Probabilistic index models

Jan De Neve

Supervisors:

Prof. dr. ir. Olivier Thas and Prof. dr. Jean–Pierre Ottoy

Dissertation submitted in fulfilment of the requirements for the degree of

Doctor in Statistical Data Analysis

Academic year 2012 - 2013

Citation information. De Neve, J. (2013). *Probabilistic index models.* Ghent University. Faculty of Sciences, Ghent, Belgium.

ISBN. 978-9-4619711-1-1.

Dankwoord

Zoals de titel aangeeft omvat deze sectie een dankwoord. Niettegenstaande het afleggen van een doctoraat voornamelijk een individueel proces is, kan het onmogelijk tot stand komen zonder de hulp van velen. Het is dan ook mijn voorrecht om deze mensen te bedanken. Ik zal mijn dankwoord kort en bondig houden, wat niet wegneemt dat mijn dankbaarheid oprecht is.

Vooreerst wens ik mijn beide promotoren te bedanken. Olivier, ik beschouw het nog steeds als een voorrecht dat ik heb mogen meewerken aan het verhaal van de PIMs zoals beschreven in dit doctoraat. Ik wens je ook te bedanken voor je enorme inspanning om me zo goed mogelijk te begeleiden doorheen de doctoraatsjaren – en uiteraard ook voor de vele aangename momenten, waar naast statistiek ook humor centraal stond. Jean–Pierre, bedankt om ervoor te zorgen dat ik naar verschillende congressen kon gaan. Ik herinner me vooral het congres in Bordeaux waar we samen naartoe zijn gegaan en waar ik de mogelijkheid heb gehad om een netwerk uit te bouwen met een invited presentation in Kopenhagen tot gevolg. Beiden uitdrukkelijk dank.

I would like to thank the members of the examination committee, Prof. Marc Aerts, Prof. Michael Akritas, Prof. Stefan Van Aelst, Prof. Mark van de Wiel, and Prof. Stijn Vansteelandt for their careful reading of this dissertation.

Uiteraard wens ik ook al mijn collega's (zowel de huidige als de voormalige) te bedanken en ik het bijzonder mijn BioStat collega's. Discussies gaande van theoretische concepten, over actuele thema's tot totaal irrelevante bedenkingen: ze dragen allemaal bij tot de zeer aangename werksfeer met een goede balans tussen inspanning en ontspanning. Een speciaal woordje wil ik richten aan mijn bureau-collega's. Yingjie, I deeply appreciate your kindness and considerateness. Kristof, jouw inzichten – die veel verder reikten dan de statistiek – alsook je gevoel voor humor, heb ik enorm geapprecieerd en vormden een echte meerwaarde.

Mijn dank gaat ook uit naar mijn familie. Sinds mijn eerste dag aan de universiteit, bijna tien

jaar geleden, hebben ze me ongelooflijk gesteund en aangemoedigd. Het lijdt geen twijfel dat dit een grote positieve impact heeft gehad en dat dit een zeer belangrijke factor is geweest in de totstandkoming van deze thesis. In het bijzonder wil ik mijn geliefde, die ik heb leren kennen op de vooravond van mijn doctoraatsperiode, bedanken. Marieke, bedankt om er samen met mij een prachtig avontuur van te maken.

Jan,

Gent, April 2013.

Voor Philip

Contents

List of symbols and abbreviations

i.i.d.	identically and independently distributed
GLM	Generalized Linear Model
$\mathrm{P}\left(Y \preccurlyeq Y'\right)$	probabilistic index, defined as $\mathrm{P}\left(Y \preccurlyeq Y'\right) := \mathrm{P}\left(Y < Y'\right) + 0.5\mathrm{P}\left(Y = Y'\right)$
PI	Probabilistic Index
PIM	Probabilistic Index Model
GEE	Generalized Estimating Equations
$\mathrm{I}\left(\cdot\right)$	indicator function, i.e. $\mathrm{I}\left(\text{true}\right) = 1$ and $\mathrm{I}\left(\text{false}\right) = 0$
$\mathrm{I}\left(Y \preccurlyeq Y'\right)$	pseudo-observation defined as $\mathrm{I}\left(Y \preccurlyeq Y'\right) := \mathrm{I}\left(Y < Y'\right) + 0.5\mathrm{I}\left(Y = Y'\right)$
$(Y, \boldsymbol{X})$	Y is the univariate outcome and $\boldsymbol{X}$ the associated d-dimensional predictor
$f_{Y\boldsymbol{X}}$	density or probability mass function of the joint distribution $(Y, \boldsymbol{X})$
$f_{Y\|\boldsymbol{X}}$	density or probability mass function of the conditional distribution of Y given $\boldsymbol{X}$
$F_{Y\|\boldsymbol{X}}$	cumulative distribution function of Y given $\boldsymbol{X}$
$\stackrel{d}{=}$	distributed according to - equality in distribution
$\stackrel{p}{\to}$	convergence in probability
$\perp\!\!\!\perp$	statistical independence
LRM	Linear Regression Model
$\mathbb{R}^p$	p-dimensional Euclidean space
δ_{ij}	Kronecker delta
$\mathrm{odds}(A)$	odds, equal to $\mathrm{P}\left(A\right)/[1 - \mathrm{P}\left(A\right)]$ for an event A
$f(x) = O[g(x)]$	there exist a constant c and an $x_0 \in \mathbb{R}$ such that $\lvert f(x)\rvert \leq c\lvert g(x)\rvert$ for $x > x_0$
$f(x) = o[g(x)]$	$\lim_{x\to\infty} f(x)/g(x) = 0$
$X_n = o_p(1)$	$X_n \stackrel{p}{\to} 0$
$X_n = O_p(1)$	X_n is bounded in probability, i.e. $\forall\varepsilon$, $\exists M_\varepsilon$, and $\exists n_\varepsilon$ such that $\mathrm{P}\left(\lvert X_n\rvert > M_\varepsilon\right) < \varepsilon, \forall n > n_\varepsilon$

$X_n = o_p(a_n)$	$a_n^{-1} X_n = o_p(1)$ for a strict positive sequence a_n
$X_n = O_p(a_n)$	$a_n^{-1} X_n = O_p(1)$ for a strict positive sequence a_n
$\Phi(\cdot)$	cumulative distribution function of the standard normal distribution
$\wedge$	AND
$\mathbf{1}$	vector with all elements equal to 1
$\boldsymbol{I}$	identity matrix: diag$(\mathbf{1})$
N(μ, σ^2)	univariate normal distribution with mean μ and variance σ^2
N($\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$)	multivariate normal distribution with mean $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$
Exponential(λ)	exponential distribution with rate λ
t_f	t-distribution with f degrees of freedom
$\boldsymbol{M}^-$	a generalized inverse of a square matrix $\boldsymbol{M}$
$(\boldsymbol{M})_{ij}$	the element on position (i, j) of matrix $\boldsymbol{M}$
$\text{logit}(x)$	logit link function, equal to $\log[x/(1-x)]$
$\text{expit}(x)$	inverse of the logit function, equal to $\exp(x)/[1+\exp(x)]$
SE	standard error
$\text{sign}(x)$	$\text{sign}(x) = 1$ if $x > 0$, $\text{sign}(x) = 0$ if $x = 0$, and $\text{sign}(x) = -1$ if $x < 0$
$\text{Cor}\,(\boldsymbol{X})$	correlation matrix of $\boldsymbol{X}$
$\text{Cov}\,(\boldsymbol{X})$	covariance matrix of $\boldsymbol{X}$

Chapter 1

Introduction

1.1 Setting

A classical problem in statistics is concerned with studying the association between a univariate outcome Y and a d-dimensional set of predictors $\boldsymbol{X}$. In Section 2.5.1, for example, we describe a case study where Y is a measure of a child's lung capacity and $\boldsymbol{X}$ contains the age, gender, and smoking status of the child. The primary focus is to understand the association between the smoking behaviour and the lung capacity, while possible confounding factors, such as gender and age, should be accounted for. In Section 2.5.2 we describe a mental health study where Y denotes a subject's mental impairment and $\boldsymbol{X}$ its life index and socio-economic status. Interest then lies in exploring the relationship between the socio-economic status and the mental impairment while controlling for the life index. As a third example, we consider a dataset where Y denotes the annual food expenditure of a household and $\boldsymbol{X}$ the annual household income. The data are used to examine Ernst Engel's hypothesis which states that the proportion of income spent on food decreases with increasing income; see Section 2.5.3 for details.

When examining the relationship between an outcome and a set of predictors it is natural to model Y mathematically as a function of $\boldsymbol{X}$, say $Y = g(\boldsymbol{X})$, for some function $g(\cdot)$. However, since Y is a random variable, this model will often be inappropriate. In the study related to the smoking behaviour, the lung capacity is affected by many other factors in addition to age, gender, and smoking status. So in general, it will be impossible to find a function $g(\cdot)$, such that $Y = g(\boldsymbol{X})$ for all children, because it is reasonable to believe that two children of the same

age and gender and with the same smoking behaviour, can still have different lung capacities. Therefore, Y will often be modelled as a function of $\boldsymbol{X}$ by means of a statistical model; for example $Y = g(\boldsymbol{X}) + \varepsilon$, where ε denotes an unobservable random variable accounting for the remaining variability which cannot be explained by the data at hand.

Once an appropriate statistical model is established, the association between Y and $\boldsymbol{X}$ can be partially examined by investigating the function $g(\cdot)$ (partially because the outcome also depends on the unobservable ε). If interest lies in studying the effect of $\boldsymbol{X}$ on the full distribution of Y, a solution consists of imposing assumptions on ε, e.g. a normal distribution with mean zero and an unknown variance which can be estimated from the data. If such an assumption is infeasible or if there is no interest in describing the effect of $\boldsymbol{X}$ on the whole outcome distribution, the statistical model is often restricted to a summary measure of Y, for example the mean. If $\mathrm{E}(\varepsilon \mid \boldsymbol{X}) = 0$, the statistical model becomes $\mathrm{E}(Y \mid \boldsymbol{X}) = g(\boldsymbol{X})$, so that $g(\cdot)$ describes the relationship between the predictors and the mean outcome. In addition to the mean, quantiles are popular summary measure as well.

Restricting Y to a summary measure often allows describing the association between outcome and the predictors more concisely. This approach inevitably results in an information loss as compared to when describing the effect of $\boldsymbol{X}$ on the whole outcome distribution. However, selecting the summary measure carefully can often still provide an informative description of the underlying process. For the smoking behaviour example it can arguably be sufficient to describe the association between the smoking status and the *average* lung capacity of children of a given age and gender, instead of describing it for each child separately.

Selecting the appropriate summary measure is important and depends on various factors: the scale and shape of the outcome, the data at hand, the research question of interest, etc. The majority of the statistical models used by data analysts focus on the mean outcome because it often has a meaningful interpretation and it has interesting mathematical properties, among other arguments. However, the mean is not always the most interesting summary measure. When considering household income, for example, the majority of the population have a low to moderate income, while only a small fraction of the population has an extremely high income (it is often said that 20% of the population owns 80% of the wealth), resulting in a skewed distribution. These high incomes have a substantial effect on the mean so that that the mean is no longer representative for the majority of the population. The median income can arguably

be a more appropriate summary measure: what income does half of the population have at least and half of the population have at most? As another example, in the mental health study, the mental impairment outcome is ordinal on a 4-level scale, with categories 1 (not impaired), 2 (mild symptom formation), 3 (moderate symptom formation), and 4 (impaired). Here the mean has no straightforward interpretation because the difference between levels 1 and 2 is not necessarily the same as the difference between levels 2 and 3 or 3 and 4. Therefore, the 4-level scale could also have been coded as 0 (not impaired), 1 (mild symptom formation), 50 (moderate symptom formation), and 1000 (impaired). Instead of considering mean impairment one can focus on, for example, the probability that the impairment score does not exceed a particular level.

In this dissertation a novel statistical model for assessing the association between Y and $\boldsymbol{X}$ is developed where the summary measure is not related to the mean or quantiles, but to the probability that the outcome increases if the predictors change. This model forms an alternative to the popular statistical models which focus on the mean or quantiles and can be used if the outcome is ordinal, interval, or ratio-scale. The model, however, is not developed to replace these existing techniques; it should merely serve as an additional tool for data analysts. In the following section, we describe the setting more formally.

1.2 Introduction to the model

Let $f_{Y\boldsymbol{X}}$ and $f_{Y|\boldsymbol{X}}$ denote the density functions of the joint distribution and the conditional distribution of Y given $\boldsymbol{X}$, respectively. For a continuous outcome Y, most statistical methods focus on the conditional mean of Y given $\boldsymbol{X}$. For example, in linear regression models $\mathrm{E}(Y \mid \boldsymbol{X}) = \boldsymbol{Z}^T\boldsymbol{\beta}$, where $\boldsymbol{Z}$ is a p-dimensional vector with elements that are functions of the covariates $\boldsymbol{X}$ and where $\boldsymbol{\beta}$ is a p-dimensional parameter vector. Sometimes the complete conditional distribution of Y given $\boldsymbol{X}$ is specified, e.g. the normal regression model, allowing for likelihood-based inference. This is often replaced by some mild assumptions on the higher-order moments of the conditional distribution so that the likelihood is no longer defined and asymptotic semiparametric theories are required for inference, e.g. estimation based on generalized estimating equations (Liang and Zeger, 1986; Zeger and Liang, 1986), of which least squares is a well known example.

In this dissertation we propose models that quantify the effects of the covariates through the *probabilistic index* (PI), which, in the present setting, is defined as

$$\mathrm{P}\left(Y \preccurlyeq Y' \mid \boldsymbol{X}, \boldsymbol{X}'\right) := \mathrm{P}\left(Y < Y' \mid \boldsymbol{X}, \boldsymbol{X}'\right) + \frac{1}{2}\mathrm{P}\left(Y = Y' \mid \boldsymbol{X}, \boldsymbol{X}'\right), \tag{1.1}$$

where $(Y, \boldsymbol{X})$ and $(Y', \boldsymbol{X}')$ are independently and identically distributed (i.i.d.) with density $f_{Y\boldsymbol{X}}$. Although we use the term density, $(Y, \boldsymbol{X})$ can also be discrete or a combination of discrete and continuous variables. Furthermore, $\boldsymbol{X}$ may also be fixed by design, but for notational convenience we will treat it as a random vector.

When Y is continuous $\mathrm{P}\left(Y = Y' \mid \boldsymbol{X}, \boldsymbol{X}'\right) = 0$ and the PI simplifies to $\mathrm{P}\left(Y \preccurlyeq Y' \mid \boldsymbol{X}, \boldsymbol{X}'\right) = \mathrm{P}\left(Y < Y' \mid \boldsymbol{X}, \boldsymbol{X}'\right)$. Definition (1.1) is also meaningful and convenient when the outcome is discrete and it implies that $\mathrm{P}\left(Y \preccurlyeq Y' \mid \boldsymbol{X} = \boldsymbol{X}'\right) = 0.5$ for both continuous and discrete outcomes.

Although the PI requires the conditional distribution $f_{Y|\boldsymbol{X}}$, here we do not make full distributional assumptions on $f_{Y|\boldsymbol{X}}$. Apart from some minimal technical assumptions we only assume that $f_{Y|\boldsymbol{X}}$ satisfies

$$\mathrm{P}\left(Y \preccurlyeq Y' \mid \boldsymbol{X}, \boldsymbol{X}'\right) = m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}), \tag{1.2}$$

in which $m(\cdot)$ is a function with range $[0, 1]$ and $\boldsymbol{\beta}$ a p-dimensional parameter vector. In Chapter 2 more details will be given. To simplify notation, we sometimes drop the condition statement in the PI and write model (1.2) as $\mathrm{P}\left(Y \preccurlyeq Y'\right) = m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta})$. Equation (1.2) implies a restriction on $f_{Y|\boldsymbol{X}}$ that describes how the covariate $\boldsymbol{X}$ affects the outcome distribution in terms of the PI. If $\Omega \subseteq \mathbb{R}$ denotes the support of the distribution function of Y, then restriction (1.2) can be explicitly expressed as a function of $f_{Y|\boldsymbol{X}}$,

$$\int_{y\in\Omega}\int_{y'\in\Omega} \mathrm{I}\left(y \preccurlyeq y'\right) f_{Y|\boldsymbol{X}}(y \mid \boldsymbol{X}) f_{Y|\boldsymbol{X}}(y' \mid \boldsymbol{X}') d\lambda(y) d\lambda(y') = m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}), \tag{1.3}$$

where $\mathrm{I}\left(y \preccurlyeq y'\right) := \mathrm{I}\left(y < y'\right) + 0.5\mathrm{I}\left(y = y'\right)$, with $\mathrm{I}\left(\cdot\right)$ the indicator function and $\lambda(\cdot)$ the counting measure for discrete outcomes and the Lebesgue measure for continuous outcomes. Because $f_{Y|\boldsymbol{X}}$ is not fully specified by (1.3), model (1.2) represents a semiparametric model which we refer to as the *Probabilistic Index Model* (PIM). Inference on the parameter vector $\boldsymbol{\beta}$ thus requires semiparametric theory which is presented in Chapters 2 and 7.

1.3 The probabilistic index

When no covariates are present, the PI has been discussed already by many authors. To our knowledge, however, there is no unambiguous terminology used throughout the literature. Acion et al. (2006) use the term *probabilistic index*, while some authors even use the notation "$\mathrm{P}(Y < Y')$" in the title of their papers; see e.g. Enis and Geisser (1971); Halperin et al. (1987); Tian (2008); Zhou (2008); Kakade et al. (2008); Browne (2010).

Others have called it the *individual exceedance probability* (Senn, 1997), *stochastic improvement* (Lehmann, 1998), *common language effect size* (McGraw and Wong, 1992), *probability of superiority* (Grissom, 1994), and in engineering science, the *reliability from stress-strength relationships* (Church and Harris, 1970).

Probabilities of the form (1.2) also appear in the analysis of ROC curves. We refer to Pepe (2003) for an overview. The PI may be interpreted as the area under the curve (AUC) of the population probability-probability plot (PP-plot), which is defined as the curve $\{(p, F_1[F_2^{-1}(p)]) \mid p \in [0,1]\}$, where F_1 and F_2 are the distribution functions of $Y \mid X = x_1$ and $Y' \mid X' = x_2$, respectively. Suppose that Y is a continuous outcome and that F_1 and F_2 have the same support Ω. Then, for fixed covariates x_1 and x_2, the AUC becomes

$$\begin{aligned}\int_0^1 F_1[F_2^{-1}(p)]dp &= \int_{y\in\Omega} F_1(y)dF_2(y) = \mathrm{E}_{Y'|x_2}\left[\mathrm{P}_{Y|x_1}\left(Y \leq y \mid y = Y', x_1\right) \mid x_1, x_2\right] \\ &= \mathrm{P}_{YY'|x_1,x_2}\left(Y \leq Y' \mid x_1, x_2\right) = \mathrm{P}\left(Y \preccurlyeq Y' \mid x_1, x_2\right), \qquad (1.4)\end{aligned}$$

with $Y \mid x_1$ and $Y' \mid x_2$ independently distributed; we will often drop the subscript $YY' \mid x_1, x_2$ from the probability operator. In the context of ROC curves, we refer to Dodd and Pepe (2003) and Brumback et al. (2006), who proposed regression models for the AUC.

The PI is also closely related to *stochastic ordering*. A distribution F_1 is said to be *stochastically smaller* than F_2 if and only if $F_1(y) \geq F_2(y)$ for all $y \in \Omega$ and with strict inequality for a non-empty subset of Ω. When F_1 is stochastically smaller than F_2, equation (1.4) immediately implies that $\mathrm{P}(Y \preccurlyeq Y' \mid x_1, x_2) > 0.5$. The implication does not hold necessarily in the other direction. Stochastic ordering is thus a stronger property than PI > 0.5.

To illustrate the interpretation of the PI consider a two-sample setting where $Y \mid (X = E)$ denotes the outcome (e.g. blood pressure) under an experimental treatment and $Y' \mid (X' = P)$

the outcome under a placebo treatment. Assume that the outcome is continuous. The PI

$$\mathrm{P}\left(Y < Y' \mid X = E, X' = P\right), \tag{1.5}$$

then gives the probability that the outcome of a randomly chosen subject of the placebo group exceeds the outcome of a randomly chosen subject of the experimental group. Many authors have argued that the PI is well suited as an effect size measure, mainly because 1) it often has an informative and intuitive interpretation which can also be understood by non-statisticians, 2) it provides a general measure for the difference between two populations, and 3) it is robust and scale-free; see, for example, Wolfe and Hogg (1971); Laine and Davidoff (1996); Acion et al. (2006); Newcombe (2006a,b); D'Agostino et al. (2006); Zhou (2008); Tian (2008); Kieser et al. (2012). The PI has also been extended to multiple outcomes; see, for example, Buyse (2010).

Despite the useful features of the PI as an effect size measure, there are settings for which it can be a misleading summary measure. For example, with lower outcomes being better, the PI (1.5) does not necessarily give the probability that for a single patient, the experimental treatment is better than placebo; instead, it compares the outcomes of two randomly selected patients. This is discussed in more detail in Section 2.6. We refer to Hand (1992); Senn (2006, 2011, 2012) for interesting discussions on the limitations of the PI as an effect size measure.

1.4 Relationship with other statistical techniques

An interesting special case arises when X is a binary $(0, 1)$ design variable which refers to two populations. With $m(X, X'; \beta) = 0.5 + \beta(X' - X)$ and $\mathrm{P}(Y_0 \preccurlyeq Y_1) := \mathrm{P}(Y \preccurlyeq Y' \mid X = 0, X' = 1)$ model (1.2) becomes

$$\mathrm{P}\left(Y_0 \preccurlyeq Y_1\right) = 0.5 + \beta,$$

which is the parameter of interest in the Wilcoxon–Mann–Whitney (WMW) test (Wilcoxon, 1945; Mann and Whitney, 1947). In particular, under the general two-sample null hypothesis $H_0 : f_{Y_0} = f_{Y_1}$, the PI equals $\mathrm{P}\left(Y_0 < Y_1\right) = 0.5$ when the outcome variable is continuous, and thus $\beta = 0$. Under mild conditions, the WMW test is consistent against the alternative $H_1 : \mathrm{P}\left(Y_0 < Y_1\right) \neq 0.5$, (see, for example, Hollander and Wolfe, 1999), which is equivalent to $H_1 : \beta \neq 0$. The class of models presented here can be considered as extensions of the WMW setting. Just as a linear regression model and the *t*-tests for testing the covariate effects in the

linear model embed the two-sample *t*-test when the linear regression model has only one $0/1$ dummy covariate, so do the tests for testing covariate effects in the PIM result in a WMW-type test in a two-sample design. In a similar fashion, PIMs embed the Kruskal–Wallis (Kruskal and Wallis, 1952) and Friedman (Friedman, 1937) rank tests for the K-sample and randomized complete block designs, respectively. This is discussed in greater detail in Chapter 4.

A PIM can also be seen as an extension of the work of Dodd and Pepe (2003) and Brumback et al. (2006), who proposed models for the PI, but with the restriction that Y and Y' are continuous outcome variables that always belong to two different populations or treatment groups. In terms of our formulation this restriction could be expressed as $\boldsymbol{X}$ and $\boldsymbol{X}'$ being distinct in at least one component which is a binary indicator for two treatment groups. They thus provide a WMW-type test for comparing two treatment groups, while controlling for one or more covariates. The methods proposed in this dissertation does not impose such a particular restriction on the covariate vector $\boldsymbol{X}$. Moreover, they further improve on Dodd and Pepe (2003) and Brumback et al. (2006) by being directly applicable to both continuous and discrete outcome variables, and by providing a consistent estimator of the variance-covariance matrix of the parameter estimators so that no computationally intensive bootstrap procedure is required.

PIMs are closely related to the pairwise ordering regression models developed by Follmann (2002). The regression model considers the pairwise ordering of patients' clinical histories and the model parameters have an interpretation which is related to the PI. More specifically, Follmann (2002) models the probability that the (possibly multidimensional) outcome of a patient is better than the outcome of another patient, where *better* can be defined in various ways.

A PIM is also related to a Bradley–Terry model (BTM) for ordinal outcomes (Bradley and Terry, 1952; Bergsma et al., 2009, 2012). Instead of the PI, a BTM models the probability

$$\mathrm{P}\left(Y > Y'\right) - \mathrm{P}\left(Y < Y'\right).$$

Bergsma et al. (2009) provide full maximum likelihood estimators for BTMs when covariates are discrete. PIMs, however, are not restricted to an ordinal outcome or discrete predictors and estimation is based on semiparametric theory instead of maximum likelihood.

As pointed out by Van Keilegom (2012), a PIM can be considered as a transformation model (Carroll and Ruppert, 1988; Linton et al., 2008). Since

$$\mathrm{P}\left(Y \preccurlyeq Y' \mid \boldsymbol{X}, \boldsymbol{X}'\right) = \mathrm{E}\left(\mathrm{P}\left(Y \preccurlyeq y \mid y = Y', \boldsymbol{X}\right) \mid \boldsymbol{X}, \boldsymbol{X}'\right),$$

PIM (1.2) can be expressed as

$$h(Y') = m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}) + \varepsilon,$$

where $h(y) = \mathrm{P}\left(Y \preccurlyeq y \mid \boldsymbol{X}\right)$, $\mathrm{E}\left(\varepsilon \mid \boldsymbol{X}, \boldsymbol{X}'\right) = 0$, and $\mathrm{Var}\left(\varepsilon \mid \boldsymbol{X}, \boldsymbol{X}'\right)$ a function of $\boldsymbol{X}$ and $\boldsymbol{X}'$. However, estimating the unknown function $h(y)$ will be difficult, especially when many predictors are present. In Chapters 2 and 7 we avoid estimating $h(\cdot)$ by considering the PIM as a restricted moment model fitted to *pseudo-observations*.

1.5 An example

To demonstrate the scope and the interpretation of the models that form the topic of this dissertation, we first introduce an example data set. In psychiatry, the mental state of a patient is often assessed by means of patient-rated questionnaires. For example, the Beck Depression Inventory (BDI) (Beck et al., 1988) is a 21-item self-report rating inventory measuring characteristic attitudes and symptoms of depression. The BDI is the sum of the scores on the 21 items; it ranges from 0 to 63, with 63 indicating severe depression. Van den Eynde et al. (2008) reported on a study in which patients with a borderline personality disorder (BPD) were treated with quetiapine, which is an antipsychotic drug. It is of interest to know how the quetiapine dose affects the patients in terms of the BDI. As the design of the original study is quite complicated, only partial results from a simplified setting are presented. The outcome variable of interest is the improvement in BDI, which is calculated as the BDI at baseline minus the BDI at the end of the study and which we denote by BD. The regressor variable is the total dose of quetiapine measured in grams (DOSE). Figure 1.1 shows a scatter-plot of the data. We consider the PIM

$$\mathrm{P}\left(\mathrm{BD} \preccurlyeq \mathrm{BD}' \mid \mathrm{DOSE}, \mathrm{DOSE}'\right) = \mathrm{expit}\left[\beta(\mathrm{DOSE}' - \mathrm{DOSE})\right], \tag{1.6}$$

with $\mathrm{expit}(x) = \exp(x)/[1 + \exp(x)]$. Using the methods described in this dissertation, we find the estimate $\hat{\beta} = 0.1711$ with estimated standard error 0.0398. The p-value for testing $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$ is smaller than 0.0001, and thus at the 5% level of significance the null hypothesis is rejected. Since $\mathrm{expit}(\beta) = \mathrm{P}\left(\mathrm{BD} \preccurlyeq \mathrm{BD}' \mid \mathrm{DOSE}' = \mathrm{DOSE} + 1\right)$, we can conclude that patients treated with a larger dose of quetiapine are more likely to show a larger improvement. In particular, when the dose is increased by 5 grams, the estimated PI equals $\mathrm{expit}(5\hat{\beta}) = 70.2\%$; that is, when comparing a group of patients treated with quetiapine

with a group that received an extra 5 grams of quetiapine, we conclude that, with probability 70.2%, the BDI of a patient from the high-dose group shows a larger improvement than for a patient from the low-dose group.

The data could as well have been analyzed with a regression model which models the mean BD instead of the PI. However, the right-hand panels of Figure 1.1 demonstrate that the dose affects not only the mean outcome, but also the variance and the skewness of the BDI distribution. Hence, the mean as an effect size is arguably not the most appropriate choice. The PI, on the other hand, acts as a quantity that summarizes the covariate effect on the outcome distribution in a meaningful effect size measure and which is not restricted to the mean.

Another important characteristic of the example is that BDI is basically an ordinal score variable. Although the BDI scale counts 64 levels, the mean BDI does not necessarily have an unambiguous interpretation and regression techniques that focus on the conditional mean of the BDI are perhaps not to be recommended. The interpretation of the PI, on the other hand, applies to ordinal, interval, and ratio-scale outcomes, since it only requires an ordering among the outcomes. Note that by subtracting the BDI at the end of the study from the baseline BDI, it is implicitly assumed that BDI is interval-scaled instead of ordinal. This is just to avoid a complicated PIM in the introduction. If the BDI is truly ordinal, then the BDI at baseline can be included as a predictor without violating its ordinal nature. This is addressed in more detail in Section 2.5.4.

Note that, instead of a PIM, cumulative logit models (McCullagh, 1980), among other techniques, may be used for the analysis of ordinal data; see, for example, Agresti (2007) or Liu and Agresti (2005) for extensive overviews on methods for ordinal data.

1.6 Some other applications

There are many examples of outcome variables measured on an ordinal scale. In pain management, for example, the effectiveness of treatments is often measured on an ordinal scale. Patients may be asked to fill out a questionnaire with questions related to their (subjective) pain experience, resulting in a pain score that has an ordinal meaning. The scale of Turk et al. (1993), for example, is a $0 - 10$ rating scale. The analysis of pain scores with PIMs would

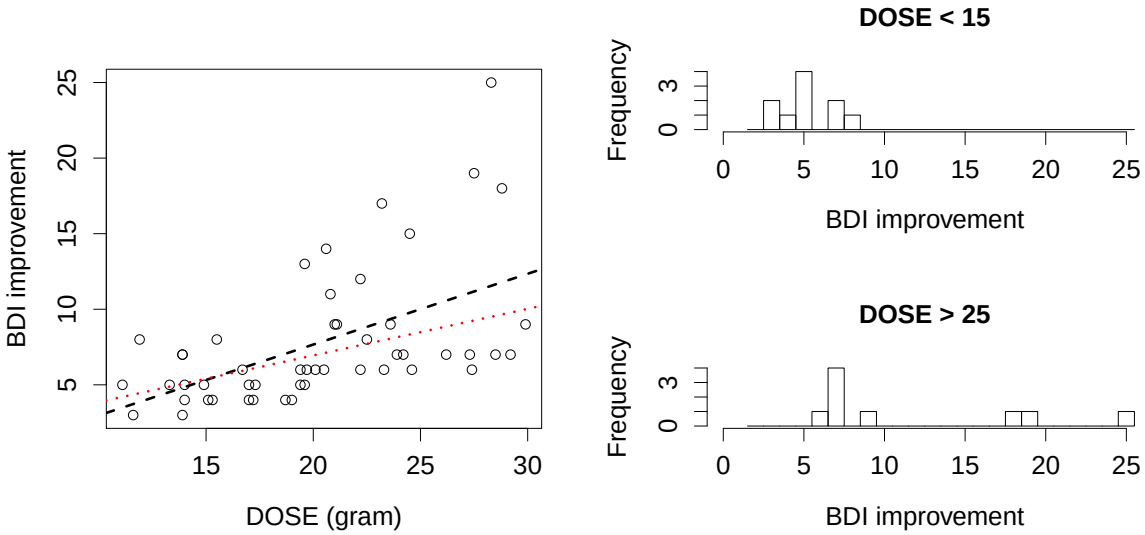

Figure 1.1: A scatter-plot of the BDI improvement versus the dose (left). The dashed and dotted lines show the linear regression model fits based on least squares and Huber's robust M-estimator, respectively. Histograms of the BDI improvements for small doses (top right) and large doses (bottom right).

result in probabilities that quantify how likely it is that the (reporting of) pain will decrease as a function of a set of covariates. Pain may also be measured on the visual analogue scale (VAS) of Wallerstein (1984), where the patient is presented with a horizontal line of 10 centimeter, anchored by the words "no pain" and "very severe pain" at the two ends. The patient is asked to mark the point on the line that best represents his or her level of pain at that moment. The distance, measured in millimeter, between the left-hand end of the line and the point marked by the patient is the numerical value used as a measure of pain. This is an example of a continuous outcome variable that may be interpreted as being ordinal, so that statements involving order comparisons, such as $\mathrm{P}\left(Y \preccurlyeq Y'\right)$, make sense. See Myles et al. (1999) for more details of the VAS scale.

PIMs may also turn out to be useful for analyzing genuine continuous outcome variables on a ratio scale for which classical regression models also seem to be appropriate. Beyerlein et al. (2008) observed that a child's body mass index (BMI) may be affected by several risk factors that, however, do not act only on the mean BMI. In particular, the skewness of the BMI distribution may change with covariate patterns. As illustrated in the BDI example, the PI summarizes the covariate effects on the shape of the outcome distribution, while retaining an informative

interpretation of the covariate effect sizes. Hence, PIMs could be a valuable alternative for BMI data. Beyerlein et al. (2008) suggested analyzing the BMI data with quantile regression methods, which forms another important class of models. It focuses on the τ^{th} quantile of the distribution of Y given $\boldsymbol{X}$, $Q_\tau(Y \mid \boldsymbol{X})$, say. Without the complete specification of the joint distribution of Y and $\boldsymbol{X}$, the τ^{th} quantile of the distribution of Y given $\boldsymbol{X}$ is modelled as $Q_\tau(Y \mid \boldsymbol{X}) = \boldsymbol{Z}^T\boldsymbol{\beta}_\tau$. These models are also semiparametric as the distribution of Y given $\boldsymbol{X}$ is not completely specified or parametrized. We refer to Koenker (2005) for an extensive overview on quantile regression.

If interest lies in assessing the effect of a regressor on several characteristics of the outcome distribution, quantile regression can arguably be the method of choice, since it allow for a rich analysis of the data by modelling multiple quantiles simultaneously. On the other hand, if it is desirable to summarize the effect of a predictor on the outcome distribution in terms of the PI, a PIM can be advocated.

These examples give already a flavour of the usefulness of the PIM. In particular when, the outcome variables are defined on an ordered scale, which can be discrete or continuous, for which the mean of the difference $Y - Y'$ does not have a proper interpretation as an effect size, but for which the PI does. More generally, the PIM may be the statisticians' method of choice whenever the PI is considered as a meaningful parameter for quantifying effect sizes. Of course, a PIM will not replace any of the existing statistical methods, it is merely a new tool in the statisticians' toolbox. Furthermore, since a PIM only considers a relative ordering among the outcomes, there can be more information loss as compared to models which exploit the richness of outcome in case it is interval or ratio-scale (van de Wiel, 2012).

1.7 Objectives and outline

The main objective of this dissertation is the development of a flexible semiparametric regression framework to model the probabilistic index: the probabilistic index model. Once this modelling framework is constructed, we are interested in studying the relationship between the PIM and a) regression methods and b) rank tests. Since the PIM is semiparametric, we will construct goodness-of-fit tools for assessing model validity. We illustrate the flexibility of a PIM by applying the model to complex genomic data. A final objective is the construction of

semiparametric efficient estimators for the PIM model parameters.

The dissertation is organized as follows.

In **Chapter 2** some popular statistical methods are reviewed and the PIM is formally introduced together with the parameter estimation and asymptotic distribution theory. The validity of the asymptotic approximations for finite samples is empirically evaluated in a simulation study and the interpretation of PIM is illustrated with several examples.

In **Chapter 3** the PIM is situated within the statistical landscape by exploring the relationships with several well-known statistical methods such as linear regression, the Cox proportional hazards model, AUC regression, rank regression, and the concordance index. The performance of a PIM and some of these methods is evaluated in a simulation study.

In **Chapter 4** the PIM methodology is situated within a broad class of rank tests. More specifically, relationships are established with the WMW, Kruskal– Wallis, and Friedman rank tests, among other. The performance of these methods relative to a PIM is evaluated in a simulation study. The PIM framework allows extending these popular rank tests to more complicated designs, while retaining an intuitive interpretation. This is illustrated with an example.

In **Chapter 5** goodness-of-fit (GOF) methods are developed for assessing the quality of the model fit of a PIM. The theoretical properties are evaluated in a simulation study and the GOF methods are illustrated on the examples of Chapters 2 and 3. Since well-established GOF methods do not apply well to PIMs, a new methodology is developed. Despite the relatively good performance, the proposed methodology should be considered as a first initiative in testing GOF of PIMs and still needs maturation.

In **Chapter 6** a case study is worked out in detail. More specifically, the PIM framework is used for the analysis of genomic reverse transcription quantitative polymerase chain reaction (RT-qPCR) data. A PIM will turn out to be appropriate for the analysis of such complex data while retaining a biologically relevant interpretation. In this chapter we summarize the most important characteristics of RT-qPCR data, without going into the biological details, so that the essence of the chapter should be understandable to data analysts unfamiliar with RT-qPCR.

In **Chapter 7** the estimation theory of Chapter 2 is revisited and new semiparameteric theory specifically constructed for PIMs is developed. A first initiative is taken towards deriving the efficient estimator and some of its properties are evaluated in restricted simulation study. This

chapter is mathematically more challenging as compared to the previous chapters and can be skipped by readers not interested in the technical details of the estimation theory.

In **Chapter 8** some conclusions are formulated and future research perspectives are discussed.

In **Appendix A** it is illustrated by examples how the R-package `pim` can be used to fit PIMs to data.

Contribution

Most of my dissertation is based on 3 published papers, one submitted paper, and a software package. The research is the result of a close collaboration with Olivier Thas, Lieven Clement, Stijn Vansteelandt, Karel Vermeulen, Nick Sabbe, and Jean–Pierre Ottoy. The idea of a PIM originates from O. Thas and it was my privilege to collaborate with all aforementioned researchers – all with their own expertise – in the development of several aspects of a PIM.

A more detailed listing of my contributions:

- Chapters 2 and 3 are based on Thas et al. (2012d). The basic construction of a PIM was the work of O. Thas. I have contributed to all aspects of the theory development and I have performed the simulation study, implemented the R code, and conducted all data analyses.

- Chapter 4 is based on a manuscript that is submitted and which is currently under review (authors: De Neve, J., Thas, O., and Ottoy, J.P.). I have been involved in all aspects of the research (model formulation, theory development, literature review, simulation studies, and data analysis). Most of the work was in close collaboration with O. Thas.

- Chapter 5 is based on De Neve et al. (2013a). I have taken the lead in this research, with guidance from O. Thas.

- Chapter 6 is based on De Neve et al. (2013c). The model formulation and the simulation set-up are a result of many discussions with L. Clement and O. Thas. I have taken the lead in writing the paper.

- Chapter 7 is the result of many discussions with, and internal reports from, O. Thas, S. Vansteelandt, and K. Vermeulen. However, the final form of Chapter 7 is from my own hand and goes beyond these internal reports.

- Appendix A is based on the R-package `pim` (De Neve and Sabbe, 2013). I have implemented all code used in this dissertation. N. Sabbe has further developed and professionalized the package and increased the applicability substantially.

Chapter 2

The probabilistic index model

The content of this chapter is primarily based on the results published in

Thas, O., De Neve, J., Clement, L., and Ottoy, J.P. (2012) Probabilistic index models (with discussion). *Journal of the Royal Statistical Society - Series B*, 74:623–671.

More specifically, it is based on sections 2, 3, 5, and 6 of the manuscript as well as on the discussions of Thomas Alexander Gerds, Stephen Senn, Lori E. Dodd, and Stijn Vansteelandt.

2.1 Outline

In Section 2.2 several popular statistical models are briefly reviewed with emphasis on estimation and interpretation. In Section 2.3 the probabilistic index model (PIM) is formally introduced, together with the parameter estimation and asymptotic theory. The validity of the asymptotic approximations for finite samples is empirically assessed in a simulation study in Section 2.4. In Section 2.5 several case studies are discussed and analyzed with a PIM as well as with more conventional statistical methods. In Section 2.6 some issues related to the interpretation of the probabilistic index (PI) are discussed and Section 2.7 gives the conclusions and discussion.

2.2 A brief review of some statistical models

This section merely serves as a concise overview of several popular statistical models. It is briefly illustrated how the model parameters can be estimated and interpreted. However, many known and important results are omitted because they fall outside the scope of this dissertation.

2.2.1 The linear regression model

The linear regression model (LRM) is defined as

$$Y = \boldsymbol{Z}^T\boldsymbol{\beta} + \varepsilon, \tag{2.1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$, $\mathrm{E}(\varepsilon) = 0$, $\boldsymbol{X} \perp\!\!\!\perp \varepsilon$, and $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \sigma^2\delta_{ij}$, i.e. the errors have a constant variance and are uncorrelated. The p-dimensional vector $\boldsymbol{Z}$ is a function of the d-dimensional covariate $\boldsymbol{X}$, for example if $d = 1$ with $\boldsymbol{X} = X$, then $\boldsymbol{Z}^T = (1, X, X^2)$ corresponds to a quadratic model with intercept which is linear in the parameters. Consider a random sample of i.i.d. observations $\{(Y_i, \boldsymbol{X}_i) \mid i = 1, \ldots, n\}$. The *Gauss–Markov theorem* states that the ordinary least squares (OLS) estimator, defined as

$$\hat{\boldsymbol{\beta}} := \mathrm{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(Y_i - \boldsymbol{Z}_i^T\boldsymbol{\beta}\right)^2,$$

is the best linear unbiased estimator of $\boldsymbol{\beta}$. Sometimes the additional assumption $\varepsilon \overset{d}{=} \mathrm{N}(0, \sigma^2)$ is required for obtaining finite sample distributional properties of $\hat{\boldsymbol{\beta}}$. However, most of these properties asymptotically hold even if this normality assumption is not fulfilled.

Model (2.1) implies

$$\mathrm{E}(Y \mid \boldsymbol{X}) = \boldsymbol{Z}^T\boldsymbol{\beta}. \tag{2.2}$$

To illustrate the interpretation consider a one-dimensional continuous predictor X and the LRM

$$\mathrm{E}(Y \mid X) = \beta_0 + \beta_1 X.$$

Then

$$\beta_1 = \mathrm{E}(Y \mid X = x + 1) - \mathrm{E}(Y \mid X = x), \tag{2.3}$$

i.e. β_1 quantifies the additive change in mean outcome if the predictor is increased by one unit. We refer to Kutner et al. (2004) for an extensive overview of linear models. Note, that if (Y, X)

and (Y', X') denote i.i.d. observations, equation (2.3) can be equivalently written as

$$\beta_1 = \mathrm{E}\left(Y' - Y \mid X = x, X' = x + 1\right).$$

Throughout this dissertation we often use this notation since it is closely related to the notation of effect sizes in terms of the PI. The interpretation of the LRM is also illustrated on an example dataset in Section 2.5.1.

2.2.2 The binary regression model

If the outcome variable Y is binary, when modelling, for example, success or failure, model (2.1) is no longer appropriate. Without loss of generally, let Y take values in $\{0, 1\}$, where $Y = 1$ denotes success and $Y = 0$ failure. The binary regression model is given by

$$g\left[\mathrm{P}\left(Y = 1 \mid \boldsymbol{X}\right)\right] = \boldsymbol{Z}^T\boldsymbol{\beta}, \tag{2.4}$$

where $g(\cdot)$ is a link function, required to assure that the predictions are within the unit interval. Usually the logit $g(x) = \log(x/[1-x])$ or probit $g(x) = \Phi^{-1}(x)$ link function is considered. The corresponding models are referred to as logistic and probit regression models, respectively. An estimator of $\boldsymbol{\beta}$ can be obtained by maximizing the likelihood, i.e.

$$\hat{\boldsymbol{\beta}} := \mathrm{argmax}_{\boldsymbol{\beta}} \prod_{i=1}^{n} g^{-1}(\boldsymbol{Z}_i^T\boldsymbol{\beta})^{Y_i} \left[1 - g^{-1}(\boldsymbol{Z}_i^T\boldsymbol{\beta})\right]^{1-Y_i}.$$

Consider a one-dimensional continuous predictor X and the logistic regression model

$$\mathrm{logit}[\mathrm{P}\left(Y = 1 \mid X\right)] = \beta_0 + \beta_1 X.$$

If we define the odds as $\mathrm{odds}\left(A\right) = \mathrm{P}\left(A\right)/[1 - \mathrm{P}\left(A\right)]$ for an event A, the interpretation of β_1 follows from

$$\beta_1 = \log\left(\frac{\mathrm{odds}\left(Y = 1 \mid X = x + 1\right)}{\mathrm{odds}\left(Y = 1 \mid X = x\right)}\right),$$

i.e. $\exp(\beta_1)$ quantifies the multiplicative change in odds on success if the predictor is increased by one unit. We refer to Hosmer and Lemeshow (2000) for an extensive overview of logistic regression models.

2.2.3 The restricted moment model

Both the linear and binary regression model can be embedded in a semiparametric restricted moment model (Chamberlain, 1987; Newey, 1988). Such a model is defined as

$$\mathrm{E}\left(Y \mid \boldsymbol{X}\right) = g^{-1}(\boldsymbol{Z}^T\boldsymbol{\beta}), \tag{2.5}$$

for which the model parameter $\boldsymbol{\beta}$ is estimated semiparametrically, i.e. apart from some mild regularity conditions, model (2.5) is the only restriction on the the conditional distribution of the outcome.

Model (2.5) can be extended to the setting with longitudinal or clustered data $\{(\boldsymbol{Y}_i, \boldsymbol{X}_i) \mid i = 1, \ldots, n\}$ with $\boldsymbol{Y}_i$ an m-vector of outcomes, $\boldsymbol{X}_i$ an $m \times d$ matrix of predictors, and $\mathrm{Cov}\left(\boldsymbol{Y}_i\right)$ not necessarily a diagonal matrix, indicating that the elements of $\boldsymbol{Y}_i$ can be correlated.

Let $(\boldsymbol{Y}_i, \boldsymbol{X}_i)$ be i.i.d., then the restricted moment model is expressed as

$$g\left[\mathrm{E}\left(\boldsymbol{Y}_i \mid \boldsymbol{X}_i\right)\right] = \boldsymbol{Z}_i\boldsymbol{\beta}, \tag{2.6}$$

with $\boldsymbol{Z}_i$ an $m \times p$ matrix. The model parameter $\boldsymbol{\beta}$ can be estimated by using *Generalized Estimating Equations* (GEE) (Liang and Zeger, 1986; Zeger and Liang, 1986). Let $\mathrm{Cov}\left(\boldsymbol{Y}_i\right) = \boldsymbol{A}_i^{1/2}\mathrm{Cor}\left(\boldsymbol{Y}_i\right)\boldsymbol{A}_i^{1/2}$, with $\boldsymbol{A}_i$ the diagonal matrix of marginal variances. The estimator $\hat{\boldsymbol{\beta}}$ is defined as the solution of

$$\sum_{i=1}^{n} \frac{\partial g(\boldsymbol{Z}_i\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\left(\boldsymbol{A}_i^{1/2}\boldsymbol{R}_i\boldsymbol{A}_i^{1/2}\right)^{-1}\left[\boldsymbol{Y}_i - g(\boldsymbol{Z}_i\boldsymbol{\beta})\right] = \boldsymbol{0}, \tag{2.7}$$

with $\boldsymbol{R}_i$ the *working correlation matrix* of $\boldsymbol{Y}_i$. This estimator is consistent and asymptotically normally distributed even if $\boldsymbol{R}_i$ is misspecified. Let $\tilde{\boldsymbol{\Sigma}}_i = \boldsymbol{A}_i^{1/2}\boldsymbol{R}_i\boldsymbol{A}_i^{1/2}$. A consistent sandwich estimator of $\mathrm{Cov}\left(\hat{\boldsymbol{\beta}}\right)$ is given by

$$\begin{aligned}
\hat{\boldsymbol{\Sigma}}_n(\hat{\boldsymbol{\beta}}) \;=\; & \left(\sum_{i=1}^{n} \frac{\partial g^T(\boldsymbol{Z}_i\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}}\tilde{\boldsymbol{\Sigma}}_i^{-1}\frac{\partial g(\boldsymbol{Z}_i\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}^T}\right)^{-1} \\
& \left(\sum_{i=1}^{n} \frac{\partial g^T(\boldsymbol{Z}_i\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}}\tilde{\boldsymbol{\Sigma}}_i^{-1}\left[\boldsymbol{Y}_i - g(\boldsymbol{Z}_i\hat{\boldsymbol{\beta}})\right]\left[\boldsymbol{Y}_i - g(\boldsymbol{Z}_i\hat{\boldsymbol{\beta}})\right]^T\tilde{\boldsymbol{\Sigma}}_i^{-1}\frac{\partial g(\boldsymbol{Z}_i\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}^T}\right) \\
& \left(\sum_{i=1}^{n} \frac{\partial g^T(\boldsymbol{Z}_i\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}}\tilde{\boldsymbol{\Sigma}}_i^{-1}\frac{\partial g(\boldsymbol{Z}_i\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}^T}\right)^{-1}.
\end{aligned}$$

See, for example, chapter 8 of Molenbergs and Verbeke (2005) or chapter 4 of Tsiatis (2006) for more details on restricted moments models and GEE.

2.2.4 The cumulative logit model

If the outcome variable Y is ordinal with k levels, the conditional mean $\mathrm{E}(Y \mid \boldsymbol{X})$ may not have a relevant interpretation. Consider the cumulative logit model

$$\operatorname{logit}\left[\mathrm{P}\left(Y \leq j \mid \boldsymbol{X}\right)\right] = \alpha_j + \boldsymbol{Z}^T\boldsymbol{\beta}, \quad j = 1, \ldots, k-1. \tag{2.8}$$

The likelihood function is constructed based on multinomial mass functions and is used to define the maximum likelihood estimator

$$\left(\hat{\alpha}_1, \ldots, \hat{\alpha}_{k-1}, \hat{\boldsymbol{\beta}}^T\right)^T := \operatorname{argmax}_{\alpha_j, \boldsymbol{\beta}} \prod_{i=1}^{n} \left(\prod_{j=1}^{k} \left[\operatorname{expit}(\alpha_j + \boldsymbol{Z}_i^T\boldsymbol{\beta}) - \operatorname{expit}(\alpha_{j-1} + \boldsymbol{Z}_i^T\boldsymbol{\beta})\right]^{\mathrm{I}(Y_i = j)} \right).$$

Let $\boldsymbol{X} = X$ and consider the model

$$\operatorname{logit}\left[\mathrm{P}\left(Y \leq j \mid X\right)\right] = \alpha_j + \beta X, \quad j = 1, \ldots, k-1.$$

The interpretation of β follows from

$$\beta = \log\left(\frac{\operatorname{odds}\left(Y \leq j \mid X = x+1\right)}{\operatorname{odds}\left(Y \leq j \mid X = x\right)}\right),$$

i.e. $\exp(\beta)$ quantifies the multiplicative change in odds that the outcome does not exceed a particular level if the predictor is increased by one unit. Model (2.8) is also referred to as the *proportional odds model*, since the effect of a covariate on the odds ratio is independent of the category j. These models can however be extended if the proportional odds assumption does not hold. We refer to Agresti (2010) for an extensive overview of proportional odds models. In Section 2.5.2 the interpretation of the cumulative logit model is illustrated on an example dataset.

2.2.5 The quantile regression model

For an ordinal, interval, or ratio-scale outcome Y, conditional quantiles are defined as

$$\mathrm{Q}_\tau\left(Y \mid \boldsymbol{X}\right) := \inf_y \left\{ y \mid F_{Y|\boldsymbol{X}}(y \mid \boldsymbol{X}) \geq \tau \right\}, \quad \tau \in (0, 1).$$

A linear quantile regression model (QRM) models this conditional quantile as a function of the covariates

$$\mathrm{Q}_\tau\left(Y \mid \boldsymbol{X}\right) = \boldsymbol{Z}^T\boldsymbol{\beta}_\tau. \tag{2.9}$$

The model parameter $\boldsymbol{\beta}_\tau$ depends on τ, which allows quantifying different effects sizes for quantile(s) of interest. If the covariate X affects different moments of the outcome distribution (e.g. mean, variance, skewness, etc.), the effect of X on, for example, the median ($\tau = 0.5$) is not necessarily the same as the effect on, for example, the 90% percentile ($\tau = 0.9$).

If we define a loss function as

$$\rho_\tau(u) := u[\tau - \mathrm{I}\,(u < 0)],$$

then a consistent estimator of $\boldsymbol{\beta}_\tau$ is given by

$$\hat{\boldsymbol{\beta}}_\tau := \mathrm{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho_\tau(Y_i - \boldsymbol{Z}_i^T \boldsymbol{\beta}).$$

The minimum can be found by reformulating the problem as a linear program and by using simplex methods; see chapter 6 in Koenker (2005) for more details.

Consider the QRM for a simple predictor X,

$$\mathrm{Q}_\tau\,(Y \mid X) = \beta_{0\tau} + \beta_{1\tau} X.$$

The interpretation of $\beta_{1\tau}$ follows from

$$\beta_{1\tau} = \mathrm{Q}_\tau\,(Y \mid X = x + 1) - \mathrm{Q}_\tau\,(Y \mid X = x)\,,$$

i.e. $\beta_{1\tau}$ quantifies the additive change in the τ^{th} quantile if the predictor is increased by one unit. We refer to Koenker (2005) for an extensive overview of quantile regression models. The interpretation of a QRM is also illustrated in Section 2.5.3.

2.3 The probabilistic index model

In this section we introduce a new model: the probabilistic index model.

2.3.1 Model formulation

Let $(Y, \boldsymbol{X})$ and $(Y', \boldsymbol{X}')$ be i.i.d., then a PIM is defined as

$$\mathrm{P}\,(Y \preccurlyeq Y' \mid \boldsymbol{X}, \boldsymbol{X}') = m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}), \tag{2.10}$$

where $m(\cdot)$ is a function with range $[0, 1]$ and subject to some smoothness conditions, which we address later. Here, $\boldsymbol{\beta}$ is the p-dimensional parameter vector. For the model to have a coherent interpretation, the function $m(\cdot)$ must satisfy $m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}) = 1 - m(\boldsymbol{X}', \boldsymbol{X}; \boldsymbol{\beta})$, i.e. $m(\cdot)$ must be *antisymmetric* about 1. This follows from $\mathrm{P}\left(Y \preccurlyeq Y' \mid \boldsymbol{X}, \boldsymbol{X}'\right) + \mathrm{P}\left(Y' \preccurlyeq Y \mid \boldsymbol{X}, \boldsymbol{X}'\right) = 1$, which also holds for discrete outcomes because of the definition of the PI as in (1.1). The antisymmetry condition implies that $m(\boldsymbol{X}, \boldsymbol{X}; \boldsymbol{\beta}) = 0.5$.

When $m(\cdot)$ does not satisfy the antisymmetry condition, the model may still be coherent when (2.10) is only defined for all $\boldsymbol{X} \prec \boldsymbol{X}'$ or $\boldsymbol{X} \preceq \boldsymbol{X}'$. The former refers to an order relation among the covariate patterns; so does the latter, but it includes $\boldsymbol{X} = \boldsymbol{X}'$. An order relation that we will use throughout the dissertation is the *lexicographical ordering*.

Definition 1 (lexicographical ordering)**.** *Let $\boldsymbol{X} = (X_1, X_2)^T$ and $\boldsymbol{X}' = (X_1', X_2')^T$ denote two vectors, then $\boldsymbol{X}$ is lexicographically smaller or equal to $\boldsymbol{X}'$, denoted as $\boldsymbol{X} \preceq_{lex} \boldsymbol{X}'$, if $X_1 < X_1'$, or if $X_1 = X_1'$ then $X_2 \leq X_2'$.*

By applying this definition recursively we can extend this order relation to vectors of dimension larger than two. See Fishburn (1974) for more information about the lexicographical order.

To avoid having to make throughout the dissertation always the distinction between models for which the antisymmetry condition holds and models for which an order restriction is imposed, we introduce the set $\mathcal{X}$ of elements $(\boldsymbol{X}, \boldsymbol{X}')$ for which model (2.10) is defined.

We use the notation $\mathcal{X}_0$ when no order restriction is imposed, i.e. $\mathcal{X}_0 := \{(\boldsymbol{X}, \boldsymbol{X}') \mid \forall\, \boldsymbol{X}, \boldsymbol{X}'\}$, further referred to as the no-order restriction. When the lexicographical order restriction is imposed, we use the notation $\mathcal{X}_{lex}$, i.e. $\mathcal{X}_{lex} := \{(\boldsymbol{X}, \boldsymbol{X}') \mid \boldsymbol{X} \preceq_{lex} \boldsymbol{X}'\}$.

To summarize, the PIM is defined as

$$\mathrm{P}\left(Y \preccurlyeq Y' \mid \boldsymbol{X}, \boldsymbol{X}'\right) = m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}) \quad \text{for all} \quad (\boldsymbol{X}, \boldsymbol{X}') \in \mathcal{X}. \tag{2.11}$$

This model expresses restrictions on the conditional distribution of Y given $\boldsymbol{X}$, but it does not fully specify this distribution, so that it is a semiparametric model. When $\mathrm{P}\left(Y = Y' \mid \boldsymbol{X}, \boldsymbol{X}'\right) = 0$ for all $(\boldsymbol{X}, \boldsymbol{X}') \in \mathcal{X}$ model (2.11) may just as well be defined in terms of $\mathrm{P}\left(Y < Y' \mid \boldsymbol{X}, \boldsymbol{X}'\right)$.

For the smoothness conditions, we impose the function $m(\cdot)$ to be related to a linear predictor, say, $\boldsymbol{Z}^T\boldsymbol{\beta}$ with $\boldsymbol{Z}$ a p-vector with elements that may depend on $\boldsymbol{X}$ and $\boldsymbol{X}'$. In many examples

$\boldsymbol{Z} = \boldsymbol{X}' - \boldsymbol{X}$ will be a convenient and meaningful choice. More specifically, we consider a model function $m(\cdot)$ of the form

$$m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}) = g^{-1}(\boldsymbol{Z}^T\boldsymbol{\beta}), \tag{2.12}$$

with $g(\cdot)$ a sufficiently smooth link function that maps $[0, 1]$ onto the range of $\boldsymbol{Z}^T\boldsymbol{\beta}$, which is usually the real line. In this dissertation we restrict $g(\cdot)$ to the logit, probit, and identity link. However, other link function can be used as well.

Although $\boldsymbol{Z}^T\boldsymbol{\beta}$ may include an intercept or an offset, we sometimes choose to write the linear predictor as $\beta_0 + \boldsymbol{Z}^T\boldsymbol{\beta}$, where β_0 is the intercept or offset. If the scope of the PIM includes $\boldsymbol{X} = \boldsymbol{X}'$ and the outcome is continuous, the offset β_0 must be set to a constant so that $\mathrm{P}(Y \preccurlyeq Y' \mid \boldsymbol{X} = \boldsymbol{X}') = 0.5$. The offset thus depends on the link function. For example, when $\boldsymbol{Z} = \boldsymbol{X}' - \boldsymbol{X}$ the offsets for the logit, probit, and identity link become $\beta_0 = 0$, $\beta_0 = 0$, and $\beta_0 = 0.5$, respectively.

2.3.2 Parameter estimation and inference

Define $\mathrm{I}(Y \preccurlyeq Y') := \mathrm{I}(Y < Y') + 0.5\mathrm{I}(Y = Y')$ in which $\mathrm{I}(Y < Y')$ and $\mathrm{I}(Y = Y')$ denote the usual indicator functions evaluated for the events $\{Y < Y'\}$ and $\{Y = Y'\}$, respectively. Since

$$\begin{aligned}
\mathrm{E}(\mathrm{I}(Y \preccurlyeq Y') \mid \boldsymbol{X}, \boldsymbol{X}') &= \mathrm{E}(\mathrm{I}(Y < Y') \mid \boldsymbol{X}, \boldsymbol{X}') + \frac{1}{2}\mathrm{E}(\mathrm{I}(Y = Y') \mid \boldsymbol{X}, \boldsymbol{X}') \\
&= \mathrm{P}(Y < Y' \mid \boldsymbol{X}, \boldsymbol{X}') + \frac{1}{2}\mathrm{P}(Y = Y' \mid \boldsymbol{X}, \boldsymbol{X}') \\
&= \mathrm{P}(Y \preccurlyeq Y' \mid \boldsymbol{X}, \boldsymbol{X}'),
\end{aligned}$$

and upon using (2.12), PIM (2.11) can be written as

$$\mathrm{E}(\mathrm{I}(Y \preccurlyeq Y') \mid \boldsymbol{X}, \boldsymbol{X}') = g^{-1}(\boldsymbol{Z}^T\boldsymbol{\beta}), \quad (\boldsymbol{X}, \boldsymbol{X}') \in \mathcal{X}. \tag{2.13}$$

If $\{(Y_i, \boldsymbol{X}_i) \mid i = 1, \ldots, n\}$ denotes a sample of n i.i.d. random observations, model formulation (2.13) suggests that the $\boldsymbol{\beta}$ parameter vector can be estimated using the set of *pseudo-observations* $I_{ij} := \mathrm{I}(Y_i \preccurlyeq Y_j)$ for all $i, j = 1, \ldots, n$ for which $(\boldsymbol{X}_i, \boldsymbol{X}_j) \in \mathcal{X}_n := \{(\boldsymbol{X}_i, \boldsymbol{X}_j) \mid i, j = 1, \ldots, n \wedge (\boldsymbol{X}_i, \boldsymbol{X}_j) \in \mathcal{X}\}$. In terms of the random sample, model (2.13) is equivalent to

$$\mathrm{E}(I_{ij} \mid \boldsymbol{X}_i, \boldsymbol{X}_j) = g^{-1}(\boldsymbol{Z}_{ij}^T\boldsymbol{\beta}), \quad (\boldsymbol{X}_i, \boldsymbol{X}_j) \in \mathcal{X}_n, \tag{2.14}$$

where $\boldsymbol{Z}_{ij}$ is function of $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$; e.g. $\boldsymbol{Z}_{ij} = \boldsymbol{X}_j - \boldsymbol{X}_i$.

In particular, model (2.13) resembles a semiparametric restricted moment model as discussed in Section 2.2.3, in which the conditional mean of the pseudo-observations is specified. In the spirit of generalized estimating equations (Liang and Zeger, 1986; Zeger and Liang, 1986), we propose to estimate the parameters by solving the estimating equations

$$\boldsymbol{U}_n(\boldsymbol{\beta}) = \sum_{(i,j)\in\mathcal{I}_n} \boldsymbol{U}_{ij}(\boldsymbol{\beta}) = \sum_{(i,j)\in\mathcal{I}_n} \boldsymbol{A}(\boldsymbol{Z}_{ij};\boldsymbol{\beta})[I_{ij} - g^{-1}(\boldsymbol{Z}_{ij}^T\boldsymbol{\beta})] = \mathbf{0}, \tag{2.15}$$

where $\mathcal{I}_n$ is the set of indexes (i,j) for which $(\boldsymbol{X}_i, \boldsymbol{X}_j) \in \mathcal{X}_n$, i.e. $\mathcal{I}_n := \{(i,j) \mid (\boldsymbol{X}_i, \boldsymbol{X}_j) \in \mathcal{X}_n\}$, and $\boldsymbol{A}(\boldsymbol{Z}_{ij};\boldsymbol{\beta})$ is a p-dimensional vector function of the regressors $\boldsymbol{Z}_{ij}$, subject to smoothness and regularity conditions which are discussed in greater detail in Chapter 7. For now, we often consider

$$\boldsymbol{A}(\boldsymbol{Z}_{ij};\boldsymbol{\beta}) = \frac{\partial g^{-1}(\boldsymbol{Z}_{ij}^T\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} V^{-1}(\boldsymbol{Z}_{ij};\boldsymbol{\beta}), \tag{2.16}$$

where $V(\boldsymbol{Z}_{ij};\boldsymbol{\beta}) = \nu^{-1}\mathrm{Var}\left(I_{ij} \mid \boldsymbol{Z}_{ij}\right)$ with ν a scale parameter. This choice corresponds to the generalized estimating equations (2.7) with the independent working correlation matrix. However, since $\mathrm{I}\left(Y_i \preccurlyeq Y_j\right)^2 = \mathrm{I}\left(Y_i < Y_j\right) + 0.25\mathrm{I}\left(Y_i = Y_j\right) = \mathrm{I}\left(Y_i \preccurlyeq Y_j\right) - 0.25\mathrm{I}\left(Y_i = Y_j\right)$, it follows that

$$\begin{aligned}\mathrm{Var}\left(I_{ij} \mid \boldsymbol{Z}_{ij}\right) &= \mathrm{E}(\mathrm{I}\left(Y_i \preccurlyeq Y_j\right)^2 \mid \boldsymbol{X}_i, \boldsymbol{X}_j) - \mathrm{E}(\mathrm{I}\left(Y_i \preccurlyeq Y_j\right) \mid \boldsymbol{X}_i, \boldsymbol{X}_j)^2 \\ &= g^{-1}(\boldsymbol{Z}_{ij}^T\boldsymbol{\beta})[1 - g^{-1}(\boldsymbol{Z}_{ij}^T\boldsymbol{\beta})] - \frac{1}{4}\mathrm{P}\left(Y_i = Y_j \mid \boldsymbol{X}_i, \boldsymbol{X}_j\right).\end{aligned}$$

If the outcome is continuous $\mathrm{P}\left(Y_i = Y_j \mid \boldsymbol{X}_i, \boldsymbol{X}_j\right) = 0$ and

$$V(\boldsymbol{Z}_{ij};\boldsymbol{\beta}) = \frac{1}{\nu} g^{-1}(\boldsymbol{Z}_{ij}^T\boldsymbol{\beta})[1 - g^{-1}(\boldsymbol{Z}_{ij}^T\boldsymbol{\beta})]. \tag{2.17}$$

For a discrete outcome $\mathrm{P}\left(Y_i = Y_j \mid \boldsymbol{X}_i, \boldsymbol{X}_j\right) \neq 0$, but for simplicity we still use the variance function (2.17). Let $\hat{\boldsymbol{\beta}}_n$ denote the estimator, defined as the root of (2.15), where for notational convenience we sometimes write $\hat{\boldsymbol{\beta}}$ suppressing the dependence on n.

The conditional mean in (2.13) does not refer to the mean of the conditional distribution of the outcome, but it refers to the conditional mean of the pseudo-observations. Moreover, the pseudo-observations are not mutually independent. For example,

$$\begin{aligned}\mathrm{Cov}[\mathrm{I}\left(Y_i \preccurlyeq Y_j\right), \mathrm{I}\left(Y_i \preccurlyeq Y_k\right)] &= \mathrm{P}[Y_i < \min(Y_j, Y_k)] + \frac{1}{2}\mathrm{P}\left(Y_i = Y_k \wedge Y_i < Y_j\right) + \\ &\quad \frac{1}{2}\mathrm{P}\left(Y_i = Y_j \wedge Y_j < Y_k\right) + \frac{1}{4}\mathrm{P}\left(Y_i = Y_j = Y_k\right) - \\ &\quad \mathrm{P}\left(Y_i \preccurlyeq Y_j\right)\mathrm{P}\left(Y_i \preccurlyeq Y_k\right).\end{aligned}$$

For a continuous outcome this simplifies to

$$\mathrm{Cov}[\mathrm{I}\left(Y_i < Y_j\right), \mathrm{I}\left(Y_i < Y_k\right)] = \mathrm{P}[Y_i < \min(Y_j, Y_k)] - \mathrm{P}\left(Y_i < Y_j\right)\mathrm{P}\left(Y_i < Y_k\right),$$

which is, in general, different from zero. The pseudo-observations possess a *cross-correlation* structure, i.e. if two pseudo-observations share a common outcome, they will in general not be independent. Consider three independent outcomes Y_i, $i = 1, 2, 3$, then $\mathrm{I}\left(Y_1 \preccurlyeq Y_2\right)$ is associated with $\mathrm{I}\left(Y_1 \preccurlyeq Y_3\right)$, $\mathrm{I}\left(Y_3 \preccurlyeq Y_1\right)$, $\mathrm{I}\left(Y_2 \preccurlyeq Y_3\right)$, $\mathrm{I}\left(Y_3 \preccurlyeq Y_2\right)$, and $\mathrm{I}\left(Y_2 \preccurlyeq Y_1\right)$.

Despite the close relationship between our method of estimation and generalized estimating equations, the asymptotic distributional properties of the estimator $\hat{\boldsymbol{\beta}}$ do not follow immediately from these theories, because the cross-correlation results in a different dependence structure than, for example, block independence as in clustered or longitudinal data. Neither does the correlation structure resembles the decaying associations as in time-series or as in spatial processes.

Lemmas 1 and 2 following state that the pseudo-observations possess the *sparse correlation* structure, as introduced by Lumley and Mayer-Hamblett (2003). This result makes their semi-parametric theory directly applicable to our setting. Theorems 1 and 2 following summarize the most important distribution theory results for the PIM.

Note that when $g^{-1}(\boldsymbol{Z}_{ij}^T\boldsymbol{\beta}) = 1 - g^{-1}(\boldsymbol{Z}_{ji}^T\boldsymbol{\beta})$, i.e. the model is antisymmetric about one, and for $\boldsymbol{A}(\cdot)$ as in (2.16), it follows that $\boldsymbol{A}(\boldsymbol{Z}_{ij};\boldsymbol{\beta}) = -\boldsymbol{A}(\boldsymbol{Z}_{ji};\boldsymbol{\beta})$. Furthermore, the solution of (2.15) for the no-order restriction is identical to the solution for a lexicographical order restriction. This follows from

$$\begin{aligned}
&\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{A}(\boldsymbol{Z}_{ij};\boldsymbol{\beta})[I_{ij} - g^{-1}(\boldsymbol{Z}_{ij}^T\boldsymbol{\beta})] = \mathbf{0} \\
\Leftrightarrow\ &\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\boldsymbol{A}(\boldsymbol{Z}_{ij};\boldsymbol{\beta})[I_{ij} - g^{-1}(\boldsymbol{Z}_{ij}^T\boldsymbol{\beta})] + \sum_{j=1}^{n-1}\sum_{i=j+1}^{n}\boldsymbol{A}(\boldsymbol{Z}_{ij};\boldsymbol{\beta})[I_{ij} - g^{-1}(\boldsymbol{Z}_{ij}^T\boldsymbol{\beta})] + \\
&\sum_{i=1}^{n}\boldsymbol{A}(\boldsymbol{Z}_{ii};\boldsymbol{\beta})[0.5 - 0.5] = \mathbf{0} \\
\Leftrightarrow\ &\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\boldsymbol{A}(\boldsymbol{Z}_{ij};\boldsymbol{\beta})[I_{ij} - g^{-1}(\boldsymbol{Z}_{ij}^T\boldsymbol{\beta})] = \mathbf{0}.
\end{aligned}$$

The last step follows from $I_{ij} = 1 - I_{ji}$, $g^{-1}(\boldsymbol{Z}_{ij}^T\boldsymbol{\beta}) = 1 - g^{-1}(\boldsymbol{Z}_{ji}^T\boldsymbol{\beta})$, and $\boldsymbol{A}(\boldsymbol{Z}_{ij};\boldsymbol{\beta}) = -\boldsymbol{A}(\boldsymbol{Z}_{ji};\boldsymbol{\beta})$. Therefore, when the PIM satisfies the antisymmetry condition, the lexicographical

ordering is preferred over the no-order restriction, for only half of the pseudo-observations are needed.

We start with the defining *sparse correlation* in the context of pseudo-observations. A more general definition can be found in Lumley and Mayer-Hamblett (2003).

Definition 2 (Sparse correlation)**.** *Let* $\{I_{ij} \mid (i,j) \in \mathcal{I}_n\}$ *denote a set of pseudo-observations. For each pseudo-observation* I_{ij}*, a set of pairs of indices* $\{S_{ij} \mid (i,j) \in \mathcal{I}_n\}$ *is defined such that* $(k,l) \notin S_{ij}$ *and* $(i,j) \notin S_{kl}$ *implies* I_{ij} *and* I_{kl} *are independent. Let* M_{nij} *denote the number of pairs in* S_{ij}*, let* $M_n = \max_{(i,j)\in\mathcal{I}_n} M_{nij}$ *and let* m_n *denote the size of the largest subset* T *such that* $S_{ij} \cap S_{kl} = \oslash$ *for all pairs* $(i,j),(k,l) \in T$*. Then the set of pseudo-observations is called sparsely correlated if we can choose* $\{S_{ij} \mid (i,j) \in \mathcal{I}_n\}$ *so that* $M_n m_n = O(|\mathcal{I}_n|)$*, with* $|\mathcal{I}_n|$ *the number of pseudo-observations.*

In the following lemmas we demonstrate that the pseudo-observations are sparsely correlated when the no-order restriction or the lexicographical order restriction are imposed.

Lemma 1 (Sparse correlation: no-order restriction)**.** *The no-ordered pseudo-observations possess the sparse correlation structure.*

Proof. Each pseudo-observation I_{ij} with $(i,j) \in \mathcal{I}_n = \{(i,j) \mid i \neq j \text{ and } i,j = 1,\ldots,n\}$ is correlated with $4n-7$ other pseudo-observations. Indeed, let $k = 1,\ldots,n$ with $k \neq i$ and $k \neq j$, then I_{ij} is correlated with $I_{ik}, I_{kj}, I_{ki}, I_{jk}, I_{ji}$, and with itself. Thus $M_n = M_{nij} = 4n-6$. The largest set of pseudo-observations that are mutually independent consists of any I_{ij} and all other I_{kl} with i,j,k,l mutually distinct. The size of this set is thus $\lfloor n/2 \rfloor$, i.e. the largest integer not larger than $n/2$. Suppose that n is even. Then

$$M_n m_n = (4n-6)n/2 = 2n^2 - 3n = O(n^2).$$

Since $O(|\mathcal{I}_n|) = O(n^2)$, the lemma holds for n even. Similarly, when n is odd, $M_n m_n = (4n-6)\lfloor n/2 \rfloor = O(n^2) = O(|\mathcal{I}_n|)$. □

Lemma 2 (Sparse correlation: lexicographical order restriction)**.** *The lexicographical ordered pseudo-observations possess the sparse correlation structure.*

Proof. The lexicographical pseudo-observations I_{ij} for which $\boldsymbol{X}_i \preceq_{lex} \boldsymbol{X}_j$ can be obtained by sorting the data $(Y, \boldsymbol{X})$ based on lexicographical ordering on $\boldsymbol{X}$ and then considering the

pseudo-observations I_{ij} with $(i,j) \in \mathcal{I}_n = \{(i,j) \mid i < j \text{ and } i,j = 1,\ldots,n\}$. Each pseudo-observation I_{ij} is correlated with $2n-4$ other pseudo-observations. Indeed I_{ij} is correlated with

- I_{ik} where $k = i+1,\ldots,n$ and $k \neq j$,
- I_{kj} where $k = 1,\ldots,j-1$ and $k \neq i$,
- I_{ki} where $k = 1,\ldots,i-1$,
- I_{jk} where $k = j+1,\ldots,n$,

and with itself. Thus $M_n = M_{nij} = 2n-3$. The largest set of pseudo-observations that are mutually independent consists of any I_{ij} and all other I_{kl} with $i < j, k < l$ mutually distinct. The size of this set is thus $\lfloor n/2 \rfloor$. Suppose that n is even. Then

$$M_n m_n = (2n-3)n/2 = n^2 - 3n/2 = O(n^2).$$

Since $O(|\mathcal{I}_n|) = O(n^2)$, the lemma holds for n even. Similarly, when n is odd, $M_n m_n = (2n-3)\lfloor n/2 \rfloor = O(n^2) = O(|\mathcal{I}_n|)$. □

Since the following two theorems are special cases of theorem 7 of Lumley and Mayer-Hamblett (2003) we will omit the proof. We define the *true $\boldsymbol{\beta}$ parameter*, $\boldsymbol{\beta}_0$, as the unique solution of

$$\lim_{n\to\infty} \mathrm{E}\left\{ |\mathcal{I}_n|^{-1} \sum_{(i,j)\in\mathcal{I}_n} \boldsymbol{A}(\boldsymbol{Z}_{ij};\boldsymbol{\beta}) \left[\mathrm{I}\left(Y_i \preccurlyeq Y_j\right) - g^{-1}(\boldsymbol{Z}_{ij}^T\boldsymbol{\beta}) \right] \right\} = \mathbf{0}. \tag{2.18}$$

The regularity conditions in the statement of Theorem 1 imply the existence of $\boldsymbol{\beta}_0$. For a more detailed discussion on the regularity and smoothness conditions under which the asymptotic properties of the estimator as given in Theorems 1 and 2 hold, we refer to Chapter 7, where asymptotics for PIMs are developed without relying on the sparse correlation theory.

Theorem 1 (Asymptotic normality)**.** *Consider the PIM (2.14) with predictors $\boldsymbol{Z}_{ij}$ taking values in a bounded subset of $\mathbb{R}^p$. We make the following assumptions:*

A1 the pseudo-observations are sparsely correlated as in Lemma 1 or Lemma 2;

A2 the link function $g(\cdot)$ and the variance function $V(\cdot)$ have three continuous derivatives;

A3 the true parameter $\boldsymbol{\beta}_0$, as defined by (2.18), is in the interior of a convex parameter space;

A4 there exist a vector $\boldsymbol{W}$ and positive definite matrix $\boldsymbol{T}$ such that

$$|\mathcal{I}_n|^{-1} \sum_{(i,j)\in\mathcal{I}_n} \boldsymbol{Z}_{ij} \xrightarrow{p} \boldsymbol{W} \text{ and } |\mathcal{I}_n|^{-1} \sum_{(i,j)\in\mathcal{I}_n} \boldsymbol{Z}_{ij}\boldsymbol{Z}_{ij}^T \xrightarrow{p} \boldsymbol{T};$$

A5 $\limsup n^{-1}\text{Var}\left(\sum_{(i,j)\in\mathcal{I}_n} I_{ij}\right) > 0.$

If $\hat{\boldsymbol{\beta}}_n$ is defined as the solution of (2.15) with $\boldsymbol{A}(\cdot)$ as in (2.16), then, as $n \to \infty$, $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)$ converges in distribution to a multivariate Gaussian distribution with zero mean and some positive definite variance-covariance matrix $\boldsymbol{\Sigma}$.

Theorem 2 (Consistent variance estimator)**.** *Under the regularity conditions of Theorem 1, the variance-covariance matrix $\boldsymbol{\Sigma}$ can be consistently estimated by the sandwich estimator $n\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_n}$, where*

$$\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_n} = \left(\sum_{(i,j)\in\mathcal{I}_n} \frac{\partial \boldsymbol{U}_{ij}(\hat{\boldsymbol{\beta}}_n)}{\partial \boldsymbol{\beta}^T}\right)^{-1} \left(\sum_{(i,j)\in\mathcal{I}_n}\sum_{(k,l)\in\mathcal{I}_n} \phi_{ijkl}\boldsymbol{U}_{ij}(\hat{\boldsymbol{\beta}}_n)\boldsymbol{U}_{kl}^T(\hat{\boldsymbol{\beta}}_n)\right) \left(\sum_{(i,j)\in\mathcal{I}_n} \frac{\partial \boldsymbol{U}_{ij}(\hat{\boldsymbol{\beta}}_n^T)}{\partial \boldsymbol{\beta}}\right)^{-1},$$

for which the indicator ϕ_{ijkl} is defined as $\phi_{ijkl} = 1$ if I_{ij} and I_{kl} are correlated and $\phi_{ijkl} = 0$ otherwise.

In summary, for large finite n and if the PIM (2.14) holds, the distribution of $\hat{\boldsymbol{\beta}}_n$ is approximately multivariate normal with mean $\boldsymbol{\beta}_0$ and a variance-covariance matrix which can be estimated by $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_n}$. In the following section this approximation for finite sample sizes is evaluated in a simulation study.

In the remainder of this dissertation we often drop the subscript n in $\hat{\boldsymbol{\beta}}_n$ and $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_n}$. The distributional properties of Theorems 1 and 2 allow to estimate and to construct confidence intervals for the model parameters $\boldsymbol{\beta}$. Furthermore, null hypotheses of the form

$$H_0 : \beta_i = \beta_0,$$

for some fixed constant β_0 (not to be confused with the intercept) can be tested with Wald-type statistics

$$\frac{\hat{\beta}_i - \beta_0}{\sqrt{\left(\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}\right)_{ii}}},$$

which have an approximate standard normal null distribution, where $(\boldsymbol{B})_{ii}$ denotes the i^{th} diagonal element of a matrix $\boldsymbol{B}$. Similarly, a general linear null hypothesis

$$H_0 : \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{\beta}_0,$$

with $\boldsymbol{\beta}_0$ a vector of constants and $\boldsymbol{C}$ a contrast matrix, can be tested with the generalized quadratic from

$$\left(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)^T \left(\boldsymbol{C}\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}\boldsymbol{C}^T\right)^- \left(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right),$$

which has a chi-squared null distribution with degrees of freedom equal to the rank of $\boldsymbol{C}\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}\boldsymbol{C}^T$ and where $\boldsymbol{B}^-$ denotes a generalized inverse of a square matrix $\boldsymbol{B}$.

2.4 Simulation study

A generic problem in the set-up of simulation studies for the evaluation of semiparametric methods is that a semiparametric model encompasses a large class of parametric data generating models. For example, the semiparametric linear regression model (SLRM)

$$\mathrm{E}\left(Y \mid X\right) = \alpha X, \tag{2.19}$$

encompasses the infinite class of data generating models $\{Y = \alpha X + \varepsilon \mid \mathrm{E}\left(\varepsilon \mid X\right) = 0\}$. For example, the parametric normal linear model with $\varepsilon \overset{d}{=} \mathrm{N}(0, \sigma^2)$, results in the SLRM (2.19). However, this also holds for linear models with a t-distributed error, i.e. $\varepsilon \overset{d}{=} t_f$, or for other zero mean error distributions. Thus many parametric models can result in the same semiparametric model. A similar argumentation holds for the semiparametric PIMs. However, in general the relationship between PIMs and data generating models is often more complicated than that of the SLRM.

We have chosen to generate data with a normal linear regression model and an exponential generalized linear model. Each of these parametric models embed a PIM; this is discussed in detail in Sections 3.2 and 3.3. Table 2.1 summarizes the relationships between three parametric data generating models and the induced PIM.

All computations have been performed with the R software (R Core Team, 2012) and all PIMs are defined for the lexicographical order restriction and are equivalent to the no-order restriction as they all satisfy the antisymmetry condition; see Section 2.3.2 for more information.

Table 2.1: Three parametric data generating models and their corresponding PIM

data generating model	embedded PIM	relationship
$Y \mid X$	$\mathrm{P}\left(Y \preccurlyeq Y' \mid X, X'\right)$	
$\mathrm{N}(\alpha X, \sigma^2)$	$\Phi[\beta(X' - X)]$	$\beta = \alpha/\sqrt{2\sigma^2}$
$\mathrm{N}(\alpha X, \sigma^2 X)$	$\Phi[\beta(X' - X)/\sqrt{X' + X}]$	$\beta = \alpha/\sigma$
$\mathrm{Exponential}[\exp(\alpha X)]$	$\mathrm{expit}[\beta(X' - X)]$	$\beta = -\alpha$

2.4.1 The normal linear model

We consider the model

$$Y_i = \alpha X_i + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2.20}$$

where $\varepsilon_i \mid X_i \stackrel{d}{=} \mathrm{N}[0, \sigma_\varepsilon^2(X_i)]$. Sample sizes of $n = 25$, $n = 50$, and $n = 200$ are considered. The predictor X takes equally spaced values in the interval $[0.1, u]$ where $u = 1$ or 10. The parameter α equals 1 or 10. Table 2.2 presents the results for a constant standard deviation, i.e. $\sigma_\varepsilon(X) = \sigma$, with $\sigma = 1$ or $\sigma = 5$. From Table 2.1 the corresponding PIM is given by

$$\Phi^{-1}\left[\mathrm{P}\left(Y \preccurlyeq Y' \mid X, X'\right)\right] = \beta(X' - X), \tag{2.21}$$

where $\beta = \alpha/\sqrt{2\sigma^2}$.

For each setting, 1000 Monte Carlo simulation runs are used for the empirical investigation of the distribution of the semiparametric estimator of β. The semiparametric estimator of Section 2.3.2 is denoted by $\hat{\beta}$, and it is further referred to as the PIM estimator. Table 2.2 shows for each simulation setting the true β parameter and the average of the simulated estimates. The latter is an approximation of the true mean of the estimator. The table also reports the average of the simulated sandwich variance estimates, which is an approximation of the expectation of the sandwich estimator, and the sample variance of the 1000 estimates $\hat{\beta}$, which is an approximation of the true variance of the estimator $\hat{\beta}$. The empirical coverages of 95% confidence intervals for β are also reported.

From Table 2.2 we conclude that the PIM estimator of β is nearly unbiased, particularly for sample sizes of 50 and more. A similar conclusion holds for the sandwich variance estimator. The empirical coverages of the 95% confidence intervals are relatively close to their nominal

level for sample sizes of 50 and more, except for $\alpha = 10$ and $u = \sigma = 1$ because of an underestimation of the variance. Figure 2.1 shows the normal QQ-plots of 1000 simulated estimates when $\alpha = 1$, $u = 10$, and $\sigma = 1$ or $\sigma = 5$, respectively. For $\sigma = 1$ there is a substantial deviation from normality in the right tail for $n = 25$ and $n = 50$ due to tied estimates. This can be explained as follows. For this setting, the outcome Y has low variability and the pseudo-observations $\mathrm{I}(Y \preccurlyeq Y')$ will be similar to $\mathrm{I}(X \preccurlyeq X')$. Since the covariates are fixed by design and lexicographically ordered, $\mathrm{I}(X \preccurlyeq X')$ is fixed for the 1000 simulation runs. Consequently, the pseudo-observations will be very similar over the simulation runs, explaining the tied estimates in the right tail. As the sample size increase, the distribution of the design points becomes more dense, so that, in general, $\mathrm{I}(Y \preccurlyeq Y')$ will be different from $\mathrm{I}(X \preccurlyeq X')$ because of the random error and hence the tied estimates disappear. This is illustrated in the top right panel of Figure 2.1. Furthermore, for $\sigma = 5$ the normal approximation of the estimator is reasonable, even for sample size $n = 25$.

Table 2.3 shows the results of simulations of heteroscedastic data with $\sigma_\varepsilon(X) = \sigma\sqrt{X}$ and $X > 0$, where $\sigma = 1$ or $\sigma = 5$. The corresponding PIM is given by (see Table 2.1)

$$\Phi^{-1}\left[\mathrm{P}\left(Y \preccurlyeq Y' | X, X'\right)\right] = \beta \frac{X' - X}{\sqrt{X' + X}},$$

where $\beta = \alpha/\sigma$. Similar conclusions hold as for the homoscedastic setting, except for $\alpha = 10$, $u = 1$, and $\sigma = 1$ for which the sandwich estimator consistently overestimates the true variance. This is a consequence of many tied estimates, similar as in the top left panel of Figure 2.1 (results not shown). Furthermore, the coverages of the confidence intervals are slightly worse as compared to the homoscedastic setting.

2.4.2 The exponential model

Let $Y_i \mid X_i \stackrel{d}{=} \text{Exponential}[\gamma(X_i)]$ with

$$\gamma(X_i) = \exp(\alpha X_i), \quad i = 1, \ldots, n. \tag{2.22}$$

Sample sizes of $n = 25$, $n = 50$, and $n = 200$ are considered. The predictor X takes equally spaced values in the interval $[0.1, u]$ where $u = 1$ or 10 and α takes on the value 0.1 or -2. The corresponding PIM is

$$\text{logit}\left[\mathrm{P}\left(Y \preccurlyeq Y' \mid X, X'\right)\right] = \beta(X' - X), \tag{2.23}$$

Table 2.2: Simulation results for the normal linear homoscedastic model, based on 1000 Monte Carlo runs. β is the true parameter, $\mathrm{Av}(\hat{\beta})$ the average of the β PIM estimates, $\mathrm{Var}(\hat{\beta})$ the sample variance of the simulated $\hat{\beta}$, $\mathrm{Av}(\hat{S}_{\hat{\beta}})$ the average of the sandwich variance PIM estimates, EC the empirical coverage of a 95% confidence interval for β.

α	u	σ	β	$\mathrm{Av}(\hat{\beta})$	$\mathrm{Var}(\hat{\beta})$	$\mathrm{Av}(\hat{S}_{\hat{\beta}})$	EC (%)
				$n = 25$			
1	1	1	0.707	0.736	0.33900	0.27877	92.0
1	1	5	0.141	0.130	0.32438	0.27008	92.8
1	10	1	0.707	0.721	0.00990	0.01184	93.0
1	10	5	0.141	0.149	0.00332	0.00248	90.2
10	1	1	7.071	7.309	1.55061	1.22519	85.7
10	1	5	1.414	1.463	0.40365	0.29884	88.7
				$n = 50$			
1	1	1	0.707	0.736	0.16640	0.15048	92.9
1	1	5	0.141	0.148	0.14905	0.14542	93.5
1	10	1	0.707	0.714	0.00615	0.00634	94.4
1	10	5	0.141	0.147	0.00148	0.00139	93.4
10	1	1	7.071	7.224	0.78701	0.67363	89.1
10	1	5	1.414	1.465	0.18646	0.16191	92.5
				$n = 200$			
1	1	1	0.707	0.716	0.03803	0.03942	95.3
1	1	5	0.141	0.145	0.04048	0.03817	94.8
1	10	1	0.707	0.709	0.00179	0.00170	94.3
1	10	5	0.141	0.141	0.00037	0.00036	95.6
10	1	1	7.071	7.110	0.19105	0.17489	93.2
10	1	5	1.414	1.427	0.04400	0.04308	95.0

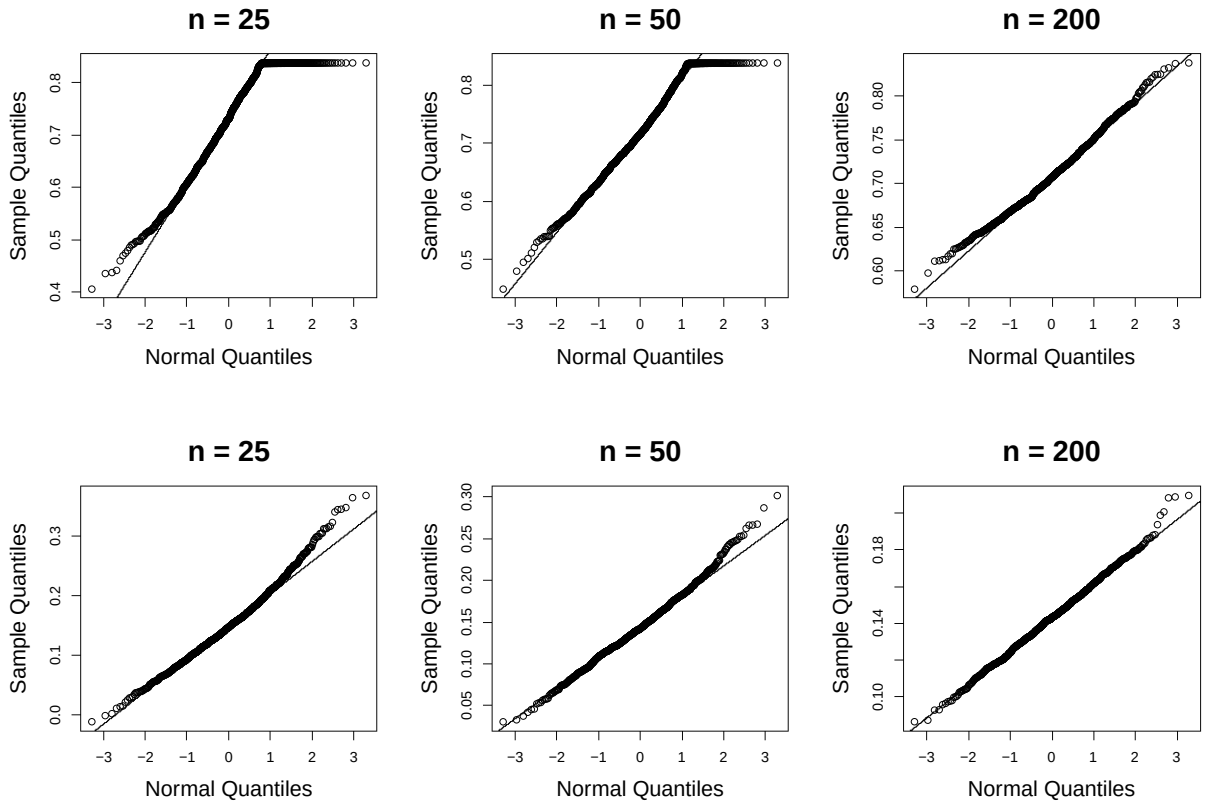

Figure 2.1: QQ-plots of $\hat{\beta}$ associated with PIM (2.21) for $\alpha = 1$, $u = 10$, $n = 25$, 50, and 200, and $\sigma = 1$ (upper panels) or $\sigma = 5$ (lower panels)

Table 2.3: Simulation results for the normal linear heteroscedastic model, based on 1000 Monte Carlo runs. β is the true parameter, $\mathrm{Av}(\hat{\beta})$ the average of the β PIM estimates, $\mathrm{Var}(\hat{\beta})$ the sample variance of the simulated $\hat{\beta}$, $\mathrm{Av}(\hat{S}_{\hat{\beta}})$ the average of the sandwich variance PIM estimates, EC the empirical coverage of a 95% confidence interval for β.

α	u	σ	β	$\mathrm{Av}(\hat{\beta})$	$\mathrm{Var}(\hat{\beta})$	$\mathrm{Av}(\hat{S}_{\hat{\beta}})$	EC (%)
				$n = 25$			
1	1	1	1	1.052	0.34771	0.27673	91.2
1	1	5	0.2	0.192	0.31399	0.26122	92.8
1	10	1	1	1.045	0.05487	0.03584	90.1
1	10	5	0.2	0.206	0.02317	0.01884	92.2
10	1	1	10	9.268	0.50991	1.75345	93.9
10	1	5	2	2.080	0.46761	0.32145	88.4
10	10	5	2	2.088	0.13541	0.10231	85.5
				$n = 50$			
1	1	1	1	1.032	0.17125	0.15259	92.9
1	1	5	0.2	0.210	0.14692	0.14205	94.4
1	10	1	1	1.025	0.02554	0.01967	90.0
1	10	5	0.2	0.208	0.01086	0.01034	94.4
10	1	1	10	9.410	0.22462	0.95066	96.0
10	1	5	2	2.063	0.20438	0.17953	92.5
10	10	5	2	2.046	0.06469	0.05539	91.4
				$n = 200$			
1	1	1	1	1.010	0.03905	0.04005	95.1
1	1	5	0.2	0.204	0.03891	0.03740	95.2
1	10	1	1	1.006	0.00568	0.00557	93.6
1	10	5	0.2	0.198	0.00271	0.00275	95.8
10	1	1	10	9.576	0.04093	0.26446	97.1
10	1	5	2	2.016	0.05006	0.04843	94.1
10	10	5	2	2.007	0.01548	0.01465	94.1

where $\beta = -\alpha$. Table 2.4 gives the results when model (2.23) is analyzed with the semiparametric PIM theory, resulting in $\hat{\beta}$.

From Table 2.4 we conclude that the PIM estimator of β and the sandwich variance estimator are nearly unbiased, particularly for sample sizes of 50 and more. The empirical coverages of the 95% confidence intervals are close to their nominal level for sample sizes of 50 and more.

Table 2.4: Simulation results for the exponential model, based on 1000 Monte Carlo runs. β is the true parameter, $\mathrm{Av}(\hat{\beta})$ the average of the β PIM estimates, $\mathrm{Var}(\hat{\beta})$ the sample variance of the simulated $\hat{\beta}$, $\mathrm{Av}(\hat{S}_{\hat{\beta}})$ the average of the sandwich variance PIM estimates, EC the empirical coverage of a 95% confidence interval for β.

α	u	σ	β	$\mathrm{Av}(\hat{\beta})$	$\mathrm{Var}(\hat{\beta})$	$\mathrm{Av}(\hat{S}_{\hat{\beta}})$	EC (%)
				$n = 25$			
-2	1	1	-2	-2.226	1.19067	0.89060	90.4
0.1	10	1	0.1	0.110	0.00902	0.00630	91.1
				$n = 50$			
-2	1	1	-2	-2.083	0.54166	0.47159	93.7
0.1	10	1	0.1	0.103	0.00337	0.00333	95.0
				$n = 200$			
-2	1	1	-2	-2.023	0.12394	0.12220	94.7
0.1	10	1	0.1	0.098	0.00090	0.00087	94.6

2.5 Examples

To illustrate the interpretation of the PIM we present several examples. In Section 2.5.1 we present the data analysis for a continuous outcome and two predictors showing an interaction effect. The example of Section 2.5.2 has an ordinal outcome variable and two predictors without an interaction effect. An example data set with a continuous outcome and a single continuous regressor is presented in Section 2.5.3. Unless stated otherwise, all PIMs are defined for the lexicographical order restriction, but since they all satisfy the antisymmetry condition, they are equivalent to the no-ordered restricted PIMs; see Section 2.3.2 for more information. To avoid

lengthy formulas we sometimes drop the conditioning in the PI notation. All hypothesis tests are performed at the 5% level of significance and all computations are performed with the R software (R Core Team, 2012).

2.5.1 The childhood respiratory disease study

The Childhood Respiratory Disease Study (CRDS) is a longitudinal study following the pulmonary function in children. We only consider the part of this study provided by Rosner (1999). The outcome is the forced expiratory volume (FEV), which is an index of pulmonary function measured as the volume of air expelled after one second of constant effort. Along with FEV (litres), the AGE (years), HEIGHT (inches), SEX, and SMOKING status (1 if the child smokes, 0 if the child does not smoke) are provided for 654 children of ages $3 - 19$. See Rosner (1999, p. 41) for more information. The primary focus is on the analysis of the effect of smoking status on the pulmonary function. The left and middle panels of Figure 2.2 display the FEV as a function of the AGE and SMOKING status; note that all very young children are non-smokers. It is believed that age may be a potential confounder, and thus the effect of smoking on FEV should be adjusted for age. Figure 2.3 shows nonparametric density estimates of the FEV distributions for several combinations of smoking status and age. This figure suggests an interaction effect between age and smoking status. In addition to the smoking effect, it is also of interest to quantify the effect of age.

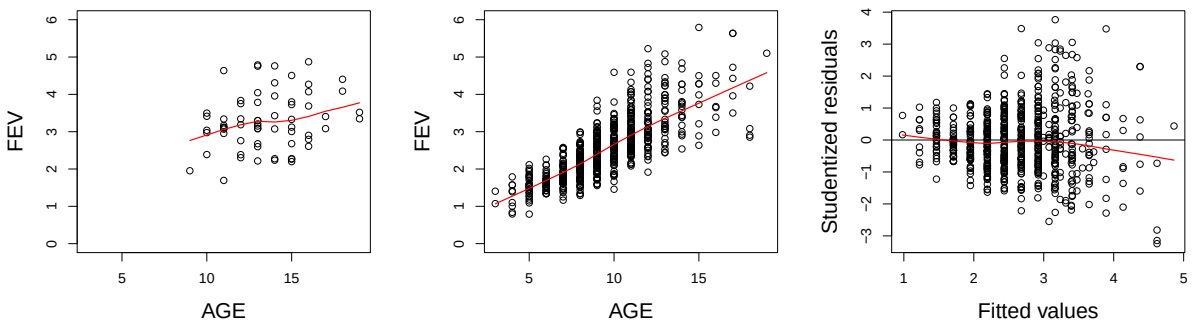

Figure 2.2: Left: FEV as a function of AGE for smokers. Middle: FEV as a function of AGE for non-smokers. Right: Studentized residuals in function of the fitted values of LRM (2.24). The solid line corresponds to a nonparametric estimate of the regression function.

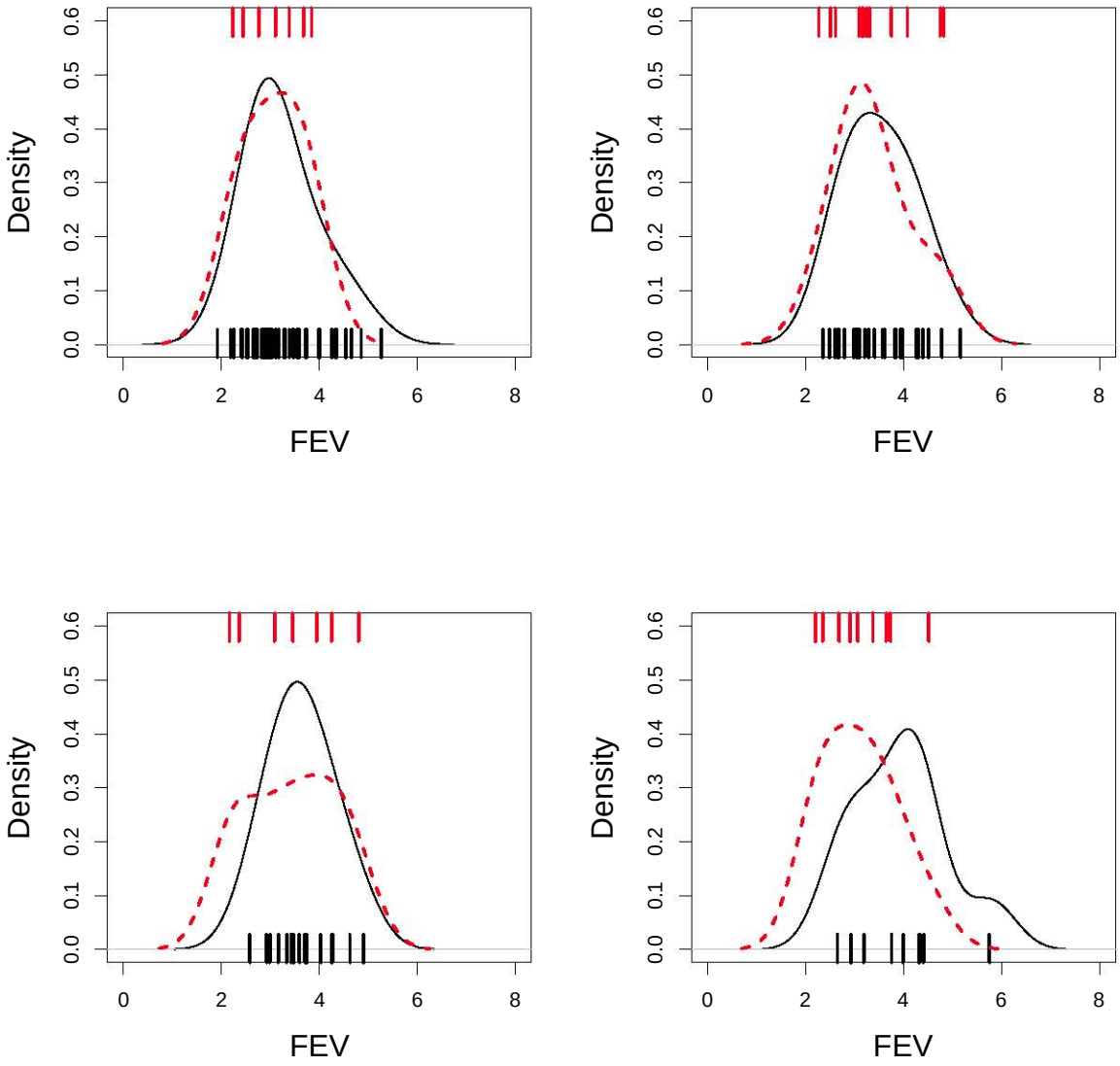

Figure 2.3: Kernel density estimates of the FEV distributions for non-smokers (solid line —) and smokers (dashed line − − −) of age 12 years (top left), 13 years (top right), 14 years (bottom left), and 15 years (bottom right). The densities are estimated using a Gaussian kernel with a bandwidth of 0.5. Beneath (non-smokers) and above (smokers) each kernel density plot is a rug plot to identity better the individual sample observations that are used for the density estimation.

For comparison purposes we first analyze the data with a linear regression model (LRM) with mean

$$\mathrm{E}(\text{FEV} \mid \text{AGE}, \text{SMOKE}) = \alpha_0 + \alpha_1 \text{AGE} + \alpha_2 \text{SMOKE} + \alpha_3 \text{AGE} * \text{SMOKE}. \tag{2.24}$$

Table 2.5 gives the model fit with ordinary least squares (OLS). Since the residual plot in the right panel of Figure 2.2 indicates non-constant variance of the error, we also fit the regression model using weighted least squares (WLS) (see Table 2.5). The weights were obtained by fitting the absolute residuals of OLS in a linear regression model with the fitted values of OLS as the regressor.

Table 2.5: Results of the OLS and WLS fits of model (2.24) and the results of the fit of the PIM (2.25)

	Estimate	SE	p-value
	LRM OLS		
intercept (α_0)	0.25	0.083	0.002
AGE (α_1)	0.24	0.008	< 0.001
SMOKE (α_2)	1.94	0.41	< 0.001
AGE*SMOKE (α_3)	-0.16	0.03	< 0.001
	LRM WLS		
intercept (α_0)	0.32	0.054	< 0.001
AGE (α_1)	0.24	0.007	< 0.001
SMOKE (α_2)	1.84	0.51	< 0.001
AGE*SMOKE (α_3)	-0.15	0.03	< 0.001
	PIM		
AGE (β_1)	0.61	0.03	< 0.001
SMOKE (β_2)	5.31	1.04	< 0.001
AGE*SMOKE (β_3)	-0.46	0.08	< 0.001

With WLS the age-specific effect of smoking on the mean level of FEV, upon using the notation

$\mathrm{E}\,(Y' - Y \mid \boldsymbol{X}, \boldsymbol{X}')$, is estimated as

$$\begin{aligned} & \hat{\mathrm{E}}\,(\text{FEV}' - \text{FEV} \mid \text{AGE} = \text{AGE}', \text{SMOKE} = 0, \text{SMOKE}' = 1) \\ = \;& \hat{\mathrm{E}}(\text{FEV} \mid \text{AGE}, \text{SMOKE} = 1) - \hat{\mathrm{E}}(\text{FEV} \mid \text{AGE}, \text{SMOKE} = 0) \\ = \;& \hat{\alpha}_2 + \hat{\alpha}_3 \text{AGE} = 1.84 - 0.15\text{AGE}. \end{aligned}$$

If we consider, for example, the age categories 12, 13, 14, and 15 of Figure 2.3, the effect of smoking on the mean FEV is estimated by 0.01, -0.14, -0.29, and -0.45, respectively, and the 95% confidence intervals are given by $[-0.19, 0.21]$, $[-0.33, 0.05]$, $[-0.49, -0.09]$, and $[-0.68, -0.21]$. Thus for the ages of 14 and 15 years the mean FEV of non-smokers is significantly larger. These estimated effects and corresponding confidence intervals are also displayed in the left panel of Figure 2.4, for ages ranging from 11 years to 16 years. This figure illustrates that the negative effect of smoking on the average lung capacity becomes more pronounced as the age increases.

When the smoking status is fixed and for an age difference of one year, the mean FEV is estimated to change by

$$\begin{aligned} & \hat{\mathrm{E}}\,(\text{FEV}' - \text{FEV} \mid \text{AGE}' = \text{AGE} + 1, \text{SMOKE} = \text{SMOKE}') \\ = \;& \hat{\alpha}_1 + \hat{\alpha}_3 \text{SMOKE} = 0.24 - 0.15\text{SMOKE}. \end{aligned}$$

For non-smokers this effect is estimated by 0.24 with a 95% confidence interval of $[0.22, 0.25]$, whereas for smokers this is 0.082 with 95% confidence interval $[0.009, 0.156]$. Figure 2.3 suggests that, while controlling for age, smoking does not only affect the mean. The effect of smoking is also visible in higher-order moments. The probabilistic index is well suited to quantify effects that do not act on a single moment of the outcome distribution.

We consider the probabilistic index model with interaction:

$$\begin{aligned} \mathrm{logit}[\mathrm{P}\,(\text{FEV} \preccurlyeq \text{FEV}')] \;=\; & \beta_1(\text{AGE}' - \text{AGE}) + \beta_2(\text{SMOKE}' - \text{SMOKE}) \\ & + \beta_3(\text{AGE}' * \text{SMOKE}' - \text{AGE} * \text{SMOKE}). \end{aligned} \tag{2.25}$$

The model has no intercept, because, when $\text{AGE}' = \text{AGE}$ and $\text{SMOKE}' = \text{SMOKE}$, the model must give $\mathrm{P}\,(\text{FEV} \preccurlyeq \text{FEV}') = \mathrm{expit}(0) = 0.5$. The parameter estimates are presented in Table 2.5. For a fixed age, the probability of having a smaller FEV for a randomly selected non-

smoker as compared to a randomly selected smoker, is estimated as

$$\begin{aligned} & \hat{\mathrm{P}}\left(\mathrm{FEV} \preccurlyeq \mathrm{FEV}' \mid \mathrm{AGE} = \mathrm{AGE}', \mathrm{SMOKE} = 0, \mathrm{SMOKE}' = 1\right) \\ = \; & \operatorname{expit}(\hat{\beta}_2 + \hat{\beta}_3 \mathrm{AGE}) = \operatorname{expit}(5.31 - 0.46 \mathrm{AGE}). \end{aligned}$$

This illustrates that the effect of smoking on the PI depends on the age. For the age categories 12, 13, 14, and 15 from Figure 2.3, the estimated probabilities of having a smaller FEV for a non-smoker are 46%, 35%, 26%, and 18%, respectively, with 95% confidence intervals $[35\%, 57\%]$, $[26\%, 45\%]$, $[18\%, 35\%]$, and $[11\%, 27\%]$. Thus if the age increases it becomes less likely that smokers have a larger FEV than non-smokers. This effect is significant at the 5% level of significance for ages of 13, 14, and 15 years. These estimated effects and corresponding confidence intervals are also displayed in the right panel of Figure 2.4, for ages ranging from 11 years to 16 years. This figure illustrates that the negative effect of smoking on the lung capacity becomes more pronounced as the age increases, but instead of focussing on the change in mean FEV, the effect size is quantified based on the PI.

On the other hand, for two randomly selected children with the same smoking status, but with a difference of one year in age, the probability that the oldest has a larger FEV is estimated as

$$\begin{aligned} & \hat{\mathrm{P}}\left(\mathrm{FEV} \preccurlyeq \mathrm{FEV}' \mid \mathrm{AGE}' = \mathrm{AGE} + 1, \mathrm{SMOKE} = \mathrm{SMOKE}'\right) \\ = \; & \operatorname{expit}(\hat{\beta}_1 + \hat{\beta}_3 \mathrm{SMOKE}) = \operatorname{expit}(0.61 - 0.46 \mathrm{SMOKE}). \end{aligned}$$

For non-smokers this probability is estimated by $\operatorname{expit}(0.61) = 65\%$ while for smokers this drops to $\operatorname{expit}(0.15) = 54\%$. The 95% confidence intervals are given by $[63\%, 66\%]$ and $[50\%, 57\%]$, respectively.

The PIM, just like any parametric or semiparametric regression model, expresses restrictions on the joint distribution of the outcome and the covariates. As for any other regression model, it is important to assess the validity of the model for a given data set. Residual plots are used to assess the goodness-of-fit (GOF) of the LRM (2.24). For PIMs, however, new GOF-tools are needed and this forms the content of Chapter 5. Therefore, the discussion of assessing the GOF of PIM (2.25) is postponed to Chapter 5.

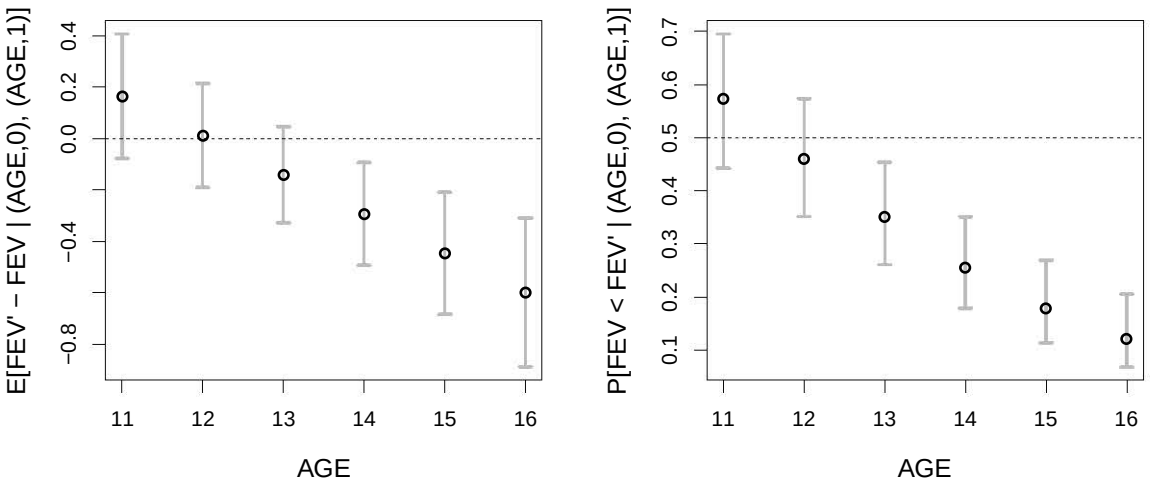

Figure 2.4: Left: estimated effect of smoking on mean FEV, while keeping AGE fixed and as a function of AGE for model (2.24), i.e. $\hat{\mathrm{E}}\left(\mathrm{FEV}' - \mathrm{FEV} \mid \mathrm{AGE} = \mathrm{AGE}', \mathrm{SMOKE} = 0, \mathrm{SMOKE}' = 1\right)$. Right: estimated effect of smoking on the PI, while keeping AGE fixed and as a function of AGE for model (2.25), i.e. $\hat{\mathrm{P}}\left(\mathrm{FEV} \preccurlyeq \mathrm{FEV}' \mid \mathrm{AGE} = \mathrm{AGE}', \mathrm{SMOKE} = 0, \mathrm{SMOKE}' = 1\right)$. The grey bars indicate the pointwise 95% confidence intervals and the dashed line $(- - -)$ represents the absence of a smoking effect.

2.5.2 The mental health study

The Mental Health Study (MHS) is a study of mental health for a random sample of 40 adult residents of Alachua County, Florida. See Agresti (2007, p. 185) for more information. The outcome is Mental Impairment (MI), which is ordinal with categories 1 (well), 2 (mild symptom formation), 3 (moderate symptom formation), and 4 (impaired). Along with the mental impairment, the life index (LI) and the socioeconomic status (SES) are reported. The SES is a binary variable coded as 0 (low SES) and 1 (high SES). The LI is a composite measure that quantifies the severity and the number of important life events such as birth of a child, death in family, divorce, etc. One of the objectives of the study is to assess whether the SES has an effect on MI. As it is believed that the LI may be a potential confounder, we consider it as a covariate. As the average MI score has no clear interpretation, a cumulative logit model can be considered

$$\operatorname{logit}[\mathrm{P}\,(\mathrm{MI} \leq j \mid \mathrm{SES}, \mathrm{LI})] = \mu_j + \alpha_1 \mathrm{SES} + \alpha_2 \mathrm{LI}, \quad j = 1, 2, 3. \tag{2.26}$$

Table 2.6 presents the parameter estimates based on the `MASS` R package (Venables and Ripley, 2002).

The cumulative logit model (2.26) gives no significant effect of SES at the 5% level of significance ($p = 0.07$). However, there is a significant effect of the life index on the cumulative logit ($p = 0.008$). Since

$$\frac{\hat{\mathrm{odds}}(\mathrm{MI} \leq j \mid \mathrm{SES}, \mathrm{LI} + 1)}{\hat{\mathrm{odds}}(\mathrm{MI} \leq j \mid \mathrm{SES}, \mathrm{LI})} = \exp(\hat{\alpha}_2) = \exp(-0.32) = 0.73,$$

it follows that the odds that the mental impairment score is not larger than a particular level decreases by an estimated factor 0.73 if the LI is one unit higher. The corresponding 95% confidence interval is $[0.56, 0.91]$. The cumulative logit model (2.26) can be further extended so that the covariate effect on the odds ratios for the events $MI \leq j$ depends on the level j. Since this more complex model does not fit significantly better (results not shown, $p = 0.68$), we keep the model with the proportional odds assumption.

Now consider the PIM

$$\operatorname{logit}\left[\mathrm{P}\left(\mathrm{MI} \preccurlyeq \mathrm{MI}'\right)\right] = \beta_1(\mathrm{SES}' - \mathrm{SES}) + \beta_2(\mathrm{LI}' - \mathrm{LI}). \tag{2.27}$$

The parameter estimates are presented in Table 2.6. The PIM analysis shows that, at the 5% level of significance, SES and LI have significant effects on the MI score in terms of the PI.

Table 2.6: Results of the fits of the cumulative logit model (2.26) and the PIM (2.27)

Parameter	Estimate	SE	p-value
	cumulative logit model		
intercept 1 (μ_1)	-0.28	0.64	0.66
intercept 2 (μ_2)	1.21	0.66	0.07
intercept 3 (μ_3)	2.21	0.72	0.002
SES (α_1)	1.11	0.61	0.07
LI (α_2)	-0.32	0.12	0.008
	PIM		
SES (β_1)	-0.74	0.34	0.03
LI (β_2)	0.20	0.07	0.006

Moreover, since

$$\hat{\mathrm{P}}\left(\mathrm{MI} \preccurlyeq \mathrm{MI}' \mid \mathrm{SES} = 0, \mathrm{SES} = 1, \mathrm{LI} = \mathrm{LI}'\right) = \mathrm{expit}(\hat{\beta}_1) = \mathrm{expit}(-0.74) = 32\%,$$

we conclude that, for two randomly chosen persons with equal LI, someone with a high SES has an estimated probability of 32% to have a larger MI score than someone with a low SES and a 95% confidence interval is given by $[20\%, 48\%]$. People with a low SES are thus more likely to be mentally impaired than others with a high SES, while all having the same LI.

Similarly, since

$$\hat{\mathrm{P}}\left(\mathrm{MI} \preccurlyeq \mathrm{MI}' \mid \mathrm{SES} = \mathrm{SES}', \mathrm{LI}' = \mathrm{LI} + 1\right) = \mathrm{expit}(\hat{\beta}_2) = \mathrm{expit}(0.2) = 55\%,$$

we conclude that, for two randomly chosen persons with the same SES and a unit difference of LI, the person with the lowest LI will have a lower MI score with an estimated probability of 55%, with a 95% confidence interval of $[51\%, 59\%]$. Thus, the larger the LI, the more likely someone is to be mentally impaired.

The PIM (2.27) can also be extended so that the effects of SES and LI on the PI do not only depend on the differences $\mathrm{SES}' - \mathrm{SES}$ and $\mathrm{LI}' - \mathrm{LI}$, but also on the covariates themselves. For

example,

$$\begin{aligned} \text{logit}\left[\text{P}\left(\text{MI} \preccurlyeq \text{MI}'\right)\right] &= \beta_1(\text{SES}' - \text{SES}) + \beta_2(\text{LI}' - \text{LI}) \\ &\quad + \beta_3\text{SES} + \beta_4\text{LI}, \end{aligned} \tag{2.28}$$

which is well defined for the strict lexicographical order restriction SES $<$ SES$'$, or SES $=$ SES$'$ and LI $<$ LI$'$. However, this more complex model did not fit significantly better, in the sense that the null hypothesis

$$H_0 : \beta_3 = \beta_4 = 0,$$

was not rejected at the 5% level of significance ($p = 0.77$). Note that the addition of β_3SES and β_4LI in model (2.28) is another way of introducing an interaction effect. However, it may not be consistent with a data generating model as it is difficult to think of a PIM that satisfies the antisymmetry condition and that reduces to (2.28) for the strict lexicographical ordering. This illustrates that the flexibility of the PIM framework may lead to models which are not coherent with an underlying data generating model. Future research should focus on establishing a solid connection between a PIM and data-generating mechanisms. However, for certain predictor values, these models can perhaps still provide a good approximation. GOF tools are useful for assessing model adequacy.

Similar as for the previous example, we postpone the GOF assessment of PIM (2.27) to Chapter 5.

2.5.3 The food expenditure study

The food expenditure data set contains data on the food expenditure (FE, in Belgian francs) and the annual household income (HI, in Belgian francs) for 235 Belgian working-class households. Ernst Engel provided these data to support his hypothesis that the proportion spent on food falls with increasing income, even if actual expenditure on food rises. The data are also used in Koenker (2005) to illustrate quantile regression and are available in the `quantreg` R package (Koenker, 2011). The left panel of Figure 2.5 plots the absolute food expenditure versus household income. The right panel plots the relative food expenditure percentage (i.e. FEP $:=$ 100FE$/$HI). These plots suggest that the absolute food expenditure increases with increasing household income, while the relative food expenditure decreases.

Consider the quantile regression model (QRM)

$$\mathrm{Q}_\tau\left(\text{FEP} \mid \text{HI}\right) = \alpha_{0\tau} + \alpha_{1\tau}\text{HI}. \tag{2.29}$$

Table 2.7 presents the parameter estimates for $\tau = 0.1$, $\tau = 0.5$, and $\tau = 0.9$. If the household income increases with 1000 Belgian francs, the 10% percentile of relative food expenditure significantly decreases with an estimate of 9% (95% confidence interval $[5\%, 13\%]$). For the median and the 90% percentile, this effect is 7% ($[5\%, 12\%]$) and 6% ($[2\%, 10\%]$), respectively. This analysis supports Engel's hypothesis that the proportion spent on food falls with increasing income. The estimated decrease is smaller for households that have a higher relative food expenditure. This difference is, however, not significant ($p = 0.6$).

Engel's hypothesis is also supported by the analysis of the data with the PIM,

$$\mathrm{logit}[\mathrm{P}\left(\text{FEP} \preccurlyeq \text{FEP}'\right)] = \beta(\text{HI}' - \text{HI}). \tag{2.30}$$

The parameter estimate is presented in Table 2.7. If the household income is, for example, 1000 Belgian francs higher, then the probability of a larger relative food expenditure percentage is estimated as

$$\hat{\mathrm{P}}(\text{FEP} \preccurlyeq \text{FEP}' \mid \text{HI}' = \text{HI} + 1000) = \mathrm{expit}(\hat{\beta}1000) = \mathrm{expit}(-0.94) = 28\%,$$

indicating it is unlikely that the household with the highest income will have a higher relative food expenditure. The 95% confidence interval is given by $[20\%, 37\%]$.

The GOF of PIM (2.30) is again postponed to Chapter 5. Furthermore, in Section 3.2 we analyze the food expenditure data with a more complex PIM.

2.5.4 The Beck depression inventory revisited

In Section 1.5 we analyzed the Beck depression inventory study where the outcome was the BDI improvement (denoted as BD), defined as the BDI at baseline (BDI_0) minus the BDI at the end of the study (BDI_1). Since the BDI is an ordinal outcome, the difference may not have a meaningful interpretation. Therefore we analyze the data with a more complicated PIM than (1.6), which does not violate the ordinal nature of the BDI. We consider the PIM

$$\mathrm{P}\left(\text{BDI}_1 \preccurlyeq \text{BDI}_1'\right) = \mathrm{expit}\Big[\alpha[\mathrm{I}\left(\text{BDI}_0 < \text{BDI}_0'\right) - \mathrm{I}\left(\text{BDI}_0 > \text{BDI}_0'\right)] + \beta(\text{DOSE}' - \text{DOSE})\Big]. \tag{2.31}$$

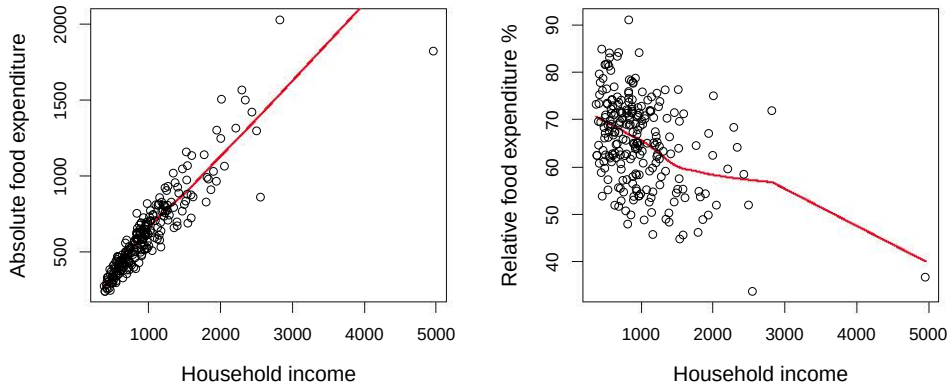

Figure 2.5: Left: absolute food expenditure as a function of the household income. Right: relative food expenditure as a function of the household income. The solid line corresponds to a nonparametric estimate of the regression function.

Table 2.7: Results of the fits of the QRM (2.29) and the PIM (2.30)

Parameter	Estimate	SE	p-value
		QRM	
		$\tau = 0.1$	
intercept (α_0)	63	2.68	< 0.001
HI (α_1)	-0.0094	0.0027	< 0.001
		$\tau = 0.5$	
intercept (α_0)	73	1.84	< 0.001
HI (α_1)	-0.0073	0.0017	< 0.001
		$\tau = 0.9$	
intercept (α_0)	82	0.9	< 0.001
HI (α_1)	-0.0062	0.0021	0.005
		PIM	
HI (β)	-0.00094	0.00021	< 0.001

In both the left and right hand side of the PIM only an ordering between the ordinal BDI's is considered. The interpretation of the model parameters follow from

$$\text{expit}(\alpha) = \text{P}\left(\text{BDI}_1 \preccurlyeq \text{BDI}_1' \mid \text{BDI}_0 < \text{BDI}_0', \text{DOSE}' = \text{DOSE}\right),$$

and

$$\text{expit}(\beta) = \text{P}\left(\text{BDI}_1 \preccurlyeq \text{BDI}_1' \mid \text{BDI}_0 = \text{BDI}_0', \text{DOSE}' = \text{DOSE} + 1\right).$$

The parameter estimates are $\hat{\alpha} = 2.47$ (SE : 0.21, $p < 0.001$) and $\hat{\beta} = -0.11$ (SE : 0.023, $p < 0.001$). We conclude that for two randomly selected patients that received the same dose of quetiapine, the estimated probability that the patient with highest BDI at baseline will also have a higher BDI at the end of the study is $\text{expit}(2.47) = 92\%$. On the other hand, if we randomly select two patients with the same baseline BDI and for which one patient receives a dose of quetiapine which is 5 gram higher, there is an estimated $\text{expit}(-0.55) = 37\%$ chance that this patient will have a higher BDI at the end of the study. This suggests that it is more likely that patients will benefit from the treatment, a conclusion similar to one obtained in Section 1.5 (recall that higher BDI indicates more depressed). Since it is difficult to think of a data generating model that implies the PIM (2.31), the parametrizations needs to be studied in more detail. This is beyond the scope of this dissertation.

The GOF of PIM (2.31) is postponed to Chapter 5.

2.6 Subject-specific probabilistic index versus population probabilistic index

In this section we briefly discuss the issue of non-collapsibility for the PI. For a similar discussion on the odds ratio we refer to, for example, Groenwold et al. (2011).

Consider a paired design and a completely randomized independent two sample design. The data generating mechanisms that we consider here are very simple because this section merely serves for pointing out potential pitfalls while interpreting the PI on population level versus interpreting the PI on subject level. A deeper discussion of non-collapsibility within the PIM framework is beyond the scope of the dissertation. We refer to Hand (1992); Senn (2011); Vansteelandt (2012) for more details.

Consider the setting where we want to assess the effect of a treatment on an outcome by means of a paired design. Let the outcome prior to treatment (no treatment N) be $Y_N \overset{d}{=} \mathrm{N}(\mu_N, \sigma^2)$, and assume that the treatment effect $D \overset{d}{=} \mathrm{N}(\delta, \sigma_D^2)$, so that the outcome after treatment, say Y_T, is given by $Y_T = Y_N + D$. For this paired design the *subject-specific* treatment effect in terms of the PI is given by

$$\begin{aligned} \mathrm{P}\left(Y_N \leq Y_T\right) &= \mathrm{P}\left(Y_N - Y_T \leq 0\right) \\ &= \mathrm{P}\left(-D \leq 0\right) \\ &= \mathrm{P}\left(Z \leq \frac{\delta}{\sigma_D}\right) \quad \text{where} \quad Z = \frac{-D+\delta}{\sigma_D} \overset{d}{=} \mathrm{N}(0,1) \\ &= \Phi\left(\frac{\delta}{\sigma_D}\right), \end{aligned}$$

while, the subject-specific treatment effect in terms of the mean is given by

$$\mathrm{E}\left(Y_T - Y_N\right) = \mathrm{E}\left(D\right) = \delta.$$

Suppose it is infeasible to set up a paired design. As an alternative strategy, the treatment can be randomized over the subjects so that half of the subjects receive the treatment while the other half remains untreated (or receives a placebo treatment).

Denote the outcomes prior to the assignment of the treatment for both groups as $\tilde{Y}_1$ and $\tilde{Y}_2$, respectively, both distributed as $\mathrm{N}(\mu_N, \sigma^2)$, with $\tilde{Y}_1$ and $\tilde{Y}_2$ statistically independent. Suppose the first group does not receive the treatment so that $\tilde{Y}_N := \tilde{Y}_1$, while for the group receiving the treatment this is $\tilde{Y}_T := \tilde{Y}_2 + D$. Consequently

$$\tilde{Y}_N \overset{d}{=} \mathrm{N}(\mu_N, \sigma^2) \quad \text{and} \quad \tilde{Y}_T \overset{d}{=} \mathrm{N}(\mu_N + \delta, \sigma^2 + \sigma_D^2).$$

For these data generating mechanisms, the *population* PI is given by

$$\begin{aligned} \mathrm{P}\left(\tilde{Y}_N \leq \tilde{Y}_T\right) &= \mathrm{P}\left(\tilde{Y}_N - \tilde{Y}_T \leq 0\right) \\ &= \mathrm{P}\left(Z \leq \frac{\delta}{\sqrt{2\sigma^2 + \sigma_D^2}}\right) \quad \text{where} \quad Z = \frac{\left(\tilde{Y}_N - \tilde{Y}_T\right) + \delta}{\sqrt{2\sigma^2 + \sigma_D^2}} \overset{d}{=} \mathrm{N}(0,1) \\ &= \Phi\left(\frac{\delta}{\sqrt{2\sigma^2 + \sigma_D^2}}\right). \end{aligned}$$

Consequently, $\mathrm{P}(\tilde{Y}_N \leq \tilde{Y}_T) \neq \mathrm{P}\left(Y_N \leq Y_T\right)$. For the population mean, on the other hand, $\mathrm{E}(\tilde{Y}_T - \tilde{Y}_N) = \mathrm{E}(\tilde{Y}_T) - \mathrm{E}(\tilde{Y}_N) = \delta$, so that $\mathrm{E}(\tilde{Y}_T - \tilde{Y}_N) = \mathrm{E}\left(Y_T - Y_N\right)$. Therefore, where a randomized trial allows to quantify the subject-specific effect in terms of the mean (since the

subject-specific effect coincides with the population effect), this does no longer hold for the PI so that, for this specific example, the Mann–Whitney estimator of the PI will be a consistent estimator for the population PI, but not for the subject-specific PI.

2.7 Discussion

A general class of semiparametric models for the PI is introduced and is referred to as a probabilistic index model (PIM). PIMs apply to ordinal, interval, and ratio-scale outcomes and the model parameters have an informative and intuitive interpretation in terms of the probabilistic index.

The asymptotic theory is based on the work of Lumley and Mayer-Hamblett (2003), using the concept of sparse correlation. The estimating equations make use of the score function of regression models under the working independence condition. Although this choice results in consistent and asymptotically normally distributed parameter estimators, it does not guarantee semiparametric efficient estimators. In Chapter 7 we improve the methods further by the construction of the asymptotic theory without making use of sparse correlation and by the construction of efficient score functions.

The results of the simulation study demonstrate that the theoretical properties of the parameter and variance estimators apply relatively well to moderately sized samples. In Chapters 3 and 4 the finite sample properties of these estimators are compared with other techniques. Several case studies are considered to illustrate the PIM. The assessment of the model adequacy is, however, postponed to Chapter 5, where goodness-of-fit methods for PIMs are constructed.

Although the PI may be considered as an intuitive effect size measure, there are some pitfall related to its interpretation in a randomized study. However, this needs to be studied in more detail.

Chapter 3

Relationship with regression models

The content of this chapter is primarily based on the results published in

Thas, O., De Neve, J., Clement, L., and Ottoy, J.P. (2012) Probabilistic index models (with discussion). *Journal of the Royal Statistical Society - Series B*, 74:623–671.

More specifically, it is based on sections 4 and 5 of the manuscript as well as on the discussions of Thomas Alexander Gerds and Joseph McKean.

3.1 Outline

The main aim of this chapter is to situate the PIM within the statistical landscape of regression methods. More specifically, the relationship with linear regression is addressed in Section 3.2, with the Cox proportional hazards model in Section 3.3, with AUC-regression in Section 3.4, with rank regression in Section 3.5, and with the cumulative logit model in Section 3.6. We also explore the relationship with the concordance index and show how it can be embedded within a PIM in Section 3.7. The performance of some of these methods is empirically assessed in a simulation study in Section 3.8. Section 3.9 gives the conclusions and discussion.

3.2 The linear regression model

Without loss of generality we limit the discussion to a one-dimensional covariate X. Consider the linear model

$$Y = \mu + \alpha X + \varepsilon, \tag{3.1}$$

where $\varepsilon \stackrel{d}{=} F_\varepsilon$, $\mathrm{E}(\varepsilon) = 0$, and $\varepsilon \perp\!\!\!\perp X$ for a continuous ε . The model can be equivalently formulated as

$$Y - (\mu + \alpha X) \mid X \sim F_\varepsilon.$$

Since Y is continuous, $\mathrm{P}(Y \preccurlyeq Y') = \mathrm{P}(Y < Y')$. Consider the conditional PI for this class of regression models,

$$\begin{aligned} \mathrm{P}(Y < Y' \mid X, X') &= \mathrm{P}(\mu + \alpha X + \varepsilon < \mu + \alpha X' + \varepsilon' \mid X, X') \\ &= \mathrm{P}(\varepsilon - \varepsilon' < \alpha(X' - X) \mid X, X') = F_\Delta[\alpha(X' - X)], \end{aligned} \tag{3.2}$$

where F_Δ is the distribution function of $\varepsilon - \varepsilon'$. Consider a PIM with link function $g(\cdot)$ and covariate pattern Z that depends on X and X'

$$\mathrm{P}(Y < Y' \mid X, X') = g^{-1}(\beta Z). \tag{3.3}$$

Combing (3.2) and (3.3) leads to the relationship

$$F_\Delta[\alpha(X' - X)] = g^{-1}(\beta Z). \tag{3.4}$$

If the linear model (3.1) holds, then relationship (3.4) suggests for PIM (3.3) to choose

$$g^{-1}(u) = F_\Delta(u) \quad \text{and} \quad Z = X' - X,$$

for which the model parameter relationship $\beta = \alpha$ is obtained. We work this out for two linear models with normal errors.

3.2.1 The homoscedastic normal linear model

Consider the normal linear regression model for which the error term $\varepsilon \stackrel{d}{=} \mathrm{N}(0, \sigma^2)$, i.e.

$$F_\varepsilon(u) = \Phi\left(\frac{u}{\sigma}\right),$$

with $\Phi(\cdot)$ the standard normal distribution function. Since $-\varepsilon \overset{d}{=} \mathrm{N}(0, \sigma^2)$ and the sum of two independently normally distributed variables is also normally distributed, if follows that $\varepsilon - \varepsilon' \overset{d}{=} \mathrm{N}(0, 2\sigma^2)$ so that

$$F_{\Delta}(u) = \Phi\left(\frac{u}{\sqrt{2\sigma^2}}\right).$$

Equation (3.4) becomes

$$\Phi\left(\frac{\alpha(X' - X)}{\sqrt{2\sigma^2}}\right) = g^{-1}(\beta Z).$$

With $Z = X' - X$ and the probit link function $g^{-1}(\cdot) = \Phi(\cdot)$, a simple relationship between α and β is established

$$\beta = \frac{\alpha}{\sqrt{2\sigma^2}},$$

which expresses that β is proportional to α. See, for example, Tian (2008) where this relationship is used to estimate the PI parametrically. From model (3.1) it follows that

$$\alpha = \mathrm{E}\left(Y \mid X + 1\right) - \mathrm{E}\left(Y \mid X\right) \quad \text{and} \quad \sigma^2 = \mathrm{Var}\left(Y \mid X\right).$$

Consequently

$$\beta = \frac{\mathrm{E}\left(Y \mid X + 1\right) - \mathrm{E}\left(Y \mid X\right)}{\sqrt{2\mathrm{Var}\left(Y \mid X\right)}}. \tag{3.5}$$

Under the normality, linearity, and homoscedasticity assumptions of the regression model we therefore conclude that β also has an interpretation in terms of the effect of X on the conditional mean of the outcome, relative to its conditional variance. Consequently, the PI does not only quantify the effect of X on the mean outcome, but also takes the variability into account. This is illustrated in Figure 3.1. For both panels the mean difference is one, while the variance in the right panel is five times the variance of the left panel. This increase in variance results in a decrease of the PI. The probability that an observation of the dashed density exceeds an observation of the solid density is 76% for the left panel, while for the right panel this decreases to 62%. When the regression model assumptions do not hold, the equivalence (3.5) does not necessarily hold, but the parameter β in the PIM is still related to the PI according to

$$\beta = g\left[\mathrm{P}\left(Y \preccurlyeq Y' \mid X, X' = X + 1\right)\right].$$

3.2.2 The heteroscedastic normal linear model

We can also establish the relationship between α and β when the residual variance σ^2 is not constant but depends on X, i.e. $\varepsilon \overset{d}{=} \mathrm{N}[0, \sigma^2(X)]$, which corresponds to $F_{\varepsilon}(u \mid X) = \Phi[u/\sigma(X)]$.

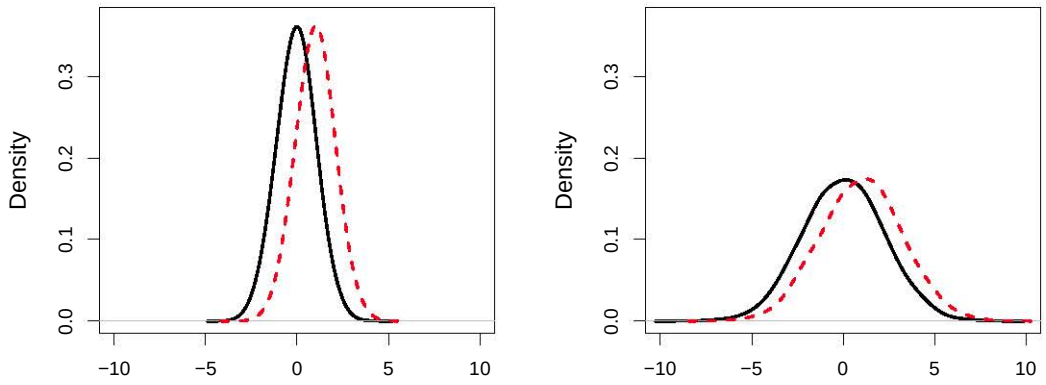

Figure 3.1: Left: densities for $\mathrm{N}(0,1)$ (solid line —) and $\mathrm{N}(1,1)$ (dashed line $- - -$). The probability that an observation of the dashed density exceeds an observation of the solid density is 76%. Right: densities for $\mathrm{N}(0,5)$ (solid line —) and $\mathrm{N}(1,5)$ (dashed line $- - -$). The probability that an observation of the dashed density exceeds an observation of the solid density is 62%.

Without loss of generality we assume that $X > 0$ and we only discuss $\sigma^2(X) = \gamma X$ as the variance function. Similar as for the homoscedastic model one can show that

$$F_\Delta(u \mid X, X') = \Phi\left(\frac{u}{\sqrt{\gamma(X'+X)}}\right).$$

Equation (3.4) becomes

$$\Phi\left(\frac{\alpha(X'-X)}{\sqrt{\gamma(X'+X)}}\right) = g^{-1}(\beta Z).$$

With $Z = (X'-X)/\sqrt{X'+X}$ and the probit link function $g^{-1}(\cdot) = \Phi(\cdot)$, a simple relationship between α and β is established

$$\beta = \frac{\alpha}{\sqrt{\gamma}}.$$

This suggests that the PIM for the heteroscadstic model should be formulated as

$$\mathrm{P}\left(Y < Y' \mid X, X'\right) = \Phi\left(\frac{X'-X}{\sqrt{X'+X}}\beta\right). \tag{3.6}$$

Model (3.6) gives a slightly different interpretation of β in terms of the PI as compared to the homoscedastic PIM. For $X' = X + 1$, we find

$$\mathrm{P}\left(Y < Y' \mid X, X' = X+1\right) = \Phi\left(\frac{\beta}{\sqrt{2X+1}}\right).$$

This expression illustrates that the effect of X on the distribution of Y diminishes as X increases, at least in terms of the PI. The increasing residual variance does not affect the covariate effect on the mean outcome, but it results in a negative effect modulation in terms of the PI. This is illustrated in Figure 3.2; the left panel shows heteroscedastic data and the related regression model $\mathrm{E}(Y \mid X) = X$. The slope of the regression model is fixed, i.e. $\mathrm{E}(Y' - Y \mid X, X' = X + 1) = 1$, implying that the effect of a unit-increase in X on $\mathrm{E}(Y \mid X)$ is independent of X, hence ignoring the effect of X on the residual variance. The right panel shows the corresponding $\mathrm{P}(Y < Y' \mid X, X' = X + 1)$, which depends on X. Indeed, for example, $\mathrm{P}(Y < Y' \mid X = 1, X' = 2) = 72\%$, while $\mathrm{P}(Y < Y' \mid X = 9, X' = 10) = 59\%$.

This phenomenon was also noticed by Brumback et al. (2006) and it suggests that one should take care in interpreting the mean effect parameter in a normal regression model with non-constant variance because the importance of the covariate effect may actually depend on the covariate value.

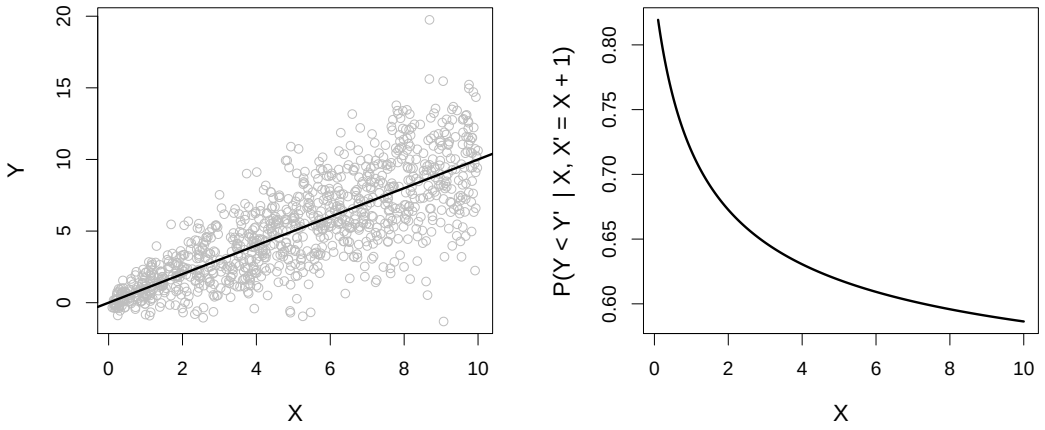

Figure 3.2: Left: the outcome Y is distributed according to a normal distribution with mean X and variance X. The solid line corresponds to the regression model $\mathrm{E}(Y \mid X) = X$. Right: the corresponding $\mathrm{P}(Y < Y' \mid X, X' = X + 1)$ as a function of X.

3.2.3 The food expenditure study revisited

To illustrate the PIM associated with heteroscedastic data, we consider the food expenditure example of Section 2.5.3. The left panel of Figure 3.3 shows the absolute food expenditure as a function of the household income. This plot has already been shown in the left panel of Figure 2.5, but now the household incomes are restricted to 1500 Belgian francs so as to have a better visualization. It is clear that the variability in absolute food expenditure increases with increasing household income. Instead of modelling the relative food expenditure percentage (FEP), we now model the absolute food expenditure (FE) with a PIM similar to (3.6):

$$\mathrm{P}\left(\mathrm{FE} \preccurlyeq \mathrm{FE}' \mid \mathrm{HI}, \mathrm{HI}'\right) = \Phi\left[\frac{(\mathrm{HI}' - \mathrm{HI})}{\sqrt{\mathrm{HI}' + \mathrm{HI}}}\beta\right]. \tag{3.7}$$

The estimated slope is $\hat{\beta} = 0.39$ (SE : 0.07) and is highly significant ($p < 0.001$). The right panel of Figure 3.3 shows the estimated PI related to a household income increase of 100 Belgian francs, as well as pointwise 95% confidence bounds. For example, if the household income is 500 Belgian francs then the probability of more food expenditure with a household income of 600 Belgian francs is estimated as 88.0% with a 95% confidence interval of $[77.4\%, 94.5\%]$. When we compare households of 1500 and 1600 Belgian francs this estimated probability drops to 75.8% with a 95% confidence interval of $[67.3\%, 82.9\%]$. This is an example of the negative effect modification of the increasing error variance.

3.3 The Cox proportional hazards model

Cox proportional hazards regression models (Cox, 1972) form a very popular class of models for the analysis of survival data, or, more generally, time-to-event data. Although the PIM was not known during the 1970s, several papers on Cox regression models appear to present results that are closely related to PIMs. For example, Holt and Prentice (1974), while studying Cox regression models for paired data, showed that the marginal likelihood of their models contains factors of the form $\mathrm{P}\left(Y_{1i} < Y_{2i} \mid X_{1i}, X_{2i}\right)$, where Y_{1i} and Y_{2i} are paired survival times (e.g. from twin studies) with covariates X_{1i} and X_{2i}. Under the assumption of proportional hazards in the absence of censored or tied data, they found that

$$\text{logit}\left[\mathrm{P}\left(Y_{1i} < Y_{2i} \mid X_{1i}, X_{2i}\right)\right] = \beta(X_{1i} - X_{2i}),$$

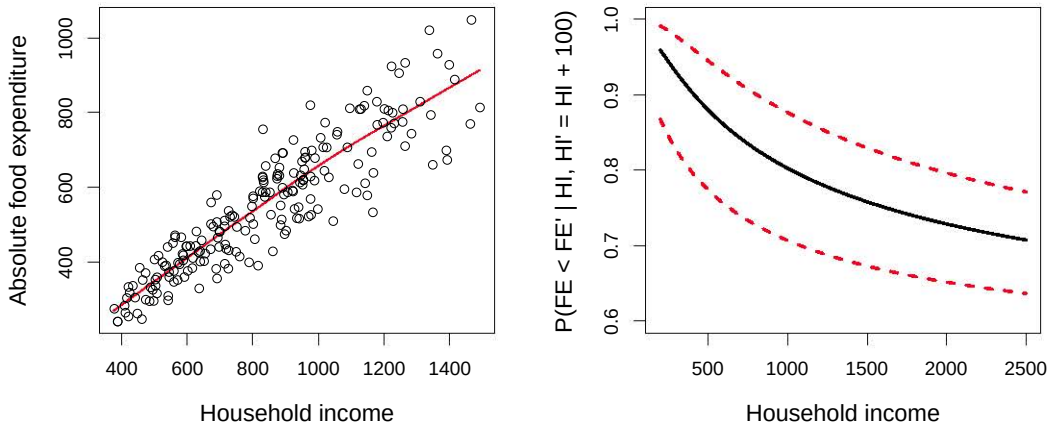

Figure 3.3: Left: absolute food expenditure (FE) in function of household income (HI). The solid line corresponds to a nonparametric estimate of the regression function. Right: estimated PI associated with a household income increase of 100 Belgian francs based on model (3.7). The dashed lines correspond to pointwise 95% confidence bounds.

in which the parameter β originates from the hazard function $\lambda(y \mid X) = \lambda_0(y)\exp(\beta X)$. Note, however, that for the PIMs presented in this dissertation, it is assumed that all observations are mutually independent, whereas Holt and Prentice (1974) developed their method for paired outcome variables (paired survival times).

Also the marginal likelihood formulation of Kalbfleish and Prentice (1973), which is related to the ranks of the survival times, is closely related to a PIM and the parameters are again interpretable in the proportional hazards model.

We will show that conditional distributions that belong to the class of proportional hazards models imply a PIM with logit link. Let $S_{Y|\boldsymbol{X}}(y \mid \boldsymbol{X}) = 1 - F_{Y|\boldsymbol{X}}(y \mid \boldsymbol{X})$ denote the survival function. The hazards function is defined as

$$\lambda(y \mid \boldsymbol{X}) = -\frac{d}{dy}\log\left[S_{Y|\boldsymbol{X}}(y \mid \boldsymbol{X})\right] = \frac{f_{Y|\boldsymbol{X}}(y \mid \boldsymbol{X})}{S_{Y|\boldsymbol{X}}(y \mid \boldsymbol{X})}. \tag{3.8}$$

In a proportional hazards model the hazards function allows a factorization of the form

$$\lambda(y \mid \boldsymbol{X}) = \lambda_0(y)\exp(\boldsymbol{X}^T\boldsymbol{\beta}), \tag{3.9}$$

in which $\lambda_0(y)$ is the baseline hazards function that does not depend on the covariate $\boldsymbol{X}$. Com-

bining (3.8) and (3.9) leads to

$$\begin{aligned} & \lambda_0(y)\exp(\boldsymbol{X}^T\boldsymbol{\beta}) = -\frac{d}{dy}\log\left[S_{Y|\boldsymbol{X}}(y \mid \boldsymbol{X})\right] \\ \Leftrightarrow \quad & \exp(\boldsymbol{X}^T\boldsymbol{\beta})\int \lambda_0(y)dy = -\log\left[S_{Y|\boldsymbol{X}}(y \mid \boldsymbol{X})\right]. \end{aligned} \tag{3.10}$$

It follows that $\int \lambda_0(y)dy = -\log\left[S_{Y|\boldsymbol{X}}(y \mid \boldsymbol{X}=\boldsymbol{0})\right]$. If we define the baseline survival function $S_0(y) := S_{Y|\boldsymbol{X}}(y \mid \boldsymbol{X}=\boldsymbol{0})$, then from (3.10) we have

$$\exp(\boldsymbol{X}^T\boldsymbol{\beta})\log[S_0(y)] = \log\left[S_{Y|\boldsymbol{X}}(y \mid \boldsymbol{X})\right].$$

Thus, within the class of proportional hazards models the survival function is of the form

$$S_{Y|\boldsymbol{X}}(y \mid \boldsymbol{X}) = [S_0(y)]^{\exp(\boldsymbol{X}^T\boldsymbol{\beta})}. \tag{3.11}$$

From (3.11) it follows that

$$\begin{aligned} \frac{dS_{Y|\boldsymbol{X}}(y \mid \boldsymbol{X})}{dy} &= \frac{\exp(\boldsymbol{X}^T\beta)}{S_0(y)}\left[S_0(y)\right]^{\exp(\boldsymbol{X}^T\beta)}\frac{dS_0(y)}{dy} \\ &= \frac{\exp(\boldsymbol{X}^T\beta)}{S_0(y)}S_{Y|\boldsymbol{X}}(y \mid \boldsymbol{X})\frac{dS_0(y)}{dy}. \end{aligned} \tag{3.12}$$

This can be equivalently expressed as

$$S_{Y|\boldsymbol{X}}(y \mid \boldsymbol{X}) = \frac{dS_{Y|\boldsymbol{X}}(y \mid \boldsymbol{X})S_0(y)\exp(-\boldsymbol{X}^T\beta)}{dS_0(y)}. \tag{3.13}$$

We substitute these expressions in the conditional PI

$$\begin{aligned} \mathrm{P}\left(Y < Y' \mid \boldsymbol{X}, \boldsymbol{X}'\right) &= \int F_{Y|\boldsymbol{X}}(y \mid \boldsymbol{X})dF_{Y|\boldsymbol{X}}(y \mid \boldsymbol{X}') \\ &= -\int \left[1 - S_{Y|\boldsymbol{X}}(y \mid \boldsymbol{X})\right]dS_{Y|\boldsymbol{X}}(y \mid \boldsymbol{X}') \\ &= 1 + \int S_{Y|\boldsymbol{X}}(y \mid \boldsymbol{X})dS_{Y|\boldsymbol{X}}(y \mid \boldsymbol{X}') \qquad (3.14) \\ &= 1 + \int \left[\frac{dS_{Y|\boldsymbol{X}}(y \mid \boldsymbol{X})S_0(y)\exp(-\boldsymbol{X}^T\beta)}{dS_0(y)}\right] \qquad (3.15) \\ &\quad \times \left[\frac{\exp(\boldsymbol{X}'^T\beta)}{S_0(y)}S_{Y|\boldsymbol{X}}(y \mid \boldsymbol{X}')\right]dS_0(y) \\ &= 1 + \exp\left[(\boldsymbol{X}' - \boldsymbol{X})^T\boldsymbol{\beta}\right]\int S_{Y|\boldsymbol{X}}(y \mid \boldsymbol{X}')dS_{Y|\boldsymbol{X}}(y \mid \boldsymbol{X}) \\ &= 1 - \exp\left[(\boldsymbol{X}' - \boldsymbol{X})^T\boldsymbol{\beta}\right]\left[1 - \mathrm{P}\left(Y > Y' \mid \boldsymbol{X}, \boldsymbol{X}'\right)\right] \\ &= 1 - \exp\left[(\boldsymbol{X}' - \boldsymbol{X})^T\boldsymbol{\beta}\right]\mathrm{P}\left(Y < Y' \mid \boldsymbol{X}, \boldsymbol{X}'\right), \qquad (3.16) \end{aligned}$$

where in equation (3.14) we substituted (3.12) and (3.13). Equation (3.16) is equivalent to

$$\begin{aligned} & \mathrm{P}\left(Y < Y' \mid \boldsymbol{X}, \boldsymbol{X}'\right)\left\{1 + \exp\left[(\boldsymbol{X}' - \boldsymbol{X})^T\boldsymbol{\beta}\right]\right\} = 1 \\ \Leftrightarrow \quad & \mathrm{P}\left(Y < Y' \mid \boldsymbol{X}, \boldsymbol{X}'\right) = \frac{1}{1 + \exp\left[-(\boldsymbol{X} - \boldsymbol{X}')^T\boldsymbol{\beta}\right]} \\ \Leftrightarrow \quad & \mathrm{P}\left(Y < Y' \mid \boldsymbol{X}, \boldsymbol{X}'\right) = \mathrm{expit}[(\boldsymbol{X} - \boldsymbol{X}')^T\boldsymbol{\beta}]. \end{aligned}$$

For survival functions satisfying (3.11), it therefore holds that

$$\mathrm{logit}\left[\mathrm{P}\left(Y < Y' \mid \boldsymbol{X}, \boldsymbol{X}'\right)\right] = (\boldsymbol{X} - \boldsymbol{X}')^T\boldsymbol{\beta}.$$

This illustrates that the PIM with a logit link and with $\boldsymbol{Z} = \boldsymbol{X} - \boldsymbol{X}'$ arises naturally from a widely applicable class of distributions. A straightforward example is the exponential distribution with rate parameter γ which has a survival function $S(y) = \exp(-\gamma y)$. Equation (3.11) is satisfied with $S_0(y) = \exp(-y)$ and $\gamma(\boldsymbol{X}) = \exp(\boldsymbol{X}^T\boldsymbol{\beta})$.

We refer to Follmann (2002) for the relationship between the PI and the proportional hazards model in the presence of censoring.

3.4 The AUC regression model

Let Y denote the continuous outcome for a binary classifier, for example, a medical test. Let X_1 denote the two states, where, for example $X_1 = 0$ corresponds to non-diseased and $X_2 = 1$ to diseased patients. For a threshold y_t, let $Y > y_t$ denote the classification into class $X_1 = 1$. The true positive rate is then defined as $\mathrm{P}\left(Y > y_t \mid X_1 = 1\right)$ and the false positive rate as $\mathrm{P}\left(Y > y_t \mid X_1 = 0\right)$. The Receiver Operating Characteristic (ROC) curve plots the true positive rate in function of the false positive rate for varying threshold y_t. Figure 3.4 shows the ROC curves for three classifiers. The solid line results in a relatively good classification: at a false positive rate of 10%, the true positive rate is 50%. The dashed line correspond to a poor classification: at a false positive rate of 50%, the true positive rate is only 10%. However, by changing the predicted labels, this classifier has the same performance as the solid line. The dotted line corresponds to a classification based on random guessing: for a false positive rate of $x\%$, the true positive rate is also $x\%$, for $x \in [0, 100]$.

As a summary of the performance of a classifier, the Area Under the Curve (AUC) of an ROC

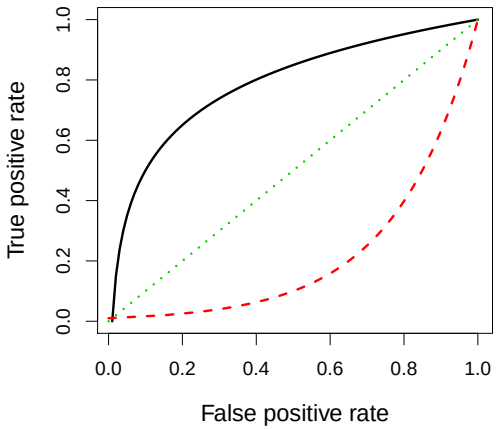

Figure 3.4: Simulated ROC curve. The solid line (—) corresponds to a relatively good classification, the dashed line (− − −) to a bad classification, and the dotted line (· · ·) to a classification based on random guessing.

curve is considered. As shown by (1.4) in Section 1.3, the AUC corresponds to the PI

$$\mathrm{P}\left(Y \preccurlyeq Y' \mid X_1 = 0, X_1' = 1\right).$$

The AUCs corresponding to the solid line and the dashed line of Figure 3.4 are 78% and 22%, respectively. For the dotted line this is 50%. Dodd and Pepe (2003) developed statistical methods which allow the AUC to depend on additional covariates, say $\boldsymbol{X}$. More specifically, their AUC regression model for a continuous outcome is given by

$$\mathrm{P}\left(Y < Y' \mid X_1 = 0, \boldsymbol{X}, X_1' = 1, \boldsymbol{X}'\right) = g^{-1}(\boldsymbol{Z}^T\boldsymbol{\beta}), \tag{3.17}$$

with $\boldsymbol{Z}$ depending on $\boldsymbol{X}$ and $\boldsymbol{X}'$. This is a special case of a PIM. Indeed, let $\tilde{\boldsymbol{X}}^T = (X_1, \boldsymbol{X}^T)$, then model (3.17) can be written as a PIM

$$\mathrm{P}\left(Y < Y' \mid \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{X}}'\right) = g^{-1}(\boldsymbol{Z}^T\boldsymbol{\beta}), \quad (\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{X}}') \in \mathcal{X},$$

where $\mathcal{X} = \{(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{X}}') \mid X_1 < X_1'\}$. Dodd and Pepe (2003) provide estimating equations similar to the estimating equation we propose, as given by (2.15). Their estimator of the variance, however, is different from the sandwich estimator in Theorem 2. More specifically, their estimator involves the conditional distribution function $F_{Y|\boldsymbol{X}}$, which must be replaced by a consistent estimator. Due to the sparseness of the covariate space this may be obstructed in real data settings. Dodd and Pepe (2003) therefore suggest to use bootstrap standard errors when covariate data are continuous or sparse.

3.5 Rank regression and the Hodges–Lehmann estimator

For the class of linear models of Section 3.2 the parameters can be estimated by means of several methods. With no full parametric assumption on the error distribution, least squares is arguably the most popular method. However, least squares suffers from the drawback that it is sensitive to outliers. Rank regression is considered as a robust alternative to least squares. We refer to McKean (2004) and McKean et al. (2009) for reviews.

Although rank regression parameter estimation can be defined in a general way, we will formulate it here only with the Wilcoxon scores. The slope parameter of the linear regression model (3.1) is estimated by minimizing

$$\sum_{i=1}^{n} \left(\frac{R[Y_i - (\mu + \alpha X_i)]}{n+1} - \frac{1}{2} \right) [Y_i - (\mu + \alpha X_i)], \tag{3.18}$$

where $R[Y_i - (\mu + \alpha X_i)]$ denotes the rank of the residual $Y_i - (\mu + \alpha X_i)$ among the n residuals. As we will see below, minimizing (3.18) is independent of μ since it drops out of the equation. The intercept is then typically estimated as $\hat{\mu} = \text{median}_i(Y_i - \hat{\alpha} X_i)$. The estimate of α is obtained by solving the estimating equation based on the partial derivative of (3.18),

$$\sum_{i=1}^{n} X_i \left(\frac{R[Y_i - (\mu + \alpha X_i)]}{n+1} - \frac{1}{2} \right) = 0. \tag{3.19}$$

The relationship with the estimating equation (2.15) of the PIM parameters becomes more transparent when the rank in (3.19) is replaced by an expression involving the pseudo-observations. We assume that there are no ties in the residuals. The rank of Y_i among the n observations can be expressed in terms of pseudo-observations as follows

$$R(Y_i) = \sum_{j=1}^{n} \mathrm{I}(Y_j \leq Y_i) = \sum_{j=1}^{n} \mathrm{I}(Y_j \preccurlyeq Y_i) + \frac{1}{2}.$$

Equation (3.19) may then be written as

$$\begin{aligned}
& \frac{1}{n+1} \sum_{i=1}^{n} X_i \left(\sum_{j=1}^{n} \mathrm{I}[Y_j - (\mu + \alpha X_j) \leq Y_i - (\mu + \alpha X_i)] - \frac{n+1}{2} \right) = 0 \\
\Leftrightarrow \quad & \frac{1}{n+1} \sum_{i=1}^{n} X_i \left(\sum_{j=1}^{n} \mathrm{I}[Y_j - (\mu + \alpha X_j) \preccurlyeq Y_i - (\mu + \alpha X_i)] - \frac{n}{2} \right) = 0 \\
\Leftrightarrow \quad & \frac{1}{n+1} \sum_{i=1}^{n} \sum_{j=1}^{n} X_i \left(\mathrm{I}[Y_j - (\mu + \alpha X_j) \preccurlyeq Y_i - (\mu + \alpha X_i)] - \frac{1}{2} \right) = 0.
\end{aligned}$$

This can be simplified to

$$\sum_{i=1}^{n}\sum_{j=1}^{n} X_i \left(\mathrm{I}[Y_j \preccurlyeq Y_i - \alpha(X_i - X_j)] - \frac{1}{2} \right) = 0. \tag{3.20}$$

To relate this to the PIM framework, consider the PIM

$$\mathrm{P}\left(Y_i \preccurlyeq Y_j \mid X_i, X_j\right) = \frac{1}{2} + \beta(X_i - X_j), \quad (X_i, X_j) \in \mathcal{X}_0,$$

and the estimating equation (2.15) with the simple index function $A(Z_{ij};\beta) = Z_{ij} = X_i - X_j$. Then

$$\begin{aligned}
& \sum_{i=1}^{n}\sum_{j=1}^{n} (X_i - X_j) \left(\mathrm{I}\left(Y_i \preccurlyeq Y_j\right) - \beta(X_i - X_j) - \frac{1}{2} \right) = 0 \\
\Leftrightarrow \quad & \sum_{i=1}^{n}\sum_{j=1}^{n} X_i \left(\mathrm{I}\left(Y_i \preccurlyeq Y_j\right) - \beta(X_i - X_j) - \frac{1}{2} \right) - \\
& \sum_{i=1}^{n}\sum_{j=1}^{n} X_j \left(\mathrm{I}\left(Y_i \preccurlyeq Y_j\right) - \beta(X_i - X_j) - \frac{1}{2} \right) = 0 \\
\Leftrightarrow \quad & \sum_{i=1}^{n}\sum_{j=1}^{n} X_i \left(\mathrm{I}\left(Y_i \preccurlyeq Y_j\right) - \beta(X_i - X_j) - \frac{1}{2} \right) + \\
& \sum_{i=1}^{n}\sum_{j=1}^{n} X_j \left(\mathrm{I}\left(Y_j \preccurlyeq Y_i\right) - \beta(X_j - X_i) - \frac{1}{2} \right) = 0 \\
\Leftrightarrow \quad & \sum_{i=1}^{n}\sum_{j=1}^{n} X_i \left(\mathrm{I}\left(Y_i \preccurlyeq Y_j\right) - \beta(X_i - X_j) - \frac{1}{2} \right) = 0.
\end{aligned} \tag{3.21}$$

By comparing the two estimating equations (3.20) and (3.21), we note that the major difference is that in rank regression the linear predictor $\alpha(X_i - X_j)$ appears within the indicator function, whereas for the PIM estimation method the linear predictor $\beta(X_i - X_j)$ appears outside the indicator function.

Another interesting observation is that the scores X_i and $X_i - X_j$ are interchangeable in the PIM estimating equation. This also holds true in the estimating equation (3.20) of the rank regression estimator. Thus pseudo-observations with equal covariate patterns do not contribute to the estimation of the parameter.

We now take a closer look at both approaches for the two-sample problem, i.e. when the covariate X is a dummy variable coding for two groups. Let $X = 1$ be used for group 1 and $X = 0$ for group 2, and suppose that the sample observations are ordered so that the first n_1 form group

1 and the last n_2 form group 2. The estimating equation (3.20) becomes

$$\begin{aligned}
&\sum_{i=1}^{n_1}\sum_{j=n_1+1}^{n}\left(\mathrm{I}\left(Y_i \preccurlyeq Y_j - \hat{\alpha}\right) - \frac{1}{2}\right) = 0 \\
\Leftrightarrow\quad &\frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=n_1+1}^{n}\mathrm{I}\left(\hat{\alpha} \preccurlyeq Y_j - Y_i\right) = \frac{1}{2} \\
\Leftrightarrow\quad &\hat{\alpha} = \mathrm{median}\{Y_j - Y_i \mid i = 1,\ldots,n_1, j = 1,\ldots,n_2\},
\end{aligned}$$

which is the Hodges–Lehmann estimator (Hodges and Lehmann, 1963). The PIM estimator is now the solution of

$$\begin{aligned}
&\sum_{i=1}^{n_1}\sum_{j=n_1+1}^{n}\left(\mathrm{I}\left(Y_i \preccurlyeq Y_j\right) - \hat{\beta} - \frac{1}{2}\right) = 0 \\
\Leftrightarrow\quad &\hat{\beta} = \frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=n_1+1}^{n}\mathrm{I}\left(Y_i \preccurlyeq Y_j\right) - \frac{1}{2},
\end{aligned}$$

which is Mann–Whitney statistic divided by $n_1 n_2$ and minus a half.

3.6 The cumulative logit model

The cumulative logit model (CLM) is briefly discussed in Section 2.2.4 and illustrated in Section 2.5.2. Here we consider the specific setting of a single predictor X so as to focus on the differences in interpretation between a CLM and a PIM. Let the outcome Y be ordinal with levels $\{1,\ldots,k\}$ and X continuous. Assume that the following CLM is appropriate

$$\mathrm{logit}[\mathrm{P}\left(Y \leq j \mid X\right)] = \mu_j + \alpha X, \quad j = 1,\ldots,k-1. \tag{3.22}$$

The interpretation of α follows from

$$\exp(\alpha) = \frac{\mathrm{odds}(Y \leq j \mid X = x+1)}{\mathrm{odds}(Y \leq j \mid X = x)},$$

i.e. $\exp(\alpha)$ quantifies the multiplicative change in odds that the outcome does not exceed a particular level if the predictor is increased by one unit. On the other hand, if the PIM

$$\mathrm{logit}[\mathrm{P}\left(Y \preccurlyeq Y' \mid X, X'\right)] = \beta(X' - X), \quad (X, X') \in \mathcal{X}_0,$$

is appropriate, the interpretation of β follows from

$$\exp(\beta) = \mathrm{odds}(Y \preccurlyeq Y' \mid X = x, X' = x+1),$$

i.e. $\exp(\beta)$ quantifies the odds that the outcome increases if the predictor is increased by one unit. Thus the parameter of the CLM is related to an odds ratio, whereas the parameter of the PIM is related to an odds and therefore quantifying the effect more directly.

If the predictor is ordinal, a linear modelling of X as in (3.22) will in general not hold. If X has l levels, say $\{1, \ldots, l\}$, then a CLM with a dummy-coded X may be more appropriate. For example

$$\text{logit}[\text{P}\,(Y \leq j \mid X)] = \mu_j + \sum_{i=1}^{l-1} \alpha_i \text{I}\,(X = i)\,, \quad j = 1, \ldots, k-1, \tag{3.23}$$

where $X = l$ is the reference group. The interpretation then follows from

$$\exp(\alpha_i) = \frac{\text{odds}(Y \leq j \mid X = i)}{\text{odds}(Y \leq j \mid X = l)}, \quad i = 1, \ldots, l-1.$$

Sometimes X can have many levels; the Beck depression inventory of Section 1.5 is ordinal and has 64 levels. The visual analogue scale of Section 1.6 is an example of a continuous ordinal variable. A dummy-coding for these types of predictors will often result in too many parameters. The PIM framework allows to include such predictors at the cost of a single parameter and without violating its ordinal nature. More specifically, if the following PIM holds

$$\text{logit}[\text{P}\,(Y \preccurlyeq Y' \mid X, X')] = \gamma[\text{I}\,(X < X') - \text{I}\,(X > X')], \quad (X, X') \in \mathcal{X}_0, \tag{3.24}$$

the interpretation follows from

$$\exp(\gamma) = \text{odds}(Y \preccurlyeq Y' \mid X < X'),$$

i.e. $\exp(\gamma)$ gives the odds that the outcome associated with a higher predictor value exceeds the outcome associated with a lower predictor value. Of course, as for any of the models presented in this section, goodness-of-fit methods are needed to assess the model adequacy. This forms the topic of Chapter 5. As we will see in the following section, PIM (3.24) is closely related to the concordance index.

3.7 The concordance index

The concordance index or C-index has been discussed by several authors; see, for example, Harrell et al. (1982, 1996). It is especially useful for discrimination of survival prediction

models; see, for example, Gerds et al. (2010); Gerds (2012); Koziol and Jia (2009). A pair of two variables (X_i, Y_i), $i = 1, 2$, are called *concordant* if

$$\text{sign}(X_1 - X_2) = \text{sign}(Y_1 - Y_2).$$

For a random sample of i.i.d. observations $\{(Y_i, X_i) \mid i = 1, \ldots, n\}$, with continuous outcome Y, the C-index is defined as the proportion of concordant pairs, i.e.

$$C = \frac{\sum_{\{i,j|X_i<X_j\}} \text{I}(Y_i < Y_j)}{\sum_{i,j} \text{I}(X_i < X_j)}.$$

Consider the PIM with identity link function

$$\text{P}(Y < Y' \mid X, X') = \frac{1}{2} + \beta[\text{I}(X < X') - \text{I}(X' > X)], \quad (X, X') \in \mathcal{X}_0.$$

The interpretation of β follows from

$$\beta = \text{P}(Y < Y' \mid X < X') - \frac{1}{2}, \tag{3.25}$$

i.e. the probability that an outcome associated with a higher X exceeds the outcome associated with a lower X, reduced with a half. It also holds that

$$\beta = \frac{1}{2} - \text{P}(Y < Y' \mid X' < X),$$

which is equivalent to (3.25), since

$$\begin{aligned} \beta &= \frac{1}{2} - \text{P}(Y < Y' \mid X' < X) \\ &= \frac{1}{2} - [1 - \text{P}(Y' < Y \mid X' < X)] \\ &= \text{P}(Y' < Y \mid X' < X) - \frac{1}{2}. \end{aligned}$$

Let $\boldsymbol{Z}_{ij}^T = (1, Z_{ij})$, with $Z_{ij} = \text{I}(X_i < X_j) - \text{I}(X_j > X_i)$, and $\boldsymbol{\beta}^T = (0.5, \beta)$. The estimating equations (2.15) with index function $\boldsymbol{A}(\boldsymbol{Z}_{ij}; \boldsymbol{\beta}) = \boldsymbol{Z}_{ij}$, become

$$\begin{aligned} & \sum_{i=1}^{n}\sum_{j=1}^{n} \boldsymbol{Z}_{ij}\left[\text{I}(Y_i < Y_j) - \boldsymbol{Z}_{ij}^T\hat{\boldsymbol{\beta}}\right] = \boldsymbol{0} \\ \Leftrightarrow \quad & \sum_{i=1}^{n}\sum_{j=1}^{n} Z_{ij}\left[\text{I}(Y_i < Y_j) - \left(\frac{1}{2} + Z_{ij}\hat{\beta}\right)\right] = 0 \\ \Leftrightarrow \quad & \sum_{\{i,j|X_i\neq X_j\}} Z_{ij}\left[\text{I}(Y_i < Y_j) - \left(\frac{1}{2} + Z_{ij}\hat{\beta}\right)\right] = 0 \\ \Leftrightarrow \quad & \sum_{\{i,j|X_i<X_j\}}\left[\text{I}(Y_i < Y_j) - \left(\frac{1}{2} + \hat{\beta}\right)\right] - \sum_{\{i,j|X_i>X_j\}}\left[\text{I}(Y_i < Y_j) - \left(\frac{1}{2} - \hat{\beta}\right)\right] = 0 \\ \Leftrightarrow \quad & \frac{\sum_{\{i,j|X_i<X_j\}} \text{I}(Y_i < Y_j)}{\sum_{i,j} \text{I}(X_i < X_j)} - \frac{1}{2} = \hat{\beta} \\ \Leftrightarrow \quad & C - \frac{1}{2} = \hat{\beta}. \end{aligned}$$

This relates the C-index to the PIM-framework. This allows to extend the C-index to multiple predictors. Consider a random sample of i.i.d. observations $\{(Y_i, \boldsymbol{X}_i = (X_{1i}, X_{2i})^T) \mid i = 1, \ldots, n\}$, and the PIM defined for the no-order restriction

$$\mathrm{P}\left(Y < Y' \mid \boldsymbol{X}, \boldsymbol{X}'\right) = \frac{1}{2} + \beta_1[\mathrm{I}\left(X_1 < X_1'\right) - \mathrm{I}\left(X_1' > X_1\right)] + \beta_2[\mathrm{I}\left(X_2 < X_2'\right) - \mathrm{I}\left(X_2' > X_2\right)].$$

The interpretation follows from

$$\beta_1 = \mathrm{P}\left(Y < Y' \mid X_1 < X_1', X_2 = X_2'\right) - \frac{1}{2},$$

i.e. the probability that an outcome associated with a higher X_1 exceeds the outcome associated with a lower X_1, reduced by $1/2$ and while keeping X_2 fixed. A similar interpretation holds for β_2.

3.8 Simulation study

In Section 2.4 the theoretical properties of Theorems 1 and 2 were evaluated in a simulation study, where data were generated according to a normal linear model with constant or varying variance and an exponential generalized linear model. Here we reconsider these simulation settings together with the cumulative logit regression model to examine the theoretical properties of the PIM estimators in more detail. The relationships with a PIM are provided in Sections 3.2, 3.3, and 3.6 respectively.

Since for each of the three settings the data-generating model is known, their parameters can also be estimated by means of maximum likelihood. Variances of the maximum likelihood estimators and powers of the Wald tests using the maximum likelihood estimators will also be reported in this section. These variances and powers need to be interpreted as optimistic benchmarks as they only give an impression of the parametric lower bound of the variances and upper bound of the powers. Moreover, it is unfair to compare variances and powers from a semiparametric method with their counterparts from a parametric method because the former methods will usually only be applied when the data-generating mechanism is unknown or incompletely specified so that no parametric statistical analysis is advised. We also remind the reader that we have introduced PIMs as a flexible class of semiparametric models to be used when the focus is on the PI as an effect-size measure. In the absence of strong parametric assumptions no parametric methods can be used for this purpose.

For each data-generating model we also consider semiparametric estimators, such as the least-squares estimator for the normal linear model and the semiparametric proportional hazards estimator for the exponential model.

All computations have been performed with the R software (R Core Team, 2012) and all PIMs are defined for the lexicographical order restriction because they all satisfy the antisymmetry condition; see Section 2.3.2 for more information.

For the reader's convenience, we summarize all data generating models in this section, but most have already been discussed in Section 2.4.

3.8.1 The normal linear model

We consider the model

$$Y_i = \alpha X_i + \varepsilon_i, \quad i = 1, \ldots, n, \tag{3.26}$$

where $\varepsilon_i \mid X_i \stackrel{d}{=} N[0, \sigma^2(X_i)]$. Sample sizes of $n = 25$, $n = 50$, and $n = 200$ are considered. The predictor X takes equally spaced values in the interval $[0.1, u]$ where $u = 1$ or 10. The parameter α equals 1 or 10. Table 3.1 presents the results for a constant standard deviation, i.e. $\sigma(X) = \sigma$, with $\sigma = 1$ or $\sigma = 5$. The corresponding PIM is given by

$$\Phi^{-1}\left[\mathrm{P}\left(Y \preccurlyeq Y' \mid X, X'\right)\right] = \beta(X' - X),$$

where $\beta = \alpha/(\sqrt{2}\sigma)$. For each setting, 1000 Monte Carlo simulation runs are used for the empirical investigation of the distributions of the semiparametric estimator of β. The semiparametric estimator of Section 2.3.2 is denoted by $\hat{\beta}$, and it is further referred to as the PIM estimator. Table 3.1 shows for each simulation setting the true β parameter and the average of the simulated estimates. The latter is an approximation of the true mean of the estimator. The table also reports the average of the simulated sandwich variance estimates, which is an approximation of the expectation of the sandwich estimator, and the sample variance of the 1000 estimates $\hat{\beta}$, which is an approximation of the true variance of the estimator $\hat{\beta}$. The empirical coverages of 95% confidence intervals are also reported. As a result of the identity $\beta = \alpha/(\sqrt{2}\sigma)$, β can also be estimated through the estimation of α and σ in (3.26) by means of least squares (LS) and maximum likelihood (ML). In the normal linear regression model LS and ML give the same point estimator of α, but their estimators of the residual variance σ^2 are different up to a factor $(n-1)/n$. Hence, the methods give different estimators of β, particularly in small samples.

From Table 3.1 we conclude that, for all sample sizes, the parametric estimators are much more efficient as compared to the PIM estimators. When α or the range of X increases, the difference in efficiency, however, decreases.

Table 3.2 shows the results of simulations of heteroscedastic data with $\sigma(X) = \sigma\sqrt{X}$, where $\sigma = 1$ or $\sigma = 5$. The corresponding PIM is given by

$$\Phi^{-1}\left[\mathrm{P}\left(Y \preccurlyeq Y' \mid X, X'\right)\right] = \beta\frac{X' - X}{\sqrt{X' + X}},$$

where $\beta = \alpha/\sigma$.

All three estimators are nearly unbiased, particularly for sample sizes of $n = 50$ or more. Surprisingly the semiparametric PIM estimator is more efficient than LS and ML when $\alpha = 10$, $u = 1$, and $\sigma = 1$. As already discussed in Section 2.4, this is a consequence of the many ties in the PIM estimates.

We also examine empirically the power of tests for testing the no-effect null hypothesis in terms of the PI. In particular, we will look at the PIM,

$$g\left[\mathrm{P}\left(Y \preccurlyeq Y' \mid X_1, X_2, X_1', X_2'\right)\right] = \beta_1(X_1' - X_1) + \beta_2(X_2' - X_2), \tag{3.27}$$

where X_1 and X_1' are $0/1$ dummies that, for example, code for two treatment groups, active treatment and placebo, say, and X_2 and X_2' refer to a continuous covariate, age, say. The no-treatment-effect null hypothesis, $H_0 : \beta_1 = 0$, is of interest. It expresses that, among patients of the same age, the chance that a treated patient's outcome is higher than the outcome of an untreated patient is 50%. To our knowledge there are hardly any statistical tests described in the literature for this problem. In Section 1.4 we have discussed the most important competitors. In this simulation study we have opted for the test of Dodd and Pepe (2003), as discussed in Section 3.4. Their test is also semiparametric, but it is limited to testing the no-treatment-effect null hypothesis in the presence of covariates, whereas our framework allows for a broad range of extensions. Their method can be embedded in a particular PIM,

$$g\left[\mathrm{P}\left(Y \preccurlyeq Y' \mid X_1 < X_1', X_2, X_2'\right)\right] = \delta_1 + \delta_2(X_2' - X_2), \tag{3.28}$$

which does not allow for $X_1 = X_1'$. Their test is based on the test statistic $B = \hat{\delta}_1/S_1$, where $\hat{\delta}_1$ is their estimator of δ_1 and S_1 is an estimator of the standard error of $\hat{\beta}_1$ which is obtained by the bootstrap. For computational reasons we limit the bootstrap procedure to 200 runs.

Table 3.1: Simulation results for the normal linear homoscedastic model, based on 1000 Monte Carlo runs. β is the true parameter, $\mathrm{Av}(\hat{\beta})$ the average of the β estimates according to the semiparametric PIM theory (PIM), $\mathrm{Var}(\hat{\beta})$ the sample variance of the simulated $\hat{\beta}$, $\mathrm{Av}(\hat{S}_{\hat{\beta}})$ the average of the sandwich variance estimates according to the semiparametric PIM theory, EC the empirical coverage of a 95% confidence interval for β, $\mathrm{Av}(\bar{\beta})$ the average of the least-squares (LS) estimates, $\mathrm{Var}(\bar{\beta})$ the sample variance of the simulated $\bar{\beta}$, $\mathrm{Av}(\tilde{\beta})$ the average of the maximum-likelihood (ML) estimates and $\mathrm{Var}(\tilde{\beta})$ the sample variance of the simulated $\tilde{\beta}$.

				PIM				LS		ML	
α	u	σ	β	$\mathrm{Av}(\hat{\beta})$	$\mathrm{Var}(\hat{\beta})$	$\mathrm{Av}(\hat{S}_{\hat{\beta}})$	EC	$\mathrm{Av}(\bar{\beta})$	$\mathrm{Var}(\bar{\beta})$	$\mathrm{Av}(\tilde{\beta})$	$\mathrm{Var}(\tilde{\beta})$
$n = 25$											
1	1	1	0.707	0.736	0.33900	0.27877	92.0	0.729	0.06814	0.744	0.07098
1	1	5	0.141	0.130	0.32438	0.27008	92.8	0.135	0.05817	0.138	0.06059
1	10	1	0.707	0.721	0.00990	0.01184	93.0	0.729	0.01214	0.745	0.01265
1	10	5	0.141	0.149	0.00332	0.00248	90.2	0.145	0.00106	0.148	0.00111
10	1	1	7.071	7.309	1.55061	1.22519	85.7	7.320	1.36451	7.471	1.42136
10	1	5	1.414	1.463	0.40365	0.29884	88.7	1.444	0.10516	1.474	0.10954
$n = 50$											
1	1	1	0.707	0.736	0.16640	0.15048	92.9	0.718	0.03465	0.725	0.03536
1	1	5	0.141	0.148	0.14905	0.14542	93.5	0.148	0.02759	0.150	0.02815
1	10	1	0.707	0.714	0.00615	0.00634	94.4	0.714	0.00568	0.721	0.00580
1	10	5	0.141	0.147	0.00148	0.00139	93.4	0.145	0.00052	0.146	0.00054
10	1	1	7.071	7.224	0.78701	0.67363	89.1	7.171	0.59224	7.244	0.60433
10	1	5	1.414	1.465	0.18646	0.16191	92.5	1.439	0.05014	1.454	0.05117
$n = 200$											
1	1	1	0.707	0.716	0.03803	0.03942	95.3	0.710	0.00798	0.712	0.00802
1	1	5	0.141	0.145	0.04048	0.03817	94.8	0.145	0.00673	0.146	0.00676
1	10	1	0.707	0.709	0.00179	0.00170	94.3	0.709	0.00128	0.710	0.00128
1	10	5	0.141	0.141	0.00037	0.00036	95.6	0.141	0.00013	0.142	0.00013
10	1	1	7.071	7.110	0.19105	0.17489	93.2	7.089	0.14540	7.107	0.14613
10	1	5	1.414	1.427	0.04400	0.04308	95.0	1.421	0.01164	1.424	0.01170

Table 3.2: Simulation results for the normal linear heteroscedastic model, based on 1000 Monte Carlo runs. β is the true parameter, $\mathrm{Av}(\hat{\beta})$ the average of the β estimates according to the semiparametric PIM theory (PIM), $\mathrm{Var}(\hat{\beta})$ the sample variance of the simulated $\hat{\beta}$, $\mathrm{Av}(\hat{S}_{\hat{\beta}})$ the average of the sandwich variance estimates according to the semiparametric PIM theory, EC the empirical coverage of a 95% confidence interval for β, $\mathrm{Av}(\bar{\beta})$ the average of the least-squares (LS) estimates, $\mathrm{Var}(\bar{\beta})$ the sample variance of the simulated $\bar{\beta}$, $\mathrm{Av}(\tilde{\beta})$ the average of the maximum-likelihood (ML) estimates and $\mathrm{Var}(\tilde{\beta})$ the sample variance of the simulated $\tilde{\beta}$.

				PIM				LS		ML	
α	u	σ	β	$\mathrm{Av}(\hat{\beta})$	$\mathrm{Var}(\hat{\beta})$	$\mathrm{Av}(\hat{S}_{\hat{\beta}})$	EC	$\mathrm{Av}(\bar{\beta})$	$\mathrm{Var}(\bar{\beta})$	$\mathrm{Av}(\tilde{\beta})$	$\mathrm{Var}(\tilde{\beta})$
$n = 25$											
1	1	1	1	1.052	0.34771	0.27673	91.2	1.097	0.12945	1.053	0.10286
1	1	5	0.2	0.192	0.31399	0.26122	92.8	0.206	0.09299	0.198	0.08389
1	10	1	1	1.045	0.05487	0.03584	90.1	1.096	0.05970	1.051	0.03285
1	10	5	0.2	0.206	0.02317	0.01884	92.2	0.219	0.01163	0.209	0.00963
10	1	1	10	9.268	0.50991	1.75345	93.9	10.987	4.94362	10.563	2.79136
10	1	5	2	2.080	0.46761	0.32145	88.4	2.169	0.27392	2.086	0.17884
10	10	5	2	2.088	0.13541	0.10231	85.5	2.209	0.23559	2.114	0.12025
$n = 50$											
1	1	1	1	1.032	0.17125	0.15259	92.9	1.044	0.06014	1.026	0.05177
1	1	5	0.2	0.210	0.14692	0.14205	94.4	0.214	0.03981	0.211	0.03839
1	10	1	1	1.025	0.02554	0.01967	90.0	1.039	0.02407	1.019	0.01525
1	10	5	0.2	0.208	0.01086	0.01034	94.4	0.212	0.00533	0.208	0.00464
10	1	1	10	9.410	0.22462	0.95066	96.0	10.471	1.99398	10.244	1.18719
10	1	5	2	2.063	0.20438	0.17953	92.5	2.093	0.11833	2.056	0.08404
10	10	5	2	2.046	0.06469	0.05539	91.4	2.089	0.08120	2.047	0.04754
$n = 200$											
1	1	1	1	1.010	0.03905	0.04005	95.1	1.010	0.01361	1.006	0.01161
1	1	5	0.2	0.204	0.03891	0.03740	95.2	0.206	0.00939	0.205	0.00921
1	10	1	1	1.006	0.00568	0.00557	93.6	1.013	0.00557	1.005	0.00345
1	10	5	0.2	0.198	0.00271	0.00275	95.8	0.201	0.00118	0.200	0.00111
10	1	1	10	9.576	0.04093	0.26446	97.1	10.098	0.47093	10.051	0.28679
10	1	5	2	2.016	0.05006	0.04843	94.1	2.022	0.02577	2.014	0.01907
10	10	5	2	2.007	0.01548	0.01465	94.1	2.020	0.01913	2.008	0.01061

Data are generated according to

$$Y_i = \alpha_1 X_{1i} + \alpha_2 X_{2i} + \varepsilon_i, \quad \varepsilon_i \stackrel{d}{=} \mathrm{N}(0,1),$$

and are analyzed by least-squares in a marginal linear model with conditional mean

$$\mathrm{E}\left(Y \mid X_1, X_2\right) = \gamma_1 X_1 + \gamma_2 X_2,$$

by the PIM (3.27) with probit link function, and by bootstrap test (BT) based on (3.28) with probit link. The LS results serve as an indication of the best powers that can be expected. The `geepack` R package (Højsgaard et al., 2005) is used to fit the marginal model which allows using sandwich variance estimates in the construction of the LS-based test.

The following design is considered. The covariate X_1 is a $0/1$ balanced dummy variable, X_2 is equally spaced over $[0.1, 10]$, α_1 takes on the values 0, 0.5, and 1 while α_2 is fixed at 1. Sample sizes of 20, 50, and 200 are considered. All tests described above are applied for testing $H_0 : \gamma_1 = 0$ versus $H_1 : \gamma_1 \neq 0$, $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$, or $H_0 : \delta_1 = 0$ versus $H_1 : \delta_1 \neq 0$. All tests are applied at the 5% level of significance. Table 3.3 shows the empirical powers based on 1000 Monte Carlo simulation runs.

For a sample size $n = 20$ the BT-based test shows complete breakdown by showing virtually no power, and the tests based on the PIM and LS are liberal. When $n = 50$ the PIM-based test has a size not too far away from the nominal level of 5%, while the LS-based test is slightly liberal and the BT-based test is still conservative. When $n = 200$ all tests are nearly unbiased. The powers of the tests in the PIM framework are generally larger than those of the BT-based test; this can perhaps be attributed to limited number of bootstrap runs. The test based on LS are slightly more powerful, as expected.

3.8.2 The exponential model

Let $Y_i \mid X_i \stackrel{d}{=}$ Exponential$[\gamma(X_i)]$ with

$$\gamma(X_i) = \exp(\alpha X_i), \quad i = 1, \ldots, n. \tag{3.29}$$

Sample sizes of $n = 25$, $n = 50$, and $n = 200$ are considered. The predictor X takes equally spaced values in the interval $[0.1, u]$ where $u = 1$ or 10 and α takes on the value 0.1 or -2. The

Table 3.3: Empirical powers (%) based on 1000 Monte Carlo runs for the normal linear model: PIM, least-squares (LS), or bootstrap (BT)

α_1	PIM	LS	BT	PIM	LS	BT	PIM	LS	BT
	$n = 20$			$n = 50$			$n = 200$		
0.0	7.6	9.5	0.0	5.7	6.4	2.0	4.7	5.3	4.2
0.5	15.0	27.3	0.0	35.3	50.6	24.4	93.4	98.0	91.0
1.0	45.9	72.3	0.2	89.5	97.5	78.7	100.0	100.0	100.0

corresponding PIM is

$$\text{logit}\left[\mathrm{P}\left(Y \preccurlyeq Y' \mid X, X'\right)\right] = \beta(X - X'), \tag{3.30}$$

where $\beta = \alpha$. Table 3.4 gives the results when model (3.30) is analyzed with the semiparametric PIM theory, resulting in $\hat{\beta}$. As a result of the identity $\beta = \alpha$, the parameter β can also be estimated based on the semiparametric proportional hazards theory, resulting in $\bar{\beta}$. The R package `survival` (Therneau and Lumley, 2010) is used for fitting the proportional hazards model. The estimator of β based on maximum likelihood theory is denoted by $\tilde{\beta}$. From Table 3.4 we conclude that the PIM estimator of β and the sandwich variance estimator are nearly unbiased for sample sizes of 50 and more. The empirical coverages of the 95% confidence intervals are close to their nominal level for sample sizes of 50 and more.

To examine the power, let $Y_i \mid (X_{1i}, X_{2i}) \stackrel{d}{=} \text{Exponential}[\gamma(X_{1i}, X_{2i})]$, with

$$\gamma(X_1, X_2) = \exp(\alpha_1 X_1 + \alpha_2 X_2).$$

The data are analyzed by partial likelihood in a proportional hazards model with hazards function

$$\lambda(X) = \exp\left(\gamma_1 X_1 + \gamma_2 X_2\right),$$

by the PIM (3.27) with logit link and by the BT test based on (3.28) with logit link. The powers with the partial-likelihood method may be considered as a semiparametric competitor of PIM, although the proportional hazards model does not coincide with the class of PIMs: they express different restrictions on the conditional outcome distribution. The same design is considered as for the power study based on the normal linear model. All tests are applied at the 5% level

Table 3.4: Simulation results for the exponential model, based on 1000 Monte Carlo runs. β is the true parameter, $\mathrm{Av}(\hat{\beta})$ the average of the β estimates using the semiparametric PIM theory (PIM), $\mathrm{Var}(\hat{\beta})$ the sample variance of the simulated $\hat{\beta}$, $\mathrm{Av}(\hat{S}_{\hat{\beta}})$ the average of the sandwich variance estimates using the semiparametric PIM theory, EC the empirical coverage of a 95% confidence interval for β, $\mathrm{Av}(\bar{\beta})$ the average of the semiparametric proportional hazards (PH) estimates, $\mathrm{Var}(\bar{\beta})$ the sample variance of the simulated $\bar{\beta}$, $\mathrm{Av}(\tilde{\beta})$ the average of the maximum-likelihood (ML) estimates and $\mathrm{Var}(\tilde{\beta})$ the sample variance of the simulated $\tilde{\beta}$.

				PIM				PH		ML	
α	u	σ	β	$\mathrm{Av}(\hat{\beta})$	$\mathrm{Var}(\hat{\beta})$	$\mathrm{Av}(\hat{S}_{\hat{\beta}})$	EC	$\mathrm{Av}(\bar{\beta})$	$\mathrm{Var}(\bar{\beta})$	$\mathrm{Av}(\tilde{\beta})$	$\mathrm{Var}(\tilde{\beta})$
$n = 25$											
-2	1	1	-2	-2.226	1.19067	0.89060	90.4	-2.178	0.87454	-1.963	0.10657
0.1	10	1	0.1	0.110	0.00902	0.00630	91.1	0.110	0.00720	0.104	0.00130
$n = 50$											
-2	1	1	-2	-2.083	0.54166	0.47159	93.7	-2.083	0.41978	-1.986	0.05564
0.1	10	1	0.1	0.103	0.00337	0.00333	95.0	0.103	0.00262	0.103	0.00060
$n = 200$											
-2	1	1	-2	-2.023	0.12394	0.12220	94.7	-2.018	0.08917	-1.999	0.01460
0.1	10	1	0.1	0.098	0.00090	0.00087	94.6	0.100	0.00072	0.100	0.00015

of significance. Table 3.5 shows the empirical powers based on 1000 Monte Carlo simulation runs. For a sample size $n = 20$ the BT based test shows complete breakdown by showing virtually no power, and the tests based on the PIM is liberal. When $n = 50$ the PIM based test is still liberal, while the PL based test is only slightly liberal and the BT based test is still conservative. When $n = 200$ all tests are nearly unbiased. The powers of the tests for $n = 200$ (i.e. when all tests correctly control the type I error) in the PIM framework are slightly larger than those of BT based test, while the test based on PL is most powerful. Note that the PL theory is semiparametrically efficient within the class of proportional hazards models, while the PIM theory is not guaranteed to be efficient within the class of PIMs. The semiparametric efficiency of PIMs is studied in more detail in Chapter 7.

Table 3.5: Empirical powers ($\%$) based on 1000 Monte Carlo runs for the exponential model: PIM, partial-likelihood (PL), or bootstrap (BT)

α_1	PIM	PL	BT	PIM	PL	BT	PIM	PL	BT
	$n = 20$			$n = 50$			$n = 200$		
0.0	9.7	4.3	0.0	8.1	6.4	3.3	4.8	4.7	4.1
0.5	22.7	16.2	0.0	30.1	38.4	17.5	77.1	93.3	75.3
1.0	42.3	44.4	0.0	76.0	89.2	57.6	100.0	100.0	100.0

3.8.3 The cumulative logit model

We consider a logistic linear model through the discretization of a continuous latent variable. In particular, the latent outcome variable is modelled as

$$Z_i = \alpha_1 X_{1i} + \alpha_2 X_{2i} + \varepsilon_i,$$

where ε_i are i.i.d. standard logistic. The latent outcome variable Z_i is discretized into four ordered categories as described in section 6.2 of Agresti (2007). The resulting ordinal outcome is denoted by Y_i. The data are analyzed by maximum likelihood in the cumulative logit model

$$\text{logit}\left[\text{P}\left(Y \leq j \mid X_1, X_2\right)\right] = \mu_j + \gamma_1 X_1 + \gamma_2 X_2,$$

and by the PIM (3.27) with logit link and by the BT test based on (3.28) with logit link. Since there is no direct relation between the PIM model parameters and the cumulative logistic model,

we can not compare efficiency of different estimators, because they estimate different population parameters. We only consider the logistic data-generating model for power comparison. The R package `MASS` (Venables and Ripley, 2002) is used to fit the cumulative logit model.

The same design is considered as for the power study based on the normal linear model. All tests are applied at the 5% level of significance. Table 3.6 shows the empirical powers based on 1000 Monte Carlo simulation runs. The PIM-based test is liberal for all sample sizes, while the BT-based test and the ML-based test have sizes close to the nominal level of 5%. The powers of all tests are comparable, especially for a sample size of $n = 200$. However, since the PIM-based test is liberal, the corresponding powers have no unambiguous interpretation.

Table 3.6: Empirical powers ($\%$) based on 1000 Monte Carlo runs for the logistic linear model: PIM, maximum-likelihood (ML), or bootstrap (BT).

α_1	PIM	ML	BT	PIM	ML	BT	PIM	ML	BT
	$n = 20$			$n = 50$			$n = 200$		
0.0	10.8	4.5	2.2	7.7	5.1	4.9	7.1	6.1	6.4
0.5	14.1	7.3	2.8	18.3	15.6	12.9	36.8	37.5	35.6
1.0	25.3	16.8	4.8	39.7	37.5	33.4	88.5	88.8	87.4

3.9 Discussion

The relationship between PIMs and several regression methods is explored. For the linear and Cox proportional hazards models there are direct relations between the model parameters. Starting from these models a PIM can be constructed, but, in general, the opposite does not hold implying that a PIM imposes less restrictions on the conditional outcome distribution.

The AUC regression model and the concordance index can be embedded within the PIM framework and can therefore be considered as special cases of a PIM. The flexible PIM modelling framework allows extending these methods to more complicated designs.

There is no direct relationship between the model parameters of a PIM and those of rank regression, but there are some interesting similarities: both estimation methods make use of pseudo-

observations. For rank regression the model parameters are within the pseudo-observations, while for PIM the model parameters are outside the pseudo-observations.

Both the PIM and the cumulative logit model are regression methods that allow analyzing ordinal outcomes. Where the former has an interpretation in terms of the odds, the latter has an interpretation in terms of the odds ratio. Both interpretations are distinct but share some similarities. The PIM also allows to include ordinal predictors with many levels at the cost of only single model parameter.

A simulation study is considered to empirically examine the performance of some of these methods. The simulation results demonstrate that the theoretical properties of the PIM parameter and variance estimators apply well to moderately sized samples, but that there is a substantial efficiency loss as compared to parametric estimators. The PIM imposes weaker restrictions on the conditional outcome distribution as compared to more parametric methods and if these parametric assumptions hold – which is the case in our simulation study – the former will often underperform as compared to the latter because it does not fully exploit all information.

Chapter 4

Relationship with rank tests

The content of this chapter is primarily based on the manuscript

De Neve, J., Thas, O., and Ottoy, J.P. (2013) A semiparametric framework for rank tests for factorial designs. *Submitted*.

4.1 Introduction

The Wilcoxon–Mann–Whitney (WMW) (Mann and Whitney, 1947; Wilcoxon, 1945) and Kruskal–Wallis (KW) (Kruskal and Wallis, 1952) tests are well known and popular rank tests to analyze two- and K-sample designs. These rank tests are distribution-free, robust, intuitively appealing, and do not necessarily focus on the mean outcome. For the WMW test, for example, the alternative hypothesis is expressed in terms of the probability $\mathrm{P}(Y_1 \preccurlyeq Y_2)$, where Y_1 (Y_2) denotes the outcome of the first (second) group. It is the probability that a random observation of the second group exceeds a random observation of the first group. The WMW null hypothesis implies $\mathrm{P}(Y_1 \preccurlyeq Y_2) = 0.5$. The alternative hypothesis of the KW test can be expressed in terms of the probabilities, $\mathrm{P}(Y_\cdot \preccurlyeq Y_i)$, where Y_i denotes the outcome in group $i = 1, \ldots, K$, and $Y_\cdot$ the outcome associated with the marginal outcome distribution; see, for example, section 9.6.1 in Thas (2009). It is the probability that a random observation in group i exceeds a random observation of the marginal distribution. Under additional assumptions, such as location-shift, the alternative can also be expressed in terms of the mean or median, and for a given family of distributions, rank tests may be constructed to be the locally most powerful rank test for

testing equality of means. For example, for the logistic distribution the WMW test is optimal in this sense. We refer to the textbooks of Hájek et al. (1999) and Lehmann (1998) for extensive overviews of these theories. Since this optimality theory requires strong parametric assumptions, and since most statisticians use rank tests when no such assumptions can be made or assessed, we will not work under location-shift, but under the less restrictive assumptions imposed by a PIM.

After the introduction of the first rank tests for the 2- and K-sample designs, a vast number of rank tests for more complicated designs have been developed; see, for example, Hollander and Wolfe (1999) for a broad overview. Despite the many papers and textbooks covering this topic, a non-experienced user is unlikely to use most of these rank tests, because 1) some of the tests have no standard name which makes finding them difficult, 2) their construction is often quite complicated and only valid for a particular design, 3) the interpretation on population level is not always understood, and 4) the majority of these tests is not implemented in standard statistical software. For classical parametric tests with focus on the mean outcome, such as the two-sample t- or ANOVA F-test, this barrier is circumvented because they arise naturally from the General Linear Model (GLM) framework. Hence, for more complicated designs the appropriate GLM may be formulated, resulting in the correct t- or F-test. Basic knowledge on GLMs is often sufficient for analyzing data from a variety of designs. Moreover, the GLM is available in most statistical software packages.

In this chapter we situate a large class of rank tests within the PIM methodology. The PIM can in a way be seen as the rank-equivalent of the GLM, but should not be confused with the rank-transform approach of Conover and Iman (1981). We will show that a transformation to the pseudo-observations is more flexible than the rank-transform, and by embedding the method in the PIM framework we can relate the tests to parameters with a well defined interpretation on population level in terms of the PI. Depending on the parametrization of the model, we can establish a simple connection between the PIM and the WMW, KW, Friedman (Friedman, 1937), Mack–Skillings (MS) (Mack and Skillings, 1980), Brown–Hettmansperger (BH) (Brown and Hettmansperger, 2002), Jonckheere–Terpstra (JT) (Jonckheere, 1954; Terpstra, 1952), and Mack–Wolfe (MW) (Mack and Wolfe, 1981) rank tests.

The PIM framework also allows for developing new rank tests for more complicated designs, even when a continuous confounder or covariate is present. In addition to hypothesis testing,

PIMs naturally model effect sizes with an informative interpretation. Estimates of the parameters assist on reporting the effect sizes. The PIM is thus the natural model to describe the restrictions on the outcome distributions for which rank tests are the natural tests for null hypotheses involving the PIM parameters.

In Section 4.2 we introduce notation and in Section 4.3 we propose a PIM parametrization for factorial designs and establish the connection with the KW, Friedman, and MS tests. We do not only generate existing rank tests, but for each of the designs considered, we also propose a different type of rank test. In Section 4.4 we consider a second PIM parametrization and demonstrate the connection with the WMW, BH, JT, and MW tests. For both sections, the PIM is restricted to factorial designs with one predictor and one blocking variable. It is demonstrated in Section 4.5 how rank tests can be extended to control for a continuous covariate. Section 4.6 extends the one-way to the two-way layout. In Section 4.8 we evaluate the performance of some new tests in a simulation study and in Section 4.9 we illustrate with an example how the model can be used for one continuous and multiple categorical predictors. Section 4.10 gives the conclusions and discussion.

4.2 Notation

For the factorial design we write the predictor as $\boldsymbol{X}^T = (X, B)$, where X is a factor variable referring to groups or treatments, and B is a blocking factor which is here considered as nuisance. Without loss of generality we assume that X takes K distinct values, say $1, \ldots, K$, and B takes L distinct values, say $1, \ldots, L$. The number of replicates for $X = i$ and $B = j$ is denoted by n_{ij} and the total sample size is denoted by $N = \sum_{i=1}^{K} \sum_{j=1}^{L} n_{ij}$. Let F_{ij} denote the distribution function of Y given $X = i$ and $B = j$. In the absence of blocks, set $B = 1$ and let n_i denote the number of replicates for $X = i$ and F_i the distribution function of Y given $X = i$.

Sometimes it will be easier to work with the classical ANOVA notation. Throughout the chapter it will be clear from the context when which notation is used; we therefore use Y again as the outcome variable. In particular, for the one way layout Y_{ij} denotes a random outcome variable in treatment group $i = 1, \ldots, K$ and block $j = 1, \ldots, B$. The index j becomes obsolete in the absence of blocks. We use $Y_{\cdot j}$ to denote the random outcome variable whose distribution is marginalized over the treatment groups, but still conditional on block j. For the reader's

convenience we resume the general PIM, as defined by (2.11), for a random sample of i.i.d. observations $\{(Y_i, \boldsymbol{X}_i) \mid i = 1, \ldots, n\}$

$$\mathrm{P}\left(Y_i \preccurlyeq Y_j \mid \boldsymbol{X}_i, \boldsymbol{X}_j\right) = m(\boldsymbol{X}_i, \boldsymbol{X}_j; \boldsymbol{\beta}) = g^{-1}(\boldsymbol{Z}_{ij}^T \boldsymbol{\beta}), \quad (\boldsymbol{X}_i, \boldsymbol{X}_j) \in \mathcal{X}_n, \tag{4.1}$$

where $\boldsymbol{Z}_{ij}$ is a function of $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ and $g(\cdot)$ a link function which, for this chapter, will often be the identity link $g(u) = u$. To distinguish between the notation and model as in (4.1) and the ANOVA form, we refer to the former as the *regression model*, whereas models with the ANOVA notation will be referred to as the *ANOVA model*. Just like with classical linear regression models, ANOVA models will have to be translated into regression models with dummy variables for the coding of the factors, before the estimation of the parameters.

4.3 The marginal probabilistic index model

As a first model we define the *marginal PIM* for the K-sample design in the absence of blocks. It is marginal in the sense that we only condition on one treatment within the PI, i.e. $\mathrm{P}\left(Y_i \preccurlyeq Y_j \mid X_j\right)$. This PI refers to the distribution of the outcome of observation j conditional on the covariate $(Y_j \mid X_j)$, and the marginal outcome distribution of an observation i (Y_i). In terms of the ANOVA notation for $X_j = k$ this becomes $\mathrm{P}\left(Y_{\cdot} \preccurlyeq Y_k\right)$, with Y_k the outcome in group k. Consider the marginal PIM model in ANOVA form,

$$\mathrm{P}\left(Y_{\cdot} \preccurlyeq Y_k\right) = \alpha_k. \tag{4.2}$$

The interpretation of α_k is immediate: it is the probability that a random observation of group k exceeds a random observation of the marginal distribution. The corresponding PIM regression model is obtained upon defining

$$\boldsymbol{Z}_{ij}^T = \left[\mathrm{I}\left(X_j = 1\right), \ldots, \mathrm{I}\left(X_j = K\right)\right] \tag{4.3}$$

for all pairs of predictors (X_i, X_j) and by considering the identity link. Let $\boldsymbol{\alpha}^T = (\alpha_1, \ldots, \alpha_K)$. Model (4.2) now becomes

$$\mathrm{P}\left(Y_i \preccurlyeq Y_j \mid X_j\right) = \boldsymbol{Z}_{ij}^T \boldsymbol{\alpha}, \quad (X_i, X_j) \in \mathcal{X}_0, \tag{4.4}$$

which we define for the no-ordering restriction $\mathcal{X}_0$. This model is closely related to the comparison mid-probability index as discussed in Parzen and Mukhopadhyay (2012a,b). Our model also follows from the nonparametric model of Akritas and Arnold (1994); see Section 4.7.

Let $\hat{\boldsymbol{\alpha}}$ denote the estimator of $\boldsymbol{\alpha}$, defined as the solution of the estimating equations (2.15), which, for a general PIM (4.1) with parameter $\boldsymbol{\beta}$, are given by

$$\sum_{(i,j)\in\mathcal{I}_n} \boldsymbol{A}(\boldsymbol{Z}_{ij};\boldsymbol{\beta})[\mathrm{I}\left(Y_i \preccurlyeq Y_j\right) - g^{-1}(\boldsymbol{Z}_{ij}^T\boldsymbol{\beta})] = \mathbf{0}. \tag{4.5}$$

Since the indentity link is used in (4.4), we suggest to set $\boldsymbol{A}(\boldsymbol{Z}_{ij};\boldsymbol{\beta}) = \boldsymbol{Z}_{ij}$ so as to obtain the ordinary least squares solution. The following lemma and corollary give the explicit form of $\hat{\boldsymbol{\alpha}}$ as a linear combination of the pseudo-observations. Note that Lemma 3 applies more generally to all regression PIMs with identity link and $\boldsymbol{A}(\boldsymbol{Z}_{ij};\boldsymbol{\beta}) = \boldsymbol{Z}_{ij}$.

Lemma 3. *The estimator of $\boldsymbol{\beta}$ in (4.1) with identity link function, defined as the solution of (4.5) with $\boldsymbol{A}(\boldsymbol{Z}_{ij};\boldsymbol{\beta}) = \boldsymbol{Z}_{ij}$, is given by*

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{I}_p,$$

with $\boldsymbol{I}_p$ the $|\mathcal{I}_n|$-vector of pseudo-observations $\mathrm{I}\left(Y_i \preccurlyeq Y_j\right)$, $(i,j) \in \mathcal{I}_n$ and $\boldsymbol{Z}$ the $|\mathcal{I}_n| \times p$ matrix with rows $\boldsymbol{Z}_{ij}^T$ corresponding to the pseudo-observations in $\boldsymbol{I}_p$. This estimator is thus an ordinary least squares *(OLS) estimator.*

The proof of Lemma 3 is immediate by recognizing that the estimation equations give the OLS.

Corollary 1. *The OLS estimator of an individual α_k in (4.2) or (4.4) may be written as*

$$\hat{\alpha}_k = \frac{1}{Nn_k}\sum_{i=1}^{N}\sum_{j=1}^{N} c_{kj}\mathrm{I}\left(Y_i \preccurlyeq Y_j\right), \tag{4.6}$$

where $c_{kj} = \mathrm{I}\left(X_j = k\right)$.

In the remainder of this section we assume that there are no ties among the sample outcome observations. This is to avoid lengthy formulas of the rank statistics. All results, however, can be extended to allow for ties.

The next lemma provides the covariance structure of the pseudo-observations. It forms the basis of many results presented later.

Lemma 4. *Let Y_i, Y_j, Y_k, and Y_l denote four i.i.d. random variables, then*

- $\mathrm{Var}[\mathrm{I}\left(Y_i \preccurlyeq Y_j\right)] = 1/4$,

- $\mathrm{Cov}[\mathrm{I}\left(Y_i \preccurlyeq Y_j\right), \mathrm{I}\left(Y_i \preccurlyeq Y_k\right)] = \mathrm{Cov}[\mathrm{I}\left(Y_i \preccurlyeq Y_j\right), \mathrm{I}\left(Y_k \preccurlyeq Y_j\right)] = 1/12$,
- $\mathrm{Cov}[\mathrm{I}\left(Y_i \preccurlyeq Y_j\right), \mathrm{I}\left(Y_k \preccurlyeq Y_i\right)] = \mathrm{Cov}[\mathrm{I}\left(Y_i \preccurlyeq Y_j\right), \mathrm{I}\left(Y_j \preccurlyeq Y_k\right)] = -1/12$,
- $\mathrm{Cov}[\mathrm{I}\left(Y_i \preccurlyeq Y_j\right), \mathrm{I}\left(Y_k \preccurlyeq Y_l\right)] = 0$.

Proof. From

$$\mathrm{P}\left(Y_i \preccurlyeq Y_j\right) = \mathrm{P}\left(Y_i < Y_j\right) = \frac{1}{2},$$

and

$$\mathrm{P}[Y_i \preccurlyeq \min(Y_j, Y_k)] = \mathrm{P}[Y_i < \min(Y_j, Y_k)] = \frac{1}{3},$$

the statement follows by recognizing that

$$\mathrm{Var}[\mathrm{I}\left(Y_i < Y_j\right)] = \mathrm{E}[\mathrm{I}\left(Y_i < Y_j\right)] - \mathrm{E}[\mathrm{I}\left(Y_i < Y_j\right)]^2,$$

and

$$\mathrm{Cov}[\mathrm{I}\left(Y_i < Y_j\right), \mathrm{I}\left(Y_i < Y_k\right)] = \mathrm{E}[\mathrm{I}\left(Y_i < Y_j\right)\mathrm{I}\left(Y_i < Y_k\right)] - \mathrm{E}[\mathrm{I}\left(Y_i < Y_j\right)]\mathrm{E}[\mathrm{I}\left(Y_i < Y_k\right)],$$

where

$$\mathrm{I}\left(Y_i < Y_j\right)\mathrm{I}\left(Y_i < Y_k\right) = \mathrm{I}[Y_i < \min(Y_j, Y_k)].$$

□

4.3.1 The K-sample design

The following lemma gives the covariance matrix of $\hat{\boldsymbol{\alpha}}$ under the null hypothesis of equal distributions. The proof follows directly from combining Corollary 1 and Lemma 4.

Lemma 5. *If $H_0 : F_1 = \cdots = F_K$ is true, then the variance of $\hat{\alpha}_l$ in (4.6) associated with PIM (4.2) or (4.4), is given by*

$$\mathrm{Var}\left(\hat{\alpha}_l\right) = \frac{(N - n_l)(N + 1)}{12N^2 n_l},$$

and the covariance by

$$\mathrm{Cov}\left(\hat{\alpha}_k, \hat{\alpha}_l\right) = -\frac{N + 1}{12N^2}, \quad k \neq l.$$

Let $\mathbf{1}$ denote the unit vector of length K. From Lemma 5 if follows that, under H_0,

$$\boldsymbol{\Sigma}_0 := \mathrm{Cov}\left(\hat{\boldsymbol{\alpha}}\right) = \frac{N + 1}{12N}\mathrm{diag}(n_1^{-1}, \ldots, n_K^{-1})\boldsymbol{M}, \tag{4.7}$$

with

$$\boldsymbol{M} = \boldsymbol{I} - \frac{1}{N}\text{diag}(n_1, \ldots, n_K)\mathbf{1}\mathbf{1}^T,$$

where $\boldsymbol{I}$ denotes the $K \times K$ identity matrix. The following theorem establishes the relationship between the marginal PIM and the KW test for which the test statistic is given by

$$\text{KW}_s := \frac{12}{N(N+1)} \sum_{l=1}^{K} n_l \left(\bar{R}_l - \frac{N+1}{2} \right)^2, \tag{4.8}$$

where $\bar{R}_l$ denotes the average rank of the sample observations in group $X = l$, for which the ranking is performed in the pooled sample. Let $\boldsymbol{B}^-$ denote a generalized inverse of a square matrix $\boldsymbol{B}$.

Theorem 3 (Kruskal–Wallis)**.** *For the K-sample design let $\hat{\boldsymbol{\alpha}}$ denote the estimator of $\boldsymbol{\alpha}^T = (\alpha_1, \ldots, \alpha_K)$ in (4.2) or (4.4), given by (4.6), and let $\boldsymbol{\Sigma}_0$ denote its covariance matrix under the null hypothesis of equal distributions (4.7), then*

$$\left(\hat{\boldsymbol{\alpha}} - \frac{1}{2}\mathbf{1} \right)^T \boldsymbol{\Sigma}_0^- \left(\hat{\boldsymbol{\alpha}} - \frac{1}{2}\mathbf{1} \right) = KW_s. \tag{4.9}$$

Proof. Since $\boldsymbol{M}$ is idempotent a generalized inverse of $\boldsymbol{\Sigma}_0$ is given by

$$\boldsymbol{\Sigma}_0^- = \frac{12N}{N+1}\boldsymbol{M}\text{diag}(n_1, \ldots, n_K).$$

Consequently

$$\left(\hat{\boldsymbol{\alpha}} - \frac{1}{2}\mathbf{1} \right)^T \boldsymbol{\Sigma}_0^- \left(\hat{\boldsymbol{\alpha}} - \frac{1}{2}\mathbf{1} \right) = A_1 - A_2,$$

where

$$A_1 = \frac{12N}{N+1} \left(\hat{\boldsymbol{\alpha}} - \frac{1}{2}\mathbf{1} \right)^T \text{diag}(n_1, \ldots, n_K) \left(\hat{\boldsymbol{\alpha}} - \frac{1}{2}\mathbf{1} \right),$$

and

$$A_2 = \frac{12}{N+1} \left(\hat{\boldsymbol{\alpha}} - \frac{1}{2}\mathbf{1} \right)^T \text{diag}(n_1, \ldots, n_K)\mathbf{1}\mathbf{1}^T\text{diag}(n_1, \ldots, n_K) \left(\hat{\boldsymbol{\alpha}} - \frac{1}{2}\mathbf{1} \right).$$

From Corollary 1 it follows that $N^{-1}\sum_{l=1}^{K} n_l\hat{\alpha}_l = 0.5$, therefore

$$\mathbf{1}^T\text{diag}(n_1, \ldots, n_K) \left(\hat{\boldsymbol{\alpha}} - \frac{1}{2}\mathbf{1} \right) = 0,$$

and hence $A_2 = 0$. Furthermore,

$$\hat{\alpha}_l = \frac{1}{Nn_l} \sum_{\{j|X_j=l\}} \left(\sum_{i=1}^{N} \text{I}\left(Y_i \leq Y_j\right) - 0.5 \right) = \frac{1}{N}\left(\bar{R}_l - 0.5\right).$$

It follows that

$$A_1 = \frac{12N}{N+1}\sum_{l=1}^{K} n_l \left(\hat{\alpha}_l - \frac{1}{2}\right)^2 = \frac{12}{N(N+1)}\sum_{l=1}^{K} n_l \left(\bar{R}_l - \frac{N+1}{2}\right)^2.$$

□

Observe that we denote the KW test statistic as KW_s. The subscript s is used to indicate that this is a *score-type* of test, in the sense that the covariance matrix $\boldsymbol{\Sigma}_0$ in (4.9) is only consistent under H_0. The PIM theory provides a sandwich estimator of the covariance matrix, given by Theorem 2, which is also consistent under the alternative and which we denote by $\hat{\boldsymbol{\Sigma}}$. It is thus straightforward to also construct a *Wald-type* KW test by replacing $\boldsymbol{\Sigma}_0$ by $\hat{\boldsymbol{\Sigma}}$ in (4.9). We refer to this statistic as KW_w. Since the marginal PIM parameters are interpretable effect sizes, $\hat{\boldsymbol{\Sigma}}$ can also be used for constructing confidence intervals for these parameters.

The WMW test is a special case of the KW test and is also embedded within the marginal PIM. However, for didactical purposes we postpone the discussion of the WMW test to Section 4.4.

4.3.2 The randomized complete block design

The marginal PIM can be extended to block designs. In ANOVA notation this becomes

$$\text{P}\left(Y_{.l} \preccurlyeq Y_{kl}\right) = \alpha_k, \tag{4.10}$$

where $k = 1, \ldots, K$ refers to the treatment group and $l = 1, \ldots, L$ to the block. The interpretation of α_k is immediate: it is the probability that a random observation in group k exceeds a random observation from the marginal distribution *within the same block*. The corresponding PIM regression model is obtained with $\boldsymbol{Z}_{ij}$ as in (4.3) and $\boldsymbol{\alpha}^T = (\alpha_1, \ldots, \alpha_K)$ as before. Model (4.10) now becomes

$$\text{P}\left(Y_i \preccurlyeq Y_j \mid B_i, X_j, B_j\right) = \boldsymbol{Z}_{ij}^T\boldsymbol{\alpha}, \quad (\boldsymbol{X}_i, \boldsymbol{X}_j) \in \mathcal{X}_n, \tag{4.11}$$

which is now only defined for $\mathcal{X}_n = \{(\boldsymbol{X}_i, \boldsymbol{X}_j) \mid B_i = B_j, i, j = 1, \ldots, N\}$, i.e. we restrict the PI to comparisons within blocks. At this point it is important to stress that the blocking does not result in extra parameters in the model, but it affects the estimating equations through a limitation on the pseudo-observations to include (expressed in the sets $\mathcal{X}_n$ and $\mathcal{I}_n$). Lemma 3 remains valid, and Corollary 2 gives the explicit form of the estimator of $\boldsymbol{\alpha}$.

Corollary 2. *The OLS estimator of an individual α_k in (4.10) or (4.11) may be written as*

$$\hat{\alpha}_k = d_k \sum_{i=1}^{N} \sum_{j=1}^{N} b_{ij} c_{kj} \mathrm{I}\left(Y_i \preccurlyeq Y_j\right), \tag{4.12}$$

where

- $b_{ij} = 1$ *if* $B_i = B_j$ *and* $b_{ij} = 0$ *otherwise,*
- $c_{kj} = \mathrm{I}\left(X_j = k\right)$*,*
- $d_k = \left(\sum_{i=1}^{N} \sum_{j=1}^{N} b_{ij} c_{kj}\right)^{-1}$.

Thas et al. (2012a) also studied statistics of the form of (4.12), but without reference to a PIM.

Consider a randomized complete block (RCB) design for which each treatment-block combination has a fixed number of replicates, i.e. $n_{ij} = n \geq 1$. For testing the null hypothesis $H_0 : F_{1j} = \cdots = F_{Kj}$ $(j = 1, \ldots, L)$, the MS test (Mack and Skillings, 1980) is an appropriate test for this design and its test statistic is given by

$$\mathrm{MS}_s := \frac{12}{K(N+L)} \sum_{l=1}^{K} \left(\bar{R}_l - \frac{N+L}{2}\right)^2, \tag{4.13}$$

where $\bar{R}_l = n^{-1} \sum_{i=1}^{L} \sum_{j=1}^{n} R_{lij}$ and R_{lij} denotes the ranking of j^{th} replicate of the outcome observation of treatment l in block i, where the ranking is performed within blocks. The test statistic asymptotically has a chi-squared null distribution with $K - 1$ degrees of freedom.

The marginal PIM is now only defined for comparisons within blocks and to establish a relationship with the MS test we need the covariance matrix of $\hat{\boldsymbol{\alpha}}$ under H_0. The proof follows directly from combining Corollary 2 and Lemma 4.

Lemma 6. *If $H_0 : F_{1j} = \cdots = F_{Kj}$, $j = 1 \ldots, L$, is true, then the variance of $\hat{\alpha}_l$ in (4.12) associated with PIM (4.10) or (4.11), is given by*

$$\mathrm{Var}\left(\hat{\alpha}_l\right) = \frac{(K-1)(K+n^{-1})}{12nLK^2},$$

and the covariance by

$$\mathrm{Cov}\left(\hat{\alpha}_k, \hat{\alpha}_l\right) = -\frac{K+n^{-1}}{12nLK^2}, \quad k \neq l.$$

The covariance matrix of the vector $\hat{\boldsymbol{\alpha}}$, under H_0, may thus be written as

$$\boldsymbol{\Sigma}_0 := \mathrm{Cov}\left(\hat{\boldsymbol{\alpha}}\right) = \frac{nK+1}{12LKn^2}\boldsymbol{M}, \tag{4.14}$$

where $\boldsymbol{M} = \boldsymbol{I} - K^{-1}\mathbf{1}\mathbf{1}^T$. The following theorem establishes the link between the marginal PIM and the MS test.

Theorem 4 (Mack–Skillings)**.** *For a RCB design, let $\hat{\boldsymbol{\alpha}}$ denote the estimator of $\boldsymbol{\alpha}$ in (4.10) or (4.11), given by (4.12), and let $\boldsymbol{\Sigma}_0$ denote its covariance matrix under the null hypothesis of equal distributions within blocks (4.14), then*

$$\left(\hat{\boldsymbol{\alpha}} - \frac{1}{2}\mathbf{1}\right)^T \boldsymbol{\Sigma}_0^- \left(\hat{\boldsymbol{\alpha}} - \frac{1}{2}\mathbf{1}\right) = MS_s.$$

Proof. Since $\boldsymbol{M}$ is idempotent, a generalized inverse is given by

$$\boldsymbol{\Sigma}_0^- = \frac{12LKn^2}{nK+1}\boldsymbol{M}.$$

Consequently

$$\left(\hat{\boldsymbol{\alpha}} - \frac{1}{2}\mathbf{1}\right)^T \boldsymbol{\Sigma}_0^- \left(\hat{\boldsymbol{\alpha}} - \frac{1}{2}\mathbf{1}\right) \;=\; A_1 - A_2,$$

where

$$A_1 = \frac{12LKn^2}{nK+1}\left(\hat{\boldsymbol{\alpha}} - \frac{1}{2}\mathbf{1}\right)^T \left(\hat{\boldsymbol{\alpha}} - \frac{1}{2}\mathbf{1}\right),$$

and

$$A_2 = \frac{12Ln^2}{nK+1}\left(\hat{\boldsymbol{\alpha}} - \frac{1}{2}\mathbf{1}\right)^T \mathbf{1}\mathbf{1}^T \left(\hat{\boldsymbol{\alpha}} - \frac{1}{2}\mathbf{1}\right).$$

Let Y_{kij} denote the sample observation of the j^{th} replicate of treatment k in block i. From Corollary 2 it follows that

$$\hat{\alpha}_l = \frac{1}{KLn^2}\sum_{k=1}^{K}\sum_{i=1}^{L}\sum_{j=1}^{n}\sum_{j'=1}^{n}\mathrm{I}\left(Y_{kij} \preccurlyeq Y_{lij'}\right).$$

Consequently $K^{-1}\sum_{l=1}^{K}\hat{\alpha}_l = 0.5$ and similar as in Theorem 3 one can show that $A_2 = 0$. Let $\bar{R}_{lij'}$ denote the ranking of sample observation $Y_{lij'}$, where the ranking is performed within blocks, then

$$\begin{aligned}
\hat{\alpha}_l &= \frac{1}{KLn^2}\sum_{i=1}^{L}\sum_{j'=1}^{n}\left(\sum_{k=1}^{K}\sum_{j=1}^{n}\mathrm{I}\left(Y_{kij} \le Y_{lij'}\right) - \frac{1}{2}\right) \\
&= \frac{1}{KLn^2}\sum_{i=1}^{L}\sum_{j'=1}^{n}\left(R_{lij'} - \frac{1}{2}\right) \\
&= \frac{1}{N}\left(\bar{R}_l - \frac{L}{2}\right),
\end{aligned}$$

where $\bar{R}_l = n^{-1} \sum_{i=1}^{L} \sum_{j'=1}^{n} R_{lij'}$. Hence

$$\begin{aligned} A_1 &= \frac{12LKn^2}{nK+1} \sum_{l=1}^{K} \left(\hat{\alpha}_l - \frac{1}{2} \right)^2 \\ &= \frac{12}{K(N+L)} \sum_{l=1}^{K} \left(\bar{R}_l - \frac{N+L}{2} \right)^2 . \end{aligned}$$

□

The Friedman test is also embedded in the marginal PIM, for it is a special case of the MS test with $n = 1$, i.e. each treatment-block combination occurs exactly once.

Corollary 3 (Friedman)**.** *If each treatment-block combination in a RCB design has one replicate and if $\hat{\boldsymbol{\alpha}}$ denotes the estimator of $\boldsymbol{\alpha}$ in (4.10) or (4.11), given by (4.12), and $\boldsymbol{\Sigma}_0$ its covariance matrix under the null hypothesis of equal distributions within blocks (4.14), then*

$$\left(\hat{\boldsymbol{\alpha}} - \frac{1}{2}\mathbf{1} \right)^T \boldsymbol{\Sigma}_0^- \left(\hat{\boldsymbol{\alpha}} - \frac{1}{2}\mathbf{1} \right) = \frac{12L}{K(K+1)} \sum_{l=1}^{K} \left(\bar{R}_l - \frac{K+1}{2} \right)^2 , \tag{4.15}$$

in which the right hand side of the equation is exactly the Friedman rank test statistic.

We refer to the Friedman statistic as F_s, where the subscript s denotes that this is a score-type of test. Similar as for the KW test, we can construct a Wald-type Friedman statistic by replacing $\boldsymbol{\Sigma}_0$ in (4.15) by the sandwich estimator $\hat{\boldsymbol{\Sigma}}$. We refer to this statistic as F_w. For completeness, the following lemma shows that the pseudo-observations associated with the marginal PIM (4.11) are sparsely correlated, so that the sandwich estimator $\hat{\boldsymbol{\Sigma}}$ is a consistent estimator of the true variance.

Lemma 7 (Sparse correlation: randomized complete blocks)**.** *The pseudo-observations associated with PIM (4.11) of a randomized complete block design possess the sparse correlation structure.*

Proof. Each pseudo-observation I_{ij} with $(i,j) \in \mathcal{I}_n = \{(i,j) \mid B_i = B_j \text{ and } i,j = 1,\ldots,N\}$ is only correlated with pseudo-observations of the same block. Each block has nK observations, thus similar as in Lemma 1 it follows that I_{ij} is correlated with $4nK - 7$ other pseudo-observations, so that $M_n = M_{nij} = 4nK - 6$. The largest set of pseudo-observations that are mutually independent consists of any I_{ij} and all other I_{kl} with i,j,k,l mutually distinct. The

size of this set is thus $\lfloor N/2 \rfloor$ (with $N = nKL$), i.e. the largest integer not larger than $N/2$. Suppose that N is even. Then

$$M_n m_n = (4nK - 6)nKL/2 = 2(nK)^2 L - 3nKL = O(n^2 K^2 L).$$

Since $O(|\mathcal{I}_n|) = O(n^2 K^2 L)$, the lemma holds for n even. Similarly, when N is odd, $M_n m_n = (4nK - 6)\lfloor nKL/2 \rfloor = O(n^2 K^2 L) = O(|\mathcal{I}_n|)$. □

4.4 The pairwise probabilistic index model

In the absence of blocks, the marginal PIM for the K-sample design is associated with rank tests which are based on the joint ranking. It refers to the comparison of the marginal outcome with an outcome in a particular treatment group, i.e. $\mathrm{P}(Y_\cdot \preccurlyeq Y_k)$. In this section we propose a PIM that models pairwise comparisons of treatment groups. In particular, for the K-sample design we propose the PIM (ANOVA notation)

$$\mathrm{P}(Y_k \preccurlyeq Y_l) = \gamma_{kl}. \tag{4.16}$$

The parameter γ_{kl} thus gives the probability that a random observation of group l exceeds a random observation of group k. The regression PIM follows from defining

$$\boldsymbol{Z}_{ij}^T = [\mathrm{I}(X_i = 1)\,\mathrm{I}(X_j = 2), \mathrm{I}(X_i = 1)\,\mathrm{I}(X_j = 3), \ldots, \mathrm{I}(X_i = K-1)\,\mathrm{I}(X_j = K)], \tag{4.17}$$

and $\boldsymbol{\gamma}$ the vector with the corresponding γ_{kl} and by considering the identity link. Then the pairwise PIM becomes (regression notation)

$$\mathrm{P}(Y_i \preccurlyeq Y_j \mid X_i, X_j) = \boldsymbol{Z}_{ij}^T \boldsymbol{\gamma}, \quad (X_i, X_j) \in \mathcal{X}_n \tag{4.18}$$

with $\mathcal{X}_n = \{(X_i, X_j) \mid X_i < X_j, i, j = 1 \ldots, N\}$, i.e. we restrict the PI to all unique treatment combinations.

The solution of the estimating equations (4.5) with $\boldsymbol{A}(\boldsymbol{Z}_{ij}, \boldsymbol{\beta}) = \boldsymbol{Z}_{ij}$ for PIM (4.18) follows immediately from Lemma 3. Corollary 4 gives the explicit formula for an individual parameter estimate.

Corollary 4. *The estimate of γ_{kl} in (4.16) or (4.18), defined as the solution of (4.5) with $\boldsymbol{A}(\boldsymbol{Z}_{ij};\boldsymbol{\beta}) = \boldsymbol{Z}_{ij}$, is of the form*

$$\hat{\gamma}_{kl} = \frac{1}{n_k n_l} \sum_{i=1}^{N} \sum_{j=1}^{N} c_{ik} c_{jl} \mathrm{I}\left(Y_i \preccurlyeq Y_j\right), \quad k < l, \tag{4.19}$$

where $c_{ik} = \mathrm{I}\left(X_i = k\right)$.

The following lemma gives the elements of the covariance matrix of $\hat{\boldsymbol{\gamma}}$, denoted as $\boldsymbol{\Sigma}_0$, under the null hypothesis of equal distributions. The proof follows directly from combining Lemma 4 and Corollary 4.

Lemma 8. *If $H_0 : F_1 = \cdots = F_K$ is true and if $\hat{\gamma}_{kl}$ denotes the estimator of γ_{kl} in PIM (4.16) or (4.18), given by (4.19), then*

- $\mathrm{Var}\left(\hat{\gamma}_{kl}\right) = (n_k + n_l + 1)(12 n_k n_l)^{-1}$, $k \neq l$,
- $\mathrm{Cov}\left(\hat{\gamma}_{kl}, \hat{\gamma}_{k'l}\right) = \mathrm{Cov}\left(\hat{\gamma}_{lk}, \hat{\gamma}_{lk'}\right) = (12 n_l)^{-1}$, $k \neq l, k' \neq l, k \neq k'$,
- $\mathrm{Cov}\left(\hat{\gamma}_{kl}, \hat{\gamma}_{lk'}\right) = \mathrm{Cov}\left(\hat{\gamma}_{lk}, \hat{\gamma}_{k'l}\right) = -(12 n_l)^{-1}$, $k \neq l, k' \neq l, k \neq k'$,
- $\mathrm{Cov}\left(\hat{\gamma}_{kl}, \hat{\gamma}_{k'l'}\right) = 0$, *if k, l, k', and l' are distinct.*

4.4.1 The two-sample design

In the following theorem we establish the relationship between the pairwise PIM and the WMW test. The proof follows immediately from Corollary 4 and Lemma 8.

Theorem 5 (Wilcoxon–Mann–Whitney)**.** *For the two-sample design let $\hat{\gamma}_{12}$ denote the estimator associated with PIM (4.16) or (4.18) and let σ_0^2 denote its variance under the null hypothesis of equal distributions, then*

$$\frac{\hat{\gamma}_{12} - 0.5}{\sigma_0} = \frac{\sum_{\{i|X_i=1\}} \sum_{\{j|X_j=2\}} \mathrm{I}\left(Y_i \preccurlyeq Y_j\right) - n_1 n_2 / 2}{\sqrt{[n_1 n_2 (n_1 + n_2 + 1)]/12}},$$

in which the right hand side of the equation is exactly the WMW statistic.

4.4.2 The three-sample design

In this section we establish the relationship between the pairwise PIM and the rank test of Brown and Hettmansperger (2002) for the three-sample design. The interpretation of their test is related to the the concept of transitivity based on the probabilistic index.

Definition 3 (PI-transitivity)**.** *Let Y_i be distributed according to F_i. A triplet (F_1, F_2, F_3) is called* PI-transitive *if* $\mathrm{P}(Y_a \preccurlyeq Y_b) \geq 0.5$ *and* $\mathrm{P}(Y_b \preccurlyeq Y_c) \geq 0.5$ *implies that* $\mathrm{P}(Y_a \preccurlyeq Y_c) \geq 0.5$*, for all* $(a, b, c) \in \{1, 2, 3\}$.

The Efron dice (Gardner, 1970; Brown and Hettmansperger, 2002) illustrate nicely that not all triplets are PI-transitive. For example, consider three dice with markings $\Omega_1 = \{2, 2, 6, 6, 7, 7\}$, $\Omega_2 = \{3, 3, 4, 4, 8, 8\}$, and $\Omega_3 = \{1, 1, 5, 5, 9, 9\}$. Let Y_i ($i = 1, 2, 3$) denote a random variable with a uniform distribution defined on Ω_i (i.e. each face of the die has the same probability $1/6$). Then $\mathrm{P}(Y_1 \preccurlyeq Y_2) = \mathrm{P}(Y_2 \preccurlyeq Y_3) = 5/9 > 0.5$, but surprisingly $\mathrm{P}(Y_1 \preccurlyeq Y_3) = 4/9 < 0.5$.

For real data examples for which the PI can be intransitive, we refer to Gillen and Emerson (2007) and Thangavelu and Brunner (2007) in the setting of multi-arm clinical trails and non-inferiority trials with active-controls and to Brown and Hettmansperger (2002) for a survey example.

The KW test has two degrees of freedom, while three pairwise comparisons can be considered. The following theorem illustrates that the KW test implicitly assumes PI-transitivity. Let $Y_\cdot$ denote the random variable with the marginal distribution of Y_1, Y_2, and Y_3. For notational convenience, let

$$P_j = \mathrm{P}(Y_\cdot \preccurlyeq Y_j) \quad \text{and} \quad P_{ij} = \mathrm{P}(Y_i \preccurlyeq Y_j). \tag{4.20}$$

Theorem 6. *Let F_i denote the distribution function associated with group $i = 1, 2, 3$. It holds that*

1. *if $P_1 = P_2 = P_3 = 0.5$ and $P_{12} = P_{13} = P_{23} = 0.5$, then the triplet (F_1, F_2, F_3) is PI-transitive,*
2. *and conversely, if $P_1 = P_2 = P_3 = 0.5$ and the triplet (F_1, F_2, F_3) is PI-transitive, then $P_{12} = P_{13} = P_{23} = 0.5$.*

Proof. If $P_{12} = P_{13} = P_{23} = 0.5$ and $P_{12} = P_{13} = P_{23} = 0.5$, then it follows that PI-transitivity is fulfilled by applying Definition 3.

If $P_1 = P_2 = P_3 = 0.5$, the system of equations $P_l = \sum_{k=1}^{3} P_{kl}/3$ simplifies to $P_{12} = P_{23}$ and $P_{13} = 1 - P_{23}$. This implies that if $P_{12} = P_{23} \geq 0.5$, then $P_{13} \leq 0.5$. If (F_1, F_2, F_3) is PI-transitive, then it follows that the system of equations has a unique solution given by $P_{12} = P_{13} = P_{23} = 0.5$. □

Since the KW test rejects in favour of the alternative $H_{1a} : P_i \neq 0.5$ for at least one $i = 1, 2, 3$, and since both $H_0 : P_1 = P_2 = P_3 = 0.5$ and $H_{1b} : P_{ij} \neq 0.5$ for some $i, j = 1, \ldots, 3$, can be true when there is PI-intransitivity, the KW-test can be insensitive to deviations from H_0 in the direction of H_{1b} when there is PI-intransitivity. Therefore, Brown and Hettmansperger (2002) proposed the statistic

$$\text{BH}_s := \text{KW}_s + \frac{3n_1n_2n_3}{N}\left(\frac{T_{12}}{n_1n_2} + \frac{T_{23}}{n_2n_3} + \frac{T_{31}}{n_3n_1}\right)^2, \tag{4.21}$$

where $T_{kl} = \sum_{\{i|X_i=k\}} \sum_{\{j|X_j=l\}} \text{sign}(Y_j - Y_i)$.

A large value of the second component of BH_s suggests PI-intransitivity. In Brown et al. (2006) they showed that BH_s has asymptotically a null distribution equal to $V_1^2 + \sqrt{3}/\pi V_2^2$, with V_1^2 distributed according to the chi-squared distribution with two degrees of freedom and with V_2 a standard logistic distributed variable. Let $\hat{\boldsymbol{\gamma}}^T = (\hat{\gamma}_{12}, \hat{\gamma}_{23}, \hat{\gamma}_{13})$ denote the estimators of the parameters in the PIM (4.18), and let $\boldsymbol{\Sigma}_0$ denote the covariance matrix of $\hat{\boldsymbol{\gamma}}$ under the null hypothesis $H_0 : F_1 = F_2 = F_3$, then the following theorem establishes the relationship between the pairwise PIM and the BH_s test.

Theorem 7 (Brown–Hettmansperger)**.** *Let $\hat{\boldsymbol{\gamma}}$ denote the estimator associated with PIM (4.18) and let $\boldsymbol{\Sigma}_0$ denote its covariance matrix under the null hypothesis of equal distributions, then*

$$\left(\hat{\boldsymbol{\gamma}} - \frac{1}{2}\mathbf{1}\right)^T \boldsymbol{\Sigma}_0^{-1} \left(\hat{\boldsymbol{\gamma}} - \frac{1}{2}\mathbf{1}\right) = \textit{BH}_s, \tag{4.22}$$

with BH_s given by (4.21).

Proof. Let $\boldsymbol{T}^T = (T_{12}, T_{23}, T_{31})$ and $\boldsymbol{V}_{\boldsymbol{T}} = \text{Cov}(\boldsymbol{T})$ under H_0, then Brown and Hettmansperger (2002) showed that an equivalent representation of BH_s is given by

$$\text{BH}_s = \boldsymbol{T}^T \boldsymbol{V}_{\boldsymbol{T}}^{-1} \boldsymbol{T}.$$

Vector $\boldsymbol{T}$ can be expressed as a function of $\hat{\boldsymbol{\gamma}}$

$$\boldsymbol{T} = 2\text{diag}(\boldsymbol{v})\hat{\boldsymbol{\gamma}} - \boldsymbol{v}^T,$$

where $\boldsymbol{v}^T = (n_1 n_2, n_2 n_3, -n_1 n_3)$ and consequently

$$\boldsymbol{V}_{\boldsymbol{T}} = 4\text{diag}(\boldsymbol{v})\boldsymbol{\Sigma}_0\text{diag}(\boldsymbol{v}).$$

Straightforward calculation shows that

$$\text{BH}_s = \left(\hat{\boldsymbol{\gamma}} - \frac{1}{2}\mathbf{1}\right)^T \boldsymbol{\Sigma}_0^{-1} \left(\hat{\boldsymbol{\gamma}} - \frac{1}{2}\mathbf{1}\right).$$

□

As the asymptotic null distribution of BH_s is not chi-squared, the null distribution of the quadratic form in the left hand side of (4.22) is not chi-squared either. This can be partially explained as follows. Let $\boldsymbol{\gamma}_0$ denote the true parameter associated with the pairwise PIM (4.18) and let $\lim_{N\to\infty} n_i/N = \lambda$, where $0 < \lambda < 1$, then $\sqrt{N}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)$ converges in distribution to a mean-zero multivariate normal distribution with covariance matrix $\boldsymbol{\Sigma}$. Under H_0 a consistent estimator of $\boldsymbol{\Sigma}$ is given by $N\boldsymbol{\Sigma}_0$, which has rank 3 for $N < \infty$, while its limit $\boldsymbol{\Sigma}_\infty := \lim_{N\to\infty} N\boldsymbol{\Sigma}_0$ has rank 2 and hence is singular. To illustrate this consider a balanced design $n := n_1 = n_2 = n_3$. The eigenvalues of $\boldsymbol{\Sigma}_0$ are given by $\lambda_1 = \lambda_2 = (3n+1)/(12n^2)$ and $\lambda_3 = 1/(12n^2)$. Only two eigenvalues of $\boldsymbol{\Sigma}_\infty$ are different from zero, and therefore $\boldsymbol{\Sigma}_\infty$ has rank two. The quadratic form $N\left(\hat{\boldsymbol{\gamma}} - 0.5\mathbf{1}\right)^T \boldsymbol{\Sigma}_\infty^- \left(\hat{\boldsymbol{\gamma}} - 0.5\mathbf{1}\right)$, has an asymptotic chi-squared null distribution with two degrees of freedom. However, since $\lim_{N\to\infty}(N\boldsymbol{\Sigma}_0)^{-1} \neq \boldsymbol{\Sigma}_\infty^-$, this is not the case for the left hand side of (4.22). Moreover, the elements of $\sqrt{N}\hat{\boldsymbol{\gamma}}$ are linearly dependent, since one can show that, under H_0,

$$\sqrt{n}\left[(\hat{\gamma}_{12} - 0.5) + (\hat{\gamma}_{23} - 0.5) - (\hat{\gamma}_{13} - 0.5)\right] \xrightarrow{p} 0.$$

We refer to Fligner (1985) and Koziol and Reid (1977) for the details.

We can force the pairwise PIM (4.18) to imply PI-transitivity by imposing the restrictions $\gamma_{kl} = \gamma'_k - \gamma'_l$ for some new parameters γ'_k. The model then simplifies to

$$\text{P}\left(Y_k \preccurlyeq Y_l\right) = \gamma'_k - \gamma'_l. \tag{4.23}$$

It is straightforward to see that this parametrization implies PI-transitivity. Furthermore, PIM (4.23) corresponds to the Bradley–Terry model; see for example Thas et al. (2012c, p. 667) and Bergsma et al. (2009, 2012).

4.4.3 Ordered and umbrella alternatives

Thus far, all tests focused on rejecting the null hypothesis $H_0 : F_1 = \ldots = F_K$ in favour of an alternative that states that some particular PIs are not equal to 0.5. However, sometimes more informative alternatives can be of interest. When the K treatments can be ordered (e.g. the dosage of a drug), one can formulate an alternative for which the outcome tends to increase or decrease with increasing treatment. Using notation (4.20), Mann (1945) defined an upward trend as

$$H_1^o : \frac{2}{K(K-1)} \sum_{k=1}^{K-1} \sum_{l=k+1}^{K} P_{lk} > \frac{1}{2}.$$

Under the location shift model

$$F_1(y - \tau_1) = \ldots = F_K(y - \tau_K), \tag{4.24}$$

H_1^o simplifies to the ordered alternative $\tau_1 \leq \ldots \leq \tau_K$, with at least one strict inequality. The Jonckheere–Terpstra (JT) test (Jonckheere, 1954; Terpstra, 1952) is consistent against H_1^o and its test statistic is given by

$$\text{JT}_s := \sigma_{JT}^{-1} \left(\sum_{k=1}^{K-1} \sum_{l=k+1}^{K} \sum_{\{i|X_i=k\}} \sum_{\{j|X_j=l\}} \text{I}\left(Y_i \preccurlyeq Y_j\right) - \mu_{JT} \right), \tag{4.25}$$

where $\mu_{JT} = (N^2 - \sum_{j=1}^{K} n_j^2)/4$ and $\sigma_{JT}^2 = [N^2(2N+3) - \sum_{j=1}^{K} n_j^2(2n_j+3)]/72$.

The JT test can also be obtained from the PIM (regression notation)

$$\text{P}\left(Y_i \preccurlyeq Y_j \mid X_i, X_j\right) = \frac{1}{2} + \delta Z_{ij}, \quad (X_i, X_j) \in \mathcal{X}_0, \tag{4.26}$$

where $Z_{ij} = \text{I}\left(X_i < X_j\right) - \text{I}\left(X_i > X_j\right)$. The interpretation of δ comes from $\delta = \text{P}(Y_i \preccurlyeq Y_j \mid X_i < X_j) - 0.5$, i.e. the probability that an outcome of a higher factor level exceeds an outcome of a lower factor level, reduced with 0.5. Equivalently, $\delta = 0.5 - \text{P}\left(Y_i \preccurlyeq Y_j \mid X_j < X_i\right)$. Note that the offset 0.5 is a consequence of the definition of Z_{ij} and the identity link. Let $\hat{\delta}$ denote the OLS estimator associated with PIM (4.26), then its variance under the null hypothesis of equal distributions, say σ_0^2, can be obtained from combining the OLS expression and Lemma 4. Indeed, if $\boldsymbol{\Sigma}_p$ denotes the matrix with elements given by Lemma 4, then $\sigma_0^2 = (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{\Sigma}_p\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}$. The hypothesis of interest can be expressed as $H_0^o : \delta = 0$. We prefer to use the regression notation here, because the factor acts as an integer-valued regressor. The following theorem establishes the relationship between PIM (4.26) and the JT test.

Theorem 8 (Jonckheere–Terpstra)**.** *Let $\hat{\delta}$ denote the OLS estimator associated with PIM (4.26) and let σ_0^2 denote its variance under the null hypothesis of equal distributions, then*

$$\frac{\hat{\delta}}{\sigma_0} = JT_s,$$

with JT_s given by (4.25).

Proof. The estimating equations (4.5) with $\boldsymbol{A}(\boldsymbol{Z}_{ij};\boldsymbol{\beta}) = \boldsymbol{Z}_{ij}$ for PIM (4.26) simplify to

$$\begin{aligned} \sum_{k=1}^{K-1}\sum_{l=k+1}^{K}\sum_{\{i|X_i=k\}}\sum_{\{j|X_j=l\}}\left[\mathrm{I}\left(Y_i \preccurlyeq Y_j\right) - \left(\frac{1}{2}+\delta\right)\right] &= 0 \\ \Leftrightarrow \hat{\delta} = \frac{1}{\tilde{N}}\left(\sum_{k=1}^{K-1}\sum_{l=k+1}^{K}\sum_{\{i|X_i=k\}}\sum_{\{j|X_j=l\}}\mathrm{I}\left(Y_i \preccurlyeq Y_j\right)\right) - \frac{1}{2}, \end{aligned}$$

where $\tilde{N} = \sum_{k=1}^{K-1}\sum_{l=k+1}^{K} n_k n_l$. The remainder of the proof follows from Hollander and Wolfe (1999, p. 209). □

The JT test is also related to the pairwise PIM; see Theorem 9. Consequently, after fitting a pairwise PIM, it is not necessary to fit the PIM (4.26) to obtain the JT test. The proof is similar to the proof of Theorem 8.

Theorem 9 (Jonckheere–Terpstra 2)**.** *Let $\hat{\boldsymbol{\gamma}}$ denote the estimator associated with the pairwise PIM (4.18) and let $\boldsymbol{\Sigma}_0$ denote its covariance matrix under the null hypothesis of equal distributions, then*

$$\frac{\boldsymbol{v}^T(\hat{\boldsymbol{\gamma}} - 1/2)}{\sqrt{\boldsymbol{v}^T\boldsymbol{\Sigma}_0\boldsymbol{v}}} = JT_s,$$

with JT_s given by (4.25) and $\boldsymbol{v}^T = (n_1 n_2, \ldots, n_{K-1}n_K)$.

Instead of an ordered alternative, an umbrella alternative can formulated. The outcome then increases (decreases) with increasing treatment up to a given factor level, say $X = P$, and then decrease (increases) with increasing factor level.

Under the location shift model (4.24) this becomes

$$\tau_1 \leq \tau_2 \leq \ldots, \leq \tau_P \geq \tau_{P+1} \geq \ldots \geq \tau_K,$$

with at least one strict inequality. In terms of the PI this can be formulated as

$$H_1^u : \frac{2}{P(P-1) + (K-P)(K-P+1)}\left(\sum_{k<l,\, l\leq P} P_{kl} + \sum_{P\leq l,\, k>l} P_{kl}\right) > \frac{1}{2}.$$

The Mack–Wolfe (MW) test (Mack and Wolfe, 1981) is consistent against H_1^u and is based on the statistic

$$\begin{aligned} \text{MW}_s \;:=\; \sigma_{MW}^{-1} \Bigg(& \sum_{k=1}^{P-1} \sum_{l=k+1}^{P} \sum_{\{i|X_i=k\}} \sum_{\{j|X_j=l\}} \mathrm{I}\left(Y_i \preccurlyeq Y_j\right) + \\ & \sum_{k=P}^{K-1} \sum_{l=k+1}^{K} \sum_{\{i|X_i=k\}} \sum_{\{j|X_j=l\}} \mathrm{I}\left(Y_i \succcurlyeq Y_j\right) - \mu_{MW} \Bigg), \end{aligned} \tag{4.27}$$

where $\mathrm{I}\left(Y_i \succcurlyeq Y_j\right) = 1 - \mathrm{I}\left(Y_i \preccurlyeq Y_j\right)$,

$$\mu_{MW} = \frac{1}{4}\left(N_1^2 + N_2^2 - \sum_{i=1}^{K} n_i^2 - n_P^2\right),$$

and

$$\begin{aligned} \sigma_{MW}^2 \;=\; & \frac{1}{72}\Bigg(2(N_1^3 + N_2^3) + 3(N_1^2 + N_2^2) - \sum_{i=1}^{K} n_i^2(2n_i + 3) - n_P^2(2n_P + 3) + \\ & 12 n_P N_1 N_2 - 12 n_P^2 N\Big), \end{aligned}$$

with $N_1 = \sum_{i=1}^{P} n_i$ and $N_2 = \sum_{i=P}^{K} n_i$. The MW test can also be obtained from the PIM framework. Let

$$Z_{ij} = \mathrm{I}\left(X_i < X_j \leq P\right) - \mathrm{I}\left(X_j < X_i \leq P\right) + \mathrm{I}\left(X_i > X_j \geq P\right) - \mathrm{I}\left(X_j > X_i \geq P\right),$$

and consider the PIM (regression notation)

$$\mathrm{P}\left(Y_i \preccurlyeq Y_j \mid X_i, X_j\right) = \frac{1}{2} + \zeta Z_{ij}, \quad (X_i, X_j) \in \mathcal{X}_0. \tag{4.28}$$

The interpretation follows from $\zeta = \mathrm{P}\left(Y_i \preccurlyeq Y_j \mid X_i < X_j \leq P\right) - 0.5$, i.e. the probability that an outcome of a higher factor level of at most P exceeds an outcome of a lower factor level reduced with 0.5. Similarly $\zeta = \mathrm{P}\left(Y_i \preccurlyeq Y_j \mid X_i > X_j \geq P\right) - 0.5$, i.e. the probability that an outcome of a lower factor level of minimal P exceeds an outcome of a higher factor level reduced with 0.5. The relationship between PIM (4.28) and the MW test is established in the following theorem.

Theorem 10 (Mack–Wolfe)**.** *Let $\hat{\zeta}$ denote the OLS estimator associated with PIM (4.28) and let σ_0^2 denote its variance under the null hypothesis of equal distributions, then*

$$\frac{\hat{\zeta}}{\sigma_0} = \mathit{MW}_s,$$

with MW_s given by (4.27).

Proof. The estimating equations (4.5) with $\boldsymbol{A}(\boldsymbol{Z}_{ij};\boldsymbol{\beta}) = \boldsymbol{Z}_{ij}$ for PIM (4.28) simplify to

$$\begin{aligned}
&\sum_{k=1}^{P-1}\sum_{l=k+1}^{P}\sum_{\{i|X_i=k\}}\sum_{\{j|X_j=l\}}\left[\mathrm{I}\left(Y_i \preccurlyeq Y_j\right) - \left(\frac{1}{2}+\zeta\right)\right] \\
+ &\sum_{l=P}^{K-1}\sum_{k=l+1}^{K}\sum_{\{i|X_i=k\}}\sum_{\{j|X_j=l\}}\left[\mathrm{I}\left(Y_i \preccurlyeq Y_j\right) - \left(\frac{1}{2}+\zeta\right)\right] = 0 \\
\Leftrightarrow \quad &\hat{\zeta} = \frac{1}{\tilde{N}_1+\tilde{N}_2}\left(\sum_{k=1}^{P-1}\sum_{l=k+1}^{P}\sum_{\{i|X_i=k\}}\sum_{\{j|X_j=l\}}\mathrm{I}\left(Y_i \preccurlyeq Y_j\right)\right. \\
+ &\left.\sum_{l=P}^{K-1}\sum_{k=l+1}^{K}\sum_{\{i|X_i=k\}}\sum_{\{j|X_j=l\}}\mathrm{I}\left(Y_i \preccurlyeq Y_j\right)\right) - \frac{1}{2}
\end{aligned}$$

where $\tilde{N}_1 = \sum_{k=1}^{P-1}\sum_{l=k+1}^{P} n_k n_l$ and $\tilde{N}_2 = \sum_{l=P}^{K-1}\sum_{k=l+1}^{K} n_k n_l$. The remainder of the proof follows from Hollander and Wolfe (1999, p. 221). □

The following theorem shows how the MW test can be obtained from the pairwise PIM. The proof is similar to the proof of Theorem 10.

Theorem 11 (Mack–Wolfe 2)**.** *Let*

$$\hat{\boldsymbol{\gamma}}^T = (\hat{\gamma}_{12},\ldots,\hat{\gamma}_{1P},\hat{\gamma}_{2P},\ldots,\hat{\gamma}_{(P-1)P},\hat{\gamma}_{1(P+1)},\ldots,\hat{\gamma}_{P(P+1)},\hat{\gamma}_{P(P+2)},\ldots,\hat{\gamma}_{(K-1)K}),$$

denote the estimator associated with (4.18) and let $\boldsymbol{\Sigma}_0$ *denote its covariance matrix under the null hypothesis of equal distributions. Let*

$$\boldsymbol{v}^T = (n_1n_2,\ldots,n_1n_P,n_2n_P,\ldots,n_{P-1}n_P,0,\ldots,0,-n_Pn_{P+1},-n_Pn_{P+2},\ldots,-n_{K-1}n_K),$$

then

$$\frac{\boldsymbol{v}^T(\hat{\boldsymbol{\gamma}}-1/2)}{\sqrt{\boldsymbol{v}^T\boldsymbol{\Sigma}_0\boldsymbol{v}}} = MW_s,$$

with MW_s *given by (4.27).*

4.4.4 Extension to block designs

The pairwise PIM can be extended to block designs. In ANOVA notation the model for the one-way layout becomes

$$\mathrm{P}\left(Y_{kj} \preccurlyeq Y_{lj}\right) = \gamma_{kl}, \tag{4.29}$$

where $i = 1, \ldots, K$ refers to the treatment group and $j = 1, \ldots, L$ to the block. The parameter γ_{kl} thus gives the probability that a random observation of group l exceeds a random observation of group k *within the same block*. The corresponding regression PIM is obtained with $\boldsymbol{Z}_{ij}$ as in (4.17) and $\boldsymbol{\gamma}$ the vector with the corresponding γ_{kl} as before. Model (4.29) now becomes

$$\mathrm{P}\left(Y_i \preccurlyeq Y_j \mid X_i, B_i, X_j, B_j\right) = \boldsymbol{Z}_{ij}^T \boldsymbol{\gamma}, \quad (\boldsymbol{X}_i, \boldsymbol{X}_j) \in \mathcal{X}_n, \tag{4.30}$$

with $\mathcal{X}_n = \{(\boldsymbol{X}_i, \boldsymbol{X}_j) \mid X_i < X_j, B_i = B_j, i, j = 1, \ldots, N\}$, i.e. the PI is restricted to comparisons within blocks and to all unique treatment combinations.

Similar as for the marginal PIM, the blocking does not result in extra parameters in the model, but if affects the estimating equations through a limitation on the pseudo-observations. Lemma 3 remains valid and Corollary 5 gives the explicit form of the estimator in the presence of blocks.

Corollary 5. *The estimate of γ_{kl} in (4.29) or (4.30), defined as the solution of (4.5) with $\boldsymbol{A}(\boldsymbol{Z}_{ij}; \boldsymbol{\beta}) = \boldsymbol{Z}_{ij}$, is of the form*

$$\hat{\gamma}_{kl} = d_{kl} \sum_{i=1}^{N} \sum_{j=1}^{N} b_{ij} c_{ik} c_{jl} \mathrm{I}\left(Y_i \preccurlyeq Y_j\right), \quad k < l, \tag{4.31}$$

where

- $b_{ij} = 1$ *if* $B_i = B_j$ *and* $b_{ij} = 0$ *otherwise,*
- $c_{ik} = \mathrm{I}\left(X_i = k\right)$*,*
- $d_{kl} = \left(\sum_{i=1}^{N} \sum_{j=1}^{N} b_{ij} c_{ik} c_{jl}\right)^{-1}$.

4.5 Correcting for continuous covariates

In the introduction of this chapter we argued that the PIM framework not only includes the classical rank tests for factorial design, but it also gives the flexibility to construct tests for more complicated designs. In this section we demonstrate briefly how a PIM may be constructed that allows for testing for a factor effect while controlling for a continuous covariate.

Consider the K-sample design, and let x denote the continuous covariate. Let Y_{kx} denote the outcome variable in group k, conditional on covariate x. The marginal PIM (4.2) may now be

extended to become (ANOVA notation)

$$\mathrm{P}\left(Y_{\cdot\cdot} \preccurlyeq Y_{kx}\right) = \alpha_k + \delta x, \tag{4.32}$$

which still belongs to the class of PIMs. The interpretation of α_k is the same as for model (4.2), but now conditional on $x = 0$ (if $x = 0$ is not within the scope of the model, the covariate may be centred first). The parameter δ has also an informative interpretation: it measures the increase of the PI $\mathrm{P}\left(Y_{\cdot\cdot} \preccurlyeq Y_{kx}\right)$ when x is increased with one unit within the same treatment group. An interaction effect between the factor and the covariate variables may be modelled by adding a term, say $\zeta_k x$ to (4.32); a restriction on the ζ_k is required to make the parameters identifiable; e.g. $\zeta_1 = 0$ or $\sum_k \zeta_k = 0$.

A potential drawback of (4.32) is that it does not result in a PI in $[0, 1]$ for all x. Therefore, it may be more appropriate the choose a logit or probit link function. For example, with a logit link model (4.32) becomes

$$\mathrm{logit}[\mathrm{P}\left(Y_{\cdot\cdot} \preccurlyeq Y_{kx}\right)] = \alpha_k + \delta x,$$

and thus $\mathrm{expit}(\alpha_k)$ has now the interpretation of α_k in (4.32), and δ is the log odds ratio of the PI for an increase of x with one unit within the same group.

The pairwise PIM (4.16) may also be extended to include the effect of x. For example, upon using the identity link,

$$\mathrm{P}\left(Y_{kx_1} \preccurlyeq Y_{lx_2}\right) = \gamma_{kl} + \eta(x_2 - x_1). \tag{4.33}$$

Thus $\gamma_{kl} = \mathrm{P}\left(Y_{kx} \preccurlyeq Y_{lx}\right)$, i.e. the probability that a random outcome of group l exceeds a random outcome of group k when both observations have the same continuous covariate $x_1 = x_2 = x$. As for the marginal model, we recommend using a logit or probit link. An example is given in Section 4.9.

4.6 The two-way layout

Consider the two-way layout where X_1 (X_2) corresponds to the first (second) factor with K_1 (K_2) levels. For the remainder of this section we consider no blocks, but all results can be generalized to block-designs by limiting the summations in (4.5) to pseudo-observations defined within the same block. All PIMs are defined for the no-order restriction.

When using the ANOVA notation, Y_{kl} denotes an outcome associated with groups $X_1 = k$ and $X_2 = l$. We use the notation $Y_{k.}$ to denote the outcome of the distribution marginalized over X_2. Similar for $Y_{.l}$ (marginalized over X_1) and $Y_{..}$ (marginalized over both X_1 and X_2). Consider the marginal PIM in ANOVA notation

$$\mathrm{P}\left(Y_{..} \preccurlyeq Y_{kl}\right) = \mu + \alpha_k + \beta_l + \gamma_{kl}. \tag{4.34}$$

Since this model is over-parametrized, restrictions are required. For example, $\alpha_1 = 0$, $\beta_1 = 0$, $\gamma_{k1} = 0$, $k = 1, \ldots, K_1$, and $\gamma_{1l} = 0$, $l = 1, \ldots, K_2$. The interpretation follows from $\alpha_k + \gamma_{kl} = \mathrm{P}\left(Y_{..} \preccurlyeq Y_{kl}\right) - \mathrm{P}\left(Y_{..} \preccurlyeq Y_{1l}\right)$, i.e. $\alpha_k + \gamma_{kl}$ gives the difference in the marginal PI of group k relative to the marginal PI of group 1 of factor X_1, while keeping X_2 fixed at group l. Since the effect depends on the level l through γ_{kl}, the latter quantifies an interaction effect.

If the sum restriction is considered, i.e. $\sum_k \alpha_k = 0$, $\sum_l \beta_l = 0$, $\sum_l \gamma_{kl} = 0$, $k = 1, \ldots, K_1$, and $\sum_k \gamma_{kl} = 0$, $l = 1, \ldots, K_2$, then, for a balanced design, the interpretation follows from $\alpha_k + \gamma_{kl} = \mathrm{P}\left(Y_{..} \preccurlyeq Y_{kl}\right) - \mathrm{P}\left(Y_{..} \preccurlyeq Y_{.l}\right)$, i.e. $\alpha_k + \gamma_{kl}$ gives the difference in the marginal PI of group k relative to the marginal PI marginalized over factor X_1, while keeping X_2 fixed at group l.

Model (4.34) may also be written as a regression PIM model by appropriate coding of dummies in vectors $\boldsymbol{Z}_{ij}$ and a corresponding stacking of the model parameters in a vector, say $\boldsymbol{\alpha}$. The model then becomes

$$\mathrm{P}\left(Y_i \preccurlyeq Y_j \mid X_{1j}, X_{2j}\right) \;=\; \boldsymbol{Z}_{ij}^T \boldsymbol{\alpha}. \tag{4.35}$$

Lemma 3 is again valid. It states that the PIM estimation theory provides the OLS estimator of $\boldsymbol{\alpha}$. Consequently, the linear PIM (4.35) can be viewed as a linear regression model fitted to the pseudo-observations instead of the original outcome observations. The variance estimator of the linear model is, however, not consistent, because the pseudo-observations are not mutually independent. As before, the general PIM theory provides a consistent sandwich estimator of the covariance matrix, but Lemma 4 may be used instead for obtaining the exact covariance matrix under the null hypothesis that neither X_1 or X_2 affects the outcome distribution.

We can also extend the pairwise PIM to the two-way layout. With the ANOVA-notation the no-interaction model becomes

$$\mathrm{P}\left(Y_{ij} \preccurlyeq Y_{kl}\right) = \frac{1}{2} + \alpha_{ik} + \beta_{jl}. \tag{4.36}$$

Since $\mathrm{P}(Y_{ij} \preccurlyeq Y_{ij}) = 0.5$, it follows that $\alpha_{ii} = 0$ and $\beta_{jj} = 0$. The interpretation follows from $\alpha_{ik} = \mathrm{P}(Y_{ij} \preccurlyeq Y_{kj}) - 0.5$, i.e. α_{ik} is the probability that a random observation of group k of factor 1 exceeds a random observation of group i of factor 1, when factor 2 is fixed at group j, reduced with 0.5.

If we impose the restrictions $\alpha_{ik} = \alpha'_i - \alpha'_k$ and $\beta_{jl} = \beta'_j - \beta'_l$, for some new parameters α'_i and β'_j, PIM (4.36) simplifies to $\mathrm{P}(Y_{ij} \preccurlyeq Y_{kl}) = 0.5 + \alpha'_i - \alpha'_k + \beta'_j - \beta'_l$, which can be considered as an extension of the Bradley–Terry model (4.23) to the two-way layout for which PI-transitivity holds.

PIM (4.36) can be extended to include interaction. We suggest

$$\mathrm{P}(Y_{ij} \preccurlyeq Y_{kl}) = \frac{1}{2} + \alpha_{ik} + \beta_{jl} + \gamma_{ikjl}. \tag{4.37}$$

Since this model is over-parametrized, additional restrictions on the model parameters or a reparametrization need to be imposed. For example, $\gamma_{ikjl} = \delta'_i \mathrm{I}(i = k) + \zeta'_j \mathrm{I}(j = l)$. Then $\alpha_{ik} + \zeta'_j = \mathrm{P}(Y_{ij} \preccurlyeq Y_{kj}) - 0.5$, i.e. $\alpha_{ik} + \zeta'_j$ is the probability that a random observation of group k of factor 1 exceeds a random observation of group i of factor 1, within group j of factor 2, reduced with 0.5.

Patel and Hoel (1973) define a measure of interaction based on the PI as follows

$$\mu_{ijkl} := \mathrm{P}(Y_{ij} \preccurlyeq Y_{il}) - \mathrm{P}(Y_{kj} \preccurlyeq Y_{kl}) \quad \text{and} \quad \mu'_{ijkl} := \mathrm{P}(Y_{ij} \preccurlyeq Y_{kj}) - \mathrm{P}(Y_{il} \preccurlyeq Y_{kl}).$$

No-interaction then corresponds to $\mu_{ijkl} = \mu'_{ijkl} = 0$. It is straightforward to see that for PIM (4.36) $\mu_{ijkl} = \mu'_{ijkl} = 0$, while for PIM (4.37) $\mu_{ijkl} = \delta'_i - \delta'_k$ and $\mu'_{ijkl} = \zeta_j - \zeta_l$, which are not necessarily equal to 0. For other definitions of interaction in a nonparametric setting, we refer to de Kroon and van der Laan (1981) and Marden and Muyot (1995).

By appropriate coding of dummies in vectors $\boldsymbol{Z}_{ij}$ and stacking the model parameters in a parameter vector, the general PIM estimation theory of Theorems 1 and 2 may once more be invoked to give OLS estimators and consistent covariance matrix estimators.

4.7 Relationship with methods of Akritas and colleagues

Akritas et al. (2000) proposed a model that forms their basis for testing for no treatment effect in the presence of a continuous covariate. In the absence of the covariate the model reduces to

the models of Akritas and Arnold (1994) and Akritas et al. (1997). Tsangari and Akritas (2004) further extend the model to more than one covariate. In the following sections we relate these methods to the marginal and pairwise PIM.

4.7.1 The one-way layout

Let F_i denote the distribution function of the outcome variable in group $i = 1, \ldots, K$ and let Y_i denote the corresponding random variable. We use $Y_\cdot$ to denote the outcome variable with the marginal distribution function. Akritas and Arnold (1994) considered the decomposition

$$F_i(y) = M(y) + A_i(y), \tag{4.38}$$

with $\sum_{i=1}^{K} A_i(y) = 0$ for all y. This restrictions implies that for equally large groups, $M(y)$ is the marginal outcome distribution. Since (4.38) specifies a conditional outcome distribution function, we can immediately obtain the probabilistic index, both for a marginal and a pairwise model.

The marginal PIM gives

$$\mathrm{P}\left(Y_\cdot \preccurlyeq Y_i\right) = \int M(y) dF_i(y) = \frac{1}{2} + \alpha_i,$$

with $\alpha_i := \int M(y) dA_i(y)$ satisfying the restriction $\sum_{i=1}^{K} \alpha_i = 0$. This model is equivalent to the PIM (4.2) after a reparameterization.

To establish the relationship with the pairwise PIM, the following lemma will be useful.

Lemma 9. *For $M(\cdot)$ and $A_i(\cdot)$ as in (4.38), it holds that*

$$\int M(y) dA_i(y) = -\int A_i(y) dM(y). \tag{4.39}$$

Proof. For notational convenience we consider a continuous outcome with support the real line. Since $F_i(\cdot)$ and $M(\cdot)$ are both distribution functions, it follows that

$$\lim_{y \to -\infty} F_i(y) = \lim_{y \to -\infty} M(y) = 0,$$

and

$$\lim_{y \to \infty} F_i(y) = \lim_{y \to \infty} M(y) = 1.$$

Consequently, from (4.38) it follows that

$$\lim_{y\to-\infty} A_i(y) = \lim_{y\to\infty} A_i(y) = 0. \tag{4.40}$$

Upon using (4.40), it holds that

$$\int_{-\infty}^{\infty} d\left[M(y)A_i(y)\right] = 0. \tag{4.41}$$

By using the product rule, the left hand side of (4.41) is also equal to

$$\int_{-\infty}^{\infty} d\left[M(y)A_i(y)\right] = \int_{-\infty}^{\infty} \left[M(y)dA_i(y) + A_i(y)dM(y)\right]. \tag{4.42}$$

Combining (4.41) and (4.42) now completes the proof.

□

The pairwise PIM becomes

$$\mathrm{P}\left(Y_i \preccurlyeq Y_j\right) = \frac{1}{2} + \alpha_j - \alpha_i + (\alpha\alpha)_{ij}, \tag{4.43}$$

where we have used the identity (4.39), as well as the notation $(\alpha\alpha)_{ij} := \int A_i(y)dA_j(y)$. When $(\alpha\alpha)_{ij} = 0$ for all i, j, model (4.43) is the Bradley–Terry-type PIM (4.23).

If $(\alpha\alpha)_{ij} = 0$ then the marginal PIM is as informative as the pairwise PIM. Indeed, the pairwise PIM can be constructed from marginal PIM as follows

$$\mathrm{P}\left(Y_i \preccurlyeq Y_j\right) = \frac{1}{2} + \mathrm{P}\left(Y_\cdot \preccurlyeq Y_j\right) - \mathrm{P}\left(Y_\cdot \preccurlyeq Y_i\right).$$

The interpretation of $(\alpha\alpha)_{ij}$ follows from

$$(\alpha\alpha)_{ij} = \mathrm{P}\left(Y_i \preccurlyeq Y_j\right) - \mathrm{P}\left(Y_\cdot \preccurlyeq Y_j\right) + \mathrm{P}\left(Y_\cdot \preccurlyeq Y_i\right) - \frac{1}{2}.$$

4.7.2 The two-way layout

For the two-way layout, Akritas and Arnold (1994) assume a decomposition of F_{ij},

$$F_{ij}(y) = M(y) + A_i(y) + B_j(y) + C_{ij}(y), \tag{4.44}$$

with restrictions $\sum_i A_i(y) = 0$, $\sum_j B_j(y) = 0$, $\sum_i C_{ij}(y) = 0$, and $\sum_j C_{ij}(y) = 0$ for all y. The marginal PIM (4.34) follows from

$$\begin{aligned} \mathrm{P}\left(Y_{\cdot\cdot} \preccurlyeq Y_{ij}\right) &= \int M(y) dF_{ij}(y) \\ &= \int M(y) d\left[M(y) + A_i(y) + B_j(y) + C_{ij}(y)\right] \\ &= \frac{1}{2} + \alpha_i + \beta_j + \gamma_{ij}, \end{aligned} \tag{4.45}$$

where $\alpha_i := \int M(y) dA_i(y)$, $\beta_i := \int M(y) dB_j(y)$, and $\gamma_{ij} := \int M(y) dC_{ij}(y)$. The restrictions imposed by Akritas and Arnold (1994) imply that $\sum_i \alpha_i = 0$, $\sum_j \beta_j = 0$, $\sum_i \gamma_{ij} = 0$, and $\sum_j \gamma_{ij} = 0$.

The interpretation of the model parameters of (4.45) can be read from

$$\begin{aligned} \mathrm{P}\left(Y_{\cdot\cdot} \preccurlyeq Y_{i\cdot}\right) &= \int M(y) dF_{i\cdot}(y) \\ &= \int M(y) d[M(y) + A_i(y)] \\ &= \frac{1}{2} + \alpha_i. \end{aligned}$$

Hence $\alpha_i = \mathrm{P}\left(Y_{\cdot\cdot} \preccurlyeq Y_{i\cdot}\right) - 0.5$. Similarly, for the interpretations of β_j and γ_{ij}, where

$$\beta_j = \mathrm{P}\left(Y_{\cdot\cdot} \preccurlyeq Y_{\cdot j}\right) - 0.5,$$

and

$$\gamma_{ij} = \mathrm{P}\left(Y_{\cdot\cdot} \preccurlyeq Y_{ij}\right) - \mathrm{P}\left(Y_{\cdot\cdot} \preccurlyeq Y_{i\cdot}\right) - \mathrm{P}\left(Y_{\cdot\cdot} \preccurlyeq Y_{\cdot j}\right) + \frac{1}{2}.$$

The pairwise PIM becomes

$$\begin{aligned} \mathrm{P}\left(Y_{ij} \preccurlyeq Y_{kl}\right) &= \int F_{ij}(y) dF_{kl}(y) \\ &= \int \left[M(y) + A_i(y) + B_j(y) + C_{ij}(y)\right] d\left[M(y) + A_k(y) + B_l(y) + C_{kl}(y)\right] \\ &= \frac{1}{2} + (\alpha_k - \alpha_i) + (\beta_l - \beta_j) + (\gamma_{kl} - \gamma_{ij}) + \\ &\quad (\alpha\alpha)_{ik} + (\beta\beta)_{jl} + (\gamma\gamma)_{ijkl} + \\ &\quad [(\alpha\beta)_{il} - (\alpha\beta)_{kj}] + [(\alpha\gamma)_{ikl} - (\alpha\gamma)_{kij}] + [(\beta\gamma)_{jkl} - (\beta\gamma)_{lij}], \end{aligned} \tag{4.46}$$

where the Greek letters α, β, and γ are used to denote the parameters originating from the Roman letters A, B, and C in (4.44), and in which we repeatedly used the property $\int U(y) dV(y) + \int V(y) dU(y) = 0$, for U and V any of terms in (4.44). Model (4.46) is a special case of the over-parametrized model (4.37). By setting some of the higher-order parameters to zero, better interpretable PIMs may be obtained; see Section 4.6.

4.7.3 The one-way layout with a continuous covariate

Let F_{kx} denote the distribution function the outcome variable in group $k = 1, \ldots, K$, conditional on the single covariate value $x \in \mathbb{R}$, and let $G(x)$ denote the distribution function or a weight function (per design) of the covariate. Without loss of generality we assume that the outcome is absolutely continuous; see Akritas et al. (2000) for details on how a minor change in the definition of F_{kx} makes their methods applicable for discrete outcomes too. The model of Akritas et al. (2000) assumes a decomposition of F_{kx},

$$F_{kx}(y) = M(y) + A_k(y) + D_x(y) + C_{kx}(y), \tag{4.47}$$

that satisfies the following restrictions: $\sum_{k=1}^{K} A_k(y) = 0$ for all y, $\int D_x(y)dG(x) = 0$ for all y, $\sum_{k=1}^{K} C_{kx}(y) = 0$ for all x and y, and $\int C_{kx}(y)dG(x) = 0$ for all k and y. Let $\boldsymbol{X}_i^T = (X_{1i}, X_{i2})$ with $X_{1i} = 1, \ldots, K$ indicating the group, and X_{2i} the continuous covariate x. Then, a marginal PIM follows from

$$\begin{aligned} \mathrm{P}\left(Y_{\cdot\cdot} \preccurlyeq Y_{kx}\right) &= \int M(y)dF_{kx}(y) \\ &= \int M(y)d\left[M(y) + A_k(y) + D_x(y) + C_{kx}(y)\right] \\ &= \frac{1}{2} + \alpha_k + \delta x + \gamma_k x, \end{aligned}$$

with $\alpha_k := \int M(y)dA_k(y)$, $\delta x := \int M(y)dD_x(y)$ and $\gamma_k x := \int M(y)dC_{kx}(y)$. This model is equivalent to our model (4.32) with the interaction term $\zeta_k x$ which was obtained without the explicit assumption that the decomposition in (4.47) holds.

Similar calculations show that (4.47) also implies a pairwise PIM. In particular,

$$\begin{aligned} \mathrm{P}\left(Y_{ix_1} \preccurlyeq Y_{kx_2}\right) &= \int F_{ix_1}(y)dF_{kx_2}(y) \\ &= \int \left[M(y) + A_i(y) + D_{x_1}(y) + C_{ix_1}(y)\right] d\left[M(y) + A_k(y) + D_{x_2}(y) + C_{kx_2}(y)\right], \end{aligned}$$

which gives 16 terms. Upon making similar assumptions as for the marginal PIM, and making use of the property $\int U(y)dV(y) + \int V(y)dU(y) = 0$, for U and V any of terms in (4.47), we find

$$\begin{aligned} &\mathrm{P}\left(Y_{ix_1} \preccurlyeq Y_{kx_2}\right) \\ = \;& \frac{1}{2} + \alpha_k - \alpha_i + (\alpha\alpha)_{ik} + \delta(x_2 - x_1) + \\ & \left[(\alpha\delta)_i + (\alpha\gamma)_{ik} + \gamma_k\right] x_2 - \left[(\alpha\delta)_k + (\alpha\gamma)_{ik} + \gamma_i\right] x_1 + \\ & \left[(\delta\gamma)_k - (\delta\gamma)_i + (\gamma\gamma)_{ik}\right] x_1 x_2, \end{aligned}$$

where the Greek letters α, γ, and δ are used to denote the parameters originating from the Roman letters A, C, and D in (4.47), and parameters formed by two Greek letters show from which integral they have resulted. The pairwise PIM may be reparametrized to

$$\mathrm{P}\left(Y_{ix_1} \preccurlyeq Y_{kx_2}\right) = \frac{1}{2} + \alpha_k - \alpha_i + (\alpha\alpha)_{ik} + \delta(x_2 - x_1) + \beta_{1ik}x_1 + \beta_{2ik}x_2 + \beta_{12ik}x_1x_2. \quad (4.48)$$

This model, in the absence of the interaction terms with the β-parameters, is equivalent to our model (4.33). Obviously, the established relationship depends on very stringent assumptions on the A, C, and D functions, particularly when the continuous covariate is involved. Similar assumptions were also used by Akritas et al. (2000) to show how their model relates to a linear model for the conditional mean outcome.

4.8 Simulation study

In this section we present the results of a simulation study to examine the empirical performance of the KW and Friedman test (KW_s and F_s), their Wald-type variants (KW_w and F_w), and the BH test (BH_s). Note that these Wald-type tests are new tests generated from a PIM. We consider balanced three-sample designs with and without blocks.

4.8.1 Empirical type I error

The empirical type I error is evaluated for observations simulated from a standard normal distribution and a t-distribution with 2 degrees of freedom. Sample sizes of $n = 5$, 25, 50, 75, and 150 per group are considered. Table 4.1 gives the empirical rejection rates at the 1%, 5%, and 10% levels of significance based on 10000 simulation runs, where both the permutation null distribution (approximated by 5000 permutations) and the asymptotic null distribution are considered. The results demonstrate that with the permutation null distribution, all tests correctly control for the type I error and with the asymptotic null distribution, both score-tests KW_s and BH_s have empirical rejection rates close to the nominal level, even for small samples. The Wald-type test KW_w, however, only correctly controls for the type I error rate if $n \geq 75$. This may be a consequence of its extra variability caused by the use of an estimated variance for standardization.

Table 4.1: Empirical type I error rates (%) at the 1%, 5%, and 10% levels of significance. Data are simulated from a standard normal distribution, $\mathrm{N}(0, 1)$, and a t-distribution with 2 degrees of freedom, t_2. The number of observations of each group is denoted by n.

n	KW_s			KW_w			BH_s		
	1%	5%	10%	1%	5%	10%	1%	5%	10%
	Permutation null distribution								
	$\mathrm{N}(0, 1)$								
5	1.04	4.97	10.18	1.17	5.43	10.17	0.90	4.95	10.26
25	0.92	5.05	10.60	0.89	5.30	10.15	0.86	5.16	10.74
50	0.86	4.68	9.26	0.95	4.73	9.32	1.08	4.48	9.53
	t_2								
5	1.07	4.77	10.09	1.07	5.18	9.78	0.91	4.74	9.90
25	0.82	5.15	10.12	0.95	5.34	10.05	0.76	5.30	10.32
50	1.08	4.89	9.57	1.13	4.96	9.49	1.10	4.63	9.81
	Asymptotic null distribution								
	$\mathrm{N}(0, 1)$								
5	0.36	4.53	8.92	17.66	25.74	31.99	0.93	4.60	9.98
25	0.85	4.76	9.68	2.52	7.82	13.90	1.57	5.59	10.49
75	0.86	4.63	9.38	1.31	5.68	10.91	1.58	5.84	10.52
150	0.99	5.11	10.26	1.30	5.72	10.85	1.49	6.16	10.97
	t_2								
5	0.24	4.26	9.27	17.70	25.48	31.67	0.87	4.48	9.70
25	0.90	4.74	9.91	2.70	7.84	13.55	1.46	5.63	10.24
75	0.83	4.82	9.48	1.26	5.78	10.57	1.31	5.11	9.83
150	0.99	5.04	10.16	1.30	5.72	10.85	1.61	6.11	10.82

4.8.2 Location-shift

To examine the empirical power under location-shift, data are generated with

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \quad i = 1, 2, 3, \quad j = 1, \ldots, n,$$

with $\varepsilon \overset{d}{=} \mathrm{N}(0, 1)$ or $\varepsilon \overset{d}{=} t_2$, and $\mu_i = i/4$. Using notation (4.20), for $\mathrm{N}(0, 1)$ it holds that $P_{12} \approx 57\%$, $P_{13} \approx 64\%$, and $P_{23} \approx 57\%$, while $P_1 \approx 43\%$, $P_2 \approx 50\%$, and $P_3 \approx 57\%$. For t_2 this is $P_{12} \approx 56\%$, $P_{13} \approx 61\%$, and $P_{23} \approx 55\%$, while $P_1 \approx 45\%$, $P_2 \approx 50\%$, and $P_3 \approx 55\%$. The permutation null distribution is used for p-value calculation and approximated by 5000 permutations. All results are based on 10000 simulation runs and testing at the 5% level of significance. Table 4.2 shows the results. KW_s and KW_w have similar powers for $n = 5$ and $n = 25$. For all sample sizes, BH_s has lower power than the KW-tests. Since location-shift implies PI-transitivity, the BH_s test suffers from a dilution effect by also including a component that aims at detecting intransitivity.

Table 4.2: Empirical powers (%) at the 5% level of significance for the location shift model. Errors are simulated from a standard normal distribution, $\mathrm{N}(0, 1)$, and a t-distribution with 2 degrees of freedom, t_2. The number of observations of each group is denoted by n.

n	KW_s	KW_w	BH_s	KW_s	KW_w	BH_s
	$\mathrm{N}(0,1)$			t_2		
5	8.42	8.63	7.42	6.90	7.13	6.51
25	31.03	31.74	26.08	17.92	18.39	16.88
50	56.40	56.17	44.97	32.25	32.11	27.17

4.8.3 No location-shift but transitive

To examine the power properties when the location-shift model does not hold, data are simulated from the standard normal distribution (referred to as group 1), Laplace distribution with location parameter 0.5 and scale parameter 1 (group 2), and the Gumbel distribution with location parameter 0 and scale parameter 1 (group 3). The left panel of Figure 4.1 shows these densities, for which $P_{12} \approx 63\%$, $P_{13} \approx 62\%$, and $P_{23} \approx 49\%$, while $P_1 = 42\%$, $P_2 = 55\%$, and

$P_3 = 54\%$. The permutation null distribution is used for p-value calculation. Table 4.3 gives the empirical powers. The KW tests have similar powers, and BH_s has the smallest power. Similar as for the location-shift model, the distributions are PI-transitive and hence the BH_s test looses power.

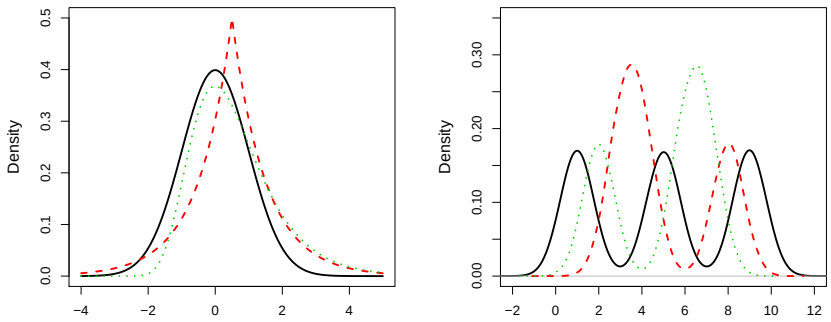

Figure 4.1: Left: Densities of the standard normal distribution (—), the Laplace distribution with location parameter 0.5 and scale parameter 1 ($- - -$), and the Gumbel distribution with location parameter 0 and scale parameter 1 ($\cdots$). Right: Densities corresponding to die Ω_1 (—), Ω_2 ($- - -$), and Ω_3 ($\cdots$).

Table 4.3: Empirical powers (%) at the 5% level of significance when the location-shift model does not hold for different group sample sizes n

n	KW_s	KW_w	BH_s
5	9.20	8.89	8.33
25	31.51	32.96	27.66
50	61.58	63.29	53.18

4.8.4 Intransitive

Brown and Hettmansperger (2002) provided an algorithm to simulate data from intransitive distributions based on the dice given in Section 4.4.2. For each die, a mixture distribution of six normal distributions with fixed variance and mean equal to a marking of the die is considered. The right panel of Figure 4.1 shows the densities when the variance is equal to 0.25. Table 4.4 shows the empirical powers when the permutation null distributions are used for p-value calculation. Both KW tests have virtually no power, because $\mathrm{P}\left(Y_{\cdot} \preccurlyeq Y_k\right) = 0.5$ for $k = 1, 2, 3$. The empirical powers are even zero at the 5% significance level, indicating that the KW tests are biased for this extreme situation. Since $\mathrm{P}\left(Y_k \preccurlyeq Y_l\right) \neq 0.5$, the BH test has non-trivial power for $n \geq 30$.

Table 4.4: Empirical powers (%) at the 5% level of significance when transitivity does not hold but $\mathrm{P}\left(Y_{\cdot} \preccurlyeq Y_k\right) = 0.5$ for $k = 1, 2, 3$, for different group sample sizes n

n	KW_s	KW_w	BH_s
12	0	0	2.70
30	0	0	70.90
60	0	0	99.43

4.8.5 Randomized complete blocks

To evaluate the performance of the Friedman test and the Wald-type version, we consider the data generating model

$$Y_{ij} = \mu + \mu_i + \nu_j + \varepsilon_{ij}, \quad i = 1, 2, 3,\ j = 1, \ldots, L,$$

with $\varepsilon \overset{d}{=} \mathrm{N}(0, 1)$ or $\varepsilon \overset{d}{=} t_2$, $\nu_j = j/L$ if $j > 1$, and $\mu_1 = \nu_1 = 0$. To empirically evaluate the type I error rate we set $\mu_i = 0$, and to examine the power we set $\mu_i = (i-1)/K$. The permutation null distribution, based on 5000 permutations, is used for p-value calculation. All results are based on 10000 simulations. Table 4.5 shows the results for several choices of L. The results show that F_s is slightly more liberal than F_w and both tests have a similar power.

Table 4.5: Empirical rejection rates (%) and empirical powers (%) at the 5% level of significance for the randomized complete block design. Errors are simulated from a standard normal distribution, $\mathrm{N}(0,1)$, and a t-distribution with 2 degrees of freedom, t_2. The number of blocks is denoted by L.

L	F_s	F_w	F_s	F_w
	$\mathrm{N}(0,1)$		t_2	
	Empirical rejection rate			
10	6.87	4.91	6.69	4.56
25	5.21	5.23	5.30	5.09
50	5.81	5.72	5.91	5.65
	Empirical power			
10	25.39	18.79	16.58	12.32
25	49.80	48.10	29.66	27.98
50	83.81	83.14	55.87	54.71

4.9 The surgical unit study

In this section we use an example data set to illustrate how a PIM can be used to construct new rank tests when the design is more complex than a K-sample study.

We consider the *surgical unit study* provided by Kutner et al. (2004), section 9.2. The data contain information on 54 patients who underwent a particular type of liver operation and it is of interest to predict the survival based on pre-operation variables. In addition to the survival time (Y, mean 702.1, St. Dev. 397.4) of each patient, several predictors are recorded. We consider: enzyme function test score (X_1: mean 77.1, St. Dev. 21.3), gender (X_2: 0: male 53.7%, 1 : female 46.3%), and history of alcohol use (X_3: 0: none 27.8%, 1: moderate 53.7%, and 2: severe 18.5%). A PIM with the identity link function is inappropriate, because the continuous predictor can cause predictions outside of the unit interval. We consider the logit link function, for which, however, the exact covariance matrix of the parameter estimators does no longer follow from Lemma 5 or Lemma 8. We thus need to rely on the Wald-type tests. We fit two PIMs to the data, as well as a linear model for comparisons purposes. To illustrate the interpretation of each model, we include the effect of the continuous predictor and the effect of

severe versus no alcohol use history. We first consider the linear model, where the outcome is log-transformed to obtain a better fit; see Kutner et al. (2004), section 9.2. In particular,

$$\mathrm{E}(\ln(Y_i) \mid \boldsymbol{X}_i) = \gamma_1 + \gamma_2 X_{1i} + \gamma_3 \mathrm{I}(X_{2i} = 1) + \gamma_4 \mathrm{I}(X_{3i} = 1) + \gamma_5 \mathrm{I}(X_{3i} = 2). \quad (4.49)$$

Table 4.6 gives the estimates, standard errors (SE), and p-values. We conclude that the mean log survival time increases with an estimate of $10\hat{\gamma}_2 = 0.14$ for an enzyme function test score of 10 units higher, while the other predictors remain fixed. Similarly, the mean log survival time is an estimated $\hat{\gamma}_5 = 0.46$ units higher for patients with a severe alcohol use history as compared to patients of the same gender and with the same enzyme function test score, but with no history of alcohol use. It has been reported that moderate alcohol consumption is associated with reduced mortality; see, for example, de Groot and Zock (1998); Foster (2010). On the other hand, this association, for example, can perhaps be caused due to a heterogeneous sample for which patients who have a history of alcohol use are not comparable to patients without a history of alcohol use. The latter group can, for example, consist of patients who are very ill and therefore need a liver operation, while the former group can consist of patients who are healthier, but need to undergo a liver operation because of their drinking habits.

Consider now a marginal PIM with the same covariates as for the linear model. Since the PI is invariant under monotonic transformations, a log transformation is not required. We write the PIM as

$$\begin{aligned} \mathrm{logit}\left[\mathrm{P}\left(Y_i \preccurlyeq Y_j \mid \boldsymbol{X}_j\right)\right] &= \alpha_1 + \alpha_2 X_{1j} + \alpha_3 \mathrm{I}(X_{2j} = 1) + \alpha_4 \mathrm{I}(X_{3j} = 1) + \\ &\quad \alpha_5 \mathrm{I}(X_{3j} = 2). \end{aligned} \quad (4.50)$$

The interpretation of α_2 follows from

$$\exp(\alpha_2) = \frac{\mathrm{odds}\left(Y_i \preccurlyeq Y_j \mid X_{1j} = x + 1, X_{2j}, X_{3j}\right)}{\mathrm{odds}\left(Y_i \preccurlyeq Y_j \mid X_{1j} = x, X_{2j}, X_{3j}\right)}.$$

The odds on a larger survival than the marginal survival of a randomly chosen patient is an estimated $\exp(10\hat{\alpha}_2) = 1.4$ times the corresponding odds of a randomly chosen patient of the same gender and with the same alcohol use history, but with an enzyme function test score which is 10 units lower. Similarly, the odds of having a larger survival than the marginal survival of a randomly chosen patient with a severe history of alcohol use is an estimated $\exp(\hat{\alpha}_5) = 3$ times the corresponding odds of a randomly chosen patient with no history of alcohol use, but with the same gender and enzyme function test score.

We also consider a pairwise PIM, which results in a different, and perhaps simpler interpretation. In particular,

$$\begin{aligned}\text{logit}\left[\mathrm{P}\left(Y_i \preccurlyeq Y_j \mid \boldsymbol{X}_i, \boldsymbol{X}_j\right)\right] &= \beta_1(X_{1j} - X_{1i}) + \beta_2 \mathrm{I}\left(X_{2i} = 0\right) \mathrm{I}\left(X_{2j} = 1\right) + \\ &\quad \beta_3 \mathrm{I}\left(X_{3i} = 0\right) \mathrm{I}\left(X_{3j} = 1\right) + \beta_4 \mathrm{I}\left(X_{3i} = 0\right) \mathrm{I}\left(X_{3j} = 2\right) + \\ &\quad \beta_5 \mathrm{I}\left(X_{3i} = 1\right) \mathrm{I}\left(X_{3j} = 2\right). \end{aligned} \tag{4.51}$$

Since

$$\text{expit}(\beta_1) = \mathrm{P}\left(Y_i \preccurlyeq Y_j \mid X_{1i} = x, X_{1j} = x + 1, X_{2i} = X_{2j}, X_{3i} = X_{3j}\right),$$

$\text{expit}(10\hat{\beta}_1) = 57\%$ is the estimated probability that the survival is larger for a randomly chosen patient as compared to a randomly chosen patient of the same gender and with the same alcohol use history, but for which enzyme function score is 10 units lower. Similarly, $\text{expit}(\beta_4) = \mathrm{P}(Y_i \preccurlyeq Y_j \mid X_{3i} = 0, X_{3j} = 2, X_{1i} = X_{1j}, X_{2i} = X_{2j})$, thus $\text{expit}(\hat{\beta}_4) = 88\%$ is the estimated probability that the survival is higher for a randomly chosen patient with a severe history of alcohol use as compared to a randomly chosen patient with no history of alcohol use, and the same enzyme function test score and gender. Model (4.51) allows us now to extend the JT test for testing versus the ordered alternative in terms of the alcohol history while accounting for the gender and enzyme function test score. The test statistic is constructed along the lines of Theorem 9, based on the standardized contrast $\hat{\beta}_3 + \hat{\beta}_4 + \hat{\beta}_5$. In particular, the null hypothesis of equal distributions is rejected in favour of the ordered alternative, which states that the survival time increases as the history of alcohol use becomes more severe (p-value: 0.008). For the JT test without the adjustment of the enzyme function test score and gender, the p-value is 0.0122.

The linear model showed no lack-of-fit (results not shown) and for both the marginal and pairwise PIM the goodness-of-fit (GOF) should be assessed. In Chapter 5, GOF methods are developed and used to assess the model adequacy of the PIMs which are fitted in the case studies of Chapters 2 and 3. However, both the marginal and pairwise PIM are, in a way, more complicated PIMs, and the current version of the software for assessing GOF of PIMs does not support these models. Therefore, although important, assessing the GOF of these models is beyond the scope of this dissertation.

Table 4.6: Parameter estimates, standard errors (SE), and p-values for models (4.49), (4.50), and (4.51)

	Linear Model				
	γ_1	γ_2	γ_3	γ_4	γ_5
Estimate	5.15	0.014	0.16	0.10	0.46
SE	0.194	0.002	0.095	0.109	0.141
p-value	< 0.001	< 0.001	0.09	0.35	0.002
	Marginal PIM				
	α_1	α_2	α_3	α_4	α_5
Estimate	-3.31	0.035	0.43	0.34	1.11
SE	0.647	0.010	0.276	0.323	0.380
p-value	< 0.001	< 0.001	0.12	0.29	0.003
	Pairwise PIM				
	β_1	β_2	β_3	β_4	β_5
Estimate	0.028	0.57	0.71	2.04	1.27
SE	0.014	0.400	0.441	0.978	0.693
p-value	0.042	0.15	0.11	0.037	0.064

4.10 Discussion

In this chapter it is shown how two parametrizations of a PIM can lead to rank tests for factorial designs. Based on the marginal PIM parametrization, the Kruskal–Wallis (KW) test statistic for a K-sample design and the Mack–Skillings and Friedman tests for a randomized complete block design, arise naturally. The pairwise PIM results in the Wilcoxon–Mann–Whitney test for the two sample design and the Brown–Hettmansperger (BH) test for the three sample design. For the pairwise PIM, the Jonckheere–Terpstra (JT) and the Mack–Wolfe (MW) tests also arise naturally. All these rank tests are score tests in the sense that their variances are obtained under the null hypothesis. The PIM theory, however, provides a sandwich estimator of the variance which is also consistent under the alternative. This allows to construct Wald-type versions of these rank tests, as well as confidence intervals for the effect sizes. A simulation study is performed for evaluating the performance of some of these tests. It is concluded that the rank tests and their Wald-type versions have similar powers. The Wald-type tests, however, are more liberal for small samples. The BH test has lowest power relative to the KW tests if PI-transitivity holds. However, for PI-intransitive data, the BH test has superior power.

The PIM representation of rank tests allows extending rank tests for more complicated designs, when, for example, a continuous confounder or multiple predictors are present. Furthermore, the PIM representation also allows to extend the BH, JT, and MW tests to block designs.

The classical rank tests are very often referred to as nonparametric tests, but this term may be misleading. Apart from some very simple settings (e.g. K-sample problem) rank tests relate to parameters of a semiparametric model which expresses restrictions on the distribution of the outcome variable. In this chapter we have demonstrated that the PIM is a natural model for rank tests. Akritas and Arnold (1994) proposed another model for which they developed rank tests, which, however, do not generally reduce to the classical tests. Their methodology was extended to several designs and to the inclusion of continuous covariates (Akritas et al., 1997, 2000; Brunner and Puri, 2002; Tsangari and Akritas, 2004). Their test statistics are rank-transform statistics, in the sense that they are functions of the rank-transformed outcome observations. Although their methods also rely on a model that expresses a restriction on the outcome distribution function, they cannot always estimate all terms in their model (Tsangari and Akritas, 2004). At this point it is also interesting to mention that the simple rank-transform methods of Conover and Iman (1981) and Hora and Conover (1984) do not always relate clearly to a

statistical model. The method consists in transforming the outcome observations to their ranks and subsequently using these transformed observations in parametric methods. For example, the two-sample *t*-test and the one-way ANOVA *F*-test applied to the rank-transformed data gives the WMW and the KW test, respectively. However, for more complicated designs Akritas (1990) showed that the parametric statistical model does no longer hold after the transformation. For example, the two-way ANOVA without interaction implies additivity of the effects on the mean outcome, but this additivity is lost with the transformation. Without explicitly referring to the probabilistic index, he made the connection. In particular, upon using asymptotic arguments, he replaced the rank-transformed outcome of Y_i with $nF(Y_i)$, with F the marginal distribution function of the outcome. When the outcome is continuous, the original parametric model that models $\mathrm{E}(Y_i \mid \boldsymbol{X}_i)$ becomes $\mathrm{E}(nF(Y_i) \mid \boldsymbol{X}_i) = n\mathrm{P}\left(Y \leq Y_i \mid \boldsymbol{X}_i\right)$, which resembles the marginal PIM. The additivity of the effects on $\mathrm{E}(Y_i \mid \boldsymbol{X}_i)$ thus becomes additive in the marginal PIM with identity link. To some extent, the PIM may also be seen as a two-stage approach in which first the n sample observations Y_i are transformed to pseudo-observations $\mathrm{I}\left(Y_i \preccurlyeq Y_j\right)$ which are subsequently used as outcome observations in a linear regression model. By restricting the set of pseudo-observations to comparisons within blocks, block designs can be analyzed. However, despite this apparently simple trick, it is not encouraged to look at it this way. Instead it is preferred to interpret the PIM within a genuine semiparametric modelling framework. This will help in ensuring correct interpretation and reporting of the analysis results.

Many rank tests are based on highly parametric models that express a location-shift effect. For example, the WMW test is the optimal rank test for detecting shifts in means when the observations show a logistic distribution; see, for example, Hájek et al. (1999). However, most statisticians choose for rank tests when no distributional assumptions can be made. Therefore, the relationship between rank tests and PIMs, as discussed in this chapter, can perhaps contribute to a better understanding of rank tests in the absence of such assumptions.

Chapter 5

Assessing the goodness-of-fit

The content of this chapter is primarily based on the results published in

De Neve, J., Thas, O., and Ottoy, J.P. (2013) Goodness-of-fit methods for probabilistic index models. *Communications in Statistics: Theory and Methods*, 42:1193–1207.

5.1 Introduction

The PIM, just like any parametric or semiparametric regression model, expresses restrictions on the joint distribution of the outcome and the covariates. It is important to assess the validity of the model for a given data set and to examine whether the proposed model is consistent with the underlying data-generating model. Consequently, formal goodness-of-fit (GOF) methods and graphical diagnostic tools are needed to assess model adequacy.

We first resume the general formulation of a PIM. Let $(Y, \boldsymbol{X})$ and $(Y', \boldsymbol{X}')$ be i.i.d., then a PIM is defined as

$$\mathrm{P}\left(Y \preccurlyeq Y' \mid \boldsymbol{X}, \boldsymbol{X}'\right) = m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}) = g^{-1}(\boldsymbol{Z}^T\boldsymbol{\beta}), \quad (\boldsymbol{X}, \boldsymbol{X}') \in \mathcal{X}, \tag{5.1}$$

with $g(\cdot)$ and link function and $\boldsymbol{Z}$ a p-vector with elements that may depend on $\boldsymbol{X}$ and $\boldsymbol{X}'$. $\mathcal{X}$ denotes the set of covariates $(\boldsymbol{X}, \boldsymbol{X}')$ for which the model is defined; throughout this section this will be the lexicographical order restriction, because all models satisfy the antisymmetry condition; see Section 2.3.2 for more information.

To illustrate our setting we consider the Childhood Respiratory Disease Study (CRDS) which

is also discussed in Section 2.5.1. The outcome variable is the forced expiratory volume (FEV in litres), and the age (AGE in years) and smoking indicator (SMOKE = 1 if the child smokes, SMOKE = 0 if the child does not smoke) are recorded for 654 children of ages $3 - 19$ years. When analyzing the effect of smoking on the lung capacity, age may be a confounder, and therefore should be taken into account. A part of the data is illustrated in Figure 5.1, which shows nonparametric density estimates of the FEV distributions for several combinations of smoking status and age. If we fit a linear PIM with logit link

$$\begin{aligned} &\text{logit}\left\{\text{P}\left[\text{FEV} \preccurlyeq \text{FEV}' \mid (\text{SMOKE}, \text{AGE}), (\text{SMOKE}', \text{AGE}')\right]\right\} \\ &= \beta_1(\text{AGE}' - \text{AGE}) + \beta_2(\text{SMOKE}' - \text{SMOKE}), \end{aligned} \tag{5.2}$$

We find $\hat{\beta}_1 = 0.56$ (SE : 0.028 and $p < 0.0001$) and $\hat{\beta}_2 = -0.46$ (SE : 0.25 and p : 0.064). The estimated probability that FEV is larger for a smoking child as compared to a non-smoker of the same age is $\hat{\text{P}}[\text{FEV} \preccurlyeq \text{FEV}' \mid \text{SMOKE} = 0, \text{SMOKE}' = 1, \text{AGE} = \text{AGE}'] = \text{expit}(-0.46) = 39\%$. It is thus unlikely that a smoker has a better pulmonary function than a non-smoker of the same age. The effect is not significant at the 5% level of significance, which is surprising, as it is expected that smoking affects a child's lungs. So perhaps the data contain no evidence for this hypothesis or the study is underpowered. However, the lack of significance may also arise when the model does not fit the data properly. Before drawing conclusions about the effect of smoking on the lung function, it is therefore important to first assess the GOF of model (5.2).

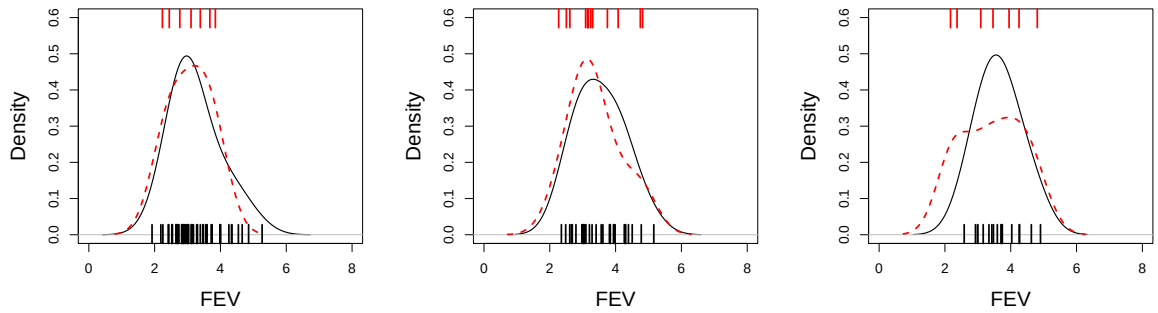

Figure 5.1: Kernel density estimates of the FEV distributions for non-smokers (—) and smokers ($- - -$) of age 12 years (left), 13 years (middle), and 14 years (right). The densities are estimated using a Gaussian kernel with a bandwidth of 0.5. Beneath (non-smokers) and above (smokers) each kernel density plot is a rug plot to identity better the individual sample observations that are used for the density estimation.

In Section 5.2 a GOF test and a graphical diagnostic tool are developed. Section 5.3 assesses the performance of the test in a simulation study, and in Section 5.4, the GOF methods are illustrated on the case studies of Sections 2.5 and 3.2.3. Section 5.5 gives the conclusions and discussion.

5.2 Goodness-of-fit methods

5.2.1 Rationale

We start by considering a single continuous predictor in a specific setting to explain the rationale; the extension to multiple predictors is addressed at the end of the section.

To formally introduce the GOF null hypothesis we denote by $m_0(X, X')$ the PIM which is consistent with the data-generating model, referred to as the *true model*, and we denote by $m(X, X'; \boldsymbol{\beta})$ the PIM that will be fitted to the data, i.e. the right hand side of (5.1), referred to as the *working model*. The GOF null hypothesis is

$$H_0 : m_0(X, X') = m(X, X'; \boldsymbol{\beta}), \quad (X, X') \in \mathcal{X}, \tag{5.3}$$

for some $\boldsymbol{\beta} \in \mathbb{R}^p$. Let the quadratic probit PIM be the true model and the linear probit PIM be the working model, i.e.

$$m_0(X, X') = \Phi\left[\beta_1(X' - X) + \beta_2\left(X'^2 - X^2\right)\right], \quad m(X, X'; \beta) = \Phi\left[\beta(X' - X)\right].$$

The null hypothesis (5.3) can now be written as

$$H_0 : \beta_2 = 0.$$

Consider the following settings: $\beta_1 = 0.3$, β_2 takes the values 0, -0.05, and -0.20 and the predictor X takes n equidistant values in $[-5, 5]$. When $\beta_2 = 0$ there is no quadratic effect and the null hypothesis (5.3) holds, while when $\beta_2 = -0.05$ ($\beta_2 = -0.20$) there is a weak (strong) quadratic effect and the null hypothesis does not hold.

Since a PIM involves a couple of predictors (X, X'), a 3-dimensional plot is needed for visualization; see Figure 5.2. Although this plot provides all information, it is difficult to interpret. We therefore restrict (X, X') to a number of values which are relevant for the interpretation. When

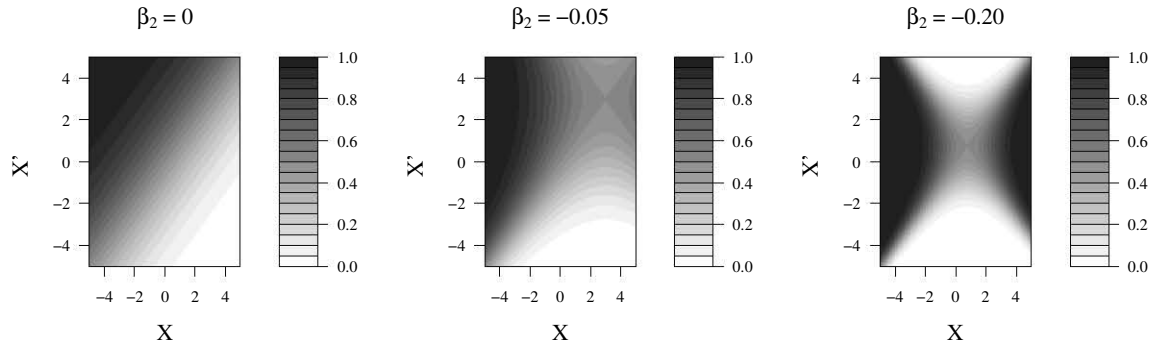

Figure 5.2: Quadratic probit PIM $\mathrm{P}\left(Y \preccurlyeq Y' \mid X, X'\right) = \Phi[\beta_1 \left(X' - X\right) + \beta_2 \left(X'^2 - X^2\right)]$, with $\beta_1 = 0.3$ as a function of X and X'. A grey coding is used to indicate the value of $\mathrm{P}\left(Y \preccurlyeq Y' \mid X, X'\right)$.

Δ denotes a fixed value, we restrict the plot to $\mathrm{P}\left(Y \preccurlyeq Y' \mid X, X' = X + \Delta\right)$, i.e. the probability that the outcome increases when the predictor is increased by Δ units. For the example setting, we can write

$$m_0(X, X' = X + \Delta) = \Phi(\tilde{\beta}_1 + \tilde{\beta}_2 X), \quad \tilde{\beta}_1 = \beta_1 \Delta + \beta_2 \Delta^2, \quad \tilde{\beta}_2 = 2\beta_2 \Delta. \tag{5.4}$$

Equation (5.4) indicates that the choice of Δ is important. As Δ increases, the difference between $m_0(X, X' = X + \Delta)$ and $m(X, X' = X + \Delta; \beta) = \Phi(\beta\Delta)$ becomes more pronounced; see Figure 5.3. Consider the left panel where $\Delta = 1$. When the linear PIM holds, i.e. $\beta_2 = 0$, $\mathrm{P}\left(Y \preccurlyeq Y' \mid X, X' = X + 1\right)$ is fixed at $\Phi(\tilde{\beta}_1) = \Phi(0.3) \approx 62\%$ and independent of X. However, with increasing magnitude of β_2, this probability depends more strongly on the predictor X. When $\beta_2 = -0.20$, for example, it holds that $\mathrm{P}\left(Y \preccurlyeq Y' \mid X, X' = X + 1\right) > 95\%$ for $X < -4$, while for $X > 4$ this becomes $\mathrm{P}\left(Y \preccurlyeq Y' \mid X, X' = X + 1\right) < 7\%$. The restricted probability provides information on the difference between a quadratic and linear PIM, while retaining a simple interpretation.

If $m_0(\cdot)$ and $\boldsymbol{\beta}$ are known the plot suggests that comparing $m_0(X, X' = X + \Delta)$ with $m(X, X' = X + \Delta; \boldsymbol{\beta})$ captures information on the adequacy of the model fit. For a point x, consider the difference $R_0 = m_0(x, x' = x + \Delta) - m(x, x' = x + \Delta; \boldsymbol{\beta})$. If the working model provides a good approximation R_0 will be close to zero; if the models differ substantially, R_0 provides information on how to improve the working model. For practical use $m_0(\cdot)$ can be replaced with a nonparametric kernel estimator, say $\hat{m}_0(\cdot)$, and $\boldsymbol{\beta}$ by a consistent estimator $\hat{\boldsymbol{\beta}}$, but a drawback

of this approach is that $\hat{m}_0(\cdot)$ may be biased (le Cessie and van Houwelingen, 1991). Therefore, we consider a kernel estimator of R_0 that is based on the residuals

$$R_{ij}(\hat{\boldsymbol{\beta}}) := \mathrm{I}\left(Y_i \preccurlyeq Y_j\right) - m(X_i, X_j; \hat{\boldsymbol{\beta}}).$$

Since the conditional expectation under H_0 is zero, there is no bias (le Cessie and van Houwelingen, 1991; Hardle and Mammen, 1993). For a fixed Δ, we obtain a graphical tool by plotting the smoothed residuals as a function of the predictor and we construct a statistical test by considering a quadratic form of these residuals.

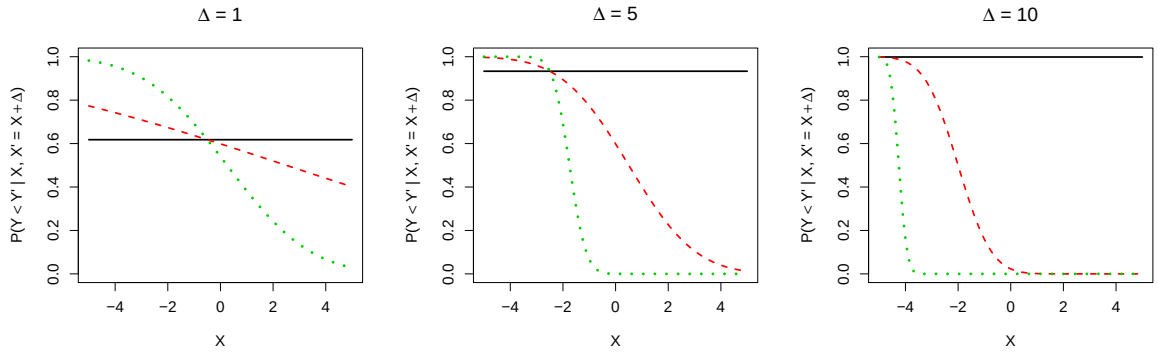

Figure 5.3: Quadratic probit PIM $\mathrm{P}\left(Y \preccurlyeq Y' \mid X, X'\right) = \Phi[\beta_1(X'-X) + \beta_2\left(X'^2 - X^2\right)]$ with predictors restricted to $X' = X + \Delta$, with $\Delta = 1$ (left), $\Delta = 5$ (middle) and $\Delta = 10$ (right), $\beta_1 = 0.3$ and $\beta_2 = 0$ (—), $\beta_2 = -0.05$ (− − −), and $\beta_2 = -0.2$ ($\cdots$)

5.2.2 The goodness-of-fit test

To construct a kernel estimator based on the residuals $R_{ij}(\hat{\boldsymbol{\beta}})$ and since a PIM depends on (X, X'), we need to define appropriate *kernels* for our setting. Consider, for example, a multivariate kernel (Silverman, 1986)

$$K_{h_1,h_2}(x, x'; X, X') = D\left(\frac{x - X}{h_1(x)}\right) D\left(\frac{x' - X'}{h_2(x')}\right), \tag{5.5}$$

where h_1 and h_2 are bandwidths that may depend on x and x' and D is a kernel function. Examples of D include the Gaussian D_G, boxcar D_B, or Epanechnikov D_E kernel function, given by

$$D_G(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right), \quad D_B(x) = \frac{1}{2}\mathrm{I}\left(|x| \leq 1\right), \quad D_E(x) = \frac{4}{3}\left(1 - x^2\right)\mathrm{I}\left(|x| \leq 1\right).$$

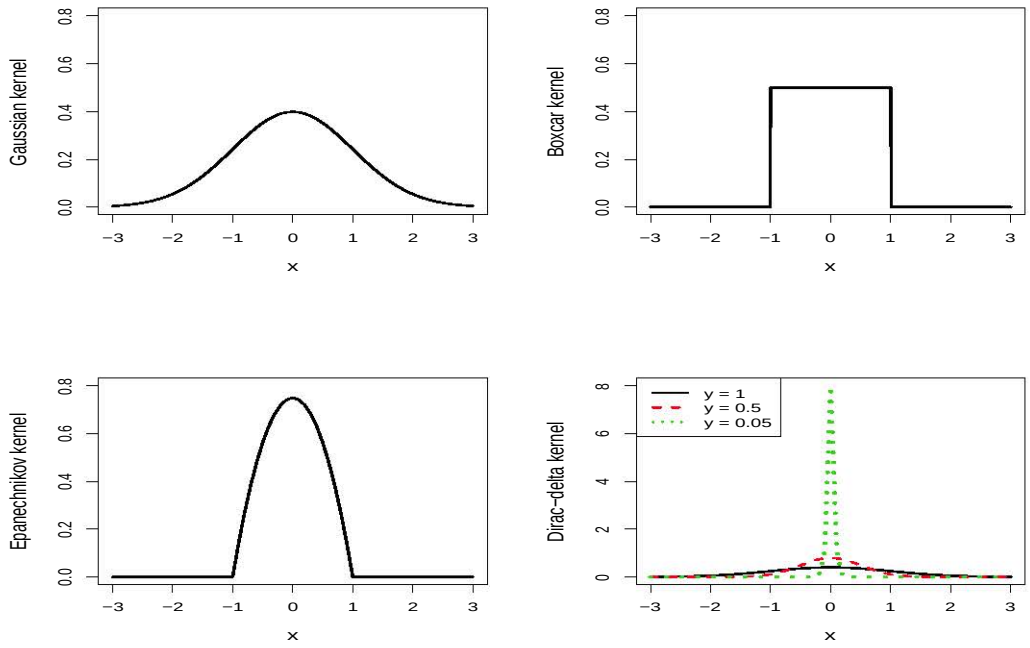

Figure 5.4: Gaussian, boxcar, and Epanechnikov kernels. The Dirac-delta function is approximated for several values of y.

Figure 5.4 shows these different kernel functions. The kernel (5.5) provides double smoothing, i.e. for each (X, X'), we consider the distance between X and x, and between X' and x'. More weight is given to couples for which simultaneously X is close to x and X' to x'. If no smoothing is desired, which, for example, may happen when a categorical predictor has sufficient replicates, we write $h_1 = h_2 = 0$ and denote by D the Dirac-delta function D_D, which can be defined as

$$D_D(x) = \lim_{y \to 0} \frac{1}{y\sqrt{2\pi}} \exp\left(-\frac{x^2}{2y^2}\right). \tag{5.6}$$

The bottom right panel of Figure 5.4 shows the right hand side of equation (5.6). As y goes to zero, the function becomes zero except at the origin for which its value is infinity.

For notional convenience we drop the dependence on h_1 and h_2 in (5.5) and write $K(x, x'; X, X')$ instead of $K_{h_1,h_2}(x, x'; X, X')$. A Nadaraya–Watson kernel estimator (Nadaraya, 1964; Watson, 1964) based on the residuals is defined by

$$\hat{R}(x, x') := \frac{\sum_{(k,l)\in\mathcal{I}_n} R_{kl}(\hat{\boldsymbol{\beta}}) K(x, x'; X_k, X_l)}{\sum_{(k,l)\in\mathcal{I}_n} K(x, x'; X_k, X_l)}. \tag{5.7}$$

To derive the asymptotic null distribution of these *smoothed residuals*, let $\boldsymbol{K}(x, x')$ denote the

$|\mathcal{I}_n|$-vector with elements

$$\frac{K(x, x'; X_k, X_l)}{\sum_{(k,l)\in\mathcal{I}_n} K(x, x'; X_k, X_l)}.$$

Furthermore, let $\boldsymbol{I}_p$ denote the $|\mathcal{I}_n|$-vector of pseudo-observations $\mathrm{I}(Y_i \preccurlyeq Y_j)$, $\boldsymbol{m}(\boldsymbol{\beta})$ the $|\mathcal{I}_n|$-vector with elements $m(X_i, X_j; \boldsymbol{\beta})$, and $\boldsymbol{V}(\boldsymbol{\beta})$ the diagonal matrix with elements $m(X_i, X_j; \boldsymbol{\beta})[1 - m(X_i, X_j; \boldsymbol{\beta})]$. Let $\boldsymbol{\beta}_0$ denote the true parameter, as defined by (2.18), and define

$$\boldsymbol{R}(\hat{\boldsymbol{\beta}}) = \boldsymbol{I}_p - \boldsymbol{m}(\hat{\boldsymbol{\beta}}), \quad \boldsymbol{R}(\boldsymbol{\beta}_0) = \boldsymbol{I}_p - \boldsymbol{m}(\boldsymbol{\beta}_0). \tag{5.8}$$

Upon using notation (5.8), we can write the smoothed residual (5.7) as

$$\hat{R}(x, x') = \boldsymbol{K}(x, x')^T \boldsymbol{R}(\hat{\boldsymbol{\beta}}).$$

The following theorem gives the asymptotic null distribution of the smoothed residual. For simplicity let $h := h_1 = h_2$ denote the bandwidth.

Theorem 12. *For $\hat{R}(x, x')$, as defined by (5.7), it holds that for a fixed x and x', as $h \to 0$ and $nh \to \infty$, under H_0,*

$$\frac{\hat{R}(x, x')}{\sqrt{\mathrm{Var}[\hat{R}(x, x')]}} \xrightarrow{d} N(0, 1),$$

with asymptotic variance

$$\mathrm{Var}\left[\hat{R}(x, x')\right] = \boldsymbol{K}(x, x')^T (\boldsymbol{I} - \boldsymbol{H}) \,\mathrm{Cov}\,(\boldsymbol{I}_p) \,(\boldsymbol{I} - \boldsymbol{H})^T \boldsymbol{K}(x, x'),$$

with $\boldsymbol{H}$ as defined in the proof by (5.12).

Proof. First note that the estimating equations (2.15) with index function (2.16) can be concisely written as

$$\boldsymbol{U}_n(\boldsymbol{\beta}) = \frac{\partial \boldsymbol{m}(\boldsymbol{\beta})^T}{\partial \boldsymbol{\beta}} \boldsymbol{V}(\boldsymbol{\beta})^{-1} [\boldsymbol{I}_p - \boldsymbol{m}(\boldsymbol{\beta})] = \boldsymbol{0}.$$

To emphasise the dependence of $\hat{\boldsymbol{\beta}}$ on the sample size n, we write $\hat{\boldsymbol{\beta}}_n$. Consider a Taylor expansion of $\boldsymbol{m}(\hat{\boldsymbol{\beta}}_n)$ about $\boldsymbol{\beta}_0$

$$\boldsymbol{m}(\hat{\boldsymbol{\beta}}_n) = \boldsymbol{m}(\boldsymbol{\beta}_0) + \frac{\partial \boldsymbol{m}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} (\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) + o_p(n^{-1/2}). \tag{5.9}$$

A Taylor expansion of $\boldsymbol{U}_n(\hat{\boldsymbol{\beta}}_n)$ leads to

$$\begin{aligned} \boldsymbol{0} &= |\mathcal{I}_n|^{-1} \boldsymbol{U}_n(\hat{\boldsymbol{\beta}}_n) = |\mathcal{I}_n|^{-1} \boldsymbol{U}_n(\boldsymbol{\beta}_0) + |\mathcal{I}_n|^{-1} \frac{\partial \boldsymbol{U}_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} (\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) + o_p(n^{-1/2}) \\ \Leftrightarrow (\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) &= -\left(\frac{\partial \boldsymbol{U}_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right)^{-1} \boldsymbol{U}_n(\boldsymbol{\beta}_0) + o_p(n^{-1/2}). \end{aligned} \tag{5.10}$$

Combining (5.9) and (5.10), it follows that

$$\boldsymbol{m}(\hat{\boldsymbol{\beta}}_n) = \boldsymbol{m}(\boldsymbol{\beta}_0) - \frac{\partial \boldsymbol{m}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \left(\frac{\partial \boldsymbol{U}_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \right)^{-1} \boldsymbol{U}_n(\boldsymbol{\beta}_0) + o_p(n^{-1/2}).$$

If we substitute $\boldsymbol{U}_n(\boldsymbol{\beta}_0)$ in this expression, then

$$\boldsymbol{m}(\hat{\boldsymbol{\beta}}_n) = \boldsymbol{m}(\boldsymbol{\beta}_0) + \boldsymbol{H}\left[\boldsymbol{I}_p - \boldsymbol{m}(\boldsymbol{\beta}_0)\right] + o_p(n^{-1/2}), \tag{5.11}$$

where

$$\boldsymbol{H} = -\frac{\partial \boldsymbol{m}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \left(\frac{\partial \boldsymbol{U}_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \right)^{-1} \frac{\partial \boldsymbol{m}(\boldsymbol{\beta})^T}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \boldsymbol{V}(\boldsymbol{\beta}_0)^{-1}. \tag{5.12}$$

From (5.8) and (5.11) it follows that

$$\begin{aligned} \hat{R}(x,x') &= \boldsymbol{K}(x,x')^T \boldsymbol{R}(\hat{\boldsymbol{\beta}}_n) \\ &= \boldsymbol{K}(x,x')^T \left(\boldsymbol{I} - \boldsymbol{H}\right) \boldsymbol{R}(\boldsymbol{\beta}_0) + o_p(n^{-1/2}). \end{aligned}$$

Consequently, under H_0 (5.3), the asymptotic expectation and variance are given by

$$\mathrm{E}\left[\hat{R}(x,x')\right] = 0, \quad \mathrm{Var}\left[\hat{R}(x,x')\right] = \boldsymbol{K}(x,x')^T \left(\boldsymbol{I} - \boldsymbol{H}\right) \mathrm{Cov}\left(\boldsymbol{I}_p\right) \left(\boldsymbol{I} - \boldsymbol{H}\right)^T \boldsymbol{K}(x,x').$$

The central limit theorem of Lumley and Mayer-Hamblett (2003) (Theorem 4, page 13) now guarantees that, under H_0 (5.3),

$$\frac{\hat{R}(x,x')}{\sqrt{\mathrm{Var}[\hat{R}(x,x')]}} \xrightarrow{d} \mathrm{N}(0,1).$$

□

Similar as in Pan (2002) and Evans and Li (2005), a consistent estimator of $\mathrm{Var}[\hat{R}(x,x')]$ can be obtained by substituting $\boldsymbol{\beta}_0$ by $\hat{\boldsymbol{\beta}}$ in $\boldsymbol{H}$ and $\mathrm{Cov}\left(\boldsymbol{I}_p\right)$ by $\hat{\boldsymbol{\Sigma}}_p$ defined as

$$\left(\hat{\boldsymbol{\Sigma}}_p\right)_{(ij),(kl)} = \begin{cases} \left[\mathrm{I}\left(Y_i \preccurlyeq Y_j\right) - m(X_i, X_j; \hat{\boldsymbol{\beta}})\right] \left[\mathrm{I}\left(Y_k \preccurlyeq Y_l\right) - m(X_k, X_l; \hat{\boldsymbol{\beta}})\right], & \text{if } \phi_{ijkl} = 1, \\ 0, & \text{if } \phi_{ijkl} = 0, \end{cases} \tag{5.13}$$

with ϕ_{ijkl} as defined in Theorem 2. For more details, see Lumley and Mayer-Hamblett (2003, p. 18).

Theorem 12 allows to construct approximate confidence bounds for each smoothed residual. Moreover, the theorem does not only hold for the Nadaraya–Watson smoothers, but it holds for *linear smoothers* in general. The definition of a linear smoother is given below and is adapted from definition 5.17 in Wasserman (2007).

Definition 4 (Linear smoother)**.** *An estimator* $\hat{R}(x, x')$ *of* $m_0(x, x') - m(x, x'; \boldsymbol{\beta}_0)$ *is a linear smoother if, for each couple* (x, x')*, there exists an* $|\mathcal{I}_n|$*-vector* $\boldsymbol{L}(x, x')$*, with elements* $L(x, x'; X_k, X_l)$*,* $(k, l) \in \mathcal{I}_n$*, such that*

$$\hat{R}(x, x') = \boldsymbol{L}(x, x')^T \boldsymbol{R}(\hat{\boldsymbol{\beta}}).$$

Instead of a local constant smoother (5.7), which suffers from design and boundary bias, local linear regression may be preferred (Fan and Gijbels, 1996; Wasserman, 2007). This is, however, beyond the scope of this dissertation.

We focus on the probability $\mathrm{P}\left(Y \preccurlyeq Y' \mid X, X' = X + \Delta\right)$ and for assessing model adequacy we plot the smoothed residuals $\hat{R}(x, x' = x + \Delta)$ as a function of x. These residuals provide information on the bias of the working model and they are bounded in $[-1, 1]$. Figure 5.5 shows such a plot, based on random samples of size $n = 150$ for the 3 settings described in the left panel of Figure 5.3 with $\Delta = 1$. The left panel of Figure 5.5 corresponds to the setting under H_0 and the residuals are close to 0. For a weak quadratic effect, the middle panel indicates that the fitted model gives biased probabilistic index estimators. For $X < -1$ the probability is underestimated, while for $X > 1$ it is overestimated. The right panel shows a strong quadratic effect for which similar conclusions hold. For each figure we also show the pointwise 95% confidence intervals. However, there is a multiplicity problem, as n confidence intervals are calculated simultaneously. Therefore, these intervals are only indicative, but they may be helpful in interpreting the graphical GOF tool.

For formal hypothesis testing we construct a single quadratic form of the smoothed residuals. The quadratic form is simplistic and it does not use all smoothed residuals. Other GOF statistics that use all residuals to form a Cramér–von Mises, Anderson–Darling, or Kolmogorov–Smirnov type of test, will very likely outperform our test. However, extending these techniques to the PIM framework is challenging because the pseudo-observations are sparsely correlated. In Appendix 5.A we provide more details on these challenges and we briefly sketch how they can be tackled; a more detailed study, however, falls beyond the scope of this dissertation.

Theorem 13. *Consider a fixed finite number of points, say* $x_1, \ldots, x_m$*, within the range of* X*, with* $m < n$*. Let* $\boldsymbol{R}_\Delta$ *denote the* m*-vector of residuals* $\hat{R}(x_i, x_i + \Delta)$ *and define the quadratic form*

$$S_\Delta := \boldsymbol{R}_\Delta^T \mathrm{Cov}\left(\boldsymbol{R}_\Delta\right)^{-1} \boldsymbol{R}_\Delta, \tag{5.14}$$

with asymptotic covariance

$$\mathrm{Cov}\left(\boldsymbol{R}_\Delta\right) = \boldsymbol{K}_\Delta(\boldsymbol{I}-\boldsymbol{H})\mathrm{Cov}\left(\boldsymbol{I}_p\right)\left(\boldsymbol{I}-\boldsymbol{H}\right)^T\boldsymbol{K}_\Delta^T$$

where $\boldsymbol{K}_\Delta$ *denotes the* $(m\times|\mathcal{I}_n|)$*-matrix of weights* $K(x_i, x_i+\Delta; X_k, X_l)/\sum_{(k,l)\in\mathcal{I}_n} K(x_i, x_i+\Delta; X_k, X_l)$ *and* $\boldsymbol{H}$ *as defined by (5.12). For a fixed finite* m*, under* H_0*, as* $h\to 0$ *and* $nh\to\infty$*,*

$$S_\Delta \xrightarrow{d} \chi^2_m. \tag{5.15}$$

Proof. Similar as in the proof of Theorem 12, the first order approximation of $\boldsymbol{R}(\hat{\boldsymbol{\beta}})$ leads to

$$\boldsymbol{R}_\Delta = \boldsymbol{K}_\Delta(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{R}(\boldsymbol{\beta}_0) + o_p(n^{-1/2}).$$

Consequently, under H_0 (5.3), the asymptotic expectation and variance are given by

$$\mathrm{E}\left(\boldsymbol{R}_\Delta\right) = \boldsymbol{0}, \quad \mathrm{Cov}\left(\boldsymbol{R}_\Delta\right) = \boldsymbol{K}_\Delta(\boldsymbol{I}-\boldsymbol{H})\mathrm{Cov}\left(\boldsymbol{I}_p\right)\left(\boldsymbol{I}-\boldsymbol{H}\right)^T\boldsymbol{K}_\Delta^T.$$

By using the Cramér–Wold device and the central limit theorem of Lumley and Mayer-Hamblett (2003, p. 13) it follows that, under H_0,

$$\boldsymbol{R}_\Delta \xrightarrow{d} \mathrm{N}[\boldsymbol{0}, \mathrm{Cov}\left(\boldsymbol{R}_\Delta\right)],$$

so that the quadratic from (5.15), asymptotically, follows a chi-squared distribution with m degrees of freedom.

□

A consistent estimator of $\mathrm{Cov}\left(\boldsymbol{R}_\Delta\right)$ can be obtained by replacing $\mathrm{Cov}\left(\boldsymbol{I}_p\right)$ with $\hat{\boldsymbol{\Sigma}}_p$ (5.13) and $\boldsymbol{\beta}_0$ by $\hat{\boldsymbol{\beta}}$ in $\boldsymbol{H}$. The quadratic form S_Δ takes the estimated correlations between the residuals $\hat{R}(x_i, x_i+\Delta)$ and $\hat{R}(x_j, x_j+\Delta)$ into account. In total $m(m-1)/2$ correlations need to be estimated. For finite sample approximations, when m is large relative to the sample size n, the estimated covariance matrix $\hat{\mathrm{Cov}}(\boldsymbol{R}_\Delta)$ is not guaranteed to be positive definite. Therefore m should be chosen small relatively to the sample size n and the design points $x_1, \ldots, x_m$ should cover the whole range of X so as to increase the likelihood of detecting departures from the underlying model.

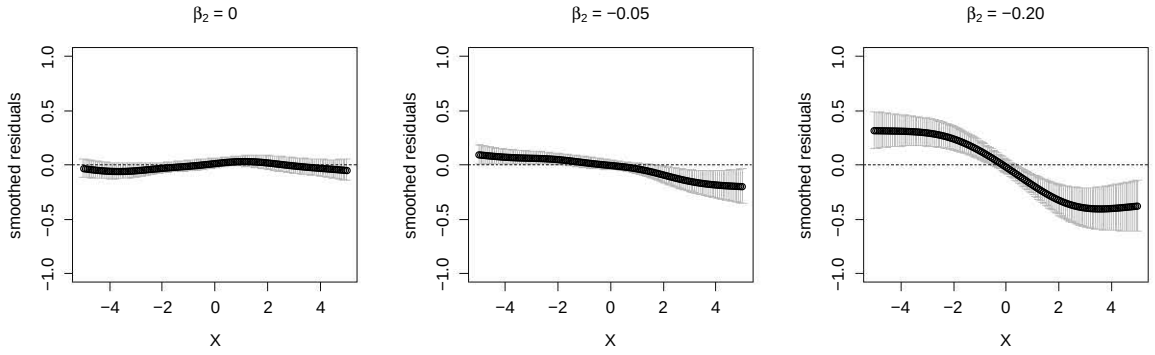

Figure 5.5: Smoothed residuals $\hat{R}(x, x+\Delta)$ as a function of x according to the different settings of the left panel of Figure 5.3 with $\Delta = 1$, for a random sample of size $n = 150$, and Gaussian kernel with $h_1 = h_2 = 1.5$. The left panel corresponds to no quadratic effect, the middle panel to a medium quadratic effect, and the right panel to a strong quadratic effect. The black dots are the smoothed residuals, and the grey bars indicate pointwise 95% confidence intervals.

5.2.3 Multiple predictors

The methods can be extended to multiple predictors, say $\boldsymbol{X}^T = (X_1, \ldots, X_d)$, by considering multivariate kernels, e.g.

$$K_{\boldsymbol{h}_1,\boldsymbol{h}_2}(\boldsymbol{x}, \boldsymbol{x}'; \boldsymbol{X}, \boldsymbol{X}') = \prod_{i=1}^{d} K_{h_{1i},h_{2i}}(x_i, x_i'; X_i, X_i'), \tag{5.16}$$

where $\boldsymbol{h}_i^T = (h_{i1}, \ldots, h_{id})$. For high-dimensional data, however, smoothers based on a multiplicative kernel are not always useful in practice due to the curse of dimensionality and the computational burden. Therefore, nonparametric smoothers can be restricted to, for example, additive models.

5.2.4 Automatic bandwidth selection

It is known that the choice of bandwidth is often more important than the choice of kernel (Wasserman, 2007). Bandwidths may be selected in a data-driven fashion by using, for example, cross-validation (CV). The properties of the leave-one-out CV for independent outcomes has been examined by many authors; see for example Wong (1983). This CV can result in poor bandwidths if outcomes are dependent. Several modifications have been proposed; see, for example, Chu and Marron (1991). We propose a modification of the leave-one-out CV score,

accounting for the sparse correlation of the pseudo-observations.

We first introduce some definitions, of which most are based on chapters 4 and 5 of Wasserman (2007). Denote the true error as

$$R_0(x, x') := m_0(x, x') - m(x, x'; \boldsymbol{\beta}_0). \tag{5.17}$$

The mean squared error (MSE) associated with the smoothing residuals is given by

$$\text{MSE}(h_1, h_2) = \text{E}\left\{ |\mathcal{I}_n|^{-1} \sum_{(i,j)\in\mathcal{I}_n} \left[\hat{R}(X_i, X_j) - R_0(X_i, X_j)\right]^2 \right\}, \tag{5.18}$$

where $\hat{R}(X_i, X_j)$ depends on the bandwidths h_1 and h_2. The optimal bandwidths, say (h_1^*, h_2^*), can, for example, be defined as the minimizer of the MSE, i.e.

$$(h_1^*, h_2^*) = \text{argmin}_{(h_1,h_2)\in\mathbb{R}_+^2} \text{MSE}(h_1, h_2).$$

However, since R_0 is unknown, this selection criterion cannot be used in practice. A intuitive solution consists of replacing $R_0(X_i, X_j)$ in (5.18) by the residual $R_{ij}(\hat{\boldsymbol{\beta}})$. This will often lead to undersmoothing because $R_{ij}(\hat{\boldsymbol{\beta}})$ is already used for obtaining $\hat{R}(X_i, X_j)$. Let $\mathcal{I}_n^{-(i,j)}$ denote the subset of $\mathcal{I}_n$ for which all elements with subscript i or j are removed, i.e.

$$\mathcal{I}_n^{-(i,j)} := \{(k, l) \mid (k, l) \in \mathcal{I}_n \wedge (i, j) \cap (k, l) = \emptyset\},$$

and

$$\hat{R}_{-(i,j)}(x, x') := \frac{\sum_{(k,l)\in\mathcal{I}_n^{-(i,j)}} R_{kl}(\hat{\boldsymbol{\beta}}) K(x, x'; X_k, X_l)}{\sum_{(k,l)\in\mathcal{I}_n^{-(i,j)}} K(x, x'; X_k, X_l)}.$$

Thus $\hat{R}_{-(i,j)}(x, x')$ corresponds to the smoothed residual obtained by omitting all residuals containing (Y_i, X_i) or (Y_j, X_j). This leads us to defining an adjusted leave-one-out cross validation score.

Definition 5 (Adjusted leave-one-out cross validation score)**.** *The adjusted leave-one-out cross validation score is defined by*

$$CV(h_1, h_2) = |\mathcal{I}_n|^{-1} \sum_{(i,j)\in\mathcal{I}_n} \left[R_{ij}(\hat{\boldsymbol{\beta}}) - \hat{R}_{-(i,j)}(X_i, X_j)\right]^2. \tag{5.19}$$

The following lemma will be useful to relate the adjusted leave-one-out cross validation score to the MSE.

Lemma 10. *It holds that*

$$\begin{aligned}\mathrm{E}\left\{\left[R_{ij}(\boldsymbol{\beta}_0)-\hat{R}_{-(i,j)}(X_i,X_j)\right]^2\right\} &= \mathrm{E}\left\{[R_{ij}(\boldsymbol{\beta}_0)-R_0(X_i,X_j)]^2\right\}+\\ &\quad \mathrm{E}\left\{\left[R_0(X_i,X_j)-\hat{R}_{-(i,j)}(X_i,X_j)\right]^2\right\}. \quad (5.20)\end{aligned}$$

Proof. If $\boldsymbol{X}$ denotes the vector of predictors X_i ($i=1,\ldots,n$), it follows that

$$\begin{aligned}&\mathrm{E}\left\{[R_{ij}(\boldsymbol{\beta}_0)-R_0(X_i,X_j)]\left[R_0(X_i,X_j)-\hat{R}_{-(i,j)}(X_i,X_j)\right]\right\}\\ =\;&\mathrm{E}\left[\mathrm{E}\left(R_{ij}(\boldsymbol{\beta}_0)-R_0(X_i,X_j)\mid\boldsymbol{X}\right)\mathrm{E}\left(R_0(X_i,X_j)-\hat{R}_{-(i,j)}(X_i,X_j)\mid\boldsymbol{X}\right)\right]\\ =\;&0\times\mathrm{E}\left[R_0(X_i,X_j)-\hat{R}_{-(i,j)}(X_i,X_j)\right]=0. \quad (5.21)\end{aligned}$$

Consequently,

$$\begin{aligned}&\mathrm{E}\left\{\left[R_{ij}(\boldsymbol{\beta}_0)-\hat{R}_{-(i,j)}(X_i,X_j)\right]^2\right\}\\ =\;&\mathrm{E}\left\{\left[R_{ij}(\boldsymbol{\beta}_0)-R_0(X_i,X_j)+R_0(X_i,X_j)-\hat{R}_{-(i,j)}(X_i,X_j)\right]^2\right\}\\ =\;&\mathrm{E}\left\{[R_{ij}(\boldsymbol{\beta}_0)-R_0(X_i,X_j)]^2+\left[R_0(X_i,X_j)-\hat{R}_{-(i,j)}(X_i,X_j)\right]^2+\right.\\ &\left.2\left[R_{ij}(\boldsymbol{\beta}_0)-R_0(X_i,X_j)\right]\left[R_0(X_i,X_j)-\hat{R}_{-(i,j)}(X_i,X_j)\right]\right\}\\ =\;&\mathrm{E}\left\{[R_{ij}(\boldsymbol{\beta}_0)-R_0(X_i,X_j)]^2\right\}+\mathrm{E}\left\{\left[R_0(X_i,X_j)-\hat{R}_{-(i,j)}(X_i,X_j)\right]^2\right\}.\end{aligned}$$

□

If we define

$$\sigma^2:=\mathrm{E}\left\{|\mathcal{I}_n|^{-1}\sum_{(i,j)\in\mathcal{I}_n}\left[R_{ij}(\hat{\boldsymbol{\beta}})-R_0(X_i,X_j)\right]^2\right\},$$

then, upon using Lemma 10, it follows that

$$\mathrm{E}[\mathrm{CV}(h_1,h_2)]\approx\sigma^2+\mathrm{MSE}(h_1,h_2), \quad (5.22)$$

relating the adjusted leave-one-out cross validation score to the MSE. Note that (5.22) is an approximation rather than an equality, since we substituted $\boldsymbol{\beta}_0$ in (5.20) by the plug-in estimator $\hat{\boldsymbol{\beta}}$ and we used the approximation $\mathrm{E}[\hat{R}_{-(i,j)}(X_i,X_j)]\approx\mathrm{E}[\hat{R}(X_i,X_j)]$.

A data-driven choice of bandwidth can therefore be obtained by choosing (h_1,h_2) which minimizes $\mathrm{CV}(h_1,h_2)$ (5.19). However, because $|\mathcal{I}_n|=O(n^2)$, we often restrict the sum in (5.19)

to a subset of $\mathcal{I}_{\text{sub}} \subset \mathcal{I}_n$ to reduce computation time. This subset will often be chosen such that for $(i, j) \in \mathcal{I}_{\text{sub}}$ it holds that $X_j - X_i \approx \Delta$.

Note that the proposed modified cross validation score is merely a first step in constructing an automatic bandwidth selection procedure. More specifically, the current approach ignores the change in distributional properties of S_Δ due to the automatic bandwidth selection. Therefore it is anticipated that the GOF test will be liberal. In Appendix 5.B we give more details on the challenges for obtaining the appropriate null distribution when the bandwidth is selected automatically.

5.3 Simulation study

The theoretical properties of S_Δ (5.15) are empirically evaluated for single and multiple predictors by means of simulations. Since the test has several tunable parameters, the effect of the choice of bandwidth and the effect of Δ on the size and power of the test are examined. The properties of the test with automatic bandwidth selection are also briefly examined.

All data-generating models are normal linear models associated with a probit PIM. However, similar conclusions hold for the exponential models associated with a logit PIM (results not shown).

5.3.1 A single predictor

Empirical sizes

To examine the empirical null distribution of S_Δ we generate data with the simple linear model

$$Y = \alpha X + \varepsilon, \quad \varepsilon \overset{d}{=} \mathrm{N}(0, \sigma^2), \tag{5.23}$$

which embeds the PIM

$$\mathrm{P}\left(Y \preccurlyeq Y' \mid X, X'\right) = m_0(X, X') = \Phi\left[\beta(X' - X)\right], \quad \beta = \frac{\alpha}{\sqrt{2\sigma^2}}. \tag{5.24}$$

The predictor X takes n equidistant values in $[-5, 5]$ and the following parameters are fixed: $\alpha = 0.9\sqrt{2}$ and $\sigma^2 = 9$. Based on 1000 Monte Carlo simulation runs, the empirical type I error

rates are calculated for the nominal significance levels of 1%, 5%, and 10%. The asymptotic chi-squared distribution (5.15) is used for p-value calculation.

The null distribution is examined for different values of Δ, sample size n, and bandwidths h_1 and h_2, which we restrict to $h_1 = h_2$ and which is further denoted by h. The statistic is based on three design points: $x_1 = -3$, $x_2 = 0$, and $x_3 = 3$ with Gaussian kernel. Table 5.1 shows the results. For a sample size $n = 100$ and a small bandwidth $h = 0.5$ the test is highly conservative, while for a large bandwidth $h = 2.5$ it is highly liberal. Best results are obtained for an intermediate bandwidth $h = 1.5$.

For $n = 250$ and $h = 0.5$ the test is too conservative for $\Delta = 1$ and slightly less conservative for $\Delta = 2$. With $h = 1.5$ the test has approximately a correct size for all Δ, while for $h = 2.5$ the test remains too liberal.

For a sample size $n = 500$ and $h = 0.5$ the test is conservative for $\Delta = 1$ and has approximately a correct size for $\Delta = 2$. For $h = 1.5$ the test has approximately a correct size, while for $h = 2.5$ the test remains liberal.

In conclusion, best results are obtained for a bandwidth of $h = 1.5$, while the choice of Δ is less important. However, for a bandwidth of $h = 0.5$ the test is conservative, while for a bandwidth of $h = 1.5$ there is an inflation of the type I error.

Table 5.1: Empirical type I error rates (%) at the 1%, 5%, and 10% levels of significance based on 1000 Monte-Carlo simulations for model (5.24)

h	Δ	$n = 100$			$n = 250$			$n = 500$		
		1%	5%	10%	1%	5%	10%	1%	5%	10%
0.5	1	0.0	0.7	3.8	0.2	3.1	7.3	0.3	3.5	7.7
0.5	2	0.0	1.7	5.1	0.3	3.2	9.0	0.5	4.9	9.9
1.5	1	0.5	4.4	8.8	0.4	5.1	9.5	1.2	4.4	11.1
1.5	2	0.3	3.6	9.3	0.6	4.7	11.2	1.2	5.8	11.7
2.5	1	3.4	9.6	15.4	2.6	8.0	14.1	2.3	7.7	13.4
2.5	2	2.3	7.4	14.4	1.9	7.8	13.8	1.8	7.5	13.0

Empirical powers

To study the power properties, we generate data according to

$$Y = \alpha_1 X + \alpha_2 f(X) + \varepsilon, \quad \varepsilon \overset{d}{=} \mathrm{N}(0, \sigma^2). \tag{5.25}$$

We fix $\alpha_1 = 0.9\sqrt{2}$ and $\sigma^2 = 9$ and consider three cases.

- A *quadratic model* with $f(X) = X^2$ and $\alpha_2 = -0.05\sqrt{2}$ or $\alpha_2 = -0.125\sqrt{2}$.
- A *sine model* with $f(X) = \sin(X)$ and $\alpha_2 = -0.6\sqrt{2}$ or $\alpha_2 = -1.2\sqrt{2}$.
- An *exponential model* with $f(X) = \exp(X)$ and $\alpha_2 = 0.02\sqrt{2}$ or $\alpha_2 = 0.04\sqrt{2}$.

The parameter values are chosen so that most empirical powers are bounded away from the trivial powers of 5% and 100%. The PIM corresponding to model (5.25) is given by

$$\mathrm{P}\left(Y \preccurlyeq Y' \mid X, X'\right) = m_0(X, X') = \Phi\left\{\beta_1(X' - X) + \beta_2[f(X') - f(X)]\right\}, \tag{5.26}$$

where $\beta_i = \alpha_i / \sqrt{2\sigma^2}$, $i = 1, 2$. We analyze the data with the incorrect working model

$$m(X, X'; \beta) = \Phi[\beta(X' - X)].$$

Figure 5.6 shows the probabilities $\mathrm{P}\left(Y \preccurlyeq Y' \mid X, X' = X + 1\right)$ associated with PIM (5.26) as a function of X for the three models and for different β_2 values.

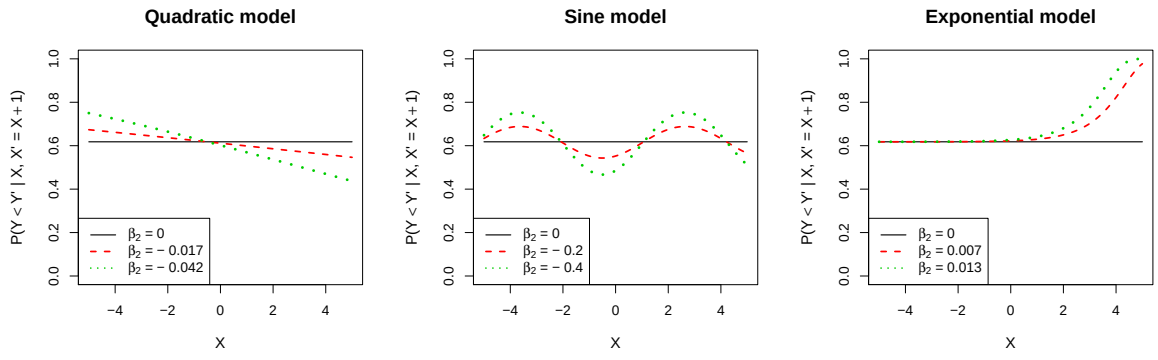

Figure 5.6: Conditional PI for $X' = X + 1$ for different values of β_2 for the quadratic, sine, and exponential versions of model (5.26)

The results in Table 5.1 suggest that empirical sizes are best controlled for a medium bandwidth. Therefore we restrict the power study to $h_1 = h_2 = 1.5$ in a Gaussian kernel with design points $x_1 = -3$, $x_2 = 0$, and $x_3 = 3$.

Table 5.2 gives the empirical rejection rates at the 5% level of significance based on 1000 Monte Carlo simulations for the different data-generating models. The test succeeds in detecting lack-of-fit (LOF). Under the conditions of the simulation study, for the quadratic and sine model, highest powers are obtained with $\Delta = 1$ while for the exponential model this is $\Delta = 2$.

Table 5.2: Empirical powers (%) at the 5% level of significance for model (5.26) based on 1000 Monte-Carlo simulations

β_2	$n = 100$		$n = 250$		$n = 500$	
	$\Delta = 1$	$\Delta = 2$	$\Delta = 1$	$\Delta = 2$	$\Delta = 1$	$\Delta = 2$
	quadratic model					
-0.017	12.0	11.0	42.1	40.7	78.2	75.5
-0.042	73.2	68.9	99.8	99.5	100.0	100.0
	sine model					
-0.2	14.6	8.7	53.6	36.3	89.2	70.5
-0.4	64.9	39.6	99.7	94.9	100.0	100.0
	exponential model					
0.007	14.0	14.2	49.6	57.2	82.4	89.7
0.013	38.1	42.1	96.9	98.6	100.0	100.0

5.3.2 Multiple predictors

Empirical sizes

Consider the data-generating model

$$Y = \alpha_1 X_1 + \alpha_2 X_2 + \varepsilon, \quad \varepsilon \stackrel{d}{=} \mathrm{N}(0, \sigma^2),$$

with embedded PIM

$$\mathrm{P}\left(Y \preccurlyeq Y' \mid \boldsymbol{X}, \boldsymbol{X}'\right) = m_0(\boldsymbol{X}, \boldsymbol{X}') = \Phi\left[\beta_1(X_1' - X_1) + \beta_2(X_2' - X_2)\right], \tag{5.27}$$

where $\boldsymbol{X} = (X_1, X_2)$ and $\beta_i = \alpha_i/\sqrt{2\sigma^2}$, $i = 1, 2$. The following parameters are fixed: $\alpha_1 = \alpha_2 = 1$ and $\sigma^2 = 9$, corresponding to $\beta_1 = \beta_2 = 0.24$. The predictor X_1 takes n equidistant values in the interval $[-5, 5]$, while $X_2 \stackrel{d}{=} \mathrm{N}(0, 4)$.

The statistic is based on three design points: $(x_{11}, x_{21}) = (-3, -2.5)$, $(x_{12}, x_{22}) = (0, 0)$, and $(x_{13}, x_{23}) = (3, 2.5)$, with Gaussian kernel and bandwidths $\boldsymbol{h}_1 = \boldsymbol{h}_2 = (1.5, 1.5)$, and different values for $\boldsymbol{\Delta}$ are considered.

Based on 1000 Monte Carlo simulation runs, the empirical rejection rates are calculated for the significance levels of 1%, 5%, and 10% and for different sample sizes n.

The results are presented in Table 5.3. For a sample size $n = 100$ the test is highly conservative, while it becomes less conservative when the sample size increases. For $n = 500$ the test has approximately a correct size for all choices of $\boldsymbol{\Delta}$.

Table 5.3: Empirical type I error rates (%) at the 1%, 5%, and 10% levels of significance for the model (5.27) based on 1000 Monte-Carlo simulations

$\boldsymbol{\Delta}$	$n = 100$			$n = 250$			$n = 500$		
	1%	5%	10%	1%	5%	10%	1%	5%	10%
$(1, 1)$	0.0	0.9	3.4	0.3	2.4	6.1	0.9	4.7	9.0
$(1, 2)$	0.0	1.1	4.5	0.2	2.3	6.9	1.1	4.3	10.2
$(2, 1)$	0.0	1.1	4.3	0.3	3.1	6.6	0.8	3.9	9.3
$(2, 2)$	0.0	0.7	5.3	0.3	3.6	6.8	0.9	3.4	10.8

Empirical powers

Consider the data-generating model with interaction

$$Y = \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_1 X_2 + \varepsilon, \quad \varepsilon \stackrel{d}{=} \mathrm{N}(0, \sigma^2). \tag{5.28}$$

We fix $\alpha_1 = \alpha_2 = 1$ and $\sigma^2 = 9$ and consider different values of α_3. Let $\boldsymbol{X} = (X_1, X_2)$, then the corresponding PIM is

$$\begin{aligned} \mathrm{P}\left(Y \preccurlyeq Y' \mid \boldsymbol{X}, \boldsymbol{X}'\right) &= m_0(\boldsymbol{X}, \boldsymbol{X}') \\ &= \Phi\left[\beta_1(X_1' - X_1) + \beta_2(X_2' - X_2) + \beta_3(X_1' X_2' - X_1 X_2)\right], \end{aligned} \tag{5.29}$$

with $\beta_i = \alpha_i/\sqrt{2\sigma^2}$. The data are analyzed with the incorrect working model

$$m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\gamma}) = \Phi\left[\gamma_1(X_1' - X_1) + \gamma_2(X_2' - X_2)\right]. \tag{5.30}$$

Figure 5.7 plots $\mathrm{P}\left(Y \preccurlyeq Y' \mid X_1' = X_1 + \Delta_1, X_2' = X_2 + \Delta_2\right)$ as a function of the sum $\Delta_2 X_1 + \Delta_1 X_2$ when $\boldsymbol{\Delta} = (1, 1)$ and for different values of β_3.

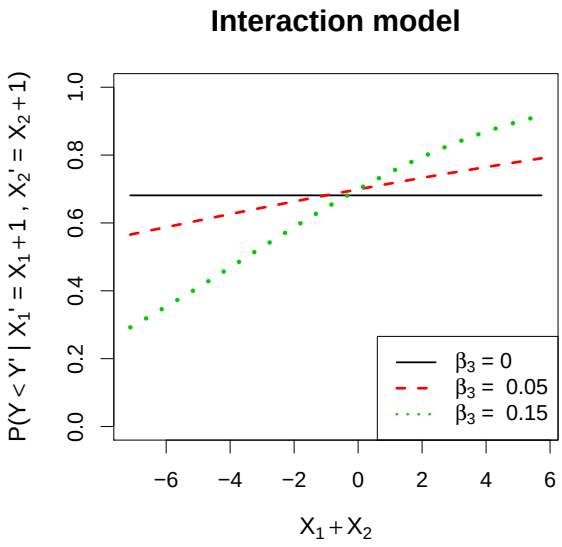

Figure 5.7: $\mathrm{P}\left(Y \preccurlyeq Y' \mid X_1' = X_1 + 1, X_2' = X_2 + 1\right)$ of model (5.29) as a function of $X_1 + X_2$ for different values of β_3

Table 5.4 gives the empirical rejection rates at the 5% significance level, based on 1000 Monte Carlo simulation runs. The statistic is based on three design points: $(x_{11}, x_{21}) = (-3, -2.5)$, $(x_{12}, x_{22}) = (0, 0)$, and $(x_{13}, x_{23}) = (3, 2.5)$, with Gaussian kernel and bandwidth $\boldsymbol{h}_1 = \boldsymbol{h}_2 = (1.5, 1.5)$.

The test succeeds in detecting an omitted interaction and under the conditions of the simulation study highest powers are obtained for $\boldsymbol{\Delta} = (1, 2)$ or $\boldsymbol{\Delta} = (2, 2)$.

5.3.3 Misspecified link function

We examine the power of detecting a misspecified link function by simulating data according to (5.23) while analyzing the data with the working model

$$m(X, X'; \gamma) = \mathrm{expit}[\gamma(X' - X)], \tag{5.31}$$

Table 5.4: Empirical powers (%) at the 5% level of significance for model (5.29) based on 1000 Monte-Carlo simulations

$\boldsymbol{\Delta}$	$(1,1)$	$(1,2)$	$(2,1)$	$(2,2)$	$(1,1)$	$(1,2)$	$(2,1)$	$(2,2)$	$(1,1)$	$(1,2)$	$(2,1)$	$(2,2)$
β_3	$n = 100$				$n = 250$				$n = 500$			
0.05	1.7	2.8	1.6	2.9	28.4	52.4	21.7	44.8	58.3	83.2	54.4	82.6
0.15	42.9	71.6	40.1	75.8	100.0	100.0	99.8	100.0	100.0	100.0	100.0	100.0

i.e. the link function is the only difference between the true and working model.

Table 5.5 shows the empirical powers of the test with $h_1 = h_2 = 1.5$, design points $x_1 = -3$, $x_2 = 0$, $x_3 = 3$ with Gaussian kernel and $\Delta = 1$ or $\Delta = 2$. The test has low power for $n = 100$ and $n = 250$ and a moderate power for $n = 500$. Best results are obtained for $\Delta = 1$.

Table 5.5: Empirical powers (%) at the 5% level of significance for model (5.31) based on 1000 Monte-Carlo simulations

$n = 100$		$n = 250$		$n = 500$	
$\Delta = 1$	$\Delta = 2$	$\Delta = 1$	$\Delta = 2$	$\Delta = 1$	$\Delta = 2$
16.8	12.4	31.9	27.6	60.0	52.1

5.3.4 Automatic bandwidth selection

To study the null distribution of S_Δ when the bandwidth is selected based on the modified cross-validation score (5.19), we reconsider the simulation set-up from Section 5.3.1 with $\Delta = 1$. We restrict the sum in (5.19) to the subset $\mathcal{I}_{\text{sub}} = \{(i,j) \mid \Delta - 0.05 < X_j - X_i < \Delta + 0.05\}$. For $n = 250$ and $n = 500$ the sum is even restricted to a random sample of size 100 from $\mathcal{I}_{\text{sub}}$. The candidate set of bandwidths is restricted to $\{0.5, 1.5, 2.5\}$ with $h_1 = h_2$. The chi-squared distribution with 3 degrees of freedom is used for p-value calculation. As mentioned in Section 5.2.4, this null distribution ignores the change in distributional properties of S_Δ due to the data-driven selection of the bandwidth.

Table 5.6 gives the empirical type I error rates. For all sample sizes the test is liberal. This

is expected since the selection of the bandwidth is not accounted for in the null distribution. As compared to Table 5.1 the results are slightly better with $h_1 = h_2 = 1.5$ and worse with $h_1 = h_2 = 0.5$ or 2.5. For $n = 500$ the empirical rejection rates are close to their nominal levels for 1% and 5% but too liberal for 10%.

To examine the empirical powers, we reconsider the quadratic model from Section 5.3.1 with $\Delta = 1$. Note that for $n = 100$ and $n = 250$ these powers can perhaps be too optimistic since the type I error is not correctly controlled. The automatic cross-validation results in a power loss as compared to Table 5.2.

Table 5.6: Empirical type I error (%) at the 1%, 5%, and 10% levels of significance and empirical powers (%) at the 5% level of significance when the bandwidth is automatically selected with the modified cross-validation score. All results are based on 1000 Monte-Carlo simulations.

$n = 100$			$n = 250$			$n = 500$			β_2	$n = 100$	$n = 250$	$n = 500$
Empirical type I error									Empirical power quadratic model			
1%	5%	10%	1%	5%	10%	1%	5%	10%	-0.017	14.5	37.5	67.9
2.9	7.4	14.1	2.1	5.5	12.0	0.9	5.8	12.6	-0.042	68.4	87.6	96.1

5.3.5 Assessing goodness-of-fit with a graphical tool

In Figure 5.8 we show the GOF plots for three simulated datasets with sample size $n = 150$ for the quadratic, sine, and exponential model of Section 5.3.1. The Gaussian kernel is used with $h = 1.5$ and $\Delta = 1$. The GOF plots show similar shapes as Figure 5.6, indicating that GOF plots are informative on how the true model differs from the working model.

Figure 5.9 shows the plots for a random sample generated by the interaction model (5.29) with $\beta_3 = 0.15$ and analyzed with PIM (5.30), as well as plots for random sample generated by (5.23) and analyzed by the PIM with incorrect link function (5.31). Similarly, the plots succeeds in indicating LOF.

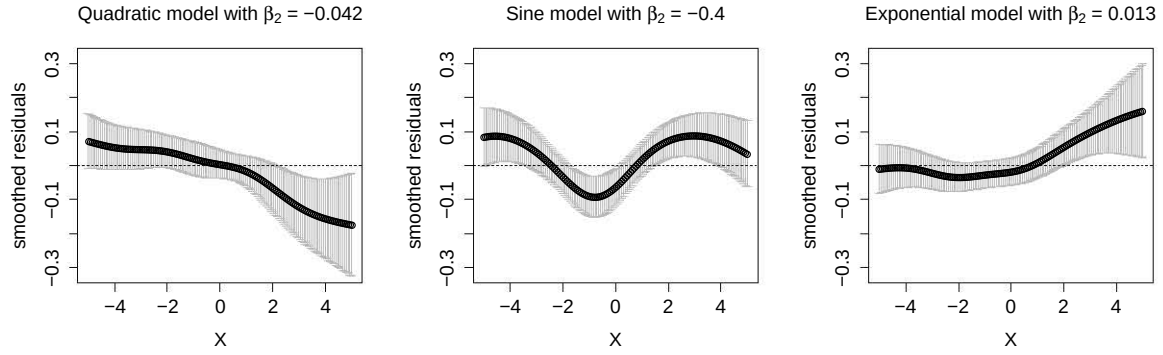

Figure 5.8: GOF plots for the quadratic (left), sine (middle), and exponential (right) models (5.26) for a random sample of size $n = 150$. The grey bars indicate the pointwise 95% confidence intervals.

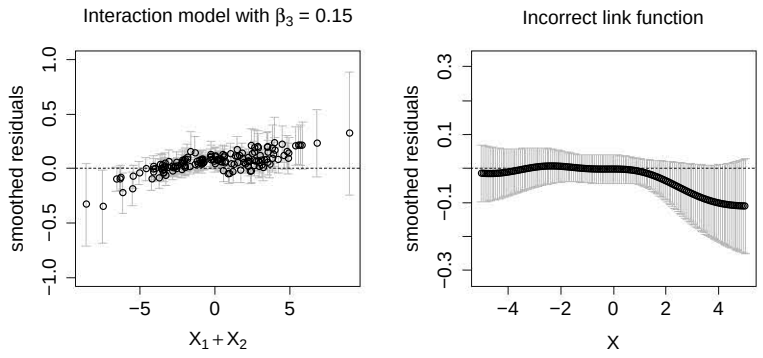

Figure 5.9: GOF plots for the interaction model (5.29) with $\beta_3 = 0.15$ (left) and the model with logit link (5.31) (right) for a random sample of size $n = 150$. The grey bars indicate the pointwise 95% confidence intervals.

5.4 Examples revisited

In this section we evaluate the GOF of the fitted PIM of Section 5.1 as well as of the fitted PIMs of Sections 2.5 and 3.2.3. All smoothed residuals are constructed with a Gaussian kernel.

5.4.1 The childhood respiratory disease study

We return to the CRDS example of Section 5.1. In model (5.2) the effect of the smoking status on the pulmonary function of a child is not significant. Smoothed residuals are constructed with $\boldsymbol{\Delta} = (1, 1)$ and bandwidths $\boldsymbol{h}_1 = \boldsymbol{h}_2 = (h, 0)$ for several values of h, for which the optimal bandwidth was selected based on the cross-validation score (5.19) with the sum restricted to $\mathcal{I}_{\text{sub}} = \{(i, j) \mid \text{AGE}_j - \text{AGE}_i = 1\}$ for computational reasons. Since the binary predictor SMOKE has sufficient replicates, smoothing is unnecessary. The left panel of Figure 5.10 shows the CV-plot where the cross-validation score is plotted as a function of the bandwidth. A minimum is attained for $h = 3.5$.

Since most (89%) of the smoking children are between 10 and 16 years old, we restrict the GOF assessment to that age class. The middle panel of Figure 5.10 plots the smoothed residuals as a function of age (AGE) for model (5.2). The plot suggest that the probability $\mathrm{P}(\text{FEV} \preccurlyeq \text{FEV}' \mid \text{SMOKE} = 0, \text{SMOKE}' = 1, \text{AGE}' = \text{AGE} + 1)$ is underestimated for younger children, while it is overestimated for the older. The GOF test confirms this: $S_\Delta = 15.7$ and $p = 0.016$. Both the plot and the test thus indicate that PIM (5.2) is inappropriate and the plot suggests that an interaction needs to be included. Therefore we fit an interaction model which takes this into account

$$\begin{aligned} \operatorname{logit}[\mathrm{P}\left(\text{FEV} \preccurlyeq \text{FEV}'\right)] \;=\; & \beta_1(\text{AGE}' - \text{AGE}) + \beta_2(\text{SMOKE}' - \text{SMOKE}) + \\ & \beta_3(\text{AGE}' * \text{SMOKE}' - \text{AGE} * \text{SMOKE}), \end{aligned} \tag{5.32}$$

with estimates $\hat{\beta}_1 = 0.61$ (SE : 0.03, $p < 0.0001$), $\hat{\beta}_2 = 5.3$ (SE : 1.04, $p < 0.0001$), and $\hat{\beta}_3 = -0.46$ (SE : 0.08, $p < 0.0001$). All effects are now highly significant. The right panel of Figure 5.10 gives the GOF plot with $h = 3.5$ and the GOF test indicated no significant evidence for LOF at the 5% level of significance: $S_\Delta = 11.6$ and $p = 0.072$. However, since the p-value is close to the significance level and the plot shows no LOF, the model can perhaps be improved by including additional predictors in addition to the age and smoking behaviour. After including

additional predictors, model adequacy should be reassessed.

The estimated effect of smoking based on PIM (5.32), in terms of the probabilistic index, is given by

$$\text{logit}\left[\hat{\text{P}}\left(\text{FEV} \preccurlyeq \text{FEV}' \mid \text{SMOKE} = 0, \text{SMOKE}' = 1, \text{AGE} = \text{AGE}'\right)\right] = 5.3 - 0.46\text{AGE}.$$

The probability for having a better pulmonary function for the smoking child decreases with increasing age. The bottom right panel of Figure 5.10 shows this probability as a function of AGE. At the age of 10, for example, the estimated probability is 68% with confidence interval $[53\%, 80\%]$. This probability indicates that the lung function is better for smoking children, which seems unreasonable. However, children who smoke at the age of ten are likely to have only just started smoking and the smoking did not affect the lungs yet. By the age of 16 this probability decreased to 12%, indicating it is highly unlikely that a smoking child has a better lung function, suggesting an adverse effect of smoking; the confidence interval for this probability is $[7\%, 21\%]$.

5.4.2 The mental health study

In Section 2.5.2 a PIM with main effects was fitted to the mental health study example

$$\text{logit}\left[\text{P}\left(\text{MI} \preccurlyeq \text{MI}'\right)\right] = \beta_1(\text{SES}' - \text{SES}) + \beta_2(\text{LI}' - \text{LI}). \tag{5.33}$$

Smoothed residuals are constructed with $\boldsymbol{\Delta} = (1, 1)$ and bandwidths $\boldsymbol{h}_1 = \boldsymbol{h}_2 = (0, h)$ with $h \in \{0.5, 1, 2, 3, 4, 5\}$, for which the optimal bandwidth is selected based on the cross-validation score (5.19). The binary predictor SES has sufficient replicates, and smoothing is unnecessary. The left panel of Figure 5.11 shows the score as a function of the bandwidth, where a minimum is attained for $h = 2$. The right panel plots the smoothed residuals as a function of the life index (LI) for model (5.33). There is no convincing evidence of LOF. The corresponding GOF test confirms this: $S_\Delta = 3.4$ and $p = 0.9$.

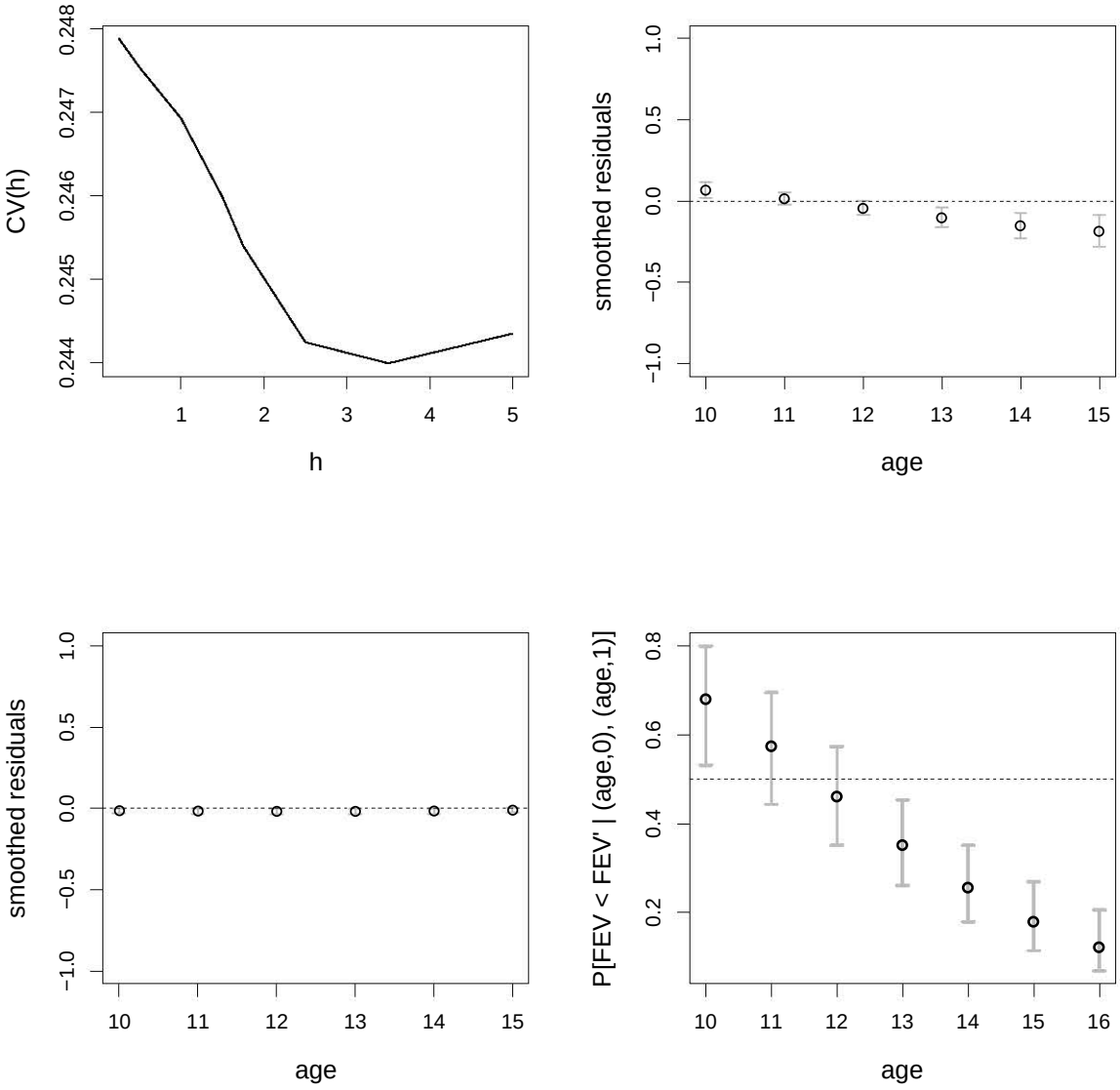

Figure 5.10: Top: cross validation score as a function of bandwidth (left) and GOF plot of the model without interaction (5.2) (right). Bottom: GOF plot of the model with interaction (5.32) (left) and $\hat{\mathrm{P}}(\mathrm{FEV} \preccurlyeq \mathrm{FEV}' \mid \mathrm{SMOKE} = 0, \mathrm{SMOKE}' = 1, \mathrm{AGE} = \mathrm{AGE}')$ as a function of AGE for model (5.32) (right). The grey bars indicate the pointwise 95% confidence intervals.

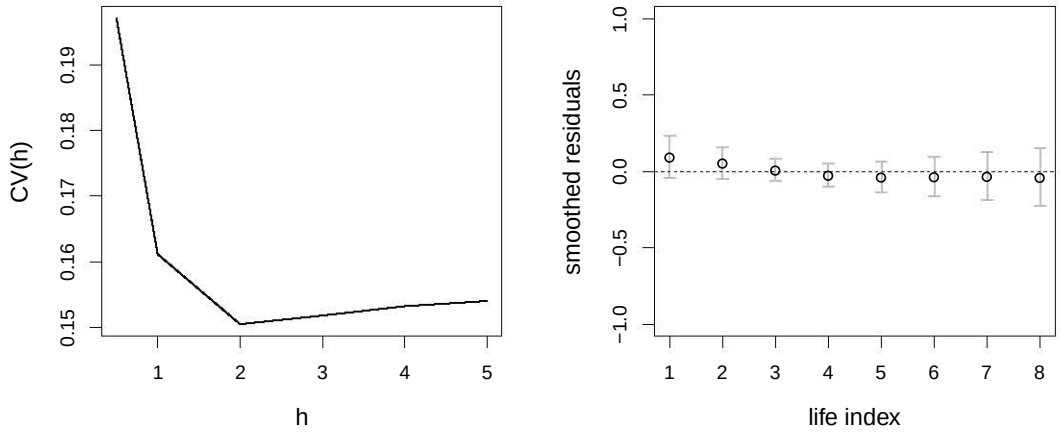

Figure 5.11: Cross validation score as a function of bandwidth (left) and the GOF plot (right) of model (5.33)

5.4.3 The food expenditure study

Two PIMs were considered for the food expenditure study. In Section 2.5.3 the effect of the household income (HI) on the relative food expenditure percentage (FEP) is examined with PIM

$$\text{logit}[\text{P}\left(\text{FEP} \preccurlyeq \text{FEP}'\right)] = \beta(\text{HI}' - \text{HI}). \tag{5.34}$$

Smoothed residuals are constructed with $\Delta = 100$ and bandwidths $h_1 = h_2 = h$ with $h \in \{100, 125, 150\}$, for which the optimal bandwidth is selected based on the cross-validation score (5.19) with the sum restricted to $\mathcal{I}_{\text{sub}} = \{(i, j) \mid 90 \leq \text{HI}_j - \text{HI}_i \leq 110\}$ for computational reasons. The top panel of Figure 5.12 shows the cross-validation score as a function of h; a minimum is attained at $h = 100$.

The middle left panel of Figure 5.12 plots the smoothed residuals as a function of the household income for model (5.34). The middle right panel shows the smoothed residuals restricted to $\text{HI} < 2000$. There is no convincing evidence for LOF and the GOF test based on the design points $\text{HI} \in \{400, 600, 800, 100, 1200, 1400, 1600\}$ confirms this: $S_\Delta = 3.33$ and $p = 0.91$.

In Section 3.2 the absolute food expenditure (FE) is examined with PIM

$$\text{P}\left(\text{FE} \preccurlyeq \text{FE}' \mid \text{HI}, \text{HI}'\right) = \Phi\left[\frac{(\text{HI}' - \text{HI})}{\sqrt{\text{HI}' + \text{HI}}}\gamma\right]. \tag{5.35}$$

The bottom left panel of Figure 5.12 shows the smoothed residuals. For low household incomes, the probability $\mathrm{P}\left(\mathrm{FE} \preccurlyeq \mathrm{FE}' \mid \mathrm{HI}' = \mathrm{HI} + 100\right)$ is overestimated with PIM (5.35). The GOF test confirms this: $S_\Delta = 84.8$ and $p < 0.0001$. As illustrated in Section 5.3.3 this can be caused by an inappropriate link function. The bottom right panel of Figure 5.12 shows the smoothed residuals for the PIM

$$\mathrm{P}\left(\mathrm{FE} \preccurlyeq \mathrm{FE}' \mid \mathrm{HI}, \mathrm{HI}'\right) = \operatorname{expit}\left[\frac{(\mathrm{HI}' - \mathrm{HI})}{\sqrt{\mathrm{HI}' + \mathrm{HI}}}\delta\right], \tag{5.36}$$

with $\hat{\delta} = 0.39$ (SE: 0.024 and $p < 0.0001$). The plot shows no LOF and the GOF test confirms this $S_\Delta = 5.7$ and $p = 0.68$. For this model, if the household income is, for example, 500 Belgian francs, the probability of larger food expenditure with a household income of 600 Belgian francs is estimated as 76.4%, while for model (5.35) this is 88.0%.

5.4.4 The Beck depression inventory

In Section 1.5 the BDI example was analyzed with PIM

$$\mathrm{P}\left(\mathrm{BD} \preccurlyeq \mathrm{BD}' \mid \mathrm{DOSE}, \mathrm{DOSE}'\right) = \operatorname{expit}\left[\beta(\mathrm{DOSE}' - \mathrm{DOSE})\right], \tag{5.37}$$

where BD denotes the BDI improvement, defined as the BDI at baseline (BDI_0) minus the BDI at the end of the study (BDI_1). Smoothed residuals are constructed with $\Delta = 1$ and bandwidths $h_1 = h_2 = h$ with $h \in \{0.5, 1, 2, 3, 4, 5\}$, for which the optimal bandwidth is selected based on the cross-validation score. The left panel of Figure 5.13 shows the cross-validation score which attains a minimum at $h = 3$. The middle panel shows the smoothed residuals. There is a weak pattern visible. The GOF test based on design points $\mathrm{BD} \in \{11, 15, 19, 23, 27\}$, however, indicates that this LOF is not significant: $S_\Delta = 6.5$ and $p = 0.26$.

In Section 2.5.4 the BDI example is analyzed with PIM

$$\mathrm{P}\left(\mathrm{BDI}_1 \preccurlyeq \mathrm{BDI}_1'\right) = \operatorname{expit}\left[\alpha[\mathrm{I}\left(\mathrm{BDI}_0 < \mathrm{BDI}_0'\right) - \mathrm{I}\left(\mathrm{BDI}_0 > \mathrm{BDI}_0'\right)] + \gamma(\mathrm{DOSE}' - \mathrm{DOSE})\right]. \tag{5.38}$$

The right panel of Figure 5.13 shows the smoothed residuals as a function of the sum $\mathrm{BDI}_0 +$ DOSE with $\boldsymbol{\Delta} = (1, 1)$ and $\boldsymbol{h}_1 = \boldsymbol{h}_2 = (3, 3)$. There is no convincing evidence of LOF. The GOF test based on design points $(\mathrm{BDI}_0, \mathrm{DOSE}) \in \{(31, 11), (39, 16), (47, 21), (55, 26)\}$ confirms this: $S_\Delta = 2.1$ and $p = 0.72$.

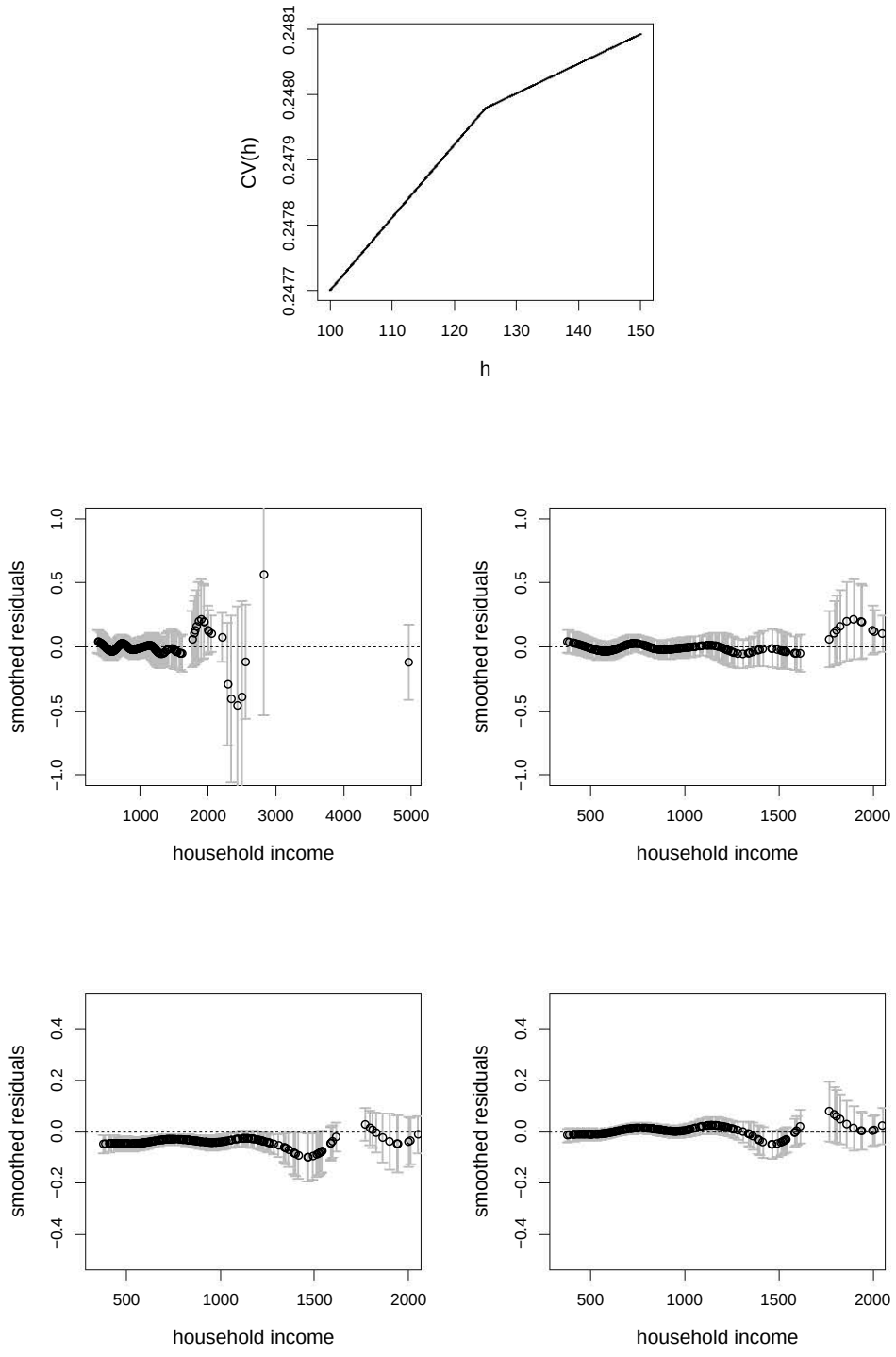

Figure 5.12: Top: cross validation score as a function of bandwidth h. Middle: GOF plot for model (5.34) (left) and the GOF plot restricted to HI < 2000 (right). Bottom: GOF plots for model (5.35) (left) and for model (5.36) (right).

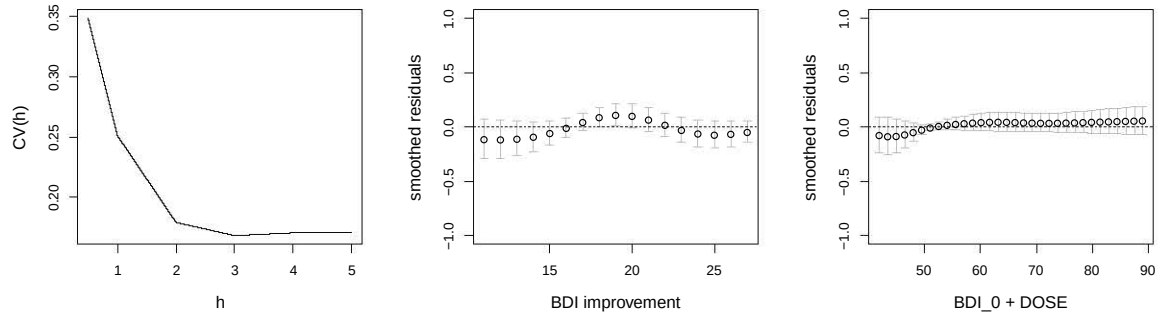

Figure 5.13: Left: cross validation score as a function of the bandwidth for model (5.37). Middle: GOF plot for model (5.37). Right: GOF plot for model (5.38).

5.5 Discussion

We constructed an informative GOF plot together with a formal GOF test for PIMs. The GOF plot provides information on how the model can be improved. The results of a power study suggested a decent performance of the test for some settings, but, however, also indicated that the test was sometimes too liberal, especially for sample sizes of 100 and 250. The GOF tools are consistent with the interpretation of a PIM, where the probability $\mathrm{P}\left(Y \preccurlyeq Y' \mid \boldsymbol{X}, \boldsymbol{X}' = \boldsymbol{X} + \boldsymbol{\Delta}\right)$ serves as a basis. The parameter $\boldsymbol{\Delta}$ should be chosen such that this probability has a meaningful interpretation; for future research it can be interesting to focus on an adaptive selection of $\boldsymbol{\Delta}$. The residuals are based on smoothers and the size of the test particularly depends on the choice of bandwidth.

We proposed a modified cross validation score to select the bandwidth automatically. The corresponding size were highly liberal, even for large sample sizes ($n = 250$). It may be of interest to extend the wild bootstrap method of Hardle and Mammen (1993) to the pseudo-observations setting, as this might improve the finite sample behaviour of the test. The test has good power for detecting an omission of a quadratic, sine, and exponential term as well as an omission of an interaction effect, while having low to moderate power for detecting a misspecified link function.

Many GOF statistics use all residuals to form a Cramér–von Mises, Anderson–Darling or Kolmogorov–Smirnov type of test. Because the pseudo-observations are sparsely correlated, the distribution theory of such test statistics is much harder than for many other types of regres-

sion models. By constructing our test statistic as a quadratic form which uses only a limited number of fixed design points, some technical difficulties are avoided. Future research may focus on extending our method so as to use all residuals. It is anticipated that this would make the method even more sensitive for detecting a wider range of model departures.

The methods constructed in this chapter can be considered as a first step towards developing GOF tools for PIMs. However, more research is required to refine these methods.

5.A Other goodness-of-fit statistics

The test statistic S_Δ, given by (5.14), is a quadratic form which uses only a limited number of fixed design points. As a result, the null distribution of S_Δ can be easily derived. Of course, other test statistics based on the smoothed residuals can be constructed, e.g. a Kolmogorov–Smirnov type of statistic

$$K_\Delta := \sqrt{n} \sup_x \left| \frac{\hat{R}(x, x+\Delta)}{\sqrt{\mathrm{Var}[\hat{R}(x, x+\Delta)]}} \right|.$$

To obtain the null distribution of K_Δ, we need to consider the smoothed residual $\hat{R}(x, x+\Delta)$ (or more generally $\hat{R}(x, x')$) as a stochastic process, which, in the context of PIMs, is not yet developed and which we postpone to future research. It is anticipated that the theory of stochastic U-processes will provide insights on how the null distribution of K_Δ can be established; see, for example, Sherman (1994). Alternative approaches for constructing appropriate test statistics can be based on residual cusum processes; see, for example, Su and Wei (1991).

5.B Automatic bandwidth selection and null distribution

Selecting the bandwidth automatically will change the distributional properties of S_Δ so that it no longer has a limiting chi-squared null distribution with m degrees of freedom, as given by Theorem 13. This is ignored in Sections 5.3.4 and 5.4, resulting in a liberal GOF test.

The appropriate null distribution can perhaps be approximated with bootstrapping techniques: for each bootstrap sample, we select the optimal bandwidth with the cross validation score

(5.19) and compute the test statistic. However, as mentioned in Section 5.5, this requires extending the wild bootstrap method of Hardle and Mammen (1993) to the pseudo-observations setting. Furthermore, since $|\mathcal{I}_n| = O(n^2)$, bootstrap and cross validation methods will be computationally very demanding. Therefore, constructing the appropriate null distribution associated with S_Δ with a data-driven choice of the bandwidth may be the topic of future research.

Chapter 6

An application to genomic data

The content of this chapter is based on the results published in

De Neve, J., Thas, O., Ottoy, J.P., and Clement, L. (2013) An extension of the Wilcoxon–Mann–Whitney test for analyzing RT-qPCR data. *Statistical Applications in Genetics and Molecular Biology (in press)*.

6.1 Introduction

Reverse transcription quantitative polymerase chain reaction (RT-qPCR) is considered as the gold standard for accurate, sensitive, and fast measurement of gene expression (Derveaux et al., 2010). The method is commonly used for the biological validation of differentially expressed genes that were discovered in large screening experiments with microarray or next generation sequencing technologies. The RT-qPCR is a cyclic process in which targeted molecules – here genes or microRNAs – are amplified and simultaneously quantified by measuring a fluorescence intensity. The raw RT-qPCR data are typically processed by plotting the fluorescence as a function of the cycle number and by summarizing this amplification curve in a single value, the quantification cycle C_q; see Figure 6.1 for an illustration. Popular procedures for calculating C_q-values are based on the number of cycles needed for the intensity to cross a certain threshold (illustrated in Figure 6.1), or on a cycle number derived from second derivatives of the amplification curve (e.g. Guescini et al., 2008). The C_q is *inversely* related to the number of target molecules (copy number): the larger the initial transcript abundance, the faster the inten-

sity grows and thus the smaller the C_q. RT-qPCR data have some typical characteristics that we introduce by examples. For more details on the biology, we refer to, for example, VanGuilder et al. (2008) and references therein.

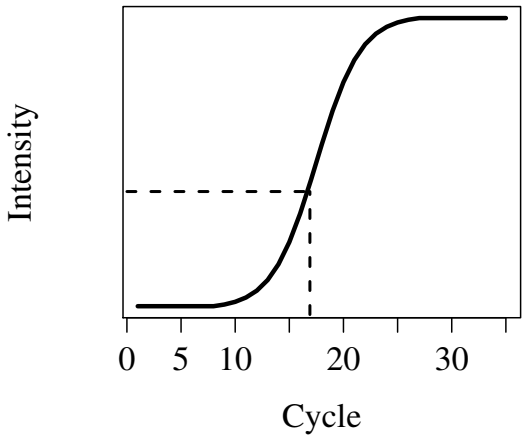

Figure 6.1: Illustrative plot of the fluorescence intensity as a function of the cycle number. The horizontal dashed line shows the threshold and corresponds to a C_q-value of 17.

We consider a housekeeping gene (which is a kind of control gene) and two microRNAs (miRNA) of two neuroblastoma studies. We refer to Section 6.4 for more details. Groups are formed based on the MYCN status which is known to be associated with neuroblastoma (e.g. Schulte et al., 2008; Alaminos et al., 2003). The left panel of Figure 6.2 shows nonparametric densities for housekeeping gene *UBC*, which is expected not to be affected by the MYCN amplification. However, the plot suggests a lower expression (thus higher C_q-values) when MYCN is amplified. This illustrates that RT-qPCR data are subject to experimentally induced variation which is not necessarily equal in both groups. This variation can be attributed to, for example, errors in the fluorescence quantification (Lalam, 2007) and differences in the amount of starting material and enzymatic efficiencies (Vandesompele et al., 2002). These errors affect the location and the tails of the densities.

The middle panel of Figure 6.2 shows the densities of *miR-17-5p* which is expected to be upregulated when MYCN is amplified (Fontana et al., 2008). Here MYCN amplification affects the location as well as the tails of the density. In cancer studies, for example, genes can sometimes only be expressed in a subsample of the populations during sampling (Tomlins et al., 2005; Thas

et al., 2012b), and consequently the tails of the density are affected.

The right panel of Figure 6.2 shows a histogram of *miR-639* when MYCN is amplified. If a feature is not expressed or the amplification step fails, the threshold is not reached. The expression is therefore undetermined and its value is set at the maximum number of cycles conducted, here 35. We refer to these values as *undetermined*. In the present setting, these undetermined values are considered as outliers.

Based on these characteristics, a test for assessing differential expression should therefore account for the experimental variation by providing a normalization constant, summarize location and tail effects with an intuitive effect size measure, and be robust to outliers. The uncertainty associated with the normalization should also be correctly propagated into the final statistical summaries for differential expression.

We propose an extension of the Wilcoxon–Mann–Whitney (WMW) test which incorporates normalization. In the microarray literature, tests that include preprocessing are often termed *unified tests*; see, for example, Wu and Irizarry (2007). Therefore we name our test the *unified WMW test* (uWMW).

The normalization constant and the effect size are defined in terms of the probabilistic index (PI) $\mathrm{P}\left(Y \preccurlyeq Y'\right)$ (where Y and Y' denote independent C_q-values) since it has an intuitive interpretation and is robust to outliers. The WMW test is a consistent rank test for testing the null hypothesis that Y and Y' coincide in distribution, against the alternative that $\mathrm{P}\left(Y \preccurlyeq Y'\right) \neq 0.5$. Fligner and Policello (1981) extended the WMW test so that it can be used for testing the less restrictive null hypothesis

$$H_0 : \mathrm{P}\left(Y \preccurlyeq Y'\right) = \frac{1}{2}.$$

The probabilistic index model (PIM) extends the Fligner and Policello WMW test by allowing for covariate adjustment. In this chapter we use the PIM framework for the construction of a WMW test for assessing differential expression, while normalizing the data simultaneously. Note that the PI is invariant under monotonic transformations, which is a desirable property for analyzing RT-qPCR data, as the relation between the number of molecules and the quantification cycle C_q depends on the PCR efficiencies which are unknown.

In Section 6.2 the uWMW test is described and Section 6.3 evaluates its performance in a simulation study. Section 6.4 illustrates the method on two case studies and Section 6.5 presents

the conclusions and discussion.

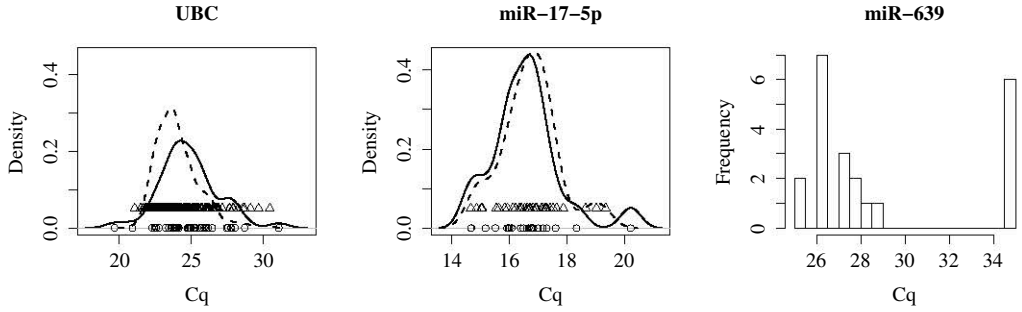

Figure 6.2: Nonparametric density estimates with Gaussian kernel for housekeeping gene *UBC* (left panel) and microRNA *miR-17-5p* (middle panel) when MYCN is amplified (—, ○) and when MYCN is normal (- - -, △). Rug plots are added to visualize the sample observations. The right panel shows the histogram of microRNA *miR-639* with limit of detection equal to 35.

6.2 The unified Wilcoxon–Mann–Whitney test

We start by studying the null hypothesis of the t-test after normalization. This null hypothesis is then reformulated in terms of the PI and a statistical test is proposed.

6.2.1 Null hypothesis

Let the random variable Y_{ijk} denote the quantification cycle C_q associated with feature $i \in \{1, \ldots, m+h\}$ (which can be a miRNA or a gene) of sample $j \in \{1, \ldots, n_k\}$ (e.g. patient or tissue) in treatment group $k \in \{1, 2\}$. The first m features are of interest and, if available, the last h features are the housekeeping features. In absence of housekeeping features set $h = 0$. Let $Y_{i.k}$ denote the C_q-value of feature i for a randomly selected sample in treatment group k. Let $Y_{..k}$ denote the C_q-value of a randomly selected feature of interest in a randomly selected sample of treatment group k. Hence, $Y_{..k}$ has a distribution function which is marginalized over all features of interest and over all samples. It will be convenient to denote the C_q-value of a randomly selected housekeeping feature in a randomly selected sample of treatment group k as $Y^*_{..k}$.

A popular normalization strategy consists of subtracting a normalization constant from the C_q-values for each sample. Vandesompele et al. (2002) consider the mean quantification cycles over stable housekeeping features and assumes that housekeeping features are, on average, not differentially expressed. We refer to this as housekeeping mean expression (HME) normalization. In absence of stable housekeeping features, Mestdagh et al. (2009) consider the mean quantification cycles over all expressed features, and assume, on average, a balance between up- and downregulation over all features. We refer to this as overall mean expression (OME) normalization.

The normalized data are given by

$$\tilde{Y}_{ijk} = Y_{ijk} - \hat{c}_{jk},$$

with $\hat{c}_{jk} = h^{-1} \sum_{i>m} Y_{ijk}$, for HME-normalization, and $\hat{c}_{jk} = m^{-1} \sum_{i \leq m} Y_{ijk}$ for OME-normalization. It is straightforward to show for feature i that the t-test based on normalized data tests the null hypothesis

$$H_0 : \mathrm{E}\left(Y_{i.1} - Y_{i.2}\right) = \mathrm{E}\left(Y_{i.1}\right) - \mathrm{E}\left(Y_{i.2}\right) = \Delta_1. \tag{6.1}$$

For HME-normalization

$$\Delta_1 \equiv \mathrm{E}\left(Y^*_{..1} - Y^*_{..2}\right),$$

i.e. Δ_1 is the mean difference in expression of the housekeeping features. Hence, testing if HME-normalized quantification cycles have, on average, a difference of 0, is equivalent to testing whether the original quantification cycles have, on average, a difference of Δ_1. A similar reasoning holds for the OME-normalization, with

$$\Delta_1 \equiv \mathrm{E}\left(Y_{..1} - Y_{..2}\right).$$

If Δ_1 is known, null hypothesis (6.1) can be tested with a classical t-test. In practice, however, Δ_1 has to be estimated first and this estimation has to be accounted for by the test procedure. The latter, however, is often ignored, so that an inflation of the type I error rate may be expected.

Hypothesis (6.1) can be reformulated in terms of the PI for constructing a null hypothesis which is more natural when adopting the WMW test:

$$H_0 : \mathrm{P}\left(Y_{i.1} \preccurlyeq Y_{i.2}\right) = \Delta_2, \tag{6.2}$$

with

$$\Delta_2 \equiv \mathrm{P}\left(Y^*_{..1} \preccurlyeq Y^*_{..2}\right), \tag{6.3}$$

or, in absence of stable housekeeping features,

$$\Delta_2 \equiv \mathrm{P}\left(Y_{\cdot\cdot 1} \preccurlyeq Y_{\cdot\cdot 2}\right). \tag{6.4}$$

The parameter (6.3) can be estimated by,

$$\hat{\Delta}_2 = \frac{1}{hn_1n_2}\sum_{i=m+1}^{m+h}\sum_{j=1}^{n_1}\sum_{j'=1}^{n_2}\mathrm{I}\left(Y_{ij1} \preccurlyeq Y_{ij'2}\right),$$

where $\mathrm{I}\left(x \preccurlyeq y\right) = \mathrm{I}\left(x < y\right) + 0.5\mathrm{I}\left(x = y\right)$, with $\mathrm{I}\left(\cdot\right)$ the indicator function. In a similar way, (6.4) can be estimated by

$$\hat{\Delta}_2 = \frac{1}{mn_1n_2}\sum_{i=1}^{m}\sum_{j=1}^{n_1}\sum_{j'=1}^{n_2}\mathrm{I}\left(Y_{ij1} \preccurlyeq Y_{ij'2}\right).$$

A naive approach for testing null hypothesis (6.2) is based on the statistic

$$\text{iWMW}_i := \frac{\sum_{j,j'}\mathrm{I}\left(Y_{ij1} \preccurlyeq Y_{ij'2}\right) - n_1n_2\hat{\Delta}_2}{\sqrt{n_1n_2(n_1+n_2+1)/12}}, \tag{6.5}$$

and using the null distribution of the classical WMW statistic. Note that iWMW_i reduces to the classical WMW statistic when replacing $\hat{\Delta}_2$ by 0.5. This method has two drawbacks. First, the test statistic is not properly standardized because the sampling variability of $\hat{\Delta}_2$ is ignored, and hence an inflation of the type I error rate may be expected. Second, it tests the more restrictive null hypothesis that the distributions of $Y_{i.1}$ and $Y_{i.2}$ coincide, instead of testing null hypothesis (6.2).

Therefore, in the next section, we extend the WMW test of Fligner and Policello (1981) for testing null hypothesis (6.2), while accounting for the estimation of Δ_2.

6.2.2 Test

PIMs are a natural framework to construct an appropriate test for (6.2). We first reprise the general formulation of PIM. Let $(Y, \boldsymbol{X})$ and $(Y', \boldsymbol{X}')$ be i.i.d., then a PIM is defined as

$$\mathrm{P}\left(Y \preccurlyeq Y' \mid \boldsymbol{X}, \boldsymbol{X}'\right) = m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}) = g^{-1}(\boldsymbol{Z}^T\boldsymbol{\beta}), \quad (\boldsymbol{X}, \boldsymbol{X}') \in \mathcal{X},$$

with $g(\cdot)$ and link function and $\boldsymbol{Z}$ a p-vector with elements that may depend on $\boldsymbol{X}$ and $\boldsymbol{X}'$. $\mathcal{X}$ denotes the set of covariates $(\boldsymbol{X}, \boldsymbol{X}')$ for which the model is defined. In our context the covariate vectors $\boldsymbol{X}$ and $\boldsymbol{X}'$ contain the information on the treatment group k and the feature i.

We restrict $\mathcal{X}$ to the couples $(\boldsymbol{X}, \boldsymbol{X}')$ which are both associated with the same feature and so that $\boldsymbol{X}$ corresponds to treatment group 1 and $\boldsymbol{X}'$ to treatment group 2. Consider the PIM with logit link

$$\mathrm{P}\left(Y_{i.1} \preccurlyeq Y_{i.2}\right) = \mathrm{expit}\left(\beta_0 + \beta_i\right). \tag{6.6}$$

Let $\mathrm{odds}\left(Y \preccurlyeq Y'\right) := \mathrm{P}\left(Y \preccurlyeq Y'\right)/[1-\mathrm{P}\left(Y \preccurlyeq Y'\right)]$. In the presence of housekeeping features, we impose the restriction $\beta_i = 0$ for $i = m+1, \ldots, m+h$, which implies

$$\beta_i = \log \frac{\mathrm{odds}\left(Y_{i.1} \preccurlyeq Y_{i.2}\right)}{\mathrm{odds}\left(Y^*_{..1} \preccurlyeq Y^*_{..2}\right)}, \quad i = 1, \ldots, m.$$

In the absence of housekeeping features, we impose the restriction $\sum_{i=1}^{m} \beta_i = 0$, leading to

$$\beta_i = \log \frac{\mathrm{odds}\left(Y_{i.1} \preccurlyeq Y_{i.2}\right)}{\mathrm{odds}\left(Y_{..1} \preccurlyeq Y_{..2}\right)}, \quad i = 1, \ldots, m. \tag{6.7}$$

Consequently, null hypothesis (6.2) is equivalent to

$$H_0 : \beta_i = 0, \tag{6.8}$$

for the two types of normalization. Theorems 1 and 2 provide the asymptotic theory for a consistent estimation of $\boldsymbol{\beta}^T = (\beta_1, \ldots, \beta_m)$. The estimator of $\boldsymbol{\beta}$, say $\hat{\boldsymbol{\beta}}$, has an asymptotic multivariate normal distribution with variance-covariance matrix $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}}$. We denote the consistent estimator for the variance-covariance matrix as $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}$. Hence, under null hypothesis (6.8),

$$\mathrm{uWMW}_i := \frac{\hat{\beta}_i}{\sqrt{\left(\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}\right)_{ii}}},$$

has an asymptotic standard normal distribution. This test is referred to as the *unified WMW test*.

Because PIM (6.6) models all data simultaneously, general linear null hypotheses that involve a subset of s features out of the m features in the experiment, can be formulated as

$$H_0 : \boldsymbol{H}\boldsymbol{\beta} = \boldsymbol{0}, \tag{6.9}$$

for some $s \times m$ matrix $\boldsymbol{H}$. The appropriate test statistic is given by

$$\mathrm{muWMW}_s := \left(\boldsymbol{H}\hat{\boldsymbol{\beta}}\right)^T \left(\boldsymbol{H}\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}\boldsymbol{H}^T\right)^- \left(\boldsymbol{H}\hat{\boldsymbol{\beta}}\right). \tag{6.10}$$

Under H_0, muWMW_s is asymptotically chi-squared distributed with degrees of freedom equal to the rank of $\boldsymbol{H}\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}\boldsymbol{H}^T$ and where $\boldsymbol{A}^-$ denotes a generalized inverse of a square matrix $\boldsymbol{A}$. This test is referred to as the *multivariate unified WMW test*.

With the offset $\beta_0 = 0$, the uWMW test simplifies to the WMW test of Fligner and Policello (1981). Note that the PI is also well defined in the presence of ties so that the test remains valid when undetermined values are substituted by the maximum number of cycles.

6.3 Simulation study

We present the results of three simulation studies to evaluate the performance of the uWMW test. The first study examines the null distribution and the second and third the performance in terms of detecting differentially expressed features.

6.3.1 Null distribution

The uWMW test is compared to the iWMW test, and to the WMW test after mean expression normalization. The test statistic of the latter can be expressed as

$$\text{nWMW}_i := \frac{\sum_{j,j'} \text{I}\left[(Y_{ij1} - \hat{c}_{j1}) \preccurlyeq (Y_{ij'2} - \hat{c}_{j'2})\right] - n_1 n_2 0.5}{\sqrt{n_1 n_2 (n_1 + n_2 + 1)/12}}, \tag{6.11}$$

and is commonly used in the qPCR literature. Note that normalization is based on the mean, while the effect size is in terms of the PI.

All p-values are calculated based on the asymptotic null distributions and data are simulated according to two distributions: the normal distribution with mean 0 and variance 4, i.e. $\text{N}(0, 4)$, and the Laplace/double exponential distribution with mean 0 and variance 2, $\text{L}(0, 2)$. The latter is chosen to illustrate that the test has a correct size for non-normal distributions too. Theoretical properties are empirically validated based on 1000 Monte-Carlo simulation runs and data are simulated from the same distribution so that $\Delta_2 = 0.5$.

In a first set-up, the design is restricted to two features: one for normalization and one for testing. Table 6.1 gives the empirical type I error rates at the 1%, 5%, and 10% significance levels, and $n = n_1 = n_2$ denotes the number of samples in each group. All results are obtained with R (R Core Team, 2012). The size of iWMW is consistently higher than its nominal level, because the estimation of Δ_2 is ignored. For $n = 10$, uWMW is slightly liberal and nWMW conservative; for $n = 25$ and $n = 50$ both tests correctly control for the type I error rate. The null distribution of WMW is conditional on the observed normalized data and is therefore condi-

tionally independent of HME-normalization. This explains the correct size of nWMW, despite the normalization is unaccounted for in the test. This is at the expense of a more restrictive null hypothesis

$$H_0 : F_{i.1} = F_{i.2},$$

with $F_{i.k}$ the cumulative distribution function of the normalized data of feature i in treatment group k.

Table 6.1: Empirical rejections rates (%) at the 1%, 5%, and 10% significance levels based on 1000 Monte-Carlo simulations. The design is restricted to two features, where the first is used for normalization and the second for testing. The Normal distribution with mean 0 and variance 4, $\mathrm{N}(0, 4)$, and the Laplace distribution with mean 0 and variance 2, $\mathrm{L}(0, 2)$, are used for simulating data for $n = 10$, $n = 25$, and $n = 75$ samples in each group.

n	uWMW			iWMW			nWMW		
	1%	5%	10%	1%	5%	10%	1%	5%	10%
	$\mathrm{N}(0, 4)$								
10	2	6.9	12.2	6.7	16.4	24.9	0.6	4.4	8.7
25	0.5	4.7	9.3	6.3	15.7	23.9	0.4	4.2	8.5
75	1.7	5.3	9.8	6.8	17.3	24.7	0.7	4.9	9.1
	$\mathrm{L}(0, 2)$								
10	1.3	6.0	12.3	6.1	17.1	24.9	0.3	3.4	8.8
25	1.6	5.3	10.4	6.8	15.4	22.2	0.8	5.3	9.2
75	1.4	4.1	8.2	5.4	15.6	25.1	1.3	3.9	8.8

In a second set-up, the number of features, say m, is set to 5 or 20, the number of samples to $n = 10$ or $n = 25$, and all features are considered for normalization. Table 6.2 gives the empirical type I error rates at the 1%, 5%, and 10% significance levels. For uWMW and nWMW similar conclusions hold as previously. The empirical type I error rate of iWMW is closer to its nominal level, because Δ_2 is now more accurately estimated by using all data. However, for $m = 5$ iWMW is conservative.

Table 6.2: Empirical rejections rates (%) at the 1%, 5%, and 10% significance levels based on 1000 Monte-Carlo simulations. The design is restricted to $m = 5$ or $m = 20$ features which are all used for normalization. The Normal distribution with mean 0 and variance 4, $\mathrm{N}(0, 4)$, and the Laplace distribution with mean 0 and variance 2, $\mathrm{L}(0, 2)$, are used for simulating data for $n = 10$ and $n = 25$ samples in each group.

(m, n)	uWMW			iWMW			nWMW		
	1%	5%	10%	1%	5%	10%	1%	5%	10%
	$\mathrm{N}(0, 4)$								
$(5, 10)$	1.0	5.7	11.0	0.2	2.4	4.8	0.3	3.9	7.6
$(5, 25)$	1.2	5.5	10.4	0.4	3.1	6.7	1.1	4.6	9.4
$(20, 10)$	1.3	7.8	13.4	0.7	6.0	10.8	0.8	5.5	10.0
$(20, 25)$	1.1	5.6	10.6	0.7	4.9	9.6	0.7	5.2	10.6
	$\mathrm{L}(0, 2)$								
$(5, 10)$	1.7	6.2	12.5	0.4	3.2	6.4	1.2	4.0	8.6
$(5, 25)$	1.3	6.6	11.9	0.1	3.3	7.5	0.8	5.2	11.1
$(20, 10)$	0.9	7.1	12.7	0.4	5.0	9.9	0.9	4.6	9.7
$(20, 25)$	1.0	5.6	11.3	0.7	4.6	9.6	1.0	5.3	9.3

6.3.2 Performance

We consider two additional simulation studies for studying the sensitivity and the specificity of uWMW. In summary, quantification cycles for 200 features are simulated over two groups, each consisting of 30 samples. Of the 200 features, 30 are differentially expressed. Different types of treatment effects are used in the simulation according to two set-ups. Appendix 6.A gives all details.

Set-up A

In a first set-up, we consider 3 types of effects:

1. differential expression for 10 features according to a *location-shift effect* which consists of adding a constant to all sample observations in one group. This corresponds to the setting where the treatment affects all subjects in the treatment group.

2. differential expression for 10 features according to a *tail effect* which consists of adding a constant to a third of the sample observations in one group. This corresponds to the setting where the treatment only affects a part of the population.

3. differential expression for 10 features according to a *contaminated location-shift effect* which consists of adding a constant to all sample observations in one group and by including outliers in the other group. This corresponds to the setting where the treatment affects all subjects in the treatment group, while for the other group, the PCR reaction failed for some subjects, resulting in high C_q-values.

We study the performance for each type of effect separately as well as for all effects combined. The latter is referred to as the *overall* effect.

For each simulated dataset, additional outliers for 10 non differentially expressed features were included. This corresponds to the setting where the PCR reaction failed, resulting in high C_q-values. These outliers allow for assessing the robustness of the normalization. Figure 6.6 in Appendix 6.A.1 gives nonparametric density estimates for several features for the different treatment effects.

1000 datasets are simulated and analyzed with a) the uWMW test using the normalization based on all features, b) the nWMW test using OME-normalization and, c) nWelch, a Welch t-test upon OME-normalization.

The analysis of each simulated dataset results in an ROC-curve and Figure 6.3 shows the average of these curves, where the average is calculated for each significance level. The false positive rate is restricted to 30%.

For the overall effect, when all 30 differentially expressed features are included, uWMW slightly outperforms nWMW, and both tests outperform nWelch.

For the separate types of differential expression, nWelch consistently underperforms uWMW and nWMW. This can be clearly seen from the bottom right panel of Figure 6.3: the outliers cancel out the location-shift on average. For the other effect types this can perhaps be explained by the non-normality of the data (see also Figure 6.6 in Appendix 6.A.1), for which it is generally known that the t-test, even under the location-shift assumption, is not necessarily the most powerful test. For the location-shift and contaminated location-shift effects, uWMW slightly outperforms nWMW. This can be explained by the sensitivity of mean expression normalization to the additional outliers. Both methods have a similar performance when the outliers for the 10 non differentially expressed features are excluded; see Figure 6.7 in Appendix 6.B.

Set-up B

We consider a second set-up to examine the impact of undetermined values, i.e. quantification cycles that did not reach the threshold and which are imputed by the maximum number of cycles (limit of detection, LOD). We first simulate differential expression for 30 features according to a location-shift effect without undetermined values; see Section 6.A.2 for details. This corresponds to the ideal setting without amplification failures. The left panel of Figure 6.4 gives the average ROC-curves based on 1000 simulated datasets. All three methods have a good and similar performance. In a second step, approximately a third of the data are randomly selected as "undetermined values" and are substituted by the LOD. The normalization of nWMW and nWelch is based on all expressed features (i.e. features which are not undetermined) following the rationale of the OME-normalization of Mestdagh et al. (2009). Normalization of uWMW is based on all features because it is robust to outliers. The right panel of Figure 6.4 gives the

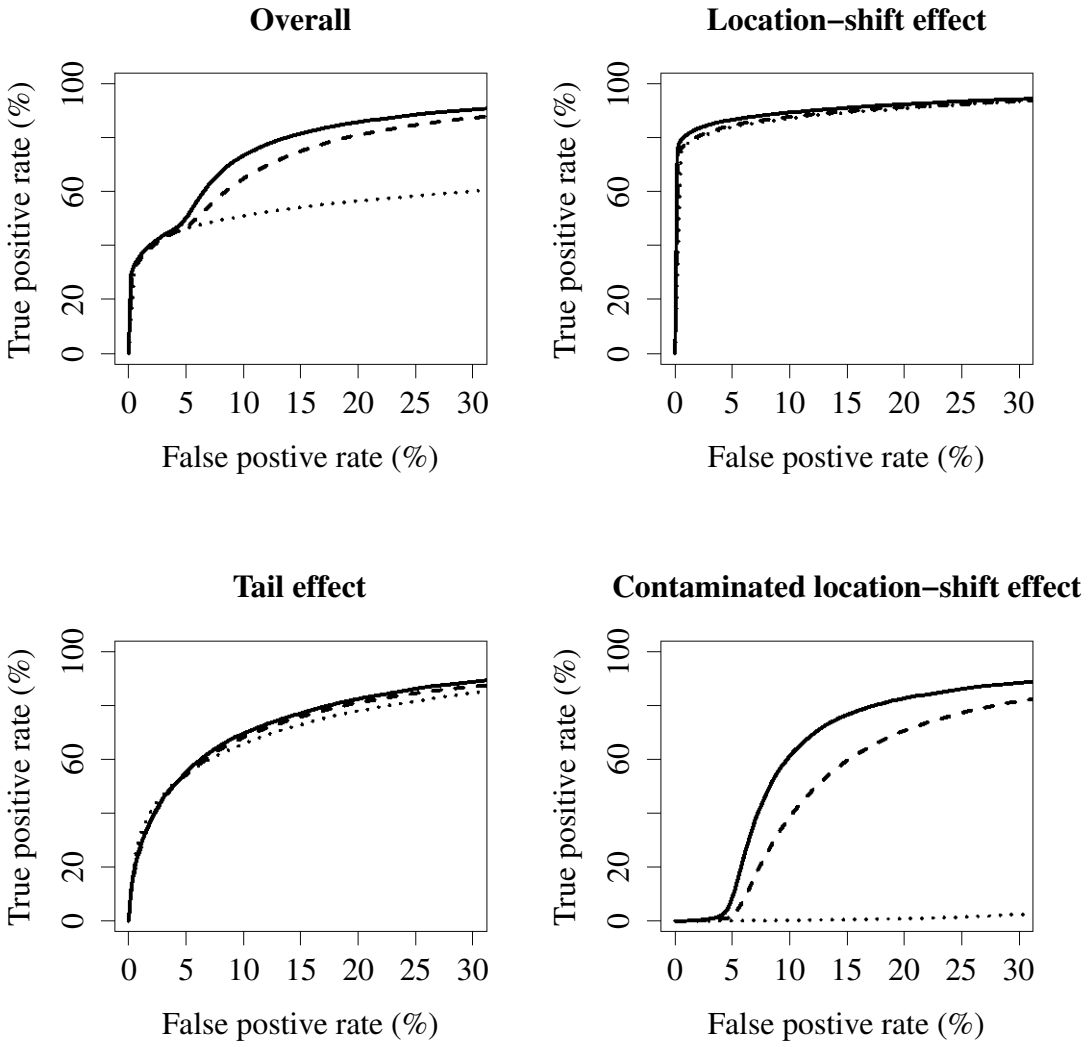

Figure 6.3: Average ROC-curves for uWMW (—), nWMW (- - -), and nWelch ($\cdots$). The top left panel shows the ROC-curve when all 30 differentially expressed features are included. The other panels show the average ROC-curve for each type of treatment effect separately, thus by only including the corresponding 10 differentially expressed features.

average ROC-curve. The performance of all three methods decreases as compared to the ideal setting without undetermined values. The performance of nWelch decreased more drastically as compared to uWMW and nWMW. This is a consequence of the sensitivity of the mean to the undetermined values. nWMW is slightly superior to uWMW since the normalization of nWMW ignores all undermined values. However, in practice, it can be difficult to distinguish between an expressed feature that has an undermined value because of a failure in the amplification and a feature that has an undetermined value because it is not expressed. The normalization of uWMW makes use of all data at the expense of a minor decrease in performance.

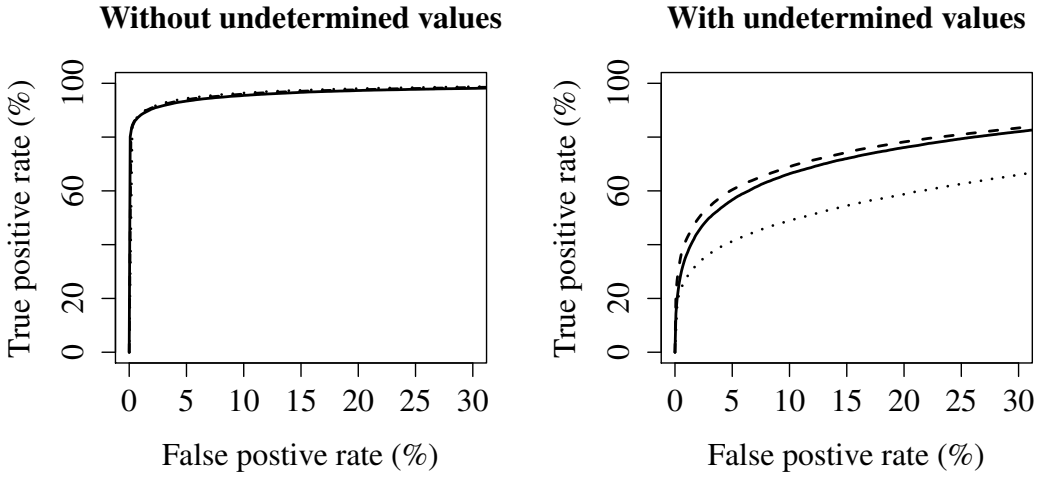

Figure 6.4: Average ROC-curves for uWMW (—), nWMW (- - -), and nWelch ($\cdots$). The left panel shows the ROC-curve for a location-shift effect without undetermined values. The right panel shows the ROC-curve based on the same data, but for which approximately a third of the data are randomly substituted by the maximum number of cycles (undetermined values).

6.4 Examples

6.4.1 The neuroblastoma microRNA study

The data are taken from Mestdagh et al. (2009). 448 miRNAs and controls are quantified in 61 neuroblastoma (NB) tumor samples: 22 MYCN amplified and 39 MYCN single copy samples. 107 miRNAs consist of at least 85% undetermined values in both groups and are removed for further analysis.

The *mir-17-92* cluster is a direct target of the MYC family of transcription factors using chromatin immunoprecipitation. In these NB cells, MYCN binds to the *mir-17-92* promoter and activates *mir-17-92* expression, and therefore differential expression is expected (Mestdagh et al., 2009; Fontana et al., 2008; O'Donnell et al., 2005).

The multivariate unified WMW test, with normalization based on all features, confirms that at least one miRNA of this cluster is differentially expressed in terms of the PI (p-value $<$ 0.00001). Table 6.3 shows the results of the uWMW test for each feature separately. The false discovery rate is controlled by the method of Benjamini and Hochberg (1995) (BH-FDR) using the `multtest` R-package (Pollard et al., 2010). At a 5% FDR, 7 of 8 miRNAs in the *mir-17-92* cluster are significantly upregulated when MYCN is amplified. We illustrate the interpretation for *miR-92*: the odds for upregulation relative to the overall odds is estimated by 5.9. When MYCN is amplified, it is thus more likely that *miR-92* is upregulated. Mestdagh et al. (2009) argued that *mir-181a* and *mir-181b* should also be differentially expressed, which is supported by our analysis; see Table 6.3. In summary, our results correspond to the findings of Mestdagh et al. (2009), who concluded that all miRNAs, except *miR-17-3p*, were differentially expressed. These results demonstrate that the uWMW test succeeds in detecting miRNAs which are believed to be differentially expressed. Table 6.3 also shows the results of the nWMW test as well as the associated effect size which is estimated by

$$\hat{\gamma}_i = \frac{1}{n_1 n_2} \sum_{j,j'} \mathrm{I}\left[(Y_{ij1} - \hat{c}_{j1}) \preccurlyeq (Y_{ij'2} - \hat{c}_{j'2})\right]. \tag{6.12}$$

Since the OME-normalization is performed within the indicator operator, the interpretation of this effect size on population level is obscured. However, both the uWMW and nWMW tests suggest an upregulation when MYCN is amplified.

6.4.2 The neuroblastoma gene study

The second neuroblastoma study is part of a larger study (Vermeulen et al., 2009). The data, quantifying 59 genes in 363 children, were used to train and to validate a multigene-expression signature study for predicting outcomes for children with neuroblastoma. In addition to gene expression, several risk factors, such as age at diagnosis, International Neuroblastoma Staging System stage, and MYCN status are reported. Housekeeping genes are provided for normalization. We focus on differential expression based on MYCN status, and because the genes are

Table 6.3: Results of the neuroblastoma miRNA study according to the uWMW test, with $\hat{\beta}$ the estimate of (6.7), SE the corresponding standard error, and with p-value adjustment according to BH-FDR, and according to the nWMW test, with $\hat{\gamma}$ as in (6.12) and with p-value adjustment according to BH-FDR.

	uWMW				nWMW	
miRNA	$\hat{\beta}$	SE	$\exp(\hat{\beta})$	adj. p-value	$\hat{\gamma}$	adj. p-value
miR-17-92						
miR-17-3p	0.19	0.31	1.2	0.6810	0.61	0.2449
miR-17-5p	0.80	0.31	2.2	0.0369	0.70	0.0279
miR-18a	0.97	0.31	2.6	0.0151	0.75	0.0052
miR-18a#	1.12	0.31	3.1	0.0040	0.83	0.0002
miR-19a	1.21	0.31	3.3	0.0022	0.82	0.0003
miR-19b	0.89	0.31	2.4	0.0208	0.75	0.0060
miR-20a	1.10	0.32	3.0	0.0056	0.79	0.0010
miR-92	1.77	0.38	5.9	0.0003	0.90	< 0.0001
miR-181						
miR-181a	1.37	0.33	3.9	0.0010	0.86	< 0.0001
miR-181b	0.90	0.31	2.5	0.0219	0.77	0.0030

selected for outcome prediction, we expect most to be differentially expressed.

For the uWMW test with housekeeping normalization, all genes are differentially expressed at a 5% BH-FDR. Figure 6.5 shows the nonparametric density estimates for gene *MRPL3*. Based on the WMW test without normalization, the odds for downregulation when MYCN is amplified is estimated by 0.81 (adjusted p-value 0.23); hence it is unlikely that this gene is downregulated. With the uWMW test, however, the odds for downregulation when MYCN is amplified relative to the overall odds of the housekeeping genes is estimated by 1.6 (adjusted p-value 0.0087). When MYCN is amplified it is now more likely that *MRPL3* is downregulated. nWMW based on housekeeping mean expression normalization confirms this (adjusted p-value < 0.00001). The effect size is given by 0.74, but, as explained in Section 6.4.1, its interpretation is not unambiguous.

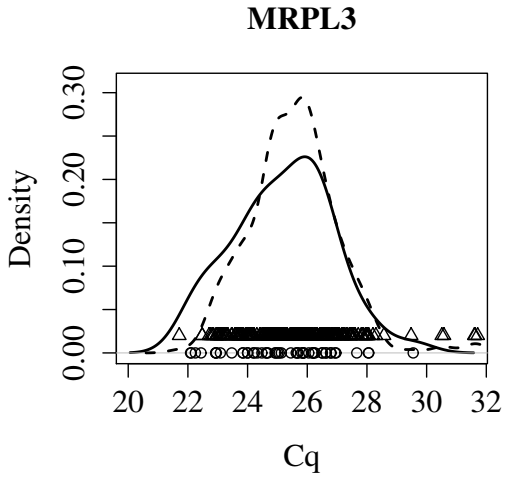

Figure 6.5: Nonparametric density estimates with Gaussian kernel for gene *MRPL3* of the neuroblastoma gene study for MYCN amplified (—, ∘) and MYCN normal (- - -, △). Rug plots are added to visualize the sample observations.

6.5 Discussion

Differential expression analysis with RT-qPCR requires normalization so as to account for technical variation which cannot be attributed to the treatments. Current methods subtract a normalization constant from the data prior to the downstream statistical analysis. When a t-test is used within the data analysis pipeline, the effect size measure has an intuitive interpretation. How-

ever, the t-test is sensitive to outliers, and whereas the treatment can affect the shape of the outcome distribution, the t-test has only power for detecting difference in means. Therefore, the Wilcoxon–Mann–Whitney (WMW) test is often preferred in practice. Applying the WMW test on normalized data, however, obscures its interpretation. It is well known that the WMW test can be interpreted in terms of the probabilistic index, but it is not clear how it can be interpreted on a population level after subtracting a normalization constant from the data.

RT-qPCR experiments often aim at validating differentially expressed features that were discovered with microarray or next generation sequencing screens. Such biological validation experiments are often an (intermediate) endpoint of a study. Hence, quantifying and interpreting the effects is very important for increasing the insight in the biological processes under study. Within this context, we extended the WMW test by incorporating the normalization in the statistical testing procedure. The method has the following properties:

- Both normalization and effect size are formulated in terms of the probabilistic index, which results in an intuitive interpretation in terms of the odds for down- or upregulation, keeps the normalization transparent, and is invariant under monotonic transformations.
- It detects location and tail effects while being robust to outliers.
- The uncertainty associated with the normalization is accounted for, so that the type I error rate is (asymptotically) correctly controlled.
- Based on the results of a simulation study with realistic settings, the method is at least competitive with classical approaches for analyzing differential expression in RT-qPCR data.
- All data are modelled simultaneously, which allows a straightforward extension towards tests on sets of features using general linear null hypotheses.
- The distributional theory is semiparametric requiring minimal assumptions and the asymptotic approximations are reasonable for moderated sample sizes.

6.A Simulation set-ups

The MYCN single copy group of the neuroblastoma miRNA study is used to set up the simulation study. This study quantifies 430 miRNAs in 39 samples of which 135 miRNAs have undetermined values in at least 50% of the samples. These miRNAs are not considered for the simulation set-up and the remaining 295 miRNAs are used to fit nonparametric densities to the expressed values (quantification cycles).
From these 295 densities 200 were selected at random for the generation of expressions for 60 samples, using the nonparametric density fits. Half of these samples are assigned to the first group and the other half to the second. One simulated dataset thus consists of 200 features and 60 samples over two groups. Differentially expressed features are then introduced by adding a constant to samples in one of the groups. The differentially expressed features are included in a way so that up and down regulation is balanced, which is an assumption of uWMW, nWMW, and nWelch.

6.A.1 Set-up A

We simulate differential expression according to a

- *location-shift effect*. Add a constant δ to the quantification cycles of all samples in group 1, where
  - $\delta = 1$ for features $1, 2, 3$ for a small treatment effect.
  - $\delta = 3$ for features $4, 5, 6$ for a moderate treatment effect.
  - $\delta = 6$ for features $7, \ldots, 10$ for a large treatment effect.
- *tail effect*. For features $11, \ldots, 20$ in group 2 add $\delta = 3$ to samples $1, \ldots, 10$, so that only a third of the samples in the second group are differentially expressed.
- *contaminated location-shift effect*. For features $21, \ldots, 30$ add $\delta = 3$ to all samples in the second group. We contaminate this location-shift effect by adding $\delta = 9$ to samples $1 \ldots, 10$ in the first group.

Figure 6.6 shows nonparametric density estimates for randomly selected features according to each type of treatment effect.

To examine robustness of the normalization procedures, we also included outliers for non-differentially expressed features: a constant $\delta = 9$ is added to samples $1, \ldots, 5$ in the first group for features $31, \ldots, 40$. These outliers make up 0.4% of the data.

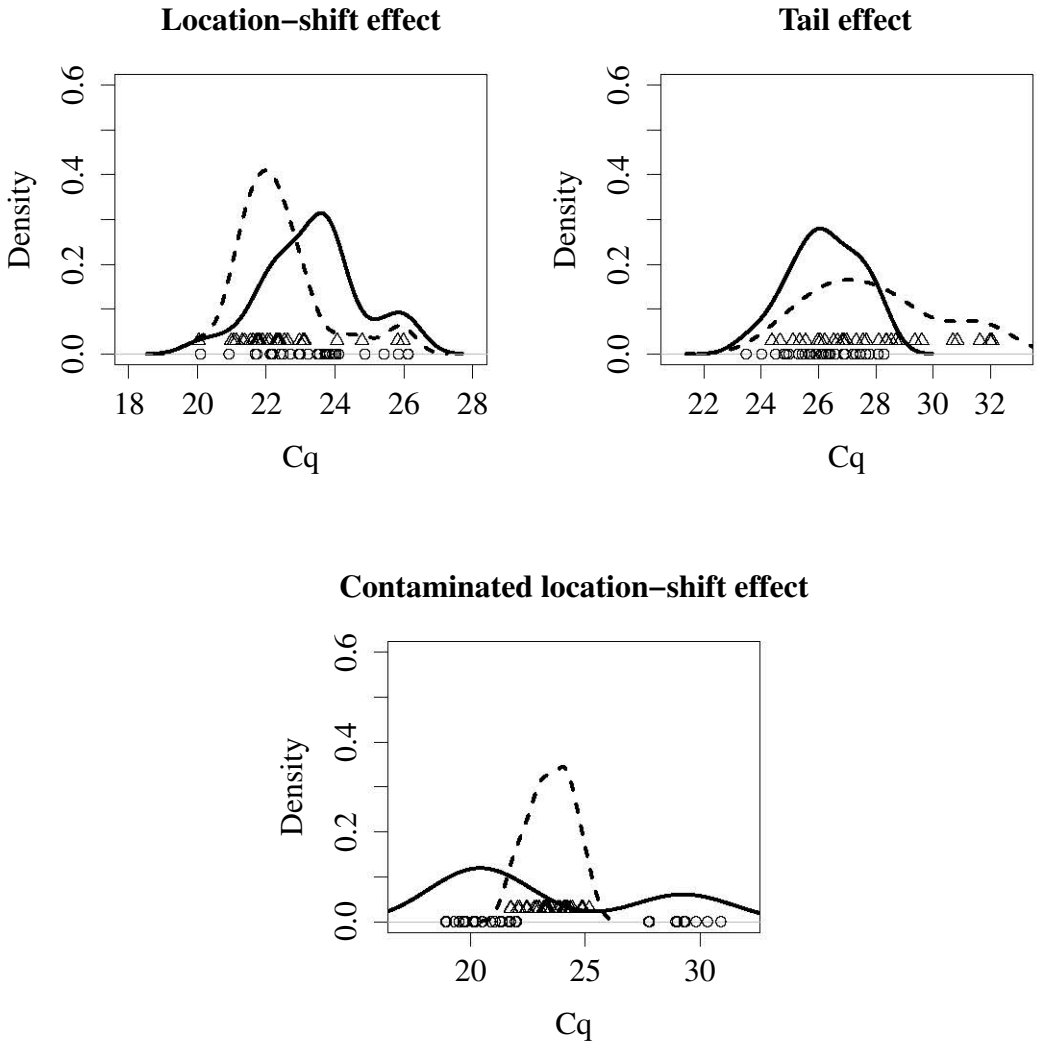

Figure 6.6: Nonparametric densities estimates of simulated data of group 1 (—, ∘) and group 2 (- - -, △) for randomly selected features according to the different types of treatment effects: location-shift effect (top left), tail effect (top right), and contaminated location-shift effect (bottom). Rug plots are added to visualize the sample observations.

6.A.2 Set-up B

In a first step we simulate data without undetermined values and include of location-shift effect for 30 of the 200 features by adding a constant

- $\delta = 1$ to the quantification cycles of all samples in group 1 for features $1, \ldots, 10$.
- $\delta = 3$ to the quantification cycles of all samples in group 1 for features $11, \ldots, 20$.
- $\delta = 6$ to the quantification cycles of all samples in group 2 for features $21, \ldots, 30$.

In a second step, 34% (which corresponds to the percentage of undetermined values of the neuroblastoma miRNA study with a detection cut-off of $C_q = 35$) of the data are randomly selected and replaced by 35 so as to represent the undetermined values.

6.B Additional simulation study

Figure 6.7 gives the average ROC-curves for the simulation study as described in Appendix 6.A.1, without the outliers in samples $1, \ldots, 5$ of the first group for features $31, \ldots, 40$. The performance of nWMW is now similar to uWMW.

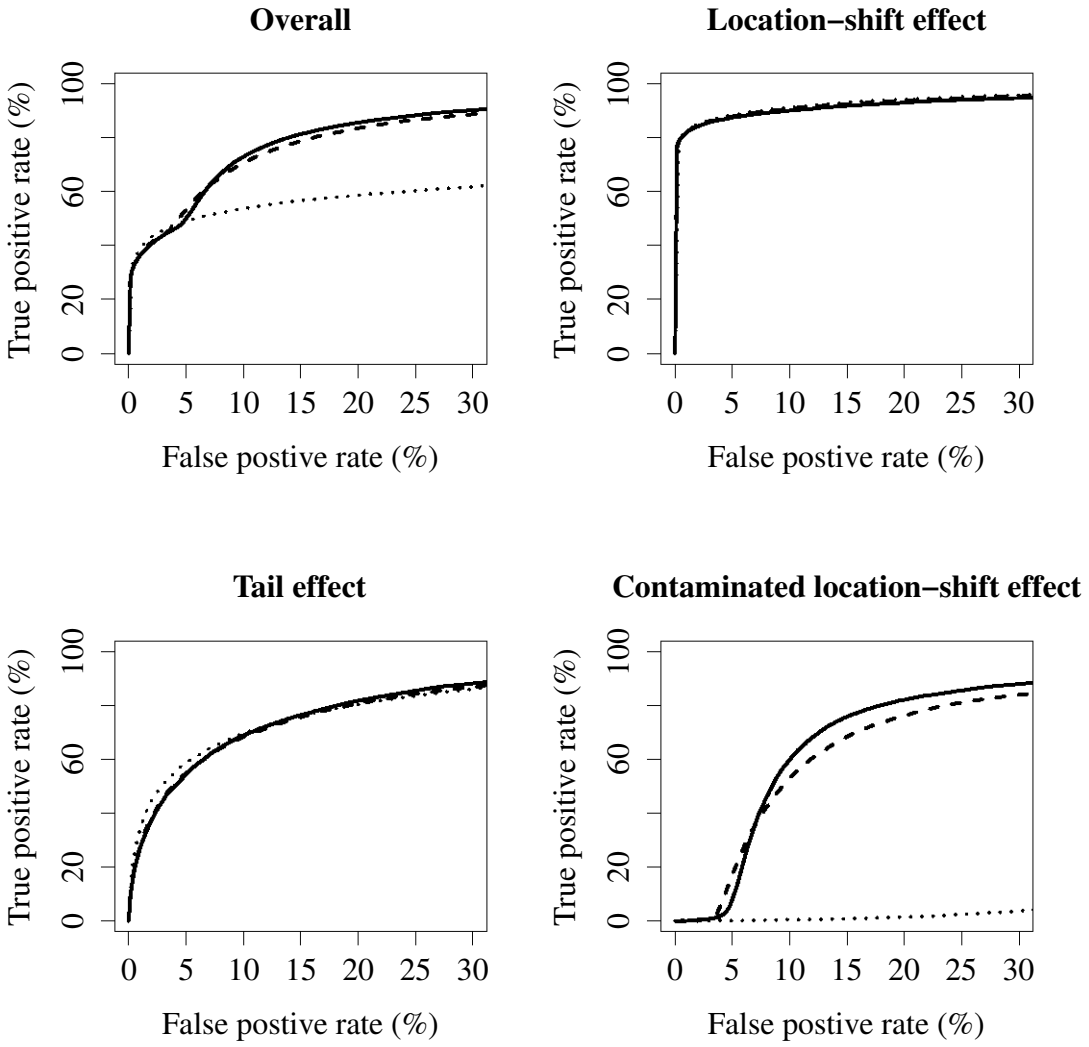

Figure 6.7: Average ROC-curves without outliers for uWMW (—), nWMW (- - -), and nWelch ($\cdots$). The top left panel shows the ROC-curve when all 30 differentially expressed features are included. The other panels show the average ROC-curve for each type of treatment effect separately, thus by only including the corresponding 10 differentially expressed features.

Chapter 7

Semiparametric efficiency

The content of this chapter is the result of many discussions with, and internal reports from, Olivier Thas, Stijn Vansteelandt, and Karel Vermeulen. However, the final form of this chapter is from my own hand and goes beyond these internal reports.

Many lemmas and theorems in Sections 7.2 and 7.3 are adapted from chapters 1-4 of Tsiatis (2006) and chapter 8 of Newey and McFadden (1994).

7.1 Motivation and outline

In Chapter 2 we proposed a semiparametric estimator for PIM-parameters with asymptotic theory based on the asymptotics of Lumley and Mayer-Hamblett (2003). Their estimating equations (2.15) make use of the independent working correlation matrix, i.e. for estimating the model parameters, we use the working assumption that the pseudo-observations are mutually independent. This working assumption is incorrect, since pseudo-observations which share a common outcome, e.g. $\mathrm{I}\left(Y \preccurlyeq Y'\right)$ and $\mathrm{I}\left(Y \preccurlyeq Y''\right)$, are generally not independent. This incorrect working assumption does not affect consistency and asymptotic normality of the estimator, nor consistency of the variance sandwich estimator (Lumley and Mayer-Hamblett, 2003). However, it can affect efficiency so that the estimator does not attain the semiparametric efficiency bound.

Furthermore, since we restrict the PIM framework to a random sample of i.i.d. observations it

is anticipated that the assumption and the regularity conditions of Lumley and Mayer-Hamblett (2003) are too strong since they hold for a more general class of data and models, i.e. marginal models for sparsely correlated data. Therefore, in this chapter, we address the following questions:

- Can we find more efficient estimators than the one proposed in Chapter 2?
- Can we find the asymptotic properties of the estimators without relying on the results of Lumley and Mayer-Hamblett (2003)?

To answer these questions we need some notion of the theory of semiparametric models. Since this is based on Hilbert spaces and parametric submodels, we start by introducing these concepts in Section 7.2. In Section 7.3 we construct the asymptotic theory for PIMs and in Section 7.4 we apply the general theory to a specific setting. Section 7.5 gives the conclusions and discussion. For literature on semiparameteric models and semiparametric estimation, we refer to Newey and McFadden (1994); Powell (1994); Bickel et al. (1998); Tsiatis (2006).

7.2 Introduction

We start by formally introducing a *semiparametric model*. Consider a random sample of i.i.d. observations $\{\boldsymbol{Z}_i = (Y_i, \boldsymbol{X}_i) \mid i = 1, \ldots, n\}$ with joint density $f_{\boldsymbol{Z}}(\boldsymbol{z}) = f_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x})$. Here $\boldsymbol{Z}$ may be continuous, discrete, or a combination, but without loss of generality we refer to $f_{\boldsymbol{Z}}(\boldsymbol{z})$ as a density. By $\mathcal{M}$ we denote the class of densities, i.e.

$$\mathcal{M} := \left\{ f_{\boldsymbol{Z}}(\boldsymbol{z}) \mid f_{\boldsymbol{Z}}(\boldsymbol{z}) \geq 0, \ \forall \boldsymbol{z} \quad \text{and} \quad \int f_{\boldsymbol{Z}}(\boldsymbol{z}) d\boldsymbol{z} = 1 \right\}.$$

Note that we use the notation $d\boldsymbol{z}$ for both the Lebesgue and counting measure.

Definition 6 (Semiparametric model)**.** *The class of densities corresponding to a semiparametric model, $\mathcal{M}_{SP} \subset \mathcal{M}$, can be described as*

$$\mathcal{M}_{SP} = \left\{ f_{\boldsymbol{Z}}(\boldsymbol{z}; \boldsymbol{\theta}) \mid \boldsymbol{\theta}^T = (\boldsymbol{\beta}^T, \boldsymbol{\eta}^T), \ \boldsymbol{\beta} \in \Theta \subset \mathbb{R}^p \right\}, \tag{7.1}$$

i.e. the density can be described by a finite-dimensional parameter $\boldsymbol{\beta}$ and a possibly infinite-dimensional nuisance parameter $\boldsymbol{\eta}$.

Hilbert spaces of random vectors form the cornerstone of semiparametric theory. The following section is based on chapter 2 of Tsiatis (2006). It provides some useful results without going into technical details.

7.2.1 Review on Hilbert spaces for random vectors

A real Hilbert space, say $\mathcal{H}$, is a complete normed linear vector space for which an inner product is defined.

Definition 7 (Inner product)**.** *For a linear vector space $\mathcal{H}$, an inner product, $\langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$, is a function satisfying, $\forall \boldsymbol{h}_1, \boldsymbol{h}_2, \boldsymbol{h}_3 \in \mathcal{H}$ and $\forall \lambda \in \mathbb{R}$,*

1. $\langle \boldsymbol{h}_1, \boldsymbol{h}_2 \rangle = \langle \boldsymbol{h}_2, \boldsymbol{h}_1 \rangle$.
2. $\langle \boldsymbol{h}_1 + \boldsymbol{h}_2, \boldsymbol{h}_3 \rangle = \langle \boldsymbol{h}_1, \boldsymbol{h}_3 \rangle + \langle \boldsymbol{h}_2, \boldsymbol{h}_3 \rangle$.
3. $\langle \lambda \boldsymbol{h}_1, \boldsymbol{h}_2 \rangle = \lambda \langle \boldsymbol{h}_1, \boldsymbol{h}_2 \rangle$.
4. $\langle \boldsymbol{h}_1, \boldsymbol{h}_1 \rangle \geq 0$ *and* $\langle \boldsymbol{h}_1, \boldsymbol{h}_1 \rangle = 0 \Leftrightarrow \boldsymbol{h}_1 = \mathbf{0}$.

Based on the inner product, a norm can be defined

$$\|\boldsymbol{h}\| := \sqrt{\langle \boldsymbol{h}, \boldsymbol{h} \rangle}, \quad \boldsymbol{h} \in \mathcal{H}.$$

The norm can be used to describe the length of a vector $\boldsymbol{h} \in \mathcal{H}$, i.e. the distance from $\boldsymbol{h}$ to the origin, denoted as $\mathbf{0}$. Furthermore, the inner product allows us to define orthogonality as $\langle \boldsymbol{h}_1, \boldsymbol{h}_2 \rangle = 0$, also denoted as $\boldsymbol{h}_1 \perp \boldsymbol{h}_2$.

We now construct a Hilbert space for random vectors. Consider a probability space $(\Omega, \mathcal{F}, \mathrm{P})$, with Ω the sample space, $\mathcal{F}$ the corresponding σ-algebra, and P the probability measure over $(\Omega, \mathcal{F})$ that generates the data $\boldsymbol{Z}_i$, $i = 1, \ldots, n$.

Consider the space of p-dimensional measurable random functions $\boldsymbol{h}$ of $\boldsymbol{Z}$ with zero mean and finite second moment

$$\mathcal{H} := \left\{ \boldsymbol{h}(\boldsymbol{Z}) \mid \boldsymbol{h}(\cdot) : \Omega \to \mathbb{R}^p,\ \mathrm{E}[\boldsymbol{h}(\boldsymbol{Z})] = \mathbf{0},\ \mathrm{E}[\boldsymbol{h}(\boldsymbol{Z})^T \boldsymbol{h}(\boldsymbol{Z})] < \infty \right\}. \tag{7.2}$$

For notational convenience, elements of $\mathcal{H}$ will be written as $\boldsymbol{h}_i$ corresponding to $\boldsymbol{h}_i(\boldsymbol{Z})$. We define an inner product as

$$\langle \boldsymbol{h}_1, \boldsymbol{h}_2 \rangle = \mathrm{E}\left(\boldsymbol{h}_1^T \boldsymbol{h}_2\right) = \mathrm{E}\left[\boldsymbol{h}_1(\boldsymbol{Z})^T \boldsymbol{h}_2(\boldsymbol{Z})\right],$$

referred to as the *covariance inner product*. Note that $\langle \boldsymbol{h}_1, \boldsymbol{h}_1 \rangle$ is a scalar and thus does not correspond to the covariance matrix which is given by $\mathrm{E}\left(\boldsymbol{h}_1 \boldsymbol{h}_1^T\right)$. One can show that $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ forms a Hilbert space; see, for example, Loève (1963). The origin of the Hilbert space is given by the function $\boldsymbol{h}$ for which $\boldsymbol{h}(\boldsymbol{Z}) = \boldsymbol{0}$. Condition 4 in Definition 7 is fulfilled for the covariance inner product if we define an equivalence class where $\boldsymbol{h}_1$ is equivalent to $\boldsymbol{h}_2$, i.e. $\boldsymbol{h}_1 \equiv \boldsymbol{h}_2$, if $\mathrm{P}\left[\boldsymbol{h}_1(\boldsymbol{Z}) \neq \boldsymbol{h}_2(\boldsymbol{Z})\right] = 0$.

For a Hilbert space we can define a linear subspace.

Definition 8 (Linear subspace)**.** *A space $\mathcal{U} \subset \mathcal{H}$ is a linear subspace if for all $\boldsymbol{u}_1, \boldsymbol{u}_2 \in \mathcal{U}$ and $\lambda_1, \lambda_2 \in \mathbb{R}$ it follows that $\lambda_1 \boldsymbol{u}_1 + \lambda_2 \boldsymbol{u}_2 \in \mathcal{U}$.*

A linear subspace is *closed* if it contains all its limit points. An important result in Hilbert spaces is given by the projection theorem. We refer to Luenberger (1969) for a proof.

Theorem 14 (Projection theorem)**.** *Let $\mathcal{U}$ denote a closed linear subspace of the Hilbert space $\mathcal{H}$, then for all $\boldsymbol{h} \in \mathcal{H}$, there exists a unique $\boldsymbol{u}_0 \in \mathcal{U}$ so that*

$$\|\boldsymbol{h} - \boldsymbol{u}_0\| \leq \|\boldsymbol{h} - \boldsymbol{u}\|, \quad \textit{for all} \quad \boldsymbol{u} \in \mathcal{U},$$

i.e. $\boldsymbol{u}_0$ is the unique element of $\mathcal{U}$ closest to $\boldsymbol{h}$. Furthermore,

$$\langle \boldsymbol{h} - \boldsymbol{u}_0, \boldsymbol{u} \rangle = 0, \quad \textit{for all} \quad \boldsymbol{u} \in \mathcal{U},$$

i.e. $\boldsymbol{h} - \boldsymbol{u}_0$ is orthogonal to $\mathcal{U}$. We denote this unique projection of $\boldsymbol{h}$ on $\mathcal{U}$, i.e. $\boldsymbol{u}_0$, as $\Pi(\boldsymbol{h} \mid \mathcal{U})$.

7.2.2 Review on parametric theory

Before introducing the theory of semiparametric models, we first introduce some results related to parametric models.

Definition 9 (Parametric model)**.** *The class of densities corresponding to a parametric model, $\mathcal{M}_P \subset \mathcal{M}$, can be described as*

$$\mathcal{M}_P = \left\{ f_{\boldsymbol{Z}}(\boldsymbol{z}; \boldsymbol{\theta}) \mid \boldsymbol{\theta}^T = (\boldsymbol{\beta}^T, \boldsymbol{\eta}^T),\ \boldsymbol{\theta}^T \in \Theta \subset \mathbb{R}^{p+r} \right\},$$

i.e. the density can be described by a finite p-dimensional parameter $\boldsymbol{\beta}$ and a finite r-dimensional nuisance parameter $\boldsymbol{\eta}$.

For both the parametric and semiparametric model, interest lies in estimating $\boldsymbol{\beta}$. If $\boldsymbol{\theta}_0^T = (\boldsymbol{\beta}_0^T, \boldsymbol{\eta}_0^T)$ denotes the truth (i.e. the model parameters corresponding to the density that generates the data), then we restrict the discussion to estimators for $\boldsymbol{\beta}$ which are asymptotically linear.

Definition 10 (Asymptotically linear estimator)**.** *An asymptotically linear (AL) estimator of $\boldsymbol{\beta}$, say $\hat{\boldsymbol{\beta}}_n$, is a p-dimensional measurable function of the sample $\{\boldsymbol{Z}_i \mid i = 1, \ldots, n\}$, so that there exists a p-dimensional measurable random function $\boldsymbol{\varphi}(\boldsymbol{Z})$, such that*

1. $\mathrm{E}[\boldsymbol{\varphi}(\boldsymbol{Z})] = \mathbf{0}$,
2. $\mathrm{E}[\boldsymbol{\varphi}(\boldsymbol{Z})\boldsymbol{\varphi}(\boldsymbol{Z})^T]$ *is finite and non-singular,*
3. *and*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{\varphi}(\boldsymbol{Z}_i) + o_p(1).$$

The function $\boldsymbol{\varphi}(\cdot)$ is named the *influence function*. Note that $\boldsymbol{\varphi}(\boldsymbol{Z}_i)$ measures the influence of the i^{th} observation on the estimator $\hat{\boldsymbol{\beta}}_n$. The following theorem states that influence functions are unique; a proof can be found in Tsiatis (2006, p. 23).

Theorem 15. *An AL estimator has a unique influence function, i.e. if $\boldsymbol{\varphi}_1$ and $\boldsymbol{\varphi}_2$ denote two influence functions associated with the AL estimator $\hat{\boldsymbol{\beta}}_n$, then*

$$\mathrm{P}\left(\boldsymbol{\varphi}_1 = \boldsymbol{\varphi}_2\right) = 1.$$

There exist AL estimators which are *super-efficient*, i.e. which are asymptotically unbiased and which have an asymptotic variance smaller than the Cràmer-Rao lower bound for some parameter values. For an example, see section 3.1 in Tsiatis (2006). These estimators, however, are unnatural and therefore we try to avoid them. This can be accomplished by defining regular estimators.

Definition 11 (Regular estimator)**.** *Consider a local data generating process, where, for each n, the data are distributed according to $\boldsymbol{\theta}_n = (\boldsymbol{\beta}_n^T, \boldsymbol{\eta}_n^T)^T$, where for some fixed parameter $\boldsymbol{\theta}^* = (\boldsymbol{\beta}^{*T}, \boldsymbol{\eta}^{*T})^T$, it holds that*

$$\sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}^*) \to \boldsymbol{c},$$

for some vector of constants $\mathbf{c}$*. Thus* $\{\boldsymbol{Z}_{i,n} \mid i = 1, \ldots, n\}$ *are i.i.d. with density* $f_{\boldsymbol{Z}}(\boldsymbol{z}; \boldsymbol{\theta}_n)$*. An estimator* $\hat{\boldsymbol{\beta}}_n$*, depending on* $\boldsymbol{Z}_{i,n}$ $(i = 1, \ldots, n)$*, is* regular *if, for each* $\boldsymbol{\theta}^*$*,*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n) \xrightarrow{d} F,$$

where F *does not depend on the local data generating process.*

For the remainder of this chapter we focus on *regular asymptotically linear (RAL)* estimators. The notion of a score vector will help us in finding RAL estimators.

Definition 12 (Score vector)**.** *The score vector evaluated in a fixed point* $\boldsymbol{\theta}_0$*, is defined as*

$$\boldsymbol{S}_{\boldsymbol{\theta}}(\boldsymbol{Z}; \boldsymbol{\theta}_0) := \frac{\partial \log f_{\boldsymbol{Z}}(\boldsymbol{Z}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}.$$

We can rewrite the score vector as $\boldsymbol{S}_{\boldsymbol{\theta}}(\boldsymbol{z}; \boldsymbol{\theta}_0)^T = (\boldsymbol{S}_{\boldsymbol{\beta}}(\boldsymbol{z}; \boldsymbol{\theta}_0)^T, \boldsymbol{S}_{\boldsymbol{\eta}}(\boldsymbol{z}; \boldsymbol{\theta}_0)^T)$, where

$$\boldsymbol{S}_{\boldsymbol{\beta}}(\boldsymbol{z}; \boldsymbol{\theta}_0) := \frac{\partial \log f_{\boldsymbol{Z}}(\boldsymbol{z}; \boldsymbol{\theta})}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}, \quad \boldsymbol{S}_{\boldsymbol{\eta}}(\boldsymbol{z}; \boldsymbol{\theta}_0) := \frac{\partial \log f_{\boldsymbol{Z}}(\boldsymbol{z}; \boldsymbol{\theta})}{\partial \boldsymbol{\eta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}.$$

The following theorem presents a result that allows describing the geometry of influence functions for RAL estimators. We refer to Tsiatis (2006, p. 28) for a proof.

Theorem 16. *Consider an AL estimator* $\hat{\boldsymbol{\beta}}_n$ *with influence function* $\boldsymbol{\varphi}(\boldsymbol{Z})$ *such that* $\mathrm{E}[\boldsymbol{\varphi}(\boldsymbol{Z})^T \boldsymbol{\varphi}(\boldsymbol{Z})]$ *exists and* $\mathrm{E}[\boldsymbol{\varphi}(\boldsymbol{Z})^T \boldsymbol{\varphi}(\boldsymbol{Z})]$ *is continuous in* $\boldsymbol{\theta}$ *in a neighbourhood of* $\boldsymbol{\theta}_0$*. If* $\hat{\boldsymbol{\beta}}_n$ *is RAL estimator, then*

$$\mathrm{E}\left[\boldsymbol{\varphi}(\boldsymbol{Z}) \boldsymbol{S}_{\boldsymbol{\beta}}(\boldsymbol{Z}; \boldsymbol{\theta}_0)^T\right] = \boldsymbol{I}, \tag{7.3}$$

and

$$\mathrm{E}\left[\boldsymbol{\varphi}(\boldsymbol{Z}) \boldsymbol{S}_{\boldsymbol{\eta}}(\boldsymbol{Z}; \boldsymbol{\theta}_0)^T\right] = \boldsymbol{0}. \tag{7.4}$$

One can also show that an element $\boldsymbol{\varphi} \in \mathcal{H}$ satisfying (7.3) and (7.4) is the influence function of some RAL estimator; see, for example, section 3.3 in Tsiatis (2006).

Sometimes it can be more natural to consider the parameter of interest $\boldsymbol{\beta}$ as a smooth p-dimensional function of $\boldsymbol{\theta}$, i.e. $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{\theta})$. The previous theorem can be generalized to this setting; see, for example, Theorem 3.2 in Tsiatis (2006).

Thus we can identify a RAL estimator $\hat{\boldsymbol{\beta}}_n$ through its influence function $\boldsymbol{\varphi}$ and the asymptotic distribution of $\boldsymbol{\varphi}$ determines the asymptotic distribution of $\hat{\boldsymbol{\beta}}_n$. Indeed, since $\boldsymbol{\varphi}(\boldsymbol{Z}_i)$,

$i = 1, \ldots, n$ are i.i.d. with finite non-singular second moment, the central limit theorem guarantees that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{\varphi}(\boldsymbol{Z}_i) \xrightarrow{d} \mathrm{N}[\mathbf{0}, \mathrm{E}\left(\boldsymbol{\varphi}\boldsymbol{\varphi}^T\right)].$$

If (some of) the predictors are fixed by design so that $\boldsymbol{Z}_i$ and hence $\boldsymbol{\varphi}(\boldsymbol{Z}_i)$ are not i.i.d., then the Lindeberg–Feller central limit theorem can be used to establish the asymptotic normality; see, for example, van der Vaart (1998, p. 20).

Since $\hat{\boldsymbol{\beta}}_n$ is asymptotically linear and by using Slutksy's lemma, it follows that

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{d} \mathrm{N}[\mathbf{0}, \mathrm{E}\left(\boldsymbol{\varphi}\boldsymbol{\varphi}^T\right)].$$

Different RAL estimators for $\boldsymbol{\beta}$, say $\hat{\boldsymbol{\beta}}_n^1, \hat{\boldsymbol{\beta}}_n^2, \ldots$ may differ in their asymptotic variance $\mathrm{E}\left(\boldsymbol{\varphi}_i\boldsymbol{\varphi}_i^T\right)$, where $\boldsymbol{\varphi}_i$ is the influence function of the RAL estimator $\hat{\boldsymbol{\beta}}_n^i$, $i = 1, 2, \ldots$ Therefore, if we have a set of candidate RAL estimators, we can choose the estimator with the smallest asymptotic variance, i.e. the most efficient estimator. To obtain this, two problems need to be addressed.

1. How do we extend the notion of smallest variance if the variance is as $p \times p$ covariance matrix associated with a p-dimensional estimator?
2. How can we find the RAL estimator with smallest asymptotic variance?

To answer these questions we first need to introduce some definitions and theorems. We start by defining a tangent space.

Definition 13 (Tangent space)**.** *The tangent space is defined as*

$$\mathcal{T} := \{\boldsymbol{B}\boldsymbol{S}_{\boldsymbol{\theta}}(\boldsymbol{Z}, \boldsymbol{\theta}_0) \mid \forall \textit{ real } p \times (p + r) \textit{ matrices } \boldsymbol{B}\}.$$

Definition 14 (Nuisance tangent space)**.** *The nuisance tangent space is defined as*

$$\Lambda := \{\boldsymbol{B}\boldsymbol{S}_{\boldsymbol{\eta}}(\boldsymbol{Z}, \boldsymbol{\theta}_0) \mid \forall \textit{ real } p \times r \textit{ matrices } \boldsymbol{B}\}.$$

From a geometrical point of view, condition (7.4) states that $\boldsymbol{\varphi}$ is orthogonal to the linear subspace Λ.

Definition 15 (Direct sum)**.** *If M and N denote two linear subspaces in $\mathcal{H}$, then $M \oplus N$ is a direct sum of M and N if $M \oplus N$ is a linear subspace in $\mathcal{H}$ and every element $\boldsymbol{x} \in M \oplus N$ has a unique representation of the form $\boldsymbol{x} = \boldsymbol{m} + \boldsymbol{n}$, where $\boldsymbol{m} \in M$ and $\boldsymbol{n} \in N$.*

Definition 16 (Orthogonal linear subspace)**.** *The set of elements in a Hilbert space which are orthogonal to a linear subspace M, is denoted as $M^{\perp}$ and referred to as the orthogonal complement of M.*

It holds that $M^{\perp}$ is a linear subspace and that $\mathcal{H} = M \oplus M^{\perp}$. If $\mathcal{T}_{\boldsymbol{\beta}}$ denotes the linear subspace

$$\mathcal{T}_{\boldsymbol{\beta}} := \{\boldsymbol{B}\boldsymbol{S}_{\boldsymbol{\beta}}(\boldsymbol{Z}, \boldsymbol{\theta}_0) \mid \forall \text{ real } p \times p \text{ matrices } \boldsymbol{B}\},$$

then one can show that

$$\mathcal{T} = \mathcal{T}_{\boldsymbol{\beta}} \oplus \Lambda.$$

The following definition allows generalizing the notion of efficiency to a multivariate setting.

Definition 17 (Asymptotic variance in multiple dimensions)**.** *Consider two influence functions for $\boldsymbol{\beta}$, say $\boldsymbol{\varphi}_1$ and $\boldsymbol{\varphi}_2$, both p-dimensional, then*

$$\operatorname{Cov}(\boldsymbol{\varphi}_1) \leq \operatorname{Cov}(\boldsymbol{\varphi}_2) \Leftrightarrow \forall \boldsymbol{a} \in \mathbb{R}^p : \ \operatorname{Var}\left(\boldsymbol{a}^T \boldsymbol{\varphi}_1\right) \leq \operatorname{Var}\left(\boldsymbol{a}^T \boldsymbol{\varphi}_2\right).$$

Hence $\operatorname{Cov}(\boldsymbol{\varphi}_1) \leq \operatorname{Cov}(\boldsymbol{\varphi}_2)$ *is equivalent to saying that* $\operatorname{E}\left(\boldsymbol{\varphi}_2 \boldsymbol{\varphi}_2^T\right) - \operatorname{E}\left(\boldsymbol{\varphi}_1 \boldsymbol{\varphi}_1^T\right)$ *is nonnegative definite.*

The Pythagorean theorem is crucial for finding the most efficient influence function. The theorem can be extended to multiple dimensions, for which we first need to introduce p-replicating linear spaces.

Definition 18 (p-replicating linear space)**.** *A linear subspace $\mathcal{U} \subset \mathcal{H}$ is a p-replicating linear space if $\mathcal{U}$ can be written as*

$$\mathcal{U} = \mathcal{U}^1 \times \cdots \times \mathcal{U}^1 = \left\{\mathcal{U}^1\right\}^p,$$

where $\mathcal{U}^1$ denotes a linear subspace in the Hilbert space of one-dimensional mean-zero random function of $\boldsymbol{Z}$ and where $\{\mathcal{U}^1\}^p$ consists of elements $\mathbf{h}$ such that

$$\boldsymbol{h}^T = (h_1, \ldots, h_p), \quad h_i \in \mathcal{U}^1.$$

Both $\mathcal{T}$ and Λ are p-replicating linear spaces; see, for example, Tsiatis (2006, p. 44). For these p-dimensional linear spaces the Pythagorean theorem can be extended.

Theorem 17 (Multivariate Pythagorean theorem)**.** *Let $\boldsymbol{h} \in \mathcal{U} \subset \mathcal{H}$ with $\mathcal{U}$ a p-replicating linear space, and let $\boldsymbol{h}' \in \mathcal{H}$ denote a element orthogonal to $\mathcal{U}$, i.e. $\langle \boldsymbol{h}, \boldsymbol{h}' \rangle = 0$, then*

$$\mathrm{Cov}\left(\boldsymbol{h} + \boldsymbol{h}'\right) = \mathrm{Cov}\left(\boldsymbol{h}\right) + \mathrm{Cov}\left(\boldsymbol{h}'\right).$$

Definition 19 (Linear variety)**.** *A translation of a linear subspace away from the origin is called a linear variety, say V. A linear variety can be written as $V = \boldsymbol{x} + M$, where $\boldsymbol{x} \in \mathcal{H}$, $\boldsymbol{x} \notin M$, $\|\boldsymbol{x}\| \neq 0$, with M a linear subspace.*

The following theorem gives the set of all influence functions. We refer to Tsiatis (2006, p. 45) for a proof.

Theorem 18. *The set of all influence functions of the parameter $\boldsymbol{\beta}$, i.e. elements of $\mathcal{H}$ satisfying conditions (7.3) and (7.4), corresponds to the linear variety*

$$\boldsymbol{\varphi}^*(\boldsymbol{Z}) + \mathcal{T}^{\perp},$$

where $\boldsymbol{\varphi}^*(\boldsymbol{Z})$ is any influence function of the parameter $\boldsymbol{\beta}$.

All these results allow us to find the most efficient influence function and hence the RAL estimator for $\boldsymbol{\beta}$ with smallest asymptotic variance. Consider an arbitrary influence function $\boldsymbol{\varphi}$ and define $\boldsymbol{l}_{eff} := \boldsymbol{\varphi} - \Pi(\boldsymbol{\varphi} \mid \mathcal{T})$. From Theorem 14 if follows that $\boldsymbol{l}_{eff} \in \mathcal{T}^{\perp}$. Define $\boldsymbol{\varphi}_{eff} := \Pi(\boldsymbol{\varphi} \mid \mathcal{T})$. By definition $\boldsymbol{\varphi}_{eff} \in \mathcal{T}$ so that $\langle \boldsymbol{\varphi}_{eff}, \boldsymbol{l}_{eff} \rangle = 0$. Since $\boldsymbol{\varphi}_{eff} = \boldsymbol{\varphi} - \boldsymbol{l}_{eff}$ it follows by Theorem 18 that $\boldsymbol{\varphi}_{eff}$ is a proper influence function. Because $\langle \boldsymbol{\varphi}_{eff}, \boldsymbol{l}_{eff} \rangle = 0$ and $\boldsymbol{\varphi} = \boldsymbol{\varphi}_{eff} + \boldsymbol{l}_{eff}$, the multivariate Pythagorean theorem guarantees that $\mathrm{Cov}\left(\boldsymbol{\varphi}\right) = \mathrm{Cov}\left(\boldsymbol{\varphi}_{eff}\right) + \mathrm{Cov}\left(\boldsymbol{l}_{eff}\right)$ and hence $\mathrm{Cov}\left(\boldsymbol{\varphi}_{eff}\right) \leq \mathrm{Cov}\left(\boldsymbol{\varphi}\right)$ in the multivariate sense. Since this holds for an arbitrary $\boldsymbol{\varphi}$, $\boldsymbol{\varphi}_{eff}$ is the efficient influence function. The following theorem gives an explicit formulation of $\boldsymbol{\varphi}_{eff}$.

Theorem 19. *The efficient influence function is given by*

$$\boldsymbol{\varphi}_{eff}(\boldsymbol{Z}) = \mathrm{E}\left(\boldsymbol{S}_{eff}(\boldsymbol{Z}; \boldsymbol{\theta}_0) \boldsymbol{S}_{eff}(\boldsymbol{Z}; \boldsymbol{\theta}_0)^T\right)^{-1} \boldsymbol{S}_{eff}(\boldsymbol{Z}; \boldsymbol{\theta}_0),$$

with $\boldsymbol{S}_{eff}(\boldsymbol{Z}; \boldsymbol{\theta}_0)$ the efficient score, given by

$$\boldsymbol{S}_{eff}(\boldsymbol{Z}; \boldsymbol{\theta}_0) = \boldsymbol{S}_{\boldsymbol{\beta}}(\boldsymbol{Z}; \boldsymbol{\theta}_0) - \Pi(\boldsymbol{S}_{\boldsymbol{\beta}}(\boldsymbol{Z}; \boldsymbol{\theta}_0) \mid \Lambda),$$

i.e. the residual of the score vector with respect to $\boldsymbol{\beta}$ after projecting it onto the nuisance tangent space.

The proof can be found in Tsiatis (2006, p. 46).

Once we have identified the set of influence functions associated with $\boldsymbol{\beta}$, the corresponding estimators can be constructed as follows. Let $\boldsymbol{\varphi}(\cdot; \boldsymbol{\beta}, \boldsymbol{\xi})$ denote an influence function (we explicitly state its dependence on the parameter of interest $\boldsymbol{\beta}$ and on an $\tilde{r}$-dimensional nuisance parameter $\boldsymbol{\xi}$ where $\tilde{r} \leq r$, with r the dimension of $\boldsymbol{\eta}$), then an estimator of $\boldsymbol{\beta}$ can be obtained by solving the set of equations

$$n^{-1} \sum_{i=1}^{n} \boldsymbol{\varphi}(\boldsymbol{Z}_i; \boldsymbol{\beta}, \hat{\boldsymbol{\xi}}) = \mathbf{0},$$

where $\hat{\boldsymbol{\xi}}$ denotes a *first-step* estimator, obtained by solving

$$n^{-1} \sum_{i=1}^{n} \boldsymbol{\psi}(\boldsymbol{Z}_i; \boldsymbol{\xi}) = \mathbf{0},$$

for some $\tilde{r}$-dimensional vector function $\boldsymbol{\psi}(\cdot)$. Chapter 6 of Newey and McFadden (1994) describes the asymptotic properties of such a *two-step* estimator for $\boldsymbol{\beta}$. For example, they provide primitive conditions under which the estimation of $\boldsymbol{\xi}$ does not affect the asymptotic distribution of $\hat{\boldsymbol{\beta}}$. Since we model PIMs semiparametrically, we postpone this discussion to Section 7.3.4, where it is shown how semiparametric two-step estimators can be constructed starting from an influence function.

7.2.3 Review on semiparametric theory

In this section we extend the theory of parametric models $\mathcal{M}_P$ to semiparametric models $\mathcal{M}_{SP}$ (7.1) which is based on the notion of parametric submodels.

Definition 20 (Parametric submodel)**.** *Let $f_0(\boldsymbol{z}) := f_{\boldsymbol{Z}}(\boldsymbol{z}; \boldsymbol{\beta}_0, \boldsymbol{\eta}_0) \in \mathcal{M}_{SP}$ denote the truth, i.e. the density that generated the data. A class of densities, denoted as*

$$\mathcal{M}_{\boldsymbol{\beta},\boldsymbol{\gamma}} = \left\{ f_{\boldsymbol{Z}}(\boldsymbol{z}; \boldsymbol{\beta}, \boldsymbol{\gamma}) \mid (\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T) \in \Theta \subset \mathbb{R}^{p+r} \right\},$$

is a parametric submodel if

1. $\mathcal{M}_{\boldsymbol{\beta},\boldsymbol{\gamma}} \subset \mathcal{M}_{SP}$,
2. $f_0(\boldsymbol{z}) \in \mathcal{M}_{\boldsymbol{\beta},\boldsymbol{\gamma}}$.

Thus a parametric submodel of a semiparametric model consist of a subset of $\mathcal{M}_{SP}$ with finite-dimensional parameters and which contains the truth. This allows us to define a RAL estimator for a semiparametric model.

Definition 21 (Semiparametric RAL estimator)**.** *An estimator for $\boldsymbol{\beta}$ is a RAL estimator for a semiparametric model if it is a RAL estimator for every parametric submodel.*

Based on the parametric submodels we define the nuisance tangent space for a semiparametric model.

Definition 22 (Semiparametric nuisance tangent space)**.** *The nuisance tangent space for a semiparametric model, say Λ, is defined as the mean-square closure of parametric submodel nuisance tangent spaces $\Lambda_{\boldsymbol{\gamma}}$, where*

$$\Lambda_{\boldsymbol{\gamma}} = \{\boldsymbol{B}\boldsymbol{S}_{\boldsymbol{\gamma}}(\boldsymbol{Z};\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0) \mid \forall \textit{ real } p \times r \textit{ matrices } \boldsymbol{B}\},$$

with $\boldsymbol{S}_{\boldsymbol{\gamma}}(\boldsymbol{Z};\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0)$ the score vector for the nuisance parameter $\boldsymbol{\gamma}$ for some parametric submodel $\mathcal{M}_{\boldsymbol{\beta},\boldsymbol{\gamma}}$. If we index these parametric submodels by j, then

$$\Lambda := \left\{\boldsymbol{h} \in \mathcal{H} \mid \exists \textit{ a sequence } (\boldsymbol{B}_j\boldsymbol{S}_{\boldsymbol{\gamma},j}) \textit{ such that } \|\boldsymbol{h} - \boldsymbol{B}_j\boldsymbol{S}_{\boldsymbol{\gamma},j}\|^2 \overset{j\to\infty}{\longrightarrow} 0\right\}.$$

The semiparametric nuisance tangent space consists of the union of all parametric submodel nuisance tangent spaces together with all the limit points. In general, Λ is not necessarily a linear space, but in the remainder of this chapter, it will always be linear. The notion of Λ allows us to define the semiparametric efficient score vector.

Definition 23 (Semiparametric efficient score)**.** *The semiparametric efficient score for $\boldsymbol{\beta}$ is defined as*

$$\boldsymbol{S}_{eff}(\boldsymbol{Z};\boldsymbol{\beta}_0,\boldsymbol{\eta}_0) := \boldsymbol{S}_{\boldsymbol{\beta}}(\boldsymbol{Z};\boldsymbol{\beta}_0,\boldsymbol{\eta}_0) - \Pi(\boldsymbol{S}_{\boldsymbol{\beta}}(\boldsymbol{Z};\boldsymbol{\beta}_0,\boldsymbol{\eta}_0) \mid \Lambda).$$

Definition 24 (Semiparametric efficiency bound)**.** *If*

$$\boldsymbol{S}^{eff}_{\boldsymbol{\beta},\boldsymbol{\gamma}}(\boldsymbol{Z};\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0) = \boldsymbol{S}_{\boldsymbol{\beta}}(\boldsymbol{Z};\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0) - \Pi(\boldsymbol{S}_{\boldsymbol{\beta}}(\boldsymbol{Z};\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0) \mid \Lambda_{\boldsymbol{\gamma}}),$$

denotes the efficient score of a parametric submodel, then we can define the semiparametric efficiency bound as

$$\sup_{(\textit{all parametric submodels } \mathcal{M}_{\boldsymbol{\beta},\boldsymbol{\gamma}})} \mathrm{E}\left(\boldsymbol{S}^{eff}_{\boldsymbol{\beta},\boldsymbol{\gamma}}\boldsymbol{S}^{eff^T}_{\boldsymbol{\beta},\boldsymbol{\gamma}}\right)^{-1}.$$

The following theorem relates the semiparametric efficiency bound to the semiparametric efficient score; we refer to Tsiatis (2006, p. 64) for a proof.

Theorem 20. *The semiparametric efficiency bound is equal to the inverse of the variance matrix of the semiparametric efficient score, i.e.*

$$\sup_{(\textit{all parametric submodels } \mathcal{M}_{\boldsymbol{\beta},\boldsymbol{\gamma}})} \mathrm{E}\left(\boldsymbol{S}^{eff}_{\boldsymbol{\beta},\boldsymbol{\gamma}}\boldsymbol{S}^{eff^T}_{\boldsymbol{\beta},\boldsymbol{\gamma}}\right)^{-1} = \mathrm{E}\left(\boldsymbol{S}_{eff}\boldsymbol{S}^T_{eff}\right)^{-1}.$$

Definition 25 (Semiparametric efficient influence function)**.** *The influence function of a semiparametric RAL estimator that achieves the semiparametric efficiency bound, if it exists, is named the efficient influence function.*

Similar as for the parametric models, we can characterize the influence functions based on the score vectors; see Tsiatis (2006, p. 66) for a proof.

Theorem 21. *Any semiparametric RAL estimator for $\boldsymbol{\beta}$ must have an influence function $\boldsymbol{\varphi}$ that satisfies*

1. $\mathrm{E}[\boldsymbol{\varphi}(\boldsymbol{Z})\boldsymbol{S}^T_{\boldsymbol{\beta}}(\boldsymbol{Z};\boldsymbol{\beta}_0,\boldsymbol{\eta}_0)] = \mathrm{E}[\boldsymbol{\varphi}(\boldsymbol{Z})\boldsymbol{S}^T_{eff}(\boldsymbol{Z};\boldsymbol{\beta}_0,\boldsymbol{\eta}_0)] = \boldsymbol{I}$,
2. $\Pi(\boldsymbol{\varphi}(\boldsymbol{Z}) \mid \Lambda) = \boldsymbol{0}$, *i.e.* $\boldsymbol{\varphi}(\boldsymbol{Z}) \in \Lambda^{\perp}$.

The efficient influence function is the unique element satisfying conditions 1 and 2 and whose variance-covariance matrix equals the efficiency bound, and is equal to

$$\boldsymbol{\varphi}_{eff}(\boldsymbol{Z};\boldsymbol{\beta}_0,\boldsymbol{\eta}_0) = \mathrm{E}\left(\boldsymbol{S}_{eff}\boldsymbol{S}^T_{eff}\right)^{-1}\boldsymbol{S}_{eff}(\boldsymbol{Z};\boldsymbol{\beta}_0,\boldsymbol{\eta}_0).$$

Theorem 22. *If a semiparametric RAL estimator for $\boldsymbol{\beta}$ exists, then the influence function of this estimator must belong to the linear variety*

$$\boldsymbol{\varphi}(\boldsymbol{Z}) + \mathcal{T}^{\perp},$$

with $\boldsymbol{\varphi}(\boldsymbol{Z})$ the influence function of any semiparametric RAL estimator for $\boldsymbol{\beta}$ and $\mathcal{T}$ the semiparametric tangent space, i.e. the mean-square closure of all parametric submodel tangent spaces. If the semiparametric efficient estimator exists, then the influence function must be the unique and well-defined element

$$\boldsymbol{\varphi}_{eff}(\boldsymbol{Z}) = \boldsymbol{\varphi}(\boldsymbol{Z}) - \Pi(\boldsymbol{\varphi}(\boldsymbol{Z}) \mid \mathcal{T}^{\perp}) = \Pi(\boldsymbol{\varphi}(\boldsymbol{Z}) \mid \mathcal{T}).$$

7.3 Semiparametric theory for probabilistic index models

We now use the theory of the previous section to find the semiparametric efficient estimator associated with a PIM. We start by expressing a PIM as a restricted moment model

$$\mathrm{P}\left(Y \preccurlyeq Y' \mid \boldsymbol{X}, \boldsymbol{X}'\right) = \mathrm{E}\left(\mathrm{I}\left(Y \preccurlyeq Y'\right) \mid \boldsymbol{X}, \boldsymbol{X}'\right) = m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}). \tag{7.5}$$

The major difference with the more conventional restricted model, as given by (2.5), is that the definition of a PIM involves a couple of observations $\boldsymbol{Z} = (Y, \boldsymbol{X})$ and $\boldsymbol{Z}' = (Y', \boldsymbol{X}')$. This makes the theory for semiparametric restricted moment models as described in section 4.5 of Tsiatis (2006) not directly applicable. We will, however, follow a similar strategy. In the remainder of this section we consider PIMs which are defined for the no-order restriction $\mathcal{X}_0 = \{(\boldsymbol{X}_i, \boldsymbol{X}_j) \mid i, j = 1, \ldots, n\}$ and for which the model satisfies $m(\boldsymbol{X}_i, \boldsymbol{X}_j; \boldsymbol{\beta}) + m(\boldsymbol{X}_j, \boldsymbol{X}_i; \boldsymbol{\beta}) = 1$. All results, however, can be extended to other order restrictions $\mathcal{X}_n$.

7.3.1 The semiparametric model

The model restriction (7.5) is equivalent to

$$\int \mathrm{I}\left(y \preccurlyeq y'\right) f_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x}) f_{Y|\boldsymbol{X}}(y' \mid \boldsymbol{x}') dy dy' = m(\boldsymbol{x}, \boldsymbol{x}'; \boldsymbol{\beta}),$$

and since $\int f_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x}) f_{Y|\boldsymbol{X}}(y' \mid \boldsymbol{x}') dy dy' = 1$, (7.5) is equivalent to

$$\int [\mathrm{I}\left(y \preccurlyeq y'\right) - m(\boldsymbol{x}, \boldsymbol{x}'; \boldsymbol{\beta})] f_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x}) f_{Y|\boldsymbol{X}}(y' \mid \boldsymbol{x}') dy dy' = 0.$$

We use the latter expression to characterize the semiparametric model associated with a PIM. More specifically, let $\mathcal{M}_{SP}^{PIM} \subset \mathcal{M}$ denote the class of densities for which

$$\begin{aligned} \mathcal{M}_{SP}^{PIM} : \; = \; & \Big\{ f_{\boldsymbol{Z}}(\boldsymbol{z}; \boldsymbol{\beta}, \boldsymbol{\eta}) = f_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x}; \boldsymbol{\beta}, \boldsymbol{\eta}_1) f_{\boldsymbol{X}}(\boldsymbol{x}; \boldsymbol{\eta}_2) \mid \boldsymbol{\beta} \in \Theta \subset \mathbb{R}^p \quad \text{and} \\ & \int [\mathrm{I}\left(y \preccurlyeq y'\right) - m(\boldsymbol{x}, \boldsymbol{x}'; \boldsymbol{\beta})] f_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x}; \boldsymbol{\beta}, \boldsymbol{\eta}_1) f_{Y|\boldsymbol{X}}(y' \mid \boldsymbol{x}'; \boldsymbol{\beta}, \boldsymbol{\eta}_1) dy dy' = 0 \Big\}. \end{aligned}$$

We denote the truth as

$$f_0(\boldsymbol{z}) = f_{\boldsymbol{Z}}(\boldsymbol{z}; \boldsymbol{\beta}_0, \boldsymbol{\eta}_0) = f_0(y \mid \boldsymbol{x}) f_0(\boldsymbol{x}) = f_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x}; \boldsymbol{\beta}_0, \boldsymbol{\eta}_{10}) f_{\boldsymbol{X}}(\boldsymbol{x}; \boldsymbol{\eta}_{20}),$$

where $\boldsymbol{\eta}_0^T = (\boldsymbol{\eta}_{10}^T, \boldsymbol{\eta}_{20}^T)$. Let $\mathcal{M}_{\boldsymbol{\beta},\boldsymbol{\gamma}_1,\boldsymbol{\gamma}_2} \subset \mathcal{M}_{SP}^{PIM}$ denote a parametric submodel of $\mathcal{M}_{SP}^{PIM}$, which we characterize as

$$\begin{aligned}\mathcal{M}_{\boldsymbol{\beta},\boldsymbol{\gamma}_1,\boldsymbol{\gamma}_2} \;=\; \Big\{ & f_{\boldsymbol{Z}}(\boldsymbol{z};\boldsymbol{\beta},\boldsymbol{\gamma}) = f_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x};\boldsymbol{\beta},\boldsymbol{\gamma}_1) f_{\boldsymbol{X}}(\boldsymbol{x};\boldsymbol{\gamma}_2) \mid \\ & (\boldsymbol{\beta}^T, \boldsymbol{\gamma}_1^T, \boldsymbol{\gamma}_2^T)^T \in \Theta_{\boldsymbol{\beta},\boldsymbol{\gamma}} \subset \mathbb{R}^{p+r_1+r_2} \quad \text{and} \\ & \int [\mathrm{I}\,(y \preccurlyeq y') - m(\boldsymbol{x},\boldsymbol{x}';\boldsymbol{\beta})] f_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x};\boldsymbol{\beta},\boldsymbol{\gamma}_1) f_{Y|\boldsymbol{X}}(y' \mid \boldsymbol{x}';\boldsymbol{\beta},\boldsymbol{\gamma}_1) dy dy' = 0 \Big\},\end{aligned}$$

where $\boldsymbol{\gamma}_1$ is an r_1-dimensional vector and $\boldsymbol{\gamma}_2$ an r_2-dimensional vector, and $\boldsymbol{\gamma}^T = (\boldsymbol{\gamma}_1^T, \boldsymbol{\gamma}_2^T)$ an r-dimensional vector with $r = r_1 + r_2$. As before, the truth is denoted as $(\boldsymbol{\beta}_0^T, \boldsymbol{\gamma}_0^T)$ and is assumed to be contained within the parametric submodel. Since proper densities in the parametric submodel can be defined for any combination of $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$, we say that $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ are variationally independent.

Consider the parametric submodel nuisance score vector

$$\boldsymbol{S}_{\boldsymbol{\gamma}}(\boldsymbol{Z};\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0) = \left(\boldsymbol{S}_{\boldsymbol{\gamma}_1}(\boldsymbol{Z};\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0)^T, \boldsymbol{S}_{\boldsymbol{\gamma}_2}(\boldsymbol{Z};\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0)^T\right)^T,$$

where

$$\boldsymbol{S}_{\boldsymbol{\gamma}_1}(\boldsymbol{Z};\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0) = \frac{\partial \log f_{Y|\boldsymbol{X}}(Y \mid \boldsymbol{X};\boldsymbol{\beta}_0,\boldsymbol{\gamma}_1)}{\partial \boldsymbol{\gamma}_1}\Big|_{\boldsymbol{\gamma}_1=\boldsymbol{\gamma}_{10}},$$

and

$$\boldsymbol{S}_{\boldsymbol{\gamma}_2}(\boldsymbol{Z};\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0) = \frac{\partial \log f_{\boldsymbol{X}}(\boldsymbol{X};\boldsymbol{\gamma}_2)}{\partial \boldsymbol{\gamma}_2}\Big|_{\boldsymbol{\gamma}_2=\boldsymbol{\gamma}_{20}}.$$

The former score does not depend on $\boldsymbol{\gamma}_{20}$, therefore we write $\boldsymbol{S}_{\boldsymbol{\gamma}_1}(\boldsymbol{Z};\boldsymbol{\beta}_0,\boldsymbol{\gamma}_{10})$, and similarly, the latter score does not depend on Y, $\boldsymbol{\beta}_0$, and $\boldsymbol{\gamma}_{10}$, therefore we write $\boldsymbol{S}_{\boldsymbol{\gamma}_2}(\boldsymbol{X};\boldsymbol{\gamma}_{20})$.

The parametric submodel nuisance tangent space

$$\Lambda_{\boldsymbol{\gamma}} := \{\boldsymbol{B}\boldsymbol{S}_{\boldsymbol{\gamma}}(\boldsymbol{Z};\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0) \mid \forall \text{ real } p \times r \text{ matrices } \boldsymbol{B}\},$$

can be written as a direct sum

$$\Lambda_{\boldsymbol{\gamma}} = \Lambda_{\boldsymbol{\gamma}_1} \oplus \Lambda_{\boldsymbol{\gamma}_2},$$

where

$$\Lambda_{\boldsymbol{\gamma}_1} := \left\{\boldsymbol{B}\boldsymbol{S}_{\boldsymbol{\gamma}_1}(\boldsymbol{Z};\boldsymbol{\beta}_0,\boldsymbol{\gamma}_{10}) \mid \forall \text{ real } p \times r_1 \text{ matrices } \boldsymbol{B}\right\}, \tag{7.6}$$

and

$$\Lambda_{\boldsymbol{\gamma}_2} := \left\{\boldsymbol{B}\boldsymbol{S}_{\boldsymbol{\gamma}_2}(\boldsymbol{X};\boldsymbol{\gamma}_{20}) \mid \forall \text{ real } p \times r_2 \text{ matrices } \boldsymbol{B}\right\}. \tag{7.7}$$

Lemma 11. *The space $\Lambda_{\boldsymbol{\gamma}_1}$ as defined by (7.6) is orthogonal to the space $\Lambda_{\boldsymbol{\gamma}_2}$ as defined by (7.7).*

Proof. From

$$\int f_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x}; \boldsymbol{\beta}_0, \boldsymbol{\gamma}_1) dy = 1, \quad \forall \boldsymbol{x}, \boldsymbol{\gamma}_1,$$

it follows that

$$\frac{\partial}{\partial \boldsymbol{\gamma}_1} \int f_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x}; \boldsymbol{\beta}_0, \boldsymbol{\gamma}_1) dy \Big|_{\boldsymbol{\gamma}_1 = \boldsymbol{\gamma}_{10}} = \mathbf{0}. \tag{7.8}$$

Upon using the chain-rule and by interchanging integration and differentiation evaluated at $\boldsymbol{\gamma}_{10}$, it follows that

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\gamma}_1} \int f_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x}; \boldsymbol{\beta}_0, \boldsymbol{\gamma}_1) dy \Big|_{\boldsymbol{\gamma}_1 = \boldsymbol{\gamma}_{10}} &= \int \frac{\partial}{\partial \boldsymbol{\gamma}_1} f_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x}; \boldsymbol{\beta}_0, \boldsymbol{\gamma}_1) dy \Big|_{\boldsymbol{\gamma}_1 = \boldsymbol{\gamma}_{10}} \\ &= \int \boldsymbol{S}_{\boldsymbol{\gamma}_1}(y, \boldsymbol{x}; \boldsymbol{\beta}_0, \boldsymbol{\gamma}_{10}) f_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x}; \boldsymbol{\beta}_0, \boldsymbol{\gamma}_{10}) dy. \end{aligned}$$

Substituting this last expression in (7.8) leads to

$$\mathrm{E}[\boldsymbol{S}_{\boldsymbol{\gamma}_1}(Y, \boldsymbol{X}; \boldsymbol{\beta}_0, \boldsymbol{\gamma}_{10}) \mid \boldsymbol{X}] = \mathbf{0}.$$

Consider now an arbitrary element of $\Lambda_{\boldsymbol{\gamma}_1}$, say $\boldsymbol{B}_1 \boldsymbol{S}_{\boldsymbol{\gamma}_1}(Y, \boldsymbol{X}; \boldsymbol{\beta}_0, \boldsymbol{\gamma}_{10})$, and an arbitrary element of $\Lambda_{\boldsymbol{\gamma}_2}$, say $\boldsymbol{B}_2 \boldsymbol{S}_{\boldsymbol{\gamma}_2}(\boldsymbol{X}; \boldsymbol{\gamma}_{20})$, then it follows that

$$\begin{aligned} \langle \boldsymbol{B}_1 \boldsymbol{S}_{\boldsymbol{\gamma}_1}(Y, \boldsymbol{X}; \boldsymbol{\beta}_0, \boldsymbol{\gamma}_{10}), \boldsymbol{B}_2 \boldsymbol{S}_{\boldsymbol{\gamma}_2}(\boldsymbol{X}; \boldsymbol{\gamma}_{20}) \rangle &= \mathrm{E}\left[\boldsymbol{S}_{\boldsymbol{\gamma}_1}(Y, \boldsymbol{X}; \boldsymbol{\beta}_0, \boldsymbol{\gamma}_{10})^T \boldsymbol{B}_1^T \boldsymbol{B}_2 \boldsymbol{S}_{\boldsymbol{\gamma}_2}(\boldsymbol{X}; \boldsymbol{\gamma}_{20}) \right] \\ &= \mathrm{E}\left[\mathrm{E}\left(\boldsymbol{S}_{\boldsymbol{\gamma}_1}(Y, \boldsymbol{X}; \boldsymbol{\beta}_0, \boldsymbol{\gamma}_{10})^T \mid \boldsymbol{X} \right) \boldsymbol{B}_1^T \boldsymbol{B}_2 \right. \\ &\quad \left. \boldsymbol{S}_{\boldsymbol{\gamma}_2}(\boldsymbol{X}; \boldsymbol{\gamma}_{20}) \right] \\ &= 0. \end{aligned}$$

The second equality holds because of the law of iterated expectation. □

7.3.2 The semiparametric nuisance tangent space

The semiparametric nuisance tangent space Λ is defined as the mean-square closure of all parametric submodel nuisance tangent spaces $\Lambda_{\boldsymbol{\gamma}} = \Lambda_{\boldsymbol{\gamma}_1} \oplus \Lambda_{\boldsymbol{\gamma}_2}$. Since $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ are variationally independent it follows that

$$\Lambda = \Lambda_{1s} \oplus \Lambda_{2s}, \tag{7.9}$$

where Λ_{1s} is the mean-square closure of all $\Lambda_{\boldsymbol{\gamma}_1}$ and Λ_{2s} the mean-square closure of all $\Lambda_{\boldsymbol{\gamma}_2}$. Since we can write the nuisance tangent space as a direct sum of Λ_{1s} and Λ_{2s}, we explicitly derive the elements of these spaces.

Theorem 23 (The space Λ_{2s})**.** *The nuisance tangent space with respect to $\boldsymbol{\eta}_2$ is given by*

$$\Lambda_{2s} = \{\boldsymbol{h}(\boldsymbol{X}) \in \mathcal{H}\},$$

i.e. the space of all p-dimensional mean-zero measurable functions of $\boldsymbol{X}$ with finite second moment.

Proof. Consider an arbitrary element of any parametric submodel $\Lambda_{\boldsymbol{\gamma}_2}$, say $\boldsymbol{BS}_{\boldsymbol{\gamma}_2}(\boldsymbol{X}; \boldsymbol{\gamma}_{20})$. From

$$\int f_{\boldsymbol{X}}(\boldsymbol{x}; \boldsymbol{\gamma}_2) d\boldsymbol{x} = 1, \quad \forall \boldsymbol{\gamma}_2,$$

it follows that, by using similar arguments as in the proof of Lemma 11,

$$\mathrm{E}[\boldsymbol{BS}_{\boldsymbol{\gamma}_2}(\boldsymbol{X}; \boldsymbol{\gamma}_{20})] = \mathbf{0},$$

hence $\Lambda_{\boldsymbol{\gamma}_2} \subset \Lambda_{2s}$. We now show that an arbitrary element of Λ_{2s} is either an element of $\Lambda_{\boldsymbol{\gamma}_2}$ for some parametric submodel or a limit of such elements. Consider a bounded element $\boldsymbol{h}^*(\boldsymbol{X}) \in \Lambda_{2s}$ for which we construct the parametric submodel with density

$$f_{\boldsymbol{X}}(\boldsymbol{x}; \boldsymbol{\gamma}_2) = f_0(\boldsymbol{x})[1 + \boldsymbol{\gamma}_2^T \boldsymbol{h}^*(\boldsymbol{x})],$$

where $\boldsymbol{\gamma}_2$ is a p-dimensional vector so that

$$[1 + \boldsymbol{\gamma}_2^T \boldsymbol{h}^*(\boldsymbol{x})] \geq 0, \quad \forall \boldsymbol{x}.$$

Since $\boldsymbol{h}^*(\boldsymbol{X})$ is bounded such a $\boldsymbol{\gamma}_2$ exists. Thus $f_{\boldsymbol{X}}(\boldsymbol{x}; \boldsymbol{\gamma}_2)$ is nonnegative and since

$$\begin{aligned} \int f_{\boldsymbol{X}}(\boldsymbol{x}; \boldsymbol{\gamma}_2) d\boldsymbol{x} &= \int f_0(\boldsymbol{x}) d\boldsymbol{x} + \int \boldsymbol{\gamma}_2^T \boldsymbol{h}^*(\boldsymbol{x}) f_0(\boldsymbol{x}) d\boldsymbol{x} && (7.10) \\ &= 1 + \boldsymbol{\gamma}_2^T \mathrm{E}[\boldsymbol{h}^*(\boldsymbol{X})] = 1 + 0 = 1, && (7.11) \end{aligned}$$

it follows that $f_{\boldsymbol{X}}(\boldsymbol{x}; \boldsymbol{\gamma}_2)$ is a proper density function. It is now easy to see that the score vector for this parametric submodel is given by

$$\boldsymbol{S}_{\boldsymbol{\gamma}_2}(\boldsymbol{X}; \boldsymbol{\gamma}_{20}) = \boldsymbol{h}^*(\boldsymbol{X}),$$

hence $\boldsymbol{h}^*(\boldsymbol{X}) \in \Lambda_{\boldsymbol{\gamma}_2}$. Since arbitrary $\boldsymbol{h}^*(\boldsymbol{X}) \in \Lambda_{2s}$ can always be taken as a limit of bounded mean-zero functions of $\boldsymbol{X}$, the statement follows. □

The following theorem gives the elements of the space Λ_{1s}.

Theorem 24 (The space Λ_{1s})**.** *The space Λ_{1s} is the space of all p-dimensional random functions $\boldsymbol{h}(Y, \boldsymbol{X}) \in \mathcal{H}$ that satisfy both*

1. $\mathrm{E}\left(\boldsymbol{h}(Y, \boldsymbol{X}) \mid \boldsymbol{X}\right) = \mathbf{0}$.
2. $\mathrm{E}\left\{[\mathrm{I}\left(Y \preccurlyeq Y'\right) - m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}_0)][\boldsymbol{h}(Y, \boldsymbol{X}) + \boldsymbol{h}(Y', \boldsymbol{X}')] \mid \boldsymbol{X}, \boldsymbol{X}'\right\} = \mathbf{0}$ *with* $(Y, \boldsymbol{X})$ *and* $(Y', \boldsymbol{X}')$ *i.i.d.*

Proof. For an arbitrary element of any parametric submodel Λ_{γ_1}, say $\boldsymbol{BS}_{\gamma_1}(Y, \boldsymbol{X}; \boldsymbol{\beta}_0, \boldsymbol{\gamma}_{10})$, because $\int f_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x}; \boldsymbol{\beta}_0, \boldsymbol{\gamma}_1)dy = 1 \; \forall \boldsymbol{\gamma}_1$, it follows that, by using similar arguments as in the proof of Lemma 11,

$$\mathrm{E}\left[\boldsymbol{BS}_{\gamma_1}(Y, \boldsymbol{X}; \boldsymbol{\beta}_0, \boldsymbol{\gamma}_{10}) \mid \boldsymbol{X}\right] = \mathbf{0}.$$

Hence, every element of Λ_{γ_1} satisfies condition 1 of the theorem. Furthermore, the model restriction states that

$$\int [\mathrm{I}\left(y \preccurlyeq y'\right) - m(\boldsymbol{x}, \boldsymbol{x}'; \boldsymbol{\beta}_0)] \, f_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x}; \boldsymbol{\beta}_0, \boldsymbol{\gamma}_1) f_{Y|\boldsymbol{X}}(y' \mid \boldsymbol{x}'; \boldsymbol{\beta}_0, \boldsymbol{\gamma}_1) dy dy' = 0, \; \forall \boldsymbol{x}, \boldsymbol{x}', \boldsymbol{\gamma}_1.$$

Consequently

$$\begin{aligned}
& \frac{\partial}{\partial \boldsymbol{\gamma}_1} \int [\mathrm{I}\left(y \preccurlyeq y'\right) - m(\boldsymbol{x}, \boldsymbol{x}'; \boldsymbol{\beta}_0)] \, f_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x}; \boldsymbol{\beta}_0, \boldsymbol{\gamma}_1) f_{Y|\boldsymbol{X}}(y' \mid \boldsymbol{x}'; \boldsymbol{\beta}_0, \boldsymbol{\gamma}_1) dy dy' \Big|_{\boldsymbol{\gamma}_1 = \boldsymbol{\gamma}_{10}} \\
& = \mathbf{0} \\
\Leftrightarrow \quad & \int [\mathrm{I}\left(y \preccurlyeq y'\right) - m(\boldsymbol{x}, \boldsymbol{x}'; \boldsymbol{\beta}_0)] \, \frac{\partial}{\partial \boldsymbol{\gamma}_1} \left[f_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x}; \boldsymbol{\beta}_0, \boldsymbol{\gamma}_1) f_{Y|\boldsymbol{X}}(y' \mid \boldsymbol{x}'; \boldsymbol{\beta}_0, \boldsymbol{\gamma}_1) \right] \Big|_{\boldsymbol{\gamma}_1 = \boldsymbol{\gamma}_{10}} dy dy' \\
& = \mathbf{0} \\
\Leftrightarrow \quad & \int [\mathrm{I}\left(y \preccurlyeq y'\right) - m(\boldsymbol{x}, \boldsymbol{x}'; \boldsymbol{\beta}_0)] \left[\boldsymbol{S}_{\gamma_1}(y, \boldsymbol{x}; \boldsymbol{\beta}_0, \boldsymbol{\gamma}_{10}) + \boldsymbol{S}_{\gamma_1}(y', \boldsymbol{x}'; \boldsymbol{\beta}_0, \boldsymbol{\gamma}_{10}) \right] \\
& f_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x}; \boldsymbol{\beta}_0, \boldsymbol{\gamma}_{10}) f_{Y|\boldsymbol{X}}(y' \mid \boldsymbol{x}'; \boldsymbol{\beta}_0, \boldsymbol{\gamma}_{10}) dy dy' = \mathbf{0}.
\end{aligned}$$

Since this holds for all $(\boldsymbol{x}, \boldsymbol{x}')$ it now follows that

$$\mathrm{E}\left\{ [\mathrm{I}\left(Y \preccurlyeq Y'\right) - m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}_0)] \left[\boldsymbol{S}_{\gamma_1}(Y, \boldsymbol{X}; \boldsymbol{\beta}_0, \boldsymbol{\gamma}_{10}) + \boldsymbol{S}_{\gamma_1}(Y', \boldsymbol{X}'; \boldsymbol{\beta}_0, \boldsymbol{\gamma}_{10}) \right] \mid \boldsymbol{X}, \boldsymbol{X}' \right\} = \mathbf{0},$$

and thus also

$$\mathrm{E}\left\{ [\mathrm{I}\left(Y \preccurlyeq Y'\right) - m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}_0)] \left[\boldsymbol{BS}_{\gamma_1}(Y, \boldsymbol{X}; \boldsymbol{\beta}_0, \boldsymbol{\gamma}_{10}) + \boldsymbol{BS}_{\gamma_1}(Y', \boldsymbol{X}'; \boldsymbol{\beta}_0, \boldsymbol{\gamma}_{10}) \right] \mid \boldsymbol{X}, \boldsymbol{X}' \right\} = \mathbf{0}.$$

Hence, every element of Λ_{γ_1} satisfies condition 2 of the theorem, and thus $\Lambda_{\gamma_1} \subset \Lambda_{1s}$. Now we will show that an arbitrary element of Λ_{1s} is either an element of Λ_{γ_1} for some parametric

submodel or a limit of such elements. Consider a bounded element $\boldsymbol{h}^*(Y, \boldsymbol{X}) \in \Lambda_{1s}$ for which we consider the parametric submodel $\Lambda_{\boldsymbol{\gamma}_1}$ with density

$$f_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x}; \boldsymbol{\beta}_0, \boldsymbol{\gamma}_1) = f_0(y \mid \boldsymbol{x}) \left[1 + \boldsymbol{\gamma}_1^T \boldsymbol{h}^*(y, \boldsymbol{x})\right], \tag{7.12}$$

where $\boldsymbol{\gamma}_1$ is a p-dimensional vector so that

$$\left[1 + \boldsymbol{\gamma}_1^T \boldsymbol{h}^*(y, \boldsymbol{x})\right] \geq 0, \quad \forall y, \boldsymbol{x}.$$

Similar as in the proof of Theorem 23 one can show that $f_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x}; \boldsymbol{\beta}_0, \boldsymbol{\gamma}_1)$ is a proper density function. To show that $\Lambda_{\boldsymbol{\gamma}_1}$ is a valid parametric submodel, the density (7.12) must satisfy the model restriction. Thus we need to show that

$$\int [\mathrm{I}\left(y \preccurlyeq y'\right) - m(\boldsymbol{x}, \boldsymbol{x}'; \boldsymbol{\beta}_0)] f_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x}; \boldsymbol{\beta}_0, \boldsymbol{\gamma}_1) f_{Y|\boldsymbol{X}}(y' \mid \boldsymbol{x}'; \boldsymbol{\beta}_0, \boldsymbol{\gamma}_1) dy dy' = 0. \tag{7.13}$$

We first discuss some intermediate results.

- Because $f_0(y \mid \boldsymbol{x})$ is the truth, it satisfies the model restriction

$$\int [\mathrm{I}\left(y \preccurlyeq y'\right) - m(\boldsymbol{x}, \boldsymbol{x}'; \boldsymbol{\beta}_0)] f_0(y \mid \boldsymbol{x}) f_0(y' \mid \boldsymbol{x}') dy dy' = 0. \tag{7.14}$$

- Because $\boldsymbol{h}^*(Y, \boldsymbol{X})$ is an element of Λ_{1s}, it follows that

$$\int [\mathrm{I}\left(y \preccurlyeq y'\right) - m(\boldsymbol{x}, \boldsymbol{x}'; \boldsymbol{\beta}_0)][\boldsymbol{h}^*(y, \boldsymbol{x}) + \boldsymbol{h}^*(y', \boldsymbol{x}')] f_0(y \mid \boldsymbol{x}) f_0(y' \mid \boldsymbol{x}') dy dy' = 0. \tag{7.15}$$

- Since $f_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x}; \boldsymbol{\beta}_0, \boldsymbol{\gamma}_1)$ is a proper density function, it follows that the left hand side of (7.13) can be equivalently written as $\mathrm{P}_{\boldsymbol{\gamma}_1}(Y \preccurlyeq Y') - m(\boldsymbol{x}, \boldsymbol{x}'; \boldsymbol{\beta}_0)$ where we use the subscript $\boldsymbol{\gamma}_1$ to indicate that the probability operator is defined with respect to the density (7.12). Because both the probability and the model restriction are bounded by the unit-interval, it follows that the left hand side of expression (7.13) lies within the interval $[-1, 1]$. Therefore it holds that

$$\begin{aligned}
& -1 \leq \int [\mathrm{I}\left(y \preccurlyeq y'\right) - m(\boldsymbol{x}, \boldsymbol{x}'; \boldsymbol{\beta}_0)] f_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x}; \boldsymbol{\beta}_0, \boldsymbol{\gamma}_1) f_{Y|\boldsymbol{X}}(y' \mid \boldsymbol{x}'; \boldsymbol{\beta}_0, \boldsymbol{\gamma}_1) dy dy' \leq 1 \\
\Rightarrow \quad & \mathbf{0} \leq \frac{\partial}{\partial \boldsymbol{\gamma}_1} \int [\mathrm{I}\left(y \preccurlyeq y'\right) - m(\boldsymbol{x}, \boldsymbol{x}'; \boldsymbol{\beta}_0)] f_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x}; \boldsymbol{\beta}_0, \boldsymbol{\gamma}_1) f_{Y|\boldsymbol{X}}(y' \mid \boldsymbol{x}'; \boldsymbol{\beta}_0, \boldsymbol{\gamma}_1) dy dy' \leq \mathbf{0} \\
\Rightarrow \quad & \int [\mathrm{I}\left(y \preccurlyeq y'\right) - m(\boldsymbol{x}, \boldsymbol{x}'; \boldsymbol{\beta}_0)] \frac{\partial}{\partial \boldsymbol{\gamma}_1} [f_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x}; \boldsymbol{\beta}_0, \boldsymbol{\gamma}_1) f_{Y|\boldsymbol{X}}(y' \mid \boldsymbol{x}'; \boldsymbol{\beta}_0, \boldsymbol{\gamma}_1)] dy dy' \\
& = \mathbf{0}.
\end{aligned} \tag{7.16}$$

After substituting (7.12) and upon using (7.15), one can show that (7.16) is equivalent to

$$\int [\mathrm{I}\left(y \preccurlyeq y'\right) - m(\boldsymbol{x}, \boldsymbol{x}'; \boldsymbol{\beta}_0)] \boldsymbol{\gamma}_1^T \boldsymbol{h}^*(y, \boldsymbol{x}) \boldsymbol{\gamma}_1^T \boldsymbol{h}^*(y', \boldsymbol{x}') f_0(y \mid \boldsymbol{x}) f_0(y' \mid \boldsymbol{x}') dy dy' = 0. \tag{7.17}$$

Proving that (7.13) holds is now straightforward upon combing the results (7.14), (7.15), and (7.17) after substituting $f_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x}; \boldsymbol{\beta}, \boldsymbol{\gamma}_1)$ by (7.12) in (7.13). Thus $\Lambda_{\boldsymbol{\gamma}_1}$ is a valid parametric submodel.

It holds that

$$\boldsymbol{S}_{\boldsymbol{\gamma}_1}(Y, \boldsymbol{X}; \boldsymbol{\beta}_0, \boldsymbol{\gamma}_{10}) = \boldsymbol{h}^*(Y, \boldsymbol{X}).$$

Consequently $\boldsymbol{h}^*(Y, \boldsymbol{X}) \in \Lambda_{\boldsymbol{\gamma}_1}$. Since arbitrary $\boldsymbol{h}^*(Y, \boldsymbol{X}) \in \Lambda_{1s}$ can always be taken as a limit of bounded mean-zero functions of Y and $\boldsymbol{X}$, the statement follows. □

Similar as for the parametric submodel tangent spaces $\Lambda_{\boldsymbol{\gamma}_1}$ and $\Lambda_{\boldsymbol{\gamma}_2}$, the semiparametric tangent spaces Λ_{1s} and Λ_{2s} are orthogonal.

Lemma 12. *Λ_{1s} is orthogonal to Λ_{2s}.*

Proof. Consider two arbitrary elements $\boldsymbol{h}_1(Y, \boldsymbol{X}) \in \Lambda_{1s}$ and $\boldsymbol{h}_2(\boldsymbol{X}) \in \Lambda_{2s}$. It follows that

$$\begin{aligned} \langle \boldsymbol{h}_1(Y, \boldsymbol{X}), \boldsymbol{h}_2(\boldsymbol{X}) \rangle &= \mathrm{E}\left[\boldsymbol{h}_1(Y, \boldsymbol{X})^T \boldsymbol{h}_2(\boldsymbol{X})\right] \\ &= \mathrm{E}\left[\mathrm{E}\left(\boldsymbol{h}_1(Y, \boldsymbol{X})^T \mid \boldsymbol{X}\right) \boldsymbol{h}_2(\boldsymbol{X})\right] \\ &= \mathrm{E}\left[\boldsymbol{0}^T \boldsymbol{h}_2(\boldsymbol{X})\right] = 0. \end{aligned}$$

□

The semiparametric nuisance tangent space can be written as the direct sum (7.9). Furthermore, Λ_{1s} is the intersection of the two linear subspaces

$$\Lambda_{1sa} := \{\boldsymbol{h}(Y, \boldsymbol{X}) \in \mathcal{H} \mid \mathrm{E}\left(\boldsymbol{h}(Y, \boldsymbol{X}) \mid \boldsymbol{X}\right) = \boldsymbol{0}\},$$

and

$$\begin{aligned} \Lambda_{1sb} := \{&\boldsymbol{h}(Y, \boldsymbol{X}) \in \mathcal{H} \mid \mathrm{E}\left\{[\mathrm{I}\left(Y \preccurlyeq Y'\right) - m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}_0)]\left[\boldsymbol{h}(Y, \boldsymbol{X}) + \boldsymbol{h}(Y', \boldsymbol{X}')\right] \mid \boldsymbol{X}, \boldsymbol{X}'\right\} = \boldsymbol{0}, \\ &\text{with } (Y, \boldsymbol{X}) \text{ and } (Y', \boldsymbol{X}') \text{ i.i.d.}\}. \end{aligned}$$

Thus $\Lambda_{1s} = \Lambda_{1sa} \cap \Lambda_{1sb}$ and we can write the semiparametric nuisance tangent space as $\Lambda = (\Lambda_{1sa} \cap \Lambda_{1sb}) \oplus \Lambda_{2s}$.

The following three lemmas are useful for simplifying the expression of the space Λ.

Lemma 13. *It holds that* $\Lambda_{1sa} = \Lambda_{2s}^{\perp}$.

Proof. Consider an arbitrary $\boldsymbol{h}_1(Y, \boldsymbol{X}) \in \Lambda_{1sa}$ and an arbitrary $\boldsymbol{h}_2(\boldsymbol{X}) \in \Lambda_{2s}$, then

$$\begin{aligned} \langle \boldsymbol{h}_1(Y, \boldsymbol{X}), \boldsymbol{h}_2(\boldsymbol{X}) \rangle &= \mathrm{E}\left[\boldsymbol{h}_1(Y, \boldsymbol{X})^T \boldsymbol{h}_2(\boldsymbol{X})\right] \\ &= \mathrm{E}\left[\mathrm{E}\left(\boldsymbol{h}_1(Y, \boldsymbol{X})^T \mid \boldsymbol{X}\right) \boldsymbol{h}_2(\boldsymbol{X})\right] \\ &= \mathrm{E}\left[\boldsymbol{0}^T \boldsymbol{h}_2(\boldsymbol{X})\right] = 0, \end{aligned}$$

i.e. $\boldsymbol{h}_1 \in \Lambda_{2s}^{\perp}$, and thus $\Lambda_{1sa} \subseteq \Lambda_{2s}^{\perp}$. We now show that each element $\boldsymbol{h} \in \mathcal{H}$ can be written as $\boldsymbol{h} = \boldsymbol{h}_1 \oplus \boldsymbol{h}_2$, where $\boldsymbol{h}_1 \in \Lambda_{1sa}$ and $\boldsymbol{h}_2 \in \Lambda_{2s}$. This follows immediately by taking $\boldsymbol{h}_1 = \boldsymbol{h} - \mathrm{E}(\boldsymbol{h} \mid \boldsymbol{X})$ and $\boldsymbol{h}_2 = \mathrm{E}(\boldsymbol{h} \mid \boldsymbol{X})$, for which it is straightforward to show that $\boldsymbol{h} - \mathrm{E}(\boldsymbol{h} \mid \boldsymbol{X}) \in \Lambda_{1sa}$ and $\mathrm{E}(\boldsymbol{h} \mid \boldsymbol{X}) \in \Lambda_{2s}$. These results also imply that $\Pi(\boldsymbol{h} \mid \Lambda_{2s}) = \mathrm{E}(\boldsymbol{h} \mid \boldsymbol{X})$ and $\Pi(\boldsymbol{h} \mid \Lambda_{1sa}) = \boldsymbol{h} - \mathrm{E}(\boldsymbol{h} \mid \boldsymbol{X})$. □

Lemma 14. *It holds that* $\Lambda_{2s} \subset \Lambda_{1sb}$.

Proof. Consider an arbitrary $\boldsymbol{h}(\boldsymbol{X}) \in \Lambda_{2s}$, then

$$\begin{aligned} & \mathrm{E}\left\{\left[\mathrm{I}\left(Y \preccurlyeq Y'\right) - m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}_0)\right]\left[\boldsymbol{h}(\boldsymbol{X}) + \boldsymbol{h}(\boldsymbol{X}')\right] \mid \boldsymbol{X}, \boldsymbol{X}'\right\} \\ = \; & \left\{\mathrm{E}\left[\mathrm{I}\left(Y \preccurlyeq Y'\right) \mid \boldsymbol{X}, \boldsymbol{X}'\right] - m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}_0)\right\}\left[\boldsymbol{h}(\boldsymbol{X}) + \boldsymbol{h}(\boldsymbol{X}')\right] \\ = \; & 0[\boldsymbol{h}(\boldsymbol{X}) + \boldsymbol{h}(\boldsymbol{X}')] = \boldsymbol{0}. \end{aligned}$$

The last equality follows from the model restriction $\mathrm{E}\left(\mathrm{I}\left(Y \preccurlyeq Y'\right) \mid \boldsymbol{X}, \boldsymbol{X}'\right) = m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}_0)$. Hence $\boldsymbol{h}(\boldsymbol{X}) \in \Lambda_{1sb}$. □

Lemma 15. *It holds that* $\Lambda = \Lambda_{1sb}$.

Proof. It holds that $\Lambda = (\Lambda_{1sa} \cap \Lambda_{1sb}) \oplus \Lambda_{2s}$. Consider now an arbitrary element $\boldsymbol{h}_1 \in \Lambda_{1sa} \cap \Lambda_{1sb}$ and an arbitrary $\boldsymbol{h}_2 \in \Lambda_{2s}$. By definition $\boldsymbol{h}_1 \in \Lambda_{1sb}$ and because of Lemma 14 $\boldsymbol{h}_2 \in \Lambda_{1sb}$. Therefore, $\boldsymbol{h}_1 + \boldsymbol{h}_2 \in \Lambda_{1sb}$ since Λ_{1sb} is a linear space, so that $\Lambda \subseteq \Lambda_{1sb}$.

Consider an arbitrary $\boldsymbol{h} \in \Lambda_{1sb}$. Since for each $\boldsymbol{h} \in \mathcal{H}$, $\mathrm{E}[\mathrm{E}(\boldsymbol{h} \mid \boldsymbol{X})] = \mathrm{E}(\boldsymbol{h}) = \mathbf{0}$ it holds that $\mathrm{E}(\boldsymbol{h} \mid \boldsymbol{X}) \in \Lambda_{2s}$. Because of Lemma 14 it follows that $\mathrm{E}(\boldsymbol{h} \mid \boldsymbol{X}) \in \Lambda_{1sb}$. Because Λ_{1sb} is a linear space it follows that $\boldsymbol{h} - \mathrm{E}(\boldsymbol{h} \mid \boldsymbol{X}) \in \Lambda_{1sb}$. Thus $\boldsymbol{h}$ can be written as $\mathrm{E}(\boldsymbol{h} \mid \boldsymbol{X}) + [\boldsymbol{h} - \mathrm{E}(\boldsymbol{h} \mid \boldsymbol{X})]$ where $\mathrm{E}(\boldsymbol{h} \mid \boldsymbol{X}) \in \Lambda_{2s}$ and $[\boldsymbol{h} - \mathrm{E}(\boldsymbol{h} \mid \boldsymbol{X})] \in \Lambda_{1sb}$. However, $[\boldsymbol{h} - \mathrm{E}(\boldsymbol{h} \mid \boldsymbol{X})] \in \Lambda_{1sb}$ is also an element of Λ_{1sa} and hence $[\boldsymbol{h} - \mathrm{E}(\boldsymbol{h} \mid \boldsymbol{X})] \in (\Lambda_{1sa} \cap \Lambda_{1sb})$. Thus $\Lambda_{1sb} \subseteq \Lambda$. $\square$

In summary, the semiparametric nuisance tangent space is given by

$$\begin{aligned}\Lambda \;=\; \Big\{ & \boldsymbol{h}(Y, \boldsymbol{X}) \in \mathcal{H} \mid \mathrm{E}\left\{ [\mathrm{I}(Y \preccurlyeq Y') - m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}_0)] \, [\boldsymbol{h}(Y, \boldsymbol{X}) + \boldsymbol{h}(Y', \boldsymbol{X}')] \mid \boldsymbol{X}, \boldsymbol{X}' \right\} = \mathbf{0}, \\ & \text{with } (Y, \boldsymbol{X}) \text{ and } (Y', \boldsymbol{X}') \text{ i.i.d.} \Big\}.\end{aligned}$$

We now derive the space orthogonal to Λ which will provide us the influence functions of the RAL estimators for $\boldsymbol{\beta}$.

Theorem 25. *If holds that*

$$\begin{aligned}\Lambda^{\perp} \;=\; \Big\{ & \boldsymbol{h}(Y, \boldsymbol{X}) \in \mathcal{H} \mid \boldsymbol{h}(Y, \boldsymbol{X}) = \mathrm{E}\left(\boldsymbol{B}(\boldsymbol{X}, \boldsymbol{X}') \left[\mathrm{I}(Y \preccurlyeq Y') - m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}_0) \right] \mid Y, \boldsymbol{X} \right), \\ & \textit{for a } p\textit{-dimensional function } \boldsymbol{B}(\boldsymbol{X}, \boldsymbol{X}') \;\; \textit{such that} \\ & \boldsymbol{B}(\boldsymbol{X}, \boldsymbol{X}') + \boldsymbol{B}(\boldsymbol{X}', \boldsymbol{X}) = \mathbf{0} \Big\}.\end{aligned} \tag{7.18}$$

Moreover, the projection of an arbitrary $\boldsymbol{h} \in \mathcal{H}$ onto Λ is

$$\Pi(\boldsymbol{h} \mid \Lambda) = \boldsymbol{h}(Y, \boldsymbol{X}) - \mathrm{E}\left(\boldsymbol{B}_{\boldsymbol{h}}(\boldsymbol{X}, \boldsymbol{X}') \left[\mathrm{I}(Y \preccurlyeq Y') - m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}_0) \right] \mid Y, \boldsymbol{X} \right), \tag{7.19}$$

where $\boldsymbol{B}_{\boldsymbol{h}}(\boldsymbol{X}, \boldsymbol{X}')$ is the solution of (7.22) as given in the proof, assuming it exists.

Proof. Throughout the proof we use the equalities $\mathrm{I}(Y \preccurlyeq Y') + \mathrm{I}(Y' \preccurlyeq Y) = 1$ and $m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}) + m(\boldsymbol{X}', \boldsymbol{X}; \boldsymbol{\beta}) = 1$. We first show that the spaces

$$\begin{aligned}\mathcal{S}_1 := \{ \boldsymbol{S}(\boldsymbol{X}, \boldsymbol{X}') \;\; \mid \;\; & \exists \text{ a } p\text{-dimensional function } \boldsymbol{A}(\boldsymbol{X}, \boldsymbol{X}') \text{ such that} \\ & \boldsymbol{S}(\boldsymbol{X}, \boldsymbol{X}') = \boldsymbol{A}(\boldsymbol{X}, \boldsymbol{X}') - \boldsymbol{A}(\boldsymbol{X}', \boldsymbol{X}) \},\end{aligned}$$

and

$$\begin{aligned}\mathcal{S}_2 := \{ \boldsymbol{B}(\boldsymbol{X}, \boldsymbol{X}') \;\; \mid \;\; & \boldsymbol{B}(\boldsymbol{X}, \boldsymbol{X}') \text{ is } p\text{-dimensional and} \\ & \boldsymbol{B}(\boldsymbol{X}, \boldsymbol{X}') + \boldsymbol{B}(\boldsymbol{X}', \boldsymbol{X}) = \mathbf{0} \},\end{aligned}$$

are equal. Indeed, consider an arbitrary element $\boldsymbol{S}(\boldsymbol{X}, \boldsymbol{X}') \in \mathcal{S}_1$. There exists a function $\boldsymbol{A}(\boldsymbol{X}, \boldsymbol{X}')$ such that $\boldsymbol{S}(\boldsymbol{X}, \boldsymbol{X}') = \boldsymbol{A}(\boldsymbol{X}, \boldsymbol{X}') - \boldsymbol{A}(\boldsymbol{X}', \boldsymbol{X})$ and consequently $\boldsymbol{S}(\boldsymbol{X}, \boldsymbol{X}') + \boldsymbol{S}(\boldsymbol{X}', \boldsymbol{X}) = \boldsymbol{0}$, i.e. $\boldsymbol{S}(\boldsymbol{X}, \boldsymbol{X}') \in \mathcal{S}_2$. Consider an arbitrary $\boldsymbol{B}(\boldsymbol{X}, \boldsymbol{X}') \in \mathcal{S}_2$, then it follows that $\boldsymbol{B}(\boldsymbol{X}, \boldsymbol{X}') = 0.5\boldsymbol{B}(\boldsymbol{X}, \boldsymbol{X}') - 0.5\boldsymbol{B}(\boldsymbol{X}', \boldsymbol{X})$, and thus $\boldsymbol{B}(\boldsymbol{X}, \boldsymbol{X}') \in \mathcal{S}_1$. Hence $\mathcal{S}_1 = \mathcal{S}_2$.

Consequently, the space (7.18) can be written as

$$\begin{aligned}\Lambda^{\perp} &= \Big\{\mathrm{E}\{[\boldsymbol{A}(\boldsymbol{X}, \boldsymbol{X}') - \boldsymbol{A}(\boldsymbol{X}', \boldsymbol{X})][\mathrm{I}\,(Y \preccurlyeq Y') - m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}_0)] \mid Y, \boldsymbol{X}\}, \\ &\quad \forall\, p\text{-dimensional functions } \boldsymbol{A}(\boldsymbol{X}, \boldsymbol{X}')\Big\}.\end{aligned}$$

We now show that this space is orthogonal to Λ. Consider an arbitrary $\boldsymbol{h}_1 \in \Lambda$ and $\boldsymbol{h}_2 := \mathrm{E}\left\{[\boldsymbol{A}(\boldsymbol{X}, \boldsymbol{X}') - \boldsymbol{A}(\boldsymbol{X}', \boldsymbol{X})]\left[\mathrm{I}\,(Y \preccurlyeq Y') - m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}_0)\right] \mid Y, \boldsymbol{X}\right\}$ for some arbitrary p-dimensional function $\boldsymbol{A}(\boldsymbol{X}, \boldsymbol{X}')$, i.e. $\boldsymbol{h}_2 \in \Lambda^{\perp}$. We show that $\langle \boldsymbol{h}_1, \boldsymbol{h}_2 \rangle = 0$. It follows that

$$\begin{aligned}\langle \boldsymbol{h}_1, \boldsymbol{h}_2 \rangle &= \mathrm{E}\left(\boldsymbol{h}_1(Y, \boldsymbol{X})^T \mathrm{E}\left\{[\boldsymbol{A}(\boldsymbol{X}, \boldsymbol{X}') - \boldsymbol{A}(\boldsymbol{X}', \boldsymbol{X})]\left[\mathrm{I}\,(Y \preccurlyeq Y') - m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}_0)\right] \mid Y, \boldsymbol{X}\right\}\right) \\ &= \mathrm{E}\left[\boldsymbol{h}_1(Y, \boldsymbol{X})^T \mathrm{E}\left(\boldsymbol{A}(\boldsymbol{X}, \boldsymbol{X}')\left[\mathrm{I}\,(Y \preccurlyeq Y') - m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}_0)\right] \mid Y, \boldsymbol{X}\right)\right] + \\ &\quad \mathrm{E}\left[\boldsymbol{h}_1(Y, \boldsymbol{X})^T \mathrm{E}\left(\boldsymbol{A}(\boldsymbol{X}', \boldsymbol{X})\left[-\mathrm{I}\,(Y \preccurlyeq Y') + m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}_0)\right] \mid Y, \boldsymbol{X}\right)\right] \\ &= \mathrm{E}\left[\boldsymbol{h}_1(Y, \boldsymbol{X})^T \mathrm{E}\left(\boldsymbol{A}(\boldsymbol{X}, \boldsymbol{X}')\left[\mathrm{I}\,(Y \preccurlyeq Y') - m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}_0)\right] \mid Y, \boldsymbol{X}\right)\right] + \\ &\quad \mathrm{E}\left[\boldsymbol{h}_1(Y, \boldsymbol{X})^T \mathrm{E}\left(\boldsymbol{A}(\boldsymbol{X}', \boldsymbol{X})\left[\mathrm{I}\,(Y' \preccurlyeq Y) - m(\boldsymbol{X}', \boldsymbol{X}; \boldsymbol{\beta}_0)\right] \mid Y, \boldsymbol{X}\right)\right] \\ &= \mathrm{E}\left[\boldsymbol{h}_1(Y, \boldsymbol{X})^T \mathrm{E}\left(\boldsymbol{A}(\boldsymbol{X}, \boldsymbol{X}')\left[\mathrm{I}\,(Y \preccurlyeq Y') - m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}_0)\right] \mid Y, \boldsymbol{X}\right)\right] + \\ &\quad \mathrm{E}\left[\boldsymbol{h}_1(Y', \boldsymbol{X}')^T \mathrm{E}\left(\boldsymbol{A}(\boldsymbol{X}, \boldsymbol{X}')\left[\mathrm{I}\,(Y \preccurlyeq Y') - m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}_0)\right] \mid Y', \boldsymbol{X}'\right)\right] \\ &= \mathrm{E}\left\{\boldsymbol{h}_1(Y, \boldsymbol{X})^T \boldsymbol{A}(\boldsymbol{X}, \boldsymbol{X}')\left[\mathrm{I}\,(Y \preccurlyeq Y') - m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}_0)\right]\right\} + \\ &\quad \mathrm{E}\left\{\boldsymbol{h}_1(Y', \boldsymbol{X}')^T \boldsymbol{A}(\boldsymbol{X}, \boldsymbol{X}')\left[\mathrm{I}\,(Y \preccurlyeq Y') - m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}_0)\right]\right\} \\ &= \mathrm{E}\left\{\left[\boldsymbol{h}_1(Y, \boldsymbol{X})^T + \boldsymbol{h}_1(Y', \boldsymbol{X}')^T\right] \boldsymbol{A}(\boldsymbol{X}, \boldsymbol{X}')\left[\mathrm{I}\,(Y \preccurlyeq Y') - m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}_0)\right]\right\}.\end{aligned}$$

Let h_{1i} denote the i^{th} element of $\boldsymbol{h}_1$ and $A_i(\boldsymbol{X}, \boldsymbol{X}')$ the i^{th} element of $\boldsymbol{A}(\boldsymbol{X}, \boldsymbol{X}')$. Since $\boldsymbol{h}_1 \in \Lambda$ it follows that

$$\mathrm{E}\left\{[h_{1i}(Y, \boldsymbol{X}) + h_{1i}(Y', \boldsymbol{X}')]\left[\mathrm{I}\,(Y \preccurlyeq Y') - m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}_0)\right] \mid \boldsymbol{X}, \boldsymbol{X}'\right\} = 0,\ i = 1, \ldots, p.$$

Consequently

$$\begin{aligned}&\mathrm{E}\left\{\left[\boldsymbol{h}_1(Y, \boldsymbol{X})^T + \boldsymbol{h}_1(Y', \boldsymbol{X}')^T\right] \boldsymbol{A}(\boldsymbol{X}, \boldsymbol{X}')\left[\mathrm{I}\,(Y \preccurlyeq Y') - m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}_0)\right]\right\} \\ = \ &\mathrm{E}\left(\mathrm{E}\left\{\left[\boldsymbol{h}_1(Y, \boldsymbol{X})^T + \boldsymbol{h}_1(Y', \boldsymbol{X}')^T\right] \boldsymbol{A}(\boldsymbol{X}, \boldsymbol{X}')\left[\mathrm{I}\,(Y \preccurlyeq Y') - m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}_0)\right] \mid \boldsymbol{X}, \boldsymbol{X}'\right\}\right) \\ = \ &\mathrm{E}\left(\sum_{i=1}^{p} A_i(\boldsymbol{X}, \boldsymbol{X}')\mathrm{E}\left\{[h_{1i}(Y, \boldsymbol{X}) + h_{1i}(Y', \boldsymbol{X}')]\left[\mathrm{I}\,(Y \preccurlyeq Y') - m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}_0)\right] \mid \boldsymbol{X}, \boldsymbol{X}'\right\}\right) \\ = \ &0,\end{aligned}$$

and thus $\langle \boldsymbol{h}_1, \boldsymbol{h}_2 \rangle = 0$. To show that the space (7.18) is the orthogonal complement of Λ, we must show that an arbitrary element $\boldsymbol{h} \in \mathcal{H}$ can be written as $\boldsymbol{h}_1 + \boldsymbol{h}_2$, where $\boldsymbol{h}_1$ is an element of (7.18) and $\boldsymbol{h}_2 \in \Lambda$. This is equivalent to saying that for each $\boldsymbol{h} \in \mathcal{H}$ there exists a function $\boldsymbol{B_h}(\boldsymbol{X}, \boldsymbol{X}')$ such that

$$\boldsymbol{h}^*(Y, \boldsymbol{X}) := [\boldsymbol{h}(Y, \boldsymbol{X}) - \mathrm{E}\left(\boldsymbol{B_h}(\boldsymbol{X}, \boldsymbol{X}')\left[\mathrm{I}\left(Y \preccurlyeq Y'\right) - m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}_0)\right] \mid Y, \boldsymbol{X}\right)] \in \Lambda, \tag{7.20}$$

for which $\boldsymbol{B_h}(\boldsymbol{X}, \boldsymbol{X}') + \boldsymbol{B_h}(\boldsymbol{X}', \boldsymbol{X}) = \boldsymbol{0}$. For notational convenience we write

$$\varepsilon(\boldsymbol{Z}, \boldsymbol{Z}') := \mathrm{I}\left(Y \preccurlyeq Y'\right) - m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}_0). \tag{7.21}$$

Since $\boldsymbol{h}^*(Y, \boldsymbol{X})$ must be in Λ, it must satisfy

$$\mathrm{E}\left(\varepsilon(\boldsymbol{Z}, \boldsymbol{Z}')\left[\boldsymbol{h}^*(Y, \boldsymbol{X}) + \boldsymbol{h}^*(Y', \boldsymbol{X}')\right] \mid \boldsymbol{X}, \boldsymbol{X}'\right) = \boldsymbol{0}.$$

Through (7.20) this equation depends on $\boldsymbol{B_h}(\boldsymbol{X}, \boldsymbol{X}')$. In particular, upon using (7.20), this equation becomes

$$\begin{aligned}
& \mathrm{E}\left(\varepsilon(\boldsymbol{Z}, \boldsymbol{Z}')\left[\boldsymbol{h}(Y, \boldsymbol{X}) + \boldsymbol{h}(Y', \boldsymbol{X}')\right] \mid \boldsymbol{X}, \boldsymbol{X}'\right) = \\
& \mathrm{E}\left(\varepsilon(\boldsymbol{Z}, \boldsymbol{Z}')\mathrm{E}\left(\boldsymbol{B_h}(\boldsymbol{X}, \boldsymbol{X}^*)\varepsilon(\boldsymbol{Z}, \boldsymbol{Z}^*) \mid \boldsymbol{Z}\right) \mid \boldsymbol{X}, \boldsymbol{X}'\right) + \\
& \mathrm{E}\left(\varepsilon(\boldsymbol{Z}, \boldsymbol{Z}')\mathrm{E}\left(\boldsymbol{B_h}(\boldsymbol{X}', \boldsymbol{X}^*)\varepsilon(\boldsymbol{Z}', \boldsymbol{Z}^*) \mid \boldsymbol{Z}'\right) \mid \boldsymbol{X}, \boldsymbol{X}'\right) \\
\Leftrightarrow \quad & \mathrm{E}\left(\varepsilon(\boldsymbol{Z}, \boldsymbol{Z}')\left[\boldsymbol{h}(Y, \boldsymbol{X}) + \boldsymbol{h}(Y', \boldsymbol{X}')\right] \mid \boldsymbol{X}, \boldsymbol{X}'\right) = \\
& \mathrm{E}\left(\varepsilon(\boldsymbol{Z}, \boldsymbol{Z}')\left[\boldsymbol{B_h}(\boldsymbol{X}, \boldsymbol{X}^*)\varepsilon(\boldsymbol{Z}, \boldsymbol{Z}^*) + \boldsymbol{B_h}(\boldsymbol{X}', \boldsymbol{X}^*)\varepsilon(\boldsymbol{Z}', \boldsymbol{Z}^*)\right] \mid \boldsymbol{X}, \boldsymbol{X}'\right),
\end{aligned} \tag{7.22}$$

subject to $\boldsymbol{B_h}(\boldsymbol{X}, \boldsymbol{X}') + \boldsymbol{B_h}(\boldsymbol{X}', \boldsymbol{X}) = \boldsymbol{0}$. If such a $\boldsymbol{B_h}(\boldsymbol{X}, \boldsymbol{X}')$ exists, the theorem follows.

□

According to Theorem 21, influence functions of RAL estimators for $\boldsymbol{\beta}$ are orthogonal to Λ, thus are elements of $\Lambda^\perp$. However, not all elements of $\Lambda^\perp$ correspond to influence functions, because according to Theorem 21, they must also satisfy

$$\mathrm{E}[\boldsymbol{\varphi}(Y, \boldsymbol{X})\boldsymbol{S}_{\boldsymbol{\beta}}^T(Y, \boldsymbol{X}; \boldsymbol{\beta}_0, \boldsymbol{\eta}_0)] = \boldsymbol{I},$$

i.e. they need to be properly normalized. Thus an arbitrary influence function is given by

$$\boldsymbol{\varphi}(Y, \boldsymbol{X}) = \boldsymbol{C}\mathrm{E}\left(\boldsymbol{B}(\boldsymbol{X}, \boldsymbol{X}')\left[\mathrm{I}\left(Y \preccurlyeq Y'\right) - m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}_0)\right] \mid Y, \boldsymbol{X}\right),$$

with $\boldsymbol{B}(\boldsymbol{X},\boldsymbol{X}')+\boldsymbol{B}(\boldsymbol{X}',\boldsymbol{X})=\mathbf{0}$ and $\boldsymbol{C}$ the normalization factor. It is straightforward to see that the normalization factor is given by

$$\boldsymbol{C}=\mathrm{E}\left[\boldsymbol{B}(\boldsymbol{X},\boldsymbol{X}')[\mathrm{I}\left(Y\preccurlyeq Y'\right)-m(\boldsymbol{X},\boldsymbol{X}';\boldsymbol{\beta}_0)]\boldsymbol{S}^T_{\boldsymbol{\beta}}(Y,\boldsymbol{X};\boldsymbol{\beta}_0,\boldsymbol{\eta}_0)\right]^{-1}.$$

In Section 7.3.4 we show how the elements of the space $\Lambda^{\perp}$ can be used to construct estimating equations for RAL estimator.

7.3.3 The efficient influence function

We take a first initiative towards deriving the efficient estimator. In the previous section we have shown that an arbitrary influence function for $\boldsymbol{\beta}$ is given by

$$\boldsymbol{\varphi}(Y,\boldsymbol{X})=\boldsymbol{C}\mathrm{E}\left(\boldsymbol{B}(\boldsymbol{X},\boldsymbol{X}')\left[\mathrm{I}\left(Y\preccurlyeq Y'\right)-m(\boldsymbol{X},\boldsymbol{X}';\boldsymbol{\beta}_0)\right]\mid Y,\boldsymbol{X}\right), \tag{7.23}$$

with

$$\boldsymbol{C}=\mathrm{E}\left[\boldsymbol{B}(\boldsymbol{X},\boldsymbol{X}')[\mathrm{I}\left(Y\preccurlyeq Y'\right)-m(\boldsymbol{X},\boldsymbol{X}';\boldsymbol{\beta}_0)]\boldsymbol{S}^T_{\boldsymbol{\beta}}(Y,\boldsymbol{X};\boldsymbol{\beta}_0,\boldsymbol{\eta}_0)\right]^{-1}.$$

Thus different choices of the index function $\boldsymbol{B}(\boldsymbol{X},\boldsymbol{X}')$ result in different influence functions. The following theorem gives the system of equations that needs to be solved to obtain the index function associated with the efficient influence function. First we introduce some notation. Let

$$\boldsymbol{D}(\boldsymbol{X},\boldsymbol{X}';\boldsymbol{\beta}_0)=\frac{\partial m(\boldsymbol{X},\boldsymbol{X}';\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}, \tag{7.24}$$

and

$$V(\boldsymbol{X},\boldsymbol{X}',\boldsymbol{X}'',\boldsymbol{X}''')=\mathrm{Cov}\left(\mathrm{I}\left(Y\preccurlyeq Y'\right),\mathrm{I}\left(Y''\preccurlyeq Y'''\right)\mid\boldsymbol{X},\boldsymbol{X}',\boldsymbol{X}'',\boldsymbol{X}'''\right). \tag{7.25}$$

Theorem 26. *The index function associated with the efficient influence function (7.23), say $\boldsymbol{B}_{eff}(\cdot)$, is the solution of*

$$\begin{aligned}\boldsymbol{D}(\boldsymbol{X},\boldsymbol{X}';\boldsymbol{\beta}_0) &= \mathrm{E}\left(\boldsymbol{B}(\boldsymbol{X},\boldsymbol{X}^*)V(\boldsymbol{X},\boldsymbol{X}',\boldsymbol{X},\boldsymbol{X}^*)+\right.\\ &\quad\left.\boldsymbol{B}(\boldsymbol{X}',\boldsymbol{X}^*)V(\boldsymbol{X},\boldsymbol{X}',\boldsymbol{X}',\boldsymbol{X}^*)\mid\boldsymbol{X},\boldsymbol{X}'\right),\end{aligned} \tag{7.26}$$

subject to $\boldsymbol{B}(\boldsymbol{X},\boldsymbol{X}')+\boldsymbol{B}(\boldsymbol{X}',\boldsymbol{X})=\mathbf{0}$ and assuming it exists, with $\boldsymbol{D}(\cdot)$ and $V(\cdot)$ as in (7.24) and (7.25), respectively.

Proof. By definition

$$\boldsymbol{S}_{eff}(Y, \boldsymbol{X}; \boldsymbol{\beta}_0, \boldsymbol{\eta}_0) = \boldsymbol{S}_{\boldsymbol{\beta}}(Y, \boldsymbol{X}; \boldsymbol{\beta}_0, \boldsymbol{\eta}_0) - \Pi(\boldsymbol{S}_{\boldsymbol{\beta}}(Y, \boldsymbol{X}; \boldsymbol{\beta}_0, \boldsymbol{\eta}_0) \mid \Lambda),$$

and by using Theorem 25 this becomes

$$\boldsymbol{S}_{eff}(Y, \boldsymbol{X}) = \mathrm{E}\left(\boldsymbol{B}_{eff}(\boldsymbol{X}, \boldsymbol{X}')\left[\mathrm{I}\left(Y \preccurlyeq Y'\right) - m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}_0)\right] \mid Y, \boldsymbol{X}\right), \tag{7.27}$$

for a function $\boldsymbol{B}_{eff}(\cdot)$ that satisfies (7.22) with $\boldsymbol{h}(Y, \boldsymbol{X}) = \boldsymbol{S}_{\boldsymbol{\beta}}(Y, \boldsymbol{X}; \boldsymbol{\beta}_0, \boldsymbol{\eta}_0)$ and $\boldsymbol{B}_{\boldsymbol{h}}(\cdot) = \boldsymbol{B}_{eff}(\cdot)$. The system of equations (7.22) can further simplified. From the model restriction it follows that

$$\int [\mathrm{I}\left(y \preccurlyeq y'\right) - m(\boldsymbol{x}, \boldsymbol{x}'; \boldsymbol{\beta})] f_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x}; \boldsymbol{\beta}, \boldsymbol{\eta}_{10}) f_{Y|\boldsymbol{X}}(y' \mid \boldsymbol{x}'; \boldsymbol{\beta}, \boldsymbol{\eta}_{10}) dy dy' = 0, \ \forall \boldsymbol{x}, \boldsymbol{x}', \boldsymbol{\beta},$$

so that

$$\begin{aligned}
& \frac{\partial}{\partial \boldsymbol{\beta}} \int [\mathrm{I}\left(y \preccurlyeq y'\right) - m(\boldsymbol{x}, \boldsymbol{x}'; \boldsymbol{\beta})] f_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x}; \boldsymbol{\beta}, \boldsymbol{\eta}_{10}) f_{Y|\boldsymbol{X}}(y' \mid \boldsymbol{x}'; \boldsymbol{\beta}, \boldsymbol{\eta}_{10}) dy dy' \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} = \mathbf{0} \\
\Leftrightarrow \quad & \frac{\partial m(\boldsymbol{x}, \boldsymbol{x}'; \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \\
& = \int [\mathrm{I}\left(y \preccurlyeq y'\right) - m(\boldsymbol{x}, \boldsymbol{x}'; \boldsymbol{\beta}_0)] \frac{\partial}{\partial \boldsymbol{\beta}} [f_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x}; \boldsymbol{\beta}, \boldsymbol{\eta}_{10}) f_{Y|\boldsymbol{X}}(y' \mid \boldsymbol{x}'; \boldsymbol{\beta}, \boldsymbol{\eta}_{10})]\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} dy dy'.
\end{aligned}$$

If we use the chain rule in the right hand side, it follows that

$$\frac{\partial m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} = \mathrm{E}\left\{\left[\mathrm{I}\left(Y \preccurlyeq Y'\right) - m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}_0)\right]\left[\boldsymbol{S}_{\boldsymbol{\beta}}(Y, \boldsymbol{X}) + \boldsymbol{S}_{\boldsymbol{\beta}}(Y', \boldsymbol{X}')\right] \mid \boldsymbol{X}, \boldsymbol{X}'\right\}, \tag{7.28}$$

where, for notation convenience, $\boldsymbol{S}_{\boldsymbol{\beta}}(Y, \boldsymbol{X}) = \boldsymbol{S}_{\boldsymbol{\beta}}(Y, \boldsymbol{X}; \boldsymbol{\beta}_0, \boldsymbol{\eta}_0)$.

Consequently, upon using notation (7.21) and (7.24) and plugging in (7.28) into (7.22) with $\boldsymbol{h}(Y, \boldsymbol{X}) = \boldsymbol{S}_{\boldsymbol{\beta}}(Y, \boldsymbol{X})$ and $\boldsymbol{B}_{\boldsymbol{h}}(\cdot) = \boldsymbol{B}_{eff}(\cdot)$, (7.22) is equivalent to

$$\begin{aligned}
& \boldsymbol{D}(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}_0) = \\
& \mathrm{E}\left(\varepsilon(\boldsymbol{Z}, \boldsymbol{Z}')[\boldsymbol{B}_{eff}(\boldsymbol{X}, \boldsymbol{X}^*)\varepsilon(\boldsymbol{Z}, \boldsymbol{Z}^*) + \boldsymbol{B}_{eff}(\boldsymbol{X}', \boldsymbol{X}^*)\varepsilon(\boldsymbol{Z}', \boldsymbol{Z}^*)] \mid \boldsymbol{X}, \boldsymbol{X}'\right).
\end{aligned} \tag{7.29}$$

The statement follows by using the law of iterated expectation and by recognizing that

$$\begin{aligned}
& \mathrm{E}\left(\varepsilon(\boldsymbol{Z}, \boldsymbol{Z}')\varepsilon(\boldsymbol{Z}, \boldsymbol{Z}^*) \mid \boldsymbol{X}, \boldsymbol{X}', \boldsymbol{X}^*\right) \\
= \ & \mathrm{E}\left\{[\mathrm{I}\left(Y \preccurlyeq Y'\right) - m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}_0)][\mathrm{I}\left(Y \preccurlyeq Y^*\right) - m(\boldsymbol{X}, \boldsymbol{X}^*; \boldsymbol{\beta}_0)] \mid \boldsymbol{X}, \boldsymbol{X}', \boldsymbol{X}^*\right\} \\
= \ & \mathrm{Cov}\left(\mathrm{I}\left(Y \preccurlyeq Y'\right), \mathrm{I}\left(Y \preccurlyeq Y^*\right) \mid \boldsymbol{X}, \boldsymbol{X}', \boldsymbol{X}^*\right) \\
= \ & V(\boldsymbol{X}, \boldsymbol{X}', \boldsymbol{X}, \boldsymbol{X}^*),
\end{aligned}$$

upon using notation (7.25) in the last equation.

□

Finding the most efficient influence function is not straightforward, since it requires

- solving a system of equations that involves a conditional expectation.
- modelling the covariance functions $V(\boldsymbol{X}, \boldsymbol{X}', \boldsymbol{X}, \boldsymbol{X}^*)$.

In Section 7.4 we consider a simplified setting where we solve this integral equation.

Note that for continuous Y,

$$\begin{aligned} V(\boldsymbol{X}, \boldsymbol{X}', \boldsymbol{X}, \boldsymbol{X}^*) &= \operatorname{Cov}\left(\mathrm{I}\left(Y < Y'\right), \mathrm{I}\left(Y < Y^*\right) \mid \boldsymbol{X}, \boldsymbol{X}', \boldsymbol{X}^*\right) \\ &= \mathrm{E}\left(\mathrm{I}\left(Y < Y'\right) \mathrm{I}\left(Y < Y^*\right) \mid \boldsymbol{X}, \boldsymbol{X}', \boldsymbol{X}^*\right) - \\ &\quad \mathrm{E}\left(\mathrm{I}\left(Y < Y'\right) \mid \boldsymbol{X}, \boldsymbol{X}'\right) \mathrm{E}\left(\mathrm{I}\left(Y < Y^*\right) \mid \boldsymbol{X}, \boldsymbol{X}^*\right) \\ &= \mathrm{P}\left(Y < \min(Y', Y^*) \mid \boldsymbol{X}, \boldsymbol{X}', \boldsymbol{X}^*\right) - m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}_0) m(\boldsymbol{X}, \boldsymbol{X}^*; \boldsymbol{\beta}_0). \end{aligned}$$

Consequently, finding the efficient influence function requires modelling $\mathrm{P}(Y < \min(Y', Y^*) \mid \boldsymbol{X}, \boldsymbol{X}', \boldsymbol{X}^*)$.

7.3.4 Semiparametric two-step estimators

Based on Theorem 25 and under regularity conditions, a consistent and asymptotically normally distributed estimator for $\boldsymbol{\beta}_0$ can be obtained as the solution of

$$\frac{1}{n} \sum_{i=1}^{n} \mathrm{E}\left(\boldsymbol{B}(\boldsymbol{X}_i, \boldsymbol{X}; \boldsymbol{\beta}) \left[\mathrm{I}\left(Y_i \preccurlyeq Y\right) - m(\boldsymbol{X}_i, \boldsymbol{X}; \boldsymbol{\beta})\right] \mid Y_i, \boldsymbol{X}_i\right) = \mathbf{0}, \tag{7.30}$$

for an arbitrary p-dimensional index function $\boldsymbol{B}(\cdot)$ subject to $\boldsymbol{B}(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}) + \boldsymbol{B}(\boldsymbol{X}', \boldsymbol{X}; \boldsymbol{\beta}) = \mathbf{0}$. For a proof we refer to the literature of Z-estimators; see, for example, Huber (1964); van der Vaart (1998); Stefanski and Boos (2002). However, since the estimating equation involves a conditional expectation it cannot be directly used in practice. We will first need to estimate this expectation. Consider a nuisance function $\boldsymbol{\alpha}(\cdot)$ and let

$$\boldsymbol{V}[\boldsymbol{Z}, \boldsymbol{\beta}, \boldsymbol{\alpha}(\cdot)] := \int \boldsymbol{B}(\boldsymbol{X}, \boldsymbol{x}; \boldsymbol{\beta}) [\mathrm{I}\left(Y \preccurlyeq y\right) - m(\boldsymbol{X}, \boldsymbol{x}; \boldsymbol{\beta})] d\boldsymbol{\alpha}(\boldsymbol{z}), \quad \boldsymbol{z} = (y, \boldsymbol{x}),$$

then estimating equation (7.30) can be written as

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{V}[\boldsymbol{Z}_i, \boldsymbol{\beta}, F_{\boldsymbol{Z}}(\cdot)] = \mathbf{0},$$

with $F_{\boldsymbol{Z}}(\cdot)$ the cumulative distribution function of $\boldsymbol{Z}$. If we want to estimate $\boldsymbol{\beta}_0$, we will first need to estimate the nuisance function. The empirical distribution function $\hat{\boldsymbol{\alpha}}(\boldsymbol{z}) = \hat{F}_{\boldsymbol{Z}}(\boldsymbol{z})$ forms a natural choice. If we substitute this nuisance estimator we obtain the estimating equation

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{V}[\boldsymbol{Z}_i; \boldsymbol{\beta}, \hat{F}_{\boldsymbol{Z}}(\cdot)] = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \boldsymbol{B}(\boldsymbol{X}_i, \boldsymbol{X}_j; \boldsymbol{\beta}) \left[\mathrm{I}\left(Y_i \preccurlyeq Y_j\right) - m(\boldsymbol{X}_i, \boldsymbol{X}_j; \boldsymbol{\beta})\right] = \mathbf{0}. \tag{7.31}$$

Note that this estimating equation corresponds closely to the estimating equation (2.15) of the sparse correlation theory applied to the no-ordering restriction.

We denote the solution of (7.31) as $\hat{\boldsymbol{\beta}}_n$ and the solution of (7.30) as $\bar{\boldsymbol{\beta}}_n$. From the Z-estimator literature we know that, under regularity conditions, $\bar{\boldsymbol{\beta}}_n$ is a consistent estimator of $\boldsymbol{\beta}_0$ and is asymptotically normally distributed and we can construct a consistent sandwich estimator for its variance; see, for example, section 3.2 in Tsiatis (2006). We now try to find similar results for the estimator $\hat{\boldsymbol{\beta}}_n$. For more general results on semiparametric two-step estimators, we refer to section 8 of Newey and McFadden (1994).

Consistency

The solution of (7.30) is an example of a Z-estimator (sometimes referred to as an M-estimator), which is more generally defined as the solution of

$$\boldsymbol{g}_n(\boldsymbol{\beta}) := \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{g}(\boldsymbol{Z}_i; \boldsymbol{\beta}) = \mathbf{0}, \tag{7.32}$$

for a known vector valued function $\boldsymbol{g}(\cdot)$. The following theorem can be found in van der Vaart (1998, p. 46) and gives conditions under which the solution of (7.32) is consistent. Let $\|\boldsymbol{x}\|_*^2 = \boldsymbol{x}^T\boldsymbol{x}$ denote a norm and let Θ denote the parameter space of $\boldsymbol{\beta}$.

Theorem 27 (Consistency)**.** *Let $\boldsymbol{g}_0(\boldsymbol{\beta})$ denote a fixed vector valued function of $\boldsymbol{\beta}$ such that for all $\epsilon > 0$*

1. $\sup_{\boldsymbol{\beta}\in\Theta} \|\boldsymbol{g}_n(\boldsymbol{\beta}) - \boldsymbol{g}_0(\boldsymbol{\beta})\|_* \xrightarrow{p} 0$,

2. $\inf_{\boldsymbol{\beta}:\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_*>\epsilon} \|\boldsymbol{g}_0(\boldsymbol{\beta})\|_* > 0 = \|\boldsymbol{g}_0(\boldsymbol{\beta}_0)\|_*$.

Then $\tilde{\boldsymbol{\beta}}_n$ so that $\boldsymbol{g}_n(\tilde{\boldsymbol{\beta}}_n) = o_p(1)$ is a consistent estimator of $\boldsymbol{\beta}_0$.

Loosely speaking the theorem states that if $\boldsymbol{g}_n(\boldsymbol{\beta})$ and $\boldsymbol{g}_0(\boldsymbol{\beta})$ have a unique root and if $\boldsymbol{g}_n(\boldsymbol{\beta})$ converges uniformly to $\boldsymbol{g}_0(\boldsymbol{\beta})$, then the root of former will converge in probability to the root of the latter. We are now interested in finding the solution of (7.31), which can be considered as a generalization of Z-estimators, since the function $\boldsymbol{g}(\cdot)$ now depends on a couple of observations, i.e. an estimator of $\boldsymbol{\beta}$ can be written as the solution of

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{g}(\boldsymbol{Z}_i, \boldsymbol{Z}_j; \boldsymbol{\beta}) = \mathbf{0}.$$

Newey and McFadden (1994) refer to these estimators as V-estimators, since they are closely related to V-statistics (Serfling, 1980).

The following lemma will be convenient to assess consistency for V-estimators. We refer to Newey and McFadden (1994, p. 2214) for a proof.

Lemma 16. *Let $\{\boldsymbol{Z}_i \mid i = 1, \ldots, n\}$, $\boldsymbol{Z}$, and $\boldsymbol{Z}'$ be i.i.d., $\boldsymbol{g}(\boldsymbol{Z}, \boldsymbol{Z}'; \boldsymbol{\beta})$ continuous at each $\boldsymbol{\beta} \in \Theta$ with probability one,* $\mathrm{E}\left(\sup_{\boldsymbol{\beta}\in\Theta} \|\boldsymbol{g}(\boldsymbol{Z}, \boldsymbol{Z}; \boldsymbol{\beta})\|_*\right) < \infty$, *and* $\mathrm{E}\left(\sup_{\boldsymbol{\beta}\in\Theta} \|\boldsymbol{g}(\boldsymbol{Z}, \boldsymbol{Z}'; \boldsymbol{\beta})\|_*\right) < \infty$, *then* $\mathrm{E}[\boldsymbol{g}(\boldsymbol{Z}, \boldsymbol{Z}'; \boldsymbol{\beta})]$ *is continuous in $\boldsymbol{\beta} \in \Theta$ and*

$$\sup_{\boldsymbol{\beta}\in\Theta}\left\| n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{g}(\boldsymbol{Z}_i, \boldsymbol{Z}_j; \boldsymbol{\beta}) - \mathrm{E}\left[\boldsymbol{g}(\boldsymbol{Z}, \boldsymbol{Z}'; \boldsymbol{\beta})\right]\right\|_* \xrightarrow{p} 0.$$

Define

$$\boldsymbol{U}(\boldsymbol{Z}, \boldsymbol{Z}'; \boldsymbol{\beta}) := \boldsymbol{B}(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta})\left[\mathrm{I}\left(Y \preccurlyeq Y'\right) - m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta})\right], \tag{7.33}$$

$$\boldsymbol{V}_n(\boldsymbol{\beta}) := \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{V}\left[\boldsymbol{Z}_i; \boldsymbol{\beta}, \hat{F}_{\boldsymbol{Z}}(\cdot)\right] = n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{U}(\boldsymbol{Z}_i, \boldsymbol{Z}_j; \boldsymbol{\beta}), \tag{7.34}$$

and

$$\boldsymbol{V}_0(\boldsymbol{\beta}) = \mathrm{E}\left\{\boldsymbol{V}\left[\boldsymbol{Z}; \boldsymbol{\beta}, F_{\boldsymbol{Z}}(\cdot)\right]\right\} = \mathrm{E}[\boldsymbol{U}(\boldsymbol{Z}, \boldsymbol{Z}'; \boldsymbol{\beta})]. \tag{7.35}$$

The following theorem gives the conditions under which $\hat{\boldsymbol{\beta}}_n$ will be consistent.

Theorem 28. *Let $\boldsymbol{U}(\boldsymbol{Z}, \boldsymbol{Z}'; \boldsymbol{\beta})$, $\boldsymbol{V}_n(\boldsymbol{\beta})$, and $\boldsymbol{V}_0(\boldsymbol{\beta})$ as defined in (7.33), (7.34), and (7.35) respectively. If (i) $\boldsymbol{U}(\boldsymbol{Z}, \boldsymbol{Z}'; \boldsymbol{\beta})$ is continuous at each $\boldsymbol{\beta} \in \Theta$ with probability one and Θ compact, (ii)* $\mathrm{E}(\sup_{\boldsymbol{\beta}\in\Theta} \|\boldsymbol{U}(\boldsymbol{Z}, \boldsymbol{Z}'; \boldsymbol{\beta})\|_*) < \infty$, *(iii) $\boldsymbol{V}_0(\boldsymbol{\beta})$ has a unique root, then $\hat{\boldsymbol{\beta}}_n$, defined as the root of $\boldsymbol{V}_n(\boldsymbol{\beta}) = \mathbf{0}$, is a consistent estimator for $\boldsymbol{\beta}_0$.*

Proof. Since

$$\boldsymbol{V}_n(\boldsymbol{\beta}) = n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{U}(\boldsymbol{Z}_i,\boldsymbol{Z}_j;\boldsymbol{\beta}), \quad \boldsymbol{V}_0(\boldsymbol{\beta}) = \mathrm{E}[\boldsymbol{U}(\boldsymbol{Z}_i,\boldsymbol{Z}_j;\boldsymbol{\beta})], \quad \boldsymbol{U}(\boldsymbol{Z},\boldsymbol{Z};\boldsymbol{\beta}) = \mathbf{0},$$

it follows from Lemma 16 that

$$\sup_{\boldsymbol{\beta}\in\Theta}\|\boldsymbol{V}_n(\boldsymbol{\beta}) - \boldsymbol{V}_0(\boldsymbol{\beta})\|_* \xrightarrow{p} 0.$$

As defined previously, $\boldsymbol{\beta}_0$ denotes the true parameter (i.e. the parameter corresponding to the data generating model). We need to show that $\boldsymbol{\beta}_0$ is the root of $\boldsymbol{V}_0(\boldsymbol{\beta})$. This can be seen as follows. By definition it holds that $\mathrm{E}\left(\mathrm{I}\left(Y \preccurlyeq Y'\right) \mid \boldsymbol{X},\boldsymbol{X}'\right) = m(\boldsymbol{X},\boldsymbol{X}';\boldsymbol{\beta}_0)$, so that

$$\begin{aligned}\boldsymbol{V}_0(\boldsymbol{\beta}_0) &= \mathrm{E}\left\{\boldsymbol{B}(\boldsymbol{X},\boldsymbol{X}';\boldsymbol{\beta}_0)\left[\mathrm{I}\left(Y \preccurlyeq Y'\right) - m(\boldsymbol{X},\boldsymbol{X}';\boldsymbol{\beta}_0)\right]\right\}\\ &= \mathrm{E}\left\{\boldsymbol{B}(\boldsymbol{X},\boldsymbol{X}';\boldsymbol{\beta}_0)\left[\mathrm{E}(\mathrm{I}\left(Y \preccurlyeq Y'\right) \mid \boldsymbol{X},\boldsymbol{X}') - m(\boldsymbol{X},\boldsymbol{X}';\boldsymbol{\beta}_0)\right]\right\} = \mathbf{0}.\end{aligned}$$

By Lemma 16 it follows that $\boldsymbol{V}_0(\boldsymbol{\beta})$ is continuous and since Θ is compact and $\boldsymbol{\beta}_0$ is the unique root it follows that

$$\inf_{\boldsymbol{\beta}:\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_*>\epsilon}\|\boldsymbol{V}_0(\boldsymbol{\beta})\|_* > 0 = \|\boldsymbol{V}_0(\boldsymbol{\beta}_0)\|_*.$$

Since $\boldsymbol{V}_n(\hat{\boldsymbol{\beta}}_n) = \mathbf{0}$ all assumptions of Theorem 27 are fulfilled so that $\hat{\boldsymbol{\beta}}_n$ is consistent for $\boldsymbol{\beta}_0$. □

Normality

Once consistency is established, we can use an expansion to obtain the asymptotic distribution of $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)$. It follows that

$$\begin{aligned}\mathbf{0} &= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{V}[\boldsymbol{Z}_i;\hat{\boldsymbol{\beta}}_n,\hat{F}_{\boldsymbol{Z}}(\cdot)]\\ &= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{V}[\boldsymbol{Z}_i;\boldsymbol{\beta}_0,\hat{F}_{\boldsymbol{Z}}(\cdot)] + \frac{1}{n}\sum_{i=1}^{n}\frac{\partial\boldsymbol{V}[\boldsymbol{Z}_i;\boldsymbol{\beta},\hat{F}_{\boldsymbol{Z}}(\cdot)]}{\partial\boldsymbol{\beta}^T}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_n^*}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0),\end{aligned}$$

with $\boldsymbol{\beta}_n^*$ an intermediate value between $\hat{\boldsymbol{\beta}}_n$ and $\boldsymbol{\beta}_0$. Consequently

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) = -\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\boldsymbol{V}[\boldsymbol{Z}_i;\boldsymbol{\beta},\hat{F}_{\boldsymbol{Z}}(\cdot)]}{\partial\boldsymbol{\beta}^T}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_n^*}\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{V}[\boldsymbol{Z}_i;\boldsymbol{\beta}_0,\hat{F}_{\boldsymbol{Z}}(\cdot)].$$

Since $\hat{\boldsymbol{\beta}}_n$ and $\hat{F}_{\boldsymbol{Z}}(\cdot)$ are consistent, under regularity conditions, it follows that

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\boldsymbol{V}[\boldsymbol{Z}_i;\boldsymbol{\beta},\hat{F}_{\boldsymbol{Z}}(\cdot)]}{\partial\boldsymbol{\beta}^T}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_n^*} \xrightarrow{p} \mathrm{E}\left(\frac{\partial\boldsymbol{V}[\boldsymbol{Z}_i;\boldsymbol{\beta},F_{\boldsymbol{Z}}(\cdot)]}{\partial\boldsymbol{\beta}^T}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right),$$

see, for example, section 8.2 of Newey and McFadden (1994). The asymptotic distribution of $n^{-1/2}\sum_{i=1}^{n} \boldsymbol{V}[\boldsymbol{Z}_i; \boldsymbol{\beta}_0, \hat{F}_{\boldsymbol{Z}}(\cdot)]$ is, however, more complicated, since the nuisance estimator $\hat{F}_{\boldsymbol{Z}}(\cdot)$ should be accounted for. We refer to theorem 8.1 of Newey and McFadden (1994) for more details.

The following theorem gives the asymptotic distribution of $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)$ and is a special case of Theorem 8.6 of Newey and McFadden, since $\boldsymbol{U}(\boldsymbol{Z}, \boldsymbol{Z}; \boldsymbol{\beta}) \equiv \mathbf{0}$ because $\mathrm{I}(Y \preccurlyeq Y) = m(\boldsymbol{X}, \boldsymbol{X}; \boldsymbol{\beta}) = 0.5$

Theorem 29. *Let* $\{\boldsymbol{Z}_i \mid i = 1, \ldots, n\}$ *be i.i.d.,* $\hat{\boldsymbol{\beta}}_n \xrightarrow{p} \boldsymbol{\beta}_0$*, (i)* $\mathrm{E}\left(\|\boldsymbol{U}(\boldsymbol{Z}, \boldsymbol{Z}'; \boldsymbol{\beta}_0)\|_*^2\right) < \infty$*, (ii)* $\boldsymbol{U}(\boldsymbol{Z}, \boldsymbol{Z}'; \boldsymbol{\beta}_0)$ *is continuously differentiable on a neighbourhood of* $\boldsymbol{\beta}_0$ *with probability one, and there is a neighbourhood* $\mathcal{N}$ *of* $\boldsymbol{\beta}_0$ *such that*

$$\mathrm{E}\left(\sup_{\boldsymbol{\beta}\in\mathcal{N}} \left\| \frac{\partial \boldsymbol{U}(\boldsymbol{Z}, \boldsymbol{Z}'; \boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T} \right\|_*\right) < \infty,$$

and (iii)

$$\boldsymbol{J}(\boldsymbol{\beta}_0) := \mathrm{E}\left(\frac{\partial \boldsymbol{U}(\boldsymbol{Z}, \boldsymbol{Z}'; \boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \right)$$

is nonsingular, then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}),$$

with

$$\boldsymbol{\Sigma} = 4\boldsymbol{J}(\boldsymbol{\beta}_0)^{-1}\mathrm{Cov}\left[\mathrm{E}\left(\boldsymbol{U}(\boldsymbol{Z}, \boldsymbol{Z}'; \boldsymbol{\beta}_0) \mid \boldsymbol{Z}\right)\right]\boldsymbol{J}(\boldsymbol{\beta}_0)^{-1^T}.$$

Consistent sandwich estimator

The following theorem gives a consistent estimator for the variance of the asymptotic distribution of $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)$. See Newey and McFadden (1994, p. 2203) for a proof.

Theorem 30. *If the conditions of Theorem 29 are fulfilled and if*

$$\mathrm{E}\left(\sup_{\boldsymbol{\beta}\in\mathcal{N}} \left\| \frac{\partial \boldsymbol{U}(\boldsymbol{Z}, \boldsymbol{Z}'; \boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T} \right\|_*^2\right) < \infty,$$

then with

$$\hat{\boldsymbol{J}}(\hat{\boldsymbol{\beta}}_n) := \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial \boldsymbol{U}(\boldsymbol{Z}_i, \boldsymbol{Z}_j; \boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_n},$$

and

$$\hat{\boldsymbol{K}}(\hat{\boldsymbol{\beta}}_n) := \frac{4}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} \boldsymbol{U}(\boldsymbol{Z}_i, \boldsymbol{Z}_j; \hat{\boldsymbol{\beta}}_n)\boldsymbol{U}(\boldsymbol{Z}_i, \boldsymbol{Z}_k; \hat{\boldsymbol{\beta}}_n)^T,$$

it holds that

$$\hat{\boldsymbol{J}}(\hat{\boldsymbol{\beta}}_n)^{-1}\hat{\boldsymbol{K}}(\hat{\boldsymbol{\beta}}_n)\hat{\boldsymbol{J}}(\hat{\boldsymbol{\beta}}_n)^{-1^T} \xrightarrow{p} \boldsymbol{\Sigma}. \tag{7.36}$$

7.3.5 Relationship with sparse correlation theory

In this section we relate the results of Theorems 28, 29, and 30 to the results of Section 2.3.2 for the no-order restriction $\mathcal{X}_0$. Upon using (2.15) and (2.16), the estimator discussed in Section 2.3.2 can be written as the solution of

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\partial m(\boldsymbol{X}_i,\boldsymbol{X}_j;\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\frac{[\mathrm{I}\left(Y_i \preccurlyeq Y_j\right)-m(\boldsymbol{X}_i,\boldsymbol{X}_j;\boldsymbol{\beta})]}{m(\boldsymbol{X}_i,\boldsymbol{X}_j;\boldsymbol{\beta})[1-m(\boldsymbol{X}_i,\boldsymbol{X}_j;\boldsymbol{\beta})]}=\mathbf{0}. \tag{7.37}$$

Upon using

$$\boldsymbol{B}(\boldsymbol{X}_i,\boldsymbol{X}_j;\boldsymbol{\beta})=\frac{\partial m(\boldsymbol{X}_i,\boldsymbol{X}_j;\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\left\{m(\boldsymbol{X}_i,\boldsymbol{X}_j;\boldsymbol{\beta})\left[1-m(\boldsymbol{X}_i,\boldsymbol{X}_j;\boldsymbol{\beta})\right]\right\}^{-1},$$

equation (7.37) can be equivalently written as

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{B}(\boldsymbol{X}_i,\boldsymbol{X}_j;\boldsymbol{\beta})\left[\mathrm{I}\left(Y_i \preccurlyeq Y_j\right)-m(\boldsymbol{X}_i,\boldsymbol{X}_j;\boldsymbol{\beta})\right]=\mathbf{0}.$$

Since $\boldsymbol{B}(\boldsymbol{X}_i,\boldsymbol{X}_j;\boldsymbol{\beta})+\boldsymbol{B}(\boldsymbol{X}_j,\boldsymbol{X}_i;\boldsymbol{\beta})=\mathbf{0}$, equation (7.37) is a special case of the class of estimating equations of Theorem 28.

To compare the sandwich estimator of Theorem 30 with the sandwich estimator of Theorem 2 let $\boldsymbol{U}_{ij}(\boldsymbol{\beta}):=\boldsymbol{U}(\boldsymbol{Z}_i,\boldsymbol{Z}_j;\boldsymbol{\beta})=\boldsymbol{B}(\boldsymbol{X}_i,\boldsymbol{X}_j;\boldsymbol{\beta})[\mathrm{I}\left(Y_i \preccurlyeq Y_j\right)-m(\boldsymbol{X}_i,\boldsymbol{X}_j;\boldsymbol{\beta})]$. From Theorem 2 it follows that the sandwich estimator of $\sqrt{n}(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0)$ can be written as

$$\tilde{\boldsymbol{J}}(\hat{\boldsymbol{\beta}}_n)^{-1}\tilde{\boldsymbol{K}}(\hat{\boldsymbol{\beta}}_n)\tilde{\boldsymbol{J}}(\hat{\boldsymbol{\beta}}_n)^{-1^T},$$

with

$$\tilde{\boldsymbol{J}}(\hat{\boldsymbol{\beta}}_n):=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\partial\boldsymbol{U}(\boldsymbol{Z}_i,\boldsymbol{Z}_j;\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^T}\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_n},$$

and

$$\begin{aligned}\tilde{\boldsymbol{K}}(\hat{\boldsymbol{\beta}}_n) \;:=\; & \frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{U}_{ij}(\hat{\boldsymbol{\beta}}_n)\left[\sum_{k=1}^{n}\boldsymbol{U}_{ik}(\hat{\boldsymbol{\beta}}_n)^T+\sum_{k=1}^{n}\boldsymbol{U}_{ki}(\hat{\boldsymbol{\beta}}_n)^T\right.\\ & \left.+\sum_{k=1}^{n}\boldsymbol{U}_{jk}(\hat{\boldsymbol{\beta}}_n)^T+\sum_{k=1}^{n}\boldsymbol{U}_{kj}(\hat{\boldsymbol{\beta}}_n)^T-\boldsymbol{U}_{ij}(\hat{\boldsymbol{\beta}}_n)^T\right].\end{aligned}$$

If follows that

$$\tilde{\boldsymbol{J}}(\hat{\boldsymbol{\beta}}_n)=\hat{\boldsymbol{J}}(\hat{\boldsymbol{\beta}}_n),$$

and, since $\boldsymbol{U}_{ij}(\hat{\boldsymbol{\beta}}_n) = \boldsymbol{U}_{ji}(\hat{\boldsymbol{\beta}}_n)$,

$$\tilde{\boldsymbol{K}}(\hat{\boldsymbol{\beta}}_n) = \hat{\boldsymbol{K}}(\hat{\boldsymbol{\beta}}_n) - \frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{U}_{ij}(\hat{\boldsymbol{\beta}}_n)\boldsymbol{U}_{ij}(\hat{\boldsymbol{\beta}}_n)^T, \tag{7.38}$$

with $\hat{\boldsymbol{J}}(\hat{\boldsymbol{\beta}}_n)$ and $\hat{\boldsymbol{K}}(\hat{\boldsymbol{\beta}}_n)$ as defined in Theorem 30. Since the last term in (7.38) converges in probability to zero, both sandwich estimators are asymptotically equivalent.

7.4 An example

As a first attempt to study the efficient estimator in more detail, we consider a specific setting where the outcome follows a normal distribution and is linearly related to one-dimensional discrete predictor.

7.4.1 The data-generating model

Consider the special setting where $\boldsymbol{Z} = (Y, X)$, with $X \in \{x_1, \ldots, x_K\}$ a one-dimensional discrete predictor with finite support. The outcome Y is continuous, has infinite support and is related to X according to

$$Y = \alpha X + \varepsilon, \quad \varepsilon \stackrel{d}{=} \mathrm{N}(0, \sigma^2), \tag{7.39}$$

with associated PIM

$$\mathrm{P}\left(Y \preccurlyeq Y' \mid X, X'\right) = \Phi[\beta(X' - X)].$$

To find the efficient estimator of β, we need to solve the integral equation (7.26). Since X is discrete, it follows that

$$\begin{aligned} & \mathrm{E}[B_{eff}(X, X^*)V(X, X', X, X^*) + B_{eff}(X', X^*)V(X, X', X', X^*) \mid X, X'] \quad (7.40) \\ = & \sum_{k=1}^{K}[B_{eff}(X, x_k)V(X, X', X, x_k) + B_{eff}(X', x_k)V(X, X', X', x_k)]\mathrm{P}\left(X^* = x_k\right), \end{aligned}$$

where $B_{eff}(\cdot)$ is the index function associated with the efficient estimator. If is of interest to find $B_{eff}(\cdot)$ so as to obtain the efficient score (7.27).

For notational convenience let $B_{ij} := B_{eff}(x_i, x_j)$, $D_{ij} := D(x_i, x_j; \beta_0)$, $V_{ijkl} := V(x_i, x_j, x_k, x_l)$, and $p_i = \mathrm{P}\left(X = x_i\right)$. The integral equation (7.26) is now equivalent to the linear system of

equations

$$D_{ij} = \sum_{k=1}^{K} B_{ik} V_{ijik} p_k + \sum_{k=1}^{K} B_{jk} V_{ijjk} p_k, \quad \forall i, j = 1, \ldots, K. \tag{7.41}$$

This system of equations can be further simplified. Since $V_{ijkl} = -V_{ijlk} = V_{klij}$, $B_{ij} = -B_{ji}$, $D_{ij} = -D_{ji}$, $B_{ii} = 0$, $D_{ii} = 0$, and $V_{iijk} = 0$ if follows that the equations for which $i = j$ do not contribute and that

$$\begin{aligned} D_{ji} &= \sum_{k=1}^{K} B_{jk} V_{jijk} p_k + \sum_{k=1}^{K} B_{ik} V_{jiik} p_k \\ \Leftrightarrow D_{ij} &= \sum_{k=1}^{K} B_{jk} V_{ijjk} p_k + \sum_{k=1}^{K} B_{ik} V_{ijik} p_k. \end{aligned}$$

Hence, the set of equations (7.41) reduces to

$$\begin{aligned} D_{ij} &= \sum_{k=1}^{K} B_{ik} V_{ijik} p_k + \sum_{k=1}^{K} B_{jk} V_{ijjk} p_k, \quad \forall i < j \\ \Leftrightarrow D_{ij} &= \sum_{k<i} B_{ki} V_{ijki} p_k + \sum_{k>i} B_{ik} V_{ijik} p_k + \sum_{k<j} B_{kj} V_{ijkj} p_k + \sum_{k>j} B_{jk} V_{ijjk} p_k. \end{aligned} \tag{7.42}$$

For simplicity, let $p_1 = \ldots = p_K = 1/K$. Since $V_{ijkl} = 0$ if $(i, j) \cap (k, l) = \emptyset$, (7.42) can be expressed as

$$D_{ij} = \frac{1}{K} \sum_{k=1}^{K-1} \sum_{l=k+1}^{K} B_{kl} V_{ijkl} + \frac{1}{K} B_{ij} V_{ijij}, \quad \forall i < j. \tag{7.43}$$

To solve this system of equations, we rewrite (7.43) in matrix notation. Let $\boldsymbol{D}$ ($\boldsymbol{B}_{eff}$) denote the $K(K-1)/2$-vector of elements D_{ij} (B_{ij}) with $i < j$ and $\boldsymbol{V}$ the $[K(K-1)/2] \times [K(K-1)/2]$ matrix with element V_{ijkl} with $i < j$ and $k < l$, and $\boldsymbol{V}_{ind}$ the $[K(K-1)/2] \times [K(K-1)/2]$ diagonal matrix with elements V_{ijij} with $i < j$. Equation (7.43) is equivalent to

$$K\boldsymbol{D} = \boldsymbol{B}_{eff}(\boldsymbol{V} + \boldsymbol{V}_{ind}).$$

Since $\boldsymbol{V}$ and $\boldsymbol{V}_{ind}$ are positive definite, so is $(\boldsymbol{V} + \boldsymbol{V}_{ind})$ and the index function corresponding to the most efficient estimator is given by

$$\boldsymbol{B}_{eff} = K\boldsymbol{D}(\boldsymbol{V} + \boldsymbol{V}_{ind})^{-1}. \tag{7.44}$$

Note that the index function of the estimating equations based on the independence working correlation matrix, i.e. the equations (2.15) with index function (2.16), are given by

$$\boldsymbol{B}_{ind} = \boldsymbol{D}\boldsymbol{V}_{ind}^{-1}. \tag{7.45}$$

7.4.2 Simulation results

We set up a small simulation study. Consider the data-generating model

$$Y = \alpha X + \varepsilon, \quad \varepsilon \stackrel{d}{=} \mathrm{N}[0, \sigma^2(X)], \quad \mathrm{P}\left(X = x_i\right) = \frac{1}{K},\ i = 1, \ldots, K, \tag{7.46}$$

where σ^2 can be fixed, i.e. $\sigma^2(X) = \sigma^2$, or a function of X, i.e. $\sigma^2(X) = \sigma^2 X$. Here, X is discrete and takes on $K = 5$ equidistant values in the interval $[0.1, u]$ where $u = 1$ or $u = 10$. For each simulation run a sample size of $n = 1000$ is considered.

The homoscedastic model (7.46) with $\sigma^2(X) = \sigma^2$ is associated with the PIM

$$\mathrm{P}\left(Y < Y' \mid X, X'\right) = \Phi\left[\beta(X' - X)\right], \quad \beta = \frac{\alpha}{\sqrt{2\sigma^2}},$$

while for the heteroscedastic model $\sigma^2(X) = \sigma^2 X$ this is

$$\mathrm{P}\left(Y < Y' \mid X, X'\right) = \Phi\left[\beta \frac{(X' - X)}{\sqrt{X' + X}}\right], \quad \beta = \frac{\alpha}{\sigma}.$$

The index functions $\boldsymbol{B}_{eff}$ (7.44) and $\boldsymbol{B}_{ind}$ (7.45) both depend on β and/or the nuisance parameters $\boldsymbol{V}$. For simplicity, these nuisance parameters are not estimated from the data, but are approximated based on 10000 Monte-Carlo simulations.

Table 7.1 gives the simulation results based on 1000 Monte-Carlo simulations, where $\hat{\beta}_{eff}$ ($\hat{\beta}_{ind}$) corresponds to the estimator of the estimating equations (7.31) with index function $\boldsymbol{B}_{eff}$ ($\boldsymbol{B}_{ind}$). Except for the heteroscedastic model with $\alpha = 1$, $u = 10$, and $\sigma = 1$, $\hat{\beta}_{eff}$ is more efficient. The difference is, however, negligible, suggesting that, for this simulation set-up, $\hat{\beta}_{ind}$ is a good approximation of the efficient estimator. However, more research is needed to study the properties of these estimators in more detail. Furthermore, the current approach of approximating the nuisance parameters based on the Monte-Carlo simulations is not useful in practice and methods for estimating these nuisance parameters from the data should be constructed. This estimation can potentially have an impact on the performance of the efficient estimator.

7.5 Discussion

In this chapter we have studied efficient estimators for probabilistic index models. Based on the semiparametric theory as described in Tsiatis (2006), we have constructed the efficient score

Table 7.1: Simulation results for the normal linear model, based on 1000 Monte Carlo runs. β is the true parameter, $\mathrm{Av}(\hat{\beta}_{eff})$ the average of the β estimate associated with index function (7.44), $\mathrm{Var}(\sqrt{n}\hat{\beta}_{eff})$ the sample variance of the simulated $\sqrt{n}\hat{\beta}_{eff}$, $\mathrm{Av}(\hat{\beta}_{ind})$ the average of the β estimate associated with index function (7.45), $\mathrm{Var}(\sqrt{n}\hat{\beta}_{ind})$ the sample variance of the simulated $\sqrt{n}\hat{\beta}_{ind}$.

α	u	σ	β	$\mathrm{Av}(\hat{\beta}_{eff})$	$\mathrm{Av}(\hat{\beta}_{ind})$	$\mathrm{Var}(\sqrt{n}\hat{\beta}_{eff})$	$\mathrm{Var}(\sqrt{n}\hat{\beta}_{ind})$
Homoscedastic linear model							
1	1	1	0.707	0.708	0.708	4.329	4.333
1	10	1	0.707	0.709	0.709	0.351	0.355
1	1	5	0.1414	0.1434	0.1434	5.0875	5.0876
1	10	5	0.1414	0.1415	0.1416	0.0544	0.0553
Heteroscedastic linear model							
1	1	1	1	1	1	5.056	5.119
1	10	1	1	1	1	1.194	1.165
1	1	5	0.2	0.2	0.2	4.836	4.838
1	10	5	0.2	0.2	0.2	0.346	0.349

for PIMs. However, the efficient score is not directly applicable in practice since it involves a conditional expectation and an integral equations needs to be solved to find the appropriate index function.

We have shown that the conditional expectation can be replaced by a sample average resulting in a semiparametric two-step estimator as described by Newey and McFadden (1994). The estimator is consistent and asymptotically normally distributed. We also provide a consistent estimator for its asymptotic variance. Most results correspond closely to the results of Chapter 2 which are based on the sparse correlation theory of Lumley and Mayer-Hamblett (2003).

We have briefly examined the performance of the efficient estimator for a specific setting where the predictor is discrete so that the integral equation reduces to a set of equations. The results of the simulation study indicate that the variance of the efficient estimator is similar to the variance of the estimator based on the independent working correlation matrix, suggesting that the latter is a good choice in practice. However, this should be examined in more detail.

Moreover, solving the integral equation in general seems to be complicated and should also be studied in more detail. Primitive conditions should be developed under which the integral equations has a solution. Furthermore, the integral equation depends on nuisance parameters for which estimators should be constructed and a strategy for solving the integral equations should also be developed. A first solution can perhaps be obtained by writing the integral equation as a conditional expectation and by replacing the conditional mean with a sample average. Alternatively, one can try to apply the results of the V-estimation theory as developed by Newey (1989).

Chapter 8

Discussion and future research perspectives

Regression methods form an important, flexible, and powerful tool for the analysis of data. Statisticians can choose out of a variety of methods to select the most appropriate one(s); a choice that mainly depends on the research question(s) and, in all its facets, on the data at hand.

Based on the research question and/or the data one can select a summary measure of interest; e.g. is the mean outcome sufficient informative or do quantiles describe the underlying process more accurately? Once this measure is selected, one can choose the type of regression model; e.g. are the predictors linearly related to the summary measure or should one consider more flexible models such as generalized additive models?

Without any doubt, the collection of regression methods is vast and provides sufficient tools for the analysis of most datasets. As stated by Thomas Gerds in his discussion of the probabilistic index models paper: *...experienced statisticians [who] know how to apply the wrong tools and still arrive at sound conclusions* (Gerds, 2012). So even applying the wrong method can lead to correct conclusions, as long as the statistician sufficiently understands the potential and limitations of the method applied.

Despite the richness of the existing regression methods, we believe that there is still room for introducing new regression models. More specifically, in this dissertation we have proposed a regression model for a summary measure different from the mean and quantiles. The measure of interest is the *probabilistic index* (PI), i.e. the probability $\mathrm{P}\left(Y \preccurlyeq Y'\right) := \mathrm{P}\left(Y < Y'\right) +$

$0.5\mathrm{P}\,(Y = Y)$, where Y and Y' denote two independent outcomes. As argued in Chapter 1, this measure has been promoted by many authors as an informative and intuitive effect size measure and is applicable to ordinal, interval, and ratio-scale outcomes. However, as for all summary measures, it inevitably results in information loss and can sometimes be misunderstood.

The PI is a well-known statistic for those familiar with the Wilcoxon–Mann–Whitney (WMW) test, since the WMW two-sided alternative hypothesis states that $H_1 : \mathrm{P}\,(Y_1 \preccurlyeq Y_2) \neq 0.5$, with Y_1 (Y_2) a random outcome of the first (second) group. This two-sample setting can be translated into a regression context by defining a predictor which indicates the two groups, e.g. $X = 1$ for the first group and $X = 2$ for the second. Let Y denote the outcome associated with X and Y' the outcome associated with X'. If we model the PI as follows

$$\mathrm{P}\,(Y \preccurlyeq Y' \mid X = 1, X' = 2) = \beta(X' - X) \equiv \beta,$$

then the alternative hypothesis of the WMW test can be expressed as $H_1 : \beta \neq 0.5$. Thus the WMW test can be reformulated as a regression problem. A natural next step consists of extending the regression model for the PI to more complicated designs than the two-sample problem, e.g. the K-sample problem, a setting with continuous predictors, etc. This extension has been the focus of this dissertation. We have proposed a flexible modelling framework for the PI, named *probabilistic index models* (PIM). If Y denotes the outcome associated with the (possibly multidimensional) predictor $\boldsymbol{X}$ and Y' the outcome associated with the predictor $\boldsymbol{X}'$ such that $(Y, \boldsymbol{X})$ and $(Y', \boldsymbol{X}')$ are independently and identically distributed, a PIM is defined as

$$\mathrm{P}\,(Y \preccurlyeq Y' \mid \boldsymbol{X}, \boldsymbol{X}') = m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}), \quad (\boldsymbol{X}, \boldsymbol{X}') \in \mathcal{X}, \tag{8.1}$$

for a function $m(\cdot)$ subject to regularity and smoothness conditions, where $\mathcal{X}$ denotes the set of couples of predictors for which the model is defined, and $\boldsymbol{\beta}$ is the unknown model parameter.

In Chapter 2 we have developed semiparametric asymptotic theory for PIMs upon using the concept of sparse correlation as introduced by Lumley and Mayer-Hamblett (2003). The results of a simulation study demonstrate that the theoretical properties of the asymptotic theory apply well to moderately sized samples. For future research it can be interesting to improve the asymptotic approximations for small samples. More specifically, extending bootstrap schemes to the PIM framework can perhaps lead to better small sample approximations.

Since a PIM involves a couple of observations $(Y, \boldsymbol{X})$ and $(Y', \boldsymbol{X}')$, the relationship with well

known models forms an important first step in understanding the type of functional relation between the PI and the predictors $\boldsymbol{X}$ and $\boldsymbol{X}'$. More specifically, how should we choose the function $m(\cdot)$ in (8.1)? In Chapter 3 we have shown that there is direct relationship between the model parameters of a PIM and the model parameters of the normal linear model and the Cox proportional hazards model. More specifically, these relationships suggest a functional form

$$m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}) = g^{-1}\left[(\boldsymbol{X}' - \boldsymbol{X})^T \boldsymbol{\beta}\right],$$

with $g(\cdot)$ the logit or probit link. This choice of $m(\cdot)$ results in PIM parameters with an informative interpretation. As an illustration, consider a one-dimensional predictor X and the PIM

$$\mathrm{P}\left(Y \preccurlyeq Y' \mid X, X'\right) = \mathrm{expit}[(X' - X)\beta], \quad \mathrm{expit}(u) = \frac{\exp(u)}{1 + \exp(u)}.$$

It follows that $\mathrm{expit}(\beta) = \mathrm{P}\left(Y \preccurlyeq Y' \mid X = x, X' = x + 1\right)$, i.e. when comparing a group of subjects for which $X = x$ with a group for which $X = x+1$, $\mathrm{expit}(\beta)$ gives the probability that the outcome of a randomly selected subject with $X = x+1$ exceeds the outcome of a randomly selected subject with a predictor values which is one unit lower.

There is also a close relationship between the PIM on the one hand and the AUC-regression model, the concordance index, the Hodges–Lehmann estimator, and the rank regression model on the other hand. For future research it would be interesting to extend the Hodges–Lehmann estimator so as to account for confounders. We briefly sketch how this could be done. Let $\boldsymbol{X}^T = (X, \boldsymbol{C}^T)$ where $X \in \{0, 1\}$ is a dummy variable indicating two groups and where $\boldsymbol{C}$ denotes a set of confounders. Consider the PIM

$$\mathrm{P}\left(Y \preccurlyeq Y' - \alpha \mid X = 0, X' = 1, \boldsymbol{C}, \boldsymbol{C'}\right) = \mathrm{expit}[(\boldsymbol{C'} - \boldsymbol{C})^T \boldsymbol{\beta}],$$

where α is the parameter of interest and $\boldsymbol{\beta}$ is nuisance. For $\boldsymbol{C} = \boldsymbol{C'}$, the model reduces to

$$\mathrm{P}\left(Y \preccurlyeq Y' - \alpha \mid X = 0, X' = 1, \boldsymbol{C} = \boldsymbol{C'}\right) = \frac{1}{2},$$

where the estimator of α will be an extension of the Hodges–Lehmann estimator, but now controlling for the confounding variables.

In Chapter 4 we have studied the relationship between the PIM and many popular rank tests. It turns out that the PIM plays the role of a general linear model (GLM) in the rank world; whereas a GLM facilitates the generalization of two-sample t-tests and ANOVA F-tests to more complicated designs, a PIM does the same trick for the WMW, Kruskal–Wallis, Friedman and

many more rank tests. Embedding these rank tests into the statistical modelling framework of a PIM allows for a better understanding of the hypotheses tested and allows for constructing confidence intervals for the associated effect sizes. Embedding all these rank tests in a single modelling framework can perhaps make these tests more accessible to non-experienced users and boost its popularity.

Since the PIM (8.1) is a semiparametric model, the adequacy of the proposed model $m(\cdot)$ should be assessed. Therefore, in Chapter 5 we have developed goodness-of-fit (GOF) methods: a statistical test as well as a graphical diagnostic tool. The GOF plot provides information on how the model can be improved. The results of a power study suggested a decent performance of the test, but, however, also indicated that the test was sometimes too liberal. Both methods are consistent with the interpretation of a PIM and are based on the probability $\mathrm{P}\left(Y \preccurlyeq Y' \mid \boldsymbol{X}, \boldsymbol{X}' = \boldsymbol{X} + \boldsymbol{\Delta}\right)$ for a fixed constant $\boldsymbol{\Delta}$. Smoothed residuals are used for both the construction of the test and for the graphical tool. A modified cross-validation score is proposed to select the bandwidth automatically. The test based on this automatic bandwidth selection, however, showed a decrease in performance and an inflated type I error rate. For future research, it can be interesting to focus on an adaptive selection of the parameter $\boldsymbol{\Delta}$. Furthermore, the automatic choice of bandwidth is computationally intensive so that the development of alternative criteria for an automatic selection is desirable. The test is constructed based on a subset of the smoothed residuals so as to avoid theoretical difficulties. However, it is anticipated that using all residuals to form a Cramér–von Mises, Anderson–Darling, or Kolmogorov–Smirnov type of test will make the method more sensitive for detecting a wider range of model departures.

In Chapter 6 we have worked out a case study in detail. More specifically, the PIM methodology is used to analyze genomic differential expression studies based on reverse transcription quantitative polymerase chain reaction (RT-qPCR). The data generated by RT-qPCR techniques require normalization so as to account for technical variation which cannot be attributed to the treatments. Furthermore, since RT-qPCR experiments often aim at validating differentially expressed genes that were discovered by microarrays or next generation sequencing screens – and RT-qPCR biological validation experiments are often an (intermediate) endpoint of a study – quantifying and interpreting the effects is important for increasing the insight in the biological processes under study. The PIM turns out to be appropriate for both goals: it allows for normal-

izing the data in a straightforward fashion, while keeping an intuitive interpretation in terms of the odds for down- or upregulation.

In Chapter 7 we have studied efficient estimators for PIMs in a semiparametric setting. The index function associated with the efficient score corresponds to the solution of an integral equation. The results of a small simulation study indicate that the variance of the efficient estimator is similar to the variance of the estimator based on the independence working correlation matrix. However, this needs to be studied in more detail: more research needs to done so as obtain primitive conditions under which the integral equation has a solution and how this solution should be obtained. Furthermore, the efficient score depends on nuisance parameters which need to be estimated. These nuisance parameters are related to a kind of *second order PIM*

$$\mathrm{P}\left[Y \preccurlyeq \min(Y', Y'') \mid \boldsymbol{X}, \boldsymbol{X}', \boldsymbol{X}''\right] = \bar{m}(\boldsymbol{X}, \boldsymbol{X}', \boldsymbol{X}''; \boldsymbol{\gamma}),$$

for a function $\bar{m}(\cdot)$. These models are also interesting beyond the efficiency context. For example, consider the setting where there are three treatments: a new treatment (X), the standard treatment (X'), and a placebo treatment (X'') and let lower outcomes be better. The second order PIM models the probability that a randomly selected subject treated with the new treatment will be better off as compared to both a randomly selected subject treated with the standard treatment and a randomly selected subject treated with placebo.

Based on the theory of semiparametric two-step estimators as described by Newey and McFadden (1994), in Chapter 7 we have also developed asymptotic theory for PIMs without relying on the sparse correlation theory of Lumley and Mayer-Hamblett (2003).

Since the PIMs form a new class of regression models, there are many extensions which can to be developed. We just name a few.

- *Functional data analysis.* Let $Y(t)$ denote a random outcome function and $\boldsymbol{X}(t)$ the associated random predictor function. Then the PIM methodology can be extended to

$$\mathrm{P}[Y(t) \preccurlyeq Y'(t) \mid \boldsymbol{X}(t), \boldsymbol{X}'(t)] = m[\boldsymbol{X}(t), \boldsymbol{X}'(t); \boldsymbol{\beta}(t)], \quad t \in \mathcal{T}. \tag{8.2}$$

  Since the estimation of such a model will involve binary processes $\mathrm{I}[Y(t) \preccurlyeq Y'(t)]\, t \in \mathcal{T}$, with $\mathrm{I}\,(\cdot)$ the indicator function, as a first step, a simplified version of model (8.2) can be studied. For example, $Y(t)$ can be considered as a random outcome independent of t, i.e. $Y(t) = Y$. In their discussion of the probabilistic index model paper, Inácio et al. (2012) sketch how this problem can be tackled.

- *Additive modelling of the predictors.* In Chapter 5 we have used smoothing techniques to construct GOF methods. This can form the basis to extend PIMs so as to allow a more flexible modelling of the predictors. More specifically, an additive PIM can be defined as

$$\mathrm{P}\left(Y \preccurlyeq Y' \mid \boldsymbol{X}, \boldsymbol{X}'\right) = g^{-1}\left[\beta_0 + \sum_{j=1}^{d} f_j(X_j, X_j')\right],$$

  where X_j denotes the j^{th} element of the d-dimensional predictor $\boldsymbol{X}$ and with $f_j(\cdot)$ functions which are adaptively estimated. This additive modelling will considerably enlarge to applicability of the PIMs.

- *Censoring.* When censoring occurs, the asymptotic theory as developed in Chapters 2 and 7 does no longer apply. Cheng et al. (1995) describe how inverse weighting can be used to estimate parameters of models similar to PIMs. This may be used as a good starting point.

- *Robustness.* The PIM clearly has some robustness properties, e.g. it is robust in the outcome space since it models merely a relative ordering $\mathrm{P}\left(Y \preccurlyeq Y'\right)$. The modelling flexibility further allows a robust modelling in the predictor space by, for example, including predictors as $\mathrm{I}\left(X < X'\right)$. The robustness properties of the associated estimators can be studied in more detail and can be compared to some robust Z-estimators.

To end this dissertation, I would like to emphasis once more that, although the PIM can be the method of choice for some settings, other regression techniques, such as quantile regression, can provide a much richer and detailed analysis of the data. So the PIM is not intended to replace any statistical method, it is merely a new tool for the data analyst.

Appendix A

Probabilistic index models in R

The R-package `pim` is originally developed by Jan De Neve up to version 1.0.2. From version 1.1.0 the package has been further developed by Nick Sabbe who increased the flexibility and applicability substantially.

De Neve, J. and Sabbe, N. (2013). *pim: Probabilistic Index Models.* R package version 1.1.0.6/r22.

A.1 Installing the package

The package is available on R-forge `https://r-forge.r-project.org/R/?group_id=1120` and can be downloaded and installed as follows:

```
install.packages("pim", repos="http://R-Forge.R-project.org")
library("pim")
```

All outputs are generated by using R, version 2.15.1 (R Core Team, 2012).

The package has many functionalities and only a few will be discussed and illustrated in this chapter. The package is still under development and its functionality/implementation can be different in other versions. Up-to-date information can be found by consulting the help-files and vignettes in the usual way.

```
?pim
vignette("pim")
vignette("pim.legacy")
```

In Section A.2 we illustrate how the package can be used to obtain the fitted PIMs of the childhood respiratory disease study of Section 2.5.1. In Section A.3 we analyze the mental health study of Section 2.5.2, while in Section A.4 the analysis of the food expenditure study of Section 2.5.3 is considered. In Section A.5 it is shown how the package can be used to fit PIMs to factorial designs with the surgical unit study of Section 4.9 as an example.

A.2 The childhood respiratory disease study

The data can be loaded from the package as follows:

```
> data("FEVData")
> head(FEVData)
  Age   FEV Height Sex Smoke
1   9 1.708   57.0   0     0
2   8 1.724   67.5   0     0
3   7 1.720   54.5   0     0
4   9 1.558   53.0   1     0
5   9 1.895   57.0   1     0
6   8 2.336   61.0   0     0
> dim(FEVData)
[1] 654   5
```

We first consider the PIM with logit link and without interaction

$$\text{logit}[\text{P}\left(\text{FEV} \preccurlyeq \text{FEV}'\right)] \quad = \quad \beta_1(\text{AGE}' - \text{AGE}) + \beta_2(\text{SMOKE}' - \text{SMOKE}). \qquad \text{(A.1)}$$

This is an example of a *standard PIM*, i.e. a PIM with logit link and a covariate function of the form $\boldsymbol{Z} = \boldsymbol{X}' - \boldsymbol{X}$. Such a model can be fitted by using similar syntax as in the `lm()` and `glm()` functions.

```
> pim.fit1a <- pim(formula = FEV ~ Age + Smoke, data = FEVData)
> summary(pim.fit1a)

Call:
pim(formula = FEV ~ Age + Smoke, data = FEVData)

       Estimate Std. Error Z value Pr(>|z|)
Age    0.555035   0.027966 19.8466   <2e-16 ***
Smoke -0.457537   0.246376 -1.8571   0.0633 .
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1
```

It follows that $\hat{\beta}_1 = 0.56$ (SE : 0.028 and $p < 0.0001$) and $\hat{\beta}_2 = -0.46$ (SE : 0.25 and $p = 0.063$). By default the no-order restriction $\mathcal{X}_0$ is considered. Since PIM (A.1) is antisymmetric about one, the lexicographical order restriction $\mathcal{X}_{lex}$ can be used to obtain the same estimates. For a sample size $n = 654$, it holds that $|\mathcal{X}_{lex}| = n(n-1)/2 = 213531$, while for the no-order restriction this is $|\mathcal{X}_0| = n^2 = 427716$. So using the lexicographical order restriction will reduce the computation time; this can be obtained with the argument `poset = lexiposet`.

```
> pim.fit1b <- pim(formula = FEV ~ Age + Smoke, data = FEVData, poset = lexiposet)
> summary(pim.fit1b)

Call:
pim(formula = FEV ~ Age + Smoke, data = FEVData, poset = lexiposet)

       Estimate Std. Error Z value Pr(>|z|)
Age    0.555035   0.028081 19.7651  < 2e-16 ***
Smoke -0.457537   0.247016 -1.8523  0.06399 .
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1
```

The PIM with interaction

$$\begin{aligned} \text{logit}[\text{P}\left(\text{FEV} \preccurlyeq \text{FEV}'\right)] \;\; = \;\; & \gamma_1(\text{AGE}' - \text{AGE}) + \gamma_2(\text{SMOKE}' - \text{SMOKE}) + \\ & \gamma_3(\text{AGE}' * \text{SMOKE}' - \text{AGE} * \text{SMOKE}), \end{aligned} \tag{A.2}$$

can be fitted as follows:

```
> pim.fit2 <- pim(formula = FEV ~ Age * Smoke, data = FEVData)
> summary(pim.fit2)

Call:
pim(formula = FEV ~ Age * Smoke, data = FEVData)

           Estimate Std. Error Z value  Pr(>|z|)
Age        0.607600   0.029993 20.2582 < 2.2e-16 ***
Smoke      5.306885   1.040823  5.0987 3.419e-07 ***
Age:Smoke -0.455388   0.078275 -5.8178 5.963e-09 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1
```

It follows that $\hat{\gamma}_1 = 0.61$ (SE : 0.030 and $p < 0.0001$), $\hat{\gamma}_2 = 5.31$ (SE : 1.04 and $p < 0.0001$), and $\hat{\gamma}_3 = -0.46$ (SE : 0.078 and $p < 0.0001$).

A.3 The mental health study

We load the data.

```
> data("MHData")
> head(MHData)
  mental ses life
1      1   1    1
2      1   1    9
3      1   1    4
4      1   1    3
5      1   0    2
6      1   1    0
> dim(MHData)
[1] 40  3
```

We fit the logit PIM with main effects

$$\text{logit}\left[\text{P}\left(\text{MI} \preccurlyeq \text{MI}'\right)\right] = \beta_1(\text{SES}' - \text{SES}) + \beta_2(\text{LI}' - \text{LI}), \tag{A.3}$$

where MI = mental, SES = ses, and LI = life.

```
> pim.fit3 <- pim(formula = mental ~ ses + life, data = MHData)
> summary(pim.fit3)

Call:
pim(formula = mental ~ ses + life, data = MHData)

      Estimate Std. Error Z value Pr(>|z|)
ses  -0.740163   0.330491 -2.2396 0.025118 *
life  0.201179   0.070893  2.8378 0.004543 **
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1
```

If follows that $\hat{\beta}_1 = -0.74$ (SE : 0.33 and $p = 0.025$) and $\hat{\beta}_2 = 0.20$ (SE : 0.07 and $p = 0.0045$). Consider the more complicated PIM which is not of the standard form

$$\text{logit}\left[\text{P}\left(\text{MI} \preccurlyeq \text{MI}'\right)\right] = \gamma_1(\text{SES}' - \text{SES}) + \gamma_2(\text{LI}' - \text{LI}) + \gamma_3\text{SES} + \gamma_4\text{LI}, \tag{A.4}$$

defined for the strict lexicographical order restriction. This model can be fitted as follows:

```
> form.tmp <- mental ~ ses + life + L(ses) + L(life) - 1
> pim.fit4 <- pim(formula = form.tmp, data = MHData, poset = lexiposet, interpretation = "
    regular")
> summary(pim.fit4)

Call:
```

```
pim(formula = form.tmp, data = MHData, poset = lexiposet, interpretation = "regular")

           Estimate Std. Error Z value Pr(>|z|)
ses_R-_L  -0.670723   0.382665 -1.7528 0.079642 .
life_R-_L  0.205459   0.069989  2.9356 0.003329 **
ses_L     -0.034676   0.163157 -0.2125 0.831693
life_L    -0.021601   0.039843 -0.5422 0.587711
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1
```

The formula statement is now more complicated for PIM (A.4) since the covariate function cannot be written as $\boldsymbol{Z} = \boldsymbol{X}' - \boldsymbol{X}$. If we use the option `interpretation = "regular"`, then we can specify the right hand side of (A.4) by using the functions `L()` and `R()`, to indicate the covariate corresponding to the outcome in the left and right hand side of the inequality within the probability operator of the probabilistic index. For example, `L(ses)` is associated with SES while for `R(ses)` this is SES$'$. Predictors without the `L()` or `R()` function will be automatically converted to $\boldsymbol{Z} = \boldsymbol{X}' - \boldsymbol{X}$. Since the PIM (A.4) has no intercept, we add `-1` in the formula statement. Furthermore, the PIM is only defined for the strict lexicographical order restriction so that we use the option `poset = lexiposet`.

In the output, the functional form is explicitly shown by using underscores. For example, `ses_R-_L` stands for the difference of SES associated with the outcome in right hand side of the inequality in the PI minus the SES value associated with the outcome in the left hand side: SES$'$ $-$ SES. Similarly `ses_L` corresponds to SES.

It follows that $\hat{\gamma}_1 = -0.67$ (SE : 0.38 and $p = 0.080$), $\hat{\gamma}_2 = 0.21$ (SE : 0.07 and $p = 0.003$), $\hat{\gamma}_3 = -0.035$ (SE : 0.16 and $p = 0.83$), and $\hat{\gamma}_4 = -0.022$ (SE : 0.04 and $p = 0.59$). The general linear null hypothesis

$$H_0 : \gamma_3 = \gamma_4 = 0,$$

can be tested as follows:

```
> stat <- t(coef(pim.fit4)[3:4])%*%solve(vcov(pim.fit4)[3:4,3:4])%*%coef(pim.fit4)[3:4]
> p.tmp <- 1 - pchisq(stat, 2)
> stat
          [,1]
[1,] 0.5339321
> p.tmp
          [,1]
[1,] 0.7656991
```

A.4 The food expenditure study

We load the data and create a new variable: the relative food expenditure defined as FEP $=$ $100\text{FE}/\text{HI}$, with FE the food expenditure and HI the household income.

```
> data("Engeldata")
> head(Engeldata)
    income  foodexp
1 420.1577 255.8394
2 541.4117 310.9587
3 901.1575 485.6800
4 639.0802 402.9974
5 750.8756 495.5608
6 945.7989 633.7978
> dim(Engeldata)
[1] 235   2
> Engeldata$relfoodexp <- Engeldata$foodexp/Engeldata$income*100
```

Consider the standard PIM

$$\text{logit}[\text{P}\left(\text{FEP} \preccurlyeq \text{FEP}'\right)] = \beta(\text{HI}' - \text{HI}), \tag{A.5}$$

which is antisymmetric about one, so it is sufficient to fit the model according to the lexicographical order restriction.

```
> pim.fit5 <- pim(formula = relfoodexp ~ income, data = Engeldata, poset = lexiposet)
Warning message:
In .handleError(paste("Fit could not be obtained: nonconvergence of the algorithm:",  :
  Fit could not be obtained: nonconvergence of the algorithm: x-values within tolerance 'xtol'
> summary(pim.fit5)

Call:
pim(formula = relfoodexp ~ income, data = Engeldata, poset = lexiposet)

          Estimate  Std. Error Z value  Pr(>|z|)
income -0.00094061  0.00021243 -4.4279 9.516e-06 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1
```

It follows that $\hat{\beta} = -0.00094$ (SE : 0.0002 and $p < 0.0001$). There is a warning that the algorithm to find the roots of the estimating equations (2.15) did not convergence within the default tolerance (which is 10^{-6}). Therefore, we have a look at the function value of the estimating equation evaluated at the estimate:

```
> pim.fit5$morefitinfo$fvec
[1] -3.128659e-06
```

Since this value is close to zero, the non-convergence is not problematic.

Consider the more complicated heteroscedastic PIM of Section 3.2.3 with probit link

$$\mathrm{P}\left(\mathrm{FE} \preccurlyeq \mathrm{FE}' \mid \mathrm{HI}, \mathrm{HI}'\right) = \Phi\left[\frac{(\mathrm{HI}' - \mathrm{HI})}{\sqrt{\mathrm{HI}' + \mathrm{HI}}}\gamma\right]. \tag{A.6}$$

Since this PIM is not of the form $\boldsymbol{Z} = \boldsymbol{X}' - \boldsymbol{X}$ we use the `interpretation="regular"` option and the `L()` and `R()` functions in the formula statement. Since the right hand side of (A.6) involves some mathematical operators, we use the `I()` operator in the formula statement. With `link = "probit"` we can choose for the probit link instead of the default logit link. The PIM is antisymmetric about one so computational time can be gained by restricting the model to the lexicographical order restriction.

```
> form.tmp <- foodexp ~ I((R(income)-L(income))/sqrt(R(income)+L(income)))-1
> pim.fit6a <- pim(formula = form.tmp, data = Engeldata, link = "probit", interpretation = "
    regular", poset = lexiposet)
Warning message:
In .handleError(paste("Fit could not be obtained: nonconvergence of the algorithm:",  :
  Fit could not be obtained: nonconvergence of the algorithm: No better point found (algorithm
       has stalled)
> summary(pim.fit6a)

Call:
pim(formula = form.tmp, data = Engeldata, link = "probit", poset = lexiposet,
    interpretation = "regular")

                                                 Estimate Std. Error Z value  Pr(>|z|)
I((income_R-income_L)/sqrt(income_R+income_L)) 0.1324129  0.0079613  16.632 < 2.2e-16 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1
```

Once again the algorithm did not converge. Let us have a look at the function values of estimating equation evaluated at the estimate:

```
> pim.fit6a$morefitinfo$fvec
[1] 1.340781e+154
```

The estimate is clearly no root of the estimating equation. The standard algorithm to find the roots is `estimator.nleqslv()` of the `nleqslv` package (Hasselman, 2012). However, other algorithms are available, e.g. `estimator.glm()` of the `glm()` function.

```
> pim.fit6b <- pim(formula = form.tmp, data = Engeldata, link = "probit", estimator =
    estimator.glm(), interpretation = "regular", poset = lexiposet)
Warning message:
In eval(expr, envir, enclos) : non-integer #successes in a binomial glm!
```

```
> summary(pim.fit6b)

Call:
pim(formula = form.tmp, data = Engeldata, link = "probit", poset = lexiposet,
    interpretation = "regular", estimator = estimator.glm())

                                                Estimate Std. Error Z value  Pr(>|z|)
I((income_R-income_L)/sqrt(income_R+income_L)) 0.389705   0.071588  5.4437 5.218e-08 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1
> pim.fit6b$morefitinfo$converged
[1] TRUE
```

The warning message is a consequence of the definition of the pseudo-observations $\mathrm{I}(Y \preccurlyeq Y') := \mathrm{I}(Y < Y') + 0.5\mathrm{I}(Y = Y')$ which can take on three values: 0, 0.5, and 1. It follows that $\hat{\gamma} = 0.39$ (SE : 0.072 and $p < 0.0001$). The name of the estimated coefficient in the output (here `I((income_R-income_L)/sqrt(income_R+income_L))`) can sometimes be too long, especially for complicated PIMs. These names can be summarized by the user with `extra.nicenames` option.

```
> pim.fit7 <- pim(formula = form.tmp, data = Engeldata, link = "probit", estimator = estimator
    .glm(),
+ interpretation = "regular", extra.nicenames = data.frame(org = "I((R(income)-L(income))/sqrt
    (R(income)+L(income)))", nice = "Z"), poset = lexiposet)
Warning message:
In eval(expr, envir, enclos) : non-integer #successes in a binomial glm!
> summary(pim.fit7)

Call:
pim(formula = form.tmp, data = Engeldata, link = "probit", poset = lexiposet,
    interpretation = "regular", estimator = estimator.glm(),
    extra.nicenames = data.frame(org = "I((R(income)-L(income))/sqrt(R(income)+L(income)))",
        nice = "Z"))

  Estimate Std. Error Z value  Pr(>|z|)
Z 0.389705   0.071588  5.4437 5.218e-08 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1
```

A.5 The surgical unit study

We load the data:

```
> data("SUData")
> head(SUData)
```

```
  EnT Gender Alcohol SurvivalTime
1  81      0       1          695
2  66      0       0          403
3  83      0       0          710
4  41      0       0          349
5 115      0       2         2343
6  72      1       1          348
> dim(SUData)
[1] 54  4
```

We rename the variables to be consistent with the notation of Section 4.9.

```
> names(SUData) <- c("X1", "X2", "X3", "Y")
> summary(SUData)
       X1          X2     X3           Y
 Min.   : 23.00   0:29   0:15   Min.   : 181.0
 1st Qu.: 67.25   1:25   1:29   1st Qu.: 482.0
 Median : 79.00          2:10   Median : 605.5
 Mean   : 77.11                 Mean   : 702.1
 3rd Qu.: 89.50                 3rd Qu.: 750.5
 Max.   :119.00                 Max.   :2343.0
```

The marginal PIM

$$\begin{aligned}\text{logit}\left[\mathrm{P}\left(Y_i \preccurlyeq Y_j \mid \boldsymbol{X}_j\right)\right] &= \alpha_1 + \alpha_2 X_{1j} + \alpha_3 \mathrm{I}\left(X_{2j}=1\right) + \alpha_4 \mathrm{I}\left(X_{3j}=1\right) + \\ &\quad \alpha_5 \mathrm{I}\left(X_{3j}=2\right), \end{aligned} \tag{A.7}$$

can be fitted with the argument `interpretation = "marginal"`. The factors will be automatically converted to dummy-variables.

```
> pim.marginal <- pim(formula = Y~ X1 + X2 + X3, data = SUData, interpretation = "marginal")
> summary(pim.marginal)

Call:
pim(formula = Y ~ X1 + X2 + X3, data = SUData, interpretation = "marginal")

             Estimate Std. Error Z value  Pr(>|z|)
(Intercept) -3.3099991  0.6478452 -5.1092 3.235e-07 ***
X1_R         0.0352055  0.0096093  3.6637 0.0002486 ***
X2_R1        0.4272539  0.2757351  1.5495 0.1212595
X3_R1        0.3427110  0.3233750  1.0598 0.2892382
X3_R2        1.1090179  0.3799358  2.9190 0.0035120 **
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1
```

It follows that $\hat{\alpha}_1 = -3.3$ (SE : 0.65 and $p < 0.0001$), $\hat{\alpha}_2 = 0.035$ (SE : 0.0.0096 and $p = 0.0002$), $\hat{\alpha}_3 = 0.43$ (SE : 0.28 and $p = 0.12$), $\hat{\alpha}_4 = 0.34$ (SE : 0.32 and $p = 0.29$), and

$\hat{\alpha}_5 = 1.11$ (SE : 0.38 and $p = 0.0035$).

To fit the pairwise PIM

$$\begin{aligned}\text{logit}\left[\mathrm{P}\left(Y_i \preccurlyeq Y_j \mid \boldsymbol{X}_i, \boldsymbol{X}_j\right)\right] &= \beta_1(X_{1j} - X_{1i}) + \beta_2 \mathrm{I}\left(X_{2i} = 0\right) \mathrm{I}\left(X_{2j} = 1\right) + \\ &\quad \beta_3 \mathrm{I}\left(X_{3i} = 0\right) \mathrm{I}\left(X_{3j} = 1\right) + \beta_4 \mathrm{I}\left(X_{3i} = 0\right) \mathrm{I}\left(X_{3j} = 2\right) + \\ &\quad \beta_5 \mathrm{I}\left(X_{3i} = 1\right) \mathrm{I}\left(X_{3j} = 2\right), \end{aligned} \tag{A.8}$$

we use the `F()` function in the formula statement to indicate that we want to consider all unique pairwise comparisons of the factor and set the interpretation to `interpretation = "regular"`. The model is defined for the strict lexicographical order restriction which is obtained by setting `poset = lexiposet`. Since the lexicograpical order restriction assumes that the predictors can be ordered, we first need to redefine them.

```
> SUData$X2 <- factor(SUData$X2, ordered = TRUE)
> SUData$X3 <- factor(SUData$X3, ordered = TRUE)
> form.tmp <- Y ~ X1 + F(X2) + F(X3) - 1
> pim.pairwise <- pim(formula = form.tmp, data=SUData, interpretation = "regular",
+                     poset = lexiposet)
> summary(pim.pairwise)

Call:
pim(formula = form.tmp, data = SUData, poset = lexiposet, interpretation = "regular")

           Estimate Std. Error Z value Pr(>|z|)
X1_R-_L    0.028222   0.013865  2.0354  0.04181 *
X2_L_R_0_1 0.574692   0.399510  1.4385  0.15029
X3_L_R_0_1 0.714924   0.441137  1.6206  0.10510
X3_L_R_0_2 2.040005   0.978036  2.0858  0.03700 *
X3_L_R_1_2 1.269355   0.693014  1.8316  0.06700 .
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1
```

It follows that $\hat{\beta}_1 = 0.028$ (SE : 0.014 and $p = 0.042$), $\hat{\beta}_2 = 0.57$ (SE : 0.40 and $p = 0.15$), $\hat{\beta}_3 = 0.71$ (SE : 0.44 and $p = 0.11$), $\hat{\beta}_4 = 2.04$ (SE : 0.98 and $p = 0.037$), and $\hat{\beta}_5 = 1.27$ (SE : 0.69 and $p = 0.064$).

Bibliography

Acion, L., Peterson, J., Temple, S., and Arndt, S. (2006). Probabilistic index: an intuitive non-parametric approach to measuring the size of treatment effects. *Statistics in Medicine*, 25:591–602.

Agresti, A. (2007). *An Introduction to Categorical Data Analysis*. Wiley, New Jersey, USA.

Agresti, A. (2010). *Analysis of Ordinal Categorical Data*. Wiley, New Jersey, USA.

Akritas, M. (1990). The rank transform method in some two factor designs. *Journal of the American Statistical Association*, 85:73–78.

Akritas, M. and Arnold, S. (1994). Fully nonparametric hypotheses for factorial designs I: multivariate repeated measures designs. *Journal of the American Statistical Association*, 89:336–343.

Akritas, M., Arnold, S., and Brunner, E. (1997). Nonparametric hypotheses and rank statistics for unbalanced factorial designs. *Journal of the American Statistical Association*, 92:258–265.

Akritas, M., Arnold, S., and Du, Y. (2000). Nonparametric models and methods for nonlinear analysis of covariance. *Biometrika*, 87:507–526.

Alaminos, M., Mora, J., Cheung, N., Smith, A., Qin, J., Chen, L., and Gerald, W. L. (2003). Genome-wide analysis of gene expression associated with MYCN in humana neuroblastoma. *Cancer Research*, 63:4538–4546.

Beck, A., Steer, R., and Garbin, M. (1988). Psychometric properties of the beck depression inventory: twenty-five years of evaluation. *Clinical Psychology Review*, 8:77–100.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society - Series B*, 57:289–300.

Bergsma, W., Croon, M., and Hagenaars, J. (2009). *Marginal Models for Dependent, Clustered and Longitudinal Categorical Data*. Springer, New York, USA.

Bergsma, W., Croon, M., Hagenaars, J., and van der Ark, A. (2012). Discussion of "Probabilistic Index Models" by O.Thas, J. De Neve, L. Clement and J.P. Ottoy. *Journal of the Royal Statistical Society - Series B*, 74:623–671.

Beyerlein, A., Fahrmeir, L., Mansmann, U., and Toschke, A. (2008). Alternative regression models to assess increase in childhood BMI. *BMC Medical Research Methodology*, 8.

Bickel, P., Klaassen, C., Ritov, Y., and Wellner, J. (1998). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, New Jersey, USA.

Bradley, R. A. and Terry, M. R. (1952). The rank analysis of incomplete block designs. I. The method of paired comparisons. *Biometrika*, 39:324–345.

Brown, B., Hettmansperger, R., and Hettmansperger, T. (2006). *The logistic distribution and a rank test for non-transitivity. Random Walk, Sequential Analysis, and Related Topics. A Festschrift in Honor of Yuan-Shih Chow*. World Scientific, New Jersey, USA.

Brown, B. and Hettmansperger, T. (2002). Kruskal–Wallis, multiple comparisons and Efron dice. *Australian and New Zealand Journal of Statistics*, 44:427–438.

Browne, R. (2010). The t-test p value and its relationship to the effect size and $P(X > Y)$. *The American Statistician*, 64:30–33.

Brumback, L., Pepe, M., and Alonzo, T. (2006). Using the ROC curve for gauging treatment effect in clinical trials. *Statistics in Medicine*, 25:575–590.

Brunner, E. and Puri, M. (2002). A class of rank-score tests in factorial designs. *Journal of Statistical Planning and Inference*, 103:331–360.

Buyse, M. (2010). Generalized pairwise comparisons of prioritized outcomes in the two-sample problem. *Statistics in Medicine*, 29:3245–3257.

Carroll, R. J. and Ruppert, D. (1988). *Transformation and Weighting in Regression*. Chapman and Hall, New York, USA.

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, 34:305–334.

Cheng, S., Wei, L., and Ying, Z. (1995). Analysis of transformation models with censored data. *Biometrika*, 92:835–845.

Chu, C. and Marron, J. (1991). Comparison of two bandwidth selectors with dependent errors. *Annals of Statistics*, 19:1906–1918.

Church, J. D. and Harris, B. (1970). The estimation of reliability from stress-strength relationships. *Technometrics*, 12:49–54.

Conover, W. and Iman, R. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician*, 35:124–133.

Cox, D. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society - Series B*, 34:187–220.

D'Agostino, R. B., Campbell, M., and Greenhouse, J. (2006). The Mann–Whitney statistic: continuous use and discovery. *Statistics in Medicine*, 25:541–542.

de Groot, L. and Zock, P. (1998). Moderate alcohol intake and mortality. *Nutrition Reviews*, 56:25–26.

de Kroon, J. and van der Laan, P. (1981). Distribution-free test procedures in two-way layouts: a concept of rank interaction. *Statistica Neerlandica*, 35:189–213.

De Neve, J. and Sabbe, N. (2013). *pim: Probabilistic Index Models*. R package version 1.1.0.6/r22.

De Neve, J., Thas, O., and Ottoy, J.P. (2013a). Goodness-of-fit methods for probabilistic index models. *Communications in Statistics - Theory and Methods*, 42:1193–1207.

De Neve, J., Thas, O., and Ottoy, J.P. (2013b). A semiparametric framework for rank tests for factorial designs. *Submitted*.

De Neve, J., Thas, O., Ottoy, J.P., and Clement, L. (2013c). An extension of the Wilcoxon–Mann–Whitney test for analyzing RT-qPCR data. *Statistical Applications in Genetics and Molecular Biology (in press)*.

Derveaux, S., Vandesompele, J., and Hellemans, J. (2010). How to do successful gene expression analysis using real-time PCR. *Methods*, 50:227 – 230.

Dodd, L. and Pepe, M. (2003). Semi-parametric regression for the area under the receiver operating characteristics curve. *Journal of the American Statistical Association*, 98:409–417.

Enis, P. and Geisser, S. (1971). Estimation of the probability that $Y < X$. *Journal of the American Statistical Association*, 66:162–168.

Evans, S. and Li, L. (2005). A comparison of goodness of fit tests for the logistic GEE model. *Statistics in Medicine*, 24:1245–1261.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. New York: Chapman & Hall.

Fishburn, P. C. (1974). Lexicographic orders, utilities and decision rules: a survey. *Management Science*, 20:1442–1471.

Fligner, M. (1985). Pairwise versus joint ranking: another look at the Kruskal–Wallis statistic. *Biometrika*, 72:705–709.

Fligner, M. and Policello, G. (1981). Robust rank procedures for the Behrens–Fisher problem. *Journal of the American Statistical Association*, 76:162–168.

Follmann, D. (2002). Regression analysis based on pairwise ordering of patients' clinical histories. *Statistics in Medicine*, 21:3353–3367.

Fontana, L., Fiori, M., Albini, S., Cifaldi, L., Giovinazzi, S., Forloni, M., Boldrini, R., Donfrancesco, A., Federici, V., Giacommi, P., Peschele, C., and Fruci, D. (2008). Antagomir-17-5p abolishes the growth of therapy-resistant neuroblastoma through p21 and BIM. *PLoS ONE*, 3:e2236.

Foster, D. (2010). A reduced mortality and moderate alcohol consumption: the phospholipase D-mTOR connection. *Cell Cycle*, 9:1291–1294.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32:675–701.

Gardner, M. (1970). The paradox of the nontransitive dice and the elusive principle of indifference. *Scientific American*, 223:110–114.

Gerds, T. (2012). Discussion of "Probabilistic Index Models" by O.Thas, J. De Neve, L. Clement and J.P. Ottoy. *Journal of the Royal Statistical Society - Series B*, 74:623–671.

Gerds, T., Kattan, M., Schumacher, M., and Yu, C. (2010). Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. Technical Report 10/7, Department of Biostatistics - University of Copenhagen.

Gillen, D. and Emerson, S. (2007). Nontransitivity in a class of weighted logrank statistics under nonproportional hazards. *Statistics and Probability Letters*, 77:123–130.

Grissom, R. (1994). Probability of the superior outcome of one treatment over another. *Journal of Applied Psychology*, 79:314–316.

Groenwold, R., Moons, K., Peelen, L., Knol, M., and Hoes, A. (2011). Reporting of treatment effects from randomized trials: a plea for multivariable risk ratios. *Contemporary Clinincal Trials*, 32:399–402.

Guescini, M., Sisti, D., Rocchi, M., Stocchi, L., and Stocchi, V. (2008). A new real-time PCR method to overcome significant quantitative inaccuracy due to slight amplification inhibition. *BMC Bioinformatics*, 9:326.

Hájek, J., Šidák, Z., and Sen, P. K. (1999). *Theory of Rank Tests*. Academic Press, San Diego, USA.

Halperin, M., Gilbert, P., and Lachin, J. (1987). Distribution-free confidence intervals for $Pr(X_1 < X_2)$. *Biometrics*, 43:71–80.

Hand, D. (1992). On comparing two treatments. *The American Statistician*, 46:190–192.

Hardle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *Annals of Statistics*, 21:1926–1947.

Harrell, F., Califf, R., Pryor, D., Lee, K., and Rosati, R. (1982). Evaluating the yield of medical tests. *Journal of the American Medical Association*, 247:2543–2546.

Harrell, F., Lee, K., and Mark, D. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15:361–387.

Hasselman, B. (2012). *nleqslv: Solve systems of non linear equations*. R package version 1.9.3.

Hodges, J. and Lehmann, E. (1963). Estimation of location based on ranks. *Annals of Mathematical Statistics*, 34:598–611.

Højsgaard, S., Halekoh, U., and Yan, J. (2005). The R package geepack for generalized estimating equations. *Journal of Statistical Software*, 15:1–11.

Hollander, M. and Wolfe, D. (1999). *Nonparametric Statistical Methods*. Wiley, New York, USA.

Holt, J. and Prentice, R. (1974). Survival analysis in twin studies and matched pair experiments. *Biometrika*, 61:17–30.

Hora, S. and Conover, W. (1984). The F statistic in the two-way layout with rank-score transformed data. *Journal of the American Statistical Association*, 79:668–673.

Hosmer, D. W. and Lemeshow, S. (2000). *Applied logistic regression (Wiley Series in probability and statistics)*. Wiley Interscience, New York; USA.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35:73–101.

Inácio, V., de Carvalho, M., and Turkman, A. A. (2012). Discussion of "Probabilistic Index Models" by O.Thas, J. De Neve, L. Clement and J.P. Ottoy. *Journal of the Royal Statistical Society - Series B*, 74:623–671.

Jonckheere, A. R. (1954). A distribution-free K-sample test against ordered alternatives. *Biometrika*, 41:133–145.

Kakade, C., Shirke, D., and Kundu, D. (2008). Inference for $P(Y < X)$ in exponentiated gumbel distribution. *Journal of Statistics and Applications*, 3:121–133.

Kalbfleish, J. and Prentice, R. (1973). Marginal likelihoods based on Cox's regression and life model. *Biometrika*, 60:267–278.

Kieser, M., Friede, T., and Gondan, M. (2012). Assessment of statistical significance and clinical relevance. *Statistics in Medicine*, doi:10.1002/sim.5634.

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press.

Koenker, R. (2011). *quantreg: Quantile Regression*. R package version 4.54.

Koziol, J. and Jia, Z. (2009). The concordance index C and the Mann–Whitney parameter $\Pr(X > Y)$ with randomly censored data. *Biometrical Journal*, 51:467–474.

Koziol, J. and Reid, N. (1977). On the asymptotic equivalence of two ranking methods for K-sample linear rank statistics. *Annals of Statistics*, 5:1099–1106.

Kruskal, W. and Wallis, W. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47:583–621.

Kutner, M. H., Nachtsheim, C. J., and Neter, J. (2004). *Applied Linear Regression Models*. McGraw-Hill/Irwin, fourth international edition.

Laine, C. and Davidoff, F. (1996). Patient-centered medicine: a professional evolution. *Journal of the American Medical Association*, 275:152–156.

Lalam, N. (2007). Statistical inference for quantitative polymerase chain reaction using a hidden Markov model: a Bayesian approach. *Statistical Applications in Genetics and Molecular Biology*, 1:1.

le Cessie, S. and van Houwelingen, J. (1991). A goodness-of-fit test for binary regression models, based on smoothing methods. *Biometrics*, 47:1267–1282.

Lehmann, E. (1998). *Nonparametrics. Statistical methods based on ranks*. Prentice Hall, Upper Saddle River, New Jersey, USA.

Liang, K. and Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13–22.

Linton, O., Sperlich, S., and Van Keilegom, I. (2008). Estimation of a semiparametric transformation model. *Annals of Statistics*, 36:686–718.

Liu, I. and Agresti, A. (2005). The analysis of ordered categorical data: an overview and a survey of recent developments. *Test*, 14:1–73.

Loève, M. (1963). *Probability Theory (third edition)*. Springer-Verlag, Berlin, Germany.

Luenberger, D. (1969). *Optimization by Vector Space Methods*. Wiley, New York, USA.

Lumley, T. and Mayer-Hamblett, N. (2003). Asymptotics for marginal generalized linear models with sparse correlations. Technical Report 207, UW Biostatistics Working Paper Series, University of Washington.

Mack, G. and Skillings, J. (1980). A Friedman-type rank test for main effects in a two-factor ANOVA. *Journal of the American Statistical Association*, 75:947–951.

Mack, G. and Wolfe, D. (1981). K-sample rank tests for umbrella alternatives. *Journal of the American Statistical Association*, 76:175–181.

Mann, H. (1945). Non-parametric tests against trend. *Econometrica*, 13:254–259.

Mann, H. and Whitney, D. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18:50–60.

Marden, J. and Muyot, M. (1995). Rank tests for main and interaction effects in analysis of variance. *Journal of the American Statistical Association*, 90:1388–1398.

McCullagh, P. (1980). Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society - Series B*, 42:109–142.

McGraw, K. and Wong, S. (1992). A common language effect size statistic. *Psychological Bulletin*, 111:361–365.

McKean, J. (2004). Robust analysis of linear models. *Statistical Science*, 19:562–570.

McKean, J., Terpstra, J., and Kloke, J. (2009). Computational rank-based statistics. *WIREs computational statistics*, 1:132–140.

Mestdagh, P., Van Vlierberghe, P., De Weer, A., Muth, D., Westermann, F. Speleman, F., and Vandesompele, J. (2009). A novel and universal method for microRNA RT-qPCR data normalization. *Genome Biology*, 10:R64.

Molenbergs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. Springer Series in Statistics, Springer-Verlag, New-York,.

Myles, P., Troedel, S., Boquest, M., and Reeves, M. (1999). The pain visual analog scale: is it linear or nonlinear? *Anesthesia and Analgesia*, 89:1517–1520.

Nadaraya, E. (1964). On estimating regression. *Theory of Probability and its Applications*, 9:141–142.

Newcombe, R. (2006a). Confidence intervals for an effect size measure based on the Mann–Whitney statistic. Part 1: general issues and tail-area-based methods. *Statistics in Medicine*, 25:543–557.

Newcombe, R. (2006b). Confidence intervals for an effect size measure based on the Mann–Whitneystatistic. Part 2: asymptotic methods and evaluation. *Statistics in Medicine*, 25:559–573.

Newey, W. (1988). Adaptive estimation of regression models via moment restrictions. *Journal of Econometrics*, 38:301–339.

Newey, W. (1989). *Locally Efficient, Residual-based Estimation of Nonlinear Simultaneous Equations*. Research memorandum. Econometric Research Program, Princeton University.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. In Engle, R. F. and McFadden, D. L., editors, *Handbook of Econometrics*, volume 4, pages 2111–2245. Elsevier.

O'Donnell, K., Wentzel, E., Zeller, K., Dang, C., and Mendell, J. (2005). c-Myc-regulated microRNAs modulate E2F1 expression. *Nature*, 435:839–843.

Pan, W. (2002). Goodness-of-fit tests for GEE with correlated binary data. *Scandinavian Journal of Statistics*, 29:101–110.

Parzen, E. and Mukhopadhyay, S. (2012a). Discussion of "Probabilistic Index Models" by O.Thas, J. De Neve, L. Clement and J.P. Ottoy. *Journal of the Royal Statistical Society - Series B*, 74:623–671.

Parzen, E. and Mukhopadhyay, S. (2012b). Modeling, dependence, classification, united statistical science, many cultures. Technical report, Preprint arXiv: 1204.4699.

Patel, K. and Hoel, D. (1973). A nonparametric test for interaction in factorial experiments. *Journal of the American Statistical Association*, 68:615–620.

Pepe, M. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press, Oxford, UK.

Pollard, K. S., Gilbert, H. N., Ge, Y., Taylor, S., and Dudoit, S. (2010). *multtest: Resampling-based multiple hypothesis testing*. R package version 2.5.14.

Powell, J. (1994). Estimation of semiparametric models. In Engle, R. F. and McFadden, D. L., editors, *Handbook of Econometrics*, volume 4, pages 2443–2521. Elsevier.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rosner, B. (1999). *Fundamentals of Biostatistics*. Pacific Grove: Duxbury.

Schulte, J. H., Horn, S., Otto, T., Samans, B., Heukamp, L. C., Eilers, Ursula-Christa, Krause, M., Astrahantseff, K., Klein-Hitpass, L., Buettner, R., Schramm, A., Christiansen, H., Eilers, M., Eggert, A., and Berwanger, B. (2008). MYCN regulates oncogenic MicroRNAs in neuroblastoma. *International Journal of Cancer*, 122:699–704.

Senn, S. (1997). Testing for individual and population equivalence based on the proportion of similar responses. *Statistics in Medicine*, 16:1303–1306.

Senn, S. (2006). Probabilistic index: an intuitive non-parametric approach to measuring the size of treatment effects by L. Acion, J. Peterson, S. Temple and S. Arndt. *Statistics in Medicine*, 25:3944–3948.

Senn, S. (2011). U is for unease: reasons for mistrusting overlap measures for reporting clinical trials. *Statistics in Biopharmaceutical Research*, 3:302–309.

Senn, S. (2012). Discussion of "Probabilistic Index Models" by O.Thas, J. De Neve, L. Clement and J.P. Ottoy. *Journal of the Royal Statistical Society - Series B*, 74:623–671.

Serfling, R. (1980). *Approximation Theorems of Mathematical Statistics*. New York: Wiley.

Sherman, R. (1994). U-processes in the analysis of a generalized semiparametric regression estimator. *Econometric Theory*, 10:372–395.

Silverman, B. (1986). *Density Estimation*. Chapman and Hall. London, U.K.

Stefanski, L. A. and Boos, D. D. (2002). The calculus of M-estimation. *The American Statistician*, 56:29–38.

Su, J. and Wei, L. (1991). A lack-of-fit test for the mean function in a generalized linear model. *Journal of the American Statistical Association*, 86:420–426.

Terpstra, T. (1952). The asymptotic normality and consistency of Kendall's test against trend, when ties are present in one ranking. *Indagationes Mathematicae*, 14:327–333.

Thangavelu, K. and Brunner, E. (2007). Wilcoxon–Mann–Whitney test for stratified samples and Efron's paradox dice. *Journal of Statistical Planning and Inference*, 137:720–737.

Thas, O. (2009). *Comparing Distributions*. Springer, New York, USA.

Thas, O., Best, D., and Rayner, J. (2012a). Using orthogonal trend contrasts for testing ranked data with ordered alternatives. *Statistica Neerlandica*, 66:452–471.

Thas, O., Clement, L., Rayner, J., Carvalho, B., and Van Criekinge, W. (2012b). An omnibus consistent adaptive percentile modified wilcoxon rank sum test with applications in gene expression studies. *Biometrics*, 68:446–454.

Thas, O., De Neve, J., Clement, L., and Ottoy, J. (2012c). Discussion of "Probabilistic Index Models" by O.Thas, J. De Neve, L. Clement and J.P. Ottoy. *Journal of the Royal Statistical Society - Series B*, 74:623–671.

Thas, O., De Neve, J., Clement, L., and Ottoy, J.P. (2012d). Probabilistic index models (with discussion). *Journal of the Royal Statistical Society - Series B*, 74:623–671.

Therneau, T. and Lumley, T. (2010). *survival: Survival analysis, including penalised likelihood.* R package version 2.36-2.

Tian, L. (2008). Confidence intervals for $P(Y_1 > Y_2)$ with normal outcomes in linear models. *Statistics in Medicine*, 27:4221–4237.

Tomlins, S. A., Rhodes, D. R., Perner, S., Dhanasekaran, S. M., Mehra, R., Sun, Xiao-Wei, Varambally, S., Cao, X., Tchinda, J., Kuefer, R., Lee, C., Montie, J. E., Shah, R. B., Pienta, K. J., Rubin, M. A., and Chinnaiyan, A. M. (2005). Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science*, 310:644–648.

Tsangari, H. and Akritas, M. (2004). Nonparametric ANCOVA with two and three covariates. *Journal of Multivariate Analysis*, 88:298–319.

Tsiatis, A. (2006). *Semiparametric Theory and Missing Data*. Springer, New York, USA.

Turk, D., Rudy, T., and Sorkin, B. (1993). Neglected topics in chronic pain treatment outcome studies: determination of success. *Pain*, 53:3–16.

van de Wiel, M. (2012). Discussion of "Probabilistic Index Models" by O.Thas, J. De Neve, L. Clement and J.P. Ottoy. *Journal of the Royal Statistical Society - Series B*, 74:623–671.

Van den Eynde, F., Senturk, V., Naudts, K., Vogels, C., Bernagie, K., Thas, O., van Heeringen, C., and Audenaert, K. (2008). Efficacy of quetiapine for impulsivity and affective symptoms in borderline personality disorder. *Journal of Clinical Psychopharmacology*, 28:147–155.

van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge, United Kingdom.

Van Keilegom, I. (2012). Discussion of "Probabilistic Index Models" by O.Thas, J. De Neve, L. Clement and J.P. Ottoy. *Journal of the Royal Statistical Society - Series B*, 74:623–671.

Vandesompele, J., De Preter, K., Pattyn, F., Poppe, B., Van Roy, N., De Paepe, A., and Speleman, F. (2002). Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biology*, 3 (7):0034.1–0034.11.

VanGuilder, H., Vrana, K., and Freeman, W. (2008). Twenty-five years of quantitative PCR for gene expression analysis. *BioTechniques*, 44:619–626.

Vansteelandt, S. (2012). Discussion of "Probabilistic Index Models" by O.Thas, J. De Neve, L. Clement and J.P. Ottoy. *Journal of the Royal Statistical Society - Series B*, 74:623–671.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, USA, fourth edition. ISBN 0-387-95457-0.

Vermeulen, J., Preter, K. D., Naranjo, A., Vercruysse, L., Roy, N. V., Hellemans, J., Swerts, K., Bravo, S., Scaruffi, P., Tonini, G. P., Bernardi, B. D., Noguera, R., Piqueras, M., Caete, A., Castel, V., Janoueix-Lerosey, I., Delattre, O., Schleiermacher, G., Michon, J., Combaret, V., Fischer, M., Oberthuer, A., Ambros, P. F., Beiske, K., Bnard, J., Marques, B., Rubie, H.,

Kohler, J., Ptschger, U., Ladenstein, R., Hogarty, M. D., McGrady, P., London, W. B., Laureys, G., Speleman, F., and Vandesompele, J. (2009). Predicting outcomes for children with neuroblastoma using a multigene-expression signature: a retrospective SIOPEN/COG/GPOH study. *Lancet Oncology*, 7:663–71.

Wallerstein, S. (1984). *Scaling Clinical Pain and Pain Relief*. Elsevier, New York, USA.

Wasserman, L. (2007). *All of Nonparametric Statistics*. Springer, New York, USA.

Watson, G. (1964). Smooth regression analysis. *Sankhya: The Indian Journal of Statistics, Series A*, 26:359–372.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1:80–83.

Wolfe, D. and Hogg, R. (1971). On constructing statistics and reporting data. *The American Statistician*, 25:27–30.

Wong, W. H. (1983). On the consistency of cross-validation in kernel nonparametric regression. *Annals of Statistics*, 11:1136–1141.

Wu, Z. and Irizarry, R. A. (2007). A statistical framework for the analysis of microarray probe-level data. *The Annals of Applied Statistics*, 1:333–357.

Zeger, S. and Liang, K. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 42:121–130.

Zhou, W. (2008). Statistical inference for $P(X < Y)$. *Statistics in Medicine*, 27:257–279.

Samenvatting

Een klassiek probleem in statistiek bestudeert de associatie tussen een enkelvoudige uitkomstvariabele Y en een d-dimensionale onafhankelijke variabele $\boldsymbol{X}$. De variabele Y kan bijvoorbeeld een maat zijn voor de longinhoud van een minderjarige, terwijl $\boldsymbol{X}$ de leeftijd, het geslacht en de rookstatus van de minderjarige voorstelt. De focus ligt dan op het bestuderen van de associatie tussen het rookgedrag en de longinhoud, terwijl mogelijke confouding factoren, zoals geslacht en leeftijd, in rekening moeten worden gebracht. Men kan een bepaald type regressiemodel gebruiken om dergelijke onderzoeksvraag te bestuderen. Een populaire keuze is het regressiemodel dat de gemiddelde uitkomst modelleert. Het gemiddelde is echter niet altijd relevant of andere samenvattingen van de uitkomstvariabele kunnen interessant zijn. Beschouw als voorbeeld de mentale toestandsscore van een patiënt (Y), met $\boldsymbol{X}$ zijn/haar levensindex en socio-economische status. Men wenst de relatie tussen de socio-economische status en de mentale toestand te onderzoeken terwijl men ook rekening wenst te houden met de levensindex. De mentale toestandsscore kan uitgedrukt worden op een schaal met 4 niveaus: $Y = 1$ (een gezonde mentale toestand), $Y = 2$ (een milde psychische aandoening), $Y = 3$ (een gematigde psychische aandoening) en $Y = 4$ (een sterke psychische aandoening). De gemiddelde mentale toestandsscore heeft geen eenduidige interpretatie vermits de uitkomstvariabele ordinaal is.

In de thesis stellen we een nieuw regressieraamwerk voor dewelke toelaat de associatie tussen Y and $\boldsymbol{X}$ te bestuderen voor zowel ordinale, interval en ratio-schaal uitkomstvariabelen Y. In het bijzonder modelleren we de kans dat de uitkomstvariabele toeneemt in functie van de onafhankelijke variabelen. Deze kans wordt de *probabilistic index* (PI) genoemd. Indien $(Y, \boldsymbol{X})$ en $(Y', \boldsymbol{X}')$ onafhankelijk en gelijk verdeelde variabelen voorstellen, dan is de PI gedefinieerd als

$$\mathrm{P}\left(Y \preccurlyeq Y' \mid \boldsymbol{X}, \boldsymbol{X}'\right) := \mathrm{P}\left(Y < Y' \mid \boldsymbol{X}, \boldsymbol{X}'\right) + \frac{1}{2}\mathrm{P}\left(Y = Y' \mid \boldsymbol{X}, \boldsymbol{X}'\right).$$

Een *probabilistic index model* (PIM) modelleert deze kans in functie van de onafhankelijke

variabelen $\boldsymbol{X}$ en $\boldsymbol{X}'$. Meer bepaald is een PIM gedefinieerd als

$$\mathrm{P}\left(Y \preccurlyeq Y' \mid \boldsymbol{X}, \boldsymbol{X}'\right) = m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}), \quad (\boldsymbol{X}, \boldsymbol{X}') \in \mathcal{X}, \tag{A.9}$$

met $m(\cdot)$ een functie met bereik $[0, 1]$, $\mathcal{X}$ de verzameling van koppels onafhankelijke variabelen waarvoor het model gedefinieerd is en $\boldsymbol{\beta}$ is de p-dimensionale parameter vector.

In deze thesis zijn zeven contributies aan de PIM methodologie voorgesteld.

1. We introduceren de PIM op een formele wijze samen met semiparametrische schatters en asymptotische distributietheorie. De resultaten van een simulatiestudie geven aan dat de asymptotische benaderingen geldig zijn voor eindige steekproefgroottes. Verschillende datasets zijn geanalyseerd geweest met behulp van een PIM om de interpretatie en flexibiliteit van de methode te illustreren.

2. We hebben de PIM gesitueerd binnen het landschap van statistische methodes door middel van de relatie tussen een PIM en meer conventionele technieken te onderzoeken.

   Om de functionele vorm $m(\cdot)$ in (A.9) beter te begrijpen, hebben we de relatie bestudeerd tussen een PIM en normale lineaire modellen (NLM) en Cox proportionele hazards modellen (CPHM). Er volgt dat er een directe relatie is tussen de modelparameters van een PIM en de modelparameters van een NLM, respectievelijk CPHM. Deze verbanden suggereren een functionele vorm $m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}) = g^{-1}[(\boldsymbol{X}' - \boldsymbol{X})^T\boldsymbol{\beta}]$ waar $g(\cdot)$ de logit of probit linkfunctie is. Deze keuze resulteert in een intuïtieve interpretatie van de modelparameters. Voor een enkelvoudige continue onafhankelijke variabele X volgt dat $g(\beta) = \mathrm{P}\left(Y \preccurlyeq Y' \mid X = x, X' = x + 1\right)$, i.e. de kans dat de uitkomstvariabele waarvoor $X = x + 1$ groter is dan de uitkomstvariabele waarvoor $X = x$.

   De PIM voorziet ook een regressieraamwerk voor de concordantie index en breidt de AUC-regressie methodologie uit. Verder zijn er ook interessante gelijkenissen en verschillen tussen een PIM en rankregressie, respectievelijk cumulatieve logit modellen.

3. We hebben de verbanden bestudeerd tussen een PIM en populaire ranktesten. Onder andere de Wilcoxon–Mann–Whitney, Kruskal–Wallis en Friedman ranktest kunnen ingebed worden in het PIM raamwerk. Deze inbedding laat toe om deze ranktesten uit te breiden naar meer complexe designs terwijl een intuïtieve interpretatie behouden blijft. Verder laat deze inbedding ook toe om voor de overeenkomstige effectmaten betrouwbaarheidsintervallen te construeren.

4. We hebben *Goodness-Of-Fit* (GOF) methoden ontwikkeld om de geschiktheid van het vooropgesteld model $m(\cdot)$ in (A.9) na te gaan. Een grafisch diagnostische figuur toont hoe het model kan verbeterd worden en een test om formeel de GOF te testen is beschikbaar. Beide methoden zijn gerelateerd aan de interpretatie met een PIM en zijn gebaseerd op de kans $\mathrm{P}\left(Y \preccurlyeq Y' \mid \boldsymbol{X}, \boldsymbol{X}' = \boldsymbol{X} + \boldsymbol{\Delta}\right)$ voor een vaste waarde $\boldsymbol{\Delta}$.

5. We hebben een toepassing in detail uitgewerkt. Meer bepaald hebben we de PIM methodologie gebruikt om genomische studies gebaseerd op *reverse transcription quantitative polymerase chain reaction* (RT-qPCR) te analyseren. Data gegenereerd door RT-qPCR vereisen een normalisatie om zo technische variatie in rekening te brengen. Verder worden RT-qPCR experimenten vaak gebruikt om differentieel geexpresseerde genen – die bijvoorbeeld ontdekt zijn door microarrays – te valideren. Dit impliceert dat het kwantificeren en interpreteren van de resultaten belangrijk is om inzicht te verwerven in de biologische processen die bestudeerd worden. Een PIM blijkt geschikt te zijn voor beide doelen: het laat toe de data te normaliseren en heeft een intuïtieve interpretatie in termen van de waarschijnlijkheid op neer- en opregulatie.

6. We hebben efficiënte schatters voor de PIM binnen een semiparametrische context bestudeerd. De indexfunctie die geassocieerd is met de efficiënte score komt overeen met de oplossing van een integraalvergelijking. De resultaten van een beknopte simulatiestudie geven aan dat de variantie van de efficiënte schatter ongeveer gelijk is aan de variantie van de schatter die gebruikt maakt van de onafhankelijke werk-correlatie matrix. Deze resultaten moeten echter nog verder in detail worden bestudeerd.

7. We hebben een R-pakket ontwikkeld waarmee de meeste voorbeelden van deze thesis kunnen geanalyseerd worden.

Niettegenstaande in deze thesis een aanzienlijk aantal van de basis technieken om een PIM te fitten zijn ontwikkeld, moeten nog veel uitbreidingen onderzocht worden zodanig de PIMs een breder toepassingsdomein kunnen bestrijken.

Summary

A classical problem in statistics is concerned with studying the association between a univariate outcome Y and a d-dimensional set of predictors $\boldsymbol{X}$. As an example, Y can be a measure of a child's lung capacity and $\boldsymbol{X}$ contains the age, gender, and smoking status of the child. The primary focus is to understand the association between the smoking behaviour and the lung capacity, while possible confounding factors, such as gender and age, should be accounted for. To address these questions, a regression model can be used. Popular choices are regression models which model the mean outcome. However, the mean may not be the only useful summary measure or sometimes the mean may not have a relevant interpretation. To illustrate this, consider an example where Y denotes a person's mental impairment and $\boldsymbol{X}$ its life index and socio-economic status. Interest lies in studying the relationship between the socio-economic status and the mental impairment while controlling for the life index. The mental impairment is an ordinal outcome on a 4-level scale with categories $Y = 1$ (not impaired), $Y = 2$ (mild symptom formation), $Y = 3$ (moderate symptom formation), and $Y = 4$ (impaired). The mean mental impairment has no straightforward interpretation since the outcome is ordinal and not interval-scale. This implies that regression models which focus on the mean can be inappropriate.

In this dissertation, we propose a new regression framework for assessing the association between Y and $\boldsymbol{X}$ which can be used for ordinal, interval, and ratio-scale outcomes Y. More specifically, we model the probability that the outcome increases as a function of the predictors. We refer to this probability as the probabilistic index (PI). Formally, if $(Y, \boldsymbol{X})$ and $(Y', \boldsymbol{X}')$ denote independently and identically distributed random variables, then the PI is defined as

$$\mathrm{P}\left(Y \preccurlyeq Y' \mid \boldsymbol{X}, \boldsymbol{X}'\right) := \mathrm{P}\left(Y < Y' \mid \boldsymbol{X}, \boldsymbol{X}'\right) + \frac{1}{2}\mathrm{P}\left(Y = Y' \mid \boldsymbol{X}, \boldsymbol{X}'\right).$$

A probabilistic index model (PIM) then models the PI as a function of the predictors $\boldsymbol{X}$ and

$\boldsymbol{X}'$. More specifically, a PIM is defined as

$$\mathrm{P}\left(Y \preccurlyeq Y' \mid \boldsymbol{X}, \boldsymbol{X}'\right) = m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}), \quad (\boldsymbol{X}, \boldsymbol{X}') \in \mathcal{X}, \tag{A.10}$$

where $m(\cdot)$ is a function with range $[0, 1]$, $\mathcal{X}$ denotes the set of couples of predictors for which the model is defined, and $\boldsymbol{\beta}$ is a p-dimensional parameter vector.

In this dissertation, seven contributions to the PIM methodology are presented.

1. We have formally introduced PIMs together with a semiparametric parameter estimation and asymptotic distribution theory. The results of a simulation study showed that the asymptotic approximations are valid for finite samples. Several example datasets were analyzed with a PIM to illustrate its interpretation and flexibility.

2. We have situated the PIM within the statistical landscape by exploring the relationships with several well-known statistical methods.

   More specifically, to understand the functional form of $m(\cdot)$ in (A.10), we studied the relationship between a PIM and the normal linear regression model (NLRM) as well as the Cox proportional hazards model (CPHM). It turns out that there is a direct relationship between the model parameters of a PIM and the model parameters of a NLRM and a CPHM, respectively. These relationships suggest a functional form $m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}) = g^{-1}[(\boldsymbol{X}' - \boldsymbol{X})^T\boldsymbol{\beta}]$, with $g(\cdot)$ the well-known logit or probit link function. This choice results in an intuitive interpretation of the model parameters. For a univariate continuous predictor X, it follows that $g(\beta) = \mathrm{P}\left(Y \preccurlyeq Y' \mid X = x, X' = x + 1\right)$, i.e. the probability that the outcome for which $X = x + 1$ exceeds the outcome for which $X = x$.

   A PIM also provides a regression framework for the concordance index and extends the AUC-regression methodology. There are also interesting similarities and disparities between a PIM and rank regression and the cumulative logit model, respectively.

3. We have studied the relationship between a PIM and popular rank tests. The Wilcoxon–Mann–Whitney, Kruskal–Wallis, and Friedman rank test, among others, can be embedded within the PIM framework. This embedding allows to extend these rank tests to more complicated designs, while retaining an intuitive interpretation. Furthermore, it allows to construct confidence intervals for the associated effects sizes. Embedding all these rank tests in a single modelling framework can perhaps make these test more accessible to non-experienced users.

4. We have developed goodness-of-fit (GOF) methods to assess the adequacy of the proposed model $m(\cdot)$. A graphical diagnostic tool shows how the model can be improved and a test allows for formal hypothesis testing. Both methods are related to the interpretation of a PIM and are based on the probability $\mathrm{P}\left(Y \preccurlyeq Y' \mid \boldsymbol{X}, \boldsymbol{X}' = \boldsymbol{X} + \boldsymbol{\Delta}\right)$ for a fixed value $\boldsymbol{\Delta}$.

5. We have worked out a case study in detail. More specifically, the PIM methodology is used to analyze genomic differential expression studies based on reverse transcription quantitative polymerase chain reaction (RT-qPCR). The data generated by RT-qPCR techniques require normalization so as to account for technical variation which cannot be attributed to the treatments under study. Furthermore, since RT-qPCR experiments often aim at validating differentially expressed genes that were discovered by microarrays or next generation sequencing screens – and RT-qPCR biological validation experiments are often an (intermediate) endpoint of a study – quantifying and interpreting the effects is important for increasing the insight in the biological processes under study. The PIM turns out to be appropriate for both goals: it allows for normalizing the data in a straightforward fashion, while keeping an intuitive interpretation in terms of the odds for down- or upregulation.

6. We have studied efficient estimators for PIMs in a semiparametric setting. The index function associated with the efficient score corresponds to the solution of an integral equation. The results of a small simulation study indicated that the variance of the efficient estimator is similar to the variance of the estimator based on the independence working correlation matrix. However, these results need to be studied in more detail.

7. We have written an R-package with which most of the examples studied in this dissertation can be analyzed.

Although most of the basic tools to fit a PIM to data are developed and studied in this dissertation, many extension still need to be constructed so as to increase to applicability of PIMs.