MULTIVARIATE CHANGE OF MEASURE AS CORRECTION METHOD IN ETHICAL PRICING

ÉLOI D'AMOUR BIZIMANA

A Thesis

in

The Department

of

Mathematics and Statistics

Presented in Partial Fulfillment of the Requirements

for the Degree of Master of Arts (Mathematics) at

Concordia University

Montreal, Quebec, Canada

December 2023

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: Éloi D'Amour Bizimana

Entitled: Multivariate Change of Measure as Correction Method in Ethical Pricing

and submitted in partial fulfillment of the requirements for the degree of

Master of Arts (Mathematics)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final Examining Committee:

_____ Thesis Supervisor

Dr. P. Gaillardetz

_____ Thesis Supervisor

Dr. M. Mailhot

_____ Examiner

Dr. Y. Lu

Approved by _____

Dr. L. Popovic (Graduate Program Director)

_____

Dr. P. Sicotte (Dean of Faculty)

_____

Date

Abstract

Multivariate Change of Measure as Correction Method in Ethical Pricing

Éloi D'Amour Bizimana

In recent years, multiple global events have drawn society's attention to fairness-related issues and various societal movements resulted from them. For many fields, the impact was immediate and substantial, but for others it has been much more timid. Insurance is one of the latter. More specifically, the way fairness is implemented in algorithms used to calculate insurance premiums has not changed in decades due in part to the lack of modernization from regulators and in part to the complexity of the issue. Nonetheless, in preparation for society's growing expectations, researchers have developed many ways to implement *algorithmic fairness*. An exposition is made on this concept, including qualitative and quantitative definitions of fairness as well as approaches to its implementation found in the literature. In particular, the method developed by Lindholm et al. (2022) [8] is discussed in detail and followed up by the introduction of our own novel approach. This approach is demonstrated on simulated data, and it is shown that it can significantly reduce unfairness according to pre-determined metrics.

Acknowledgments

First and foremost, I extend my deepest love and gratitude to my family, who have motivated me, more than they know, to keep pushing this project to the end. This thesis may be the result of my work, but I could not have done it without each and every one of you. *Murakoze cyane.*

I thank my friends for their support, counsel and for helping take my mind off of the "serious stuff" every once in a while, keeping me focused and refreshed for more work.

I am deeply appreciative of my supervisors, Dr. Gaillardetz and Dr. Mailhot, for the opportunity they have given me and for the time and effort they have invested in me in the past few years. This journey could not have begun nor reached its conclusion without either of them, and for that I will always be grateful.

Last but not least, I want to thank me for sticking with me all the way and for never giving up on me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Insurers use a plethora of variables to make various decisions relating to their policyholders, such as determining insurance premiums, classifying risks or allocating capital. There is no questioning the use of the majority of these variables. For example, none would argue that proximity of a fire station would help determine the price of a fire insurance policy. However, that is not the case for all variables that are at the insurer's disposition. Variables such as age, gender and postal code are very often provided to insurers and discriminating with respect to these variables is far from being approved unanimously by experts in algorithmic fairness. This is why it is important that there is legislation in place that considers the impact, and more specifically the discriminatory impact, of variables that may be considered "sensitive".

In Canada, legislation tends to be binary, generally limiting itself to allowing or forbidding the use of certain variables. For instance, gender can be used as an auto insurance pricing variable in Alberta, Ontario and Quebec but not in other Canadian provinces. When a government decides to prohibit the use of a sensitive variable in insurance pricing models, it is certainly with the intention that no discrimination is made with respect to that variable, but that approach has long been criticized by researchers. Zliobaite and Custers [14] not only show

that simply disusing a variable does very little to prevent discrimination with respect to that variable, but they even demonstrate a way of using the variable adequately that is much more efficient in that regard.

More formally, consider a *treatment* Y obtained from *explanatory variables* $\mathbf{X}$ and *protected variables* $\mathbf{P}$, with $(\mathbf{X}, \mathbf{P})$ making up all the information available to the insurer. In the current context of insurance pricing, treatment will be analogous to premium charged to policyholders. For protected variables, there is no consensus on their definition, but we will consider them to be variables against which the insurer does not want to discriminate. Note that the use of the word "against" here does not imply that the discrimination is disadvantageous.

We are interested in how we can treat individuals fairly in insurance. Most dictionaries will define fairness as the quality of treating people equally or in a way that is reasonable, but, in insurance, "treating" everyone equally is often inappropriate, as charging everyone the same premium goes against the actuarial principle that a premium should be representative of the associated risk. Therefore, we start by discussing various notions of fairness that are pertinent when implementing data-driven algorithms, *i.e. algorithmic fairness*. Algorithmic fairness is a very broad subject, and we only present some of its concepts, with a focus on the ones pertinent to our work. Tremblay (2022) [11] and Wang et al. (2022) [12] provide a more comprehensive and modern review of the subject.

1.1 Types of Fairness

In algorithmic fairness, two main ideologies stand out: individual fairness and group fairness. Note that, although we will use mathematical statements to describe notions of fairness, they are much more complex than mathematics could ever hope to encompass and should be appropriately combined with ethical considerations when put into practice. Also, these mathematical statements may present these ideologies as conflicting, but conceptually there is good argument

that they not only should, but can be applied concurrently, as is elaborated by Binns (2020) [2] from a more theoretical standpoint.

1.1.1 Individual Fairness

Individual fairness defends that, if two individuals are similar with respect to their explanatory variables $\mathbf{X}$, then they should be treated similarly, *i.e.* have a similar Y. For a more mathematical interpretation, consider individuals A and B with treatment, explanatory and protected variables $(Y_A, \mathbf{X}_A, \mathbf{P}_A)$ and $(Y_B, \mathbf{X}_B, \mathbf{P}_B)$, respectively. Then, individual fairness could be represented by

$$M(\mathbf{X}_A, \mathbf{X}_B) < \delta \implies m(Y_A, Y_B) < \epsilon,$$

for small $\delta, \epsilon > 0$, where M is a distance measure for occurrences of $\mathbf{X}$ and m is a distance measure for occurrences of Y. Determining the treatment similarity function m is typically straightforward (*e.g.*, difference between premiums, whether loans were approved). The challenge usually arises when determining the similarity function for explanatory variables M.

Counterfactual fairness is one approach that would fall under the umbrella of individual fairness. It asserts that an individual should receive the same treatment if they were in a "counterfactual universe" where they had the same explanatory variables, but different protected variables, or, mathematically:

$$\mathbf{X}_A = \mathbf{X}_B \implies Y_A = Y_B \quad \text{even when} \quad \mathbf{P}_A \neq \mathbf{P}_B.$$

While counterfactual fairness is one of the easier approaches to implement in individual fairness, there are multiple arguments against its use. One such argument is that protected variables are typically not available to the insurer, hindering its implementation. Also, counterfactual fairness implicitly assumes that protected variables are reduced to their observation, *e.g.* $P \in \{0, 1\}$. However, protected variables often have impacts that go beyond what datasets can register. The

reader is pointed to Kohler-Hausmann (2019) [7] for an extensive conceptual discussion of the matter.

Access to $\mathbf{P}$ is not always necessary to implement individual fairness. For instance, Dwork et al. (2011) [5] establish a Lipschitz condition that is respected when the distance between outcomes Y is bounded by the distance between the explanatory variables $\mathbf{X}$, without any regard to protected variables $\mathbf{P}$.

1.1.2 Group Fairness

Group fairness defends that, overall, categories of a sensitive variable should be treated similarly, independently of the explanatory variables. Group fairness approaches will be concerned with the combination of the treatment and protected variables $(Y, \mathbf{P})$, with no regard for explanatory variables $\mathbf{X}$. Let $\mathrm{Dom}(\mathbf{P})$ represent the domain of $\mathbf{P}$. Then, it could be said that group fairness is respected with regards to P when the following equation is true:

$$\forall \mathbf{p}_1, \mathbf{p}_2 \in \mathrm{Dom}(\mathbf{P}) \quad \mathrm{E}(g(Y) \mid \mathbf{P} = \mathbf{p}_1) = \mathrm{E}(g(Y) \mid \mathbf{P} = \mathbf{p}_2),$$

where g is some function of the treatment. An equality is often difficult to achieve, and bounding a difference or a ratio is often more manageable:

$$|\mathrm{E}(g(Y) \mid P = 0) - \mathrm{E}(g(Y) \mid P = 1)| < \epsilon \quad \text{or} \quad 1-\epsilon < \left| \frac{\mathrm{E}(g(Y) \mid P = 0)}{\mathrm{E}(g(Y) \mid P = 1)} \right| < \frac{1}{1-\epsilon},$$

for small $\epsilon > 0$.

Detection of Group Unfairness

Multiple authors have developed ways of assessing whether unfairness is present. Zliobaite (2017) [13] presents various statistical tests that have as null hypothesis that there is no discrimination. One such test is the *regression slope test*, which fits an ordinary least squares regression to the data, using Y as the response and $(\mathbf{X}, \mathbf{P})$ as the covariates. If the coefficient β_P is significantly different from zero,

then the null hypothesis is rejected and the model may be unfair when it comes to sensitive variable P, where $p \in \mathbf{P}$.

A more popular approach to detecting group unfairness is assessing whether the model respects a certain condition. Corbett-Davies et al. (2017) [3] suggest statistical parity (also known as demographic parity), which we define below for a single protected variable P.

Definition 1.1.1 (Statistical Parity)**.** *In a classification context, a function f satisfies* statistical parity *if the treatment $Y = f(\cdot)$ is independent from the protected variable $P \in \mathrm{Dom}(P) = \{p_0, p_1, \ldots, p_n\}$ with $n < \infty$. That is,*

$$\forall y \in \mathrm{Dom}(Y) \quad \mathbb{P}(Y = y \mid P = p_0) = \cdots = \mathbb{P}(Y = y \mid P = p_n),$$

where $\mathrm{Dom}(Y)$ is finite.

In a regression context, where Y is continuous, the condition becomes

$$\forall y \in \mathrm{Dom}(Y) \quad \mathbb{P}(Y \leq y \mid P = p_0) = \cdots = \mathbb{P}(Y \leq y \mid P = p_n), \tag{1.1.1}$$

where $\mathrm{Dom}(Y)$ is uncountably infinite.

An implication of statistical parity is statistical parity in expectation.

Definition 1.1.2 (Statistical parity in expectation)**.** *A function f satisfies statistical parity in expectation if the expectation of the treatment $Y = f(\cdot)$ is independent from the protected variable $P \in \mathrm{Dom}(P) = \{p_0, p_1, \ldots, _n\}$ with $n < \infty$. That is,*

$$\mathrm{E}(Y \mid P = p_0) = \mathrm{E}(Y \mid P = p_1) = \cdots = \mathrm{E}(Y \mid P = p_n) = \mathrm{E}(Y). \tag{1.1.2}$$

Statistical parity is the go-to measure of group fairness, and some even treat it as the definition of group fairness itself. However, statistical parity on its own is not a sufficient notion of fairness. Dwork et al. (2011) [5] give the "self-

fulfilling prophecy" as an example. The self-fulfilling prophecy considers a loan analyst tasked with determining whether individuals should receive a loan. The analyst should only base their decision on the individual's explanatory variables $\mathbf{X}$, but is pernicious and also uses the protected variable $P \in \{0, 1\}$. Instead of blatantly denying loans to individuals with $P = 0$, the analyst grants loans to an equivalent proportion of individuals with $P = 0$ and $P = 1$, but, among individuals with $P = 0$, selects those who are more likely to default according to the explanatory variables $\mathbf{X}$. In doing so, the analyst does achieve statistical parity, but effectively builds a case against individuals with $P = 0$. Since these individuals are more likely to default from the start, it will be possible to make the observation that individuals with $P = 0$ are more likely to default than individuals with $P = 1$.

1.2 Correction of Unfairness

When it comes to correcting unfairness, four types of approaches are present in the literature: ignorance, pre-treatment, during treatment, and post-treatment.

Ignorance

Ignorance, as its name implies, consists in ignoring the sensitive variables $\mathbf{P}$, *i.e.* not explicitly using them as arguments of the model function. This approach is, for good reason, criticised by many. Indeed, it is often the case that explanatory variables $\mathbf{X}$ and sensitive variables $\mathbf{P}$ will have some level of dependence, such that there is information about sensitive variables included in the explanatory variables. For example, many home insurance models use postal code (or some transformation of postal code) as an explanatory variable, but postal code and ethnicity can be highly dependent. Multiple cities have neighbourhoods where a majority of residents are of the same ethnicity. In such a scenario, although ethnicity is not directly used, enough of its information is contained in the postal code that there would not be much difference between results drawn from directly

using either one of the variables. Ignorance distinguishes itself from the other three approaches in that the other three will typically use sensitive variables in order to attain a fairer model. As mentioned before, many authors criticize the ignorance approach because a proper use of protected variables achieve much better results when it comes to introducing fairness.

Pre-treatment

Pre-treatment consists in adjusting the data before applying the model function. A pre-treatment approach would be used when it is believed that bias is already present in the data being used. Possible adjustments include but are not limited to data selection (for example, removing observations from the data to obtain as many observations with $S = 0$ than observations with $S = 1$), and data modification.

Intra-treatment

Intra-treatment consists in changing the model function itself. This can be done in many ways and at many levels, from adopting a neural network instead of a linear model to adjusting a loss function. This method is model specific, such that developing an industry-wide standard for during-treatment approaches to fairness is difficult. Also, due to the complexity of model functions, this method is generally the hardest one to implement. Fitzsimons et al. (2019) [6] impose fairness constraints on kernel regression which allows for regression trees that satisfy statistical parity in expectation.

Post-treatment

Post-treatment consists in making adjustments directly to the model outputs such that certain requirements are met. For example, statistical parity could be easier to attain under post-treatment approaches, since it may depend on the output variable Y and a protected variable P which is not an input of the model function. An advantage of post-treatment is that the model function is not

required to implement it. Petersen et al. (2021) [10] consider the treatment Y and a similarity graph between individuals, undertaking the problem of individual fairness in a classification context through a graph smoothing approach.

Our Approach

In the literature, the majority of the work on algorithmic fairness is concerned with classification problems, such as advertisement or job applicant selection, and even within the work that is done in regression, not all of it is applicable to insurance. In the following, we will bring our focus to group fairness in insurance pricing. In particular, our objective will be to implement a post-treatment correction to premiums that have already been calculated such that they are more fair, according to some metrics. We can ignore the self-fulfilling prophecy described in Section 1.1.2 because we are using a post-treatment approach, and the treatment, *i.e.* the calculation of premiums, will have made a proper use of the explanatory variables.

Although the present paper and its developments apply to group fairness, this is not a statement that group fairness should be preferred to individual fairness. Such considerations should be tackled by regulators, and we simply provide a method that would satisfy these needs should they arise in the industry.

We begin by explaining the intra-treatment approach of Lindholm et al. (2022) [8] to a discrimination-free premium in Chapter 2. We discuss both advantages and disadvantages to illustrate some of the work done in the literature for group fairness in a regression context. Then, in Chapter 3, we generalize work from Pesenti et al (2018) [9]. This generalization will serve as the mathematical backbone behind our own correction method, which we elaborate and demonstrate in Chapter 4. Finally, we conclude and consider some open problems in Chapter 5.

1.3 Preliminaries

We briefly introduce the reader to measure theory. In the following chapters, we will be considering random variables through a measure theory lens rather than a probability theory lens, and the definitions in this section will equip the reader with necessary knowledge. This assumes that the reader is already well-versed in probability, and we refer to Chapters 2, 5 and 12 of Axler (2019) [1] for a thorough exposition.

Definition 1.3.1 (σ-algebra)**.** *Let Z be some set and let 2^Z represent its power set,* i.e. *the set of all its subsets. Then, a subset $\Sigma \subseteq 2^Z$ is called a* σ-algebra on *Z if and only if it satisfies the following three conditions:*

1. $Z \in \Sigma$
2. $E \in \Sigma \implies Z \setminus E \in \Sigma$ (Σ *is closed under complement*)
3. $E_1, E_2, E_3, \cdots \in \Sigma \implies E_1 \cup E_2 \cup E_3 \cup \cdots \in \Sigma$ (Σ *is closed under countable union*).

From Conditions 2 and 3, De Morgan's laws imply that Σ is also closed under countable intersection.

Definition 1.3.2 (Measurable space, measurable set)**.** *Let Z be some set and Σ be a σ-algebra on Z. Then, (Z, Σ) is a* measurable space*. If $E \in \Sigma$, then E is called a* Σ-measurable set*, or simply a measurable set when Σ is clear from the context.*

Definition 1.3.3 (Measure, probability measure)**.** *Let (Z, Σ) be a measurable space. Then, a set function $\mu : \Sigma \to [-\infty, \infty]$ is called* a measure on (Z, Σ) *if and only if the following conditions hold:*

1. $\mu(\emptyset) = 0$
2. *Non-negativity:* $\forall E \in \Sigma,\ \mu(E) \geq 0$
3. *Countable additivity: For all countable collections of sets $\{E_i\}_{i=1}^{\infty}$ in Σ such*

that $\forall j, k \geq 1,\ E_j \cap E_k = \emptyset$*:*

$$\mu\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mu\left(E_i\right).$$

If $\mu(Z) = 1$*, then* μ *is called a* probability measure*.*

Definition 1.3.4 (σ-finite measure)**.** *Let* (Z, Σ, μ) *be a measure space. The measure* μ *is said to be a* σ-finite measure *if and only if there exists a sequence of sets* $\{E_i\}_{i=1}^{\infty} \in \Sigma$ *with* $\mu(E_i) < \infty$ *for all* $i \in \mathbb{N}$ *such that* $\bigcup_{i\in\mathbb{N}} E_i = Z$*.*

Definition 1.3.5 (Null set)**.** *Let* (Z, Σ, μ) *be a measure space. If* $\mu(E) = 0$*, then* E *is called a* μ-null set*, or simply a null set when* μ *is clear from the context.*

Definition 1.3.6 (Measure space, probability space)**.** *Let* (Z, Σ) *be a measurable space and* μ *be a measure on* (Z, Σ)*. Then,* (Z, Σ, μ) *is a* measure space*. If* μ *is a probability measure, we say* (Z, Σ) *is a* probability space*.*

Definition 1.3.7 (Product measure)**.** *Let* (Z_1, Σ_1, μ_1) *and* (Z_2, Σ_2, μ_2) *be two measure spaces. Let* $\Sigma_1 \otimes \Sigma_2$ *be the* σ*-algebra on* $Z_1 \times Z_2$*. Then,* $\mu_1 \times \mu_2$ *is* a product measure *on the measurable space* $(Z_1 \times Z_2, \Sigma_1 \otimes \Sigma_2)$ *if and only if*

$$\forall (E_1 \times E_2) \in \Sigma_1 \otimes \Sigma_2 \quad (\mu_1 \times \mu_2)(E_1 \times E_2) = \mu_1(E_1)\mu_2(E_2).$$

Definition 1.3.8 (Absolute Continuity)**.** *Let* (Z, Σ) *be a measurable space on which the measures* μ *and* ν *are defined. We say that* ν *is* absolutely continuous *with respect to* μ *if and only if*

$$\forall E \in \Sigma \quad \mu(E) = 0 \implies \nu(E) = 0.$$

We write $\nu \ll \mu$*.*

We also state the Radon-Nikodym theorem here, as it will be useful in Chapter 3.

Theorem 1.3.9 (Radon-Nikodym)**.** *Let μ and ν be two measures on the measurable space (Z, Σ). If $\nu \ll \mu$, then there exists a Σ-measurable function $f : Z \to [0, \infty)$ such that for any $E \in \Sigma$,*

$$\nu(E) = \int_E f \, \mathrm{d}\mu.$$

Chapter 2

The Discrimination-free Premium

In this chapter, we outline the approach to a *discrimination-free premium* proposed by Lindholm et al. (2022) [8]. Using notation previously introduced, we have:

- Y represents treatment, in this case insurance premium
- X, observed, represents an explanatory variable used in the pricing model
- P represents a *protected* variable, a variable the insurer does not want to discriminate against, such as gender, religion, ethnicity, etc. Because insurers seldom request that sort of information from potential policyholders, P is typically not observed, but we make the assumption that the information is available.

There can be, and most often will be, multiple explanatory variables $\mathbf{X}$ and protected variables $\mathbf{P}$, but we limit ourselves to one of each for simplicity. However, it is important to note that the results presented in this and following chapter extend naturally to multivariate $\mathbf{X}$ and/or $\mathbf{P}$.

2.1 Working Out the Discrimination-free Premium

Define the *best-estimate premium* as follows.

Definition 2.1.1 (Best-estimate premium)**.** *The* best-estimate premium Y^{BE} *with respect to* (X, P) *is*

$$\tau(X, P) := \mathrm{E}(L \mid X, P). \tag{2.1.1}$$

The best-estimate premium is what the insurer would charge using all the information at their disposal. It is clearly discriminatory, as a change in P would directly result in a change in $\tau(X, P)$.

As mentioned prior, a regulator may forbid the use of P in the model function to avoid discrimination. In that case, insurers would charge an *unawareness premium*, defined as follows.

Definition 2.1.2 (Unawareness premium)**.** *The* unawareness premium Y^{U} *with respect to* X *is*

$$\tau(X) := \mathrm{E}(L \mid X). \tag{2.1.2}$$

Using the unawareness premium rather than the best-estimate premium is a method that falls under the ignorance approach to fairness. The insurer blinds itself toward P and makes its best-estimate of Y without it. While the unawareness premium does not directly discriminate with respect to P, expressing it in the following way shows that it does not avoid discrimination altogether:

$$\tau(X) = \int_{p} \tau(X, p)\, \mathrm{d}\mathbb{P}(p \mid X). \tag{2.1.3}$$

So, the unawareness premium indirectly depends on the distribution of P conditioned on X. This means that if any information on P can be drawn from X, it will be reflected in the calculated premium, as illustrated in the below.

Consider an insurer calculating premiums using past experience on a portfolio of 20 individuals as a basis. For each individual, we observe the previous year's loss L, the region of residence $X \in \{A, B, C\}$ and the immigration status $P \in \{0, 1\}$ ($P = 1$ if they have immigrated, $P = 0$ otherwise). The portfolio is presented in Table 2.1 and illustrated in Figure 2.1.

ID	Loss	Region	Status	ID	Loss	Region	Status
1	135.93	A	0	11	341.66	B	1
2	212.69	A	0	12	187.21	B	1
3	26.23	A	0	13	131.16	C	0
4	25.16	A	0	14	468.84	C	0
5	39.27	A	1	15	399.31	C	1
6	260.73	A	1	16	392.00	C	1
7	277.99	B	0	17	710.37	C	1
8	122.01	B	0	18	247.92	C	1
9	235.00	B	1	19	127.58	C	1
10	36.13	B	1	20	222.83	C	1

Table 2.1: Mock portfolio

To assess whether there is any degree of empirical dependence between X and P, we wish to calculate some conditional probabilities. To do so, we first calculate empirical joint probabilities for the region X and the immigration status P, which are obtained simply by counting the number of individuals respecting a given region-status combination and dividing by the total number of observations. For example, to calculate $\mathbb{P}(X = A, P = 0)$, count the number of individuals residing in region A that have not immigrated. Four individuals respect that condition (IDs #1-4). Dividing by 20 total observations yields $\mathbb{P}(X = A, P = 0) = 4/20 = 0.2$. Results for all combinations are presented in table 2.2.

$\mathbb{P}(\cdot, \cdot)$	$P = 0$	$P = 1$	$P \in \{0, 1\}$
$X = A$	0.2	0.1	0.3
$X = B$	0.1	0.2	0.3
$X = C$	0.1	0.3	0.4
$X \in \{A, B, C\}$	0.4	0.6	1

Table 2.2: Empirical joint distribution $\mathbb{P}$ of the region X and the immigration status P for the mock portfolio.

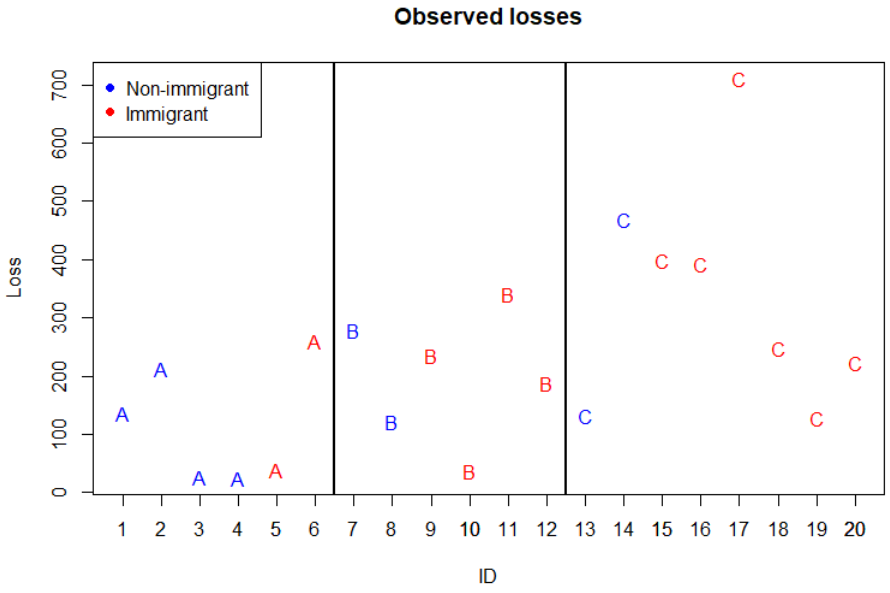

Figure 2.1: Losses for each individual by ID. Letters indicate the region of residence of the individual and colors indicate immigration status. Vertical lines segment the individuals by region of residence.

The following conditional probabilities are obtained:

$$\mathbb{P}(P = 1 \mid X = A) = \frac{\mathbb{P}(P = 1, X = A)}{\mathbb{P}(X = A)} = \frac{2/20}{6/20} = \frac{1}{3},$$

$$\mathbb{P}(P = 1 \mid X = B) = \frac{\mathbb{P}(P = 1, X = B)}{\mathbb{P}(X = B)} = \frac{4/20}{6/20} = \frac{2}{3},$$

$$\mathbb{P}(P = 1 \mid X = C) = \frac{\mathbb{P}(P = 1, X = C)}{\mathbb{P}(X = C)} = \frac{6/20}{8/20} = \frac{3}{4}.$$

These conditional probabilities show that knowledge on X provides knowledge on P. For example, knowing that an individual resides in region C would make the insurer quite confident that that same individual is an immigrant because the portfolio suggests that 3 in 4 individuals residing in region C are immigrants[1].

[1]While this portfolio is fabricated, it represents a realistic scenario, as touched on in Section 1.2.

If X and P were independent, all conditional probabilities would be equal. The fact that is not the case indicates that there is empirical dependence between X and P. That in and of itself does not necessarily imply unfairness, but this dependence may add unduly dependence Y and P, due to Y being a function of X.

Say the insurer wishes to calculate a newcomer's premium. The insurer assumes that losses in the coming year will be identical in distribution to losses in the past year, and decides to calculate premiums using the *equivalence principle*, that is:

$$\begin{aligned} \text{E(Cash in-flow)} &= \text{E(Cash out-flow)}, \\ \text{E}(Y) &= \text{E}(L). \end{aligned}$$

The equivalence principle is important because it ensures that, on average, the insurer will have enough capital to cover the insured's claims.

The newcomer resides in region C and is an immigrant ($P = 1$). The insurer's first attempt is the best-estimate premium (2.1.1):

$$\begin{aligned} Y_{NEW}^{BE} &= \text{E}(L \mid X = x_{NEW}, P = p_{NEW}) \\ &= \text{E}(L \mid X = C, P = 1) \\ &= \sum_{i=1}^{20} l_i \cdot \frac{\mathbb{P}(L = l_i, X_i = C, P_i = 1)}{\mathbb{P}(X = C, P = 1)} \\ &= 399.31 \cdot \frac{1/20}{6/20} + 392.00 \cdot \frac{1/20}{6/20} + \cdots + 222.83 \cdot \frac{1/20}{6/20} \\ &= 350. \end{aligned}$$

Observe that the result is exactly an average of the losses of individuals 15 through 20. This is not a coincidence. Indeed, the best-estimate premium will simply average across all observations in the portfolio with the same combination (X, P) as the newcomer, set to $(C, 1)$. Because of this, it is clear that the best-estimate

premium discriminates against P. If the newcomer had not been an immigrant, they would have obtained a premium that is the average of all observations with $(X, P) = (C, 0)$, which a simple calculation yields to be 300. Table 2.3 also presents average premiums conditionally on each region, each immigration status and each combination of region and immigration status.

$\mathrm{E}(L \mid \cdot, \cdot)$	$P = 0$	$P = 1$	$P \in \{0, 1\}$
$X = A$	100.00	150.00	116.67
$X = B$	200.00	200.00	200.00
$X = C$	300.00	350.00	337.50
$X \in \{A, B, C\}$	175.00	266.67	230.00

Table 2.3: Expected values of the losses conditionally on each region, each immigration status and each combination of region and immigration status.

To avoid the kind of discrimination exhibited by the best-estimate premium, the insurer opts for the unawareness premium (2.1.2):

$$\begin{aligned} Y_{NEW}^{U} &= \mathrm{E}(L \mid X = x_{NEW}) \\ &= \mathrm{E}(L \mid X = C) \\ &= \sum_{i=1}^{20} l_i \cdot \frac{\mathbb{P}(L = l_i, X_i = C)}{\mathbb{P}(X = C)} \\ &= 131.16 \cdot \frac{1/20}{8/20} + 468.84 \cdot \frac{1/20}{8/20} + \cdots + 222.83 \cdot \frac{1/20}{8/20} \\ &= 337.50. \end{aligned}$$

Not too differently from the best-estimate premium, the unawareness premium is obtained by averaging across all observations with $X = C$. The insurer disregards P and now only considers the region X. Because no direct use of P was made in this calculation, it may appear as though Y_{NEW}^{U} is not discriminatory and, in fact, even if the newcomer had not been an immigrant (*i.e.* if they had $P = 0$ instead of $P = 1$), they would have been charged the same premium. However,

the unawareness premium can also be calculated using (2.1.3):

$$\begin{aligned} Y^U_{NEW} &= \sum_{p=0}^{1} \mathrm{E}(L \mid X = x_{NEW}, P = p) \cdot \mathbb{P}(P = p \mid X = x_{NEW}) \\ &= \mathrm{E}(L \mid X = C, P = 0) \cdot \mathbb{P}(P = 0 \mid X = C) \\ &\quad + \mathrm{E}(L \mid X = C, P = 1) \cdot \mathbb{P}(P = 1 \mid X = C) \\ &= 300 \cdot \frac{1}{4} + 350 \cdot \frac{3}{4} \\ &= 337.50. \end{aligned}$$

Using (2.1.3) to obtain the unawareness premium shows that it does indeed discriminate against P, but it is nebulous *how* exactly that is the case. The unawareness premium depends not on the immigration status of the newcomer, but rather implicitly depends on the *distribution* of immigration status conditional on region of residence. In this case, because the newcomer resides in region C, the insurer will use individuals also residing in region C (IDs #13-20) to calculate the premium. That is demonstrated by the explicit use of $\mathbb{P}(P = 0 \mid X = C)$ and $\mathbb{P}(P = 1 \mid X = C)$ in calculations. (2.1.3) thus exposes the unawareness premium as a weighted average of the best estimate premiums for

1. a non-immigrant newcomer residing in region C: $\tau(X = C, P = 0) = 300$
2. an immigrant newcomer residing in region C: $\tau(X = C, P = 1) = 350$.

The weights are the proportions of non-immigrants and immigrants residing in region C ($\frac{2}{8} = \frac{1}{4}$ and $\frac{6}{8} = \frac{3}{4}$, respectively). The unawareness premium will be the same for all newcomers residing in region C. We point out two key observations:

1. The unawareness premium, as a discrimination-free alternative to the best-estimate premium, does result in an improvement as it is free of direct discrimination, but it is not completely discrimination-free. Indeed, it performs *indirect discrimination*, which is explained in more detail in later;

2. Note how the newcomer's unawareness premium attributes more weight to immigrant individuals than to non-immigrant individuals because there are more of the former in region C. It could be considered unfair for a non-immigrant newcomer to be charged a premium that is more representative of an immigrant's risk. This grouping effect is not uncommon in actuarial pricing, but it is desirable only for explanatory variables.

Now that some understanding of direct discrimination is built, it is properly defined here.

Definition 2.1.3 (Direct discrimination)**.** *A premium $\tau'(X)$ avoids direct discrimination if it can be written as*

$$\tau'(X) = \mathrm{E}'(Y \mid X),$$

where the expectation $\mathrm{E}'(\cdot)$ is taken with respect to a probability measure $\mathbb{P}'$ such that it (the expectation) exists.

Conceptually, this definition means that a premium is free of direct discrimination if it can be obtained using as information *only* the explanatory variable X. In particular, the unawareness premium obtained for the newcomer does not perform direct discrimination as it was obtained using only the region of residence $X = C$.

The question now becomes: can a premium that is free of both direct and indirect discrimination be constructed? As answer to this question, Lindholm et al. (2022) [8] present the discrimination-free premium, defined as follows.

Definition 2.1.4 (Discrimination-free premium)**.** *The* discrimination-free premium *Y^{DF} with respect to (X, P) is:*

$$h'(X) := \int_p \tau(X, p)\, \mathrm{d}\mathbb{P}'(p), \tag{2.1.4}$$

where the distribution $\mathbb{P}'(p)$ is defined on the same range as the marginal distri-

bution of the discriminatory variable $\mathbb{P} \sim \mathbb{P}(p)$.

The choice of $\mathbb{P}'$ itself is not particularly crucial, the key step rather being the *marginalization* with respect to $\mathbf{X}$. More precisely, the requirement is that X and P are independent under the chosen distribution $\mathbb{P}'$. If it were not for that distinction, (2.1.3) shows that the unawareness premium would be a special case of (2.1.4) with $\mathbb{P}'(P = p) = \mathbb{P}(P = p \mid X)$. A natural choice is $\mathbb{P}'(P = p) = \mathbb{P}(P = p)$, the marginal empirical distribution of P and we define the resulting premium below.

Definition 2.1.5 (Margin-based premium)**.** *The* margin-based premium Y^{MB} *with respect to* X *is:*

$$h(X) := \int_p \tau(X, p)\, \mathrm{d}\mathbb{P}(p). \tag{2.1.5}$$

It is a special case of the discrimination-free premium, with $\mathbb{P}' = \mathbb{P}$ *in (2.1.4).*

Consider now the newcomer's margin-based premium:

$$\begin{aligned} Y^{MB}_{NEW} &= \sum_{p=0}^{1} \mathrm{E}(L \mid X = x_{NEW}, P = p) \cdot \mathbb{P}(P = p) \\ &= \mathrm{E}(L \mid X = C, P = 0) \cdot \mathbb{P}(P = 0) \\ &\quad + \mathrm{E}(L \mid X = C, P = 1) \cdot \mathbb{P}(P = 1) \\ &= 300 \cdot \frac{8}{20} + 350 \cdot \frac{12}{20} \\ &= 330. \end{aligned}$$

Like the unawareness premium, the margin-based premium averages the best-estimate premiums of all occurrences of immigration status in the newcomer's region of residence C. The difference comes down to the weights attributed to each best-estimate premium. Recall $\mathbb{P}$ is but one of (possibly infinitely) many choices for $\mathbb{P}'$. Essentially, any choice of $\mathbb{P}'$ will correspond to a possible weighing of the best-estimate premiums.

Remark 2.1.6. *The unawareness premium of (2.1.3) and the discrimination-*

free premium of (2.1.4) are quite similar. The key difference is in the employed distribution of P. In (2.1.3), the empirical distribution of P conditional on X is used, and is the reason behind the indirect discrimination. It implicitly introduces dependence between the explanatory variable and the protected variable. Because the premium inevitably depends on the explanatory variable, there is a second-order transitive effect which results in the premium being dependent on the protected variable as well.

The choice of $\mathbb{P}'$ being up to the insurer gives them a lot of power. For example, a greedy insurer could try to maximize the premium[2]. In wanting to attribute all the weight to the higher best-estimate premium, they would select $\mathbb{P}'(P = 1) = 1$ and obtain, for the newcomer:

$$\begin{aligned} Y_{NEW}^{DF} &= \sum_{p=0}^{1} \mathrm{E}(L \mid X = x_{NEW}, P = p) \cdot \mathbb{P}(P = p) \\ &= \mathrm{E}(L \mid X = C, P = 0) \cdot \mathbb{P}(P = 0) \\ &\quad + \mathrm{E}(L \mid X = C, P = 1) \cdot \mathbb{P}(P = 1) \\ &= 300 \cdot 0 + 350 \cdot 1 \\ &= 350. \end{aligned}$$

It would also be possible to minimize the premium, by selecting $\mathbb{P}'(P = 1) = 0$:

$$\begin{aligned} Y_{NEW}^{DF} &= \sum_{p=0}^{1} \mathrm{E}(L \mid X = x_{NEW}, P = p) \cdot \mathbb{P}(P = p) \\ &= \mathrm{E}(L \mid X = C, P = 0) \cdot \mathbb{P}(P = 0) \\ &\quad + \mathrm{E}(L \mid X = C, P = 1) \cdot \mathbb{P}(P = 1) \\ &= 300 \cdot 1 + 350 \cdot 0 \\ &= 300. \end{aligned}$$

[2]Although we illustrate this idea on the premium of a single newcomer, it would be more complex, but still feasible, to maximize the total premiums for a portfolio.

These extrema can be represented, formulaically, as follows:

$$h^+(X) := \sup_{\mathbb{P}'} \int_p f(X,p)\,\mathrm{d}\mathbb{P}'(p),$$
$$h^-(X) := \inf_{\mathbb{P}'} \int_p f(X,p)\,\mathrm{d}\mathbb{P}'(p).$$

From these extrema, the following inequalities can be obtained:

$$h^-(X) \leq h'(X),\, h(X),\, \tau(X) \leq h^+(X), \tag{2.1.6}$$
$$\mathrm{E}\left[h^-(X)\right] \leq \mathrm{E}\left[h'(X)\right],\, \mathrm{E}\left[h(X)\right],\, \mathrm{E}\left[Y\right] \leq \mathrm{E}\left[h^+(X)\right]. \tag{2.1.7}$$

The marginalization of $\mathbb{P}'$ with respect to X can ensure that no indirect discrimination takes place, but many choices will respect that criterion. Furthermore, the choice of $\mathbb{P}'$ can have a significant impact on the calculated premiums, as demonstrated by the newcomer's maximal and minimal premiums. Therefore, it is important that $\mathbb{P}'$ is chosen with care and can be justified. To that effect, instead of considering only a single newcomer's premium, the insurer now turns their gaze to the premiums of their initial portfolio. Figure 2.2 presents the best-estimate, unawareness and margin-based premiums for the 20 individuals of the original portfolio. In the figure, it is explicit that best-estimate premium suffer from direct discrimination, while unawareness and margin-based premiums do not. Indeed, in regions A and C, best-estimate premiums are higher for immigrants than for non-immigrants (because immigrants tended to have higher losses in the past year) and both unawareness and margin-based premiums do not vary. It is not obvious from the figure how unawareness and margin-based premiums indirectly discriminate, but comparing the two sets of premiums within regions may help build some intuition.

Consider first region A. When $X = A$, the two best-estimate premiums are $\tau(A,0) = 100$ and $\tau(A,1) = 150$. In the unawareness premium calculation, which is a weighted average of the two, they get weights of $\mathbb{P}(P = 0 \mid X = A) = 2/3$

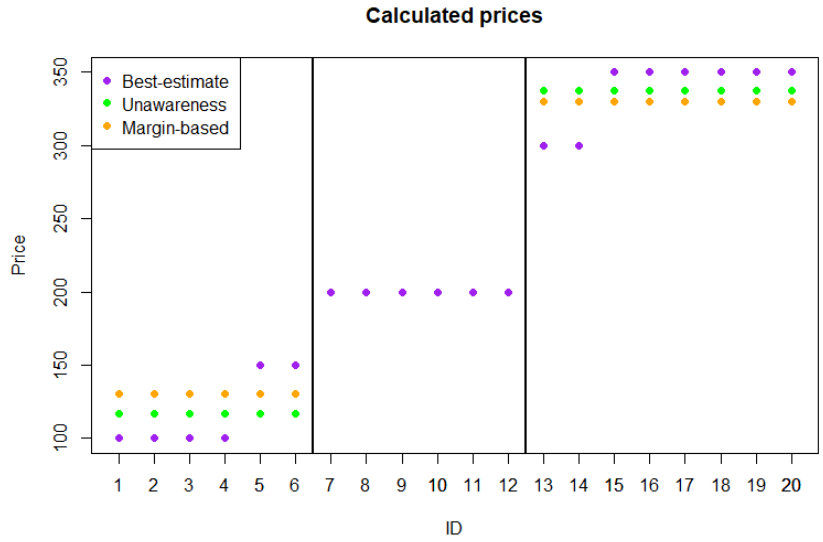

Figure 2.2: Best-estimate, unawareness and margin-based premiums for all individuals of the mock portfolio. Vertical lines segment the individuals by region of residence, with the left-most section being region A, the middle section being region B and the right-most section being region C.

and $\mathbb{P}(P = 1 \mid X = A) = 1/3$, respectively. The unawareness premium attributes more weight to non-immigrants than to immigrants because more non-immigrants reside in region A. The margin-based premium attributes to $\tau(A, 0)$ and $\tau(A, 1)$ weights of $\mathbb{P}(P = 0) = 12/20 = 3/5$ and $\mathbb{P}(P = 1) = 8/20 = 2/5$, respectively. Since the margin-based approach attributes more weight to the larger $\tau(A, 1)$ than the unawareness approach, it results in the higher premium. The unawareness premium is so low because it uses information on the immigration status gained from its dependence with the *considered* region of residence. The margin-based premium deviates from the unawareness premium by using information on P gained from its dependence with all *regions*. Therefore, the margin-based premium still depends on the immigration status, but in a way that is consistent across the whole portfolio.

Now consider region C. When $X = C$, the two best-estimate premiums are $\tau(C, 0) = 300$ and $\tau(C, 1) = 350$. Here, the weight attributed to the greater best-estimate premium $\tau(C, 1)$ is $\mathbb{P}(P = 1) = 12/20 = 3/5$, which, contrary to

region A, is *lower* than its weight of $\mathbb{P}(P = 1 \mid X = C) = 6/8 = 3/4$ for the unawareness premium. Hence, for region C, the margin-based premium is lower than the unawareness premium and the interpretation is exactly the same as for region A. Information on immigration status is used but in the same way as in region A.

In region B, because $\tau(B, 0) = \tau(B, 1) = 200$ (the two best-estimate premiums are equal), the resulting premium will be the same no matter the weighing used. This means the unawareness premium is equal to the margin-based premium, as supported by Figure 2.2. In fact, the premium obtained using any other distribution $\mathbb{P}'$ in (2.1.4) would always be 200.

Comparing the unawareness and margin-based premiums in regions A and C shows that the latter premium does succeed in reducing discrimination when compared to the unawareness premium. Although it depends on a distribution of the immigration status $\mathbb{P}'$, that distribution is not affected by the region of residence. Therefore, margin-based premiums $h(X)$, while still drawing explanatory power from the region of residence X through the best-estimate premiums $\tau(X, \cdot)$, are free of second-order, or indirect, discrimination toward the immigration status due to the region of residence. This leads to a proper definition of indirect discrimination:

Definition 2.1.7 (Indirect discrimination)**.** *A premium $h'(X)$ that avoids direct discrimination is said to avoid indirect discrimination if X and P are independent under $\mathbb{P}'$.*

Remark 2.1.8. *By Definition 2.1.7, avoiding direct discrimination is a prerequisite to avoiding indirect discrimination.*

Having defined indirect discrimination, a rigorous verification that the margin-based premium $h(X)$ avoids it is made. The definition states that X and P must be independent under $\mathbb{P}'$, meaning that the joint distribution $\mathbb{P}'(X, P)$ is required. Up to this point, the margin-based premium was said to have $\mathbb{P}' = \mathbb{P}$,

but that is not completely accurate, as X and P are not independent under their empirical joint distribution $\mathbb{P}(X = x, P = p)$. Actually, the margin-based premium must have $\mathbb{P}' = \mathbb{Q}$ which is such that $\mathbb{Q}(P = p) = \mathbb{P}(P = p)$, but $\mathbb{Q}(X = x, P = p) \neq \mathbb{P}(X = x, P = p)$. Again, many $\mathbb{Q}$ will satisfy these conditions, but a natural choice is constructed here.

Since Definition 2.1.7 requires that X and P be independent under $\mathbb{Q}$, we must have $\mathbb{Q}(X = x, P = p) = \mathbb{Q}(X = x) \cdot \mathbb{Q}(P = p)$. For the margin-based premium, we require $\mathbb{Q}(P = p) = \mathbb{P}(P = p)$ but $\mathbb{Q}(X = x)$ is practically arbitrary, as its choice will not impact the resulting premium (see that (2.1.5) does not use $\mathbb{P}'(X = x)$ in any way). Despite the choice of $\mathbb{Q}(X = x)$ not impacting the margin-based premium, $\mathbb{Q}(X = x) = \mathbb{P}(X = x)$ is most natural, and the resulting distribution $\mathbb{Q}(X = x, P = p)$ for the mock portfolio is presented in Table 2.4.

$\mathbb{Q}(\cdot, \cdot)$	$P = 0$	$P = 1$	$P \in \{0, 1\}$
$X = A$	0.12	0.18	0.3
$X = B$	0.12	0.18	0.3
$X = C$	0.16	0.24	0.4
$X \in \{A, B, C\}$	0.4	0.6	1

Table 2.4: Joint distribution $\mathbb{Q}$ of the region X and the immigration status P for the mock portfolio.

Comparing the empirical joint distribution of (X, P) of Table 2.2 to the newly obtained joint distribution of Table 2.4, it can be seen that marginal probabilities (the bottom rows for X and right-most columns for P) are the same, but that the remainders of the tables differ. Furthermore, in Table 2.4, each joint entry is the product of the corresponding marginal probabilities. We dub this method the *Product-of-Marginals* (PoM) method. For example:

$$\mathbb{Q}(X = A, P = 0) = \mathbb{Q}(X = A)\mathbb{Q}(P = 0) = \mathbb{P}(X = A)\mathbb{P}(P = 0) = 0.3 \cdot 0.4 = 0.12$$

Now, the following conditional probabilities are obtained:

$$\mathbb{Q}(P = 1 \mid X = A) = \frac{\mathbb{Q}(P = 1, X = A)}{\mathbb{Q}(X = A)} = \frac{0.18}{0.3} = \frac{3}{5}$$

$$\mathbb{Q}(P = 1 \mid X = B) = \frac{\mathbb{Q}(P = 1, X = B)}{\mathbb{Q}(X = B)} = \frac{0.18}{0.3} = \frac{3}{5}$$

$$\mathbb{Q}(P = 1 \mid X = C) = \frac{\mathbb{Q}(P = 1, X = C)}{\mathbb{Q}(X = C)} = \frac{0.24}{0.4} = \frac{3}{5}$$

All these probabilities being equal validates that no information on P can be gained from knowledge of X.

Another important observation is that the (x, p) combinations with $\mathbb{Q}(X = x, P = p) > 0$ are the same as those with $\mathbb{P}(X = x, P = p) > 0$. This may seem trivial here because all combinations have non-zero probability under $\mathbb{P}$, however suppose $\mathbb{P}(X = A, P = 0) = 0$ had been observed, with $\mathbb{P}(X = A) = 0.3$ and $\mathbb{P}(P = 0) = 0.4$ being unchanged. In that case, the PoM method would still yield $\mathbb{Q}(X = A, P = 0) = \mathbb{P}(X = A)\mathbb{P}(P = 0) = 0.3 \cdot 0.4 = 0.12$, but this would suppose that an impossible combination under $\mathbb{P}$ – the empirical distribution – was possible under the distribution $\mathbb{Q}$ used for pricing which is not acceptable. It would not be reasonable to obtain the premium under assumptions for which there was absolutely no basis.

Recall, this property of $\mathbb{Q}$ to only attribute non-zero probabilities to combinations that already had non-zero probabilities under $\mathbb{P}$ was defined as *absolute continuity* (see Definition 1.3.8), and it is key. When considering $\mathbb{Q}$ as an "adjustment" for $\mathbb{P}$, having $\mathbb{Q} \ll \mathbb{P}$ can provide some comfort in using $\mathbb{Q}$, especially when there already is some level of confidence in $\mathbb{P}$.

Now, considering that the margin-based premium has $\mathbb{P}' = \mathbb{Q}$ (not $\mathbb{P}' = \mathbb{P}$), it is straightforward to verify it avoids indirect discrimination:

1. It avoids direct discrimination;
2. X and P are independent under $\mathbb{P}' = \mathbb{Q}$, by construction of $\mathbb{Q}$.

This construction of the margin-based premium can sometimes guarantee the existence of a discrimination-free premium, as per the following proposition.

Proposition 2.1.9. *Assume there exists a product measure* $\mathbb{P}'(\mathbf{X}, \mathbf{P}) = \mathbb{P}'(\mathbf{X})\mathbb{P}'(\mathbf{P})$ *on* (Z, Σ) *such that* $\mathbb{P}' \ll \mathbb{P}$*. Then, there exists a premium* $h'(\mathbf{X})$ *that avoids indirect discrimination.*

Proof. Absolute continuity implies that every $\mathbb{P}$-null set is also a $\mathbb{P}'$-null set. Therefore, the best-estimate premium $\tau(X, P)$ is well-defined on all sets where (X, P) has a positive $\mathbb{P}'$-probability mass. Since the latter is a product measure, the discrimination-free premium $h'(X)$ can be calculated by integrating $\tau(X, P)$ over $\mathrm{d}\mathbb{P}'(p \mid X) = \mathrm{d}\mathbb{P}'(p)$. This completes the proof. □

The joint distribution $\mathbb{Q}$, constructed using the PoM method on the empirical joint distribution $\mathbb{P}$, is one measure respecting Proposition 2.1.9, and it leads to the margin-based premium. It is important to note that $\mathbb{Q}$ is not the only joint distribution leading to the margin-based premium, as there very well could be others having marginals equivalent to $\mathbb{P}(P = \cdot)$. It is also reiterated here that the margin-based premium is *not* the only possible discrimination-free premium, as marginals other than $\mathbb{P}(P = \cdot)$ can be used in (2.1.4).

2.2 Adjusting the Discrimination-free Premium

With the margin-based premium having exemplified the discrimination-free premium, a flaw of the latter is presented here. Observe that

$$\mathrm{E}\left[h'(X)\right] = \mathrm{E}\left[\int_p \tau(X, p)\,\mathrm{d}\mathbb{P}'(p)\right] \neq \mathrm{E}\left[\int_p \tau(X, p)\,\mathrm{d}\mathbb{P}(p)\right] = \mathrm{E}\left[\mathrm{E}\left(Y \mid X\right)\right] = \mathrm{E}\left[Y\right].$$

This means that the discrimination-free premium will, in general, not be unbiased. It is desirable for the discrimination-free premium to be unbiased because that would ensure that, in aggregate, the insurer has enough capital to cover all their obligations. If the choice of $\mathbb{P}'$ in (2.1.4) leads to a premium that is smaller, on average, than $\mathrm{E}\left[Y\right]$, the insurer is *expecting* to be unable to pay for certain claims.

Let B' be the bias of the discrimination-free premium h':

$$B' = \mathrm{E}\left[h'(X)\right] - \mathrm{E}\left[Y\right]. \tag{2.2.1}$$

As per (2.1.7), B' can be positive, zero or negative. To correct for this bias, two simplistic approaches are proposed.

The first is an additive approach. Let the *uniformly adjusted $\mathbb{P}'$-discrimination-free premium* be

$$h'^{,u}(X) = h'(X) - B'. \tag{2.2.2}$$

This premium applies the same flat correction to all individuals, independently of both the explanatory and the protected variables. While this approach does ensure that the resulting distribution of premiums will be unbiased, it is possible that it produces negative premiums, which should surely be avoided. It is demonstrated below that the uniformly adjusted $\mathbb{P}'$-discrimination-free premium is unbiased:

Proof.

$$\begin{aligned}
\mathrm{E}\left[h'^{,u}(X)\right] &= \mathrm{E}\left[h'(X) - B'\right] \\
&= \mathrm{E}\left[h'(X)\right] - \mathrm{E}\left[B'\right] \\
&= \mathrm{E}\left[h'(X)\right] - B' \\
&= \mathrm{E}\left[h'(X)\right] - \left(\mathrm{E}\left[h'(X)\right] - \mathrm{E}\left[Y\right]\right) \\
&= \mathrm{E}\left[Y\right].
\end{aligned}$$

□

The second approach is a multiplicative approach. Let the *proportionally adjusted $\mathbb{P}'$-discrimination-free premium* be

$$h'^{,u}(X) = h'(X)\frac{\mathrm{E}\left[Y\right]}{\mathrm{E}\left[Y\right] + B'} = h'(X)\frac{\mathrm{E}\left[Y\right]}{\mathrm{E}\left[h'(X)\right]}. \tag{2.2.3}$$

This premium distributes the bias according to the size of the discrimination-free premium. This approach also ensures that the distribution of the discrimination-free premium will be unbiased, however individuals with a bigger discrimination-free premium will suffer from a bigger effect (in absolute value) of the bias correction. For example, suppose that Alice has a discrimination-free premium of 100, Bob has a discrimination-free premium of 180, $\mathrm{E}[Y] = 200$ and $\mathrm{E}[h'(\mathbf{X})] = 160$. Then, Alice and Bob's proportionally adjusted discrimination-free premium would be $100 \cdot \frac{200}{160} = 125$ and $180 \cdot \frac{200}{160} = 225$, respectively. This would mean Alice's premium increases by $125 - 100 = 25$ while Bob's premium increases by $225 - 180 = 45$ to correct for the bias. The proportional impact – an increase of 25% – is the same for both, but bigger premiums having bigger impacts could be considered unfair.

With the additive and multiplicative approaches both having their disadvantages, a more sophisticated approach is proposed. A desirable property of $\mathbb{P}'$ would be that it is close to $\mathbb{P}$. The closer $\mathbb{P}$ is to $\mathbb{P}'$, the more "realistic" it is. Quantifying the difference between two distributions requires a new tool, defined below.

Definition 2.2.1 (Kullback-Leibler Divergence)**.** *Consider two probability measures $\mathbb{P}_1$ and $\mathbb{P}_2$ defined on the same measurable space (Z, Σ). Then, the* Kullback-Leibler (KL) divergence from $\mathbb{P}_1$ to $\mathbb{P}_2$ *is the following*

$$D_{\mathrm{KL}}(\mathbb{P}_2 \parallel \mathbb{P}_1) = \mathrm{E}^{\mathbb{P}_1}\left(\frac{\mathbb{P}_2(X)}{\mathbb{P}_1(X)} \log\left(\frac{\mathbb{P}_2(X)}{\mathbb{P}_1(X)}\right)\right).$$

It is always non-negative and vanishes if and only if $\mathbb{P}_1 = \mathbb{P}_2$.

If $\mathbb{P}_2 \ll \mathbb{P}_1$, then $\frac{\mathbb{P}_2(X)}{\mathbb{P}_1(X)} = \xi(X)$ where ξ is the change of measure function allowing to go from $\mathbb{P}_1$ to $\mathbb{P}_2$, and so

$$D_{\mathrm{KL}}(\mathbb{P}_2 \parallel \mathbb{P}_1) = \mathrm{E}^{\mathbb{P}_1}\left[\xi(X) \log\left(\xi(X)\right)\right].$$

We note that the KL divergence is a convex function of the probability measure $\mathbb{P}_2$.

Remark 2.2.2. *The KL divergence is* not *a distance, so it is not symmetric. That is, the KL divergence from* $\mathbb{P}_1$ *to* $\mathbb{P}_2$ *is not equal to the KL divergence from* $\mathbb{P}_2$ *to* $\mathbb{P}_1$*:*

$$D_{\mathrm{KL}}\left(\mathbb{P}_2 \parallel \mathbb{P}_1\right) \neq D_{\mathrm{KL}}\left(\mathbb{P}_1 \parallel \mathbb{P}_2\right).$$

Due to Csiszár [4], it is possible to minimize KL divergence subject to the mean being unchanged. First, denote

$$\psi(P) = \int_x \tau(x, P)\, \mathrm{d}\mathbb{P}(x).$$

The above expression is the counterpart of $h(X)$ in the sense that the best-estimate premium is averaged over the distribution of X instead of the distribution of P. Then, due to independence of X and P under $\mathbb{P}'$, the following is true:

$$\mathrm{E}\left[h'(X)\right] = \mathrm{E}'\left[\psi(P)\right].$$

Proof.

$$\begin{aligned}
\mathrm{E}\left[h'(X)\right] &= \mathrm{E}\left[\int_p \tau(X, p)\, \mathrm{d}\mathbb{P}'(p)\right] \\
&= \int_x \left[\int_p \tau(X, p)\, \mathrm{d}\mathbb{P}'(p)\right] \mathrm{d}\mathbb{P}(x) \\
&= \int_p \int_x \tau(X, p)\, \mathrm{d}\mathbb{P}(x) \mathrm{d}\mathbb{P}'(p) \\
&= \int_p \psi(P) \mathrm{d}\mathbb{P}'(p) \\
&= \mathrm{E}'\left[\psi(P)\right].
\end{aligned}$$

□

Expressing the expectation in terms of E' rather than E is a trick that allows the use of Csiszár's methodology. Consider the following optimization problem. It is required to find

$$\arg\min_{\mathbb{P}'} \mathrm{E}\left[\frac{\mathrm{d}\mathbb{P}'}{\mathrm{d}\mathbb{P}'}\log\left(\frac{\mathrm{d}\mathbb{P}'}{\mathrm{d}\mathbb{P}'}\right)\right], \quad \text{such that} \quad \mathrm{E}'\left[\psi(P)\right] = \mathrm{E}\left[Y\right]. \tag{2.2.4}$$

Csiszár (1975) [4] presents the solution to this problem as $\mathbb{P}'$ such that

$$\mathbb{P}'(p) = \mathrm{E}\left[\mathbf{1}_{\{P\leq p\}}\frac{e^{\beta\psi(P)}}{\mathrm{E}\left[e^{\beta\psi(P)}\right]}\right],$$

where $\mathbf{1}_{\{.\}}$ is the indicator function (1 when the argument is true, 0 otherwise) and β is a parameter suitably chosen such that the constraint $\mathrm{E}'\left[\psi(P)\right] = \mathrm{E}\left[Y\right]$ is fulfilled. The resulting premium is defined properly below.

Definition 2.2.3 (KL-based premium)**.** *The* KL-based premium Y^{KL} *with respect to* X *is:*

$$h'^{,KL}(x) = h'(x) = \mathrm{E}\left[\tau(x,P)\frac{e^{\beta\psi(P)}}{\mathrm{E}\left[e^{\beta\psi(P)}\right]}\right].$$

It is also a special case of the discrimination-free premium, with $\mathbb{P}'$ *being the solution to the optimization problem given by (2.2.4).*

Applying this methodology to our mock portfolio results in

$$\mathbb{P}'(P=0) = 0.4285714 \quad \text{and} \quad \mathbb{P}'(P=1) = 0.5714286.$$

Note that this is quite close to the empirical distribution $\mathbb{P}$, which had $\mathbb{P}(P = 0) = 0.4$ and $\mathbb{P}(P = 1) = 0.6$. Figure 2.3 updates the premiums already obtained for the mock portfolio with the KL-based premiums. As could be expected due to how close $\mathbb{P}$ and $\mathbb{P}'$ are with respect to the KL divergence, the margin-based and KL-based premiums are close. Table 2.5 shows that the KL-based premium is lower than the margin-based premium in regions A and C. That is because the two regions had in common that the best-estimate premium was lower for $P = 0$ than for $P = 1$ (see Table 2.3). Recall the unawareness, margin-based and KL-based premiums are all weighted averages of best-estimate premiums within a region, meaning that $\mathbb{P}'(P = 0) > \mathbb{P}(P = 0)$ implies the KL-based premiums will always be lower than the margin-based premiums for regions A and C of this

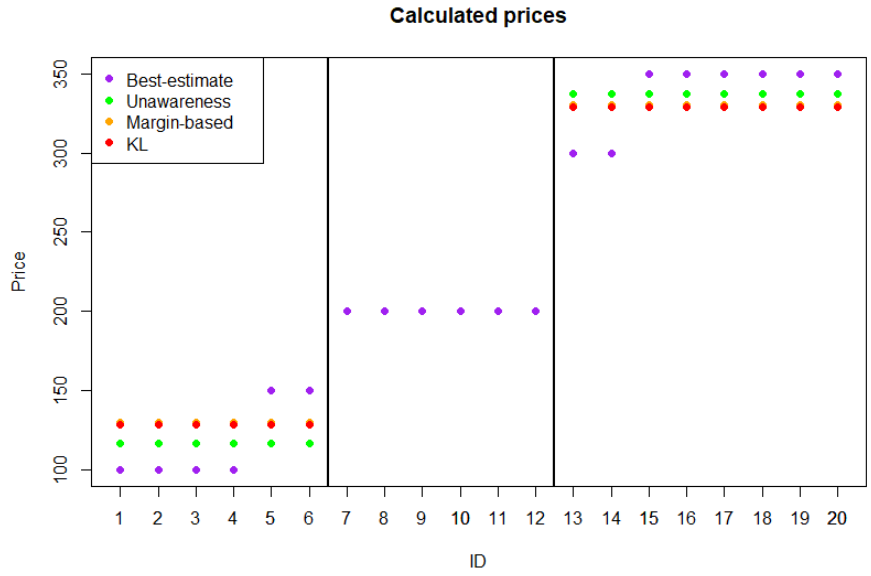

Figure 2.3: Best-estimate, unawareness and margin-based premiums for all individuals of the mock portfolio. Vertical lines segment the individuals by region of residence, with the left-most section being region A, the middle section being region B and the right-most section being region C.

portfolio. As for region B, all premiums are 200 independently of the weighing, since the two best-estimate premiums for that region are 200.

	Unawareness	Margin-based	KL-based
A	116.67	130.00	128.57
B	200.00	200.00	200.00
C	337.50	330.00	328.57

Table 2.5: Unawareness, margin-based and KL-based premiums for all regions of the mock portfolio

To summarize, three methods have been suggested to correct the bias introduced by the discrimination-free premium:

1. The first is an additive approach. The uniformly adjusted $\mathbb{P}'$-discrimination-free premium of (2.2.2) has the disadvantage of possibly producing negative premiums;

2. The second is a multiplicative approach. The proportionally adjusted $\mathbb{P}'$-discrimination-free premium of (2.2.3) has the disadvantage of applying a

bigger chance to bigger initial premiums;

3. The third approach is more complex. It solves an optimization problem requiring that the distribution of the resulting premiums is unbiased, while ensuring that it is as close as possible to the empirical distribution $\mathbb{P}$ with respect to the KL divergence.

Among these approaches, the third is to be preferred and it is considered an integral part of the implementation of the discrimination-free premium.

2.3 Advantages and Disadvantages to the Discrimination-free Premium

The discrimination-free premium has many advantages, the main one being removing the dependence of Y on P due to $\mathbf{X}$. It is an elegant, easily justifiable and intuitive way of obtaining fairer premiums. Also, the fact that the discrimination-free premium is unbiased guarantees that the insurer has, on average, enough capital to cover their claims.

However, it also has a few disadvantages. As an intra-treatment approach to fairness, it requires *all* the data available to the insurer, namely Y, $\mathbf{X}$ and $\mathbf{P}$, but this is not unreasonable, as it is likely the insurer themselves would implement this approach on their portfolio. Additionally, up until now, we have assumed that the model function f was simply the expected value based on the empirical measure $\mathbb{P}$, but insurers are using increasingly complicated algorithms to calculate insurance premiums. To implement this approach, these algorithms need to be modified to take $\mathbf{P}$ as inputs, and that may not be trivial work.

Furthermore, once the model function has been adequately altered, there is a subsequent increase to computational cost. If we refer to Section 2.1, one calculation of a newcomer's discrimination-free premium required on its own one intermediary calculation per category of P, which, at the very least, doubles

computation time. When there are multiple protected variables $\mathbf{P}$, this becomes one intermediary calculation *per possible state* of $\mathbf{P}$, further increasing computation time.

Moreover, recall changing measures from the empirical $\mathbb{P}$ to a $\mathbb{P}'$ under which X and P are independent is a requirement to the discrimination-free premium. By construction and necessity of the discrimination-free premium, $\mathbb{P}'$ will be a distribution under which the marginal distribution of P will have changed, which is a very strong assumption. Indeed, protected variables can be very static, and not change over the course of an individual's life. Premiums based on a different probabilistic structure of protected variables may be considered too unrealistic depending on the nature of those variables, despite minimization of the KL divergence from $\mathbb{P}$ to $\mathbb{P}'$.

Additionally, Lindholm et al. (2022) [8] do not discuss the use of a quantifier that could help assess whether the goal of the discrimination-free premium was attained after implementation.

Chapter 3

Grid-based Change of Measure

Pesenti et al. (2018) [9] use a change of measure on a single random variable to perform sensitivity analysis at a given quantile level of the random variable. In this chapter, we generalize that approach to a change of measure at various quantile levels of multiple random variables simultaneously.

3.1 Framework

Let $\mathbf{Z} = (Z_1, \ldots, Z_D)$ be a vector of real-valued random variables, where D is the number of dimensions. Each Z_d has a corresponding *split-vector* (or a vector of splits) $\mathbf{t}_d = (t_{d,0}, t_{d,1}, \ldots, t_{d,S_d}, t_{d,S_d+1})$ where $\inf \mathrm{Dom}(Z_d) = t_{d,0} < t_{d,1} < \cdots < t_{d,S_d} < t_{d,S_d+1} = \sup \mathrm{Dom}(Z_d)$ for $d = 1, \ldots, D$ and S_d is number of chosen splits of $\mathbf{t}_d$, *i.e.* the number of splits for the d^{th} random variable X_d. Note that $t_{d,0}$ and t_{d,S_d+1} cannot be "chosen", as they are imposed by the domain of the random variable Z_d. We also let $\mathcal{T} = \{\mathbf{t}_1, \ldots, \mathbf{t}_D\}$ be the set of split-vectors.

For a single Z_d, the domain $\mathrm{Dom}(Z_d)$ is partitioned by the vector $\mathbf{t}_d$ into $S_d + 1$ intervals that are numbered from 0 to S_d in increasing order and denoted $H_{d,0}, H_{d,1}, \ldots, H_{d,S_d}$. For example, say $Z_1 \sim \mathrm{Unif}(0, 1)$ and let $\mathbf{t}_1 = (0, 0.4, 0.9, 1)$. Then, the number of chosen splits is $S_1 = 2$ resulting in $(S_1 + 1) = 3$ intervals,

numbered from 0 to 2. The 0^{th} interval of Z_1 would be $H_{1,0} = [0, 0.4]$, the 1^{st} interval of Z_1 would be $H_{1,1} = (0.4, 0.9]$ and the 2^{nd} interval of Z_1 would be $H_{1,2} = (0.9, 1]$. The three intervals form a partition of $\text{Dom}(Z_1) = [0, 1]$.

We use as convention that all sub-intervals are left-open and right-closed, with the exception of the 0^{th} intervals which are closed. The intervals we consider will each be of one of two forms:

1. $H_{d,0} = [t_{d,0}, t_{d,1}]$ for some $d \in \{1, \dots D\}$;
2. $H_{d,i} = (t_{d,i}, t_{d,i+1}]$ for some $d \in \{1, \dots D\}$ and some $i \in \{1, \dots, S_d\}$.

Naturally, if $t_{d,0} = \inf \text{Dom}(Z_d) = -\infty$, then the 0^{th} interval of the partition for Z_d will be left-open and right closed rather than closed. Similarly, if $t_{d,S_d+1} = \sup \text{Dom}(Z_d) = \infty$, then the S_d^{th} interval of the partition for Z_d will be open rather than left-open and right-closed.

For the D-tuple $\mathbf{Z}$, these $N_S = \sum_{d=1}^{D} S_d$ splits separate $\text{Dom}(\mathbf{Z}) = \times_{d=1}^{D} \text{Dom}(Z_d)$ in $N_R = \prod_{d=1}^{D}(S_d + 1)$ regions. Let $\mathcal{I}_\mathcal{T}$ be the set of (non-random) D-tuples with d^{th} element in $\{0, \dots, S_d\}$. Then, $\mathcal{I}_\mathcal{T}$ has exactly N_R elements. Each of these elements will be used to denote one of the regions of $\text{Dom}(\mathbf{X})$ delimited by the set of split-vectors $\mathcal{T}$. For any integer number of dimensions $D \geq 1$, we define a *region* to be one of the subsets of $\text{Dom}(\mathbf{Z})$ delimited by $\mathcal{T}$ and $\mathcal{R}_\mathcal{T}$ to be the set of such regions. For any $\mathbf{i} \in \mathcal{I}_\mathcal{T}$, the region $R_\mathbf{i} \in \mathcal{R}_\mathcal{T}$ corresponds to the cross-product of the sub-intervals designated by each element of $\mathbf{i}$.

For example, if $\mathbf{i} = (0, \dots, 0)$, then $R_\mathbf{i} \subset \text{Dom}(\mathbf{Z})$ denotes the cross-product of all 0^{th} sub-intervals, *i.e.* $R_\mathbf{i} = [0, t_{1,1}] \times \cdots \times [0, t_{D,1}]$. As another example, if $\mathbf{j} = (1, 0, \dots, 0)$, then $R_\mathbf{j} \subset \text{Dom}(\mathbf{Z})$ denotes the cross-product of the region where Z_1 is in its 1^{st} sub-interval, but all other risks are in their 0^{th} sub-interval, *i.e.* $R_\mathbf{j} = (t_{1,1}, t_{1,2}] \times [0, t_{2,1}] \times \cdots \times [0, t_{D,1}]$. Note that all elements of $\mathcal{R}_\mathcal{T}$ are pairwise-disjoint and form a partition of the domain $\text{Dom}(\mathbf{Z})$.

Consider the special case where $Z_i \sim \text{Unif}(0, 1)$ for $i = 1, 2, 3$ and the splits are equidistant within their split-vector, *i.e.* $\forall d \in \{1, 2, 3\}\ \forall k \in \{0, \dots, S_d\} \quad t_{d,k+1} -$

$t_{d,k} = c > 0$. Then, we have the following:

1. In 1 dimension ($D = 1$), we have split the interval $[0, 1]$ in $S_1 + 1$ sub-intervals that all have the same length c;
2. In 2 dimensions ($D = 2$), we have split the square $[0, 1]^2$ in $(S_1 + 1)(S_2 + 1)$ sub-squares that all have the same area c^2;
3. In 3 dimensions ($D = 3$), we have split the cube $[0, 1]^3$ in $(S_1 + 1)(S_2 + 1)(S_3 + 1)$ sub-cubes that all have the same volume c^3.

Now, consider the probability of $\mathbf{Z}$ to lie in any given region. Let $\mathbb{P}$ be the (known) probability measure for $\mathbf{Z}$. Then, we define $A = \{\alpha_{\mathbf{i}} \mid \mathbf{i} \in \mathcal{I}_{\mathcal{T}}\}$ such that

$$\alpha_{\mathbf{i}} = \mathbb{P}(\mathbf{Z} \in R_{\mathbf{i}}).$$

Figure 3.1 illustrates the segmentation of $\mathrm{Dom}(\mathbf{Z})$ with respect to $\mathcal{R}_{\mathcal{T}}$ and $\mathbb{P}$, and it will be referred to as the *A-grid*.

Since $\mathcal{R}_{\mathcal{T}}$ is a partition of the domain, we have that $\sum_{\mathbf{i} \in \mathcal{T}} \alpha_{\mathbf{i}} = 1$ and we impose the restriction that $\mathcal{T}$ must be chosen such that $\forall \mathbf{i} \in \mathcal{T}\ \alpha_{\mathbf{i}} > 0$. Such a $\mathcal{T}$ always exists since it is always possible to have no chosen splits for every element of $\mathbf{Z}$, resulting in the following single region:

$$\begin{aligned} R_{0,\ldots,0} &= H_{1,0} \times \cdots \times H_{D,0} \\ &= [t_{1,0}, t_{1,1}] \times \cdots \times [t_{D,0}, t_{D,1}] \\ &= [\inf \mathrm{Dom}(Z_1), \sup \mathrm{Dom}(Z_1)] \times \cdots \times [\inf \mathrm{Dom}(Z_D), \sup \mathrm{Dom}(Z_D)] \\ &= \mathrm{Dom}(\mathbf{Z}). \end{aligned}$$

Suppose the probability measure $\mathbb{P}$ is not satisfactory and a different probability measure, say $\mathbb{P}^*$, with $\mathbb{P}^*(\mathbf{Z} \in R_{\mathbf{i}}) = \kappa_{\mathbf{i}}$, is adequate. We define the set $\mathcal{Q}_{\mathcal{T}}$ as follows.

$$\mathcal{Q}_{\mathcal{T}} = \{\mathbb{P}^* \mid \forall \mathbf{i} \in \mathcal{I}_{\mathcal{T}}\ \ \mathbb{P}^*(\mathbf{Z} \in R_{\mathbf{i}}) = \kappa_{\mathbf{i}}\}. \tag{3.1.1}$$

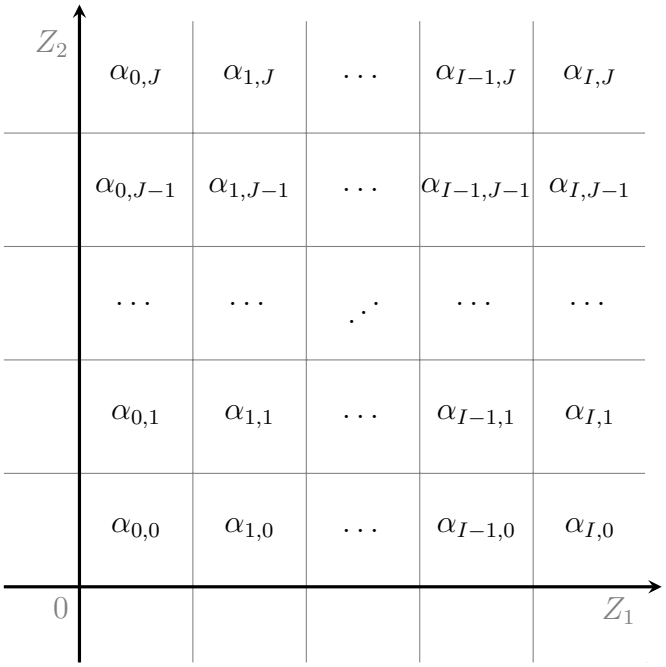

Figure 3.1: Graphical representation of the segmentation of $\mathrm{Dom}(\mathbf{Z})$ when $D = 2$ and $Z_i \sim \mathrm{Unif}(0,1)$ for a general choice of splits. The $\alpha_{(i,j)}$ represent the probability of the random vector $\mathbf{Z} = (Z_1, Z_2)$ of lying in the corresponding region $R_{(i,j)}$ for $i \in \{0, 1, \ldots, I\}$ and $j \in \{0, 1, \ldots, J\}$ where $I = S_1$ and $J = S_2$.

Our objective is then to perform a change of measure from $\mathbb{P}$ to any element of $\mathcal{Q}_{\mathcal{T}}$. To that effect, a change of measure function is necessary. Proposition 3.1.1 gives the required function and introduces the resulting probability measure $\mathbb{Q}$, whose use is the crux of this work.

Proposition 3.1.1. *Let $\mathbf{Z} = (Z_1, \ldots, Z_D)$ be a D-dimensional random vector having domain $Dom(\mathbf{Z})$, $\mathcal{T} = \{\mathbf{t}_1, \ldots, \mathbf{t}_D\}$ be the set of split-vectors, $\mathcal{R}_{\mathcal{T}}$ be the set of regions in the domain partitioned by $\mathcal{T}$ and $\mathcal{I}_{\mathcal{T}}$ be the set of indices resulting from $\mathcal{T}$. Also, let $\mathbb{P}$ be such that $\mathbb{P}(\mathbf{Z} \in R_{\mathbf{i}}) = \alpha_{\mathbf{i}}$. Recall $\mathcal{T}$ is chosen such that $\forall \mathbf{i} \in \mathcal{I}_{\mathcal{T}}\ \alpha_{\mathbf{i}} > 0$. Finally, let $\mathcal{Q}_{\mathcal{T}}$ be defined as in (3.1.1). Then, the following function is a change of measure function allowing to go from $\mathbb{P}$ to another measure $\mathbb{Q} \in \mathcal{Q}_{\mathcal{T}}$:*

$$\gamma(\mathbf{Z}) = \sum_{\mathbf{i} \in \mathcal{I}_{\mathcal{T}}} \frac{\kappa_{\mathbf{i}}}{\alpha_{\mathbf{i}}} \mathbf{1}_{\{\mathbf{Z} \in R_{\mathbf{i}}\}}.$$

Proof. Let $\mathbb{Q}$ be the probability measure obtained from applying the change of

measure function γ to $\mathbb{P}$. Then, for any $\mathbf{i} \in \mathcal{I}_{\mathcal{T}}$, the region $R_{\mathbf{i}} \in \mathcal{R}_{\mathcal{T}}$, has the following probability under $\mathbb{Q}$:

$$
\begin{aligned}
\mathbb{Q}(\mathbf{Z} \in R_{\mathbf{i}}) &= \mathrm{E}^{\mathbb{Q}}\left(\mathbf{1}_{\{\mathbf{Z} \in R_{\mathbf{i}}\}}\right) \\
&= \mathrm{E}^{\mathbb{P}}\left(\gamma\left(\mathbf{Z}\right)\mathbf{1}_{\{\mathbf{Z} \in R_{\mathbf{i}}\}}\right) \\
&= \sum_{\mathbf{j} \in \mathcal{I}_{\mathcal{T}}} \mathrm{E}^{\mathbb{P}}\left(\gamma\left(\mathbf{Z}\right)\mathbf{1}_{\{\mathbf{Z} \in R_{\mathbf{i}}\}} \;\middle|\; \mathbf{Z} \in R_{\mathbf{j}}\right)\mathbb{P}(\mathbf{Z} \in R_{\mathbf{j}}) \\
&= \mathrm{E}^{\mathbb{P}}\left(\gamma\left(\mathbf{Z}\right)\mathbf{1}_{\{\mathbf{Z} \in R_{\mathbf{i}}\}} \;\middle|\; \mathbf{Z} \in R_{\mathbf{i}}\right)\mathbb{P}(\mathbf{Z} \in R_{\mathbf{i}}) \\
&\quad + \sum_{\mathbf{j} \in \mathcal{I}_{\mathcal{T}} \setminus \mathbf{i}} \mathrm{E}^{\mathbb{P}}\left(\gamma\left(\mathbf{Z}\right)\mathbf{1}_{\{\mathbf{Z} \in R_{\mathbf{i}}\}} \;\middle|\; \mathbf{Z} \in R_{\mathbf{j}}\right)\mathbb{P}(\mathbf{Z} \in R_{\mathbf{j}}) \\
&= \mathrm{E}^{\mathbb{P}}\left(\gamma\left(\mathbf{Z}\right)\mathbf{1}_{\{\mathbf{Z} \in R_{\mathbf{i}}\}} \;\middle|\; \mathbf{Z} \in R_{\mathbf{i}}\right)\alpha_{\mathbf{i}} \\
&\quad + \sum_{\mathbf{j} \in \mathcal{I}_{\mathcal{T}} \setminus \mathbf{i}} \mathrm{E}^{\mathbb{P}}\left(\gamma\left(\mathbf{Z}\right)\mathbf{1}_{\{\mathbf{Z} \in R_{\mathbf{i}}\}} \;\middle|\; \mathbf{Z} \in R_{\mathbf{j}}\right)\alpha_{\mathbf{j}} \\
&= \mathrm{E}^{\mathbb{P}}\left(\gamma\left(\mathbf{Z}\right)\mathbf{1}_{\{\mathbf{Z} \in R_{\mathbf{i}}\}} \;\middle|\; \mathbf{Z} \in R_{\mathbf{i}}\right)\alpha_{\mathbf{i}} + 0 \\
&= \mathrm{E}^{\mathbb{P}}\left[\left(\frac{\kappa_{\mathbf{i}}}{\alpha_{\mathbf{i}}}\mathbf{1}_{\{\mathbf{Z} \in R_{\mathbf{i}}\}} + \sum_{\mathbf{k} \in \mathcal{I}_{\mathcal{T}} \setminus \mathbf{i}}\left(\frac{\kappa_{\mathbf{k}}}{\alpha_{\mathbf{k}}}\mathbf{1}_{\{\mathbf{Z} \in R_{\mathbf{k}}\}}\right)\right)\mathbf{1}_{\{\mathbf{Z} \in R_{\mathbf{i}}\}} \;\middle|\; \mathbf{Z} \in R_{\mathbf{i}}\right]\alpha_{\mathbf{i}} \\
&= \mathrm{E}^{\mathbb{P}}\left[\left(\frac{\kappa_{\mathbf{i}}}{\alpha_{\mathbf{i}}} \cdot 1 + \sum_{\mathbf{k} \in \mathcal{I}_{\mathcal{T}} \setminus \mathbf{i}}\left(\frac{\kappa_{\mathbf{k}}}{\alpha_{\mathbf{k}}} \cdot 0\right)\right) \cdot 1\right]\alpha_{\mathbf{i}} \\
&= \mathrm{E}^{\mathbb{P}}\left(\frac{\kappa_{\mathbf{i}}}{\alpha_{\mathbf{i}}} \cdot 1 \cdot 1\right)\alpha_{\mathbf{i}} \\
&= \frac{\kappa_{\mathbf{i}}}{\alpha_{\mathbf{i}}} \cdot \alpha_{\mathbf{i}} \\
&= \kappa_{\mathbf{i}},
\end{aligned}
$$

where the following fact is used throughout:

$$\forall R_{\mathbf{i}}, R_{\mathbf{j}} \in \mathcal{R}_{\mathcal{T}} \;\; R_{\mathbf{i}} \neq R_{\mathbf{j}} \implies R_{\mathbf{i}} \cap R_{\mathbf{j}} = \emptyset \implies \mathbf{1}_{\{\mathbf{Z} \in R_{\mathbf{i}}\}} \mid \mathbf{Z} \in R_{\mathbf{j}} = 0.$$

□

There may be many elements in the set $\mathcal{Q}_{\mathcal{T}}$, and we will use $\mathbb{Q}$ to refer to the one that is obtained from Proposition 3.1.1. However, in the proposition, we only

gave the expression of the probability measure $\mathbb{Q}$ evaluated at sets contained in $\mathcal{R}_{\mathcal{T}}$. We define it here for a general set $E \subset \mathrm{Dom}(\mathbf{Z})$.

Proposition 3.1.2. *Suppose $E \subseteq Dom(\mathbf{Z})$. Let $\mathcal{I}_{\mathcal{T}}^0 = \{\mathbf{j} \in \mathcal{I}_{\mathcal{T}} \mid E \cap R_{\mathbf{j}} = \emptyset\}$, $\mathcal{I}_{\mathcal{T}}^R = \{\mathbf{j} \in \mathcal{I}_{\mathcal{T}} \mid E \cap R_{\mathbf{j}} = R_{\mathbf{j}}\}$ and $\mathcal{I}_{\mathcal{T}}^E = \mathcal{I}_{\mathcal{T}} \setminus \left(\mathcal{I}_{\mathcal{T}}^0 \cup \mathcal{I}_{\mathcal{T}}^R\right)$. Then, the expression of the probability measure $\mathbb{Q}$ attained from $\mathbb{P}$ through the change of measure function defined in Proposition 3.1.1 evaluated at E is:*

$$\mathbb{Q}(\mathbf{Z} \in E) = \sum_{\mathbf{j} \in \mathcal{I}_{\mathcal{T}}} \frac{\kappa_{\mathbf{j}}}{\alpha_{\mathbf{j}}} \mathbb{P}\left(\mathbf{Z} \in E \cap R_{\mathbf{j}}\right), \tag{3.1.2}$$

and it can be decomposed into:

$$\mathbb{Q}(\mathbf{Z} \in E) = \sum_{\mathbf{j} \in \mathcal{I}_{\mathcal{T}}^R} \kappa_{\mathbf{j}} + \sum_{\mathbf{j} \in \mathcal{I}_{\mathcal{T}}^E} \frac{\kappa_{\mathbf{j}}}{\alpha_{\mathbf{j}}} \mathbb{P}\left(\mathbf{Z} \in E \cap R_{\mathbf{j}}\right). \tag{3.1.3}$$

Proof. The proof begins similarly to that of Proposition 3.1.1.

$$
\begin{aligned}
\mathbb{Q}(\mathbf{Z} \in E) &= \mathrm{E}^{\mathbb{Q}}\left(\mathbf{1}_{\{\mathbf{Z}\in E\}}\right) \\
&= \mathrm{E}^{\mathbb{P}}\left(\gamma\left(\mathbf{Z}\right)\mathbf{1}_{\{\mathbf{Z}\in E\}}\right) \\
&= \sum_{\mathbf{j}\in\mathcal{I}_{\mathcal{T}}} \mathrm{E}^{\mathbb{P}}\left(\gamma\left(\mathbf{Z}\right)\mathbf{1}_{\{\mathbf{Z}\in E\}} \;\middle|\; \mathbf{Z}\in R_{\mathbf{j}}\right)\mathbb{P}(\mathbf{Z}\in R_{\mathbf{j}}) \\
&= \sum_{\mathbf{j}\in\mathcal{I}_{\mathcal{T}}} \mathrm{E}^{\mathbb{P}}\left[\left(\frac{\kappa_{\mathbf{j}}}{\alpha_{\mathbf{j}}}\mathbf{1}_{\{\mathbf{Z}\in R_{\mathbf{j}}\}} + \sum_{\mathbf{k}\in\mathcal{I}_{\mathcal{T}}\setminus\mathbf{j}}\left(\frac{\kappa_{\mathbf{k}}}{\alpha_{\mathbf{k}}}\mathbf{1}_{\{\mathbf{Z}\in R_{\mathbf{k}}\}}\right)\right)\mathbf{1}_{\{\mathbf{Z}\in E\}} \;\middle|\; \mathbf{Z}\in R_{\mathbf{j}}\right]\alpha_{\mathbf{j}} \\
&= \sum_{\mathbf{j}\in\mathcal{I}_{\mathcal{T}}} \mathrm{E}^{\mathbb{P}}\left(\frac{\kappa_{\mathbf{j}}}{\alpha_{\mathbf{j}}}\mathbf{1}_{\{\mathbf{Z}\in R_{\mathbf{j}}\}}\mathbf{1}_{\{\mathbf{Z}\in E\}} + \sum_{\mathbf{k}\in\mathcal{I}_{\mathcal{T}}\setminus\mathbf{j}}\left(\frac{\kappa_{\mathbf{k}}}{\alpha_{\mathbf{k}}}\mathbf{1}_{\{\mathbf{Z}\in R_{\mathbf{k}}\}}\mathbf{1}_{\{\mathbf{Z}\in E\}}\right) \;\middle|\; \mathbf{Z}\in R_{\mathbf{j}}\right)\alpha_{\mathbf{j}} \\
&= \sum_{\mathbf{j}\in\mathcal{I}_{\mathcal{T}}} \mathrm{E}^{\mathbb{P}}\left(\frac{\kappa_{\mathbf{j}}}{\alpha_{\mathbf{j}}}\mathbf{1}_{\{\mathbf{Z}\in E\cap R_{\mathbf{j}}\}} + \sum_{\mathbf{k}\in\mathcal{I}_{\mathcal{T}}\setminus\mathbf{j}}\left(\frac{\kappa_{\mathbf{k}}}{\alpha_{\mathbf{k}}}\mathbf{1}_{\{\mathbf{Z}\in E\cap R_{\mathbf{k}}\}}\right) \;\middle|\; \mathbf{Z}\in R_{\mathbf{j}}\right)\alpha_{\mathbf{j}} \\
&= \sum_{\mathbf{j}\in\mathcal{I}_{\mathcal{T}}} \mathrm{E}^{\mathbb{P}}\left(\frac{\kappa_{\mathbf{j}}}{\alpha_{\mathbf{j}}}\left(\mathbf{1}_{\{\mathbf{Z}\in E\cap R_{\mathbf{j}}\}} \mid \mathbf{Z}\in R_{\mathbf{j}}\right) + \sum_{\mathbf{k}\in\mathcal{I}_{\mathcal{T}}\setminus\mathbf{j}}\left(\frac{\kappa_{\mathbf{k}}}{\alpha_{\mathbf{k}}}\cdot 0\right)\right)\alpha_{\mathbf{j}} \\
&= \sum_{\mathbf{j}\in\mathcal{I}_{\mathcal{T}}} \mathrm{E}^{\mathbb{P}}\left(\frac{\kappa_{\mathbf{j}}}{\alpha_{\mathbf{j}}}\left(\mathbf{1}_{\{\mathbf{Z}\in E\cap R_{\mathbf{j}}\}} \mid \mathbf{Z}\in R_{\mathbf{j}}\right)\right)\alpha_{\mathbf{j}} \\
&= \sum_{\mathbf{j}\in\mathcal{I}_{\mathcal{T}}} \frac{\kappa_{\mathbf{j}}}{\alpha_{\mathbf{j}}}\left[\alpha_{\mathbf{j}}\right]\mathrm{E}^{\mathbb{P}}\left(\mathbf{1}_{\{\mathbf{Z}\in E\cap R_{\mathbf{j}}\}} \mid \mathbf{Z}\in R_{\mathbf{j}}\right) \\
&= \sum_{\mathbf{j}\in\mathcal{I}_{\mathcal{T}}} \kappa_{\mathbf{j}}\mathbb{P}\left(\mathbf{Z}\in E\cap R_{\mathbf{j}} \mid \mathbf{Z}\in R_{\mathbf{j}}\right) \\
&= \sum_{\mathbf{j}\in\mathcal{I}_{\mathcal{T}}} \kappa_{\mathbf{j}}\frac{\mathbb{P}\left(\mathbf{Z}\in E\cap R_{\mathbf{j}}, \mathbf{Z}\in R_{\mathbf{j}}\right)}{\mathbb{P}\left(\mathbf{Z}\in R_{\mathbf{j}}\right)} \\
&= \sum_{\mathbf{j}\in\mathcal{I}_{\mathcal{T}}} \kappa_{\mathbf{j}}\frac{\mathbb{P}\left(\mathbf{Z}\in E\cap R_{\mathbf{j}}\right)}{\mathbb{P}\left(\mathbf{Z}\in R_{\mathbf{j}}\right)} \\
&= \sum_{\mathbf{j}\in\mathcal{I}_{\mathcal{T}}} \kappa_{\mathbf{j}}\frac{\mathbb{P}\left(\mathbf{Z}\in E\cap R_{\mathbf{j}}\right)}{\alpha_{\mathbf{j}}} \\
&= \sum_{\mathbf{j}\in\mathcal{I}_{\mathcal{T}}} \frac{\kappa_{\mathbf{j}}}{\alpha_{\mathbf{j}}}\mathbb{P}\left(\mathbf{Z}\in E\cap R_{\mathbf{j}}\right),
\end{aligned}
$$

which proves the first equation.

Furthermore, we have that

$$\mathbb{P}\left(\mathbf{Z} \in E \cap R_{\mathbf{j}}\right) = \begin{cases} 0, & \text{if } E \cap R_{\mathbf{j}} = \emptyset \\ \alpha_{\mathbf{j}}, & \text{if } E \cap R_{\mathbf{j}} = R_{\mathbf{j}} \\ \mathbb{P}\left(\mathbf{Z} \in E \cap R_{\mathbf{j}}\right), & \text{otherwise.} \end{cases}$$

Consequently,

$$\begin{aligned} \mathbb{Q}(\mathbf{Z} \in E) &= \sum_{\mathbf{j} \in \mathcal{I}_{\mathcal{T}}} \frac{\kappa_{\mathbf{j}}}{\alpha_{\mathbf{j}}} \mathbb{P}\left(\mathbf{Z} \in E \cap R_{\mathbf{j}}\right) \\ &= \sum_{\mathbf{j} \in \mathcal{I}_{\mathcal{T}}^{0}} \frac{\kappa_{\mathbf{j}}}{\alpha_{\mathbf{j}}} \mathbb{P}\left(\mathbf{Z} \in E \cap R_{\mathbf{j}}\right) + \sum_{\mathbf{j} \in \mathcal{I}_{\mathcal{T}}^{R}} \frac{\kappa_{\mathbf{j}}}{\alpha_{\mathbf{j}}} \mathbb{P}\left(\mathbf{Z} \in E \cap R_{\mathbf{j}}\right) + \sum_{\mathbf{j} \in \mathcal{I}_{\mathcal{T}}^{E}} \frac{\kappa_{\mathbf{j}}}{\alpha_{\mathbf{j}}} \mathbb{P}\left(\mathbf{Z} \in E \cap R_{\mathbf{j}}\right) \\ &= \sum_{\mathbf{j} \in \mathcal{I}_{\mathcal{T}}^{0}} \frac{\kappa_{\mathbf{j}}}{\alpha_{\mathbf{j}}} 0 + \sum_{\mathbf{j} \in \mathcal{I}_{\mathcal{T}}^{R}} \frac{\kappa_{\mathbf{j}}}{\alpha_{\mathbf{j}}} \mathbb{P}\left(\mathbf{Z} \in R_{\mathbf{j}}\right) + \sum_{\mathbf{j} \in \mathcal{I}_{\mathcal{T}}^{E}} \frac{\kappa_{\mathbf{j}}}{\alpha_{\mathbf{j}}} \mathbb{P}\left(\mathbf{Z} \in E \cap R_{\mathbf{j}}\right) \\ &= \sum_{\mathbf{j} \in \mathcal{I}_{\mathcal{T}}^{R}} \frac{\kappa_{\mathbf{j}}}{\alpha_{\mathbf{j}}} \alpha_{\mathbf{j}} + \sum_{\mathbf{j} \in \mathcal{I}_{\mathcal{T}}^{E}} \frac{\kappa_{\mathbf{j}}}{\alpha_{\mathbf{j}}} \mathbb{P}\left(\mathbf{Z} \in E \cap R_{\mathbf{j}}\right) \\ &= \sum_{\mathbf{j} \in \mathcal{I}_{\mathcal{T}}^{R}} \kappa_{\mathbf{j}} + \sum_{\mathbf{j} \in \mathcal{I}_{\mathcal{T}}^{E}} \frac{\kappa_{\mathbf{j}}}{\alpha_{\mathbf{j}}} \mathbb{P}\left(\mathbf{Z} \in E \cap R_{\mathbf{j}}\right), \end{aligned}$$

which proves the second equation.

The main difference between the two equations is that the first considers all $\mathbf{j} \in \mathcal{I}_{\mathcal{T}}$ while the second considers only the non-zero terms of the sum.

□

The probability measure $\mathbb{Q}$ attained by the proposed change of measure function is one of possibly infinitely many in $\mathcal{Q}_{\mathcal{T}}$ and before motivating this particular selection we present some of its properties.

Corollary 3.1.3. *The probability measure $\mathbb{Q}$ is absolutely continuous with respect to $\mathbb{P}$.*

Proof. Let E be such that $\mathbb{P}\left(\mathbf{Z} \in E\right) = 0$. Then, since $\mathcal{R}_{\mathcal{T}}$ is a partition of the

domain of $\mathbf{Z}$

$$\forall \mathbf{i} \in \mathcal{I}_{\mathcal{T}} \quad \mathbb{P}\left(\mathbf{Z} \in E \cap R_{\mathbf{i}}\right) = 0.$$

And, using (3.1.2), we have that

$$\begin{aligned}
\mathbb{Q}(\mathbf{Z} \in E) &= \sum_{\mathbf{j} \in \mathcal{I}_{\mathcal{T}}} \frac{\kappa_{\mathbf{j}}}{\alpha_{\mathbf{j}}} \mathbb{P}\left(\mathbf{Z} \in E \cap R_{\mathbf{j}}\right) \\
&= \sum_{\mathbf{j} \in \mathcal{I}_{\mathcal{T}}} \frac{\kappa_{\mathbf{j}}}{\alpha_{\mathbf{j}}} \cdot 0 \\
&= 0.
\end{aligned}$$

□

$\mathbb{Q}$ being absolutely continuous with respect to $\mathbb{P}$ is an appealing property when it comes to real-world applications. When going from $\mathbb{P}$ to $\mathbb{Q}$, we want to change the probability of regions of the domain but it may be inappropriate or even unrealistic to include additional regions into the domain. Absolute continuity ensures that cannot happen. We claim that measures in $\mathcal{Q}_{\mathcal{T}}$ that are absolutely continuous with respect to $\mathbb{P}$ are particularly reasonable choices when it comes to deviating from $\mathbb{P}$, and so we restrict our attention to the set $\mathcal{Q}'_{\mathcal{T}}$, defined as follows

$$\mathcal{Q}'_{\mathcal{T}} = \mathcal{Q}_{\mathcal{T}} \cap \{\mathbb{P}^* \mid \mathbb{P}^* \ll \mathbb{P}\}.$$

Proposition 3.1.4. *As per Proposition 3.1.1 and Corollary 3.1.3,* $\mathbb{Q} \in \mathcal{Q}'_{\mathcal{T}}$.

Ensuring that $\mathbb{Q}$ does not expand the initial domain of $\mathbf{U}$ is not the only step we can make toward its reasonability. The Kullback-Leibler divergence can be used to determine how much two probability distributions differ from each other.

Proposition 3.1.5. *Among all elements of* $\mathcal{Q}'_{\mathcal{T}}$*, the probability distribution* $\mathbb{Q}$ *as obtained through the change of measure function* γ *defined in Proposition 3.1.1 minimizes the Kullback-Leibler divergence from* $\mathbb{P}$.

Proof. First, we have that

$$\begin{aligned}
D_{KL}\left(\mathbb{Q} \parallel \mathbb{P}\right) &= \mathrm{E}^{\mathbb{P}}\left[\gamma(\mathbf{U}) \log\left(\gamma(\mathbf{U})\right)\right] \\
&= \sum_{\mathbf{j} \in \mathcal{I}_{\mathcal{T}}} \mathrm{E}^{\mathbb{P}}\left[\gamma(\mathbf{U}) \log\left(\gamma(\mathbf{U})\right) \mid \mathbf{U} \in R_{\mathbf{j}}\right] \mathbb{P}\left(\mathbf{U} \in R_{\mathbf{j}}\right) \\
&= \sum_{\mathbf{j} \in \mathcal{I}_{\mathcal{T}}} \mathrm{E}^{\mathbb{P}}\left[\sum_{\mathbf{k} \in \mathcal{I}_{\mathcal{T}}} \frac{\kappa_{\mathbf{k}}}{\alpha_{\mathbf{k}}} \mathbf{1}_{\{\mathbf{U} \in R_{\mathbf{k}}\}} \log\left(\sum_{\mathbf{k} \in \mathcal{I}_{\mathcal{T}}} \frac{\kappa_{\mathbf{k}}}{\alpha_{\mathbf{k}}} \mathbf{1}_{\{\mathbf{U} \in R_{\mathbf{k}}\}}\right) \,\middle|\, \mathbf{U} \in R_{\mathbf{j}}\right] [\alpha_{\mathbf{j}}] \\
&= \sum_{\mathbf{j} \in \mathcal{I}_{\mathcal{T}}} \mathrm{E}^{\mathbb{P}}\left[\frac{\kappa_{\mathbf{j}}}{\alpha_{\mathbf{j}}} \log\left(\frac{\kappa_{\mathbf{j}}}{\alpha_{\mathbf{j}}}\right)\right] [\alpha_{\mathbf{j}}] \\
&= \sum_{\mathbf{j} \in \mathcal{I}_{\mathcal{T}}} \frac{\kappa_{\mathbf{j}}}{\alpha_{\mathbf{j}}} \log\left(\frac{\kappa_{\mathbf{j}}}{\alpha_{\mathbf{j}}}\right) [\alpha_{\mathbf{j}}] .
\end{aligned}$$

If $\mathbb{Q}$ is the only element of $\mathcal{Q}'_{\mathcal{T}}$, then the proof is complete.

If there is at least one other element in $\mathcal{Q}'_{\mathcal{T}}$, let $\mathbb{P}^*$ be any such element and $\xi^{\mathbb{P}^*}$ be the change of measure function allowing to go from $\mathbb{P}$ to $\mathbb{P}^*$. We know $\xi^{\mathbb{P}^*}$ exists due to the Radon-Nikodym Theorem 1.3.9. Then,

$$\begin{aligned}
D_{KL}\left(\mathbb{P}^* \parallel \mathbb{P}\right) &= \mathrm{E}^{\mathbb{P}}\left[\xi^{\mathbb{P}^*}(\mathbf{U}) \log\left(\xi^{\mathbb{P}^*}(\mathbf{U})\right)\right] \\
&= \sum_{\mathbf{j} \in \mathcal{I}_{\mathcal{T}}} \mathrm{E}^{\mathbb{P}}\left[\xi^{\mathbb{P}^*}(\mathbf{U}) \log\left(\xi^{\mathbb{P}^*}(\mathbf{U})\right) \mid \mathbf{U} \in R_{\mathbf{j}}\right] \mathbb{P}\left(\mathbf{U} \in R_{\mathbf{j}}\right) \\
&= \sum_{\mathbf{j} \in \mathcal{I}_{\mathcal{T}}} \mathrm{E}^{\mathbb{P}}\left[\xi^{\mathbb{P}^*}(\mathbf{U}) \log\left(\xi^{\mathbb{P}^*}(\mathbf{U})\right) \mid \mathbf{U} \in R_{\mathbf{j}}\right] [\alpha_{\mathbf{j}}] \\
&\geq \sum_{\mathbf{j} \in \mathcal{I}_{\mathcal{T}}} \mathrm{E}^{\mathbb{P}}\left[\xi^{\mathbb{P}^*}(\mathbf{U}) \mid \mathbf{U} \in R_{\mathbf{j}}\right] \cdot \log\left(\mathrm{E}^{\mathbb{P}}\left[\xi^{\mathbb{P}^*}(\mathbf{U}) \mid \mathbf{U} \in R_{\mathbf{j}}\right]\right) [\alpha_{\mathbf{j}}] .
\end{aligned}$$

The inequality is due to Jensen's inequality on the convex function $g(x) = x \log x$.

Next, we exploit the following property of $\mathbb{P}^*$ (which is shared with all elements

of $\mathcal{Q}'_{\mathcal{T}}$):

$$\begin{aligned}
\mathbb{P}^*\left(\mathbf{U} \in R_{\mathbf{j}}\right) &= \mathrm{E}^{\mathbb{P}^*}\left(\mathbf{1}_{\mathbf{U}\in R_{\mathbf{j}}}\right) \\
&= \mathrm{E}^{\mathbb{P}}\left(\xi^{\mathbb{P}^*}(\mathbf{U})\,\mathbf{1}_{\{\mathbf{U}\in R_{\mathbf{j}}\}}\right) \\
&= \sum_{\mathbf{k}\in\mathcal{I}_{\mathcal{T}}} \mathrm{E}^{\mathbb{P}}\left(\xi^{\mathbb{P}^*}(\mathbf{U})\,\mathbf{1}_{\{\mathbf{U}\in R_{\mathbf{j}}\}} \;\middle|\; \mathbf{U}\in R_{\mathbf{k}}\right)\mathbb{P}\left(\mathbf{U}\in R_{\mathbf{k}}\right) \\
&= \mathrm{E}^{\mathbb{P}}\left(\xi^{\mathbb{P}^*}(\mathbf{U})\,\mathbf{1}_{\{\mathbf{U}\in R_{\mathbf{j}}\}} \;\middle|\; \mathbf{U}\in R_{\mathbf{j}}\right)\mathbb{P}\left(\mathbf{U}\in R_{\mathbf{j}}\right),
\end{aligned}$$

which leads to

$$\begin{aligned}
\frac{\mathbb{P}^*\left(\mathbf{U}\in R_{\mathbf{j}}\right)}{\mathbb{P}\left(\mathbf{U}\in R_{\mathbf{j}}\right)} &= \mathrm{E}^{\mathbb{P}}\left(\xi^{\mathbb{P}^*}(\mathbf{U})\,\mathbf{1}_{\{\mathbf{U}\in R_{\mathbf{j}}\}} \;\middle|\; \mathbf{U}\in R_{\mathbf{j}}\right) \\
\frac{\mathbb{P}^*\left(\mathbf{U}\in R_{\mathbf{j}}\right)}{\mathbb{P}\left(\mathbf{U}\in R_{\mathbf{j}}\right)} &= \mathrm{E}^{\mathbb{P}}\left(\xi^{\mathbb{P}^*}(\mathbf{U}) \;\middle|\; \mathbf{U}\in R_{\mathbf{j}}\right).
\end{aligned}$$

By definition $\mathbb{P}\left(\mathbf{U}\in R_{\mathbf{j}}\right) = \alpha_{\mathbf{j}}$ and $\mathbb{P}^*\left(\mathbf{U}\in R_{\mathbf{j}}\right) = \kappa_{\mathbf{j}}$, hence

$$\mathrm{E}^{\mathbb{P}}\left(\xi^{\mathbb{P}^*}(\mathbf{U}) \;\middle|\; \mathbf{U}\in R_{\mathbf{j}}\right) = \frac{\kappa_{\mathbf{j}}}{\alpha_{\mathbf{j}}},$$

which we plug in the earlier inequality to obtain that

$$D_{KL}\left(\mathbb{P}^* \parallel \mathbb{P}\right) \geq \sum_{\mathbf{j}\in\mathcal{I}_{\mathcal{T}}} \frac{\kappa_{\mathbf{j}}}{\alpha_{\mathbf{j}}} \cdot \log\left(\frac{\kappa_{\mathbf{j}}}{\alpha_{\mathbf{j}}}\right)[\alpha_{\mathbf{j}}] = D_{KL}\left(\mathbb{Q} \parallel \mathbb{P}\right).$$

□

We have shown that $\mathbb{Q}$ is absolutely continuous with respect to $\mathbb{P}$ and that, among all absolutely continuous measures that satisfy our required probabilities $\kappa_{\mathbf{i}}$, it is the one that minimizes KL divergence.

3.2 Illustrative Example

We present here a toy example to familiarize the reader with the change of measure and the resulting probability measure.

Example 3.2.1. *We begin with a simple example in one dimension and with two splits of the domain to illustrate the change of measure. Consider* $\mathbf{X} = X_1 \sim \text{Unif}(0,1)$ *with* $\mathcal{T} = \{\mathbf{t}_1\}$ *and* $\mathbf{t}_1 = (0, 0.4, 0.9, 1)$*. Then,*

- *The number of dimensions is* $D = 1$
- *The number of chosen splits for variable* X_1 *is* $S_1 = 2$ *and we have effectively split the domain* $[0,1]$ *in* $(S_1 + 1) = 3$ *intervals:*
  - 0^{th} *interval:* $[0, 0.4]$
  - 1^{st} *interval:* $(0.4, 0.9]$
  - 2^{nd} *interval:* $(0.9, 1]$
- *The set of indices is* $\mathcal{I}_\mathcal{T} = \{0, 1, 2\}$ *and the set of regions is* $\mathcal{R}_\mathcal{T} = \{R_0, R_1, R_2\} = \{[0, 0.4], (0.4, 0.9], (0.9, 1]\}$

Since X_1 *is a standard uniform random variable under the baseline probability measure* $\mathbb{P}$*, we have that*

- $\mathbb{P}(U_1 \in R_0) = \mathbb{P}(U_1 \in [0, 0.4]) = 0.4 = \alpha_0$
- $\mathbb{P}(U_1 \in R_1) = \mathbb{P}(U_1 \in (0.4, 0.9]) = 0.5 = \alpha_1$
- $\mathbb{P}(U_1 \in R_2) = \mathbb{P}(U_1 \in (0.9, 1]) = 0.1 = \alpha_2$

Now, say we require a distribution $\mathbb{Q}$ *for which*

- $\mathbb{Q}(U_1 \in R_0) = \mathbb{Q}(U_1 \in [0, 0.4]) = 0.2 = \kappa_0$
- $\mathbb{Q}(U_1 \in R_1) = \mathbb{Q}(U_1 \in (0.4, 0.9]) = 0.45 = \kappa_1$
- $\mathbb{Q}(U_1 \in R_2) = \mathbb{Q}(U_1 \in (0.9, 1]) = 0.35 = \kappa_2$

Then, through Proposition 3.1.1, we can use the following change of measure function to reach $\mathbb{Q}$ *from* $\mathbb{P}$*:*

$$\gamma(U_1) = \sum_{\mathbf{i} \in \mathcal{I}_\mathcal{T}} \frac{\kappa_\mathbf{i}}{\alpha_\mathbf{i}} \mathbf{1}_{\{U_1 \in R_\mathbf{i}\}} = \frac{0.2}{0.4} \mathbf{1}_{\{U_1 \in R_0\}} + \frac{0.45}{0.5} \mathbf{1}_{\{U_1 \in R_1\}} + \frac{0.35}{0.1} \mathbf{1}_{\{U_1 \in R_2\}}.$$

So far, we have looked at the change of measure through probability measures. We use the present example to illustrate the idea through the perspective of cumulative distribution functions (CDFs). Let F_{U_1} be the CDF of U_1 under $\mathbb{P}$. Then, $F_{U_1}(u) = \mathbb{P}(U_1 \in [0,u]) = u$ for any $u \in [0,1]$. Also, let G_{U_1} be the CDF of U_1 under $\mathbb{Q}$ such that $G_{U_1}(u) = \mathbb{Q}(U_1 \in [0,u])$. Then, using (3.1.2), we have

$$
\begin{aligned}
G_{U_1}(u) &= \frac{\kappa_0}{\alpha_0}\mathbb{P}\left(U_1 \in [0,u] \cap R_0\right) + \frac{\kappa_1}{\alpha_1}\mathbb{P}\left(U_1 \in [0,u] \cap R_1\right) + \frac{\kappa_2}{\alpha_2}\mathbb{P}\left(U_1 \in [0,u] \cap R_2\right) \\
&= \frac{0.2}{0.4}\mathbb{P}\left(U_1 \in [0,u] \cap [0,0.4]\right) + \frac{0.45}{0.5}\mathbb{P}\left(U_1 \in [0,u] \cap (0.4,0.9]\right) \\
&+ \frac{0.35}{0.1}\mathbb{P}\left(U_1 \in [0,u] \cap (0.9,1]\right) \\
&= \begin{cases} \frac{0.2}{0.4}\mathbb{P}\left(U_1 \in [0,u]\right) + \frac{0.45}{0.5}\mathbb{P}\left(U_1 \in \emptyset\right) + \frac{0.35}{0.1}\mathbb{P}\left(U_1 \in \emptyset\right) & , \textit{if } u \leq 0.4 \\ \frac{0.2}{0.4}\mathbb{P}\left(U_1 \in [0,0.4]\right) + \frac{0.45}{0.5}\mathbb{P}\left(U_1 \in (0.4,u]\right) + \frac{0.35}{0.1}\mathbb{P}\left(U_1 \in \emptyset\right) & , \textit{if } 0.4 < u \leq 0.9 \\ \frac{0.2}{0.4}\mathbb{P}\left(U_1 \in [0,0.4]\right) + \frac{0.45}{0.5}\mathbb{P}\left(U_1 \in (0.4,0.9]\right) + \frac{0.35}{0.1}\mathbb{P}\left(U_1 \in (0.9,u]\right) & , \textit{otherwise} \end{cases} \\
&= \begin{cases} \frac{0.2}{0.4} \cdot u + \frac{0.45}{0.5} \cdot 0 + \frac{0.35}{0.1} \cdot 0 & , \textit{if } u \leq 0.4 \\ \frac{0.2}{0.4} \cdot 0.4 + \frac{0.45}{0.5}\left(u - 0.4\right) + \frac{0.35}{0.1} \cdot 0 & , \textit{if } 0.4 < u \leq 0.9 \\ \frac{0.2}{0.4} \cdot 0.4 + \frac{0.45}{0.5} \cdot 0.5 + \frac{0.35}{0.1}\left(u - 0.9\right) & , \textit{otherwise} \end{cases} \\
&= \begin{cases} \frac{0.2}{0.4} \cdot u & , \textit{if } u \leq 0.4 \\ 0.2 + \frac{0.45}{0.5}\left(u - 0.4\right) & , \textit{if } 0.4 < u \leq 0.9 \\ 0.65 + \frac{0.35}{0.1}\left(u - 0.9\right) & , \textit{otherwise.} \end{cases}
\end{aligned}
$$

The reader is invited to compare this final expression to (3.1.3) for each of the cases on the value of u.

Chapter 4

The Inverted Premium

In this chapter, we introduce the *inverted premium*, the result of our correction method for fairness that can be applied on any kind of data, but particularly on a continuous premium with a binary protected variable.

4.1 Framework

Consider N individuals who wish to be insured. For each of these individuals, a continuous premium $Y > 0$ is calculated by applying a model function f to explanatory variables $\mathbf{X}$, *i.e.* $Y = f(\mathbf{X})$. Now, suppose we observe one categorical protected variable $P \in \{0, 1, 2, \ldots, J\}$ for each individual. Since no discrimination should be made with respect to P, we would hope that $Y \perp\!\!\!\perp P$. However, this may not be the case for various reasons. In particular, if there is dependence between P and one or more of the explanatory variables, it is likely that there will be dependence between P and Y as well, causing them to appear dependent to some degree. No matter the reason behind the observed dependence between Y and P, we may want to attenuate it.

In the following, we assume:

1. For N insured, we observe only the premium Y and the protected variable

P;

2. N is large enough that it is reasonable to use the empirical distribution $\mathbb{P}$ for the observed data.

Let $\mathbf{Z} = (Y, P)$ and $\mathbb{P}$ be the empirical distribution on the observed premiums and protected variable. Suppose a set of split-vectors $\mathcal{T} = (\mathbf{t}_1, \mathbf{t}_2) = (\mathbf{t}_Y, \mathbf{t}_P)$ is selected with the numbers of chosen splits for each variable being $S_1 = S_Y = I$ and $S_2 = S_P = J$[1]. Then,

- The interval $H_{Y,i}$ is the i^{th} interval for Y
- The interval $H_{P,j}$ is the j^{th} interval for P
- The set of indices $\mathcal{I}_\mathcal{T}$ is $\{0, 1, \ldots, I\} \times \{0, 1, \ldots, J\}$
- The set of regions is

$$\mathcal{R}_\mathcal{T} = \{H_{Y,i} \times H_{P,j} \mid (i, j) \in \mathcal{I}_\mathcal{T}\}.$$

For P, we suppose that each chosen element of $\mathbf{t}_P$ corresponds to a value between each category of P, such that the intervals will each contain exactly one of those categories. This means we have $\mathbf{t}_P = \{t_{P,0}, t_{P,1}, \ldots, t_{P,J}, t_{P,J+1}\} = \{0, 0.5, 1.5, \ldots, J - 1.5, J - 0.5, J\}$, resulting in:

- $H_{P,0} = [0, 0.5] \implies \mathbb{P}(P \in H_{P,0}) = \mathbb{P}(P = 0)$,
- $H_{P,j} = (j - 0.5, j + 0.5] \implies \mathbb{P}(P \in H_{P,j}) = \mathbb{P}(P = j)$ for $j = 1, \ldots, J - 1$ and
- $H_{P,J} = (J - 0.5, J] \implies \mathbb{P}(P \in H_{P,J}) = \mathbb{P}(P = J)$

for a total of $1 + (J - 1) + 1 = J + 1$ intervals, numbered from 0 to J.

Remark 4.1.1. *Grouping of categories of P is possible. For instance, instead of $\mathbf{t}_P = \{0, 0.5, 1.5, \ldots, J - 1.5, J - 0.5, J\}$, an option is $\mathbf{t}_P^* = \{0, 1.5, 2.5, \ldots, J -$*

[1]For convenience to the reader, we change indices from 1 and 2 to Y and P in this chapter to make the association with the corresponding variables clear.

$1.5, J-0.5, J\}$. *This would result in a total of* J *intervals instead of* $J+1$. *The* 0^{th} *interval would become* $H^*_{P,0} = H_{P,0} \cup H_{P,1} = [0, 1.5]$ *and result in* $\mathbb{P}(P \in H^*_{P,0}) = \mathbb{P}(P \in \{0, 1\})$. *For the other intervals, there would be a shift of indices such that* $H^*_{P,j} = H_{P,j+1}$ *for* $j \in \{1, \ldots, J-1\}$.

4.1.1 Correction Test

To assess the observed dependence between Y and P, conditional probabilities can be used. In general, we can calculate the probability of Y to lie in its i^{th} interval conditionally on P lying in its j^{th} interval as follows. Denote:

$$\begin{aligned}\alpha_{i|j} &= \mathbb{P}\left(Y \in H_{Y,i} \mid P \in H_{P,j}\right) \\ &= \frac{\mathbb{P}\left(Y \in H_{Y,i}, P \in H_{P,j}\right)}{\mathbb{P}\left(P \in H_{P,j}\right)}.\end{aligned}$$

This quantity can be expressed in terms of the $\alpha_{(i,j)}$ only:

$$\begin{aligned}\alpha_{i|j} &= \frac{\mathbb{P}\left(Y \in H_{Y,i}, P \in H_{P,j}\right)}{\mathbb{P}\left(P \in H_{P,j}\right)} \\ &= \frac{\mathbb{P}\left(Y \in H_{Y,i}, P \in H_{P,j}\right)}{\sum_{i=0}^{I} \mathbb{P}\left(Y \in H_{Y,i}, P \in H_{P,j}\right)} \\ &= \frac{\mathbb{P}\left(\mathbf{Z} \in R_{(i,j)}\right)}{\sum_{i=0}^{I} \mathbb{P}\left(\mathbf{Z} \in R_{(i,j)}\right)} \\ &= \frac{\alpha_{(i,j)}}{\sum_{i=0}^{I} \alpha_{(i,j)}} \\ &= \frac{\alpha_{(i,j)}}{\alpha_{\bullet,j}},\end{aligned}$$

where we introduce the notation $\alpha_{\bullet,j} = \sum_{i=0}^{I} \alpha_{(i,j)}$. We also denote $\alpha_{i,\bullet} = \sum_{j=0}^{J} \alpha_{(i,j)}$.

If Y and P are independent, then the following must hold:

$$\forall i \in \{0, 1, \ldots, I\} \quad \mathbb{P}\left(Y \in H_{Y,i} \mid P \in H_{P,0}\right) = \cdots = \mathbb{P}\left(Y \in H_{Y,i} \mid P \in H_{P,J}\right),$$

or, equivalently

$$\forall i \in \{0, 1, \ldots, I\} \quad \alpha_{i|0} = \cdots = \alpha_{i|J}. \tag{4.1.1}$$

This condition means that the conditional probabilities of Y to lie in a given interval is the same no matter which category of P is being conditioned on. As an example, if we only consider the 0^{th} interval of Y in the 0^{th} and the 1^{st} categories of P, $\alpha_{0|0} = \alpha_{0|1}$ means that Y is as likely to lie in $H_{Y,0}$ conditionally on P being in $H_{P,0}$ than it is to lie in $H_{Y,0}$ conditionally on P being in $H_{P,1}$. Along with (4.1.1), this means the condition is respected across all intervals (categories) of P for all $i \in \{0, 1, \ldots, I\}$. Note however that although (4.1.1) must hold, it is not necessarily true that $\alpha_{i_1|j} = \alpha_{i_2|j}$ for $(i_1, j), (i_2, j) \in \mathcal{I}_{\mathcal{T}}$, *i.e.* it may not, and usually will not, be the case that Y is as likely to lie in any of its intervals conditionally on P being in its j^{th} category.

Because the conditional probabilities of (4.1.1) must be equal under independence of Y and P, how different they actually are from each other can give a sense of whether there is a need for correction in the first place. For each of the intervals of Y (for each $i \in \{0, 1, \ldots, I\}$), we measure:

$$\Delta_i^{\mathbb{P}} = \max_{0 \le j \le J} \alpha_{i|j} - \min_{0 \le j \le J} \alpha_{i|j}.$$

$\Delta_i^{\mathbb{P}}$ represents the greatest absolute difference in the conditional probabilities of Y being in $H_{Y,i}$ across the categories of P. If $\Delta_i^{\mathbb{P}}$ is very small, then the conditional probabilities $\alpha_{i|j}$ are all very close to each other, providing some comfort that Y and P are somewhat independent. On the other hand, if $\Delta_i^{\mathbb{P}}$ is very large, then there are at least two categories of P for which the conditional probabilities of Y lying in $H_{Y,i}$ are far apart, meaning that there may be some dependence between Y and P. $\Delta_i^{\mathbb{P}}$ is inspired by statistical parity (see (1.1.1)) and we demonstrate its calculation in Example 4.1.2.

Example 4.1.2 (Calculating $\Delta_i^{\mathbb{P}}$)**.** *Suppose* $\mathbf{Z} = (Y, P)$ *is such that* $Y \in [0, 10]$ *and* $P \in \{0, 1, 2\}$*. Say* $\mathbf{t}_1 = (0, 3, 8, 10)$ *and* $\mathbf{t}_2 = (0, 0.5, 1.5, 2)$*, with probabilities*

presented in Table 4.1:

$\mathbb{P}(\cdot,\cdot)$	$P \in [0, 0.5]$	$P \in (0.5, 1.5]$	$P \in (1.5, 2]$	$\mathbf{P \in [0, 2]}$
$Y \in [0, 3]$	0.05	0.6	0.06	**0.71**
$Y \in (3, 8]$	0.07	0.07	0.03	**0.17**
$Y \in (8, 10]$	0.08	0.03	0.01	**0.12**
$\mathbf{Y \in [0, 10]}$	**0.2**	**0.7**	**0.1**	**1**

Table 4.1

Then, we can calculate conditional probabilities. For example,

$$\begin{aligned}
\alpha_{0|2} &= \mathbb{P}(Y \in H_{Y,0} \mid P \in H_{P,2}) \\
&= \mathbb{P}(Y \in [0, 5] \mid P \in (1.5, 2]) \\
&= \mathbb{P}(Y \in [0, 5] \mid P = 2) \\
&= \frac{\mathbb{P}(Y \in [0, 5], P = 2)}{\mathbb{P}(P = 2)} \\
&= \frac{0.08}{0.1} \\
&= 0.8.
\end{aligned}$$

Proceeding similarly for other combinations of $H_{Y,i}$ and $H_{P,j}$ gives Table 4.2:

$\alpha_{i\|j}$	$j = 0$	$j = 1$	$j = 2$
$i = 0$	0.25	0.86	0.6
$i = 1$	0.35	0.1	0.3
$i = 2$	0.4	0.04	0.1
Total	**1**	**1**	**1**

Table 4.2: Conditional probabilities

Remark 4.1.3. Observe that the sum of each column of Table 4.2 is 1. This is because, conditionally on P being any of its categories, Y *will* lie somewhere in $\bigcup_{i=0}^{I} H_{Y,i} = \mathrm{Dom}(Y)$. Mathematically, we have that $\mathbb{P}(Y \in \mathrm{Dom}(Y) \mid P \in E) = 1$, where E is any set including at least one element of $\mathrm{Dom}(P)$.

Using these conditional probabilities, we can calculate the $\Delta_i^{\mathbb{P}}$ *as:*

$$\begin{aligned}\Delta_0^{\mathbb{P}} &= \max_{0\le j\le J} \alpha_{0|j} - \min_{0\le j\le J} \alpha_{0|j} \\ &= \max\{0.25, 0.86, 0.6\} - \min\{0.25, 0.86, 0.6\} \\ &= 0.86 - 0.25 \\ &= 0.61.\end{aligned}$$

$$\begin{aligned}\Delta_1^{\mathbb{P}} &= \max_{0\le j\le J} \alpha_{1|j} - \min_{0\le j\le J} \alpha_{1|j} \\ &= \max\{0.35, 0.1, 0.3\} - \min\{0.35, 0.1, 0.3\} \\ &= 0.35 - 0.1 \\ &= 0.25.\end{aligned}$$

$$\begin{aligned}\Delta_2^{\mathbb{P}} &= \max_{0\le j\le J} \alpha_{2|j} - \min_{0\le j\le J} \alpha_{2|j} \\ &= \max\{0.4, 0.04, 0.1\} - \min\{0.4, 0.04, 0.1\} \\ &= 0.4 - 0.04 \\ &= 0.36.\end{aligned}$$

Proposition 4.1.4 considers a special case for $\Delta_i^{\mathbb{P}}$.

Proposition 4.1.4. *When* $S_Y = 1$*, we have that* $\Delta_0^{\mathbb{P}} = \Delta_1^{\mathbb{P}}$*, for* any *probability measure* $\mathbb{P}$*.*

Proof. Without loss of generality, let j_2 be the j for which $\alpha_{0|j}$ is maximized and j_1 be the j for which $\alpha_{0|j}$ is minimized. Then, because $\alpha_{0|j} = 1 - \alpha_{1|j}$ (due to Y only having two intervals), we have that j_2 is the j for which $\alpha_{1|j}$ is minimized and j_1 is the j for which $\alpha_{1|j}$ is maximized. As such, we have the following:

$$\Delta_0^{\mathbb{P}} = \alpha_{0|j_2} - \alpha_{0|j_1} \quad \text{and} \quad \Delta_1^{\mathbb{P}} = \alpha_{1|j_1} - \alpha_{1|j_2}.$$

Therefore,

$$\begin{aligned}\Delta_0^{\mathbb{P}} &= \alpha_{0|j_2} - \alpha_{0|j_1} \\ &= (1 - \alpha_{1|j_2}) - (1 - \alpha_{1|j_1}) \\ &= \alpha_{1|j_1} - \alpha_{1|j_2} \\ &= \Delta_1^{\mathbb{P}}.\end{aligned}$$

□

When $S_Y > 1$, there may be some intervals of Y for which $\Delta_i^{\mathbb{P}}$ is much greater than others. These intervals would be seemed problematic. Proposition 4.1.4 implies that when there are only two intervals of Y, they are both equally problematic.

After having calculated the $\Delta_i^{\mathbb{P}}$, a simple threshold test, which we dub the Δ-*test*, is carried out to determine whether a correction of premiums should be applied:

$$\exists \Delta_i^{\mathbb{P}} \quad \text{such that} \quad \Delta_i^{\mathbb{P}} > \epsilon \implies \text{Apply correction}, \tag{4.1.2}$$

for some $\epsilon > 0$. In other words, we apply the correction if there is any interval of Y in which "a lot" more individuals of some category of P lie than individuals of another category of P, where the measure of "a lot" is controlled by ϵ.

4.1.2 Correction

Finding K^*

If it is determined that the correction should be applied, our goal is to find a measure $\mathbb{Q}$ such that $\Delta_i^{\mathbb{Q}} < \epsilon$ for $i \in \{0, 1, \ldots, I\}$. Let $A = \{\alpha_{(i,j)} \mid (i,j) \in \mathcal{I}_{\mathcal{T}}\}$. Our first key step is two-fold:

1. Find a set $K^* = \{\kappa^*_{(i,j)} \mid (i,j) \in \mathcal{I}_{\mathcal{T}}\}$ such that

$$\forall i \in \{0, 1, \ldots, I\} \; \forall j_1, j_2 \in \{0, 1, \ldots, J\} \quad |\kappa^*_{i|j_2} - \kappa^*_{i|j_1}| < \epsilon.$$

2. Using Propositions 3.1.1 and 3.1.2, apply a change of measure from $\mathbb{P}$ to $\mathbb{Q}^*$ with $\mathbb{Q}^*$ such that

$$\forall (i,j) \in \mathcal{I}_{\mathcal{T}} \quad \mathbb{Q}^* \left(\mathbf{Z} \in R_{(i,j)} \right) = \kappa^*_{(i,j)}.$$

Figure 4.1 illustrates the segmentation of $\text{Dom}(\mathbf{Z})$ with respect to some general $K = \{\kappa_{(i,j)} \mid (i,j) \in \mathcal{I}_{\mathcal{T}}\}$, and it will be referred to as the *K-grid*.

To simplify the search for K^*, suppose that, instead of $|\kappa^*_{i|j_2} - \kappa^*_{i|j_1}| < \epsilon$, we impose $|\kappa^*_{i|j_2} - \kappa^*_{i|j_1}| = 0$. This forces the following condition on K^*:

$$\forall j_1, j_2 \in \{0, 1, \ldots, J\} \quad (\kappa^*_{(0,j_1)}, \kappa^*_{(1,j_1)}, \ldots, \kappa^*_{(I,j_1)}) = c \cdot (\kappa^*_{(0,j_2)}, \kappa^*_{(1,j_2)}, \ldots, \kappa^*_{(I,j_2)}). \tag{4.1.3}$$

for some $c > 0$. In terms of the K-grid of Figure 4.1, this means that all rows of the grid will be multiples of each other, allowing us to select a *starting vector* $\mathbf{v} = (v_0, v_1, \ldots, v_I)$ of which all rows will be multiples. The only restriction on $\mathbf{v}$ is that the sum of its elements must be 1. This is not a very constraining condition however, because any vector of length $I + 1$ that is not the zero-vector can be normalized to have a sum of 1. This starting vector is then allocated to each row in the following way:

$$\forall j \in \{0, 1, \ldots, J\} \quad (\kappa^*_{(0,j)}, \kappa^*_{(1,j)}, \ldots, \kappa^*_{(I,j)}) = \alpha_{\bullet,j} \cdot \mathbf{v}.$$

Because all rows are multiples of each other, behavior of Y will be similar across all categories of P. Furthermore, allocating $\mathbf{v}$ in this way makes it so that the row sums of the K-grid are the same as the row sums of the A-grid, due to the starting vector $\mathbf{v}$ having a sum of 1. This means that marginal probabilities of P will not be affected by the change of measure, which is desirable. Indeed, while we could want to tune the marginal distribution of the premiums Y, it may not make sense to disturb the marginal distribution of the protected variable, as it typically does not change in individuals.

Remark 4.1.5. *Due to its categorical nature, P will see no effect to its marginal distribution caused by the change of measure. However, because Y is continuous, its marginal distribution will be affected, except at locations of splits. Recall that $Y = f(\mathbf{X})$, such that a change to the marginal distribution of Y also results in changes to the distribution of* $\mathbf{X}$, even though $\mathbf{X}$ is not considered at all in the procedure. *If impacts to some of the marginal distributions of* $\mathbf{X}$ *also need to be minimized, then those variables should be brought into the analysis and additional restrictions should be stated with respect to their probabilities under $\mathbb{Q}^*$, resulting in both a greater number of dimensions and a more complex construction of K^*.*

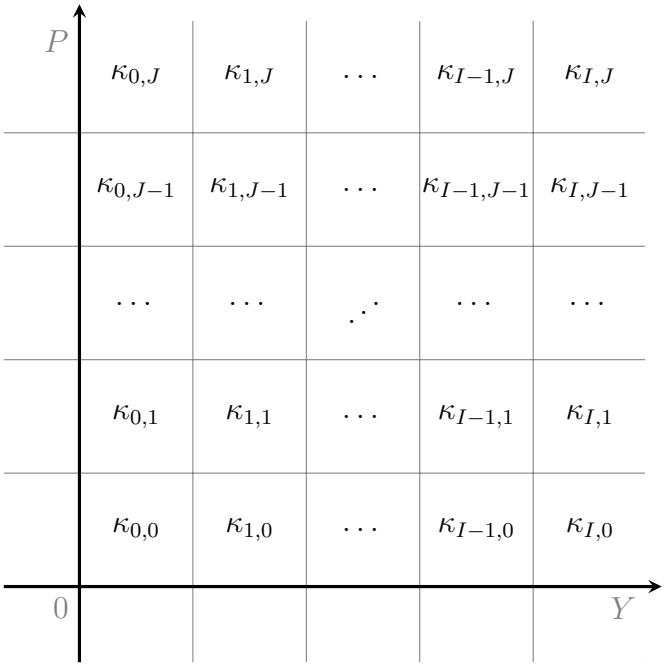

Figure 4.1: Graphical representation of the K-grid for a general K. The $\kappa_{(i,j)}$ represent the probability of the random vector $\mathbf{Z}$ of lying in the corresponding region $R_{(i,j)}$ for $(i,j) \in \mathcal{I}_{\mathcal{T}}$ under the measure $\mathbb{Q}$ obtained from K.

While there are infinitely many possibilities for $\mathbf{v}$, the most natural one is

$$\dot{\mathbf{v}} = (\alpha_{0,\bullet}, \alpha_{1,\bullet}, \ldots, \alpha_{I,\bullet}) = (\mathbb{P}(Y \in H_{Y,0}), \mathbb{P}(Y \in H_{Y,1}), \ldots, \mathbb{P}(Y \in H_{Y,I})).$$

It is trivial to validate the the sum of the elements of $\dot{\mathbf{v}}$ is 1. This choice of vector will make it so the marginal behavior of Y is spread across all categories of P.

This is equivalent to the Product-of-Marginals method from Section 2.1. Indeed, this allocation results in $\kappa_{(i,j)} = \alpha_{i,\bullet} \cdot \alpha_{\bullet,j}$, leading to no information on Y being obtainable from P.

Remark 4.1.6. *A nuance is made here that we are dealing with the continuous variable Y at discrete splits $\mathbf{t}_Y$. When $y \in \mathbf{t}_Y$, the conditional probabilities $\mathbb{P}(Y \leq y \mid P = j)$ may all be equal for $j \in \{0, 1, \ldots, J\}$, but that generally will not be the case when $y \notin \mathbf{t}_Y$ because K imposes no restrictions at those values. The differences will be smaller than they were under $\mathbb{P}$, but they may not be exactly zero.*

Choosing $\mathbf{v} = \dot{\mathbf{v}}$ leads to the following properties of the K^*-grid:

$$\forall i \in \{0, 1, \ldots, I\} \; \kappa^*_{i,\bullet} = \alpha_{i,\bullet}, \tag{4.1.4}$$

$$\forall j \in \{0, 1, \ldots, J\} \; \kappa^*_{\bullet,j} = \alpha_{\bullet,j}. \tag{4.1.5}$$

In other words, the sums of the columns of the K^*-grid are the same as that of the A-grid and the sums of the rows of the K^*-grid are the same that of the A-grid. This is an ideal property, as it implies a minimal impact to the marginal distributions of both Y and P. Indeed, recall that

- $\alpha_{i,\bullet} = \mathbb{P}(Y \in H_{Y,i})$,
- $\alpha_{\bullet,j} = \mathbb{P}(Y \in H_{P,j})$,
- $\kappa^*_{i,\bullet} = \mathbb{Q}^*(Y \in H_{Y,i})$,
- $\kappa^*_{\bullet,j} = \mathbb{Q}^*(Y \in H_{P,j})$.

Generalization to Multivariate P

In the previous section, the construction of K^* is made with respect to a single protected categorical variable P with $J+1$ categories $0, 1, \ldots, J$. This simplistic assumption conveniently avoids dimensionality issues, which we explore in this section.

We generalize the previous section's methodology to a multivariate vector of protected variables $\mathbf{P} = (P_1, P_2, \ldots, P_B)$ with $B > 1$, each having J_b+1 categories $0, 1, \ldots, J_b$ for $b = 1, 2, \ldots, B$. There are $D = B + 1$ variables in total (the premium Y and the B protected variables).

The goal remains the same, *i.e.* find a measure $\mathbb{Q}^*$ such that $\Delta^{\mathbb{Q}^*} < \epsilon$. As in the previous section, we aim for $\Delta^{\mathbb{Q}^*} = 0$. When $\mathbf{P}$ is multivariate, the condition is more complex and some additional notation is required. Let $\mathcal{I}_{\mathcal{T}}^{\mathbf{P}}$ be the set of (non-random) B-tuples with b^{th} element in $\{0, \ldots, J_b\}$ for $b = 1, \ldots, B$. Correspondingly, let $\mathcal{I}_{\mathcal{T}}^{Y} = \{0, 1, \ldots, I\}$. Denote $H_{\mathbf{P},\mathbf{m}} = H_{P_1,j_1} \times \cdots \times H_{P_B,j_B}$ with $\mathbf{m} = (j_1, \ldots, j_B) \in \mathcal{I}_{\mathcal{T}}^{\mathbf{P}}$. Then, we must find a $\mathbb{Q}^*$ such that

$$\forall i \in \mathcal{I}_{\mathcal{T}}^{Y} \; \forall \mathbf{m}_1, \mathbf{m}_2 \in \mathcal{I}_{\mathcal{T}}^{\mathbf{P}} \quad \mathbb{Q}^*(Y \in H_{Y,i} \mid \mathbf{P} \in H_{\mathbf{P},\mathbf{m}_1}) = \mathbb{Q}^*(Y \in H_{Y,i} \mid \mathbf{P} \in H_{\mathbf{P},\mathbf{m}_2}).$$

The condition is simply a generalization of (4.1.1), stating that, the probability of Y lying in one of its intervals remains the same no matter the given state of the protected variables $\mathbf{P}$.

To use κ^* notation, let $\mathbf{m} = (m_1, \ldots, m_B) \in \mathcal{I}_{\mathcal{T}}^{\mathbf{P}}$, $\mathbf{n} = (n_1, \ldots, n_B) \in \mathcal{I}_{\mathcal{T}}^{\mathbf{P}}$ and ${\kappa^*}_{i,\mathbf{m}} = {\kappa^*}_{i,m_1,\ldots,m_B}$. Then, we must have

$$\forall i \in \mathcal{I}_{\mathcal{T}}^{Y} \; \forall \mathbf{m}, \mathbf{n} \in \mathcal{I}_{\mathcal{T}}^{\mathbf{P}} \quad \frac{{\kappa^*}_{i,\mathbf{m}}}{{\kappa^*}_{\bullet,\mathbf{m}}} = \frac{{\kappa^*}_{i,\mathbf{n}}}{{\kappa^*}_{\bullet,\mathbf{n}}}.$$

Again, enforcing such a condition on K^* imposes a strong relationship between groups of its elements. In particular, (4.1.3) generalizes to

$$\forall \mathbf{m}, \mathbf{n} \in \mathcal{I}_{\mathcal{T}}^{\mathbf{P}} \quad (\kappa^*_{0,\mathbf{m}}, \kappa^*_{1,\mathbf{m}}, \ldots, \kappa^*_{I,\mathbf{m}}) = c \cdot (\kappa^*_{0,\mathbf{n}}, \kappa^*_{1,\mathbf{n}}, \ldots, \kappa^*_{I,\mathbf{n}})$$

for some $c > 0$. So, in the case of multivariate $\mathbf{P}$, we can also simplify the search for K^* to choosing an appropriate starting vector $\mathbf{v}$ and allocating it to all dimensions of $\mathbf{P}$ adequately. In line with the approach already presented for

univariate P, a recommendation is the vector of marginal probabilities of Y:

$$\dot{\mathbf{v}} = (\alpha_{0,\bullet}, \alpha_{1,\bullet}, \ldots, \alpha_{I,\bullet}) = (\mathbb{P}(Y \in H_{Y,0}), \mathbb{P}(Y \in H_{Y,1}), \ldots, \mathbb{P}(Y \in H_{Y,I})),$$

where here $\alpha_{i,\bullet}$ is simplified notation for $\alpha_{i,\bullet,\ldots,\bullet}$.

The allocation is made with respect to each possible state of $\mathbf{P}$, such that the κ^* are given by

$$\begin{aligned} \forall i \in \mathcal{I}_{\mathcal{T}}^{Y} \ \forall \mathbf{m} \in \mathcal{I}_{\mathcal{T}}^{\mathbf{P}} \quad {\kappa^*}_{i,\mathbf{m}} &= \alpha_{i,\bullet} \cdot \alpha_{\bullet,\mathbf{m}} \\ &= \mathbb{P}(Y \in H_{Y,i}) \cdot \mathbb{P}(\mathbf{P} \in H_{\mathbf{P},\mathbf{m}}). \end{aligned}$$

Thus, increasing the number of dimensions or the number of splits does not make the search for K^* much more complicated, at least conceptually.

However, computationally speaking, the number of operations performed increases significantly with the number of variables and the number of chosen splits for each variable. Despite the simplicity of the operations, their sheer number can be the cause of an important computational cost. In general, because we need each element of the K^*-hypergrid, the number of required values is the number of regions:

$$N_R = (S_Y + 1) \prod_{b=1}^{B} (S_{P_b} + 1) = \prod_{d=1}^{D} (S_d + 1)$$

Table 4.3 shows that the number of regions can get very high very fast, which poses a computational risk.

N_R	$(S_d+1)=2$	$(S_d+1)=3$	$(S_d+1)=5$	$(S_d+1)=10$
$D=2$	4	9	25	100
$D=3$	8	27	125	1,000
$D=5$	32	243	3,125	100,000
$D=10$	1,024	59,049	9,765,625	10,000,000,000

Table 4.3: Total number of regions with respect to number of variables and number of chosen splits for each variable. The number of splits is the same for all variables.

A way to reduce this risk is to group categories of variables when possible, as discussed in Remark 4.1.1. However, as the table shows, the number of regions is much more sensitive to the number of variables D than to the number of splits for each variable S_d. To reduce D, some preliminary analysis should be done on the data to evaluate whether any variables can be dropped.

Adjusting K^*

In general, assessing the corrections to premiums due to K^* before performing the premium inversion is difficult. In some cases, despite best efforts to minimize the impact of the correction, the changes may be too pronounced. For example, using K^* may result in lower premiums, effectively reducing income to the insurer. For such situations, some flexibility can be helpful, and we elaborate here on how to introduce it.

We return to the univariate P assumption. Consider the set $K^* - A = \{\kappa^*_{(i,j)} - \alpha_{(i,j)} \mid (i,j) \in \mathcal{I}_\mathcal{T}\}$. Since both K^* and A will always have a sum of 1, it is expected that $K^* - A$ will have a sum of 0. We further observe that, because the column and row sums of K^* and A are the same due to properties (4.1.4) and (4.1.5), the column and row sums of the $(K^* - A)$-grid will also all be 0. This last statement will be true not only for $K^* - A$, but also for any multiple of $K^* - A$. Thus, we can introduce the *strength parameter* $\lambda \in [0, 1]$ and let

$$K(\lambda) = A + \lambda(K^* - A) = \{\alpha_{(i,j)} + \lambda(\kappa^*_{(i,j)} - \alpha_{(i,j)}) \mid (i,j) \in \mathcal{I}_\mathcal{T}\}. \tag{4.1.6}$$

We have as special cases $K(0) = A$ and $K(1) = K^*$. Also, denoting $\mathbb{Q}^\lambda$ as the measure obtained from $K(\lambda)$, we get $\mathbb{Q}^0 = \mathbb{P}$ and $\mathbb{Q}^1 = \mathbb{Q}^*$.

The advantage of $K(\lambda)$ is that it retains the minimal impact on the marginal distributions of Y and P while also being flexible in how much it diverts from $\mathbb{P}$ *and* in how close to zero the $\Delta_i^{\mathbb{Q}^\lambda}$ will be. As λ increases from 0 to 1:

- The marginal probabilities of Y at each element of $\mathbf{t}_Y$ and of P at each

element of $\mathbf{t}_P$ are unaffected, *i.e.*

$$\forall \lambda \in [0,1]\ \forall i \in \{0,1,\ldots,I\} \quad \mathbb{Q}^\lambda(Y \in H_{Y,i}) = \mathbb{P}(Y \in H_{Y,i})$$

and

$$\forall \lambda \in [0,1]\ \forall j \in \{0,1,\ldots,J\} \quad \mathbb{Q}^\lambda(P \in H_{P,j}) = \mathbb{P}(P \in H_{P,j}).$$

- $D_{KL}\left(\mathbb{Q}^\lambda \parallel \mathbb{P}\right)$ increases from 0
- For all $i \in \{0,1,\ldots,I\}$ $\Delta_i^{\mathbb{Q}^\lambda}$ decreases from $\Delta_i^{\mathbb{P}}$ to 0

Intuitively, and as will be demonstrated in Section 4.2, this also means that the more λ increases, the more premiums will change from their initial values. Hence, λ can be used to control various statistics on the premiums, such as total premiums received, mean of premiums received, *etc.* Tuning λ can help diminish the overall reduction of premiums.

Proposition 4.1.7 shows that, under specific conditions, the decrease of $\Delta_i^{\mathbb{Q}^\lambda}$ from $\Delta_i^{\mathbb{P}}$ to 0 can be linear in λ.

Proposition 4.1.7. *When $S_Y = S_P = 1$, $\Delta_i^{\mathbb{Q}^\lambda}$ is a linear function of $\lambda \in [0,1]$. In particular,*

$$\Delta_i^{\mathbb{Q}^\lambda} = (1-\lambda)\Delta_i^{\mathbb{P}}.$$

Proof. When there are only two categories of P ($S_P = 1$), we have that $\max_{0\le j\le 1} \kappa_{i|j}$ and $\min_{0\le j\le 1} \kappa_{i|j}$ will each be one of $\kappa_{i|0}$ or $\kappa_{i|1}$. Without loss of generality, assume

that $\max_{0\leq j\leq 1}\kappa_{i|j}=\kappa_{i|0}$ and $\min_{0\leq j\leq 1}\kappa_{i|j}=\kappa_{i|1}$. Then,

$$\begin{aligned}
\Delta_i^{\mathbb{Q}^\lambda} &= \kappa_{i|0}-\kappa_{i|1}\\
&= \frac{\kappa_{(i,0)}}{\kappa_{\bullet,0}}-\frac{\kappa_{(i,1)}}{\kappa_{\bullet,1}}\\
&= \frac{\alpha_{(i,0)}+\lambda\cdot\left(\alpha_{i,\bullet}\cdot\alpha_{\bullet,0}-\alpha_{(i,0)}\right)}{\alpha_{\bullet,0}}-\frac{\alpha_{(i,1)}+\lambda\cdot\left(\alpha_{i,\bullet}\cdot\alpha_{\bullet,1}-\alpha_{(i,1)}\right)}{\alpha_{\bullet,1}}\\
&= \frac{(1-\lambda)\cdot\alpha_{(i,0)}+\lambda\cdot\alpha_{i,\bullet}\cdot\alpha_{\bullet,0}}{\alpha_{\bullet,0}}-\frac{(1-\lambda)\cdot\alpha_{(i,1)}+\lambda\cdot\alpha_{i,\bullet}\cdot\alpha_{\bullet,1}}{\alpha_{\bullet,1}}\\
&= (1-\lambda)\left[\frac{\alpha_{(i,0)}}{\alpha_{\bullet,0}}-\frac{\alpha_{(i,1)}}{\alpha_{\bullet,1}}\right]+\lambda\left[\alpha_{i,\bullet}-\alpha_{i,\bullet}\right]\\
&= (1-\lambda)\Delta_i^{\mathbb{P}},
\end{aligned}$$

where, in the third line, we make use of (4.1.6) in the numerators and of the fact that column sums of the $K(\lambda)$-grid are the same as that of the A-grid in the denominators. □

We can leverage (4.1.2) to determine an ideal value of λ

$$\lambda=\inf\{\omega\in[0,1]\mid\forall i\in\{0,1,\ldots,I\}\ \Delta_i^{\mathbb{Q}^\omega}<\epsilon\}. \tag{4.1.7}$$

This selection of λ makes it so that all $\Delta_i^{\mathbb{Q}^\lambda}$ will be below ϵ – effectively bringing conditional distributions of Y across categories of P closer – but not all of them will have been so affected that they go to 0. Also, taking the infimum of the set reduces how much $\mathbb{Q}^\lambda$ strays from $\mathbb{P}$.

We note that selecting λ in this way is a suggestion, and other options can be considered based on the situation, as will be exemplified in Section 4.2.2. However, if ϵ were to be imposed by a regulator, then (4.1.7) can be seen as dictating how much $\mathbb{Q}^\lambda$ *must* diverge from $\mathbb{P}$ to be considered fair.

Inverting the Premium

We now suppose an adjusted set K has been selected to perform a change of measure from $\mathbb{P}$ to $\mathbb{Q}$. Let Y_n be the initially calculated premium for individuals $n = 1, 2, \ldots, N$. Then, for individual n, there are two steps to obtaining the corrected premium Y_n^c from $\mathbb{Q}$:

1. Find $u_n = \mathbb{P}(Y \leq Y_n)$;
2. Invert the marginal distribution of Y under $\mathbb{Q}$ at u_n, *i.e.* compute the inverted premium $Y_n^c = \inf\{y > 0 \mid \mathbb{Q}(Y \leq y) \geq u_n\}$.

By this method, we transfer quantiles of $\mathbb{P}$ to quantiles of $\mathbb{Q}$, *i.e.* the individual who had the $100u^{\text{th}}$ quantile of Y under $\mathbb{P}$ will be the individual who has the $100u^{\text{th}}$ quantile of Y under $\mathbb{Q}$, as illustrated in Figure 4.2 for the value $u = 0.8$. We note two of the properties of this method:

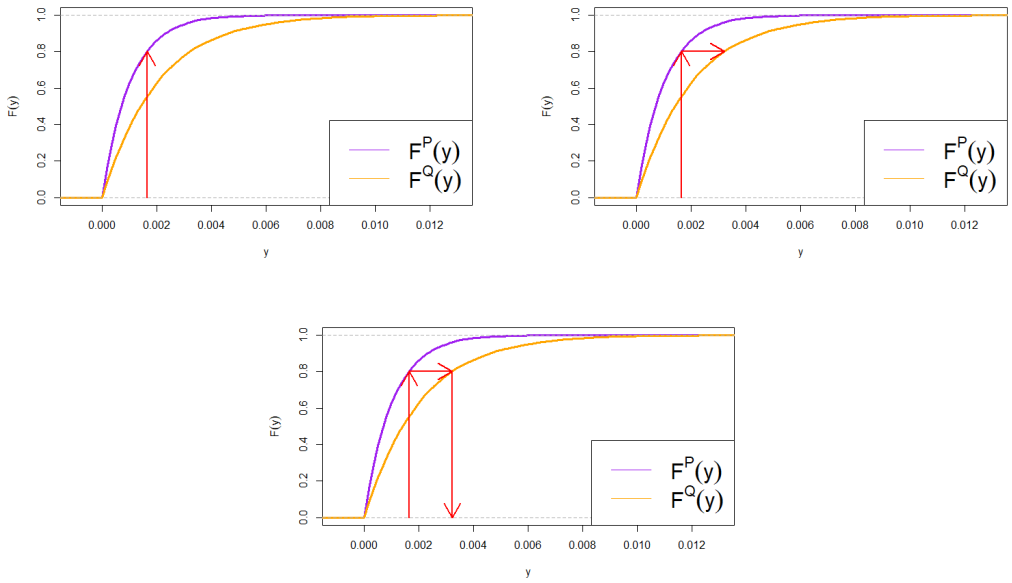

Figure 4.2: Illustration of the calculation of a new premium when both $\mathbb{P}$ and $\mathbb{Q}$ are available. The red arrow illustrates that the 80^{th} quantile under $\mathbb{P}$ becomes the 80^{th} quantile under $\mathbb{Q}$.

1. *Domain preservation*: Due to $\mathbb{Q}$ being absolutely continuous with respect to $\mathbb{P}$, the domain of Y under $\mathbb{Q}$ is a subset of the domain of Y under $\mathbb{P}$. As

such, drawing quantiles from $\mathbb{Q}$ ensures that we do not attribute a premium that would have been "impossible" under $\mathbb{P}$ (outside of its domain). Recall that initial premiums are such that $Y = f(\mathbf{X})$ for some model function f, thus there is already sufficient statistical motivation behind them.

2. *Order preservation*: If, for two individuals A and B, we had $Y_A \leq Y_B$ under $\mathbb{P}$, then $Y_A^c \leq Y_B^c$ will be true under $\mathbb{Q}$ as well. This is directly due to this method being a "transfer of quantiles" of sort. Order preservation is a desirable property because individuals may not be interested in a "fairer" premium if it meant that they now had a higher premium than someone who used to have a higher premium than them.

4.2 Illustrating the Correction Method

4.2.1 The Data

For the illustration, we simulate $N = 10000$ observations with 8000 of them having $P = 0$ and 2000 of them having $P = 1$. The simulations are such that $Y \mid P = 0 \sim N(\mu_0 = 1000, \sigma_0 = 250)$ and $Y \mid P = 1 \sim N(\mu_1 = 1300, \sigma_1 = 250)$. The difference $\mu_1 - \mu_0$ is exaggerated here for illustrative purposes. Figure 4.3 shows the CDFs for premiums and for all categories of P. As expected, $\forall y\ \mathbb{P}(Y \leq y \mid P = 0) \geq \mathbb{P}(Y \leq y) \geq \mathbb{P}(Y \leq y \mid P = 1)$. The distribution of the global premiums is some weighted average of the two conditional distributions, so it lies between them. Also, because there are much more observations with $P = 0$, the global distribution (black) is closer to the distribution for observations with $P = 0$ (red) than to the the distribution for observations with $P = 1$ (green).

These observations are also supported by Table 4.4. Quartiles and central measures for Y always lie between that for $Y \mid P = 0$ and $Y \mid P = 1$.

Remark 4.2.1. *For reasons detailed in the below, it is required that observations*

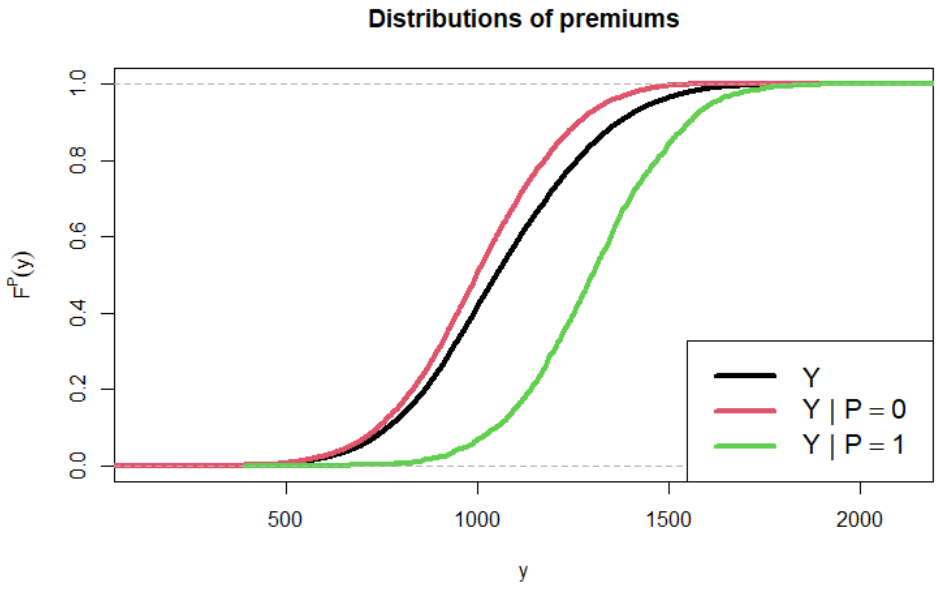

Figure 4.3: CDFs of global premiums (black), premiums for observations with $P = 0$ (red) and premiums for observations with $P = 1$ (green) under the empirical measure $\mathbb{P}$.

$\mathbb{P}$	**Min.**	**1st qrtl**	**Median**	**EV**	**3rd qrtl**	**Max.**	**SD**
Y	265.7	898.0	1,047.6	1,058.7	1,217.1	1,975.4	235.4
$Y \mid P = 0$	265.7	864.0	996.5	998.7	1,137.0	1,762.1	203.8
$Y \mid P = 1$	610.0	1,170.4	1,298.8	1,298.7	1,429.6	1,975.4	197.1

Table 4.4: Summary of premiums, premiums for $P = 0$ and premiums for $P = 1$ under the empirical distribution $\mathbb{P}$.

of $Y \mid P = 0$ *and* $Y \mid P = 1$ *have some overlap, which is the case in our simulated data.*

We have $\mathbf{Z} = (Y, P)$ and so the number of dimensions is $D = 2$. Also, let $\mathcal{T} = \{\mathbf{t}_Y, \mathbf{t}_P\} = \{(0, y_{0.65}, \infty), (0, 0.5, 1)\}$, such that the domain of Y is split at its 65^{th} quantile $y_{0.65}$ and the domain of P is split between 0 and 1. Thus, the domain of $\mathbf{Z}$ is split in four regions:

- $R_{0,0} = [0, y_{0.65}] \times [0, 0.5] \sim [0, y_{0.65}] \times \{0\}$,
- $R_{0,1} = [0, y_{0.65}] \times (0.5, 1] \sim [0, y_{0.65}] \times \{1\}$,
- $R_{1,0} = (y_{0.65}, \infty) \times [0, 0.5] \sim (y_{0.65}, \infty) \times \{0\}$,

- $R_{1,1} = (y_{0.65}, \infty) \times (0.5, 1] \sim (y_{0.65}, \infty) \times \{1\}$.

Table 4.5 summarizes how many observations lie in each of the four regions.

	$Y \leq y_{0.65}$	$Y > y_{0.65}$	Total
$P = 1$	415	1585	2000
$P = 0$	6085	1915	8000
Total	6500	3500	10000

Table 4.5: Contingency table of the combination of premiums below or above $y_{0.65}$ and of the category of the protected variable.

Following up on Remark 4.2.1, note that there is no zero in the contingency table, which means there will be no zero-valued $\alpha_{(i,j)}$. This is an important requirement, as the absence of zero-valued $\alpha_{(i,j)}$ is a key assumption to Proposition 3.1.1. When applying this to real data, the set of splits $\mathcal{T}$ should always be selected such that this is respected.

Under the empirical distribution, we simply divide the contingency table by the number of observations $N = 10000$ to obtain the probabilities under $\mathbb{P}$, $\alpha_{(i,j)}$ for $(i, j) \in \mathcal{I}_{\mathcal{T}}$. Figure 4.4 illustrates the segmentation of the domain of $\mathbf{Z}$ and shows the obtained probabilities.

A few validations can be made:

- $\alpha_{0,0} + \alpha_{0,1} = \alpha_{0,\bullet} = 0.65 = \mathbb{P}(Y \leq y_{0.65})$,
- $\alpha_{1,0} + \alpha_{1,1} = \alpha_{1,\bullet} = 0.35 = \mathbb{P}(Y > y_{0.65})$,
- $\alpha_{0,0} + \alpha_{1,0} = \alpha_{\bullet,0} = 0.8 = \mathbb{P}(P = 0)$,
- $\alpha_{0,1} + \alpha_{1,1} = \alpha_{\bullet,1} = 0.2 = \mathbb{P}(P = 1)$.

4.2.2 Correcting the Premiums

Now that we have the data, our first step is to use the Δ-test to determine whether there is a need for correction. Preliminary calculations are the $\alpha_{i|j}$, presented in

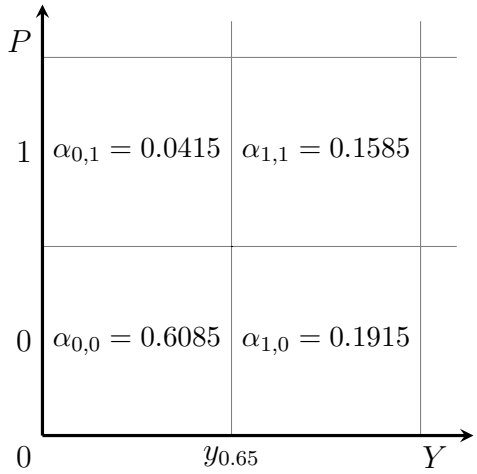

Figure 4.4: Graphical representation of the segmentation of $\mathrm{Dom}(V)$. The $\alpha_{(i,j)}$ for $i, j \in \{0, 1\}$ represent the probability of the random vector $V = (Y, P)$ of lying in the corresponding region.

Table 4.6.

$\alpha_{i\|j}$	$j = 0$	$j = 1$
$i = 0$	0.7606	0.2075
$i = 1$	0.2394	0.7925

Table 4.6: Conditional probabilities

Given Table 4.6, we can calculate the $\Delta_i^{\mathbb{P}}$:

$$\begin{aligned}
\Delta_0^{\mathbb{P}} &= \max_{0 \le j \le J} \alpha_{0|j} - \min_{0 \le j \le J} \alpha_{0|j} \\
&= \max\{0.7606, 0.2075\} - \min\{0.7606, 0.2075\} \\
&= 0.7606 - 0.2075 \\
&= 0.5531.
\end{aligned}$$

$$\begin{aligned}
\Delta_1^{\mathbb{P}} &= \max_{0 \le j \le J} \alpha_{1|j} - \min_{0 \le j \le J} \alpha_{1|j} \\
&= \max\{0.2394, 0.7925\} - \min\{0.2394, 0.7925\} \\
&= 0.7925 - 0.2394 \\
&= 0.5531.
\end{aligned}$$

Note that these results agree with Proposition 4.1.4 as we are in the special case $S_Y = 1$.

For the Δ-test, we choose $\epsilon = 0.1$ and have that

$$\Delta_0^{\mathbb{P}} = 0.5531 > 0.1 \quad \text{and} \quad \Delta_1^{\mathbb{P}} = 0.5531 > 0.1.$$

The Δ-test suggests that there is a need for correction, which we apply in the following.

Using K^*

The first step to the correction method is finding K^*. As starting vector, we set $\mathbf{v} = \dot{\mathbf{v}} = (\alpha_{0,\bullet}, \alpha_{1,\bullet})$, the sum of the columns of the A-grid shown in Figure 4.4. This means we are using the Product-of-Marginals method:

$$\kappa^*_{(i,j)} = \alpha_{i,\bullet} \cdot \alpha_{\bullet,j}$$

The K^*-grid is illustrated in Figure 4.5. We can validate that column and row sums are as for the A-grid:

- ${\kappa^*}_{\bullet,0} = {\kappa^*}_{(0,0)} + {\kappa^*}_{(1,0)} = 0.52 + 0.28 = \mathbf{0.8} = \alpha_{\bullet,0}$
- ${\kappa^*}_{\bullet,1} = {\kappa^*}_{(0,1)} + {\kappa^*}_{(1,1)} = 0.13 + 0.07 = \mathbf{0.2} = \alpha_{\bullet,1}$
- ${\kappa^*}_{0,\bullet} = {\kappa^*}_{(0,0)} + {\kappa^*}_{(0,1)} = 0.52 + 0.13 = \mathbf{0.65} = \alpha_{0,\bullet}$
- ${\kappa^*}_{1,\bullet} = {\kappa^*}_{(1,0)} + {\kappa^*}_{(1,1)} = 0.28 + 0.07 = \mathbf{0.35} = \alpha_{\bullet,0}$

Another important validation is that $\Delta_i^{\mathbb{Q}^*} = 0$ for $i = 0, 1$. We first calculate the ${\kappa^*}_{i|j} = \frac{{\kappa^*}_{(i,j)}}{\kappa^*_{\bullet,j}}$ and present them in Table 4.7.

${\kappa^*}_{i\|j}$	$j=0$	$j=1$
$i=0$	0.65	0.65
$i=1$	0.35	0.35

Table 4.7: Conditional probabilities under $\mathbb{Q}^*$

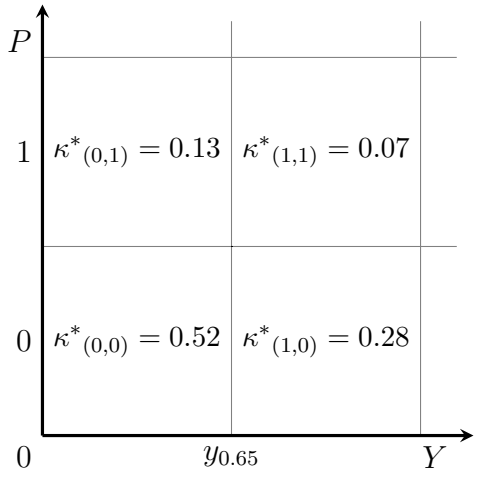

Figure 4.5: K^*-grid for the illustration.

Using Table 4.7, we can calculate the $\Delta_i^{\mathbb{Q}^*}$. For $i = 0$, we have:

$$\begin{aligned}
\Delta_0^{\mathbb{Q}^*} &= \max_{0\leq j\leq 1} {\kappa^*}_{0|j} - \min_{0\leq j\leq 1} {\kappa^*}_{0|j} \\
&= \max\{0.65, 0.65\} - \min\{0.65, 0.65\} \\
&= 0.65 - 0.65 \\
&= 0,
\end{aligned}$$

and for $i = 1$, we have:

$$\begin{aligned}
\Delta_1^{\mathbb{Q}^*} &= \max_{0\leq j\leq 1} {\kappa^*}_{1|j} - \min_{0\leq j\leq 1} {\kappa^*}_{1|j} \\
&= \max\{0.35, 0.35\} - \min\{0.35, 0.35\} \\
&= 0.35 - 0.35 \\
&= 0.
\end{aligned}$$

Thus, our validation is satisfied.

We use K^* along with Propositions 3.1.1 and 3.1.2 to produce the measure $\mathbb{Q}^*$ on $\mathbf{Z}$. Then, using our inversion method, we can calculate new premiums. Table 4.8 presents some statistics on premiums obtained from $\mathbb{Q}^*$. We first compare expected premiums and observe that $\mathrm{E}^{\mathbb{Q}^*}(Y) - \mathrm{E}^{\mathbb{P}}(Y) = 2.0$. An important

nuance must be made here that Table 4.8 presents a summary of the corrected premiums under their distribution of origin $\mathbb{Q}^*$, such that the expected value is equal to:

$$\mathrm{E}^{\mathbb{Q}^*}(Y) = \sum_{n=1}^{N} y \cdot \mathbb{Q}^*\left(Y = y\right) = 1060.7.$$

It may seem intuitive to multiply this value by N to obtain the total premiums received by the insurer under $\mathbb{Q}^*$, but that would be incorrect. The insurer would simply receive the sum of the corrected premiums:

$$\sum_{n=1}^{N} Y^c = N\left(\frac{1}{N}\sum_{n=1}^{N} Y^c\right) \neq N\left(\sum_{n=1}^{N} y \cdot \mathbb{Q}^*\left(Y = y\right)\right) = N \cdot \mathrm{E}^{\mathbb{Q}^*}(Y).$$

The inequality shows that it would indeed be wrong to multiply $\mathrm{E}^{\mathbb{Q}^*}(Y)$ by N to obtain the total corrected premiums received. The fact that the total premiums received are calculated on an "empirical" basis makes it important to compare Y^c and Y on that same basis as well. Table 4.9 presents some statistics and shows, in particular, that the average change to premiums is an increase of approximately 0.6$, which is very small relative to their scale.

$\mathbb{Q}^*$	**Min.**	**1st**	**Median**	**EV**	**3rd**	**Max.**	**SD**
Y	292.1	933.7	1,077.9	1,060.7	1,192.2	1,975.4	197.9
$Y \mid P = 0$	292.1	909.5	1,040.2	1,138.2	1,186.5	1,724.9	197.4
$Y \mid P = 1$	626.0	1,033.5	1,142.6	1,274.0	1,245.3	1,975.4	178.0

Table 4.8: Summary of premiums, premiums for $P = 0$ and premiums for $P = 1$ under the new distribution $\mathbb{Q}^*$.

	Min.	**1st**	**Median**	**Mean**	**3rd**	**Max.**
$Y^c - Y$	-82.5	-13.7	14.4	0.6	17.1	57.3
Y^c/Y (%)	95.2	98.9	101.5	100.5	102.0	120.0

Table 4.9: Summary of the differences between premiums Y and corrected premiums Y^c on an empirical basis.

As for the quartiles of each distribution of Y under $\mathbb{Q}^*$, we still have that quantities for Y lie between the corresponding quantities for $Y \mid P = 0$ and $Y \mid P = 1$,

providing a good sense check. Also, the standard deviations have greatly reduced from the standard deviations under $\mathbb{P}$. In wanting to bring the distributions of $Y \mid P = 0$ and $Y \mid P = 1$ closer to that of Y, we have also made the distributions themselves more compact. We also note that the standard deviation of Y is much closer to that of $Y \mid P = 0$ than to that of $Y \mid P = 1$ because, again, there are many more individuals with $P = 0$ than with $P = 1$.

Reverting back to the small average increase, there are two caveats:

1. A very small average increase does not imply that every individual increase is also very small. It may simply be the case that individuals with very large increases are offset by individuals with very large decreases. This means that a minimal impact to the total premiums can hide very large impacts to policyholder premiums, which may lead to multiple policyholders simply leaving the portfolio.

2. Because premiums are an important part of revenue for the insurer, a small change to average premium can lead to an important change to total premium received. This is particularly worrisome if the change of measure results in an average decrease of premiums and leads to a significant loss of revenue.

After having considered summary statistics of the corrected premiums, we analyze them more diligently through their distributions. Figure 4.6 shows the distributions of premiums under $\mathbb{Q}^*$. The first observation is that all distributions – the global distribution, the distribution conditional on $P = 0$ and the distribution conditional on $P = 1$ – are joined at $y_{0.65}$, meaning that, under $\mathbb{Q}^*$, the global distribution and the two conditional distributions of Y all have $y_{0.65}$ as 65^{th} quantile. This is directly due to the choice of K^*, as demonstrated by table 4.7.

Comparing Figures 4.3 and 4.6, a visual interpretation is that the conditional distributions are *pulled* towards the global distribution *along* the vertical red line

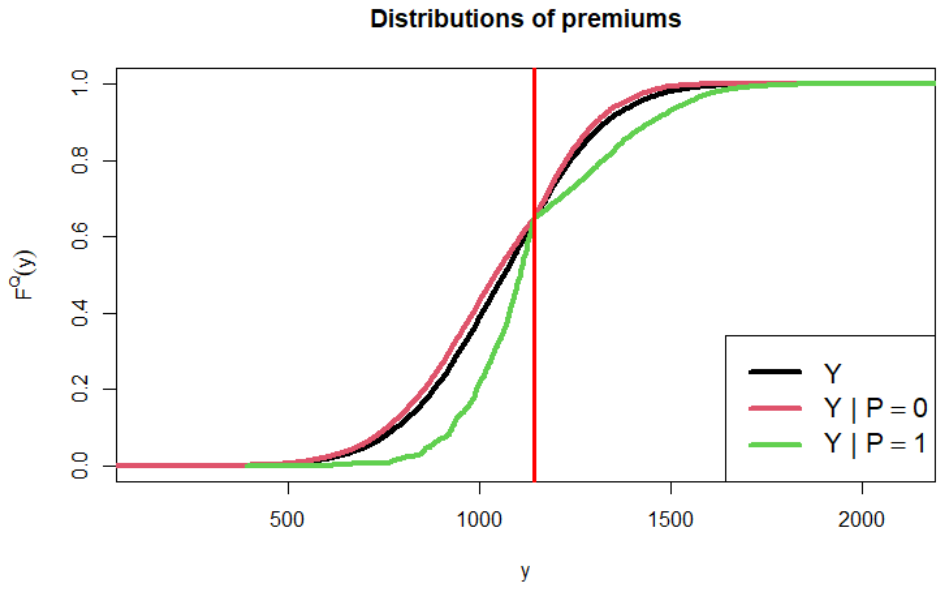

Figure 4.6: CDFs of global premiums (black), premiums for observations with $P = 0$ (red) and premiums for observations with $P = 1$ (green) under the corrected measure $\mathbb{Q}$. The vertical red line represents the 65^{th} quantile of Y, $y_{0.65} = 1143.06$.

at $y_{0.65}$. Recall, when building K^*, we used as starting vector the marginal Y probabilities $\dot{\mathbf{v}}$ (A-grid column sums) and allocated them to the K^*-grid rows with respect to marginal P probabilities (A-grid row sums). This implies that the marginal behavior of Y at elements of $\mathbf{t}_Y$ will be replicated by the conditional distributions as well, and this in reflected in Figure 4.6.

Despite them being close, the global distributions under $\mathbb{P}$ and $\mathbb{Q}$ are not the same. They are compared graphically in Figure 4.7. As already demonstrated, the global distributions are joined at $y_{0.65}$. For $y < y_{0.65}$, $\mathbb{Q}^*(Y \leq y)$ is smaller than $\mathbb{P}(Y \leq y)$, while the opposite is true for $y < y_{0.65}$. That is because, under $\mathbb{P}$, premiums under $y_{0.65}$ mostly came from individuals with $P = 0$ (6085 out of 6500 individuals, as per the contingency Table 4.5). Under $\mathbb{Q}^*$, we are trying to bring all premiums closer to the marginal distribution of the premiums, and so premiums for individuals with $P = 0$ will mostly increase, because their distribution is stochastically dominated by that of the marginal premiums.

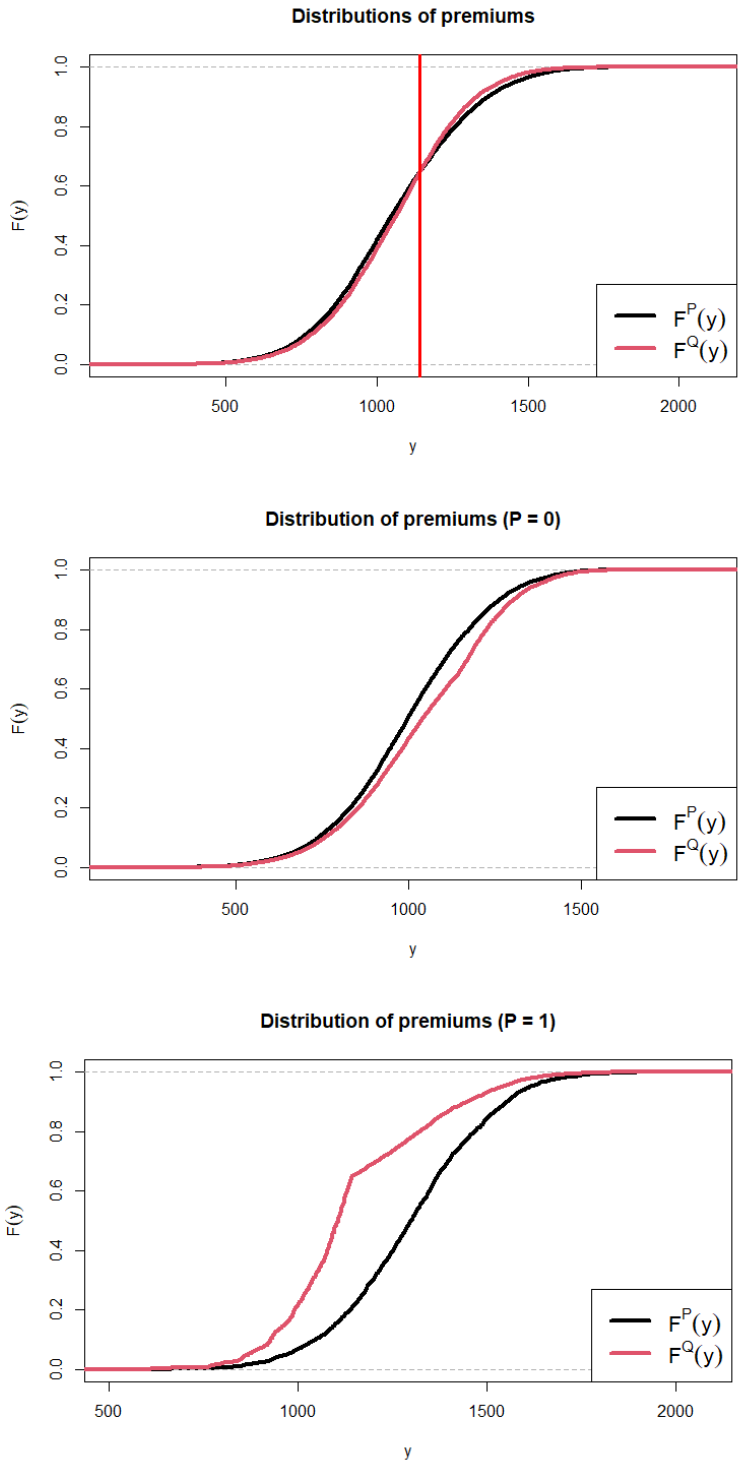

Figure 4.7: Distributions of Y (top), $Y \mid P = 0$ (middle) and $Y \mid P = 1$ (bottom) under $\mathbb{P}$ (black) and $\mathbb{Q}^*$ (red).

After comparing the distributions of premiums, we look at the corrections made to the premiums in Figure 4.8. The first observation is that corrections seem to form a continuous function, say g, of the initial premium, with some noise closer to the smaller and larger initial premiums. This function g is positive, zero and negative when the initial premium is smaller than, equal to and greater than $y_{0.65}$, respectively. This follows from the distributions of Y under $\mathbb{P}$ and $\mathbb{Q}^*$, as explained when considering Figure 4.7 (top). All premiums below $y_{0.65}$ will increase while all premiums above $y_{0.65}$ will decrease. Also, a clear change in the behavior of g occurs at the blue line in the figure, located at the initial mean $\mathrm{E}^{\mathbb{P}}(Y) = 1058.7$. Before the initial mean, corrections are slowly increasing, while they are sharply decreasing after the initial mean. Our inversion method seeks to bring the means of $Y \mid P = 0$ and $Y \mid P = 1$ closer to that of Y. Because $\mathrm{E}^{\mathbb{P}}(Y \mid P = 0) = 998.7$ is much closer to $\mathrm{E}^{\mathbb{P}}(Y)$ than $\mathrm{E}^{\mathbb{P}}(Y \mid P = 1) = 1298.7$, the corrections will, in aggregate, be less severe for $P = 0$ than for $P = 1$. However, due to the inversion being made on the distribution $\mathbb{Q}^*(Y \leq \cdot)$ independently of P, any correction will depend solely on the premium itself, and corrections being less severe for individuals with $P = 0$ translates to corrections being less severe for *small* premiums, since premiums for individuals with $P = 0$ tend to be small.

Remark 4.2.2. *It is possible to modify the inversion such that inverted premiums also depend on the protected variable. Instead of inverting from $\mathbb{Q}^*(Y \leq \cdot)$, one could invert from $\mathbb{Q}^*(Y \leq \cdot \mid P = p)$ for individuals with $P = p$.*

In addition, corrections are, for the most part, small relative to the initial premiums. As per Table 4.9, they vary from roughly -80 to 60, which translates to a multiplicative correction going from approximately 95% to 120%, with only 6 individuals out of 10000 exceeding a 105% correction. Recall the disparity between premiums under $P = 0$ and $P = 1$ was significant ($\mu_1 - \mu_0 = 300$ being a relative difference of 30%). Despite this important bias, the corrections not exceeding 100 in absolute value attests that the scale of the solution will not attain

the scale of the problem, which is a comforting notion. This also demonstrates that the minimization of the KL divergence and the choice of $\dot{\mathbf{v}}$ as starting vector effectively reduce the impact of the change of measure.

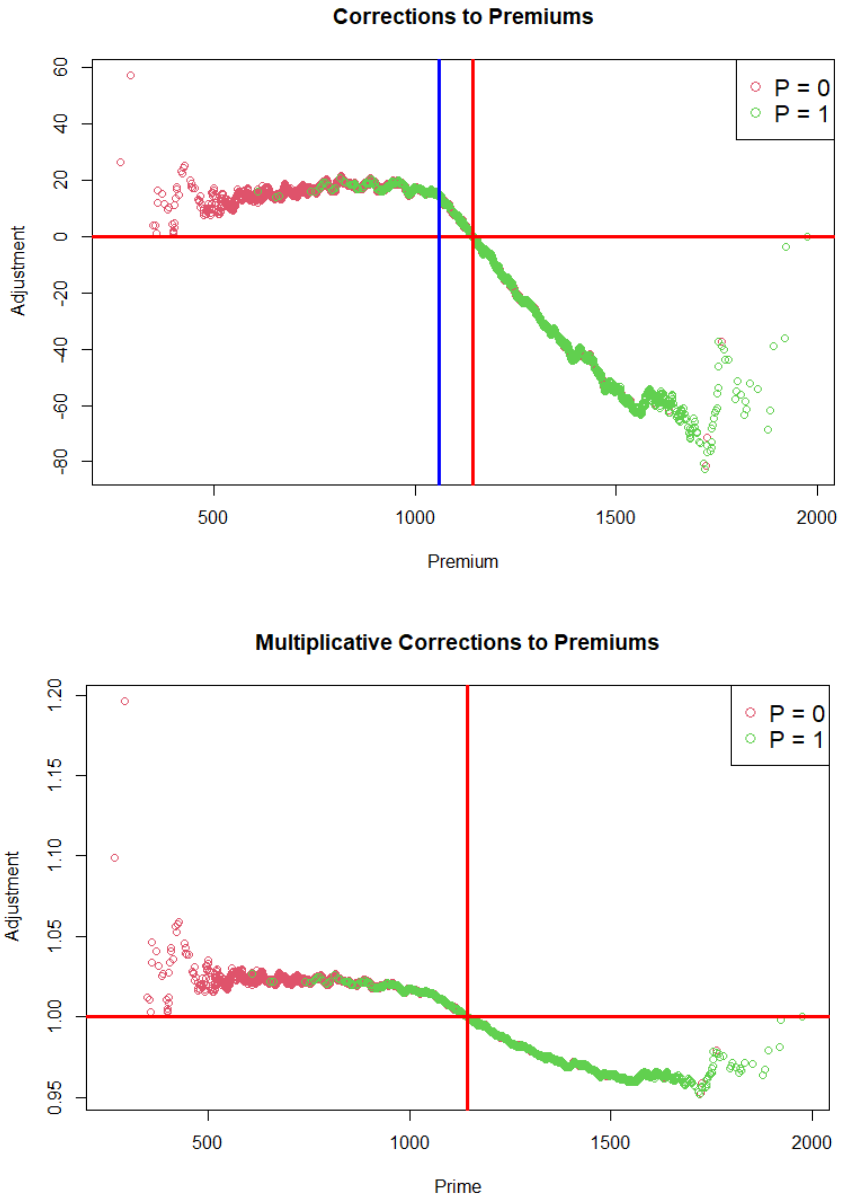

Figure 4.8: Corrections (top) and multiplicative corrections (bottom) made to premiums after inversion from the distribution $\mathbb{Q}^*$ for $P = 0$ (red) and $P = 1$ (green). The red lines pinpoint the neutrality point at an initial premium of $y_{0.65}$, and the blue line indicates the mean of the initial premiums.

Adjusting K^*

In the previous section, we showed that the overall impact of the inversion method on the initial premiums was relatively small. One of the caveats we mentioned in the previous section is that a small overall impact does not guarantee that all individual corrections will be small as well. We use this as motivation to demonstrate here how the insurer could use the set $K(\lambda)$ to minimize this effect.

Recall (4.1.6), in which we defined $K(\lambda)$ as a generalization of K^* which retains the column sums of A (and so the marginal Y probabilities under $\mathbb{P}$) but results in $\Delta_i^{\mathbb{Q}}$ values which will not necessarily be equal to 0:

$$K(\lambda) = A + \lambda(K^* - A) = \{\alpha_{(i,j)} + \lambda(\kappa_{(i,j)} - \alpha_{(i,j)}) \mid (i,j) \in \mathcal{I}_{\mathcal{T}}\}.$$

The strength parameter λ controls how much $K(\lambda)$ and $\mathbb{Q}^\lambda$ will stray from A and $\mathbb{P}$, respectively. It stands to reason that, as λ increases from 0 to 1, other quantities will also gradually move away from their original values under $\mathbb{P}$.

Before discussing the adjustment to K^*, we consider how λ has an effect on $\mathbb{Q}^\lambda$. Figure 4.9 shows how both the KL divergence of $\mathbb{Q}^\lambda$ with respect to $\mathbb{P}$ (top) and $\Delta_i^{\mathbb{Q}^\lambda}$ (bottom) evolve as λ increases from 0 to 1[2]. Recall the KL divergence is a measure of how much $\mathbb{Q}^\lambda$ diverges from $\mathbb{P}$, and $\Delta_i^{\mathbb{Q}^\lambda}$ is a measure of how different behavior of Y is across categories of P under $\mathbb{Q}^\lambda$.

Remark 4.2.3. *Figure 4.9 presents results obtained from performing the change of measure at multiple values of λ. Closed-form expressions for these quantities as functions of λ are difficult to obtain in general.*

As expected, the KL divergence increases along with λ. As stated in Definition 2.2.1, it is a convex function of $\mathbb{Q}^\lambda$. It is 0 only when $\lambda = 0 \implies \mathbb{Q}^\lambda = \mathbb{P}$ and it reaches a maximum at $\lambda = 1 \implies \mathbb{Q}^\lambda = \mathbb{Q}^*$ on the range $\lambda \in [0,1]$. As for $\Delta_i^{\mathbb{Q}^\lambda}$, it also behaves as expected, decreasing linearly from $\Delta_0^{\mathbb{P}} = \Delta_1^{\mathbb{P}} = 0.5531$ at $\lambda = 0$

[2]We need not specify a value of i for $\Delta_i^{\mathbb{Q}^\lambda}$ since $\Delta_0^{\mathbb{Q}^\lambda} = \Delta_1^{\mathbb{Q}^\lambda}$ due to $S_Y = 1$ (see Proposition 4.1.4)

to 0 at $\lambda = 1$. The linear decrease to 0 is a result of having $S_Y = S_P = 1$ (see Proposition 4.1.7). In other cases, the decrease may not be linear, but $\Delta_i^{\mathbb{Q}^\lambda}$ will always reach 0 at $\lambda = 1$.

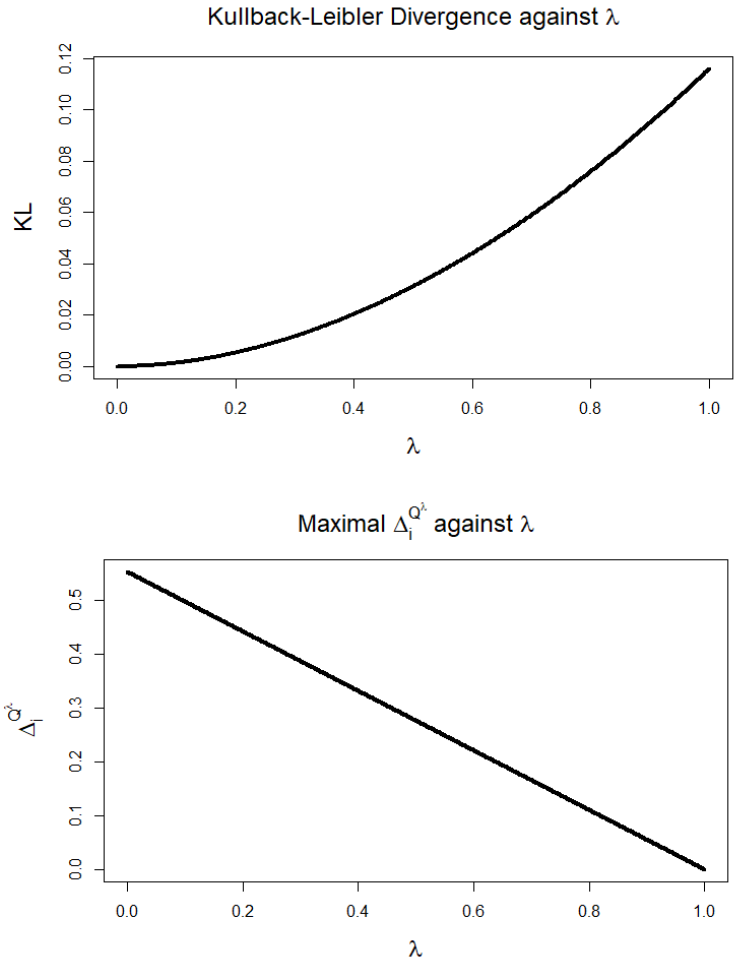

Figure 4.9: Kullback-Leibler divergence (top) and $\Delta_i^{\mathbb{Q}^\lambda}$ (bottom) as functions of λ.

Now, to decrease the individual corrections, the insurer decides to limit the mean absolute correction. The mean absolute correction is preferred to the maximum absolute correction because it will be less sensitive to outliers. In our current illustration, if one individual had an correction of say 500, capping corrections at 100 would certainly decrease that individual's correction, but would have nearly no effect on all other corrections, which range from -80 to 60. Reducing the mean

absolute difference is a way to work around this issue and affect all corrections rather than only the more extreme ones. Figure 4.10 illustrates the mean absolute corrections as a function of λ. If the insurer chose to cap the mean absolute correction at 12.5, they would simply need to use the greatest λ that produces a mean absolute correction greater than or equal to 12.5.

Remark 4.2.4. *While it was expected that the mean absolute correction would increase with λ, the fact the increase seems linear comes as a surprise. We conjecture that it is due to the linear evolution of $K(\lambda)$ from A to K^*. We do not present them here, but multiple other statistics, such as the mean correction, evolve linearly with λ.*

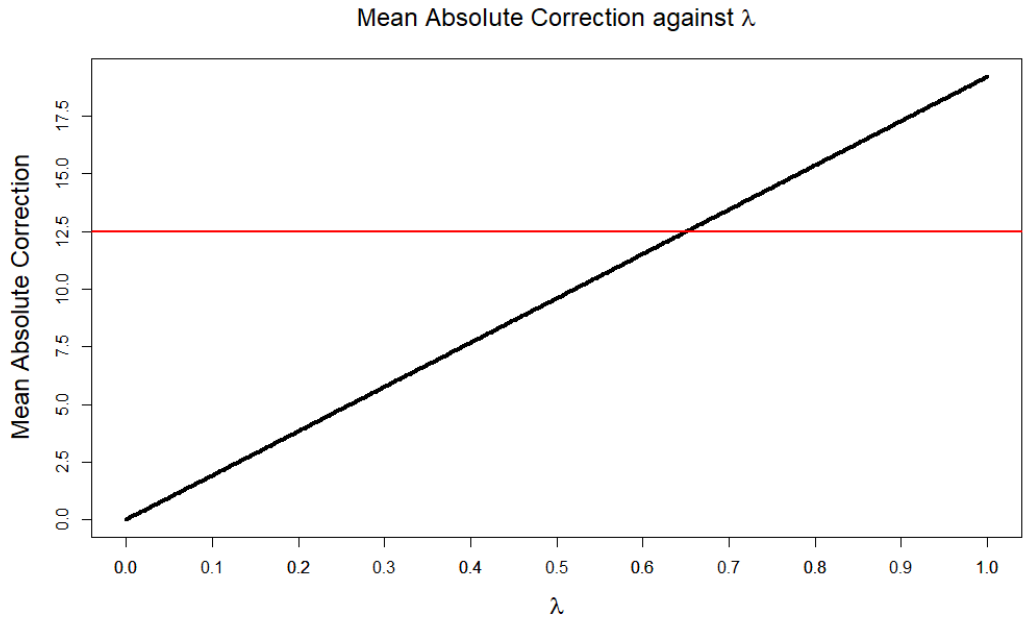

Figure 4.10: Mean absolute corrections applied to premiums as a function of λ. The red line represents the limit of 12.5 determined by the insurer.

In general, the adjustment to K^* can be made using any statistic, or even combination of statistics, as a function of λ, and choosing what limit needs to be imposed. The downside to this method is that complete sets of corrections need to be calculated for each value of λ, which is not computationally efficient. However, doing the exercise for only a few values, say $\lambda = 0.2, 0.4, 0.6, 0.8, 1$, may be

enough to make a selection.

Chapter 5

Conclusion

This thesis concludes by first summarizing the motivation behind our work and our approach to a hopefully more ethical insurance premium, as well as listing its advantages and disadvantages. We then consider some open problems that have been raised from our method.

In the past few decades, insurance regulators have treated the use of protected variables as binary, either allowing or prohibiting their use in their pricing models. However, recent research has repeatedly shown that simply ignoring a variable will not eliminate discrimination towards it. It is actually the case that, to do so effectively, it is better to use the variable appropriately. Because society is becoming more and more concerned about algorithmic fairness, it may not be long before its expectations are turned toward fairness in insurance, and this may be the push needed for insurance regulators to consider different perspectives toward fairness. In preparation for this eventuality, many researchers have developed methods to implement algorithmic fairness in insurance, with respect to different notions of fairness, namely individual fairness and group fairness.

We presented two group fairness approaches to insurance pricing. The first is the discrimination-free premium of Lindholm et al. (2022) [8] which we exposed in Chapter 2. It removes the discriminatory impact of $\mathbf{P}$ on Y due to observable

dependence between $\mathbf{X}$ and $\mathbf{P}$. It is an intra-treatment approach which requires the addition of $\mathbf{P}$ to the model function inputs as well as the weighted averaging of multiple outputs of the resulting function. Although it is very efficient and has the notable advantage of being unbiased, it can drastically increase computation time, particularly for large portfolios.

The second approach we discussed is our own inverted premium as a post-treatment approach to group fairness, as seen in Chapter 4. It hinges on the grid-based change of measure laid out in Chapter 3, a generalization of the change of measure used by Pesenti et al. (2018) [9]. We introduced $\Delta_i^{\mathbb{P}}$ as a quantitative indicator of unfairness based on statistical parity and have shown how a change of measure to a judiciously chosen measure $\mathbb{Q}$ can reduce this quantity to zero. The strength parameter λ can be used to allow for some flexibility and recommendations were made on its selection. The main advantages of our method are that changes to global quantities, such as the mean and the median, are relatively small and that changes to individual premiums do not attain the scale of the initially observed discrepancy. However, a disadvantage is that the selection of λ may be complex and require multiple complete iterations of the process.

Finally, we reiterate that algorithmic fairness is only one of the endeavours that should be undertaken to attain fairness in insurance. The implementation of fairness is a complex issue, and it cannot be completely solved through statistical or mathematical methods alone. Nonetheless, it is important that these methods are as refined as possible in case their implementation is ever required by legislators. This work is intended to supplement these approaches and provide insurers with additional options to group fairness.

5.1 Open Problems

The method we have constructed for ethical insurance premiums relies on the grid-based change of measure of Chapter 3. Due to its flexibility, it is not far-

fetched to say that it can be applied in numerous fields other than insurance. In fact, changes of measure are frequently used in finance, along with the Radon-Nikodym theorem, to price various types of financial products. An interesting endeavour would be to identify other fields that would benefit from the use of a different measure than the empirical one.

In Remark 4.1.6, we noted that we could only control the probability distribution of the premiums at discrete values (the chosen splits). However, if we were to let the number of chosen splits increase to infinity, we would gain more and more control over $\mathbb{Q}$ – potentially forcing independence between Y and P – but stray further and further away from the empirical measure $\mathbb{P}$. Visually, this would correspond to the A-grid (see Figure 3.1) having an infinitely large amount of infinitely small rectangles. We expect that doing so would allow us to fully choose the quantiles of the distribution $\mathbb{Q}$, but would lead to a KL divergence that explodes to infinity, since the $\alpha_{\mathbf{i}} \in A$ would be near-zero (see Definition 2.2.1). This is all conjecture, and future work could shed some light on this line of thought.

In Section 4.2, we illustrated our correction method on simulated data. Our data was built such that there was quite a large discrepancy between means of the premiums for each category of the protected variable, and we showed that we are able to tighten the gap not only for the means, but also for the quantiles of premiums of each category. While this is representative of one type of issue that may be observed from empirical data, many more types of issues remain. For instance, the treatment of data for which means across categories of P are close but variances are very different would likely be very different than what we've presented. Exploring the effect our method has on different types of discrepancies would definitely be insightful.

Figure 4.10 seemed to indicate that there is a direct link between the linear reallocation of probabilities across regions by the strength parameter λ and the change in premium resulting from the inversion method.

In Section 4.1.1, we suggested the Δ-test (see (4.1.2)) as a way to determine whether to apply any correction to the outputs of the initial model function. Recall the test compares the $\Delta_i^{\mathbb{P}}$ to a pre-determined ϵ, and if any $\Delta_i^{\mathbb{P}}$ exceeds ϵ, then the correction should be applied. Because the value of ϵ is so impactful, it should be chosen carefully. However, to properly assess what constitutes an adequate value of ϵ, multiple studies should be made on real data.

Bibliography

[1] S. Axler. *Measure, Integration and Real Analysis*. Springer Cham, 2019. ISBN: 978-3-030-33143-6. URL: https://link.springer.com/book/10.1007/978-3-030-33143-6.

[2] R. Binns. "On the Apparent Conflict Between Individual and Group Fairness". In: *Conference on Fairness, Accountability, and Transparency* (Jan. 2020). DOI: 10.1145/3351095.3372864.

[3] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq. "Algorithmic decision making and the cost of fairness". In: (June 2017). DOI: 10.48550/arXiv.1701.08230.

[4] I. Csiszár. "I-Divergence Geometry of Probability Distributions and Minimization Problems". In: *The Annals of Probability* 3.1 (1975), pp. 146–158.

[5] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. "Fairness Through Awareness". In: (Nov. 2011). DOI: 10.48550/arXiv.1104.3913.

[6] J. Fitzsimons, A. Al Ali, M. Osborne, and S Roberts. "A General Framework for Fair Regression". In: *Entropy* 21.8 (July 2019). DOI: 10.3390/e21080741.

[7] I. Kohler-Hausmann. "Eddie Murphy and the Dangers of Counterfactual Causal Thinking about Detecting Racial Discrimination". In: *Northwestern University Law Review* 113.5 (Mar. 2019), pp. 1163–1228. DOI: 10.2139/ssrn.3050650.

[8] M. Lindholm, R. Richman, A. Tsanakas, and M. V. Wüthrich. "Discrimination-free Insurance Pricing". In: *ASTIN Bulletin: The Journal of the IAA* 52.1 (2022), pp. 55–89. DOI: 10.1017/asb.2021.23.

[9] S. M. Pesenti, P. Millossovich, and A. Tsanakas. "Reverse Sensitivty Testing: What Does It Take to Break the Model?" In: *European Journal of Operational Research* 274.2 (Oct. 2018). DOI: 10.1016/j.ejor.2018.10.003.

[10] F. Petersen, D. Mukherjee, Y. Sun, and M Yurochkin. "Post-processing for Individual Fairness". In: (Oct. 2021). DOI: 10.48550/arXiv.2110.13796.

[11] V. Tremblay. "Équité algorithmique: perspective interdisciplinaire et recommandations pour statisticiens et autres scientifiques de données". In: (2022). hal-03663226.

[12] X. Wang, Y. Zhang, and R. Zhu. "A brief review on algorithmic fairness". In: *Management System Engineering* 1.7 (Nov. 2022), pp. 55–89. DOI: 10.1007/s44176-022-00006-z.

[13] I. Zliobaite. "Measuring discrimination in algorithmic decision making". In: *Data Mining and Knowledge Discovery* 31 (July 2017), pp. 1060–1089. DOI: 10.1007/s10618-017-0506-1.

[14] I. Zliobaite and B. Custers. "Using Sensitive Personal Data May Be Necessary for Avoiding Discrimination in Data-Driven Decision Models". In: *Artificial Intelligence and Law* 24 (June 2016), pp. 183–201. URL: https://ssrn.com/abstract=3047233.