Let

$$\tilde{v}_{jk}(\boldsymbol{\theta}) := \frac{\partial^2 \widetilde{f}}{\partial \tilde{x}_{jk}^2}(\widetilde{\boldsymbol{\theta}}) = n \sum_{k'=k}^{M_j} \exp(\theta_{jk'}) \qquad \text{and} \qquad \begin{aligned} \tilde{\gamma}_{jk}(\boldsymbol{\theta}) &:= \widetilde{\theta}_{jk} - \tilde{v}_{jk}^{-1} \frac{\partial \widetilde{f}}{\partial \tilde{x}_{jk}}(\widetilde{\boldsymbol{\theta}}) \\ &= T_{jk}(\boldsymbol{\theta}) + \tilde{v}_{jk}(\boldsymbol{\theta})^{-1} \underline{w}_{jk} - 1, \end{aligned}$$

Then the proposed search direction is the solution to the following optimization problem

$$\psi(\boldsymbol{\theta}) := \operatorname*{arg\,min}_{\tilde{\mathbf{x}}\in\widetilde{\Theta}} \sum_{(j,k)\in\mathcal{P}} \tilde{v}_{jk}(\boldsymbol{\theta}) \left(\tilde{x}_{jk} - \tilde{\gamma}_{jk}(\boldsymbol{\theta}) \right)^2.$$

The above problem is solvable by using PAVA algorithm on $(x_{jk})_{j=\ell_k}^{L_{k-1}}$ for each $1 < k \leq m$.

After the algorithm proposes a search direction, the algorithm then computes a new point for the next iteration. Let $\boldsymbol{\theta}_s$ be the output of the algorithm at s-th iteration. Then, the new point is $\boldsymbol{\theta}_{s+1} = (1-t)\boldsymbol{\theta}_s + t\psi(\boldsymbol{\theta}_s)$, which remains in $\widetilde{\Theta}$ due to its convexity property. The goal is to find the best t, such that the objective function f in (4.19) decreases. We omit the details of how to find the suitable t.

5 Minimization of The Empirical Risk with The Likelihood Ratio Order Constraint

In Chapter 4, Mösching and Dümbgen (2024) estimate the conditional distributions by maximizing an empirical likelihood function. In this chapter, we propose another method to estimate conditional distribution functions. This method uses the empirical risk as the objective function with CRPS as the loss function. We use CRPS as the scoring rule because it is proper and is strictly proper when the probability measure has a finite first moment. We then require the minimizer of the empirical risk to satisfy the likelihood ratio order constraint.

To this end, we start with computing the empirical risk and formulating the constraints (Section 5.1). It turns out that the objective function is convex, and we give a counterexample that proves the non-convexity of the feasible set (Section 5.2). To ensure that the feasible set is convex, we apply a log transformation on the estimand with a price of losing the convexity property of the objective function as will be seen in Section 5.3.

Before we start, let us define and recall several notions and notations that we will use throughout this chapter. Similar to at the beginning of Chapter 4, we use $\mathcal{D}_n = (X_i, Y_i)_{i=1}^n$ to denote a data set. For any $1 \leq i \leq n$, let $Y(X_i) = Y_i := [Y|X = X_i]$ and suppose it has a finite first moment and it follows an unknown distribution $G_{Y|X_i}$. We further require that for any $x_i \leq x_j$ and $1 \leq i \neq j \leq n$, we have

$$Y(X_i) \leq_{\mathrm{lr}} Y(X_j).$$

We will construct the "best" estimator $F_{Y|X_i}$ of $G_{Y|X_i}$, in a sense that it minimizes the excepted CRPS using the estimator $F_{Y|X_i}$, while $Y(X_i)$ has a true distribution $G_{Y|X_i}$, for each $i = 1, \dots, n$. Formulated differently, we require $F_{Y|X_i}$ to have the following property

$$\mathbb{E}_{G_{Y|X_i}}[\mathrm{CRPS}(F_{Y|X_i}, Y(X_i))] \approx \mathbb{E}_{G_{Y|X_i}}[\mathrm{CRPS}(G_{Y|X_i}, Y(X_i))], \quad \text{for each } i = 1, \dots, n.$$

Instead of taking the expectation of $\mathrm{CRPS}(F_{Y|X}, Y(X))$ w.r.t. the distribution $G_{Y|X}$, we compute the risk w.r.t. to the joint distribution of (X, Y). Assume that $(X, Y) \sim P$ and $\mathcal{D}_n$ consists of n realizations of (X, Y). Let P_X be the marginal distribution of X. Then,

$$\begin{aligned}
\mathbb{E}_{(X,Y)\sim P}[\mathrm{CRPS}(F_{Y|X}, Y(X)] &= \mathbb{E}_{P_X}\left[\mathbb{E}_{G_{Y|X}}\left[\mathrm{CRPS}(F_{Y|X}, Y(X))\right]\right] \\
&\geq \mathbb{E}_{P_X}\left[\mathbb{E}_{G_{Y|X}}\left[\mathrm{CRPS}(G_{Y|X}, Y(X))\right]\right] \\
&= \mathbb{E}_{(X,Y)\sim P}[\mathrm{CRPS}(G_{Y|X}, Y(X))].
\end{aligned}$$

Due to the (strict) propriety property of CRPS, the inequality above is equal if and only if $F_{Y|X} = G_{Y|X}$. Since the joint distribution P is unknown, we will use the empirical risk to estimate the expectation, i.e.

$$\mathbb{E}_{(X,Y)\sim P}[\mathrm{CRPS}(F_{Y|X}, Y(X)] \approx \frac{1}{n}\sum_{i=1}^{n} \mathrm{CRPS}(F_{Y|X_i}, Y(X_i)) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{CRPS}(F_{Y|X_i}, Y_i). \tag{5.1}$$

The above quantity will be computed by choosing a specific model for $F_{Y|X_i}$ of each $1 \leq i \leq n$, which allows us to formulate the function that we wish to minimize.

5.1. Formulating the objective function and the constraints

The model choice of the conditional distributions is the same as in Mösching and Dümbgen (2024). Before the estimation procedure, we sort the covariate and the response values in $\mathcal{D}_n$ from the smallest to the largest value:

$$\{X_1, \ldots, X_n\} = \{x_1, \ldots, x_\ell\} \qquad \text{and} \qquad \{Y_1, \ldots, Y_n\} = \{y_1, \ldots, y_m\},$$

where $(x_j)_{j=1}^{\ell}$ and $(y_k)_{k=1}^{m}$ are strictly increasing sequences. Let $w_{jk} := \#\{i : (X_i, Y_i) = (x_j, y_k)\}$. We then model the conditional distribution by assuming that $F_{Y|x_j}$ has a support $\{y_1, \ldots, y_m\}$ with probability mass function

$$f(y_k|x_j) := q_{jk} \qquad \text{for all } 1 \leq j \leq \ell, 1 \leq k \leq m.$$

Therefore, the conditional distributions for any $1 \leq j \leq \ell$ are

$$F(y|x_j) = \sum_{k=1}^{m} q_{jk} \mathbb{1}\{y_k \leq y\},$$

which means that the empirical risk in (5.1) becomes

$$\frac{1}{n} \sum_{j=1}^{\ell} \sum_{k=1}^{m} w_{jk} \mathrm{CRPS}(F_{Y|x_j}, y_k).$$

We denote the expression above with $R(\mathbf{q})$, where $\mathbf{q}$ is a vector of size $\ell \cdot m$ and

$$\mathbf{q} = (q_{11}, \ldots, q_{1m}, \ldots, q_{j1}, \ldots, q_{jm}, \ldots, q_{\ell 1}, \ldots, q_{\ell m})^\top.$$

We compute the score when we predict the probability of observing y_k using $F_{Y|x_j}$. For any $1 \leq j \leq \ell$ and $1 \leq k \leq m$, we have

$$\mathrm{CRPS}(F_{Y|x_j}, y_k) = \int_{\mathbb{R}} \left[\sum_{k'=1}^{m} q_{jk'} \mathbb{1}\{y_{k'} \leq z\} - \mathbb{1}\{y_k \leq z\} \right]^2 dz.$$

We write $\mathbb{R}$ as a set of disjoint intervals:

$$\mathbb{R} = (-\infty, y_1) \cup \left(\bigcup_{k=1}^{m-1} [y_k, y_{k+1}) \right) \cup [y_m, \infty).$$

Let $M > 0$, $F_{jt} := \sum_{k'=1}^{t} q_{jk'}$ and $\Delta y_t = y_{t+1} - y_t$, we then have

$$\begin{aligned}
\mathrm{CRPS}(F_{Y|x_j}, y_k) &= \sum_{t=1}^{m-1} \int_{y_t}^{y_{t+1}} \left[\sum_{k'=1}^{m} q_{jk'} \mathbb{1}\{y_{k'} \leq z\} - \mathbb{1}\{y_k \leq z\} \right]^2 dz + \lim_{M \to \infty} \int_{y_m}^{y_m + M} \left[F_{jm} - 1 \right]^2 dz \\
&= \sum_{t=1}^{m-1} \left\{ \left[F_{jt} - \mathbb{1}\{t \geq k\} \right]^2 \Delta y_t \right\} + \lim_{M \to \infty} M \left[F_{jm} - 1 \right]^2 \\
&= \sum_{t=1}^{k-1} F_{jt}^2 \Delta y_t + \sum_{t=k}^{m-1} (1 - F_{jt})^2 \Delta y_t + \lim_{M \to \infty} M \left[F_{jm} - 1 \right]^2 \\
&= \sum_{t=1}^{m-1} F_{jt}^2 \Delta y_t - 2 \sum_{t=k}^{m-1} F_{jt} \Delta y_t + (y_m - y_k) + \lim_{M \to \infty} M \left[F_{jm} - 1 \right]^2.
\end{aligned}$$

Let $M > 0$ be sufficiently large. The empirical risk is then

$$R(\mathbf{q}) = \frac{1}{n}\sum_{j=1}^{\ell}\sum_{k=1}^{m} w_{jk}\left\{\sum_{t=1}^{m-1} F_{jt}^2 \Delta y_t - 2\sum_{t=k}^{m-1} F_{jt}\Delta y_t + (y_m - y_k)\right\} + \frac{M}{n}\sum_{j=1}^{\ell}\sum_{k=1}^{m} w_{jk}[F_{jm} - 1]^2.$$

The optimization problem is therefore

$$\min_{\mathbf{q}\in\mathcal{C}} \sum_{j=1}^{\ell}\sum_{k=1}^{m} w_{jk}\left\{\sum_{t=1}^{m-1} F_{jt}^2 \Delta y_t - 2\sum_{t=k}^{m-1} F_{jt}\Delta y_t\right\} + M\sum_{j=1}^{\ell} w_{j+}[F_{jm} - 1]^2, \tag{5.2}$$

where $w_{j+} = \sum_{k=1}^{m} w_{jk}$ and $\mathcal{C}$ is some feasible region.

At last, we determine the feasible region $\mathcal{C}$. For that, we discuss the constraints we impose for this minimization problem. Firstly, we want that for any $1 \leq j \leq \ell$ and $1 \leq k \leq \ell$, the function F_{jk} is a well-defined conditional distribution function, that is, $F_{jm} = 1$ for all $1 \leq j \leq \ell$. This constraint is accounted for by the second term of $R(\mathbf{q})$. If $\mathbf{q} \in [0,\infty)^{\ell\cdot m}$ and $M > 0$ is large, then

$$F_{jm} = \sum_{k'=1}^{m} q_{jk'} \approx 1, \qquad \forall 1 \leq j \leq \ell,$$

and $q_{jk} \in [0,1]$ for all $1 \leq j \leq \ell, 1 \leq k \leq m$. Furthermore, we add the constraint to ensure that the likelihood ratio is increasing, i.e.

$$q_{j_1k_1}q_{j_2k_2} \geq q_{j_2k_1}q_{j_1k_2} \ \forall 1 \leq j_1 < j_2 \leq \ell, \forall 1 \leq k_1 < k_2 \leq m.$$

This means that the feasible region is

$$\mathcal{C} := \left\{\mathbf{q} \in [0,\infty)^{\ell\cdot m} : q_{j_1k_1}q_{j_2k_2} \geq q_{j_2k_1}q_{j_1k_2} \ \forall 1 \leq j_1 < j_2 \leq \ell, \forall 1 \leq k_1 < k_2 \leq m\right\}.$$

Note that for any $1 \leq j_1 < j_2 \leq \ell$ and $1 \leq k_1 < k_2 \leq m$,

$$q_{j_1k_1}q_{j_2k_2} \geq q_{j_2k_1}q_{j_1k_2} \quad \Leftrightarrow \quad q_{j_1k_1}q_{j_2k_2} - q_{j_2k_1}q_{j_1k_2} \geq 0 \quad \Leftrightarrow \quad \det\begin{pmatrix} q_{j_1k_1} & q_{j_1k_2} \\ q_{j_2k_1} & q_{j_2k_2} \end{pmatrix} \geq 0.$$

Therefore, the set $\mathcal{C}$ may also be formulated as follows

$$\mathcal{C} := \left\{\mathbf{q} \in [0,\infty)^{\ell\cdot m} : \det\begin{pmatrix} q_{j_1k_1} & q_{j_1k_2} \\ q_{j_2k_1} & q_{j_2k_2} \end{pmatrix} \geq 0 \ \forall 1 \leq j_1 < j_2 \leq \ell, \forall 1 \leq k_1 < k_2 \leq m\right\}. \tag{5.3}$$

In the subsequent section, we investigate the convexity of the objective function and the feasible set.

5.2. The convexity of the objective function and the feasible set

In an optimization problem, we wish that the objective function is strictly convex and the feasible set is convex. It ensures that if the objective function has a minimum in the feasible set, the minimizer is unique. Therefore, we investigate the convexity of the objective function and the feasible set for our problem. It turns out that the objective function in (5.2) has a minimum and is strictly convex in the domain $[0,\infty)^{\ell\cdot m}$. The convexity property of the objective function is proven in Section 5.2.1. After that, we show that the objective function for the unconstrained problem has a minimum (Section 5.2.2). Unfortunately, the feasible set in (5.3) is a non-convex set, which we illustrate with a counterexample in Section 5.2.3.

5.2.1. The objective function is convex

To show that the objective function (5.2) is convex, we will show that the Hessian matrix is positive definite. We prove this claim by using the LU-decomposition of the Hessian matrix. It will turn out that the pivots of the upper triangular matrix are all positive, which means that all of its eigenvalues are positive.

Let $\phi_M(\mathbf{q})$ be the objective function in (5.2), i.e.

$$\phi_M(\mathbf{q}) := \sum_{j=1}^{\ell}\sum_{k=1}^{m} w_{jk}\left\{\sum_{t=1}^{m-1}\left(\sum_{k'=1}^{t} q_{jk'}\right)^2 \Delta y_t - 2\sum_{t=k}^{m-1}\left(\sum_{k'=1}^{t} q_{jk'}\right)\Delta y_t\right\} + M\sum_{j=1}^{\ell} w_{j+}\left[\sum_{k'=1}^{m} q_{jk'} - 1\right]^2 .$$

We have for any $1 \le j \le \ell$ and $1 \le k \le m-1$,

$$\begin{aligned}
\frac{\partial \phi_M}{\partial q_{jk}} &= 2w_{j+}\left\{\sum_{t=k}^{m-1}\left(\sum_{k'=1}^{t} q_{uk'}\right)\Delta y_t\right\} - 2(y_m - y_k)\sum_{k'=1}^{k} w_{jk'} - 2\sum_{k'=k+1}^{m-1} w_{jk'}(y_m - y_{k'}) \\
&\quad + 2Mw_{j+}\left(\sum_{k'=1}^{m} q_{jk'} - 1\right) \\
&= 2w_{j+}\left\{(y_m - y_k)\left(\sum_{k'=1}^{k-1} q_{jk'} + q_{jk}\right) + \sum_{k'=k+1}^{m-1} q_{jk'}(y_m - y_{k'})\right\} \\
&\quad - 2(y_m - y_k)\sum_{k'=1}^{k} w_{jk'} - 2\sum_{k'=k+1}^{m-1} w_{uk'}(y_m - y_{k'}) + 2Mw_{j+}\left(\sum_{k'=1}^{m} q_{jk'} - 1\right), \\
\frac{\partial \phi_M}{\partial q_{jm}} &= 2Mw_{j+}\left(\sum_{k'=1}^{m} q_{jk'} - 1\right).
\end{aligned}$$

As for the second derivatives, the Hessian matrix is an $\ell \cdot m \times \ell \cdot m$. The diagonal entries are for all $1 \le j \le \ell$ and $1 \le k \le m-1$

$$\begin{aligned}
\frac{\partial^2 \phi_M}{\partial q_{jk}^2} &= 2w_{j+}(y_m - y_k) + 2Mw_{j+} \\
\frac{\partial^2 \phi_M}{\partial q_{jm}^2} &= 2Mw_{j+}
\end{aligned}$$

As for the other entries, first note that for any $1 \le j \ne j' \le \ell$ and $1 \le k, k' \le m$,

$$\frac{\partial^2 \phi_M}{\partial q_{j'k'}\partial q_{jk}} = 0 \qquad \text{and} \qquad \frac{\partial^2 \phi_M}{\partial q_{jk'}\partial q_{jk}} \ne 0.$$

For any $1 \le j \le \ell$, let $1 \le k \ne k' \le m$, then

$$\frac{\partial^2 \phi_M}{\partial q_{jk'}\partial q_{jk}} = \begin{cases} 2w_{j+}(y_m - y_{k'}) + 2Mw_{j+} & \text{if } 1 \le k < k' \le m-1 \\ 2w_{j+}(y_m - y_k) + 2Mw_{j+} & \text{if } 1 \le k' < k \le m-1 \\ 2Mw_{j+} & \text{if } k = m, 1 \le k' \le m-1 \\ 2Mw_{j+} & \text{if } 1 \le k \le m-1, k' = m. \end{cases}$$

This means that the Hessian matrix is a diagonal block matrix, in which each block is an $m \times m$ matrix. The Hessian matrix of ϕ_M is then

$$\nabla^2 \phi_M = \begin{pmatrix} A_{1,(1:m)} & 0 & 0 & \cdots & 0 & 0 \\ 0 & A_{2,(1:m)} & 0 & \cdots & 0 & 0 \\ 0 & 0 & A_{3,(1:m)} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \vdots & A_{\ell,(1:m)} \end{pmatrix}$$

where for any $1 \le j \le \ell$, the matrix $A_{j,(1:m)} \in \mathbb{R}^{m\times m}$ and is defined as follows,

$$\begin{pmatrix} 2w_{j+}(y_m - y_1) + 2Mw_{j+} & 2w_{j+}(y_m - y_2) + 2Mw_{j+} & \cdots & 2w_{j+}(y_m - y_{m-1}) + 2Mw_{j+} & 2Mw_{j+} \\ 2w_{j+}(y_m - y_2) + 2Mw_{j+} & 2w_{j+}(y_m - y_2) + 2Mw_{j+} & \cdots & 2w_{j+}(y_m - y_{m-1}) + 2Mw_{j+} & 2Mw_{j+} \\ 2w_{j+}(y_m - y_3) + 2Mw_{j+} & 2w_{j+}(y_m - y_3) + 2Mw_{j+} & \cdots & 2w_{j+}(y_m - y_{m-1}) + 2Mw_{j+} & 2Mw_{j+} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 2w_{j+}(y_m - y_{m-1}) + 2Mw_{j+} & 2w_{j+}(y_m - y_{m-1}) + 2Mw_{j+} & \cdots & 2w_{j+}(y_m - y_{m-1}) + 2Mw_{j+} & 2Mw_{j+} \\ 2Mw_{j+} & 2Mw_{j+} & \cdots & 2Mw_{j+} & 2Mw_{j+} \end{pmatrix}$$

Note that all of the entries of the matrix above are positive. Furthermore, let $c_i = c_i(j) := 2w_{j+}(y_m - y_i) + 2Mw_{j+}$, then $c_1 > c_2 > \cdots > c_m$. Each block $A_{j,(1:m)}$ is a symmetric matrix and so $\nabla^2\phi_M$ is symmetric. For each $1 \le j \le \ell$, the matrix $A_{j,(1:m)}$ is positive definite and therefore so is $\nabla^2\phi_M$. The positive definiteness of $A_{j,(1:m)}$ is shown in the following lemma.

Lemma 5.2.1. *Let $A \in \mathbb{R}^{m\times m}$ be a matrix. Let a_{ik} denote the entries of A, where for all $1 \le i, k \le m$,*

$$a_{ik} = \begin{cases} c_k & \text{if } i < k, \\ c_i & \text{if } i \ge k, \end{cases}$$

and $(c_k)_{k=1}^m$ is an arbitrary real-valued sequence such that $c_1 > c_2 > \cdots > c_m > 0$. The matrix A is positive definite.

Proof. The matrix A is defined as follows,

$$\mathbf{A} = \begin{pmatrix} c_1 & c_2 & c_3 & \cdots & c_{m-1} & c_m \\ c_2 & c_2 & c_3 & \cdots & c_{m-1} & c_m \\ c_3 & c_3 & c_3 & \cdots & c_{m-1} & c_m \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ c_{m-1} & c_{m-1} & c_{m-1} & \cdots & c_{m-1} & c_m \\ c_m & c_m & c_m & \cdots & c_m & c_m \end{pmatrix}.$$

To show that A is positive definite, we show that A has an LU-decomposition with positive pivots (Meyer, 2023, p. 559). Let

$$\mathbf{L} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ c_2/c_1 & 1 & 0 & \cdots & 0 & 0 \\ c_3/c_1 & c_3/c_2 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ c_{m-1}/c_1 & c_{m-1}/c_2 & c_{m-1}/c_3 & \cdots & 1 & 0 \\ c_m/c_1 & c_m/c_2 & c_m/c_3 & \cdots & c_m/c_{m-1} & 1 \end{pmatrix}$$

and

$$\mathbf{U} = \begin{pmatrix} c_1 & c_2 & c_3 & \cdots & c_{m-1} & c_m \\ 0 & \frac{c_2}{c_1}(c_1 - c_2) & \frac{c_3}{c_1}(c_1 - c_2) & \cdots & \frac{c_{m-1}}{c_1}(c_1 - c_2) & \frac{c_m}{c_1}(c_1 - c_2) \\ 0 & 0 & \frac{c_3}{c_2}(c_2 - c_3) & \cdots & \frac{c_{m-1}}{c_2}(c_2 - c_3) & \frac{c_m}{c_2}(c_2 - c_3) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \frac{c_{m-1}}{c_{m-2}}(c_{m-1} - c_{m-2}) & \frac{c_m}{c_{m-2}}(c_{m-1} - c_{m-2}) \\ 0 & 0 & 0 & \cdots & 0 & \frac{c_m}{c_{m-1}}(c_{m-1} - c_m) \end{pmatrix}.$$

The entries of matrices of L and U are denoted as $(\ell_{ij})_{i,j}$ and $(u_{ij})_{i,j}$ respectively, where

$$\ell_{ij} = \begin{cases} c_i/c_j & \text{if } i \ge j, \\ 0 & \text{if } i < j, \end{cases} \qquad \text{and} \qquad u_{ij} = \begin{cases} c_j & \text{if } i = 1, \\ c_j/c_{i-1}(c_{i-1} - c_i) & \text{if } i \le j, \\ 0 & \text{if } i > j. \end{cases}$$

Now, we compute the entries of the outcome LU. We have for any $1 \le k \le m$,

$$\sum_{j=1}^{m} \ell_{1j}u_{jk} = \ell_{11}u_{1k} = u_{1k} = c_k = a_{1k}$$

If $1 < i \leq m$ and $1 \leq k \leq m$ such that $i < k$, then

$$\begin{aligned}
\sum_{j=1}^{m} \ell_{ij}u_{jk} &= \ell_{i1}u_{1k} + \sum_{j=2}^{m} \ell_{ij}u_{jk} \\
&= \frac{c_i}{c_1}c_k + \sum_{2\leq j\leq i} \ell_{ij}u_{jk} + \sum_{i<j\leq m} \ell_{ij}u_{jk} \\
&= \frac{c_i}{c_1}c_k + \sum_{2\leq j\leq i} \ell_{ij}u_{jk} \\
&= \frac{c_i}{c_1}c_k + \sum_{j=2}^{i} \frac{c_i}{c_j}\frac{c_k}{c_{j-1}}(c_{j-1} - c_j) \\
&= \frac{c_i}{c_1}c_k + \sum_{j=2}^{i-1} \frac{c_i}{c_j}\frac{c_k}{c_{j-1}}(c_{j-1} - c_j) + \frac{c_k}{c_{i-1}}(c_{i-1} - c_i) \\
&= \frac{c_i}{c_1}c_k + c_ic_k\sum_{j=2}^{i-1}\left(\frac{1}{c_j} - \frac{1}{c_{j-1}}\right) + c_k - \frac{c_i}{c_{i-1}}c_k \\
&= \frac{c_i}{c_1}c_k + c_ic_k\left(\frac{1}{c_{i-1}} - \frac{1}{c_1}\right) + c_k - \frac{c_i}{c_{i-1}}c_k = c_k = a_{ik}.
\end{aligned}$$

In case $i \geq k$, we do similar computations as before,

$$\begin{aligned}
\sum_{j=1}^{m} \ell_{ij}u_{jk} &= \frac{c_i}{c_1}c_k + \sum_{2\leq j\leq k} \ell_{ij}u_{jk} \\
&= \frac{c_i}{c_1}c_k + c_ic_k\sum_{j=2}^{k-1}\left(\frac{1}{c_j} - \frac{1}{c_{j-1}}\right) + \frac{c_i}{c_{k-1}}(c_{k-1} - c_k) \\
&= \frac{c_i}{c_1}c_k + c_ic_k\left(\frac{1}{c_{k-1}} - \frac{1}{c_1}\right) + c_i - \frac{c_i}{c_{k-1}}c_k = c_i = a_{ik}.
\end{aligned}$$

Hence, $\mathbf{A} = \mathbf{LU}$. Because $c_1 > c_2 > \cdots > c_m > 0$, all pivots of matrix $\mathbf{U}$ are positive. Therefore, the matrix $\mathbf{A}$ is positive definite. □

5.2.2. The existence and uniqueness of the minimum

The strict convexity property ensures that the objective function has a unique minimum if it exists. In this section, we confirm that the function does have a minimum. The idea is to first show that for any fixed $\mathbf{q} \in [0,\infty)^{\ell\cdot m}$, the function value ϕ_M goes to infinity as we moves increasingly farther from $\mathbf{q}$. Then, we restrict ϕ_M on a bounded domain and use continuity and the compactness of the set to guarantee the existence of the minimum.

The following lemma shows the existence and the uniqueness of the minimum of ϕ_M.

Lemma 5.2.2. *The function $\phi_M(\mathbf{q})$ has a minimum in $[0,\infty)^{\ell m}$ and it is unique.*

Proof. Let $c > 1$ and fix $M > 0$. Choose $\mathbf{q} \in [0,\infty)^{\ell m}$ such that for all $1 \leq j \leq \ell$, we have $\sum_{k=1}^{m} q_{jk} = 1$, then,

$$\begin{aligned}
\phi_M(c\mathbf{q}) = \sum_{j=1}^{\ell}\sum_{k=1}^{m}&\left[w_{jk}\left\{c^2\sum_{t=1}^{m-1}\left(\sum_{k'=1}^{t} q_{jk'}\right)^2 \Delta y_t - 2c\sum_{t=k}^{m-1}\left(\sum_{k'=1}^{t} q_{jk'}\right)\Delta y_t\right\}\right] \\
&+ M\sum_{j=1}^{\ell}\left[w_{j+}\left(\sum_{k'=1}^{m} cq_{jk'} - 1\right)^2\right]
\end{aligned}$$

$$
\begin{aligned}
&\geq \sum_{j=1}^{\ell}\sum_{k=1}^{m}\left[w_{jk}\left\{M\left(\sum_{k'=1}^{m} cq_{jk'}-1\right)^2 - 2c\sum_{t=k}^{m-1}\left(\sum_{k'=1}^{t} q_{jk'}\right)\Delta y_t\right\}\right]\\
&\geq \sum_{j=1}^{\ell}\sum_{k=1}^{m}\left[w_{jk}\left\{M(c-1)^2 - 2c\sum_{t=k}^{m-1}\left(\sum_{k'=1}^{m} q_{jk'}\right)\Delta y_t\right\}\right]\\
&\geq \sum_{j=1}^{\ell}\sum_{k=1}^{m}\left[w_{jk}\left\{M(c-1)^2 - 2c(y_m-y_k)\right\}\right]\\
&\geq \sum_{j=1}^{\ell}\sum_{k=1}^{m}\left[w_{jk}\left\{M(c-1)^2 - 2c(y_m-y_1)\right\}\right] = n(c-1)^2M - 2c(y_m-y_1)\ell m
\end{aligned}
$$

Note that the lower bound of $\widetilde{\phi}_M(c\mathbf{q})$ is independent on the choice of $\mathbf{q}$. Furthermore, this bound goes to infinity as $c \to \infty$. In other words, there exists $c' > 1$ such that for any $c > c'$, we have

$$\widetilde{\phi}_M(c\mathbf{q}) \geq \widetilde{\phi}_M(\mathbf{q}).$$

Now consider the following bounded set

$$\mathcal{S} := \left\{\mathbf{q} \in [0,\infty)^{\ell\cdot m} : \sum_{k=1}^{m} q_{jk} \leq c', \text{ for all } 1 \leq j \leq \ell\right\},$$

and we restrict the domain of $\widetilde{\phi}_M$ in $\mathcal{S}$. Because the function $\widetilde{\phi}_M$ cannot attain its minimum outside this set and the function is continuous, then a minimum exists on the compact set $\mathcal{S}$. The uniqueness is guaranteed due to $\widetilde{\phi}$ being strictly convex in $[0,\infty)^{\ell\cdot m}$.

□

5.2.3. Counterexample illustrating the non-convexity of the feasible set

In this section, we give a counterexample that shows why the feasible set $\mathcal{C}$ in (5.3) is a non-convex set. For the sake of consistency, we replace $[0,\infty)^{\ell\cdot m}$ with $[0,\infty)^{\ell\cdot m}$. So, the feasible set is then

$$\mathcal{C} := \left\{\mathbf{q} \in [0,\infty)^{\ell\cdot m} : \det\begin{pmatrix} q_{j_1k_1} & q_{j_1k_2}\\ q_{j_2k_1} & q_{j_2k_2}\end{pmatrix} \geq 0 \ \forall 1 \leq j_1 < j_2 \leq \ell, \forall 1 \leq k_1 < k_2 \leq m\right\}. \tag{5.4}$$

To prove that a set $\mathcal{C}$ is convex, we want that for any $\mathbf{x},\mathbf{y} \in \mathcal{C}$, we have $(1-t)x + ty \in \mathcal{C}$ for any $t \in [0,1]$. We choose $\ell = m = 2$ and consider the following vectors

$$\mathbf{q} = (0.1, 0.7, 0, 0.2)^\top \qquad \text{and} \qquad \mathbf{p} = (0.4, 0.2, 0.2, 0.1)^\top.$$

Then,

$$\det\begin{pmatrix} 0.1 & 0.7\\ 0 & 0.2\end{pmatrix} = 0.02 \qquad \text{and} \qquad \det\begin{pmatrix} 0.4 & 0.2\\ 0.2 & 0.1\end{pmatrix} = 0,$$

which means that $\mathbf{p},\mathbf{q} \in \mathcal{C}$. Now let $t \in [0,1]$ and consider $(1-t)\mathbf{p} + t\mathbf{q}$, then we have

$$\det\begin{pmatrix} 0.1(1-t) + 0.4t & 0.7(1-t) + 0.2t\\ 0.2t & 0.2(1-t) + 0.1t\end{pmatrix}.$$

For $t = 1/2$, the determinant of the above expression is

$$(0.05 + 0.2)(0.1 + 0.05) - 0.1(0.35 + 0.1) = 0.0375 - 0.045 = -0.0075 < 0.$$

There exists $t \in (0,1)$ such that $(1-t)\mathbf{p} + t\mathbf{q} \notin \mathcal{C}$, therefore the set $\mathcal{C}$ is not convex.

These results imply that the objective function have local minima. Therefore, we decided to transform the feasible set so that it is convex, with a price of changing the parameter of the objective function. In the next section, we will see that the objective function with the transformed parameter loses its strict convexity property.

5.3. Log transformation of the parameter and quadratic form

We have seen in Section 5.2 that the objective function is convex, but the feasible set is non-convex. To fix the non-convexity, we transform the parameters so that the feasible set becomes convex. We then have to estimate the transformed parameter. To ensure that we obtain the original parameter back uniquely, we choose a transformation that is bijective. To this end, we follow a similar method as in Mösching and Dümbgen (2024). This results in estimating the transformed parameter with a price that the objective function loses its convexity property. In this section, we show the truthiness of this statement by producing an example.

5.3.1. Transforming a non-convex set into a convex set

Recall that we wish the minimizer $\mathbf{q}$ of the objective function $\phi_M(\mathbf{q})$ to satisfy

$$q_{j_1k_1}q_{j_2k_2} \geq q_{j_1k_2}q_{j_2k_1} \quad \forall\ 1 \leq j_1 < j_2 \leq \ell,\ 1 \leq k_1 < k_2 \leq m,$$

where $\mathbf{q} \in [0,\infty)^{\ell\cdot m}$. Let $\boldsymbol{\theta} = \log(\mathbf{q}) = (\theta_{jk})_{j,k}$ and take log in both sides of the equation. The inequality above is then equivalent to

$$\theta_{j_1k_1} + \theta_{j_2k_2} - \theta_{j_1k_2} - \theta_{j_2k_1} \geq 0 \quad \forall\ 1 \leq j_1 < j_2 \leq \ell,\ 1 \leq k_1 < k_2 \leq m,$$

where $\boldsymbol{\theta} \in \mathbb{R} \cup \{-\infty\}$. Let $\mathbb{R}_{-\infty} := \mathbb{R} \cup \{-\infty\}$, the transformed feasible set that we are considering now is the set

$$\widetilde{\mathcal{C}} := \left\{\boldsymbol{\theta} \in \mathbb{R}_{-\infty}^{\ell m} : \theta_{j_1k_1} - \theta_{j_1k_2} + \theta_{j_2k_2} - \theta_{j_2k_1} \geq 0,\ \forall\ 1 \leq j_1 < j_2 \leq \ell, 1 \leq k_1 < k_2 \leq m\right\}. \tag{5.5}$$

For convention, we allow writing $\log(0) = -\infty$ and therefore $\exp(-\infty) = 0$. Note that $\mathbf{0} \in \mathcal{C}$, but $\mathbf{0} \notin \widetilde{\mathcal{C}}$. This is because we may obtain an indeterminate form $\infty - \infty$. To allow a vector with entries of infinity as well, we allow $\infty - \infty = -\infty + \infty = 0$ as another convention. Lastly, we also allow the following operations:

$$\begin{aligned}
\infty + c = c + \infty &= \infty & \forall c &\in (-\infty,\infty] \\
-\infty + c = c - \infty &= -\infty & \forall c &\in [-\infty,\infty) \\
c \cdot \infty = \infty \cdot c &= \infty & \forall c &\in (0,\infty] \\
c \cdot \infty = \infty \cdot c &= -\infty & \forall c &\in [-\infty,0) \\
\infty \cdot 0 = 0 \cdot \infty = \frac{c}{\infty} = \frac{c}{-\infty} &= 0 & \forall c &\in (-\infty,\infty)
\end{aligned}$$

Next, we show that $\widetilde{\mathcal{C}}$ is a convex set.

Lemma 5.3.1. *The set $\widetilde{\mathcal{C}} \subset \mathbb{R}^{\ell m}$ is a convex set*

Proof. Take any $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ from $\widetilde{\mathcal{C}}$ and let $t \in [0,1]$. Now define $\boldsymbol{\theta} := (1-t)\boldsymbol{\theta}_1 + t\boldsymbol{\theta}_2$ and denote $\theta_{jk}^{(1)}$ and $\theta_{jk}^{(2)}$ as the entries of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ respectively. Then for any $\forall\ 1 \leq j_1 < j_2 \leq \ell, 1 \leq k_1 < k_2 \leq m$, it is clear that we have

$$(1-t)\left[\theta_{j_1k_1}^{(1)} - \theta_{j_1k_2}^{(1)} + \theta_{j_2k_2}^{(1)} - \theta_{j_2k_1}^{(1)}\right] + t\left[\theta_{j_1k_1}^{(2)} - \theta_{j_1k_2}^{(2)} + \theta_{j_2k_2}^{(2)} - \theta_{j_2k_1}^{(2)}\right] \geq 0.$$

Hence $\boldsymbol{\theta} \in \widetilde{\mathcal{C}}$, and we finish the proof. □

5.3.2. Reparametrization of the objective function

Now that the feasible set is a convex set, it remains to check what happens to the objective function if it is a function of $\boldsymbol{\theta}$. Furthermore, the objective function ϕ_M is a double sums of quadratic function. We will rewrite this function so that it is easier to implement it as a code. In particular, we rewrite ϕ_M in a quadratic form. Besides being easier to implement, we also obtain information about the global minimum of the unconstrained problem. The example that ϕ_M is a non-convex function in $\widetilde{\mathcal{C}}$ is shown here as well.

We start with the empirical risk that we want to minimize:

$$\frac{1}{n}\sum_{j=1}^{\ell}\sum_{k=1}^{m} w_{jk}\mathrm{CRPS}(F_{Y|x_j}.y_k).$$

Fix a large $M > 0$ and recompute the CRPS of $F_{Y|x_j}$ when y_k is observed :

$$\begin{aligned}\mathrm{CRPS}(F_{Y|x_j}, y_k) &= \sum_{t=1}^{m-1}\left\{\left(\sum_{k'=1}^{t} q_{jk'} - \mathbb{1}(t \geq k)\right)^2 (y_{t+1} - y_t)\right\} + M\left[\sum_{k'=1}^{m} q_{jk'} - 1\right]^2 \\ &= \sum_{t=1}^{m-1}\left\{\left(F_{jt}(\mathbf{q}) - \mathbb{1}(t \geq k)\right)^2 (y_{t+1} - y_t)\right\} + \left[F_{jm}(\mathbf{q}) - \mathbb{1}(m \geq k)\right]^2 (y_m + M - y_m) \\ &= \sum_{t=1}^{m}\left\{\left(F_{jt}(\mathbf{q}) - \mathbb{1}(t \geq k)\right)^2 (y_{t+1} - y_t)\right\} \\ &= \sum_{t=1}^{m}\left\{\left(F_{jt}(\mathbf{q}) - \mathbb{1}(t \geq k)\right)^2 \Delta y_t\right\}.\end{aligned}$$

where $F_{jt}(\mathbf{q}) := \sum_{k'=1}^{t} q_{jk}$, $\Delta y_t := y_{t+1} - y_t$ and we set $y_{m+1} := y_m + M$. Now let

$$\mathbf{F}(\mathbf{q}) := (F_{11}(\mathbf{q}), \dots, F_{1m}(\mathbf{q}), \dots, F_{j1}(\mathbf{q}), \dots, F_{jm}(\mathbf{q}), \dots, F_{\ell 1}(\mathbf{q}), \dots, F_{\ell m}(\mathbf{q}))^\top \in \mathbb{R}^{\ell m}$$

and let $\mathbf{F}_j(\mathbf{q}) := (F_{j1}(\mathbf{q}), \dots, F_{jm}(\mathbf{q}))^\top \in \mathbb{R}^m$ for all $1 \leq j \leq \ell$, then

$$\mathbf{F}(\mathbf{q}) = (\mathbf{F}_1(\mathbf{q}), \dots, \mathbf{F}_\ell(\mathbf{q}))^\top.$$

So, the risk function that we want to minimize is

$$R_M(\mathbf{q}) := \sum_{j=1}^{\ell}\sum_{k=1}^{m}\left\{\sum_{t=1}^{m}\left(F_{jt}(\mathbf{q}) - \mathbb{1}(t \geq k)\right)^2 \frac{w_{jk}\Delta y_t}{n}\right\}.$$

Now, the goal is to write the risk function as

$$R_M(\mathbf{q}) = (\mathbf{F}(\mathbf{q}) - \mathbf{c})^\top \mathbf{A}(\mathbf{F}(\mathbf{q}) - \mathbf{c}) + K.$$

First, let us write the summand as a quadratic form. Fix $1 \leq j \leq \ell$ and $1 \leq k \leq m$ and let

$$Q_{jk}(\mathbf{F}) := \sum_{t=1}^{m}\left(F_{jt} - \mathbb{1}(t \geq k)\right)^2 \frac{w_{jk}\Delta y_t}{n}.$$

Let $\mathbf{c}_j(k) := (\mathbb{1}(1 \geq k), \mathbb{1}(2 \geq k), \dots, 1)^\top$ and $\mathbf{a}_j(k) := (\mathbf{0}, \dots, \mathbf{c}_j(k), \dots, \mathbf{0})^\top$, where $\mathbf{c}_j$ is from the $[(j-1)m+1]$-th until jm-th location and $\mathbf{0} \in \mathbb{R}^m$. Let $\mathbf{A}_{jk} \in \mathbb{R}^{\ell m \times \ell m}$ be a matrix and let $\mathbf{C}_{j,(1:m)}(k)$ be an $m \times m$ matrix defined as follows

$$\mathbf{C}_{j,(1:m)}(k) := \begin{pmatrix} (w_{jk}/n)\Delta y_1 & 0 & \cdots & 0 \\ 0 & (w_{jk}/n)\Delta y_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & (w_{jk}/n)\Delta y_m \end{pmatrix}.$$

The matrix $\mathbf{A}_{jk}$ is then

$$\mathbf{A}_{jk} := \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & \mathbf{C}_{j,(1:m)}(k) & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 \end{pmatrix},$$

where $\mathbf{C}_{j,(1:m)}(k)$ is at the j-th diagonal block. So then

$$R_M(\mathbf{q}) = \sum_{j=1}^{\ell}\sum_{k=1}^{m}(\mathbf{F}(\mathbf{q}) - \mathbf{a}_j(k))^\top \mathbf{A}_{jk}(\mathbf{F}(\mathbf{q}) - \mathbf{a}_j(k)).$$

Computing the above equation so that we can complete the squares:

$$\begin{aligned} R_M(\mathbf{q}) &= \sum_{j=1}^{\ell}\sum_{k=1}^{m} \mathbf{F}(\mathbf{q})^\top \mathbf{A}_{jk}\mathbf{F}(\mathbf{q}) - 2\mathbf{a}_j(k)^\top \mathbf{A}_{jk}\mathbf{F}(\mathbf{q}) + \mathbf{a}_j(k)^\top \mathbf{A}_{jk}\mathbf{a}_j(k) \\ &= \mathbf{F}(\mathbf{q})^\top\left(\sum_{j=1}^{\ell}\sum_{k=1}^{m}\mathbf{A}_{jk}\right)\mathbf{F}(\mathbf{q}) - 2\left(\sum_{j=1}^{\ell}\sum_{k=1}^{m}\mathbf{a}_j(k)^\top\mathbf{A}_{jk}\right)\mathbf{F}(\mathbf{q}) + \sum_{j=1}^{\ell}\sum_{k=1}^{m}\mathbf{a}_j(k)^\top\mathbf{A}_{jk}\mathbf{a}_j(k). \end{aligned}$$

Let

$$\mathbf{A} := \sum_{j=1}^{\ell}\sum_{k=1}^{m}\mathbf{A}_{jk},$$

$$\mathbf{c} := \sum_{j=1}^{\ell}\sum_{k=1}^{m}\mathbf{A}_{jk}\mathbf{a}_j(k),$$

$$K := \sum_{j=1}^{\ell}\sum_{k=1}^{m}\mathbf{a}_j(k)^\top\mathbf{A}_{jk}\mathbf{a}_j(k) - \mathbf{c}^\top\mathbf{A}^{-1}\mathbf{c}.$$

Note that matrix $\mathbf{A} \in \mathbb{R}^{\ell m\times \ell m}$ is defined as followed

$$\mathbf{A} = \sum_{k=1}^{m}\begin{pmatrix} \mathbf{C}_{1,(1:m)}(k) & 0 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \mathbf{C}_{2,(1:m)}(k) & 0 & \cdots & 0 & \cdots & 0 \\ 0 & 0 & \mathbf{C}_{3,(1:m)}(k) & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & \mathbf{C}_{j,(1:m)}(k) & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & \mathbf{C}_{\ell,(1:m)}(k) \end{pmatrix},$$

where for any $1 \le j \le \ell$,

$$\sum_{k=1}^{m}\mathbf{C}_{j,(1:m)}(k) = \begin{pmatrix} (w_{j+}/n)\Delta y_1 & 0 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & (w_{j+}/n)\Delta y_2 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & 0 & (w_{j+}/n)\Delta y_3 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & (w_{j+}/n)\Delta y_{k'} & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & (w_{j+}/n)\Delta y_m \end{pmatrix}.$$

This matrix is invertible since it is a diagonal matrix and the entries are positive. Hence,

$$R(\mathbf{q}) = (\mathbf{F}(\mathbf{q}) - \mathbf{A}^{-1}\mathbf{c})^\top\mathbf{A}(\mathbf{F}(\mathbf{q}) - \mathbf{A}^{-1}\mathbf{c}) - \mathbf{c}^\top\mathbf{A}^{-1}\mathbf{c} + \sum_{j=1}^{\ell}\sum_{k=1}^{m}\mathbf{a}_j(k)^\top\mathbf{A}_{jk}\mathbf{a}_j(k). \tag{5.6}$$

Lastly, note that we can write $\mathbf{F}(\mathbf{q})$ as a product of an invertible matrix times a vector $\mathbf{q}$. Indeed, for any $1 \le j \le \ell$,

$$\mathbf{F}_j(\mathbf{q}) = \begin{pmatrix} F_{j1}(\mathbf{q}) \\ F_{j2}(\mathbf{q}) \\ F_{j3}(\mathbf{q}) \\ \vdots \\ F_{j,m-1}(\mathbf{q}) \\ F_{jm}(\mathbf{q}) \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ 1 & 1 & 0 & \cdots & 0 & 0 \\ 1 & 1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 1 & 1 & \cdots & 1 & 0 \\ 1 & 1 & 1 & \cdots & 1 & 1 \end{pmatrix}\begin{pmatrix} q_{j1} \\ q_{j2} \\ q_{j3} \\ \vdots \\ q_{j,m-1} \\ q_{jm} \end{pmatrix} =: \mathbf{L}_j\mathbf{q}_j$$

where $\mathbf{L}_j$ is $m \times m$ lower-triangular matrix and $\mathbf{q}_j$ is a vector of length m. Therefore,

$$\mathbf{F}(\mathbf{q}) = \begin{pmatrix} \mathbf{F}_1(\mathbf{q}) \\ \mathbf{F}_2(\mathbf{q}) \\ \vdots \\ \mathbf{F}_\ell(\mathbf{q}) \end{pmatrix} = \begin{pmatrix} \mathbf{L}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{L}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{L}_\ell \end{pmatrix} \begin{pmatrix} \mathbf{q}_1 \\ \mathbf{q}_2 \\ \vdots \\ \mathbf{q}_\ell \end{pmatrix} =: \mathbf{L}\mathbf{q},$$

where the diagonal block matrix $\mathbf{L} \in \mathbb{R}^{\ell m \cdot \ell m}$. The matrix $\mathbf{L}$ is still invertible because it is a diagonal block matrix in which each block is invertible. We can therefore write (5.6) as follows

$$\begin{aligned} R_M(\mathbf{q}) &= (\mathbf{L}\mathbf{q} - \mathbf{L}\mathbf{L}^{-1}\mathbf{A}^{-1}\mathbf{c})^\top \mathbf{A}(\mathbf{L}\mathbf{q} - \mathbf{L}\mathbf{L}^{-1}\mathbf{A}^{-1}\mathbf{c}) + K \\ &= (\mathbf{q} - \mathbf{L}^{-1}\mathbf{A}^{-1}\mathbf{c})^\top \mathbf{L}^\top \mathbf{A}\mathbf{L}(\mathbf{q} - \mathbf{L}^{-1}\mathbf{A}^{-1}\mathbf{c}) + K. \end{aligned}$$

As a function of $\boldsymbol{\theta}$, we then have

$$R_M(\boldsymbol{\theta}) = (e^{\boldsymbol{\theta}} - \mathbf{L}^{-1}\mathbf{A}^{-1}\mathbf{c})^\top \mathbf{L}^\top \mathbf{A}\mathbf{L}(e^{\boldsymbol{\theta}} - \mathbf{L}^{-1}\mathbf{A}^{-1}\mathbf{c}) + K.$$

The function $R_M(\boldsymbol{\theta})$ is not in a quadratic form anymore because $\boldsymbol{\theta} \mapsto e^{\boldsymbol{\theta}}$ is a non-linear transformation of $\boldsymbol{\theta}$.

The function $R_M(\boldsymbol{\theta})$ is furthermore not convex as illustrated in Figure 5.1. We generate $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^4$ randomly, where each entry of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are independently sampled from $\mathcal{N}(-3, 0)$ and standard normal distribution respectively. The entries of these vectors should satisfy the likelihood ratio order constraint

$$\theta_{11} - \theta_{12} + \theta_{22} - \theta_{21} \geq 0.$$

From this simulation, we discover an example that shows $R_M(\boldsymbol{\theta})$ is not convex. Let $\widetilde{\phi}_M(\boldsymbol{\theta}) := R_M(\boldsymbol{\theta}) - K$,

$$\boldsymbol{\theta}_1 := (-2.6, -4.1, -3.1, -3.5)^\top \qquad \text{and} \qquad \boldsymbol{\theta}_2 := (-0.3, -0.3, -2, -1.4)^\top.$$

Then $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ satisfy the likelihood ratio order constraint, but there exists $\lambda \in [0, 1]$ such that

$$\widetilde{\phi}_M((1-\lambda)\boldsymbol{\theta}_1 + \lambda\boldsymbol{\theta}_2) > (1-\lambda)\widetilde{\phi}_M(\boldsymbol{\theta}_1) + \lambda\widetilde{\phi}_M(\boldsymbol{\theta}_2).$$

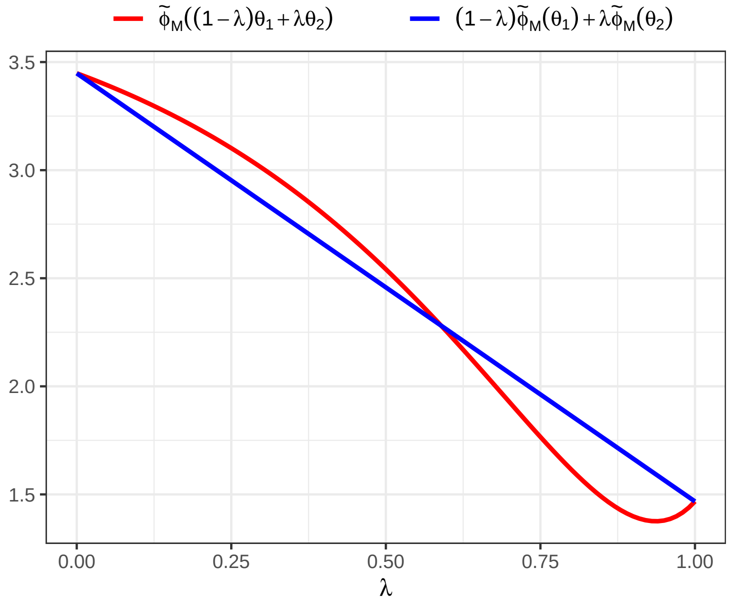

Figure 5.1: A plot illustrating why $R_M(\boldsymbol{\theta})$ is a non-convex function. Define $\widetilde{\phi}_M(\boldsymbol{\theta}) := R_M(\boldsymbol{\theta}) - K$. The red and blue curve represent the function $\lambda \mapsto \widetilde{\phi}_M((1-\lambda)\boldsymbol{\theta}_1 + \lambda\boldsymbol{\theta}_2)$ and $\lambda \mapsto \widetilde{\phi}_M(\boldsymbol{\theta}_1) + \lambda\widetilde{\phi}_M(\boldsymbol{\theta}_2)$ respectively. We choose $\boldsymbol{\theta}_1 := (-2.6, -4.1, -3.1, -3.5)^\top$ and $\boldsymbol{\theta}_2 := (-0.3, -0.3, -2, -1.4)^\top$.

6

Numerical Minimization of The Empirical Risk: The Gradient Projection Method

In Chapter 5, we have shown that the objective function is convex, but the feasible set is non-convex. After transforming the parameters into a logarithmic scale, we achieved the convexity of the feasible set. Unfortunately, the objective function with the transformed parameters loses its convexity property. Consequently, applying a numerical optimization method to our minimization problem no guarantees a global minimizer. Additionally, due to a large number of constraints, the minimization problem is impractical to program.

Despite these problems, we still attempt to numerically solve the objective function with the likelihood ratio order constraint. The impracticality problem can be fixed by reducing the number of constraints drastically. This is due to the linear dependence of the constraints (Section 6.1). We will see that the constraint is rewritten as a matrix inequality. Due to the presence of this inequality, we solve the minimization problem by the gradient projection method described in Section 6.2. In Section 6.3, we modify the algorithm in Section 6.2 to sequentially approximates the conditional distribution functions. In Section 6.4, we do a small simulation study where we visually compare the performance of the proposed algorithms, compared to the one proposed by Mösching and Dümbgen (2024).

6.1. Reduction of the number of constraints

Recall from Chapter 5 that the feasible set $\widetilde{\mathcal{C}}$ of the minimization problem is defined in (5.5). For later computation, we now let $\widetilde{\mathcal{C}}$ as follows

$$\widetilde{\mathcal{C}} := \left\{ \boldsymbol{\theta} \in \mathbb{R}_{-\infty}^{\ell m} : -\theta_{j_1 k_1} + \theta_{j_1 k_2} + \theta_{j_2 k_1} - \theta_{j_2 k_2} \le 0, \ \forall\ 1 \le j_1 < j_2 \le \ell, 1 \le k_1 < k_2 \le m \right\}.$$

Because the constraints are linear combinations of $\boldsymbol{\theta}$, we can express the constraints as a matrix inequality, which has a size of $\binom{\ell}{2}\binom{m}{2} \times \ell m$. Due to memory limits in a computer, it is not possible to store a large matrix. For instance, suppose $\ell = 100$ and $m = 100$, then `R` requires approximately 1800 Gb memories to store the matrix [1]. In this section, we show that the number of constraints is reducible to a matrix of size $(\ell - 1)(m - 1) \times \ell m$.

Let us start with an example of a matrix that represents $\widetilde{\mathcal{C}}$ with small ℓ and m.

Example 6.1.1. Let $\ell = 2$ and $m = 4$, then the matrix inequality that represents $\widetilde{\mathcal{C}}$ is of size 6×8.

[1] In `R`, we can use `object.size()` to approximate the memory size needed to store an `R`-object.

Below is the matrix inequality

$$\begin{pmatrix} -1 & 1 & 0 & 0 & 1 & -1 & 0 & 0 \\ -1 & 0 & 1 & 0 & 1 & 0 & -1 & 0 \\ -1 & 0 & 0 & 1 & 1 & 0 & 0 & -1 \\ 0 & -1 & 1 & 0 & 0 & 1 & -1 & 0 \\ 0 & -1 & 0 & 1 & 0 & 1 & 0 & -1 \\ 0 & 0 & -1 & 1 & 0 & 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} \theta_{11} \\ \theta_{12} \\ \theta_{13} \\ \theta_{14} \\ \theta_{21} \\ \theta_{22} \\ \theta_{23} \\ \theta_{24} \end{pmatrix} \leq \mathbf{0}.$$

Let $\mathbf{T}$ be the matrix above, then note that $\mathbf{T}$ is not a full-rank matrix. Let $\mathbf{R}_u$ be the row vector of the u-th row of matrix $\mathbf{T}$. Then

$$\mathbf{R}_2 = \mathbf{R}_1 + \mathbf{R}_4; \qquad \mathbf{R}_3 = \mathbf{R}_1 + \mathbf{R}_5; \qquad \mathbf{R}_5 = \mathbf{R}_4 + \mathbf{R}_6.$$
$$\Updownarrow$$
$$\mathbf{R}_3 = \mathbf{R}_1 + \mathbf{R}_4 + \mathbf{R}_6;$$

Now define a matrix $\widetilde{\mathbf{T}} \in \mathbb{R}^{3\times 8}$, in which each row consists of entries of $\mathbf{R}_1, \mathbf{R}_4$ and $\mathbf{R}_6$:

$$\widetilde{\mathbf{T}} = \begin{pmatrix} -1 & 1 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 & 1 & -1 \end{pmatrix}.$$

Then $\widetilde{\mathbf{T}}$ is a full-rank matrix. If $\widetilde{\mathbf{T}}\boldsymbol{\theta} \leq \mathbf{0}$, then $\mathbf{T}\boldsymbol{\theta} \leq \mathbf{0}$; and vice versa. Let $\langle \mathbf{x}, \mathbf{y} \rangle$ be the usual inproduct in $\mathbb{R}^{\ell m}$. Note that the converse of the statement follows immediately. Suppose $\widetilde{\mathbf{T}}\boldsymbol{\theta} \leq \mathbf{0}$, then for $u \in \{1, 4, 6\}$ we have $\langle \mathbf{R}_u^\top, \boldsymbol{\theta} \rangle \leq 0$. Furthermore,

$$\begin{aligned} \langle \mathbf{R}_2^\top, \boldsymbol{\theta} \rangle &= \langle \mathbf{R}_1^\top, \boldsymbol{\theta} \rangle + \langle \mathbf{R}_4^\top, \boldsymbol{\theta} \rangle \leq 0, \\ \langle \mathbf{R}_3^\top, \boldsymbol{\theta} \rangle &= \langle \mathbf{R}_1^\top, \boldsymbol{\theta} \rangle + \langle \mathbf{R}_4^\top, \boldsymbol{\theta} \rangle + \langle \mathbf{R}_6^\top, \boldsymbol{\theta} \rangle \leq 0, \\ \langle \mathbf{R}_5^\top, \boldsymbol{\theta} \rangle &= \langle \mathbf{R}_4^\top, \boldsymbol{\theta} \rangle + \langle \mathbf{R}_6^\top, \boldsymbol{\theta} \rangle \leq 0, \end{aligned}$$

Therefore, we can replace the matrix $\mathbf{T}$ with $\widetilde{\mathbf{T}}$, which has a smaller number of rows and it is a full-rank matrix. The matrix inequality $\widetilde{\mathbf{T}}\boldsymbol{\theta} \leq \mathbf{0}$ is equivalent to the following constraint for $\boldsymbol{\theta}$,

$$-\theta_{j-1,k-1} + \theta_{j-1,k} + \theta_{j,k-1} - \theta_{j,k} \leq 0 \quad \forall 1 < j \leq \ell, 1 < k \leq m.$$

△

Let

$$\widetilde{\mathcal{C}}_{\text{red.}} := \left\{ \boldsymbol{\theta} \in \mathbb{R}_{-\infty}^{\ell m} : -\theta_{j-1,k-1} + \theta_{j-1,k} + \theta_{j,k-1} - \theta_{j,k} \leq 0 \quad \forall 1 < j \leq \ell, 1 < k \leq m \right\},$$

for a general ℓ and m. Following from Example 6.1.1, we have that the constraint in $\widetilde{\mathcal{C}}$ is equivalent to the constraint in $\widetilde{\mathcal{C}}_{\text{red.}}$. This has been proved in Mösching and Dümbgen (2024) for $\boldsymbol{\theta} \in \mathbb{R}^{\mathcal{P}}$. In our case, we have $\boldsymbol{\theta} \in \mathbb{R}_{-\infty}^{\ell m}$, and the constraints are still equivalent by following the arguments from the aforementioned article. This implies that the number of constraints are reduced from $\binom{\ell}{2}\binom{m}{2}$ to $(\ell - 1)(m - 1)$ number of constraints. We state this result in the following lemma.

Lemma 6.1.2. *Let* $\boldsymbol{\theta} \in \mathbb{R}_{-\infty}^{\ell m}$. *Then*

$$-\theta_{j_1 k_1} + \theta_{j_1 k_2} + \theta_{j_2 k_1} - \theta_{j_2 k_2} \leq 0, \ \forall \ 1 \leq j_1 < j_2 \leq \ell, 1 \leq k_1 < k_2 \leq m \tag{6.1}$$

is equivalent to

$$-\theta_{j-1,k-1} + \theta_{j-1,k} + \theta_{j,k-1} - \theta_{j,k} \leq 0 \quad \forall 1 < j \leq \ell, 1 < k \leq m. \tag{6.2}$$

Proof. Take any $\boldsymbol{\theta} \in \mathbb{R}_{-\infty}^{\ell m}$. If $\boldsymbol{\theta}$ satisfies (6.1), then it is immediate that $\boldsymbol{\theta}$ satisfies (6.2). Indeed, the inequality in (6.2) is still true by letting $j_1 = j - 1, j_2 = j$ and $k_1 = k - 1$ and $k_2 = k$ for any $1 < j \leq \ell$ and $1 < k \leq m$.

Now suppose that $\boldsymbol{\theta}$ satisfies (6.2) and let $1 \le j_1 < j_2 \le \ell$ and $1 \le k_1 < k_2 \le m$ be arbitrary. Let $j \in \{j_1 + 1, \dots, j_2\}$ and $k \in \{k_1 + 1, \dots, k_2\}$. Then we have

$$-\theta_{j-1,k_1} + \theta_{j-1,k_2} + \theta_{j,k_1} - \theta_{j,k_2} = \sum_{k=k_1+1}^{k_2} -\theta_{j-1,k-1} + \theta_{j-1,k} + \theta_{j,k-1} - \theta_{j,k} \le 0. \tag{6.3}$$

The inequality is true because of (6.2) and $j \in \{j_1 + 1, \dots, j_2\}$ and $k \in \{k_1 + 1, \dots, k_2\}$. Next, by (6.3),

$$-\theta_{j_1 k_1} + \theta_{j_1 k_2} + \theta_{j_2 k_1} - \theta_{j_2 k_2} = \sum_{j=j_1+1}^{j_2} -\theta_{j-1,k_1} + \theta_{j-1,k_2} + \theta_{j,k_1} - \theta_{j,k_2} \le 0.$$

Hence if $\boldsymbol{\theta}$ satisfies (6.2), then it also satisfies (6.1). □

Let $\widetilde{\mathbf{T}} \in \mathbb{R}^{(\ell-1)(m-1)\times \ell m}$ such that $\widetilde{\mathbf{T}}\boldsymbol{\theta} \le \mathbf{0}$ is equivalent to $\boldsymbol{\theta} \in \widetilde{\mathcal{C}}_{\text{red.}}$. The matrix $\widetilde{\mathbf{T}}$ has a rank $(\ell - 1)(m - 1)$, which means that $\widetilde{\mathbf{T}}$ is a full-rank matrix. Indeed, recall that

$$\boldsymbol{\theta} = (\theta_{11}, \dots, \theta_{1k}, \dots, \theta_{1m}, \dots, \theta_{j1}, \dots, \theta_{jk}, \dots, \theta_{jm}, \dots, \theta_{\ell 1}, \dots, \theta_{\ell k}, \dots, \theta_{\ell m})^\top.$$

At row $\mathbf{R}_u$ of $\widetilde{\mathbf{T}}$, where $1 \le u \le (\ell - 1)(m - 1)$, the first non-zero element is at $(j - 2)m + k - 1$ for $1 < j < \ell$ and $1 < k \le m$. If we construct $\widetilde{\mathbf{T}}$ as in Example 6.1.1, then the first non-zero entry is shifted to the right as u increases. It means that the elements below the first non-zero entry are zero, which makes it the pivot of $\mathbf{R}_u$. Each row has such pivot and hence the matrix $\widetilde{\mathbf{T}}$ is a full-ranked matrix.

6.2. The implementation of the gradient projection method

Now that the number of constraints is reduced, we proceed to implement an algorithm that solves the following minimization problem

$$\hat{\boldsymbol{\theta}} := \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\mathbb{R}^{\ell m}_{-\infty}} \widetilde{\phi}_M(\boldsymbol{\theta}) = \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\mathbb{R}^{\ell m}_{-\infty}} (e^{\boldsymbol{\theta}} - \mathbf{L}^{-1}\mathbf{A}^{-1}\mathbf{c})^\top \mathbf{L}^\top \mathbf{A}\mathbf{L}(e^{\boldsymbol{\theta}} - \mathbf{L}^{-1}\mathbf{A}^{-1}\mathbf{c})$$
$$\text{s.t.} \quad \widetilde{\mathbf{T}}\boldsymbol{\theta} \le \mathbf{0}.$$

To solve this, we choose the modification of the gradient projection method of Rosen (1960), proposed by Du and Zhang (1989). This iterative method ensures that in each iteration, the new proposed point stays feasible and it minimizes the objective function. The modification of this method ensures that the iterative method converges to a Karush-Kuhn-Tucker (KKT) point or it generates a sequence of $\boldsymbol{\theta}_s$ such that every accumulation point is a KKT point. For details of the unmodified version of the method we refer the reader to Appendix D. The solution of the algorithm, however, is not necessarily a global minimum because $\widetilde{\phi}_M$ is a non-convex function.

Table 6.1: A list of parameters that we need to create matrices for computing the objective function $\widetilde{\phi}_M$

Parameter	Description
n	the number of observations
ℓ	the number of unique covariate values
m	the number of unique response values
$\mathbf{x}$	a vector of size ℓ such that $x_1 < \cdots < x_\ell$
$\mathbf{y}$	a vector of size m such that $y_1 < \cdots < y_m$
$\Delta\mathbf{y}$	a vector of size m such that for $1 \le k < m$, the k-th entry is $y_{k+1} - y_k$ and the m-th entry is a large positive number M
$\mathbf{w}$	an $\ell \times m$ matrix such that w_{jk} is the number of observed (x_j, y_k)
$\mathbf{w}_{j+}$	a vector of size ℓ, where the j-th entry is $\sum_{k=1}^{m} w_{jk}$

Before applying the algorithm, we prepare the dataset by computing the necessary parameters for the objective function $\widetilde{\phi}_M$. For instance, we need to extract ℓ and m from the dataset to determine the size of the matrices. We further need to sort the covariate and the response values from the lowest to the highest values. With the response values, we also need to compute the difference between each consecutive response value. In Table 6.1, we summarize all of the necessary parameters of the dataset we extract before proceeding to the main algorithm. With these parameters, we can then construct matrices $\mathbf{A}$, $\mathbf{L}$, a vector $\mathbf{c}$, as defined in Section 5.3.2 and the matrix $\widetilde{\mathbf{T}}$ that represents the constraints for $\boldsymbol{\theta}$ in $\widetilde{\mathcal{C}}_{\text{red.}}$.

Once the necessary matrices and vectors are constructed, we enter the initialization step and proceed with the main procedure of the gradient projection method which consists of two steps. We describe the procedures in Algorithm 1 and Algorithm 2. In these algorithms we use $s_{\max}$ to stop the algorithm early if required.

We add several remarks on the Algorithm 1. First of all, the modification of the gradient projection method occurs in the choice of the direction vector $\mathbf{d}$. In the unmodified version, the algorithm uses $\mathbf{d}_s^{II}$ for the direction vector if $\mathbf{w} \ngeq \mathbf{0}$ and there is no $\mathbf{d}_s^{I}$. Further, if $c = 0$ and $\mathbf{d}_s^I \neq \mathbf{0}$, then $\mathbf{d}_s^I$ will always be assigned to $\mathbf{d}_s$. This coincides with the unmodified gradient projection method in Appendix D. If $\mathbf{M}$ is an empty matrix or if $\mathbf{d}_s^I = \mathbf{0}$, then the algorithm is also similar to the unmodified gradient projection method. Regarding the gradient of $\widetilde{\phi}_M$, we compute it in the following way:

$$\nabla\widetilde{\phi}_M(\boldsymbol{\theta}) = 2\mathbf{E}(\boldsymbol{\theta})(\mathbf{L}^\top\mathbf{A}\mathbf{L})(e^{\boldsymbol{\theta}} - \mathbf{L}^{-1}\mathbf{A}^{-1}\mathbf{c}),$$

where $\mathbf{E}(\boldsymbol{\theta})$ is a diagonal $\ell m \times \ell m$ matrix with $e^{\boldsymbol{\theta}}$ as the diagonal entries[2].

To comment further on the Algorithm 2, if $\widehat{\mathbf{d}} \leq \mathbf{0}$ we use the L-BFGS-B algorithm to compute λ_s, otherwise we use a method that combines both the golden section search and successive parabolic interpolation[3]. The L-BFGS-B algorithm uses the gradient projection method and the quasi-Newton method. The former is used for finding the set of binding constraints and the latter is used to approximate the Hessian matrix of $\widetilde{\phi}_M$. We refer the reader to Byrd et al. (1995) for more detail on the L-BFGS-B algorithm. In the alternative case, the golden section search method is used to reduce the length of the interval for finding the minimum. Meanwhile, the successive parabolic interpolation ensures faster convergence to the minimum. For these two methods we refer the reader to Brent (2003, p. 146) for a description of the algorithm. We also recommend Bazaraa et al. (2006, p. 350) for an the explanation of the golden section search method.

The last part that we need to address is the method for generating a feasible point $\boldsymbol{\theta}_0$. To this end, we use the uniquely sorted response values from the given dataset. Next, we sample independent new response values $\mathbf{y}_0 := (y_k^{(0)})_{k=1}^m$ from a uniform distribution on the interval $[y_1, y_K]$. For each $j = 1, \dots, \ell$ and $x_j \in \mathbf{x}$, we then compute

$$f_{Y|X}^{(0)}(y|x_j) = \frac{1}{\xi(x_j)} y^{\xi(x_j)-1} e^{-y} \mathbb{1}\{y > 0\},$$

where $\xi(x_j) = x_j + a$ and $a \in \{10, 20\}$. Here, the function $f_{Y|X}^{(0)}$ is a density function of a gamma distribution with shape function $\xi(x_j)$ and scale 1. The shape function $\xi(x_j)$ is an isotonic function and the scale is a constant. Therefore, the densities admit the likelihood ratio order if $x_j < x_{j'}$ and $1 \leq j, j' \leq \ell$. Thus, $\boldsymbol{\theta}_0 = (\log f_{Y|X}^{(0)}(y_k^{(0)}|x_j))_{j,k} \in \mathbb{R}^{\ell m}$, where $1 \leq j \leq \ell$ and $1 \leq k \leq m$, is a feasible starting point.

6.3. Sequentially estimating the conditional distribution functions

In Algorithm 2, the algorithm determines λ, which tells us how far to move from $\boldsymbol{\theta}_s$ in the direction of $\mathbf{d}$ such that the function value decreases. Using a common λ, we move all entries of $\boldsymbol{\theta}_s$ in the direction of $\mathbf{d}$ and obtain a new point for the next iteration, $\boldsymbol{\theta}_{s+1}$. In this section, we adjust the Algorithms 1 and 2 so that for any fixed $1 \leq j \leq m$, we apply the algorithm only on a vector $\boldsymbol{\theta}_j := (\theta_{j1}, \dots, \theta_{jm})^\top$. Further, for each $1 \leq j \leq m$, we obtain λ_j instead of λ. For $j = 1$, we solve the unconstrained minimization

[2] Let $\mathbf{f}(\mathbf{x})$ be a function, where $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{A} \in \mathbb{R}^{d\times d}$ is a symmetric matrix, then $\frac{d}{d\mathbf{x}}\mathbf{f}(\mathbf{x})^\top\mathbf{A}\mathbf{f}(\mathbf{x}) = \frac{d}{d\mathbf{x}}\mathbf{f}(\mathbf{x})(\mathbf{A} + \mathbf{A}^\top)\mathbf{f}(\mathbf{x}) = 2\frac{d}{d\mathbf{x}}\mathbf{f}(\mathbf{x})\mathbf{A}\mathbf{f}(\mathbf{x})$. Here, $\frac{d}{d\mathbf{x}}\mathbf{f}(\mathbf{x})$ is the Jacobian matrix of $\mathbf{f}(\mathbf{x})$.

[3] If $\widehat{\mathbf{d}} \leq \mathbf{0}$, then we use `optim()` and choose `method = 'L-BFGS-B'`. If $\widehat{\mathbf{d}} \nleq \mathbf{0}$, then we use `optimize()`

Algorithm 1: The main step that returns the estimated minimizer $\widehat{\boldsymbol{\theta}}$

```
Input: Dataset, the maximum number of iteration $s_{\max}$, $M > 0$, $c > 0$ and $\mathbf{b}$
Compute the parameters that are listed in Table 6.1 ;
Construct the matrix constraint $\widetilde{\mathbf{T}}$ of size $(\ell - 1)(m - 1) \times \ell m$;
Generate a starting point $\boldsymbol{\theta}_0 \in \mathbb{R}^{\ell \cdot m}$ such that $\widetilde{\mathbf{T}}\boldsymbol{\theta}_0 \leq \mathbf{0}$;
Decompose $\widetilde{\mathbf{T}}$ and $\mathbf{b}$ into $(\widetilde{\mathbf{T}}_1, \widetilde{\mathbf{T}}_2)$ and $(\mathbf{b}_1, \mathbf{b}_2)$ respectively such that $\widetilde{\mathbf{T}}_1\boldsymbol{\theta}_0 = \mathbf{b}_1$ and $\widetilde{\mathbf{T}}_2\boldsymbol{\theta}_0 < \mathbf{b}_2$;
$s \leftarrow 0$;
$\boldsymbol{\theta}_s \leftarrow \boldsymbol{\theta}_0$;
while TRUE do
    if $s = s_{\max}$ then
        break
    end
    $\mathbf{M} \leftarrow \widetilde{\mathbf{T}}_1$;
    Compute $\nabla\widetilde{\phi}_M(\boldsymbol{\theta}_s)$;
    if $\mathbf{M}$ is an empty matrix then
        if $\nabla\widetilde{\phi}_M(\boldsymbol{\theta}_s) = \mathbf{0}$ then
            break
        end
        else
            $\mathbf{d}_s \leftarrow -\nabla\widetilde{\phi}_M(\boldsymbol{\theta}_s)$;
            Do Algorithm 2 ;
            $s \leftarrow s + 1$;
        end
    end
    else
        $\mathbf{P} \leftarrow \mathbf{I} - \mathbf{M}^\top(\mathbf{M}\mathbf{M}^\top)^{-1}\mathbf{M}$;
        $\mathbf{d}_s^I \leftarrow -\mathbf{P}\nabla\widetilde{\phi}_M(\boldsymbol{\theta}_s)$;
        $\mathbf{w} \leftarrow -[(\mathbf{M}\mathbf{M}^\top)^{-1}\mathbf{M}]\nabla\widetilde{\phi}_M(\boldsymbol{\theta}_s)$;
        if $\mathbf{w} \geq \mathbf{0}$ then
            if $\mathbf{d}_s^I = \mathbf{0}$ then
                break
            end
            else
                $\mathbf{d}_s \leftarrow \mathbf{d}_s^I$;
                Do Algorithm 2;
                $s \leftarrow s + 1$;
            end
        end
        else
            $w_h \leftarrow \min_j w_j$;
            Remove the $h$-th row from $\widetilde{\mathbf{T}}_1$ and call it $\widehat{\mathbf{M}}$;
            $\widehat{\mathbf{P}} \leftarrow \mathbf{I} - \widehat{\mathbf{M}}^\top(\widehat{\mathbf{M}}\widehat{\mathbf{M}}^\top)^{-1}\widehat{\mathbf{M}}$;
            $\mathbf{d}_s^{II} \leftarrow -\widehat{\mathbf{P}}\nabla\widetilde{\phi}_M(\boldsymbol{\theta}_s)$;
            $\mathbf{d}_s \leftarrow \mathbf{d}_s^I$ if $||\mathbf{d}_s^I||/|w_h| > c$; otherwise $\mathbf{d}_s^{II}$;
            Do Algorithm 2;
            $s \leftarrow s + 1$;
        end
    end
end
Output: The estimated minimizer $\boldsymbol{\theta}_s$
```

Algorithm 2: The line search step that returns a new point $\boldsymbol{\theta}_s$, $\mathbf{b}_2$, and the matrices $\widetilde{\mathbf{T}}_1$ and $\widetilde{\mathbf{T}}_2$

```
Input: Step $s \geq 0$, $\widetilde{\mathbf{T}}_2$, $\widetilde{\mathbf{T}}$, a vector $\boldsymbol{\theta}$ and $\mathbf{b}_2$ from Algorithm 1, and a direction vector $\mathbf{d}$
$\widehat{\mathbf{b}} \leftarrow \mathbf{b}_2 - \widetilde{\mathbf{T}}_2\boldsymbol{\theta}$;
$\widehat{\mathbf{d}} \leftarrow \widetilde{\mathbf{T}}_2\mathbf{d}$;
if $\widehat{\mathbf{d}} \leq \mathbf{0}$ then
    $\lambda_s \leftarrow \arg\min_\lambda \widetilde{\phi}_M(\boldsymbol{\theta}_s + \lambda\mathbf{d})$ such that $\lambda \geq 0$;
end
else
    $\lambda_{\max} \leftarrow \min\{\widehat{b}_j/\widehat{d}_j : \widehat{d}_j > 0\}$ ;
    $\lambda_s \leftarrow \arg\min_\lambda \widetilde{\phi}_M(\boldsymbol{\theta}_s + \lambda\mathbf{d})$ such that $0 \leq \lambda \leq \lambda_{\max}$;
end
$\boldsymbol{\theta}_s \leftarrow \boldsymbol{\theta}_s + \lambda_s\mathbf{d}$;
Decompose $\widetilde{\mathbf{T}}$ into $\widetilde{\mathbf{T}}_1$ and $\widetilde{\mathbf{T}}_2$ such that $\widetilde{\mathbf{T}}_1\boldsymbol{\theta}_s = \mathbf{0}$ and $\widetilde{\mathbf{T}}_2\boldsymbol{\theta}_s < \mathbf{0}$;
Output: The new $\boldsymbol{\theta}_s$, $\mathbf{b}_2$, and the matrices $\widetilde{\mathbf{T}}_1$ and $\widetilde{\mathbf{T}}_2$
```

problem. For $j > 1$, we use the gradient projection method so that the likelihood ratio order constraint is satisfied. To this end, we start by discussing the changes in the objective function. Subsequently, we also change the constraint and adjust the algorithms from the previous section.

6.3.1. The adjustment of the objective function and the constraint

Let us recall the matrices and vectors required to construct the objective function $\widetilde{\phi}_M$ in Section 6.2. We refer the reader to Section 5.3.2 for the definitions of these matrices and vectors. We first have $\boldsymbol{\theta} \in \mathbb{R}^{\ell\cdot m}_{-\infty}$, which the entries are defined as

$$\boldsymbol{\theta} = (\theta_{11}, \dots, \theta_{1k}, \dots, \theta_{1m}, \dots, \theta_{j1}, \dots, \theta_{jk}, \dots, \theta_{jm}, \dots, \theta_{\ell 1}, \dots, \theta_{\ell k}, \dots, \theta_{\ell m})^\top.$$

Let $\boldsymbol{\theta}_j := (\theta_{j1}, \dots, \theta_{jm})^\top \in \mathbb{R}^m_{-\infty}$ for any $j = 1, \dots, \ell$, then $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_\ell)^\top$. The matrix $\mathbf{L}$ is a diagonal block matrix of size $\ell m \times \ell m$, where each block is an $m \times m$ lower triangular matrix, denoted as $\mathbf{L}_j$. The matrix $\mathbf{A}$ has the same type and size of matrix as in $\mathbf{L}$, but each block in $\mathbf{A}$ is an $m \times m$ matrix defined as

$$\mathbf{A}_j := \begin{pmatrix} (w_{j+}/n)\Delta y_1 & 0 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & (w_{j+}/n)\Delta y_2 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & 0 & (w_{j+}/n)\Delta y_3 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & (w_{j+}/n)\Delta y_{k'} & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & (w_{j+}/n)\Delta y_m \end{pmatrix}.$$

Lastly, a vector $\mathbf{c} \in \mathbb{R}^{\ell m}$ consists of vectors $\mathbf{c}_j \in \mathbb{R}^m$ where

$$\mathbf{c}_j := \sum_{k=1}^{m} \begin{pmatrix} (w_{jk}/n)\Delta y_1 & 0 & \cdots & 0 \\ 0 & (w_{jk}/n)\Delta y_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & (w_{jk}/n)\Delta y_m \end{pmatrix} \begin{pmatrix} \mathbb{1}(1 \geq k) \\ \mathbb{1}(2 \geq k) \\ \vdots \\ 1 \end{pmatrix} = \frac{1}{n} \begin{pmatrix} w_{j1}\Delta y_1 \\ (w_{j1} + w_{j2})\Delta y_2 \\ \vdots \\ w_{j+}\Delta y_m. \end{pmatrix}.$$

We use the vectors $\boldsymbol{\theta}_j$, $\mathbf{c}_j$, and the matrices $\mathbf{A}_j$ and $\mathbf{L}_j$ to compute

$$\widetilde{\phi}_M(\boldsymbol{\theta}_j) = (e^{\boldsymbol{\theta}_j} - \mathbf{L}_j^{-1}\mathbf{A}_j^{-1}\mathbf{c}_j)^\top \mathbf{L}_j^\top \mathbf{A}_j \mathbf{L}_j (e^{\boldsymbol{\theta}_j} - \mathbf{L}_j^{-1}\mathbf{A}_j^{-1}\mathbf{c}_j).$$

Minimizing $\widetilde{\phi}_M(\boldsymbol{\theta}_J)$ corresponds to minimizing

$$\frac{1}{n}\sum_{k=1}^{m} w_{jk}\mathrm{CRPS}(F_{Y|x_j}, y_k),$$

for a fixed $1 \leq j \leq m$.

Regarding the constraint for $\boldsymbol{\theta}$, the inequality in (6.2) is equivalent to

$$\theta_{j,k-1} - \theta_{j,k} \leq \theta_{j-1,k-1} - \theta_{j-1,k}$$

for all $1 < j \leq \ell$ and $1 < k \leq m$. For the adjusted algorithm, we first solve

$$\widehat{\boldsymbol{\theta}}_1 = \operatorname*{arg\,min}_{\boldsymbol{\theta}_1 \in \mathbb{R}^m_{-\infty}} \widetilde{\phi}_M(\boldsymbol{\theta}_1)$$

numerically by the method of steepest descent method. For $1 < j \leq m$, we solve

$$\widehat{\boldsymbol{\theta}}_j = \operatorname*{arg\,min}_{\boldsymbol{\theta}_j \in \widetilde{\mathcal{C}}_{\mathrm{red.,j}}} \widetilde{\phi}_M(\boldsymbol{\theta}_j)$$

with a feasible convex set

$$\widetilde{\mathcal{C}}_{\mathrm{red.},j} := \{\boldsymbol{\theta}_j \in \mathbb{R}^m_{-\infty} : \theta_{j,k-1} - \theta_{j,k} \leq \widehat{\theta}_{j-1,k-1} - \widehat{\theta}_{j-1,k} \quad 1 < k \leq m\}.$$

Let $\breve{\mathbf{T}}$ be a matrix of size $(m-1) \times m$, such that its main diagonal entries are 1, its superdiagonal entries are -1 and 0 otherwise. Then, for any $1 < j \leq \ell$,

$$\breve{\mathbf{T}}\boldsymbol{\theta}_j \leq \breve{\mathbf{T}}\widehat{\boldsymbol{\theta}}_{j-1} \quad \text{if and only if} \quad \boldsymbol{\theta}_j \in \widetilde{\mathcal{C}}_{\mathrm{red.},j}.$$

The matrix $\breve{\mathbf{T}}$ is a full-rank matrix, with rank $m-1$.

As a remark, for each $j = 1, \ldots, \ell$, we need to propose an initial feasible point $\boldsymbol{\theta}_j^{(0)}$. To reduce the work on programming, we reuse the initial point generated for Algorithm 1. However, from our numerical experiment, we discover that for this new algorithm the starting point might be infeasible. To correct the initial point so that it becomes feasible, we do as follows. Fix $1 < j \leq m$, and let

$$-\Delta\boldsymbol{\theta}_j^{(0)} := \begin{pmatrix} \theta_{j1}^{(0)} - \theta_{j2}^{(0)} \\ \theta_{j2}^{(0)} - \theta_{j3}^{(0)} \\ \vdots \\ \theta_{j,m-1}^{(0)} - \theta_{jm}^{(0)} \end{pmatrix} = \breve{\mathbf{T}}\boldsymbol{\theta}_j^{(0)} \qquad \text{and} \qquad -\Delta\widehat{\boldsymbol{\theta}}_{j-1} := \begin{pmatrix} \widehat{\theta}_{j-1,1} - \widehat{\theta}_{j-1,2} \\ \widehat{\theta}_{j-1,2} - \widehat{\theta}_{j-1,3} \\ \vdots \\ \widehat{\theta}_{j-1,m-1} - \widehat{\theta}_{j-1,m} \end{pmatrix} = \breve{\mathbf{T}}\widehat{\boldsymbol{\theta}}_{j-1},$$

which are vectors of size $m-1$. Suppose $-\Delta\boldsymbol{\theta}_j^{(0)} \nleq -\Delta\widehat{\boldsymbol{\theta}}_{j-1}$, then there exists a non-empty set of index $I \subseteq \{1, \ldots, m-1\}$ such that for any $i \in I$,

$$-\Delta\theta_{ji}^{(0)} = \theta_{j,i}^{(0)} - \theta_{j,i+1}^{(0)} > \widehat{\theta}_{j-1,i} - \widehat{\theta}_{j-1,i+1} = -\Delta\widehat{\theta}_{j-1,i}.$$

Now we replace $-\Delta\boldsymbol{\theta}_j^{(0)}$ with $-\Delta\breve{\boldsymbol{\theta}}_j^{(0)}$ where its entries are defined as followed:

$$-\Delta\breve{\theta}_{ji}^{(0)} = \begin{cases} -\Delta\widehat{\theta}_{j-1,i} & \text{, if } i \in I \\ -\Delta\theta_{ji}^{(0)} & \text{, if } i \notin I. \end{cases}$$

Then the new initial point is

$$\breve{\boldsymbol{\theta}}_j^{(0)} := \begin{pmatrix} \theta_{j1}^{(0)} \\ \theta_{j1}^{(0)} + \Delta\breve{\theta}_{j1}^{(0)} \\ \theta_{j1}^{(0)} + \Delta\breve{\theta}_{j1}^{(0)} + \Delta\breve{\theta}_{j2}^{(0)} \\ \vdots \\ \theta_{j1}^{(0)} + \sum_{k=1}^{m-1} \breve{\theta}_{jk}^{(0)} \end{pmatrix},$$

which is feasible because $\breve{\mathbf{T}}\breve{\boldsymbol{\theta}}_j^{(0)} = -\Delta\breve{\boldsymbol{\theta}}_j^{(0)} \leq \breve{\mathbf{T}}\widehat{\boldsymbol{\theta}}_{j-1}$ by definition of $-\Delta\breve{\boldsymbol{\theta}}_j^{(0)}$.

6.3.2. The description of the adjusted algorithm

The first algorithm that we describe is for solving the minimization problem

$$\widehat{\boldsymbol{\theta}}_1 = \operatorname*{arg\,min}_{\boldsymbol{\theta}_1 \in \mathbb{R}^m_{-\infty}} \widetilde{\phi}_M(\boldsymbol{\theta}_1).$$

For this problem, we use the method of steepest descent. This iterative method uses $\mathbf{d} = -\nabla\widetilde{\phi}_M$ for the direction vector. It then performs a line search to determine how far we should move from a given point in the direction of $\mathbf{d}$. More information on this method can be found in Bazaraa et al. (2006, p. 384). The implementation of the steepest descent method is in Algorithm 3.

At the beginning of Algorithm 3, we construct a large size of matrices $\mathbf{L}$, $\mathbf{L}^{-1}$, $\mathbf{A}$, $\mathbf{A}^{-1}$, and a vector $\mathbf{c}$. From these matrices and vector, we extract a sub-matrix and sub-vector that correspond to computing $\widetilde{\phi}_M(\boldsymbol{\theta}_j)$. The locations of these sub-matrices and sub-vectors are already determined in Section 5.3.2, which is the $(j-1)*m+1$-th through (jm)-th entries. Since the matrices are diagonal block matrices, we only need these blocks to compute the objective function and apply the steepest descent method.

Algorithm 3: The steepest descent algorithm for approximating $\widehat{\boldsymbol{\theta}}_1$

```
Input: Dataset, the maximum number of iteration $s_{\max}$, $M > 0$, $c > 0$
Compute the parameters that are listed in Table 6.1 ;
Construct the matrices $\mathbf{L}$, $\mathbf{L}^{-1}$, $\mathbf{A}$, $\mathbf{A}^{-1}$ and a vector $\mathbf{c}$ ;
Generate a starting point $\boldsymbol{\theta}^{(0)} \in \mathbb{R}^{\ell m}$ ;
$s \leftarrow 0, \quad j \leftarrow 1, \quad \mathrm{loc} \leftarrow ((j-1)*m+1):(jm)$ ;
$\mathbf{A}_j \leftarrow \mathbf{A}[\mathrm{loc},\mathrm{loc}], \quad \mathbf{A}_j^{-1} \leftarrow \mathbf{A}^{-1}[\mathrm{loc},\mathrm{loc}], \quad \mathbf{L}_j \leftarrow \mathbf{L}[\mathrm{loc},\mathrm{loc}], \quad \mathbf{L}_j^{-1} \leftarrow \mathbf{L}^{-1}[\mathrm{loc},\mathrm{loc}], \quad \mathbf{c}_j \leftarrow$
  $\mathbf{c}[\mathrm{loc}]$ ;
$\boldsymbol{\theta}_j^{(0)} \leftarrow \boldsymbol{\theta}^{(0)}[\mathrm{loc}]$ ;
$\mathbf{d}_s \leftarrow -\nabla\widetilde{\phi}_M(\boldsymbol{\theta}_j^{(0)})$ ;
while TRUE do
    $\boldsymbol{\theta}_j^{(s)} \leftarrow \boldsymbol{\theta}_j^{(0)}$;
    if $s = s_{\max}$ then
        break
    end
    if $||\mathbf{d}_s|| < \varepsilon$ then
        break
    end
    else
        $\lambda_s \leftarrow \arg\min_\lambda \widetilde{\phi}_M(\boldsymbol{\theta}_j^{(s)} + \lambda\mathbf{d}_s)$ such that $\lambda \geq 0$;
        $s \leftarrow s+1$;
        $\boldsymbol{\theta}_j^{(s)} \leftarrow \boldsymbol{\theta}_j^{(s)} + \lambda_s\mathbf{d}_s$;
        $\mathbf{d}_s \leftarrow -\nabla\widetilde{\phi}_M(\boldsymbol{\theta}_j^{(s)})$ ;
    end
end
Output: The approximated $\widehat{\boldsymbol{\theta}}_1$
```

As for Algorithm 4, we apply the gradient projection method for each $j = 2,\ldots,m$. For each j, we extract the necessary matrices for computing $\widetilde{\phi}_M$ in the same manner as in Algorithm 3. We then correct the initial starting point if it is infeasible, as discussed in Section 6.3.1. Afterwards, we enter the while loop as described in Algorithm 1.

6.4. Simulation studies

In this section, we compare the estimator proposed by Mösching and Dümbgen (2024) with the one developed in this thesis. For comparison purposes, we generate a dataset using the same density function as in the aforementioned article. We then visually assess the performance of the estimated conditional distribution functions, contrasting it with the estimator proposed by Mösching and Dümbgen (2024). Before we present the results, we explain how the dataset is generated which we follow from Mösching and Dümbgen (2024).

Algorithm 4: The method of gradient projection for approximating $\widehat{\boldsymbol{\theta}}_j$ for each $j = 2, \dots, m$. This is the continuation of Algorithm 3

```
Input: Dataset, the maximum number of iteration $s_{\max}$, $M > 0$, and $c > 0$
For every $j = 2, 3, \dots, m$ do
    $\mathrm{loc} \leftarrow ((j-1) * m + 1) : (jm)$ ;
    $\mathbf{A}_j \leftarrow \mathbf{A}[\mathrm{loc}, \mathrm{loc}]$,   $\mathbf{A}_j^{-1} \leftarrow \mathbf{A}^{-1}[\mathrm{loc}, \mathrm{loc}]$,   $\mathbf{L}_j \leftarrow \mathbf{L}[\mathrm{loc}, \mathrm{loc}]$,   $\mathbf{L}_j^{-1} \leftarrow \mathbf{L}^{-1}[\mathrm{loc}, \mathrm{loc}]$,   $\mathbf{c}_j \leftarrow$
      $\mathbf{c}[\mathrm{loc}]$ ;
    $\boldsymbol{\theta}_j^{(0)} \leftarrow \boldsymbol{\theta}^{(0)}[\mathrm{loc}]$ ;
    $-\Delta\boldsymbol{\theta}_j^{(0)} \leftarrow \breve{\mathbf{T}}\boldsymbol{\theta}_j^{(0)}$;
    $-\Delta\widehat{\boldsymbol{\theta}}_{j-1} \leftarrow \breve{\mathbf{T}}\widehat{\boldsymbol{\theta}}_{j-1}$;
    if $-\Delta\boldsymbol{\theta}_j^{(0)} \nleq -\Delta\widehat{\boldsymbol{\theta}}_{j-1}$ then
        $I \leftarrow \{i : -\Delta\theta_{ji}^{(0)} > -\Delta\widehat{\theta}_{j-1,i}\}$;
        $-\Delta\breve{\boldsymbol{\theta}}_j^{(0)} \leftarrow -\Delta\boldsymbol{\theta}_j^{(0)}$;
        $-\Delta\breve{\boldsymbol{\theta}}_j^{(0)}[I] \leftarrow -\Delta\widehat{\boldsymbol{\theta}}_{j-1}[I]$;
        $\boldsymbol{\theta}_j^{(0)} \leftarrow$ the cumulative sum of a vector $(\theta_{j1}^{(0)}, \Delta\boldsymbol{\theta}_j^{(0)\top})^\top$ ;
    end
    Decompose $\breve{\mathbf{T}}$ and $\mathbf{b}$ into $(\breve{\mathbf{T}}_1, \breve{\mathbf{T}}_2)$ and $(\mathbf{b}_1, \mathbf{b}_2)$ respectively such that $\breve{\mathbf{T}}_1\boldsymbol{\theta}_j^{(0)} = \mathbf{b}_1$ and
      $\breve{\mathbf{T}}_2\boldsymbol{\theta}_j^{(0)} < \mathbf{b}_2$;
    Do the while loop from Algoritm 1 with constraint matrix $\breve{\mathbf{T}}$, the binding constraint matrix $\breve{\mathbf{T}}_1$,
      the starting point $\boldsymbol{\theta}_j^{(0)}$ and a vector $\mathbf{b}_2$. Save the result in $\widehat{\boldsymbol{\theta}}_j$;
end
Output: The approximated $\widehat{\boldsymbol{\theta}}_j$ for $j = 2, \dots, m$.
```

6.4.1. Generating the dataset

In order to generate the dataset, we first generate the covariate values, which we then use to generate the corresponding response values. To generate the covariate values $X_1, \dots, X_n$, we draw each X_i independently for $i = 1, \dots, n$, where each X_i takes value in a set

$$\mathcal{X} := 1 + \frac{3}{\ell_0} \cdot \{1, 2, \dots, \ell_0\},$$

and $\ell_0 \in \mathbb{N}$. The sampling is done with replacement so that it is possible to obtain ties among the X_i.

To generate the response values given $X_1, \dots, X_n$, we sample $Y_1, \dots, Y_n$ from the gamma distribution with the following density function

$$g_{Y|X}(y|x) := \frac{1}{\Gamma(k(x))b(x)^{k(x)}} y^{k(x)-1} e^{-y/b(x)} \mathbb{1}\{y > 0\} \tag{6.4}$$

where the shape function $k(x) : \mathcal{X} \to \mathbb{R}_{\geq 0}$ and the shape function $b(x) : \mathcal{X} \to \mathbb{R}_{\geq 0}$ are defined as follows: $k(x) := 2 + (x+1)^2$ and $b(x) := 1 - e^{-10x}$. Both $k(x)$ and $b(x)$ are isotonic if $x \geq 0$. Therefore by Example 2.3.5, we have that $[Y|X = x_i] \leq_{\mathrm{lr}} [Y|X = x_{i'}]$ if $x_i < x_{i'}$. The sample size of the generated data set is $n = 15$ and we set $\ell_0 = 1000$. To allow the possibility of ties the response values are rounded to one decimal place.

6.4.2. Comparing the results of the estimated conditional distributions

For consistent comparison, we use the same generated dataset when using algorithms in Section 6.2 and Section 6.3. The generated dataset has $\ell = m = 15$. The starting feasible points for these algorithms are not necessary. We choose $c = 1000$ for both algorithms. Regarding the parameter M, we set $M = 1000$ for Algorithm 1 and $M = 10^4$ for Algorithm 3 and Algorithm 4.

In this simulation study, we aim to plot the estimated conditional distribution functions. Each figure we produce includes four functions of y: the true conditional distribution function with density (6.4), the estimated conditional distribution using algorithms in Section 6.2 or in Section 6.3, the distribution function computed using the initial feasible point, and the estimated conditional distribution using algorithms from Mösching and Dümbgen (2024). We denote these functions as $G_{Y|x_j}$, $\widehat{F}_{Y|x_j}$, $\breve{F}_{Y|x_j}$, $F^{(0)}_{Y|x_j}$ and $\overline{F}_{Y|x_j}$ respectively. Let $\mathbb{F}_{Y|x_j} \in \{\widehat{F}_{Y|x_j}, \breve{F}_{Y|x_j}, F^{(0)}_{Y|x_j}, \overline{F}_{Y|x_j}\}$, then with the output of the algorithms, we compute $\mathbb{F}_{Y|x_j}$ in the following manner:

$$\mathbb{F}_{Y|x_j}(y_k|x_j) = \frac{1}{\sum_{k'=1}^{m} \exp(\vartheta_{jk'})} \sum_{k'=1}^{k} \exp(\vartheta_{jk'}) \quad \forall 1 \le j \le \ell, 1 \le k \le m,$$

where $\vartheta_{jk} \in \{\widehat{\theta}_{jk}, \breve{\theta}_{jk}, \theta^{(0)}_{jk}, \overline{\theta}_{jk}\}$. We still normalize $\mathbb{F}_{Y|x_j}$ since $\sum_{k'=1}^{m} \exp(\vartheta_{jk'})$ is not exactly equal to one or smaller than one for $\mathbb{F}_{Y|x_j} \in \{\widehat{F}_{Y|x_j}, \breve{F}_{Y|x_j}, F^{(0)}_{Y|x_j}\}$. Additionally, we assume that the conditional distributions has support $\{y_1, \ldots, y_m\}$ for any $1 \le j \le \ell$. As a remark, the computation of $\overline{F}_{Y|x_j}$ and therefore the normalization, is already computed from the `LRDistReg` package. Finally, we plot $G_{Y|x_j}$ and $\mathbb{F}_{Y|x_j}$, where $j = 1, 7, 15$. The corresponding covariate values are $x_1 = 1.054, x_7 = 1.624$ and $x_{15} = 3.94$.

We first present the results of using algorithms in Section 6.2 in Figure 6.1. We set $s_{\max} = 3000$ and $a = 20$ for generating the initial feasible point $\boldsymbol{\theta}_0$. We observe that $\widehat{F}_{Y|x_7}$ and $\overline{F}_{Y|x_7}$ are closely aligned. However, the algorithm of Mösching and Dümbgen algorithm performs better for estimating $G_{Y|x_j}$, for $j = 1, 7, 15$. While $\overline{F}_{Y|x_1}$ over estimates $G_{Y|x_1}$, the estimator $\widehat{F}_{Y|x_1}$ severely underestimate $G_{Y|x_1}$, especially on approximately $y \in [0, 6]$. For $j = 15$, the estimator $\overline{F}_{Y|x_{15}}$ lies close to the true distribution. However, the estimator $\widehat{F}_{Y|x_{15}}$ performs poorly in estimating $G_{Y|x_{15}}$.

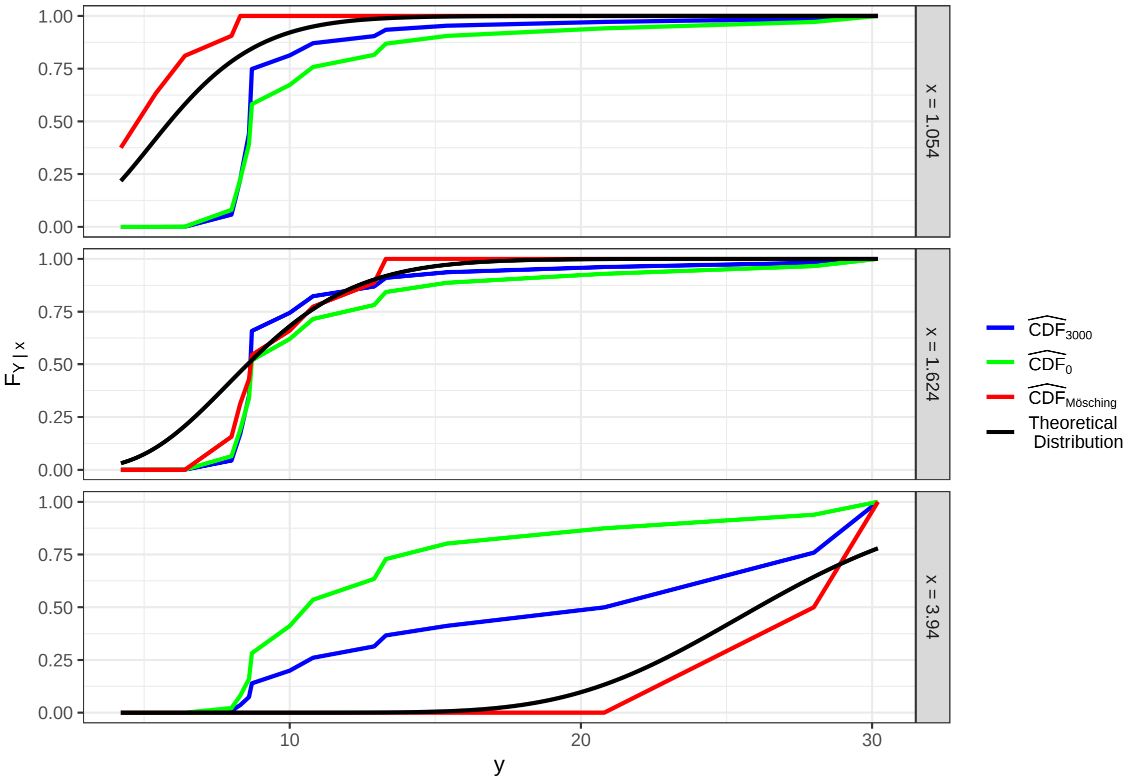

Figure 6.1: The plot of $\{\widehat{F}_{Y|x_j}, F^{(0)}_{Y|x_j}, \overline{F}_{Y|x_j}\}$ with the true distributions $G_{Y|x_j}$ for $j = 1, 7, 15$. The function $F^{(0)}_{Y|x_j}$ is the gamma distribution with shape $x_j + 20$ and scale is 1. The number iterations needed to produce $\widehat{F}_{Y|x_j}$ is 3000.

We then increase the number of maximal iterations $s_{\max}$ to $2 \cdot 10^4$, with the results displayed in Figure 6.2. Compared to when $s_{\max} = 3000$, the estimators $\widehat{F}_{Y|x_j}$ for $j = 1, 7, 15$, are closer to the true distribution $G_{Y|x_j}$. However, we observe that the bias of the estimator $\widehat{F}_{Y|x_1}$ remains large for approximately $y \in [0, 6]$. Furthermore, there is a notable improvement in $\widehat{F}_{Y|x_{15}}$ compared to the result in Figure 6.1. In this instance, the estimator $\widehat{F}_{Y|x_{15}}$ fits $G_{Y|x_{15}}$ better, while $\overline{F}_{Y|x_{15}}$ overall underestimates it.

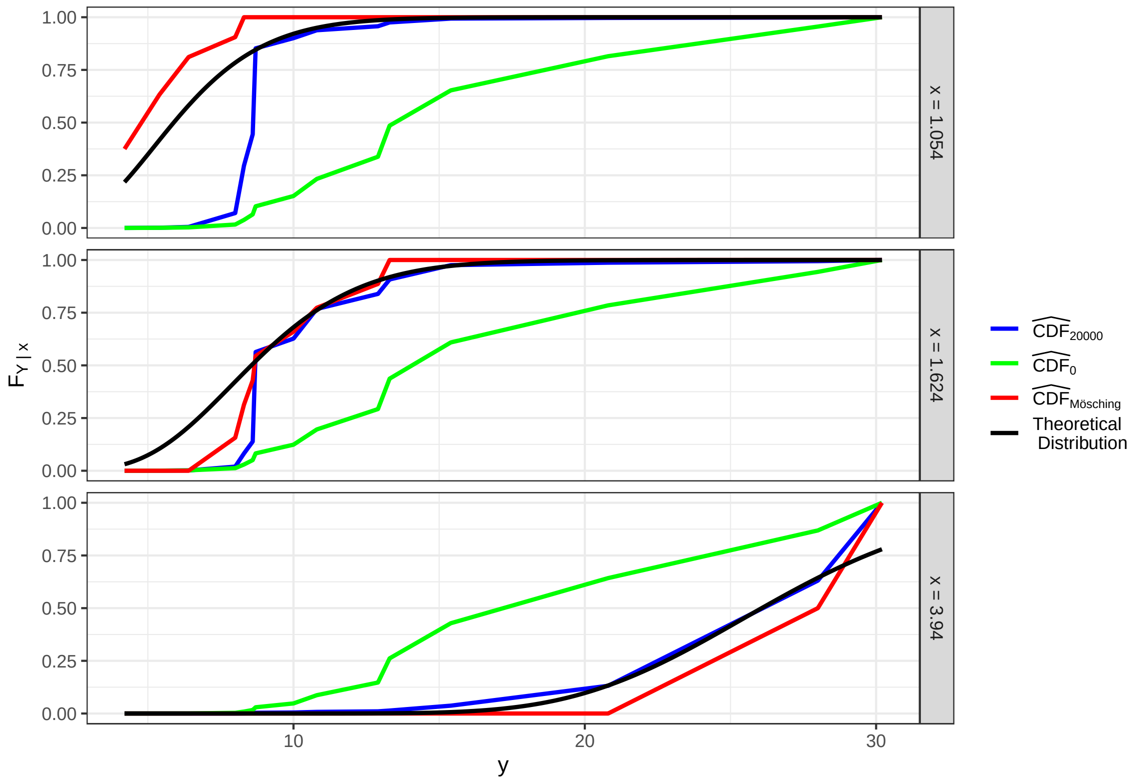

Figure 6.2: The plot of $\{\widehat{F}_{Y|x_j}, F^{(0)}_{Y|x_j}, \overline{F}_{Y|x_j}\}$ with the true distributions $G_{Y|x_j}$ for $j = 1, 7, 15$. The function $F^{(0)}_{Y|x_j}$ is the gamma distribution with shape $x_j + 20$ and scale is 1. The number iterations needed to produce $\widehat{F}_{Y|x_j}$ is $2 \cdot 10^4$.

For more figures of $\widehat{F}_{Y|x_j}$ and $j = 1, \ldots, \ell$, we refer the reader to Appendix F. Here, we also present the results of $\widehat{F}_{Y|x_j}$ when $s_{\max}$ is set to $5 \cdot 10^4$. We observe in these figures that the bias of $\widehat{F}_{Y|x_j}$ for $j = 1$ is smaller for approximately $y \in [0, 6]$. Recall from Figure 6.2, it illustrates the poor performance of the estimator $\widehat{F}_{Y|x_1}$ within this interval.

Lastly, we present the results of estimating $G_{Y|x_j}$ using the algorithm in Section 6.3. Here, we set $s_{\max} = 1000$.In Figure 6.3, we observe overall poor performance of $\check{F}_{Y|x_j}$. Notably, the function $\check{F}_{Y|x_1}$, which is the result of the unconstrained minimization, appears similar to $\widehat{F}_{Y|x_1}$ in Figure 6.1 and Figure 6.2. While the function $\check{F}_{Y|x_1}$ fits $G_{Y|x_1}$ reasonably well over a certain interval, for $j = 7, 15$ the function $\check{F}_{Y|x_j}$ performs poorly compared to $\widehat{F}_{Y|x_j}$ from Figure 6.2 and $\overline{F}_{Y|x_j}$.

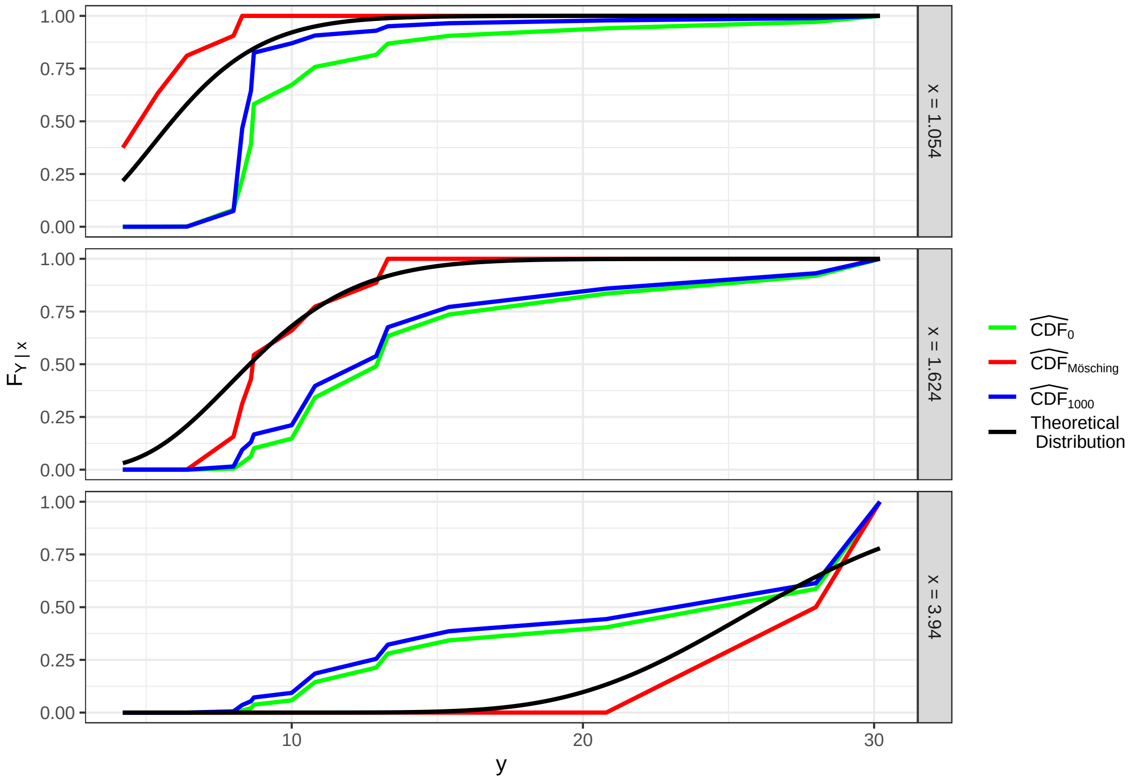

Figure 6.3: The plot of $\{\breve{F}_{Y|x_j}, F^{(0)}_{Y|x_j}, \overline{F}_{Y|x_j}\}$ with the true distributions $G_{Y|x_j}$ for $j = 1, 7, 15$. The function $F^{(0)}_{Y|x_j}$ is the gamma distribution with shape $x_j + 20$ and scale is 1. The number iterations needed to produce $\breve{F}_{Y|x_j}$ is 1000.

7 Conclusion

The purpose of this thesis was to implement an algorithm that estimates conditional distribution functions with a likelihood ratio order constraint. The estimated conditional distribution functions minimize the empirical risk with the CRPS as the loss function while satisfying the imposed constraint. Before we implemented the algorithm to solve this minimization problem, we transformed the estimand into a logarithmic scale. This transformation converts the non-convex feasible set into a convex one. However, the objective function becomes non-convex within this set. To solve the problem numerically, we use the gradient projection method to find a local minimum that is feasible. During the implementation of the gradient projection method, we also modified it to estimate the distribution functions sequentially.

In the initial version of the algorithm, we observed that the performance of the estimators are improved as we extend the run time. This observation was based on comparing the initial proposed function and the estimated function after the algorithm was stopped. Visually, the algorithm accurately fit the estimated distribution function to the true distribution over some interval. However, we observe underestimation as well on some parts of the distribution function. Notably, for the distribution function given the smallest covariate value, the algorithm of Mösching and Dümbgen (2024) tended to overestimate the distribution, while the algorithm implemented in this thesis exhibited underestimation. This underestimation might be caused by the early stopping of the algorithm.

Regarding the modified version of the algorithm, its performance falls short compared to the initial version. Although under performance is explainable by the early stopping ($s_{\max} = 1000$ for each $j = 1, \dots, \ell$), extending the running time for each j would results in a longer total running time. Moreover, due to the early stopping, poor performance for any $j - 1$ and $1 < j \le \ell$ negatively impacts the performance of the subsequent estimators. These effects accumulate, which leads to overall suboptimal estimators.

In contrast, the initial version of the algorithm estimates the distribution functions simultaneously, which avoids the poor performance of the estimators to cascade. However, the modified algorithm offers an advantage. Since the algorithm estimates the distribution functions sequentially, the number of constraints is reduced. Specifically, the size of the matrix constraint is $(m - 1) \times m$ instead of $(\ell - 1)(m - 1) \times \ell m$. This might be beneficial when working with a larger dataset, despite the trade-off in the estimators performance.

7.1. Limitations of the study

We now discuss several limitations of our studies. First, though the running time of the proposed algorithms were not exactly measured, it took approximately 3 to 4 hours and 8 to 10 hours to obtain the estimators when $s_{\max} = 2 \cdot 10^4$ and when $s_{\max} = 5 \cdot 10^4$ respectively. We also observed the poor performances of the estimators, when $s_{\max} = 1000$. Given these results, we believed that there is no value in making a predictive performance comparison, since the performance was visibly inferior to the method proposed by Mösching and Dümbgen (2024). However, performance did improved when $s_{\max} = 2 \cdot 10^4$. Despite this, the length computational time made it impractical to conduct Monte-Carlo simulations, even with small sample size ($n = 15$). Therefore, we are unable to investigate whether the algorithm proposed by Mösching and Dümbgen (2024) or the one proposed here performs better.

Another important limitation is that we did not verify whether the estimated probabilities are indeed a global minimizer of the empirical risk. Unlike the method proposed by Mösching and Dümbgen (2024), where the authors have proven that the negative log-likelihood function is strictly convex in a convex feasible set, our optimization problem does not have this property. As a consequence, the algorithms return a local minimum. If one wants to verify that the algorithms return a global minimizer, one can generate random initial feasible points and solve the optimization problem using these starting points. Then compute the value of the objective function for each approximated solution. The smallest value of the objective function is possibly the approximated global minimum. We decided not to pursue this investigation as well, due the same reason as why we did not conduct Monte Carlo simulations.

7.2. Concluding remarks and future direction

Considering the limitations and the results presented in this thesis, we conclude that the estimation procedure in this thesis is impractical to use, compared to the procedure from Mösching and Dümbgen (2024). Even though, the estimators using a small dataset performs reasonably well.

The main problem lies on the running time to obtain the estimated conditional distribution functions While the estimator of Mösching and Dümbgen (2024) produce estimates of the distribution functions in less than a minute, the algorithm in this thesis may took 8 to 10 hours when using a small dataset. Therefore, we would like to explore other algorithms that could reduce the computational time.

Despite these disadvantages, the CRPS scoring rule is one evaluation tool for probabilistic forecasts. A well-predicted event that uses the 'correct' distribution corresponds to a small loss, otherwise the loss is large. We hypothesize that on average, the loss of forecasting new events using the estimators presented here is smaller than Mösching and Dümbgen (2024). However, due to the absence of convexity property of the objective function, this may not be necessarily the true. If we have developed an algorithm that solves the presented optimization problem efficiently, we can conduct an investigation that is presented in the previous section.

Bibliography

Baringhaus, L., & Franz, C. (2004). On a new multivariate two-sample test. *Journal of Multivariate Analysis*, *88*(1), 190–206. https://doi.org/10.1016/S0047-259X(03)00079-4

Barlow, R. E., Bartholomew, D. J., Bremner, J. M., & Brunk, H. D. (1972). *Statistical inference under order restrictions: The theory and application of isotonic regression*. John Wiley & Sons.

Bauer, H. (2001). *Measure and integration theory*. Walter de Gruyter.

Bazaraa, M., Sherali, H., & Shetty, C. (2006). Methods of feasible directions. In *Nonlinear programming* (pp. 537–653). John Wiley & Sons, Ltd. https://doi.org/10.1002/0471787779.ch10

Brent, R. P. (2003). *Algorithms for minimization without derivatives*. Dover Publications.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1–3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

Busam, R., & Freitag, E. (2009). *Complex analysis* (2nd ed.). Springer Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-93983-2

Byrd, R. H., Lu, P., Nocedal, J., & Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, *16*(5), 1190–1208. https://doi.org/10.1137/0916069

Cover, T. M., & Thomas, J. A. (2005). Entropy, relative entropy, and mutual information. In *Elements of information theory* (pp. 13–55). John Wiley & Sons, Ltd. https://doi.org/10.1002/047174882X.ch2

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(2), 187–202. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

Du, D.-Z., & Zhang, X.-S. (1989). Global convergence of rosen's gradient projection method. *Mathematical Programming*, *44*(1), 357–366. https://doi.org/10.1007/BF01587098

Dümbgen, L., & Mösching, A. (2023). On stochastic orders and total positivity. *ESAIM: Probability and Statistics*, *27*, 461–481. https://doi.org/10.1051/ps/2023005

Epstein, E. S. (1969). A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology and Climatology*, *8*(6), 985–987. https://doi.org/10.1175/1520-0450(1969)008<0985:ASSFPF>2.0.CO;2

Fahrmeir, L., Kneib, T., Lang, S., & Marx, B. (2013). *Regression: Models, methods and applications*. Springer Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-34333-9

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–378. https://doi.org/10.1198/016214506000001437

Gneiting, T., & Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, *1*(1), 125–151. https://doi.org/10.1146/annurev-statistics-062713-085831

Gneiting, T., & Raftery, A. E. (2005). Weather forecasting with ensemble methods. *Science*, *310*(5746), 248–249. https://doi.org/10.1126/science.1115255

Gneiting, T., Raftery, A. E., Westveld, A. H., & Goldman, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum crps estimation. *Monthly Weather Review*, *133*(5), 1098–1118. https://doi.org/10.1175/MWR2904.1

Grünwald, P. D., & Dawid, A. P. (2004). Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *The Annals of Statistics*, *32*(4), 1367–1433. https://doi.org/10.1214/009053604000000553

Henzi, A., Ziegel, J. F., & Gneiting, T. (2021). Isotonic distributional regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *83*(5), 963–993. https://doi.org/10.1111/rssb.12450

Kalnay, E. (2002). *Atmospheric modeling, data assimilation and predictability*. Cambridge University Press.

Karim, M. R., & Islam, M. A. (2019). *Reliability and survival analysis*. Springer Singapore. https://doi.org/10.1007/978-981-13-9776-9

Karlin, S. (1968). *Total positivity* (Vol. 1). Stanford University Press.

Matheson, J. E., & Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management Science*, *22*(10), 1087–1096. https://doi.org/10.1287/mnsc.22.10.1087

Meyer, C. D. (2023). *Matrix analysis and applied linear algebra, second edition*. Society for Industrial; Applied Mathematics. https://doi.org/10.1137/1.9781611977448

Mösching, A., & Dümbgen, L. (2022). *LRDistReg: Estimates a family of distributions under likelihood ratio order constraint* [R package version 1.0, commit 5b22f8d2e442f9a9cfa37eaccdcf53bfeca53257]. https://github.com/AlexandreMoesching/LRDistReg

Mösching, A., & Dümbgen, L. (2024). Estimation of a likelihood ratio ordered family of distributions. *Statistics and Computing*, *34*(1). https://doi.org/10.1007/s11222-023-10370-9

Murphy, A. H. (1966). A note on the utility of probabilistic predictions and the probability score in the cost-loss ratio decision situation. *Journal of Applied Meteorology and Climatology*, *5*(4), 534–537. https://doi.org/10.1175/1520-0450(1966)005<0534:ANOTUO>2.0.CO;2

Owen, A. B. (2001). *Empirical likelihood*. Boca Raton, Fla : Chapman & Hall/CRC.

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Robertson, T., Wright, F. T., & Dykstra, R. L. (1988). *Order restricted statistical inference*. John Wiley & Sons.

Rosen, J. B. (1960). The gradient projection method for nonlinear programming. part i. linear constraints. *Journal of the Society for Industrial and Applied Mathematics*, *8*(1), 181–217. Retrieved July 24, 2024, from http://www.jstor.org/stable/2098960

Sason, I. (2022). Divergence measures: Mathematical foundations and applications in information-theoretic and statistical problems. *Entropy*, *24*(5). https://doi.org/10.3390/e24050712

Schefzik, R., Thorarinsdottir, T. L., & Gneiting, T. (2013). Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statistical Science*, *28*(4), 616–640. https://doi.org/10.1214/13-STS443

Selten, R. (1998). Axiomatic characterization of the quadratic scoring rule. *Experimental Economics*, *1*(1), 43–61. https://doi.org/10.1023/A:1009957816843

Shaked, M., & Shanthikumar, J. G. (2007). *Stochastic orders*. Springer New York, NY. https://doi.org/10.1007/978-0-387-34675-5

van der Hofstad, R. (2016). *Random graphs and complex networks* (Vol. 1). Cambridge University Press. https://doi.org/10.1017/9781316779422

von Holstein, C.-A. S. S. (1970). A family of strictly proper scoring rules which are sensitive to distance. *Journal of Applied Meteorology and Climatology*, *9*(3), 360–364. https://doi.org/10.1175/1520-0450(1970)009<0360:AFOSPS>2.0.CO;2

Winkler, R. L., & Murphy, A. H. (1968). “Good” probability assessors. *Journal of Applied Meteorology and Climatology*, *7*(5), 751–758. https://doi.org/10.1175/1520-0450(1968)007<0751:PA>2.0.CO;2

Wu, Y., & Westling, T. (2023). Nonparametric inference under a monotone hazard ratio order. *Electronic Journal of Statistics*, *17*(2), 3181–3225. https://doi.org/10.1214/23-EJS2173

Yeung, R. W. (2002). *A first course in information theory*. Springer New York, NY. https://doi.org/10.1007/978-1-4419-8608-5

A

Additional Information about The Ranked Probability Score

In this appendix, we explain the construction of the ranked probability score (RPS) in detail. The reason for that is to explain why there is a term ”ranked” in the name of this scoring rule. The first lemma in this appendix is found in Epstein (1969). We add small details of the results to understand how the author arrived at their calculations. The second lemma in this appendix is not in Epstein (1969). The lemma is necessary so that it is clear that the combination of these two lemma results in the RPS.

The construction of the RPS uses the utility framework. Consider the following situation given in Murphy (1966). Let $\Omega = \{W, NW\}$, where W is the event of adverse weather and NW is the event of non-adverse weather. Let P and DNP be the action to protect and not to protect respectively. Let C be the cost of taking protective measures and L be the loss of not taking the protective measure. If W occurs, then we should take action P which will cost C. If P is not taken then there will be a loss L. If NW occurs, then protection is unnecessary which yields zero loss and zero cost. If the action to protect is taken anyway, then there will be a cost of C. These scenarios can be summarized in the following cost-loss matrix:

	W	NW
P	C	C
DNP	L	0

If W occurs with probability p, then the expected cost when action P is taken is $pC + (1-p)C = C$. If action DNP is taken, then the expected loss is pL. We then take action depending on which expectation is bigger. If the expected loss is bigger than the expected cost, i.e. $pL > C$ or equivalently $p > C/L$, then we follow the course of action P. If $p < C/L$, then the action DNP is taken.

The quantity C/L is called the cost-loss ratio and it is defined on the interval $(0,1)$ for a decision situation to exist. Indeed, if $C/L > 1$, then we will always take the course of action DNP. If $C/L \leq 0$, then the loss is always larger than the cost and so we always take the action P.

For convenience, we can transform the cost-loss matrix into a utility matrix, so that the preferred outcome has utility $+1$, and otherwise it is zero. Then the cost-loss matrix becomes the following utility matrix:

	W	NW
P	$1-C/L$	$1-C/L$
DNP	0	1

If c_{ij} and u_{ij} are the entries of the cost-loss matrix and the utility matrix respectively. Then the transformation of c_{ij} is defined by $u_{ij} := 1 - c_{ij}/L$. According to this utility matrix, the action of DNP while the weather is W is the least desirable outcome. The most favourable outcome is when the weather is non-adverse (NW) and no protective action is taken (DNP). The utility of taking action P is $1 - C/L$, which is close to $+1$ if the loss L is very large. Lastly, by using the utility matrix and the decision rule, we can define the utility as the following:

$$U = (1 - C/L)\mathbb{1}(p \geq C/L) + \delta\mathbb{1}(p < C/L), \tag{A.1}$$

where $\delta = 1$ if NW occurs, and equal to 0 otherwise.

Epstein (1969) then generalizes the situation described above for a finite sample space. Let $\Omega = \{W_1, \ldots, W_r\}$, such that $r \geq 2$ and the elements are ranked such that the event W_1 is the most adverse weather and W_r is non-adverse weather. Let $A_j, j \in \Omega$ be the action taken such that the protection measure is successively less complete. Instead of a 2×2 cost-loss matrix, we now have a $r \times r$ cost-loss matrix. The entries of this matrix are defined as follows:

$$c_{ij} = \begin{cases} \dfrac{C(r-i)}{r-1}, & \text{if } i \leq j, \\ \\ \dfrac{C(r-i)+L(i-j)}{r-1}, & \text{if } i > j. \end{cases} \tag{A.2}$$

If the decision is to take a full protective measure, while the observed weather is less adverse than predicted, then the total cost comes only from the protection. If the protective measure is not sufficient, since the observed weather is more severe than predicted, then there is a cost from the protection, in addition to the loss. The loss increases as the action i lies further than the action j, which should have been taken when weather j occurs.

The decision rule and the utility matrix are defined in a similar matter as in the case when $r = 2$. The cost-loss matrix in the general case is transformed into the utility matrix by defining the following entries $u_{ij} = 1 - c_{ij}/L$, where c_{ij} is (A.2). The decision rule of taking action k would be

$$\sum_{j=1}^{k-1} p_j < C/L < \sum_{j=1}^{k} p_j.$$

Generalizing (A.1), we express the utility U_j if weather W_j occurs as a function of the utility matrix, a vector $\mathbf{p} = (p_1, \ldots, p_r) \in [0,1]^r$ and $0 < C/L < 1$, which yield

$$U_j = \sum_{i=1}^{r} u_{ij} d_i(\mathbf{p}, C/L),$$

where

$$d_i(\mathbf{p}, C/L) = \begin{cases} 1, & \text{if } \displaystyle\sum_{k=1}^{i-1} p_k < C/L \leq \sum_{k=1}^{i} p_k, \\ 0, & \text{otherwise.} \end{cases}$$

Treating C/L as a random variable which follows a uniform distribution on $(0,1)$, we can then compute the expected utility when the weather W_j occurs. We state the result in the following lemma.

Lemma A.1. *Let C/L be uniformly distributed on interval $(0,1)$. Then,*

$$\mathbb{E}[U_j] = 1 - \frac{1}{2(r-1)} \sum_{i=1}^{r-1} \left[\left(\sum_{k=1}^{i} p_k \right)^2 \right] - \frac{1}{r-1} \sum_{i=j+1}^{r} (i-j) p_i.$$

The maximum value of the expected utility when W_j is observed, is $(r+j-2)/2(r-1)$. Its minimum value is $(j-1)/(r-1)$ in the region $j \leq (r+1)/2$, and it becomes $1/2$ for $j > (r+1)/2$.

Proof. Let $X := C/L$, then,

$$u_{ij} = \begin{cases} 1 - \dfrac{X(r-i)}{r-1}, & \text{if } i \leq j, \\ \\ 1 - \dfrac{X(r-i)+(i-j)}{r-1}, & \text{if } i > j. \end{cases} \tag{A.3}$$

we have

$$\mathbb{E}[U_j] = \int_0^1 \sum_{i=1}^{r} u_{ij} d_i(\mathbf{p}, x)\, dx$$

$$= \frac{1}{r-1}\int_0^1 \left\{\sum_{i=1}^{j}[r-1-x(r-i)]d_i(\mathbf{p},x) + \sum_{i=j+1}^{r}[r-1-x(r-i)+j-i]d_i(\mathbf{p},x)\right\}dx$$

$$= \frac{1}{r-1}\left\{\sum_{i=1}^{j}\int_0^1 [r-1-x(r-i)]d_i(\mathbf{p},x)\,dx + \sum_{i=j+1}^{r}\int_0^1 [r-1-x(r-i)+j-i]d_i(\mathbf{p},x)\,dx\right\}$$

$$= \frac{1}{r-1}\left\{\sum_{i=1}^{r}\int_0^1 [r-1-x(r-i)]d_i(\mathbf{p},x)\,dx + \sum_{i=j+1}^{r}\int_0^1 [j-i]d_i(\mathbf{p},x)\,dx\right\}. \tag{A.4}$$

Let

$$G_k = \sum_{i=1}^{k} p_i, \qquad \text{and therefore} \qquad 1 - G_k = \sum_{i=k+1}^{r} p_i, \tag{A.5}$$

then combining (A.4) and (A.5) yield

$$\mathbb{E}[U_j] = \frac{1}{r-1}\left\{\sum_{i=1}^{r}\int_{G_{i-1}}^{G_i}[r-1-x(r-i)]\,dx + \sum_{i=j+1}^{r}\int_{G_{i-1}}^{G_i}[j-i]\,dx\right\}$$

$$= \frac{1}{r-1}\left\{\sum_{i=1}^{r}\left[(r-1)x - \frac{1}{2}x^2(r-i)\right]_{x=G_{i-1}}^{x=G_i} + \sum_{i=j+1}^{r}[(j-i)x]_{x=G_{i-1}}^{x=G_i}\right\}$$

$$= \frac{1}{r-1}\left\{\sum_{i=1}^{r}\left[(r-1)(G_i - G_{i-1}) - \frac{1}{2}(G_i^2 - G_{i-1}^2)(r-i)\right] + \sum_{i=j+1}^{r}[(j-i)(G_i - G_{i-1})]\right\}$$

$$= 1 - \frac{1}{2(r-1)}\sum_{i=1}^{r}(r-i)\left[G_i^2 - G_{i-1}^2\right] - \frac{1}{r-1}\sum_{i=j+1}^{r}(i-j)p_i$$

$$= 1 - \frac{1}{2(r-1)}\sum_{i=1}^{r-1}\left[\left(\sum_{k=1}^{i}p_k\right)^2\right] - \frac{1}{r-1}\sum_{i=j+1}^{r}(i-j)p_i.$$

The last equality is obtained by the following computation

$$\begin{aligned}\sum_{i=1}^{r}(r-i)\left[G_i^2 - G_{i-1}^2\right] &= (r-1)G_1^2 + (r-2)\left[G_2^2 - G_1^2\right] \\ &\quad + (r-3)\left[G_3^2 - G_2^2\right] + \cdots + 2\left[G_{r-2}^2 - G_{r-3}^2\right] + \left[G_{r-1}^2 - G_{r-2}^2\right] \\ &= G_1^2(r-1-(r-2)) + G_2^2(r-2-(r-3)) + \cdots + G_{r-2}^2(2-1) + G_{r-1}^2 \\ &= \sum_{i=1}^{r-1}G_i^2 \\ &= \sum_{i=1}^{r-1}\left(\sum_{k=1}^{i}p_k\right)^2.\end{aligned}$$

To analyze how $\mathbb{E}[U_j]$ behaves, we first suppose that we predict the weather W_j with 100% confidence and the weather W_j is observed for any $j \in \Omega$. This means for every j we perfectly predict W_j, which corresponds to finding the maximum of the expected utility. Let $(p_1, \dots, p_r)$ be the vector probability such that it is 1 at the j-th position and 0 otherwise. Then, the last term in $\mathbb{E}[U_j]$ vanishes because $p_i = 0$ for any $j+1 \le i \le r$. Further, $G_i = 1$ whenever $j \le i \le r-1$ and $G_i = 0$ otherwise. Hence,

$$\mathbb{E}[U_j] = 1 - \frac{1}{2(r-1)}\sum_{i=j}^{r-1}1 = 1 - \frac{r-j}{2(r-1)} = \frac{r+j-2}{2(r-1)}.$$

As for the minimum value, suppose we predict the observed weather W_j incorrectly with 100% confidence. Consider the following two cases:

Case 1: Let $(p_1, \ldots, p_r)$ be a probability vector such that $p_1 = 1$ and $p_i = 0$ for any $1 < i \leq r$. Suppose we observe an event $1 < j \leq r$. Then,

$$\mathbb{E}[U_j] = 1 - \frac{1}{2(r-1)} \sum_{i=1}^{r-1} 1 = \frac{1}{2}.$$

Case 2: Let $(p_1, \ldots, p_r)$ be a probability vector such that $p_r = 1$ and $p_i = 0$ for any $1 \leq i < r$. Suppose we observe an event $1 \leq j < r$. Then,

$$\mathbb{E}[U_j] = 1 - \frac{1}{r-1} \sum_{i=j+1}^{r} (i-j)p_i = 1 - \frac{r-j}{r-1} = \frac{j-1}{r-1}.$$

Combining these two cases yields the minimum expected utility, which is $\min\{(j-1)/(r-1), 1/2\}$. This value increases with j in the region $j \leq (r+1)/2$, and it becomes a constant for $j > (r+1)/2$. □

From the Lemma A.1, we observe that the maximum and the minimum of the expected utility depend on the subsequent observed weather. The maximum value of $\mathbb{E}[U_j]$ increases as j increases. This means that the utility is at its best when we perfectly predict the most non-adverse weather. Perfectly predicting the worst weather does not increase the utility.

For this reason, Epstein (1969) adjusts the scoring scheme so that it is less dependent on what type of weather occurs. To achieve this, the elements of the sample space are ranked differently. Recall that in the beginning, we rank them from the most adverse to the least adverse weather. Now we rank the elements of the sample space from the least adverse to the most adverse weather. The action sequence $(A_j)_{j=1}^{r}$ is now an action to protect and is subsequently more complete. Then we define u_{ij}^{+} to be the entries of the new utility matrix

$$u_{ij}^{+} = \begin{cases} 1 - \dfrac{C(r-i)}{L(r-1)}, & \text{if } i > j, \\[2ex] 1 - \dfrac{C(r-i) + L(j-i)}{L(r-1)}, & \text{if } i \leq j. \end{cases} \tag{A.6}$$

This is similar to u_{ij}, which is defined in (A.3). The cost-loss ratio C/L is still assumed to be random and it follows from a uniform distribution $(0, 1)$. We then define the following random variable that is similar to U_j:

$$U_j^{+} = \sum_{i=1}^{r} u_{ij}^{+} d_i^{+}(\mathbf{p}, C/L),$$

where

$$d_i^{+}(\mathbf{p}, C/L) = \begin{cases} 1, & \text{if } \displaystyle\sum_{k=i+1}^{r} p_k < C/L \leq \sum_{k=i}^{r} p_k, \\ 0, & \text{otherwise.} \end{cases}$$

Equivalently, $d_i^{+}(\mathbf{p}, C/L) = 1$ if and only if $(1 - G_i) < C/L \leq (1 - G_{i-1})$, and 0 otherwise. We now state what the expected utility U_j^{+} is.

Lemma A.2. *Let C/L be uniformly distributed on interval $(0, 1)$. Then,*

$$\mathbb{E}[U_j^{+}] = 1 - \frac{1}{2(r-1)} \sum_{i=1}^{r-1} \left[\left(\sum_{k=i+1}^{r} p_k \right)^2 \right] - \frac{1}{r-1} \sum_{i=1}^{j} (j-i)p_i.$$

The maximum value of the expected utility when W_j is observed, is $(2r - j - 1)/2(r-1)$. Its minimum value is $(r-j)/(r-1)$ in the region $j \geq (r+1)/2$, and it is $1/2$ for $j < (r+1)/2$.

Proof. The computation in this proof is similar to the ones in Lemma A.1. Let $X := C/L$, then we have

$$\begin{aligned}
\mathbb{E}[U_j^+] &= \int_0^1 \sum_{i=1}^{r} u_{ij}^+ d_i^+(\mathbf{p},x)\,dx \\
&= \frac{1}{r-1}\int_0^1 \left\{ \sum_{i=1}^{j} [r-1-x(r-i)+i-j] d_i^+(\mathbf{p},x) + \sum_{i=j+1}^{r} [r-1-x(r-i)] d_i^+(\mathbf{p},x) \right\} dx \\
&= \frac{1}{r-1}\left\{ \sum_{i=1}^{j} \int_0^1 [i-j] d_i^+(\mathbf{p},x)\,dx + \sum_{i=1}^{r} \int_0^1 [r-1-x(r-i)] d_i^+(\mathbf{p},x)\,dx \right\} \\
&= \frac{1}{r-1}\left\{ \sum_{i=1}^{j} \int_{1-G_i}^{1-G_{i-1}} [i-j]\,dx + \sum_{i=1}^{r} \int_{1-G_i}^{1-G_{i-1}} [r-1-x(r-i)]\,dx \right\} \\
&= \frac{1}{r-1}\left\{ \sum_{i=1}^{j} [i-j](G_i - G_{i-1}) + \sum_{i=1}^{r} \left[(r-1)(G_i - G_{i-1}) - \frac{1}{2}((1-G_i)^2 - (1-G_{i-1})^2)(r-i) \right] \right\} \\
&= \frac{1}{r-1}\left\{ \sum_{i=1}^{j} (i-j)p_i + (r-1) - \frac{1}{2}\sum_{i=1}^{r} \left[((1-G_i)^2 - (1-G_{i-1})^2)(r-i) \right] \right\} \\
&= 1 - \frac{1}{2(r-1)} \sum_{i=1}^{r-1} \left[\left(\sum_{k=i+1}^{r} p_k \right)^2 \right] - \frac{1}{r-1} \sum_{i=1}^{j} (j-i)p_i.
\end{aligned}$$

To confirm the maximum and the minimum value of the expected utility U_j^+, we follow the step that is in Lemma A.1. For the maximum expected utility, let $(p_1, \ldots, p_r)$ be the vector probability such that it is 1 at the j-th position and 0 otherwise. Then, the last term in $\mathbb{E}[U_j]$ vanishes because $p_i = 0$ for any $j+1 \le i \le r$. If the observed outcome is $1 < j \le r$, then,

$$\mathbb{E}[U_j^+] = 1 - \frac{1}{2(r-1)} \sum_{i=1}^{j-1} 1 = 1 - \frac{j-1}{2(r-1)} = \frac{2r-j-1}{2(r-1)}.$$

If $j = 1$, then $\mathbb{E}[U_j^+]$ is simply equal to 1. As for the minimum value, suppose we predict the observed weather W_j incorrectly with 100% confidence. Consider the following two cases:

Case 1: Let $(p_1, \ldots, p_r)$ be a probability vector such that $p_1 = 1$ and $p_i = 0$ for any $1 < i \le r$. Suppose we observe an event $1 < j \le r$. Then, the second term of $\mathbb{E}[U_j^+]$ vanishes. So,

$$\mathbb{E}[U_j^+] = 1 - \frac{1}{r-1} \sum_{i=1}^{j-1} 1 = \frac{r-j}{r-1}.$$

Case 2: Let $(p_1, \ldots, p_r)$ be a probability vector such that $p_r = 1$ and $p_i = 0$ for any $1 \le i < r$. Suppose we observe an event $1 \le j < r$. Then, the last term of $\mathbb{E}[U_j^+] = 0$. So,

$$\mathbb{E}[U_j] = 1 - \frac{1}{2(r-1)} \sum_{i=1}^{r-1} 1 = \frac{1}{2}.$$

Combining these two cases yields the minimum expected utility, which is $\min\{(r-j)/(r-1), 1/2\}$. This value is a constant when $j \le (r+1)/2$, and it decreases in j linearly for $j > (r+1)/2$. □

We observe from Lemma A.2 that the maximal expected utility also depends on the outcome of the weather. To make the score independent of the type of observed weather, Epstein (1969) combines

Lemma A.1 and A.2. To be more precise, the ranked probability score is the result of the two lemmas in this appendix, which we subtract this result from $1 - \mathbb{E}[C/L]$ (von Holstein, 1970). That is,

$$S_{\mathrm{RPS}}(\mathbf{p}, j) = \mathbb{E}[U_j] + \mathbb{E}[U_j^+] - (1 - \mathbb{E}[C/L]).$$

Indeed,

$$\begin{aligned} S_{\mathrm{RPS}}(\mathbf{p}, j) = 1 - \frac{1}{2(r-1)} \sum_{i=1}^{r-1} \left[\left(\sum_{k=1}^{i} p_k \right)^2 \right] - \frac{1}{r-1} \sum_{i=j+1}^{r} (i-j) p_i \\ + 1 - \frac{1}{2(r-1)} \sum_{i=1}^{r-1} \left[\left(\sum_{k=i+1}^{r} p_k \right)^2 \right] - \frac{1}{r-1} \sum_{i=1}^{j} (j-i) p_i - \frac{1}{2} \\ = \frac{3}{2} - \frac{1}{2(r-1)} \sum_{i=1}^{r-1} \left[\left(\sum_{k=1}^{i} p_k \right)^2 + \left(\sum_{k=i+1}^{r} p_k \right)^2 \right] - \frac{1}{r-1} \sum_{i=j+1}^{r} (i-j) p_i - \frac{1}{r-1} \sum_{i=1}^{j} (j-i) p_i \\ = \frac{3}{2} - \frac{1}{2(r-1)} \sum_{i=1}^{r-1} \left[\left(\sum_{k=1}^{i} p_k \right)^2 + \left(\sum_{k=i+1}^{r} p_k \right)^2 \right] - \frac{1}{r-1} \sum_{i=1}^{r} |i-j| p_i. \end{aligned}$$

If C/L is uniformly distributed, then $1 - \mathbb{E}[C/L] = 1/2$. In a case that the distribution of C/L is unknown, von Holstein (1970) showed how to generate a family of ranked probability scores.

B

On Totally Positive Distribution

Recall from (4.1) that we want the density function of $[Y|X = x_s]$ and $[Y|X = x_t]$ to satisfy the following relation:

$$f(y_1|x_s)f(y_2|x_t) - f(y_2|x_s)f(y_1|x_t) \geq 0, \tag{B.1}$$

for all $x_s < x_t$ and $y_1 < y_2$. Let

$$\mathbf{K} := \begin{pmatrix} f(y_1|x_s) & f(y_1|x_t) \\ f(y_2|x_s) & f(y_2|x_t) \end{pmatrix}$$

Then (B.1) is equivalent to

$$\det(\mathbf{K}) \geq 0.$$

This looks similar to a notion of total positivity, which was introduced by Karlin (1968, p. 11).

Definition B.1 (Totally positive function)**.** *Let $E, F \subseteq \mathbb{R}$ and let $K : E \times F \to \mathbb{R}$ be a function. The function K is totally positive of order r (denoted as TP$_r$) if for any*

$$x_1 < \cdots < x_m, y_1 < \cdots < y_m, \qquad x_i \in E, y_i \in F, 1 \leq m \leq r,$$

we have

$$\begin{vmatrix} K(x_1, y_1) & K(x_1, y_2) & \cdots & K(x_1, y_m) \\ K(x_2, y_1) & K(x_2, y_2) & \cdots & K(x_2, y_m) \\ \vdots & \vdots & \ddots & \vdots \\ K(x_m, y_1) & K(x_m, y_2) & \cdots & K(x_m, y_m) \end{vmatrix} \geq 0.$$

To connect the notion of likelihood ratio order and a function being totally positive with order 2, Dümbgen and Mösching (2023) define what is a totally positive distribution of order 2.

Definition B.2 (Totally positive distribution of order 2)**.** *Let $A_1 < A_2$ and $B_1 < B_2$ (element-wise) be any Borel sets in $\mathbb{R}^2$. A probability distribution $\mathbb{P}$ is TP$_2$ distribution if*

$$\mathbb{P}(A_2 \times B_1)\mathbb{P}(A_1 \times B_2) \leq \mathbb{P}(A_1 \times B_1)\mathbb{P}(A_2 \times B_2).$$

Let f be the density function of a probability distribution $\mathbb{P}$ of (X, Y). If f is a TP$_2$ function, then $\mathbb{P}$ is a TP$_2$ distribution. The converse is not true however, unless the (X, Y) is a joint discrete random variable. This is stated in the following theorem, but it is stated as a corollary in Dümbgen and Mösching (2023) with no proof.

Theorem B.3. *Let (X, Y) be a joint discrete random variable with a joint probability mass function $h(x, y) := \mathbb{P}(X = x, Y = y)$. The following statements are equivalent:*

1. *For any $x_1 < x_2$ with $\mathbb{P}(X = x_1), \mathbb{P}(X = x_2) > 0$,*

$$[Y|X = x_1] \leq_{\mathrm{lr}} [Y|X = x_2].$$

2. *The function h is TP$_2$.*

Proof. This follows directly by using the Bayes's theorem. Let $h(y|x) := \mathbb{P}(Y = y|X = x)$. Then statement (1) implies for any $y_1 < y_2$,

$$h(y_1|x_1)h(y_2|x_2) \geq h(y_1|x_2)h(y_2|x_1).$$

This is equivalent to

$$h(y_1, x_1)\mathbb{P}(X = x_1)h(x_2, y_2)\mathbb{P}(X = x_2) \geq h(y_1, x_2)\mathbb{P}(X = x_2)h(x_1, y_2)\mathbb{P}(X = x_1),$$

and so, by removing the common factors, we obtain that h is TP$_2$. The converse can also be shown using the Bayes's theorem, i.e. $h(x, y) = h(y|x)\mathbb{P}(X = x)$. □

C

Isotonic Regression

In this appendix, we review a regression method that models a set of observations such that the fitted curve is non-decreasing. This regression is called the isotonic regression. We will first define isotonic regression and we explain this notion graphically as well. The sources that we use in this appendix are mainly from Barlow et al. (1972) and Robertson et al. (1988).

Before we define isotonic regression, consider the following regression problem. Let $(x_i, y_i)_{i=1}^n$ be a data set with n independent observations, and suppose for all $1 \leq i \leq n$, the random variable $X_i|Y_i \sim \mathcal{N}(\mu(x_i), \sigma^2)$ and known σ^2. The goal is to estimate $\mu(x_i)$ for all $1 \leq i \leq n$, with the restriction that

$$\mu(x_{(1)}) \leq \mu(x_{(2)}) \leq \cdots \leq \mu(x_{(n)}) \quad \text{where } x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}.$$

Let $y_j(x_i)$ be the j-th observed response at x_i and n_i be the number of observed response values at x_i. In the unrestricted case, we can estimate $\mu(x_{(i)})$ by minimizing the following negative log-likelihood function

$$-\log L(\mu(x)|(x_i, y_i)_{i=1}^n, \sigma^2) = \frac{n}{2}\log(2\pi) + \frac{n}{2}\log\sigma^2 + \frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{n_i}(y_i - \mu(x_i))^2.$$

The minimizer of the negative log-likelihood is

$$\mu(x_{(i)}) = \frac{1}{n_i}\sum_{j=1}^{n_i} y_j(x_{(i)}) =: \overline{y}_i \quad \text{for } 1 \leq i \leq k,$$

which is the average response values at $x_{(i)}$. As for the restricted case (i.e. with the monotonicity constraint), we have seen in Chapter 4 that the minimization problem of the unrestricted problem can be rewritten as follows

$$\widehat{\mu(x)} := \underset{\mu(x)}{\arg\min}\left\{\sum_{i=1}^{n} n_i(\overline{y}_i - \mu(x_{(j)}))^2\right\}$$
$$\text{s.t.} \quad \mu(x_{(i)}) \leq \mu(x_{(i+1)}) \qquad \forall\ 1 \leq i < n.$$

This is a particular case of isotonic regression. Below is a formal definition of isotonic regression taken from Robertson et al. (1988, p. 14).

Definition C.1. *Let g be a function on X and w be a positive function on X. An isotonic function g^* on X is an* isotonic regression *of g with weights w if g^* minimizes*

$$\sum_{x \in X}[g(x) - f(x)]^2 w(x),$$

in the class of all isotonic functions f on X. That is, for all isotonic functions f on X,

$$\sum_{x\in X}[g(x) - g^*(x)]^2 w(x) \leq \sum_{x\in X}[g(x) - f(x)]^2 w(x).$$

If we go back to the example where $Y|X$ is assumed to follow a normal distribution, we have $g(x) = \overline{y}_j$ and $w(x) = n_i$. To obtain g^*, we first explain it through graphical construction of g^* (Section C.1. Then, we give an example of a well-known algorithm that determines the isotonic regression (Section C.2).

C.1. Graphical construction: Greatest Convex Minorant

In this section we use Barlow et al. (1972, p. 9 - 13) for the explanation of the graphical construction of g^*. As for the explanation about the greatest convex minorant (GCM), we use Robertson et al. (1988, p. 7).

Let

$$G_j = \sum_{i=1}^{j} g(x_i)w(x_i) \quad \text{and} \quad W_j = \sum_{i=1}^{j} w(x_i),$$

where $x_1 < x_2 < ... < x_n$. For all $1 \leq j \leq n$, plot the points $P_j = (W_j, G_j)$, and $P_0 := (0,0)$ and connect these points. The slope between P_j and P_{j-1} is

$$\frac{G_j - G_{j-1}}{W_j - W_{j-1}} = g(x_j).$$

The plot with segments that join the points P_j and P_{j-1} is called the *cumulative sum diagram* (CSD). It turns out that g^* is the slope of the *greatest convex minorant* (GCM) of the CSD. Let $G^*(t)$ be the GCM of the CSD on $[0, W_n]$, then at t, the function value $G^*(t)$ is the supremum of all convex functions on $[0, W_n]$ such that $G^*(t) \leq G(t)$. Let $G^*(W_j) := G_j^*$ and let $P_j^* := (W_j, G_j^*)$, then the CSD and GCM at P_j are the same, i.e. $G_j^* = G_j$. An example of a CSD and its GCM is given in Figure C.1. To obtain $g^*(x_j)$ from this plot, we simply compute the slope of the segment between P_j^* and P_{j-1}^*, i.e.

$$\frac{G_j^* - G_{j-1}^*}{W_j - W_{j-1}} = g^*(x_j).$$

Note that if $G_j^* < G_j$, then $g^*(x_{j+1}) = g^*(x_j)$ for $j = 1, \dots, n-1$ and the last point $G_n = G_n^*$. In Figure C.1, we therefore have $g^*(x_1) = g^*(x_2) = g^*(x_3) < g^*(x_4) = g^*(x_5)$.

The following result shows that the isotonic regression g^* of g on X is unique (see Theorem 1.1 in Barlow et al. (1972) or Theorem 1.2.1 in Robertson et al. (1988)).

Theorem C.1. *Let $x_1 < \cdots < x_n$. Then, g^* is the isotonic regression of g. Indeed, if $f(x)$ is isotonic on x, then*

$$\sum_{i=1}^{n}[g(x_i) - f(x_i)]^2 w(x_i) \geq \sum_{i=1}^{n}[g(x_i) - g^*(x_i)]^2 w(x_i) + \sum_{i=1}^{n}[g^*(x_i) - f(x_i)]^2 w(x_i). \tag{C.1}$$

The isotonic regression is unique.

Proof. We have

$$\begin{aligned}\sum_{i=1}^{n}[g(x_i) - f(x_i)]^2 w(x_i) &= \sum_{i=1}^{n}[g(x_i) - g^*(x_i) + g^*(x_i) - f(x_i)]^2 w(x_i)\\ &= \sum_{i=1}^{n}[g(x_i) - g^*(x_i)]^2 w(x_i) + \sum_{i=1}^{n}[g^*(x_i) - f(x_i)]^2 w(x_i)\\ &\quad + 2\sum_{i=1}^{n}[g(x_i) - g^*(x_i)][g^*(x_i) - f(x_i)]w(x_i).\end{aligned}$$

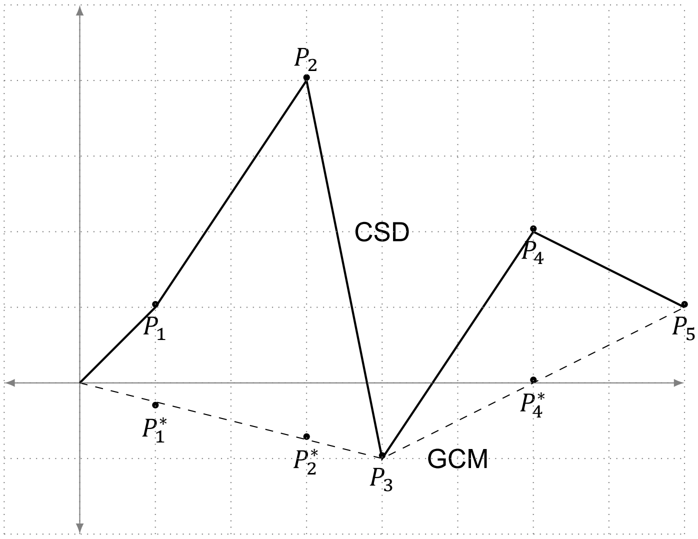

Figure C.1: The solid line is the cumulative sum diagram (CSD) and the dashed line is the greatest convex minorant (GCM) of the CSD. The coordinates of P_j's are (W_j, G_j) and the coordinates of P_j^*'s are (W_j, G_j^*), for all $j \in \{1, 2, 3, 4, 5\}$. In this instance, we have $P_3^* = P_3$ and $P_5^* = P_5$.

So we need to show that

$$\sum_{i=1}^{n} [g(x_i) - g^*(x_i)][g^*(x_i) - f(x_i)]w(x_i) \geq 0, \tag{C.2}$$

for all isotonic functions f. We apply Abel's partial sum formula, which is also stated and proven in Lemma E.3 in Appendix E. We let

$$A_i = \sum_{k=1}^{i-1} [g(x_k) - g^*(x_k)]w(x_k) = G_{i-1} - G_{i-1}^*,$$

and $b_i = [g^*(x_i) - f(x_i)]$, for $1 \leq i \leq n$. Then,

$$\sum_{i=1}^{n} [g(x_i) - g^*(x_i)][g^*(x_i) - f(x_i)]w(x_i) = \sum_{i=1}^{n} [G_{i-1} - G_{i-1}^*]\{[f(x_i) - f(x_{i-1})] - [g^*(x_i) - g^*(x_{i-1})]\}$$
$$+[G_n - G_n^*][g^*(x_n) - f(x_n)],$$

where we set $f(x_0) = g^*(x_0) = G_0 = G_0^* = 0$. Because $P_n = P_n^*$, we have that $G_n = G_n^*$. Because G_{i-1}^* is the GCM of a CSD, we have $G_{i-1}^* < G_{i-1}$ which implies $g^*(x_i) = g^*(x_{i-1})$. Hence, $[G_{i-1} - G_{i-1}^*][g^*(x_i) - g^*(x_{i-1})] = 0$ for all $1 \leq i \leq n$. Lastly, $G_{i-1}^* \leq G_{i-1}$ and $f(x_{i-1}) \leq f(x_i)$ for all $1 \leq i \leq n$, the inequality (C.2) is true.

To verify the uniqueness, let $g^\#$ to be another isotonic regression that minimizes

$$\sum_{i=1}^{n} [g(x_i) - f(x_i)]^2 w(x_i),$$

over all isotonic functions f on x. Since both g^* and $g^\#$ minimizes the sum of squares,

$$\sum_{i=1}^{n} [g(x_i) - g^*(x_i)]^2 w(x_i) = \sum_{i=1}^{n} [g(x_i) - g^\#(x_i)]^2 w(x_i). \tag{C.3}$$

By (C.1),

$$\sum_{i=1}^{n}[g(x_i) - g^{\#}(x_i)]^2 w(x_i) \geq \sum_{i=1}^{n}[g(x_i) - g^{*}(x_i)]^2 w(x_i) + \sum_{i=1}^{n}[g^{*}(x_i) - g^{\#}(x_i)]^2 w(x_i). \tag{C.4}$$

Subtract both sides of the inequality in (C.4) by $\sum_{i=1}^{n}[g(x_i) - g^{*}(x_i)]^2 w(x_i)$ and use (C.3) yields,

$$\sum_{i=1}^{n}[g^{*}(x_i) - g^{\#}(x_i)]^2 w(x_i) \leq 0.$$

Because $w(x_i) > 0$ and $[g^{*}(x_i) - g^{\#}(x_i)]^2 \geq 0$, we have $g^{*}(x_i) = g^{\#}(x_i)$ for all $1 \leq i \leq n$. □

C.2. The Pool-Adjacent-Violator Algorithm (PAVA)

From Figure C.1, we observe that the construction of the GCM of the CSD can be done by replacing line segments with larger slope with a smaller slope on an interval $[W_{j-2}, W_j]$ such that $G_{j-1} > G_j$. Indeed, on $[W_3, W_5]$, the point P_4 lies higher than P_5, which means $G_4 > G_5$. The GCM of on this interval is the line segment P_3P_5. On $[W_1, W_3]$, we have $G_2 > G_3$, which means that we can replace the segment P_2P_3 by a line segment P_1P_3. However, the slope of the line segment P_1P_3 is still larger than the line segment P_0P_3. Hence, the GCM of this particular CSD consists of the line segments P_0P_3 and P_3P_5. The slopes of these line segments are the isotonic regression.

The PAVA algorithm is widely used to determine the isotonic regression. It starts with the given g, which is the solution if g is isotonic. Otherwise, there exists an i such that $g(x_{i-1}) > g(x_i)$. The idea is to pool x_{i-1} and x_i to create a block $\{x_{i-1}, x_i\}$. Then, compute the weighted average

$$\mathrm{Av}(i-1, i) := \frac{g(x_{i-1})w(x_{i-1}) + g(x_i)w(x_i)}{w(x_{i-1}) + w(x_i)} = \frac{G_i - G_{i-2}}{W_i - W_{i-2}},$$

which is equivalently the slope of the line segment $P_{i-2}P_i$. If after the first iteration, there exists another $j < i$ such that g_j is, say, strictly larger than $\mathrm{Av}(i-1, i)$, then we compute again the weighted average but with weight $w(x_{i-1}) + w(x_i)$. This value is in fact equivalent as the slope of the line segment $P_{j-1}P_i$. The procedure is repeated until there is no more violator.

Example C.1. To illustrate this algorithm, we give an example of data set that is used to produce the CSD in Figure C.1. We already sort the values $(x_i)_{i=1}^{5}$ from the smallest to the largest. The data set is given in Table C.1. We are unable to set $g(x_j) = g^{*}(x_j)$ for all j because it is not isotonic. We have

Table C.1: Example data set used for plotting the CSD in Figure C.1

j	$w(x_j)$	$g(x_j)$	W_j	G_j
1	1	1	1	1
2	2	3/2	3	4
3	1	-5	4	-1
4	2	3/2	6	2
5	2	-1/2	8	1

$g(x_2) > g(x_3)$ and $g(x_4) > g(x_5)$ which violate the isotonicity constraint. So we pool data points that are non-increasing. In our examples, the pools are $\{2, 3\}$ and $\{4, 5\}$. Now, we compute the weighted average of the pooled data points:

$$\frac{g(x_2)w(x_2) + g(x_3)w(x_3)}{w(x_2) + w(x_3)} = \frac{G_3 - G_1}{W_3 - W_1} = -\frac{2}{3}$$
$$\frac{g(x_4)w(x_4) + g(x_5)w(x_5)}{w(x_4) + w(x_5)} = \frac{G_5 - G_3}{W_5 - W_3} = \frac{1}{2}$$

These values are the slopes that connect the segments P_1P_3 and P_3P_5 respectively. Next, we replace $g(x_2)$ and $g(x_3)$ with $g_1^{*}(x_2) = g_1^{*}(x_3) := -2/3$, $g(x_4)$ and $g(x_5)$ with $g_1^{*}(x_4) = g_1^{*}(x_5) := 1/2$ and

Table C.2: The new data points after the first iteration. There is still a violator, since the average in pool $\{1\}$ is still larger than the average in pool $\{2, 3\}$.

Pools	Weights	Average
{1}	1	1
{2,3}	3	-2/3
{4,5}	4	1/2

$g_1^*(x_j) = g(x_j)$ otherwise. The weights of the new values accumulate, and so the new weights are $w_1(x_2) = w_1(x_3) := w(x_2) + w(x_3) = 3$, and similarly, $w_1(x_4) = w_1(x_5) := w_1(x_4) + w_1(x_5) = 4$, otherwise $w_1(x_j) = w_1(x_j)$. The result of this iteration is summarized in Table C.2.

The procedure that we have done in the first iteration is repeated until the isotonicity constraint of g on x is satisfied. Since $g_1^*(x_1) > g_1^*(x_2)$ still violates the isotonicity constraint, we compute again the weighted average with the newly computed data points. We have

$$\frac{g_1^*(x_1)w_1(x_1) + g_1^*(x_2)w_1(x_2)}{w_1(x_1) + w_1(x_2)} = \frac{1 \cdot 1 + (-2/3) \cdot 3}{4} = -\frac{1}{4}.$$

Hence, the isotonic regression g^* is defined as follows

$$g^*(x_1) = g^*(x_2) = g^*(x_3) = -1/4, \text{ and } g^*(x_4) = g^*(x_5) = 1/2.$$

△

D

Optimization Theory: The Gradient Projection Method

This appendix discusses the numerical optimization method we use in Chapter 6. Recall that we want to minimize a function subject to inequality constraints. Due to these constraints, the steepest direction method may lead to infeasible points. We use the gradient projection method to ensure that the next iterated point remains in the feasible set and it improves the objective function.

We start with a recap of basic notions of Karush-Kuhn-Tucker (KKT) optimality conditions. Then we discuss the algorithm of the gradient project method. We specifically discuss the following optimization problem:

$$\begin{aligned} &\min f(\mathbf{x}) \\ &\text{s.t. } \mathbf{A}\mathbf{x} \leq \mathbf{0}, \end{aligned}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable, $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{A} \in \mathbb{R}^{n\times d}$. We denote the minimization problem above as problem P. We use $\mathbf{I}$ as the notation for the identity matrix with a suitable dimension. We use Bazaraa et al. (2006) for the main source of this appendix.

First, we define feasible direction, improving feasible direction, and active (binding) constraints.

Definition D.1. *Consider the problem* P*, and a feasible set* S*.*

- *A vector* $\mathbf{d} \neq \mathbf{0}$ *is a* feasible direction *at* $\mathbf{x} \in S$ *if there exists a* $\delta > 0$ *such that for any* $\lambda \in (0, \delta)$*, we have* $\mathbf{x} + \lambda\mathbf{d} \in S$ *(Bazaraa et al., 2006, p. 538 - 539).*
- *A vector* $\mathbf{d}$ *is an* improving feasible direction *at* $\mathbf{x} \in S$*, if* $\mathbf{d}$ *is a feasible direction, then* $f(\mathbf{x} + \lambda\mathbf{d}) < f(\mathbf{x})$ *(Bazaraa et al., 2006, p. 538 - 539).*
- *Let* $\mathbf{a}_1, \dots, \mathbf{a}_n$ *denote the row vectors of matrix* $\mathbf{A}$ *from problem* P*. The* active constraints *or* binding constraints *is the set of indices* $\{i = 1, \dots, n : \langle \mathbf{a}_i^\top, \mathbf{x}\rangle = 0\}$ *(Bazaraa et al., 2006, p. 177).*

Lastly, we state a result of the Karush-Kuhn-Tucker (KKT) optimality condition for constrained optimization problem P. We first state the necessary KKT conditions, followed by the sufficient condition. The theorem below will not be proven, and we refer the reader to Bazaraa et al. (2006, p. 190) and Bazaraa et al. (2006, p. 195) for the proofs.

Theorem D.2. *Consider the problem* P *and let* $\mathbf{x}^*$ *be a feasible solution of* P*. Let* I *be the set of binding constraints and* $(\mathbf{a}_i)_{i=1}^n$ *denote the row vectors of matrix* $\mathbf{A}$*. Suppose* $(\mathbf{a}_i)_{i\in I}$ *are linearly independent. If* $\mathbf{x}^*$ *is a local solution of* P*, then there exists* $u_i \geq 0$*,* $i \in I$ *such that*

$$\nabla f(\mathbf{x}^*) + \sum_{i\in I} u_i \mathbf{a}_i^\top = \mathbf{0}.$$

The above condition is referred to as the KKT condition. If f *is a convex function on the set* $\{\mathbf{x} : \mathbf{A}\mathbf{x} \leq \mathbf{0}\}$*, then the KKT conditions are sufficient for the optimality condition.*

D.1. Generating improving feasible direction

In an unconstrained optimization with differentiable objective function, one can solve the problem by using the steepest descent. Let $\mathbf{d} \in \mathbb{R}^d$ such that

$$\lim_{\lambda \to 0^+} \frac{f(\mathbf{x} + \lambda \mathbf{d}) - f(\mathbf{x})}{\lambda} < 0,$$

then $\mathbf{d}$ is called the direction of descent. One can show that for

$$\mathbf{d} := -\frac{\nabla f(\mathbf{x})}{||\nabla f(\mathbf{x})||},$$

then the direction is the steepest descent. In each iteration of the steepest descent algorithm, it computes the direction $-\nabla f(\mathbf{x})$ and subsequently a line search is performed along this direction.

For solving the problem P, we want that the next iterated point stays feasible. To do this, instead of moving along $-\nabla f(\mathbf{x})$, we let $\mathbf{d} := -\mathbf{P}\nabla f(\mathbf{x})$, where $\mathbf{P}$ is a suitable projection matrix. Suppose $\mathbf{P}\nabla f(\mathbf{x}) \neq \mathbf{0}$ and $\mathbf{A}^\top = (\mathbf{A}_1^\top, \mathbf{A}_2^\top)$ such that $\mathbf{A}_1\mathbf{x} = \mathbf{0}, \mathbf{A}_2\mathbf{x} < \mathbf{0}$, and $\mathbf{A}_1^\top$ is a full-rank matrix. Let $\mathbf{P} = \mathbf{I} - \mathbf{A}_1^\top(\mathbf{A}_1\mathbf{A}_1^\top)^{-1}\mathbf{A}_1$, then we observe $\mathbf{P}$ is a projection matrix, i.e. this matrix is symmetric and $\mathbf{P}^2 = \mathbf{P}$. The matrix $\mathbf{A}_1\mathbf{A}_1^\top$ is also non-singular because $\mathbf{A}_1\mathbf{A}_1^\top$ because $\mathbf{A}_1^\top$ is a full-ranked matrix. We further observe that $\mathbf{A}_1\mathbf{P} = \mathbf{0}$, meaning that $\mathbf{P}$ projects the gradient of the binding constraints into $\mathbf{0}$. Therefore, $\mathbf{A}_1\mathbf{P}\nabla f(\mathbf{x}) = \mathbf{0}$ which implies that $\mathbf{P}f(\mathbf{x})$ is a vector in the null space of $\mathbf{A}_1$. Hence the matrix $\mathbf{P}$ projects $\nabla f(\mathbf{x})$ onto the null space of $\mathbf{A}_1$.

The following lemma shows that for $\mathbf{d} = -\mathbf{P}\nabla f(\mathbf{x})$ and $\mathbf{x}$ satisfies $\mathbf{A}\mathbf{x} \leq \mathbf{0}$, the objective function at $\mathbf{x} + \lambda\mathbf{d}$ is improved and it remains feasible.

Lemma D.1. *Consider the problem* P *that we want to solve. Let* $\mathbf{x}$ *be a feasible point and let* $\mathbf{A}^\top = (\mathbf{A}_1^\top, \mathbf{A}_2^\top)$ *such that* $\mathbf{A}_1\mathbf{x} = \mathbf{0}$ *and* $\mathbf{A}_2\mathbf{x} < \mathbf{0}$. *Let* $\mathbf{P}$ *be a projection matrix such that* $\mathbf{P}\nabla f(\mathbf{x}) \neq \mathbf{0}$ *and* $\mathbf{d} := -\mathbf{P}\nabla f(\mathbf{x})$*, then there exists* $\delta > 0$ *such that*

$$f(\mathbf{x} + \lambda\mathbf{d}) < f(\mathbf{x}) \qquad \text{for all } 0 < \lambda < \delta. \tag{D.1}$$

Let $\mathbf{A}_1$ *and assume* $\operatorname{rank}(\mathbf{A}_1) = n$. *If* $\mathbf{P} = \mathbf{I} - \mathbf{A}_1^\top(\mathbf{A}_1\mathbf{A}_1^\top)^{-1}\mathbf{A}_1$*, then* $\mathbf{d}$ *satisfies* (D.1) *and* $\mathbf{x} + \lambda\mathbf{d}$ *is a feasible point for all* $\lambda \in (0, \delta)$.

Proof. Because f is differentiable, we have

$$\lim_{\lambda \to 0^+} \frac{f(\mathbf{x} + \lambda\mathbf{d}) - f(\mathbf{x})}{\lambda} = \langle \nabla f(\mathbf{x}), \mathbf{d} \rangle.$$

Indeed, by applying first-order Taylor expression of the expression in the limit around $\mathbf{x}$ we have

$$\frac{f(\mathbf{x} + \lambda\mathbf{d}) - f(\mathbf{x})}{\lambda} = \langle \nabla f(\mathbf{x}), \mathbf{d} \rangle + O(\lambda^2),$$

and let $\lambda \to 0^+$. Since $\mathbf{P}$ is idempotent and symmetric, we have

$$\langle \nabla f(\mathbf{x}), \mathbf{d} \rangle = -\nabla f(\mathbf{x})^\top \mathbf{P} \nabla f(\mathbf{x}) = -\nabla f(\mathbf{x})^\top \mathbf{P}^\top \mathbf{P} \nabla f(\mathbf{x}) = -||\mathbf{P}\nabla f(\mathbf{x})||^2 < 0.$$

This implies that $\mathbf{d} = -\mathbf{P}\nabla f(\mathbf{x})$ is an improving direction. If $\mathbf{P} = \mathbf{I} - \mathbf{A}_1^\top(\mathbf{A}_1\mathbf{A}_1^\top)^{-1}\mathbf{A}_1$, then

$$\mathbf{A}_1\mathbf{d} = -\mathbf{A}_1\mathbf{P}\nabla f(\mathbf{x}) = -\mathbf{A}_1(\mathbf{I} - \mathbf{A}_1^\top(\mathbf{A}_1\mathbf{A}_1^\top)^{-1}\mathbf{A}_1)\nabla f(\mathbf{x}) = \mathbf{0}.$$

Therefore for some $\delta > 0$, we have $\mathbf{A}_1(\mathbf{x} + \lambda\mathbf{d}) = \mathbf{0}$ for any $\lambda \in (0, \delta)$. Since $\mathbf{A}_2(\mathbf{x} + \lambda\mathbf{d}) < \mathbf{0}$ remains true for small enough λ, we conclude that $\mathbf{x} + \lambda\mathbf{d} \in S$, where $S = \{x \in \mathbb{R}^d : A\mathbf{x} \leq \mathbf{0}\}$. □

In case when $\mathbf{d} = \mathbf{0}$ at $\mathbf{x}$, then we either have a point that satisfies a KKT conditions, or we need to adjust the projection matrix to compute a new improving feasible direction. Indeed, assume $\mathbf{P}\nabla f(\mathbf{x}) = \mathbf{0}$ and let $\mathbf{u} := -(\mathbf{A}_1\mathbf{A}_1^\top)^{-1}\mathbf{A}_1\nabla f(\mathbf{x})$, then

$$\mathbf{P}\nabla f(\mathbf{x}) = (\mathbf{I} - \mathbf{A}_1^\top(\mathbf{A}_1\mathbf{A}_1^\top)^{-1}\mathbf{A}_1)\nabla f(\mathbf{x}) = \nabla f(\mathbf{x}) + \mathbf{A}_1^\top\mathbf{u} = \mathbf{0}. \tag{D.2}$$

We observe that if $\mathbf{u} \geq \mathbf{0}$, then $\mathbf{x}$ satisfies the KKT conditions. If $\mathbf{u} \ngeq \mathbf{0}$ then we will choose an entry u_j such that $u_j < 0$. The corresponding row of $\mathbf{A}_1$ is then removed to create a new matrix $\widehat{\mathbf{A}}_1$. Let $\widehat{\mathbf{P}}$ be the new projection matrix that is constructed similarly but replacing $\mathbf{A}_1$ with $\widehat{\mathbf{A}}_1$. Then the direction $-\mathbf{P}\nabla f(\mathbf{x})$ is an improving feasible direction.

Lemma D.2. *Consider the problem* P *that we want to solve. Let* $\mathbf{x}$ *be a feasible point and let* $\mathbf{A}^\top = (\mathbf{A}_1^\top, \mathbf{A}_2^\top)$ *such that* $\mathbf{A}_1\mathbf{x} = \mathbf{0}$ *and* $\mathbf{A}_2\mathbf{x} < \mathbf{0}$*. Let* $\mathbf{u} := -(\mathbf{A}_1\mathbf{A}_1^\top)^{-1}\mathbf{A}_1\nabla f(\mathbf{x})$ *and* $\mathbf{A}_1$ *is a full-rank matrix. Suppose* $\mathbf{u} \neq 0$ *and* u_j *is the negative entry of* $\mathbf{u}$*. Let* $\widehat{\mathbf{A}}_1$ *be a matrix that is obtained from* $\mathbf{A}_1$ *by removing the* j*-th row* $\mathbf{A}_1$*. If* $\widehat{\mathbf{P}} := \mathbf{I} - \widehat{\mathbf{A}}_1^\top(\widehat{\mathbf{A}}_1\widehat{\mathbf{A}}_1^\top)^{-1}\widehat{\mathbf{A}}_1$*, then* $\mathbf{d} = -\widehat{\mathbf{P}}\nabla f(\mathbf{x})$ *is an improving feasible direction.*

Proof. Assume $\mathbf{u} \ngeq \mathbf{0}$ and let $u_j < 0$, which is the j-th entry of $\mathbf{u}$. Let $\widehat{\mathbf{P}}$ be the projection matrix as defined in the statement. We start by proving that $\widehat{\mathbf{P}}\nabla f(\mathbf{x}) \neq \mathbf{0}$ by contradiction. Suppose $\widehat{\mathbf{P}}\nabla f(\mathbf{x}) = \mathbf{0}$. Let $\widehat{\mathbf{u}} = -(\widehat{\mathbf{A}}_1\widehat{\mathbf{A}}_1^\top)^{-1}\widehat{\mathbf{A}}_1\nabla f(\mathbf{x})$, then

$$\widehat{\mathbf{P}}\nabla f(\mathbf{x}) = (\mathbf{I} - \widehat{\mathbf{A}}_1^\top(\widehat{\mathbf{A}}_1\widehat{\mathbf{A}}_1^\top)^{-1}\widehat{\mathbf{A}}_1)\nabla f(\mathbf{x}) = \nabla f(\mathbf{x}) + \widehat{\mathbf{A}}_1^\top\widehat{\mathbf{u}} = \mathbf{0}.$$

Note that from (D.2), we have

$$\nabla f(\mathbf{x}) + \mathbf{A}_1^\top\mathbf{u} = \nabla f(\mathbf{x}) + \widehat{\mathbf{A}}_1^\top\mathbf{u}^* + u_j\mathbf{r}_j^\top = \mathbf{0}, \tag{D.3}$$

where $\mathbf{r}_j$ is the j-th row of $\mathbf{A}_1$. This means that

$$\nabla f(\mathbf{x}) + \widehat{\mathbf{A}}_1^\top\widehat{\mathbf{u}} - \left(\nabla f(\mathbf{x}) + \widehat{\mathbf{A}}_1^\top\mathbf{u}^* + u_j\mathbf{r}_j^\top\right) = \widehat{\mathbf{A}}_1^\top\left(\widehat{\mathbf{u}} - \mathbf{u}^*\right) - u_j\mathbf{r}_j^\top = \mathbf{0}.$$

Because $u_j \neq 0$, it implies that the rows of $\mathbf{A}_1$ are linearly dependent. It contradicts the assumption that A_1 is a full-rank matrix. So $\widehat{\mathbf{P}}\nabla f(\mathbf{x}) \neq \mathbf{0}$ and therefore $\mathbf{d} = -\widehat{\mathbf{P}}\nabla f(\mathbf{x})$ is an improving direction by Lemma D.1.

It remains to show that for some $\delta > 0$, a point $\mathbf{x} + \lambda\mathbf{d}$ satisfies $\mathbf{Ax} \leq \mathbf{0}$ for any $\lambda \in (0, \delta)$. We have that

$$\widehat{\mathbf{A}}_1\mathbf{d} = -\widehat{\mathbf{A}}_1\widehat{\mathbf{P}}\nabla f(\mathbf{x}) = -\widehat{\mathbf{A}}_1(\mathbf{I} - \widehat{\mathbf{A}}_1^\top(\widehat{\mathbf{A}}_1\widehat{\mathbf{A}}_1^\top)^{-1}\widehat{\mathbf{A}}_1)\nabla f(\mathbf{x}) = \mathbf{0}.$$

This yields

$$\mathbf{A}_1\mathbf{d} = \begin{pmatrix}\widehat{\mathbf{A}}_1\mathbf{d}\\ \mathbf{r}_j\mathbf{d}\end{pmatrix} = \begin{pmatrix}\mathbf{0}\\ \mathbf{r}_j\mathbf{d}\end{pmatrix}.$$

It is therefore sufficient to show that $\mathbf{r}_j\mathbf{d} \leq \mathbf{0}$. Multiply both sides of Equation (D.3) by $\mathbf{r}_j\widehat{\mathbf{P}}$ yields

$$\mathbf{r}_j\widehat{\mathbf{P}}\nabla f(\mathbf{x}) + \mathbf{r}_j\widehat{\mathbf{P}}\widehat{\mathbf{A}}_1^\top\mathbf{u}^* + u_j\mathbf{r}_j\widehat{\mathbf{P}}\mathbf{r}_j^\top = \mathbf{0} \Leftrightarrow -\mathbf{r}_j\mathbf{d} + u_j\mathbf{r}_j\widehat{\mathbf{P}}\mathbf{r}_j^\top = 0.$$

Because the projection matrix $\widehat{\mathbf{P}}$ is positive semi-definite and $u_j < 0$, we have that $r_j\mathbf{d} \leq \mathbf{0}$. Hence $\mathbf{A}_1\mathbf{d} \leq \mathbf{0}$. We conclude that $\mathbf{d}$ is a feasible direction. □

D.2. Line search for the next iteration

We have seen how a projection matrix is constructed such that the next iterated point stays feasible and it improves the objective function. Let $\mathbf{x}_s$ be the result of the s-th iteration such that it is feasible and $\mathbf{d}_s$ is the direction vector. The next iterated point is defined as $\mathbf{x}_{s+1} = \mathbf{x}_s + \lambda_s\mathbf{d}_s$. In this section, we formulate the optimization problem for finding λ_s.

We solve the following one-dimensional optimization problem:

$$\begin{aligned} \min\ & f(\mathbf{x}_s + \lambda\mathbf{d}_s)\\ \text{s.t}\ & \mathbf{A}(\mathbf{x}_s + \lambda\mathbf{d}_s) \leq \mathbf{0}\\ & \lambda \geq 0. \end{aligned}$$

The optimization problem can in fact be relaxed such that there is only one constraint. Let $\mathbf{A}^\top$ be decomposed into $(\mathbf{A}_1^\top, \mathbf{A}_2^\top)$ in the same way as in Lemma D.1. Since $\mathbf{A}_1\mathbf{x}_s = \mathbf{0}$, and we have in the proof of Lemma D.2 that $\mathbf{A}_1\mathbf{d} \leq \mathbf{0}$, we conclude $\mathbf{A}_1(\mathbf{x}_s + \lambda\mathbf{d}_s) \leq \mathbf{0}$ for all $\lambda \geq 0$. Next, we need $\lambda \geq 0$ such that $\mathbf{A}_2(\mathbf{x}_s + \lambda\mathbf{d}_s) \leq \mathbf{0}$, i.e. $\lambda\mathbf{A}_2\mathbf{d}_s \leq -\mathbf{A}_2\mathbf{x}_s$. Because $\mathbf{A}_2\mathbf{x}_s \leq \mathbf{0}$, if $\mathbf{A}_2\mathbf{d}_s \leq \mathbf{0}$, then $\lambda\mathbf{A}_2\mathbf{d}_s \leq -\mathbf{A}_2\mathbf{x}_s$ is true for any $\lambda \geq 0$. If $\mathbf{A}_2\mathbf{d}_s \nleq \mathbf{0}$, then we need to choose λ such that $\lambda \leq (-\mathbf{A}_2\mathbf{x}_s)_j/(\mathbf{A}_2\mathbf{d}_s)_j$ such that the j-th entry of $\mathbf{A}_2\mathbf{d}_s$ is positive. Therefore, the line search problem becomes

$$\begin{aligned} \min\ & f(\mathbf{x}_s + \lambda\mathbf{d}_s)\\ \text{s.t}\ & 0 \leq \lambda \leq \lambda_{\max}, \end{aligned}$$

where

$$\lambda_{\max} = \begin{cases} \min\{\hat{b}_i/\hat{d}_i : \hat{d}_i > 0\} & \text{if } \hat{\mathbf{d}} \nleq \mathbf{0}, \\ \infty & \text{if } \hat{\mathbf{d}} \leq \mathbf{0}. \end{cases}$$

and

$$\hat{\mathbf{b}} := -\mathbf{A}_2\mathbf{x}_s; \qquad \hat{\mathbf{d}} = \mathbf{A}_2\mathbf{d}_s.$$

E

Other Technical Computations

Lemma E.1. *Given a data set* $\mathcal{D}_n = (x_i, y_i)_{i=1}^n$, *let*

$$\{x_1, \dots, x_n\} = \{x_{(1)}, \dots, x_{(\ell)}\} \quad \textit{and} \quad \{y_1, \dots, y_n\} = \{y_{(1)}, \dots, y_{(m)}\},$$

where $x_{(1)} < \dots < x_{(\ell)}$ *and* $y_{(1)} < \dots < y_{(m)}$ *for some* $1 \leq \ell, m \leq n$. *Let* $w_{jk} = \#\{i : (x_i, y_i) = (x_{(j)}, y_{(k)})\}$. *Then*

$$\widehat{\alpha} = \overline{y} - \beta\overline{x}, \quad \widehat{\beta} = \frac{\overline{xy} - \overline{y} \cdot \overline{x}}{\overline{x^2} - \overline{x}}, \quad \textit{and} \quad \widehat{\sigma^2} = \frac{1}{n} \sum_{j=1}^{\ell} \sum_{k=1}^{m} w_{jk} (y_{(k)} - \widehat{\alpha} - \widehat{\beta} x_{(j)})^2,$$

solve the following optimization problem:

$$\operatorname*{arg\,max}_{(\alpha,\beta,\sigma^2) \in \mathbb{R}^2 \times R_{>0}} \left\{ -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{j=1}^{\ell} \sum_{k=1}^{m} w_{jk} (y_{(k)} - \alpha - \beta x_{(j)})^2 \right\}.$$

Proof. We obtain the following partial derivatives:

$$\begin{aligned}
\frac{\partial}{\partial \alpha} \mathcal{L}(\alpha, \beta, \sigma^2) &= \frac{1}{\sigma^2} \sum_{j=1}^{\ell} \sum_{k=1}^{m} w_{jk} (y_{(k)} - \alpha - \beta x_{(j)}) \\
\frac{\partial}{\partial \beta} \mathcal{L}(\alpha, \beta, \sigma^2) &= \frac{1}{\sigma^2} \sum_{j=1}^{\ell} \sum_{k=1}^{m} w_{jk} x_{(j)} (y_{(k)} - \alpha - \beta x_{(j)}) \\
\frac{\partial}{\partial \sigma^2} \mathcal{L}(\alpha, \beta, \sigma^2) &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{j=1}^{\ell} \sum_{k=1}^{m} w_{jk} (y_{(k)} - \alpha - \beta x_{(j)})^2.
\end{aligned}$$

Now, let $w_{j+} := \sum_{k=1}^m w_{jk}$ and $w_{+k} := \sum_{j=1}^{\ell} w_{jk}$, then

$$\begin{aligned}
\frac{\partial}{\partial \alpha} \mathcal{L}(\alpha, \beta, \sigma^2) = 0 &\Leftrightarrow \sum_{j=1}^{\ell} \sum_{k=1}^{m} w_{jk} (y_{(k)} - \alpha - \beta x_{(j)}) = 0 \\
&\Leftrightarrow \sum_{j=1}^{\ell} \sum_{k=1}^{m} w_{jk} (y_{(k)} - \beta x_{(j)}) - \alpha \sum_{j=1}^{\ell} \sum_{k=1}^{m} w_{jk} = 0 \\
&\Leftrightarrow \widehat{\alpha} = \frac{1}{n} \sum_{j=1}^{\ell} \sum_{k=1}^{m} w_{jk} (y_{(k)} - \beta x_{(j)})
\end{aligned}$$

$$
\begin{aligned}
&\Leftrightarrow \widehat{\alpha} = \frac{1}{n}\sum_{k=1}^{m} w_{+k}y_{(k)} - \beta\frac{1}{n}\sum_{j=1}^{\ell} w_{j+}x_{(j)}\\
&\Leftrightarrow \widehat{\alpha} = \overline{y} - \beta\overline{x}.\\
\frac{\partial}{\partial\beta}\mathcal{L}(\widehat{\alpha},\beta,\sigma^2) = 0 &\Leftrightarrow \sum_{j=1}^{\ell}\sum_{k=1}^{m} w_{jk}x_{(j)}(y_{(k)} - \widehat{\alpha} - \beta x_{(j)}) = 0\\
&\Leftrightarrow n\left(\overline{xy} - \widehat{\alpha}\overline{x} - \beta\overline{x^2}\right) = 0\\
&\Leftrightarrow \overline{xy} - \overline{y}\cdot\overline{x} + \beta\overline{x}^2 - \beta\overline{x^2} = 0\\
&\Leftrightarrow \widehat{\beta} = \frac{\overline{xy} - \overline{y}\cdot\overline{x}}{\overline{x^2} - \overline{x}}\\
\frac{\partial}{\partial\sigma^2}\mathcal{L}(\widehat{\alpha},\widehat{\beta},\sigma^2) = 0 &\Leftrightarrow \widehat{\sigma^2} = \frac{1}{n}\sum_{j=1}^{\ell}\sum_{k=1}^{m} w_{jk}(y_{(k)} - \widehat{\alpha} - \widehat{\beta}x_{(j)})^2
\end{aligned}
$$

These critical points maximize the log-likelihood and they are unique log function is strictly concave. □

Lemma E.2. *Given a data set $\mathcal{D}_n = (x_i, y_i)_{i=1}^n$, let*

$$\{x_1, \dots, x_n\} = \{x_{(1)}, \dots, x_{(\ell)}\} \quad \textit{and} \quad \{y_1, \dots, y_n\} = \{y_{(1)}, \dots, y_{(m)}\},$$

where $x_{(1)} < \dots < x_{(\ell)}$ and $y_{(1)} < \dots < y_{(m)}$ for some $1 \le \ell, m \le n$. Let $w_{jk} = \#\{i : (x_i, y_i) = (x_{(j)}, y_{(k)})\}$ and we define

$$w_{j+} := \sum_{k=1}^{m} w_{jk} \qquad \textit{and} \qquad \overline{y}_j := \frac{\sum_{k=1}^{m} w_{jk}y_{(k)}}{\sum_{k=1}^{m} w_{jk}} = \frac{\sum_{k=1}^{m} w_{jk}y_{(k)}}{w_{j+}}.$$

Then,

$$\sum_{j=1}^{\ell}\sum_{k=1}^{m} w_{jk}(y_{(k)} - \overline{y}_j)(\overline{y}_j - \mu(x_{(j)})) = 0$$

Proof. We have

$$
\begin{aligned}
\sum_{j=1}^{\ell}\sum_{k=1}^{m} w_{jk}(y_{(k)} - \overline{y}_j)(\overline{y}_j - \mu(x_{(j)})) &= \sum_{j=1}^{\ell}\left[\overline{y}_j\cdot\left(\sum_{k=1}^{m} w_{jk}y_{(k)}\right)\right] - \sum_{j=1}^{\ell}\left[\mu(x_{(j)})\cdot\left(\sum_{k=1}^{m} w_{jk}y_{(k)}\right)\right]\\
&\quad - \sum_{j=1}^{\ell}\left[\left(\overline{y}_j\right)^2\cdot\left(\sum_{k=1}^{m} w_{jk}\right)\right] + \sum_{j=1}^{\ell}\left[\overline{y}_j\mu(x_{(j)})\cdot\left(\sum_{k=1}^{m} w_{jk}\right)\right]\\
&= \sum_{j=1}^{\ell}\left[\overline{y}_j\cdot\left(\sum_{k=1}^{m} w_{jk}y_{(k)}\right)\right] - \sum_{j=1}^{\ell}\left[\mu(x_{(j)})\cdot\left(\sum_{k=1}^{m} w_{jk}y_{(k)}\right)\right]\\
&\quad - \sum_{j=1}^{\ell}\left(\overline{y}_j\right)^2 w_{j+} + \sum_{j=1}^{\ell}\overline{y}_j\mu(x_{(j)})w_{j+}\\
&= \sum_{j=1}^{\ell}\frac{\left(\sum_{k=1}^{m} w_{jk}y_{(k)}\right)^2}{w_{j+}} - \sum_{j=1}^{\ell}\left[\mu(x_{(j)})\cdot\left(\sum_{k=1}^{m} w_{jk}y_{(k)}\right)\right]\\
&\quad - \sum_{j=1}^{\ell}\frac{\left(\sum_{k=1}^{m} w_{jk}y_{(k)}\right)^2}{w_{j+}} + \sum_{j=1}^{\ell}\left[\mu(x_{(j)})\cdot\left(\sum_{k=1}^{m} w_{jk}y_{(k)}\right)\right]\\
&= 0.
\end{aligned}
$$

□

Lemma E.3 (Abel's partial summation formula, adjusted from Exercise 13 page 34 in Busam and Freitag (2009)). *Let* $(a_i)_{i=1}^{n}$ *and* $(b_i)_{i=1}^{n}$ *be real-valued sequence and define*

$$A_i = a_1 + a_2 + \cdots + a_{i-1}, \quad \text{for some } i \geq 1,$$

where we set $A_1 := 0$ *and* $b_0 := 0$*. Then, for each* $n \geq 1$*,*

$$\sum_{i=1}^{n} a_i b_i = \sum_{i=1}^{n} A_i (b_{i-1} - b_i) + A_{n+1} b_n.$$

Proof. We have,

$$\begin{aligned}
\sum_{i=1}^{n} a_i b_i &= \sum_{i=1}^{n} (A_{i+1} - A_i) b_i \\
&= \sum_{i=1}^{n} A_{i+1} b_i - \sum_{i=1}^{n} A_i b_i \\
&= \sum_{i=2}^{n+1} A_i b_{i-1} - \sum_{j=1}^{n} A_i b_i \\
&= \sum_{i=1}^{n} A_i b_{i-1} - \sum_{i=1}^{n} A_i b_i - A_1 b_0 + A_{n+1} b_n \\
&= \sum_{i=1}^{n} A_i (b_{i-1} - b_i) + A_{n+1} b_n.
\end{aligned}$$

□

Lemma E.4. *Let*

$$\widetilde{f}(\tilde{\mathbf{x}}) = \sum_{j=1}^{\ell} \sum_{k=m_j}^{M_j} \left(-\underline{w}_{jk} \tilde{x}_{jk} + n \exp\left(\sum_{k'=m_j}^{k} \tilde{x}_{jk'} \right) \right) \qquad \text{and} \qquad \nabla \widetilde{f}(\tilde{\mathbf{x}}) = \left(\frac{\partial \widetilde{f}(\tilde{\mathbf{x}})}{\partial \tilde{x}_{jk}} \right)_{(j,k) \in \mathcal{P}}.$$

Then for any $(j,k) \in \mathcal{P}$*,*

$$\frac{\partial \widetilde{f}(\tilde{x})}{\partial \tilde{x}_{jk}} = -\underline{w}_{jk} + n \sum_{k'=k}^{M_j} \exp\left(\sum_{s=m_j}^{k'} \tilde{x}_{js} \right).$$

Proof. Fix $1 \leq j \leq \ell$ and $m_j \leq k \leq M_j$. We have

$$\begin{aligned}
\sum_{k=m_j}^{M_j} \left(-\underline{w}_{jk} \tilde{x}_{jk} + n \exp\left(\sum_{k'=m_j}^{k} \tilde{x}_{jk'} \right) \right) &= (-\underline{w}_{jm_j} \tilde{x}_{jm_j} + n \exp(\tilde{x}_{jm_j})) \\
&\quad + (-\underline{w}_{j,m_j+1} \tilde{x}_{j,m_j+1} + n \exp(\tilde{x}_{jm_j} + \tilde{x}_{j,m_j+1})) + \cdots \\
&\quad + (-\underline{w}_{jk} \tilde{x}_{jk} + n \exp(\tilde{x}_{jm_j} + \cdots + \tilde{x}_{jk})) + \cdots \\
&\quad + (-\underline{w}_{jM_j} \tilde{x}_{jM_j} + n \exp(\tilde{x}_{jm_j} + \cdots + \tilde{x}_{jk} + \cdots + \tilde{x}_{jM_j})).
\end{aligned}$$

Therefore, differentiating the above expression w.r.t. $\tilde{x}_{jk}$ yields

$$\frac{\partial \widetilde{f}(\tilde{x})}{\partial \tilde{x}_{jk}} = -\underline{w}_{jk} + n \sum_{k'=k}^{M_j} \exp\left(\sum_{s=m_j}^{k'} \tilde{x}_{js} \right).$$

This is because $\tilde{x}_{jk}$ appears in the exponential $M_j - k + 1$ times. □

F

Additional Figures Produced Using Algorithms in Section 6.2

F.1. Figures with parameter $s_{\max} = 20000$

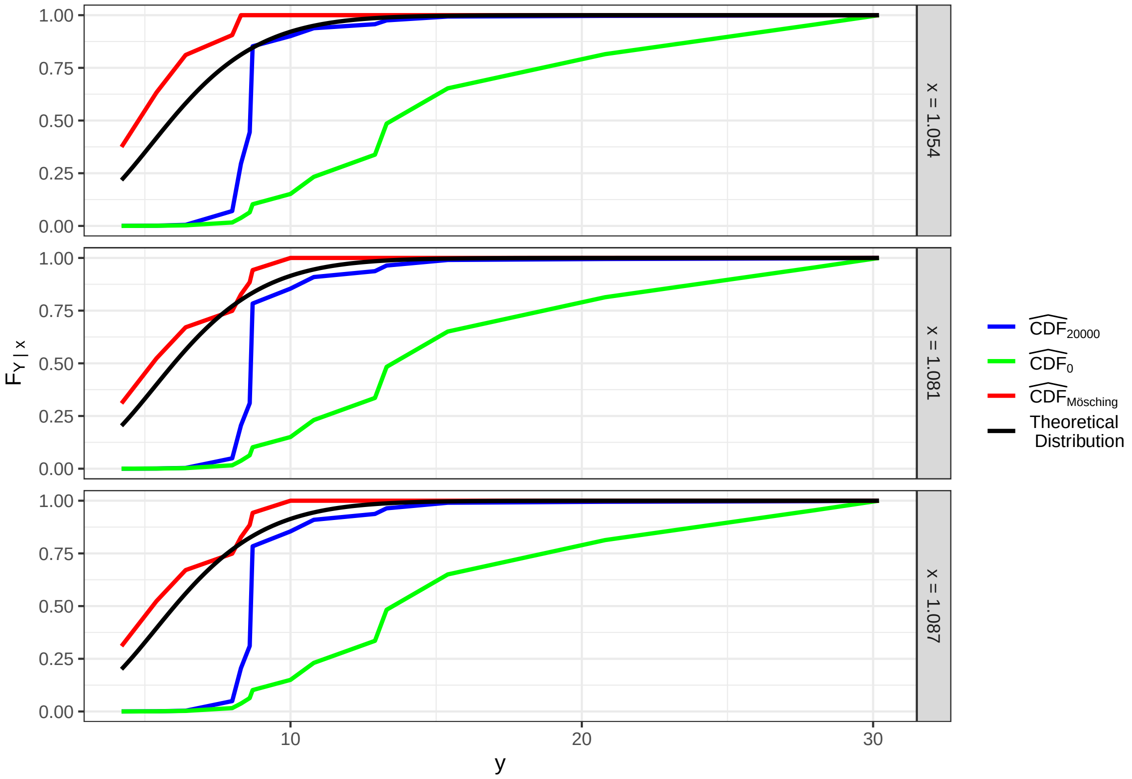

Figure F.1: The plot of $\{\widehat{F}_{Y|x_j}, F^{(0)}_{Y|x_j}, \overline{F}_{Y|x_j}\}$ with the true distributions $G_{Y|x_j}$ for $j = 1, 2, 3$. The function $F^{(0)}_{Y|x_j}$ is the gamma distribution with shape $x_j + 20$ and scale is 1. The number iterations needed to produce $\widehat{F}_{Y|x_j}$ is 20000.

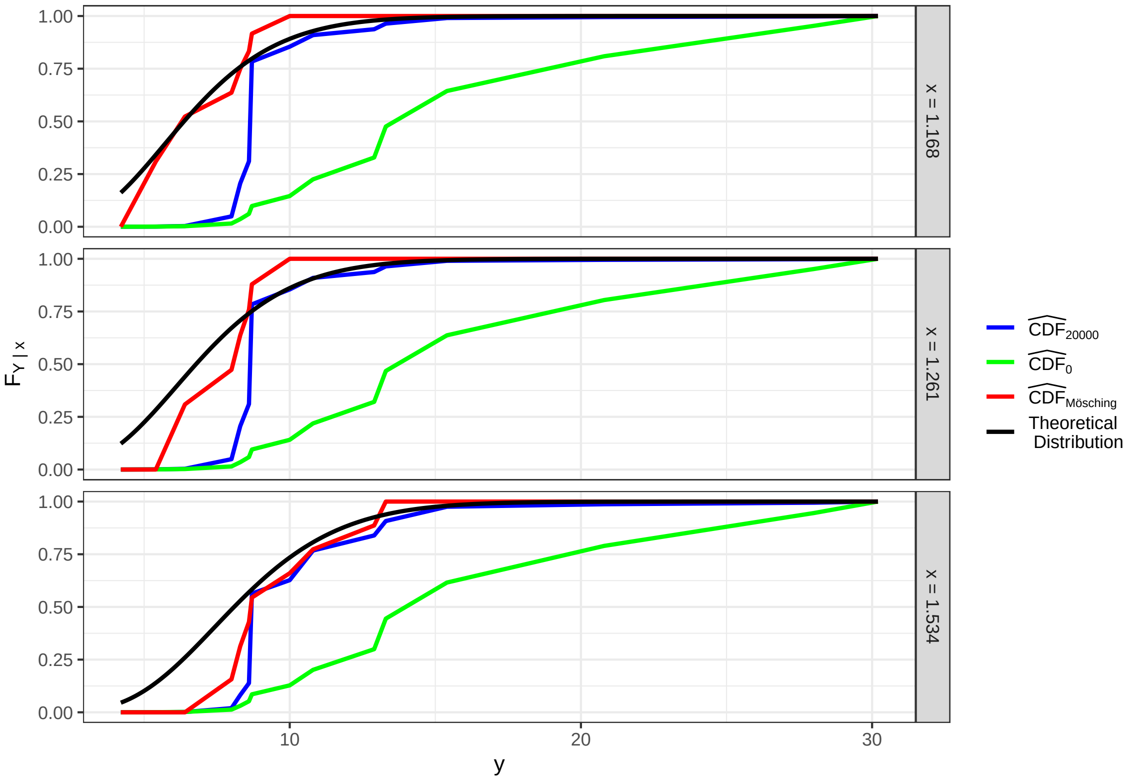

Figure F.2: The plot of $\{\widehat{F}_{Y|x_j}, F^{(0)}_{Y|x_j}, \overline{F}_{Y|x_j}\}$ with the true distributions $G_{Y|x_j}$ for $j = 4, 5, 6$. The function $F^{(0)}_{Y|x_j}$ is the gamma distribution with shape $x_j + 20$ and scale is 1. The number iterations needed to produce $\widehat{F}_{Y|x_j}$ is 20000.

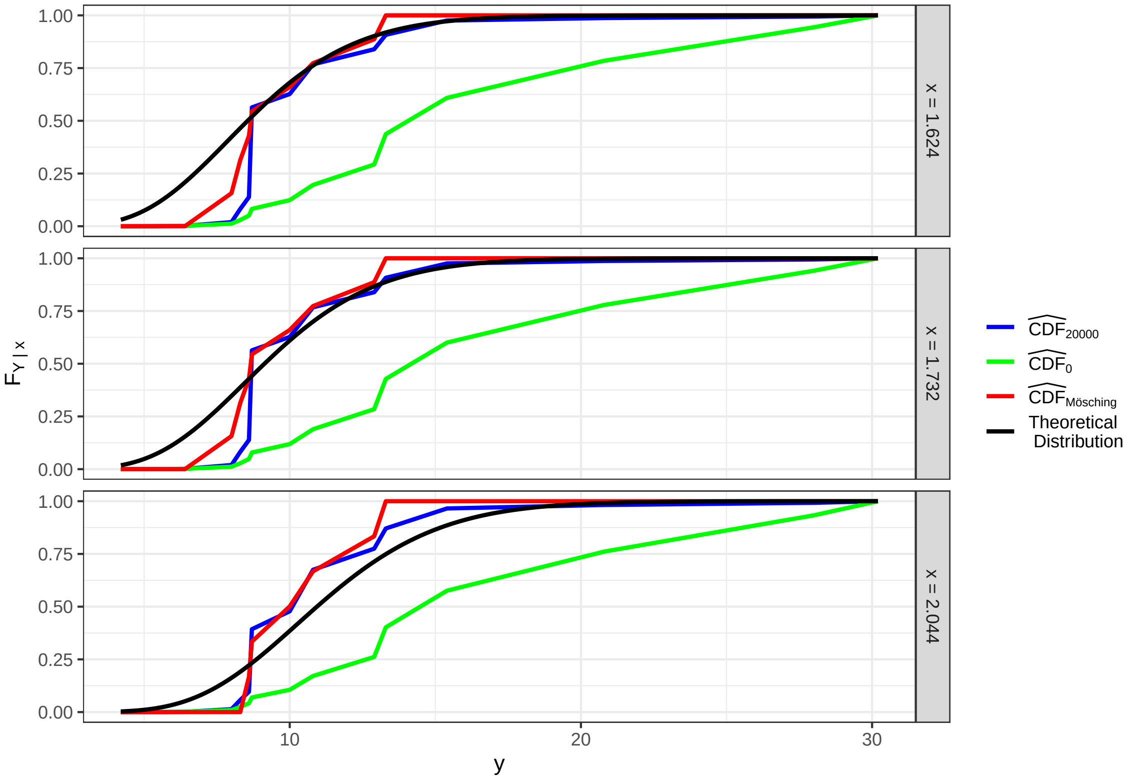

Figure F.3: The plot of $\{\widehat{F}_{Y|x_j}, F^{(0)}_{Y|x_j}, \overline{F}_{Y|x_j}\}$ with the true distributions $G_{Y|x_j}$ for $j = 7, 8, 9$. The function $F^{(0)}_{Y|x_j}$ is the gamma distribution with shape $x_j + 20$ and scale is 1. The number iterations needed to produce $\widehat{F}_{Y|x_j}$ is 20000.

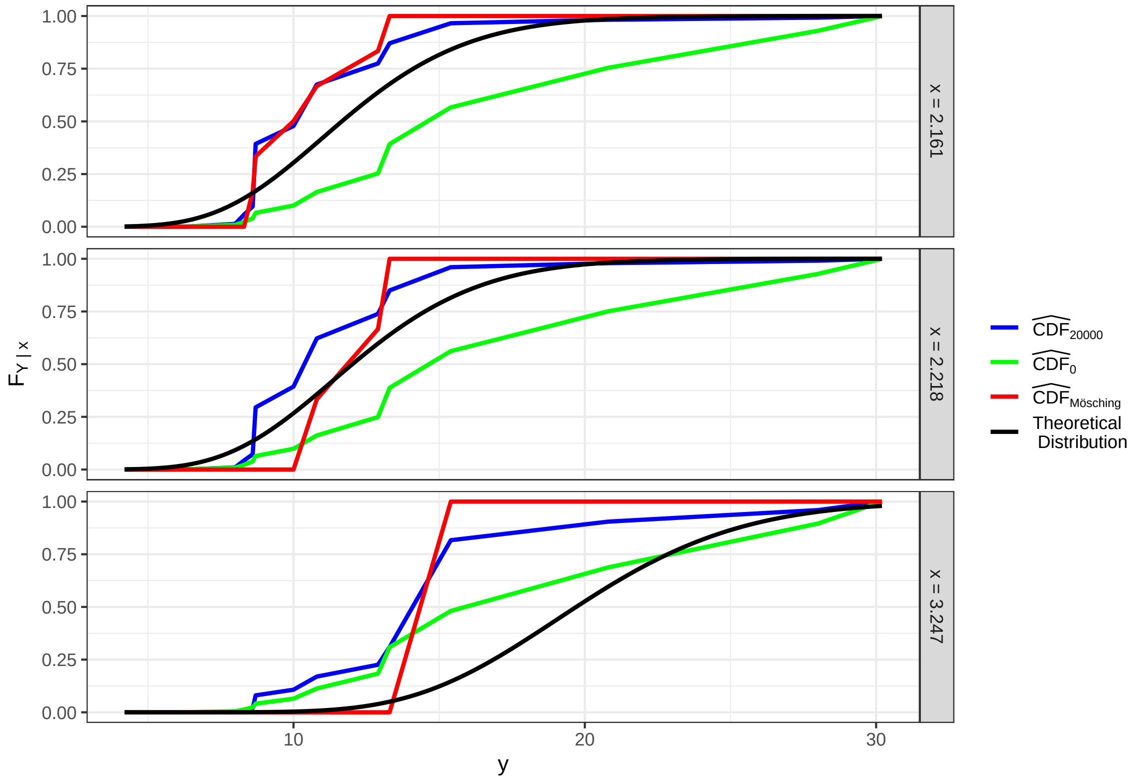

Figure F.4: The plot of $\{\widehat{F}_{Y|x_j}, F^{(0)}_{Y|x_j}, \overline{F}_{Y|x_j}\}$ with the true distributions $G_{Y|x_j}$ for $j = 10, 11, 12$. The function $F^{(0)}_{Y|x_j}$ is the gamma distribution with shape $x_j + 20$ and scale is 1. The number iterations needed to produce $\widehat{F}_{Y|x_j}$ is 20000.

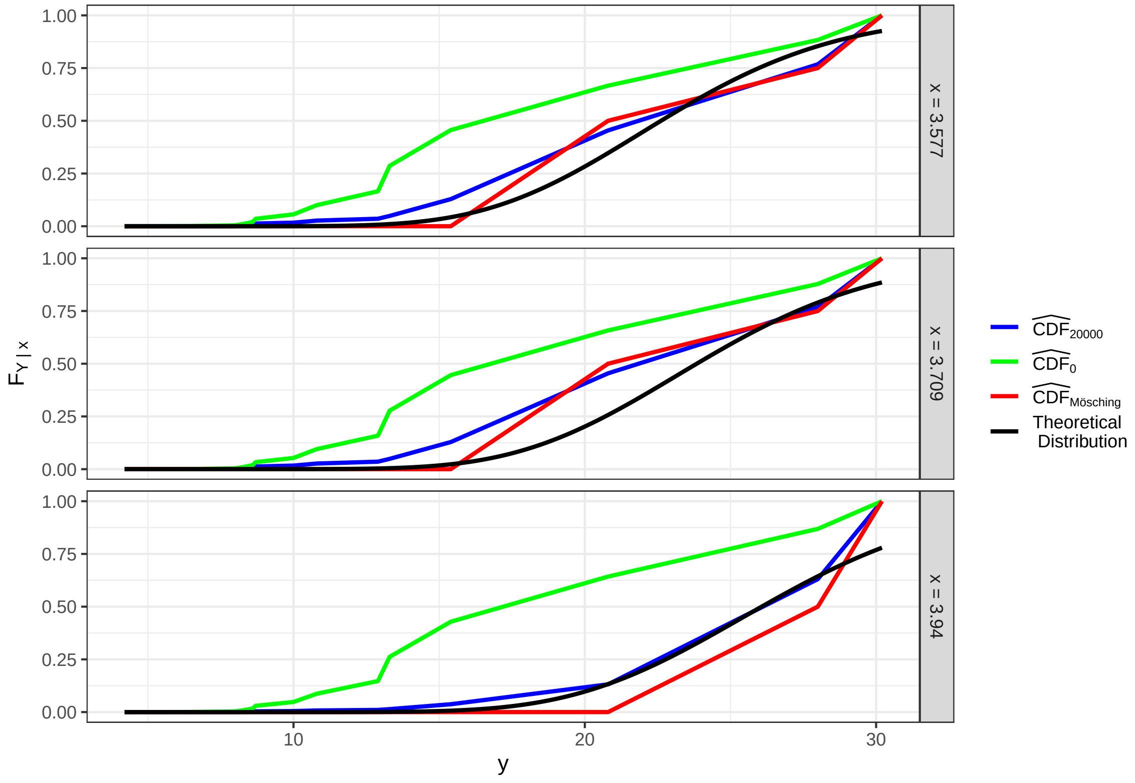

Figure F.5: The plot of $\{\widehat{F}_{Y|x_j}, F^{(0)}_{Y|x_j}, \overline{F}_{Y|x_j}\}$ with the true distributions $G_{Y|x_j}$ for $j = 13, 14, 16$. The function $F^{(0)}_{Y|x_j}$ is the gamma distribution with shape $x_j + 20$ and scale is 1. The number iterations needed to produce $\widehat{F}_{Y|x_j}$ is 20000.

F.2. Figures with parameter $s_{\max} = 50000$

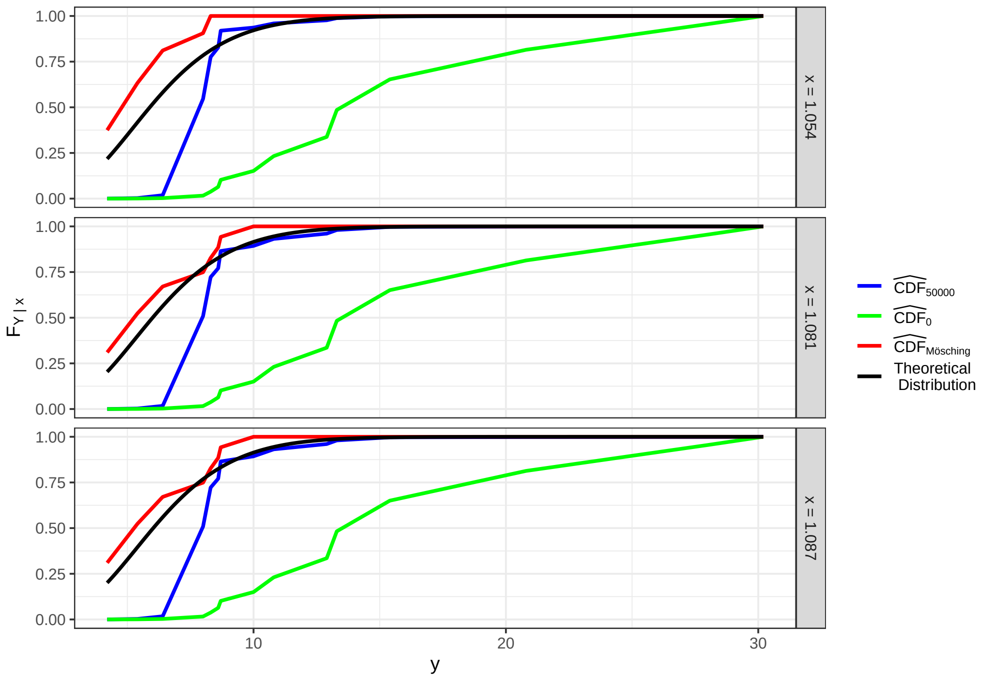

Figure F.6: The plot of $\{\widehat{F}_{Y|x_j}, F^{(0)}_{Y|x_j}, \overline{F}_{Y|x_j}\}$ with the true distributions $G_{Y|x_j}$ for $j = 1, 2, 3$. The function $F^{(0)}_{Y|x_j}$ is the gamma distribution with shape $x_j + 20$ and scale is 1. The number iterations needed to produce $\widehat{F}_{Y|x_j}$ is 50000.

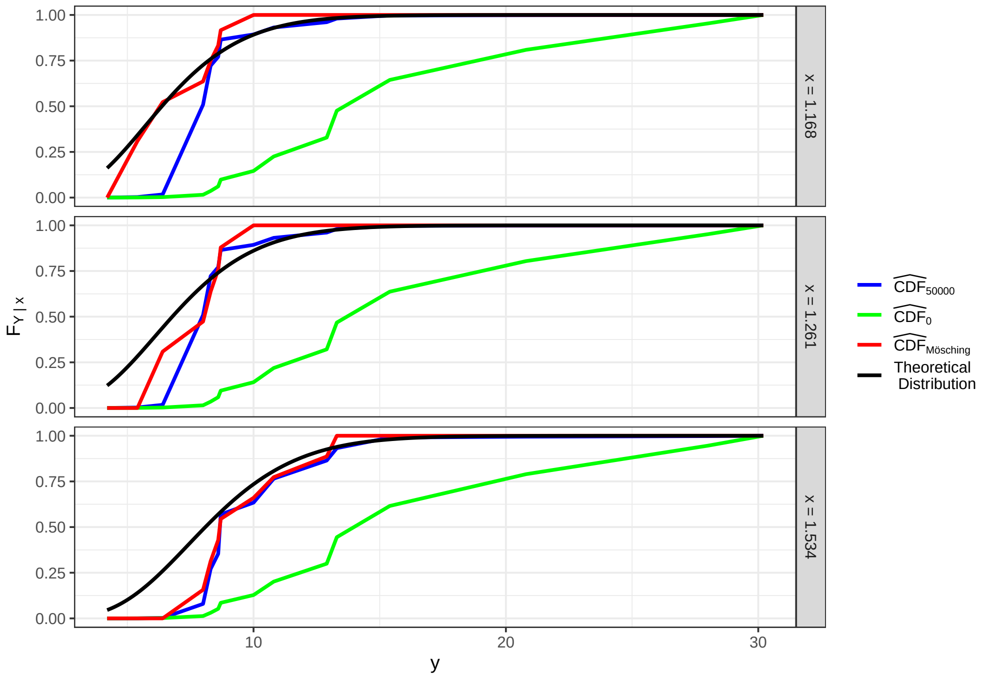

Figure F.7: The plot of $\{\widehat{F}_{Y|x_j}, F^{(0)}_{Y|x_j}, \overline{F}_{Y|x_j}\}$ with the true distributions $G_{Y|x_j}$ for $j = 4, 5, 6$. The function $F^{(0)}_{Y|x_j}$ is the gamma distribution with shape $x_j + 20$ and scale is 1. The number iterations needed to produce $\widehat{F}_{Y|x_j}$ is 50000.

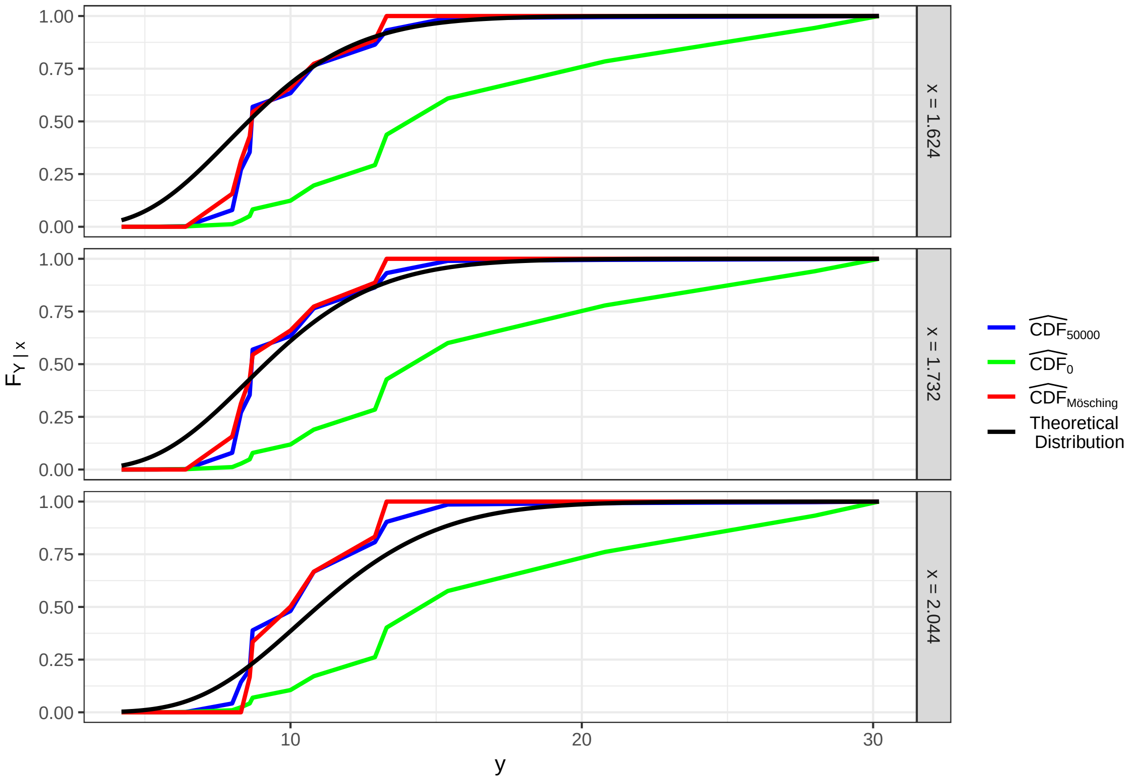

Figure F.8: The plot of $\{\widehat{F}_{Y|x_j}, F^{(0)}_{Y|x_j}, \overline{F}_{Y|x_j}\}$ with the true distributions $G_{Y|x_j}$ for $j = 7, 8, 9$. The function $F^{(0)}_{Y|x_j}$ is the gamma distribution with shape $x_j + 20$ and scale is 1. The number iterations needed to produce $\widehat{F}_{Y|x_j}$ is 50000.

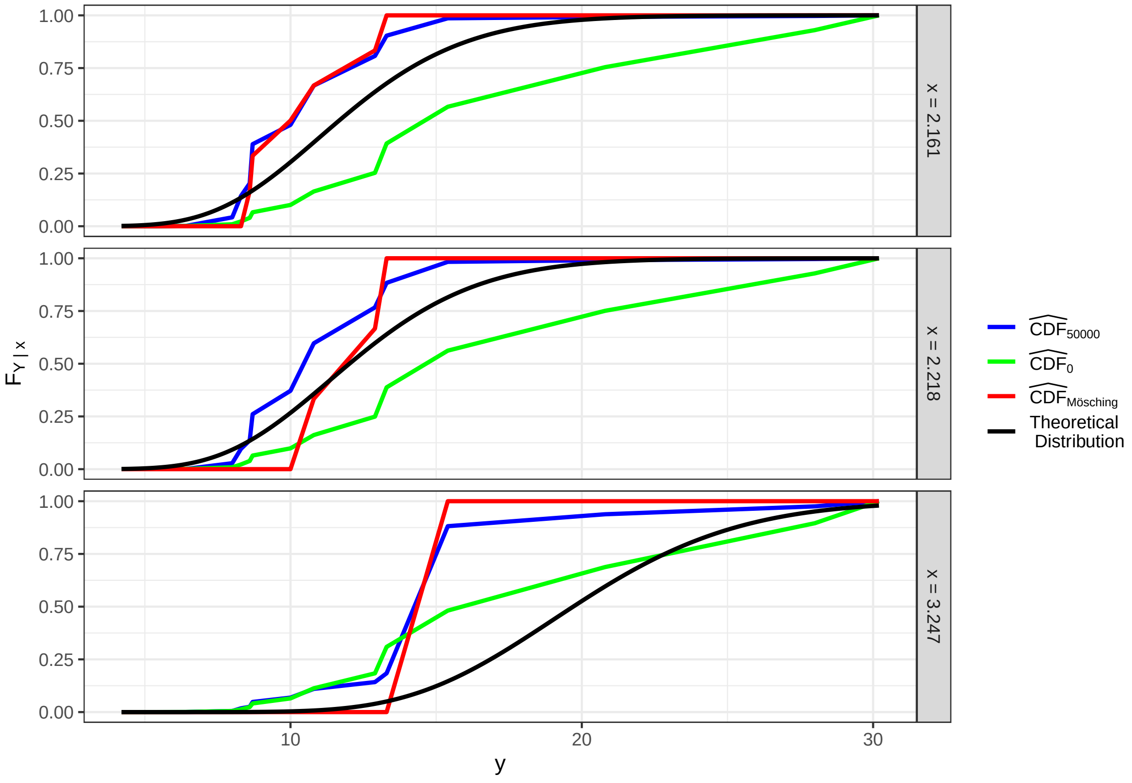

Figure F.9: The plot of $\{\widehat{F}_{Y|x_j}, F^{(0)}_{Y|x_j}, \overline{F}_{Y|x_j}\}$ with the true distributions $G_{Y|x_j}$ for $j = 10, 11, 12$. The function $F^{(0)}_{Y|x_j}$ is the gamma distribution with shape $x_j + 20$ and scale is 1. The number iterations needed to produce $\widehat{F}_{Y|x_j}$ is 50000.

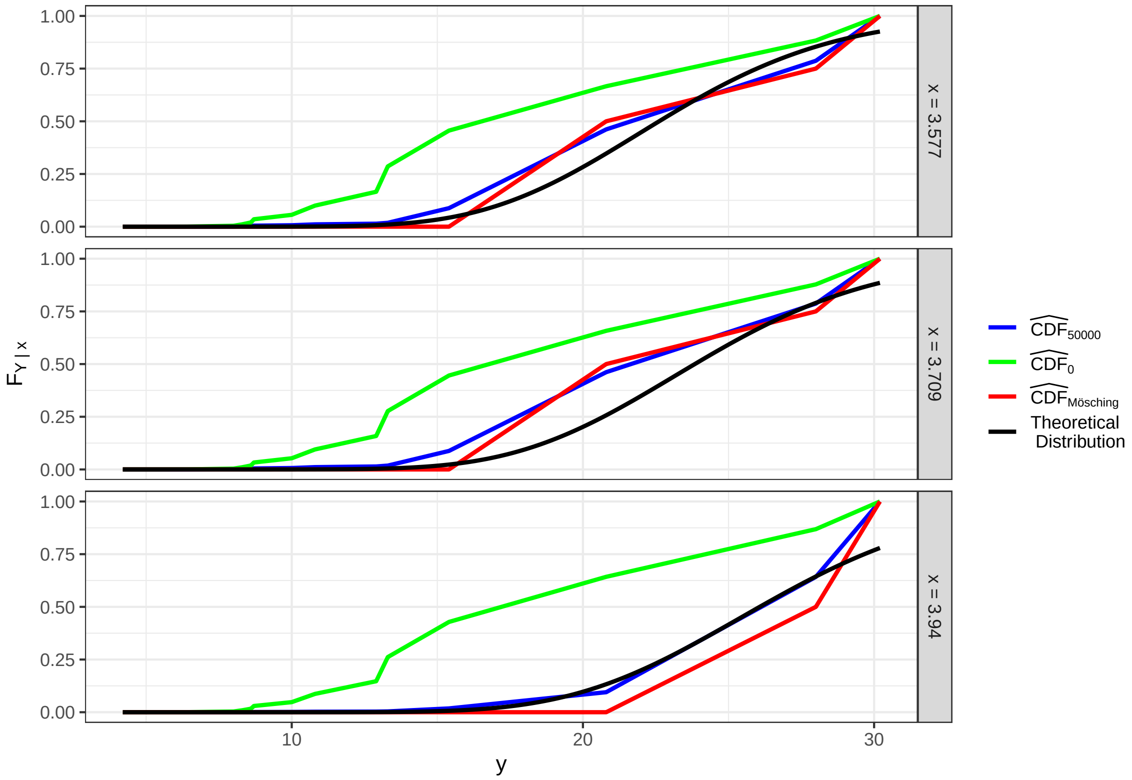

Figure F.10: The plot of $\{\widehat{F}_{Y|x_j}, F^{(0)}_{Y|x_j}, \overline{F}_{Y|x_j}\}$ with the true distributions $G_{Y|x_j}$ for $j = 13, 14, 15$. The function $F^{(0)}_{Y|x_j}$ is the gamma distribution with shape $x_j + 20$ and scale is 1. The number iterations needed to produce $\widehat{F}_{Y|x_j}$ is 50000.