Journal of the Royal Statistical Society
SERIES B Statistical Methodology

J. R. Statist. Soc. B (2012)
74, *Part* 4, *pp.* 623–671

Probabilistic index models

Olivier Thas

Ghent University, Belgium, and University of Wollongong, Australia

and Jan De Neve, Lieven Clement and Jean-Pierre Ottoy

Ghent University, Belgium

[*Read before* The Royal Statistical Society *at a meeting organized by the* Research Section *on Wednesday, February 15th, 2012*, Professor G. A. Young *in the Chair*]

Summary. We present a semiparametric statistical model for the probabilistic index which can be defined as $P(Y \leqslant Y^*)$, where Y and Y^* are independent random response variables associated with covariate patterns $\mathbf{X}$ and $\mathbf{X}^*$ respectively. A link function defines the relationship between the probabilistic index and a linear predictor. Asymptotic normality of the estimators and consistency of the covariance matrix estimator are established through semiparametric theory. The model is illustrated with several examples, and the estimation theory is validated in a simulation study.

Keywords: Area under the curve regression; Semiparametric inference; Stress–strength; Wilcoxon–Mann–Whitney test

1. Introduction

Consider the class of studies in which a single response variable is measured simultaneously with some covariates. Let Y and $\mathbf{X}$ denote the response variable and the d-dimensional covariate respectively, and let f_{YX} and $f_{Y|X}$ denote the density functions of the joint distribution and the conditional distribution of Y given $\mathbf{X}$ respectively. We use the same notation for the probability mass functions when Y or $\mathbf{X}$ are discrete variables. When Y is a continuous random variable, most statistical methods focus on the conditional mean of Y, given $\mathbf{X}$. For example, in linear regression models $E(Y|\mathbf{X}) = \mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta}$, where $\mathbf{Z}$ is a p-dimensional vector with elements that are functions of the covariates and where $\boldsymbol{\beta}$ is a p-dimensional parameter vector. Sometimes the complete conditional distribution of Y given $\mathbf{X}$ is specified (e.g. the normal regression model), allowing for likelihood-based inference, but this is often asymptotically replaced by some mild assumptions on the higher order moments of the conditional distribution so that the likelihood is no longer defined and semiparametric theories are required for inference.

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

$$P(Y < Y^*|\mathbf{X}, \mathbf{X}^*) + \tfrac{1}{2}P(Y = Y^*|\mathbf{X}, \mathbf{X}^*), \tag{1}$$

where $(Y, \mathbf{X})$ and $(Y^*, \mathbf{X}^*)$ are independently distributed with density f_{YX}. We introduce the notation $P(Y \preccurlyeq Y^*|\mathbf{X}, \mathbf{X}^*)$ for the PI as defined in expression (1). When Y is continuous $P(Y = Y^*|\mathbf{X}, \mathbf{X}^*) = 0$ and the PI simplifies to $P(Y \preccurlyeq Y^*|\mathbf{X}, \mathbf{X}^*) = P(Y < Y^*|\mathbf{X}, \mathbf{X}^*)$. Definition (1) is also

Address for correspondence: Olivier Thas, Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University, Coupure Links 653, B-9000 Gent, Belgium.
E-mail: Olivier.Thas@UGent.be

 1369–7412/12/74623

meaningful and convenient when the response is ordinal. Our definition implies that $P(Y \preccurlyeq Y^*|\mathbf{X}=\mathbf{x}, \mathbf{X}^*=\mathbf{x}) = \frac{1}{2}$ for both continuous and ordinal responses.

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

$$P(Y \preccurlyeq Y^*|\mathbf{X}, \mathbf{X}^*) = m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}), \tag{2}$$

in which m is a function with range $[0, 1]$ and $\boldsymbol{\beta}$ a p-dimensional parameter vector. In Section 2 more details will be given. Equation (2) thus implies a restriction on $f_{Y|X}$ that describes how the covariate $\mathbf{X}$ affects the response distribution in terms of the PI. Because $f_{Y|X}$ is not fully specified by assumption (2), model (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 Section 3.

An interesting special case arises when X is a binary design variable which refers to two populations. With $m(X, X^*; \beta) = 0.5 + \beta(X^* - X)$ model (2) becomes $P(Y \preccurlyeq Y^*|X=0, X^*=1) = P(Y_0 \preccurlyeq Y_1) = 0.5 + \beta$, which is the parameter of interest in the Wilcoxon–Mann–Whitney (WMW) test. In particular, under the general two-sample null hypothesis $H_0 : f_{Y|X=0} = f_{Y|X=1}$, the PI equals $P(Y_0 < Y_1) = 0.5$ when the response variable is continuous, and thus $\beta = 0$. Under mild conditions, the WMW test is consistent against the alternative $H_1 : P(Y_0 < Y_1) \neq 0.5$ or $\beta \neq 0$. The class of models that is 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. Our models also extend the work of Brumback *et al.* (2006), who proposed models for the PI, but with the restriction that Y and Y^* are continuous response variables that always belong to two different populations or treatment groups. In terms of our formulation this restriction could be expressed as $\mathbf{X}$ and $\mathbf{X}^*$ being distinct in at least one component which is a binary indicator for two treatment groups. Brumback and colleagues thus provided a WMW-type test for comparing two treatment groups, while controlling for one or more covariates. Our methods do not impose any particular restriction on the covariate vector $\mathbf{X}$. Moreover, the methods that are proposed in this paper further improve on Brumback *et al.* (2006) by being directly applicable to both continuous and ordinal response 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.

To demonstrate the scope and the interpretation of the models that form the topic of this paper, 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 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, we present here partial results from a simplified setting. The response 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). Fig. 1 shows a scatter plot of the data. We consider the PIM

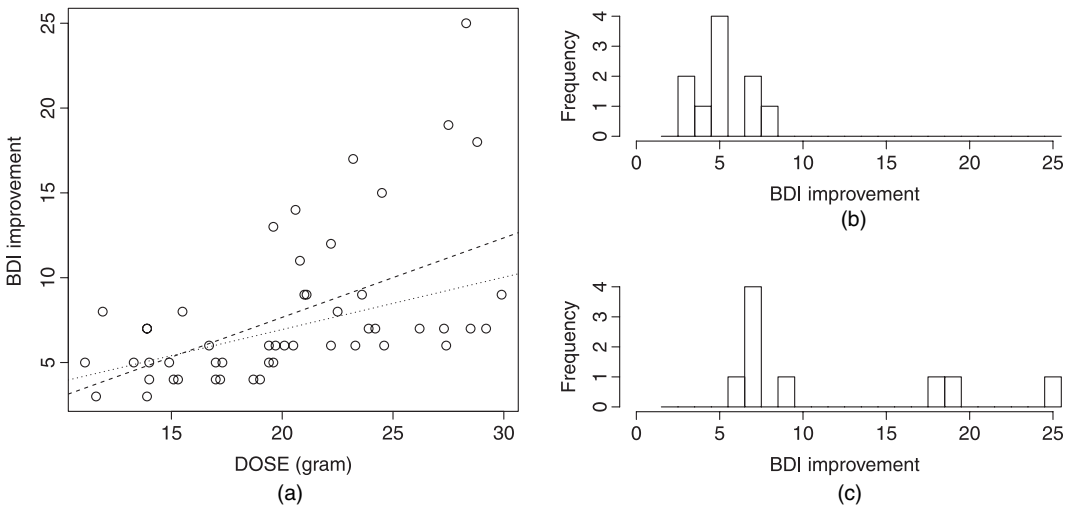

Fig. 1. (a) Scatter plot of BDI improvement *versus* dose (-------, linear regression model based on least squres; ·····, linear regression model based on Huber's robust *M*-estimator), and histograms of the improvements in BDI for (b) small doses and (c) large doses

$$P(\mathrm{BD} \preccurlyeq \mathrm{BD}^*) = \mathrm{expit}\{\beta(\mathrm{DOSE}^* - \mathrm{DOSE})\}.$$

Using the methods that are described in this paper, we find the estimate $\hat{\beta} = 0.1711$ with estimated standard deviation 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. Therefore we conclude that patients who are treated with a larger dose of quetiapine are more likely to benefit from the treatment. In particular, when the dose is increased by 5 g, the estimated PI equals $\mathrm{expit}(5\hat{\beta}) = 70.2\%$, i.e., when comparing a group of patients treated with quetiapine with a group that received an extra 5 g 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. At first sight the reader might think that the data could just as well have been analysed with a (linear) regression model, but, as illustrated in Fig. 1, the linearity assumption would be violated; a transformation or non-linear regression techniques may resolve this problem. However, Figs 1(b) and 1(c) further demonstrate that the dose affects not only the mean response, but also the variance and the skewness of the BDI distribution. The PI acts here as a quantity that summarizes the covariate effect on the response distribution in a meaningful effect size measure. 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. Regression techniques that focus on the conditional mean of the BDI are thus not to be recommended. The interpretation of the PI, in contrast, applies to both continuous and ordinal variables. Cumulative or adjacent categories logistic regression models (McCullagh, 1980) may also 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. Some other examples of response variables for which classical regression models are not the most appropriate are briefly discussed in the next paragraph.

There are many examples of response variables that are measured on an ordinal scale; we name just one more example. In pain management 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 only 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 result in probabilities that quantify how likely it is that the pain will decrease as a function of a set of covariates. Pain may also be measured on the visual analogue scale of Wallerstein (1984). For this the patient is presented with a horizontal line of 10 cm, 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, which is measured in millimetres, between the left-hand end of the line and the point marked by the patient is the numerical value that is used as a measure of pain. This is an example of a response variable that may be interpreted as being ordinal, but it may just as well be considered as a continuous response variable. However, not every variable that is measured on a continuous scale is necessarily an interval or ratio scale variable. For example, a patient with a visual analogue scale pain score of 4 does not necessarily have twice as much pain as someone with a pain score of 2. Thus, again the mean does not have a meaning, but statements involving order comparisons, such as $P(Y \preccurlyeq Y^*)$, do make sense. See Myles *et al.* (1999) for more details of the visual analogue scale.

PIMs may also turn out to be useful for analysing genuine continuous response 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 may be affected by several risk factors that, however, do not act only on the mean body mass index. In particular, the skewness of the body mass index distribution may change with covariate patterns. As illustrated in the BDI example, the PI summarizes the covariate effects on the shape of the response distribution, while remaining a very informative interpretation of the covariate effect sizes. Hence, PIMs could be a valuable alternative for body mass index data. Beyerlein *et al.* (2008) suggested analysing the body mass index data with quantile regression methods. Quantile regression (Koenker, 2005) is another important class of models. It focuses on the quantile distribution of Y given $\mathbf{X}$, $Q_{Y|X}(\cdot|\mathbf{X})$, say. Without the complete specification of the joint distribution of Y and $\mathbf{X}$, the τth quantile of the distribution of Y given $\mathbf{X}$ is modelled as $Q_{Y|X}(\tau|\mathbf{X}) = \mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta}_\tau$. These models are also semiparametric as the distribution of Y given $\mathbf{X}$ is not completely specified or parameterized.

The examples of the previous paragraphs already give a flavour of the usefulness of the PIM. In particular, the response variables were defined on an ordered scale, which could be discrete or continuous, for which the mean of the difference $Y - Y^*$ did not have a proper interpretation as an effect size, but for which the PI did. More generally, the PIM may be the statisticians' method of choice whenever the PI is considered as a meaningful parameter for quantifying effect sizes.

In Section 6 three example data sets are worked out in detail to demonstrate the scope of PIMs.

When no covariates are present, the PI has been discussed already by many researchers. To our knowledge, however, no unambiguous terminology is used throughout the literature. Some researchers even use the notation '$P(Y < Y^*)$' in the title of their papers; see for example Browne (2010), Enis and Geisser (1971), Tian (2008) and Zhou (2008). Others have called it the *individual exceedance probability*, *relative effect* or *stochastic improvement*. Thas (2009) gives an overview in his section 7.6. In engineering sciences it is known as *reliability* in the context of *stress–strength* problems; see Kotz *et al.* (2003) for an overview. Probabilities of the form (1) also appear in the analysis of receiver operating characteristic curves. We refer to Page (2003) for an excellent and relevant treatment. The PI may also be interpreted as the area under the curve of the population *PP*-plot, which is defined as the curve $\{(p, F_1\{F_2^{-1}(p)\}) : p \in [0, 1]\}$, where F_1 and F_2 are the distribution functions of $Y|X = x_1$ and $Y^*|X^* = x_2$ respectively. Suppose that Y

is a continuous response variable and that F_1 and F_2 have the same support, $\mathcal{S}$, say. Then the area under the curve becomes

$$\int_0^1 F_1\{F_2^{-1}(p)\}\,\mathrm{d}p = \int_{\mathcal{S}} F_1(y)\,\mathrm{d}F_2(y) = E_{Y^*|x_2}\{P_{Y|x_1}(Y \leqslant y | y = Y^*, x_1, x_2)\}$$
$$= P_{YY^*|x_1,x_2}(Y \leqslant Y^* | x_1, x_2) = P(Y \preccurlyeq Y^* | x_1, x_2), \tag{3}$$

with $Y|x_1$ and $Y^*|x_2$ independently distributed; usually we shall drop the index $YY^*|x_1, x_2$ from the probability operator. In the context of receiver operating characteristic curves, we refer to Dodd and Pepe (2003), who proposed regression models for the area under the curve which have formed the theoretical basis of the work of Brumback *et al.* (2006), which has been referred to earlier in this section. 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 $F_1(y) \geqslant F_2(y)$ for all $y \in \mathcal{S}$ and with strict inequality for at least a subset of $\mathcal{S}$. When F_1 is stochastically smaller than F_2, equation (3) immediately implies that $P(Y \preccurlyeq Y^*|x_1, x_2) > 0.5$. The implication does not hold necessarily in the other direction. Stochastic ordering is thus a stronger property than PI > 0.5, but the PI has the advantage of being a very informative effect size measure, as argued by many researchers; see Acion *et al.* (2006), Browne (2010), Laine and Davidoff (1996) and Zhou (2008), among others. This is further illustrated in the examples that are included in this paper.

After the class of PI models has been formally defined in Section 2 and the parameter estimation and asymptotic distribution theory are presented in Section 3, we discuss in Section 4 the relationship between the PIM and several other statistical methods such as linear regression, Cox proportional hazards regression, the WMW test, rank regression and the Hodges–Lehmann estimator. Note, however, that these connections to other statistical methods are given only to gain a better understanding of the PIMs and to motivate certain PI model formulations. We do not claim that PIMs should replace other statistical models, but they may be a valuable addition to the statisticians' toolbox, particularly when the research question allows a natural formulation with the PI as an effect size measure. The validity of the asymptotic theory is assessed in a simulation study in Section 5. More examples are presented in Section 6, and conclusions are formulated in Section 7.

2. The model and its interpretation

In its most general form the PIM is defined as

$$P(Y \preccurlyeq Y^* | \mathbf{X}, \mathbf{X}^*) = m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}), \tag{4}$$

where m is a function with range [0, 1], and $\boldsymbol{\beta}$ is a p-dimensional parameter vector. For the model to have a coherent interpretation, the function m must satisfy $m(\mathbf{X}, \mathbf{X}; \boldsymbol{\beta}) = m(\mathbf{X}^*, \mathbf{X}^*; \boldsymbol{\beta}) = 0.5$ and $m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}) = 1 - m(\mathbf{X}^*, \mathbf{X}; \boldsymbol{\beta})$, i.e. m must be antisymmetric about 1. The former restriction is guaranteed to hold because of the definition of the PI as in expression (1). When m does not satisfy the antisymmetry condition, the model may still be coherent when equation (4) is only defined for all $\mathbf{X} \prec \mathbf{X}^*$ or $\mathbf{X} \preceq \mathbf{X}^*$. The former refers to an order relation between the covariate patterns; so does the latter, but it includes $\mathbf{X} = \mathbf{X}^*$. Suppose that $\mathbf{X}^{\mathrm{T}} = (X_1, X_2)$ is a vector of dimension 2. Then an example of such an order relation is the *lexicographical ordering*, i.e. $\mathbf{X} \preceq_{\mathrm{lex}} \mathbf{X}^*$ if $X_1 < X_1^*$, or $X_1 = X_1^*$ and $X_2 \leqslant X_2^*$. By applying this definition recursively we can extend this order relation to vectors of dimension larger than 2. See Fishburn (1974) for more information about the lexicographical order. To avoid having always to make throughout the paper 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 $(\mathbf{X}, \mathbf{X}^*)$ for which model (4) is defined. We use the notation $\mathcal{X}_0$ when no order restriction is imposed, which is further referred to as the 'NO' order restriction. To summarize, the PIM is defined as

$$P(Y \preccurlyeq Y^* | \mathbf{X}, \mathbf{X}^*) = m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}) \qquad \text{for all } (\mathbf{X}, \mathbf{X}^*) \in \mathcal{X}. \tag{5}$$

This model expresses restrictions on the conditional distribution of Y given $\mathbf{X}$, but it does not fully specify this distribution. Hence, it is a semiparametric model. When $P(Y = Y^*) = 0$ model (5) may just as well be defined in terms of $P(Y < Y^* | \mathbf{X}, \mathbf{X}^*)$.

In this paper we restrict the function m to be related to a linear predictor, $\mathbf{Z}^T\boldsymbol{\beta}$, say, with $\mathbf{Z}$ a p-dimensional vector with elements that may depend on $\mathbf{X}$ and $\mathbf{X}^*$. In many examples $\mathbf{Z} = \mathbf{X}^* - \mathbf{X}$ will be a convenient and meaningful choice. We write

$$m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}) = g^{-1}(\mathbf{Z}^T\boldsymbol{\beta}), \tag{6}$$

with $g(\cdot)$ a proper link function that maps $[0, 1]$ onto the range of $\mathbf{Z}^T\boldsymbol{\beta}$, which is usually the real line. Since we basically model a probability, popular choices for g include the logit and the probit link functions. In some instances the identity link may be convenient.

Although $\mathbf{Z}^T\boldsymbol{\beta}$ may include an intercept or an offset, we sometimes choose to write the linear predictor as $\beta_0 + \mathbf{Z}^T\boldsymbol{\beta}$, where β_0 is an offset. If the scope of the PIM includes $\mathbf{X} = \mathbf{X}^*$ and the response is continuous, the offset β_0 must be set to a constant so that $P(Y \preccurlyeq Y^* | \mathbf{X} = \mathbf{x}, \mathbf{X}^* = \mathbf{x}) = 0.5$. The offset thus depends on the link function. For example, when $\mathbf{Z} = \mathbf{X}^* - \mathbf{X}$ the offsets for the logit, probit and identity link become $\beta_0 = 0$, $\beta_0 = 0$ and $\beta_0 = 0.5$ respectively.

3. Parameter estimation and statistical inference

3.1. Parameter estimation

Define $I(Y \preccurlyeq Y^*) = I(Y < Y^*) + \frac{1}{2} I(Y = Y^*)$ in which $I(Y < Y^*)$ and $I(Y = Y^*)$ denote the usual indicator functions evaluated for the events $Y < Y^*$ and $Y = Y^*$ respectively. The PIM (5) can then be written as

$$E\{I(Y \preccurlyeq Y^*) | \mathbf{X}, \mathbf{X}^*\} = P(Y \preccurlyeq Y^* | \mathbf{X}, \mathbf{X}^*) = m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}) = g^{-1}(\mathbf{Z}^T\boldsymbol{\beta}), \tag{7}$$

for $(\mathbf{X}, \mathbf{X}^*) \in \mathcal{X}$. When $(Y_1, \mathbf{X}_1), (Y_2, \mathbf{X}_2), \ldots, (Y_n, \mathbf{X}_n)$ denotes a sample of n independent identically distributed (IID) random variables with joint density function f_{YX}, model formulation (7) suggests that the $\boldsymbol{\beta}$ parameter vector can be estimated by using the set of *pseudo-observations* $I_{ij} = I(Y_i \preccurlyeq Y_j)$ for all $i, j = 1, \ldots, n$ for which $(\mathbf{X}_i, \mathbf{X}_j) \in \mathcal{X}$. In particular, model (7) resembles a conditional moment semiparametric model (see for example Chamberlain (1987), Newey (1988) or chapter 4 of Tsiatis (2006)), in which the conditional mean of the pseudo-observations is specified. We therefore propose to estimate the parameters by solving the estimating equations

$$\mathbf{U}_n(\boldsymbol{\beta}) = \sum_{(i,j) \in \mathcal{I}_n} \mathbf{A}(\mathbf{Z}_{ij}; \boldsymbol{\beta}) \{I_{ij} - g^{-1}(\mathbf{Z}_{ij}^T\boldsymbol{\beta})\} = \mathbf{0}, \tag{8}$$

where $\mathcal{I}_n$ is the set of indices (i, j) for which $(\mathbf{X}_i, \mathbf{X}_j) \in \mathcal{X}$, and $\mathbf{A}(\mathbf{Z}_{ij}; \boldsymbol{\beta})$ is a p-dimensional vector function of the regressors $\mathbf{Z}_{ij}$. Let $\hat{\boldsymbol{\beta}}$ denote the estimator. Although perhaps more efficient choices for $\mathbf{A}$ exist, we shall consider only

$$\mathbf{A}(\mathbf{Z}_{ij}; \boldsymbol{\beta}) = \frac{\partial g^{-1}(\mathbf{Z}_{ij}^T\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \mathbf{V}^{-1}\{g^{-1}(\mathbf{Z}_{ij}^T\boldsymbol{\beta})\}, \tag{9}$$

where $\mathbf{V}\{g^{-1}(\mathbf{Z}_{ij}^T\boldsymbol{\beta})\} = (1/\nu)\mathrm{var}(I_{ij} | \mathbf{Z}_{ij})$, with ν a scale parameter. This choice corresponds to

the quasi-likelihood estimating equations as used, for example, in the analysis of longitudinal data (Liang and Zeger, 1986; Zeger and Liang, 1986), where they are also referred to as *generalized estimating equations*. In the present setting, however, the conditional mean does not refer to the mean of the conditional distribution of the response, but it refers to the mean of the pseudo-observations. Moreover, 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, for the pseudo-observations I_{ij} have a more complicated dependence structure than, for example, block independence as in clustered or longitudinal data. Lemmas 1 and 2 of Section 3.2 state that the pseudo-observations have the *sparse correlation* structure of Lumley and Hamblett (2003). This result makes the semiparametric theory of Lumley and Hamblett (2003) directly applicable to our setting. Theorems 1 and 2 that we present in Section 3.3 summarize the most important distribution theory results for the PIM.

When $m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}) = 1 - m(\mathbf{X}^*, \mathbf{X}; \boldsymbol{\beta})$ the solution of equations (8) for the NO order restriction is identical to the solution for a lexicographical order restriction. Therefore, when m satisfies the antisymmetry condition, the lexicographical ordering is preferred over the NO order restriction for only half of the pseudo-observations are needed. This also demonstrates that the estimator is independent of the order in which the covariates appear in the definition of the lexicographical ordering.

3.2. Sparse correlation

In this section we shall show that the pseudo-observations are sparsely correlated, but we start with the defining *sparse correlation* in the context of pseudo-observations. A more general definition can be found in Lumley and Hamblett (2003).

Definition 1. Let I_{ij} $((i, j) \in \mathcal{I}_n)$ denote a set of pseudo-observations. For each pseudo-observation I_{ij} a set of pairs of indices S_{ij} $((i, j) \in \mathcal{I}_n)$ is defined such that $(k, l) \notin S_{ij}$ and $(i, j) \notin S_{kl}$ implies that 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 $S_{ij} \cap S_{kl} = \emptyset$ 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} $((i, j) \in \mathcal{I}_n)$ 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 no order restriction or the lexicographical order restriction is imposed.

Lemma 1 (sparse correlation: NO order restriction). The NO ordered pseudo-observations have the sparse correlation structure.

Proof. Each pseudo-observation $I_{ij} \in \mathcal{I}_n = \{(i, j) : i \neq j\}$ 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 and 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 $O(|\mathcal{I}_n|) = O(n^2)$, lemma 1 holds for n even. Similarly, when n is odd, $M_n m_n = (4n - 6) \times \lfloor n/2 \rfloor = O(n^2) = O(|\mathcal{I}_n|)$.

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

Proof. The lexicographical pseudo-observations I_{ij} for which $\mathbf{X}_i \preceq_{\text{lex}} \mathbf{X}_j$ can be obtained by sorting the data $(Y, \mathbf{X})$ on the basis of lexicographical ordering on $\mathbf{X}$ and then considering the pseudo-observations $I_{ij} \in \mathcal{I}_n = \{(i,j) : i < j \text{ and } i, j = 1, \ldots, n\}$. Each pseudo-observation I_{ij} is correlated with $2n-4$ other pseudo-observations. Indeed I_{ij} is correlated with

(a) I_{ik} where $k = i+1, \ldots, n$ and $k \neq j$,
(b) I_{kj} where $k = 1, \ldots, j-1$ and $k \neq i$,
(c) I_{ki} where $k = 1, \ldots, i-1$,
(d) 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 $O(|\mathcal{I}_n|) = O(n^2)$, lemma 2 holds for n even. Similarly, when n is odd, $M_n m_n = (2n-3) \times \lfloor n/2 \rfloor = O(n^2) = O(|\mathcal{I}_n|)$.

3.3. Asymptotic normality of the parameter estimators

Since the following two theorems are special cases of theorem 7 of Lumley and Hamblett (2003) we shall omit the proof. We only need to define the *true* $\boldsymbol{\beta}$*-parameter*, $\boldsymbol{\beta}_0$, say, in the semiparametric PIM. Instead of defining $\boldsymbol{\beta}_0$ through the independence working log-likelihood function as in Lumley and Hamblett (2003), we define $\boldsymbol{\beta}_0$ as the $\boldsymbol{\beta}$ for which

$$\lim_{n\to\infty} \left(E\left[\sum_{(i,j)\in\mathcal{I}_n} \mathbf{A}(\mathbf{Z}_{ij}; \boldsymbol{\beta}) \{ I_{ij} - g^{-1}(\mathbf{Z}_{ij}^{\mathrm{T}} \boldsymbol{\beta}) \} \right] \right) = \mathbf{0}. \tag{10}$$

The regularity conditions in the statement of theorem 1 imply the existence of $\boldsymbol{\beta}_0$.

Theorem 1 (asymptotic normality). Consider the PIM (7) with predictors $\mathbf{Z}_{ij}$ taking values in a bounded subset of $\mathbb{R}^p$. We make the following assumptions.

Assumption 1. The pseudo-observations are sparsely correlated, with m_n as in lemma 1 or lemma 2.

Assumption 2. The link function g and the variance function $\mathbf{V}$ have three continuous derivatives.

Assumption 3. The true parameter $\boldsymbol{\beta}_0$, as defined by equation (10), is in the interior of a convex parameter space.

Assumption 4. There are a vector $\mathbf{W}$ and positive definite matrix $\mathbf{T}$ such that

$$|\mathcal{I}_n|^{-1} \sum_{(i,j)\in\mathcal{I}_n} \mathbf{Z}_{ij} \to \mathbf{W} \quad \text{and} \quad |\mathcal{I}_n|^{-1} \sum_{(i,j)\in\mathcal{I}_n} \mathbf{Z}_{ij} \mathbf{Z}_{ij}^{\mathrm{T}} \to \mathbf{T}.$$

Assumption 5. $\limsup\{m_n^{-1} \operatorname{var}(\Sigma_{(i,j)\in\mathcal{I}_n} I_{ij})\} > 0$.

Then, as $n \to \infty$, $(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)\sqrt{m_n}$ 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

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

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

4. Relationship with other methods

In this section we show how the PIMs are related to other statistical methods. In Sections 4.1 and 4.2 we demonstrate that the parameters of linear regression models and Cox proportional hazard models have simple relationships with the parameters of a PIM with particularly chosen link functions and linear predictors. The connection between hypothesis tests in the semiparametric PIM framework and the WMW rank test is explored in Section 4.3, and the link between the PIM parameter estimators and rank regression is the topic of Section 4.4. We do not suggest that the PIM methodology is a direct competitor of these other methods, but by understanding these relationships the reader may gain a better appreciation of the PIMs' position in the landscape of statistical models, and he or she may find arguments for choosing one or other link function.

4.1. Linear regression models

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

$$Y = \mu + \alpha X + \varepsilon,$$

where ε is a zero-mean error term with continuous distribution function F_ε which does not depend on the covariate X. The model can be equivalently formulated as

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

Since Y is continuous, $P(Y \preccurlyeq Y^*) = P(Y < Y^*)$. Consider now the PI for this class of regression models,

$$\begin{aligned} P(Y < Y^* | X, X^*) &= P(\mu + \alpha X + \varepsilon < \mu + \alpha X^* + \varepsilon^* | X, X^*) \\ &= P\{\varepsilon - \varepsilon^* < \alpha(X^* - X)\} = F_\Delta\{\alpha(X^* - X)\}, \end{aligned}$$

where F_Δ is the distribution function of $\varepsilon - \varepsilon^*$. Thus, for a PIM with link function g we find the relationship

$$g\{P(Y < Y^* | X, X^*)\} = g[F_\Delta\{\alpha(X^* - X)\}] = \beta Z. \tag{11}$$

This relationship for linear regression models immediately suggests the link function $g(\cdot) = F_\Delta^{-1}(\cdot)$, for which a PIM with linear predictor $Z = X^* - X$ and $\beta = \alpha$ is obtained.

A simple and important example is the normal linear regression model for which the error term ε is normally distributed with mean 0 and constant variance σ^2. The distribution F_Δ is also normal with mean 0 and variance $2\sigma^2$. With Φ the distribution function of a standard normal distribution, equation (11) becomes

$$g\{P(Y < Y^* | X, X^*)\} = g\left[\Phi\left\{\frac{\alpha(X^* - X)}{\sqrt{2}\sigma}\right\}\right] = \beta Z.$$

With the probit link function ($g(\cdot) = \Phi^{-1}(\cdot)$) and with $Z = X^* - X$, a simple relationship between α and β is established: $\beta = \alpha/\sqrt{2}\sigma$, which expresses that β is proportional to α. Under the

normality, linearity and homoscedasticity assumptions of the regression model we therefore conclude that β has also an interpretation in terms of the effect of X on the conditional mean of the response. When the regression model assumptions do not hold, the parameter β in the PIM still has the interpretation in terms of the PI.

Without repeating the calculations we also give the relationship between α and β when the residual variance σ^2 is not constant in the normal linear regression model. Without loss of generality we suppose that $X > 0$. We discuss only $\sigma^2(X) = \gamma X$ as the variance function. The relationship between the regression parameters then becomes

$$\beta = \frac{\alpha}{\sqrt{\{\sigma^2(X) + \sigma^2(X^*)\}}} = \frac{\alpha}{\sqrt{\{\gamma(X^* + X)\}}},$$

which suggests that the PIM should better be formulated as

$$\Phi^{-1}\{P(Y < Y^*|X, X^*)\} = \frac{X^* - X}{\sqrt{(X^* + X)}}\beta \qquad \text{for } (X, X^*) \in \mathcal{X}_0 \tag{12}$$

so we again find a simple relationship between the parameters, $\beta = \alpha/\sqrt{\gamma}$. Model (12) gives a slightly different interpretation of β in terms of the PI. For $X^* = X + 1$, we find

$$P(Y < Y^*|X, X^* = X + 1) = \Phi\left\{\frac{\beta}{\sqrt{(2X + 1)}}\right\}.$$

This expression illustrates that the effect of X on the distribution of Y diminishes as X increases, at least in terms of the PI. In the normal regression model, the increasing residual variance does not affect the covariate effect on the mean response, whereas it results in a negative effect modulation in terms of the PI. This is further illustrated with a real data example in Section 6.3. This was also noted by Brumback *et al.* (2006) and it suggests that we should take care in interpreting the α-parameter in a normal regression model with non-constant variance because the importance of the covariate effect may actually depend on the covariate value.

4.2. Cox proportional hazard model

Cox proportional hazard 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 references 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 $P(T_{1i} < T_{2i}|X_{1i}, X_{2i})$, where T_{1i} and T_{2i} are paired survival times (e.g. from twin studies) with covariates (X_{1i}, X_{2i}). Under the assumption of proportional hazards in the absence of censored or tied data, they found that

$$\text{logit}\{P(T_{1i} < T_{2i}|X_{1i}, X_{2i})\} = \beta(X_{1i} - X_{2i}),$$

which resembles a PIM with $Z = X_1 - X_2$ in which the parameter β originates from the hazard function $\lambda(t|X) = \lambda_0(t)\exp(\beta X)$. Note, however, that in the PIMs that are presented in this paper it is assumed that all observations are mutually independent, whereas Holt and Prentice (1974) developed their method for paired response variables (paired survival times).

Also the marginal likelihood formulation of Kalbfleisch 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 hazard model.

We shall show that conditional distributions that belong to the class of proportional hazard models imply a PIM with logit link. Let $S(y|\mathbf{X}) = 1 - F(y|\mathbf{X})$ denote the survival function. The

hazard function is defined as $\lambda(y|\mathbf{X}) = -\partial \log\{S(y|\mathbf{X})\}/\partial y = f(y|\mathbf{X})/S(y|\mathbf{X})$. In a proportional hazard model the hazard function allows a factorization of the form $\lambda(y|\mathbf{X}) = \lambda_0(y)\exp(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta})$, in which $\lambda_0(y)$ is the baseline hazard function that does not depend on the covariate $\mathbf{X}$. Thus, within the class of proportional hazard models the survival function is of the form

$$S(y|\mathbf{X}) = c(\mathbf{X})\, S_0(y)^{\exp(\mathbf{X}^{\mathrm{T}}\beta)}, \tag{13}$$

where $S_0(y) = S(y|\mathbf{X}=\mathbf{0})$ is the baseline survival function and $c(\mathbf{X})$ is a normalization constant to make $S(y|\mathbf{X})$ a proper distribution function. Suppose that $\mathcal{S}$ is the support of Y. Straightforward algebra then gives

$$\begin{aligned} P(Y<Y^*|\mathbf{X},\mathbf{X}^*) &= \int_{\mathcal{S}} F(y|\mathbf{X})\,\mathrm{d}F(y|\mathbf{X}^*) = -\int_{\mathcal{S}} \{1-S(y|\mathbf{X})\}\,\mathrm{d}S(y|\mathbf{X}^*) \\ &= 1-\exp\{\boldsymbol{\beta}(\mathbf{X}^*-\mathbf{X})\}\; P(Y<Y^*|\mathbf{X},\mathbf{X}^*), \end{aligned}$$

from which we find the PIM

$$\mathrm{logit}\{P(Y<Y^*|\mathbf{X},\mathbf{X}^*)\} = \boldsymbol{\beta}(\mathbf{X}^*-\mathbf{X}).$$

This illustrates that the PIM with a logit link and with $\mathbf{Z}=\mathbf{X}^*-\mathbf{X}$ arises naturally from a widely applicable class of distributions. A straightforward example is the exponential distribution with rate parameter γ which has survival function $S(y)=\exp(-\gamma y)$. Equation (13) is satisfied with $S_0(y)=\exp(y)$ and $\gamma(\mathbf{X})=\exp(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta})$.

Equation (13) characterizes this class of distributions through its survival function, but its form immediately suggests that, for distributions for which $F(y|\mathbf{X}) = c(\mathbf{X})\, F_0(y)^{\exp(\mathbf{X}^{\mathrm{T}}\beta)}$ holds, a PIM also results.

4.3. Two-sample problem

Consider a sample of n IID random variables (Y_i, X_i), $i=1,\ldots,n$, with Y_i continuous. Without loss of generality assume that the sample of Y observations does not contain ties and that the observations are ordered so that the first n_1 observations belong to the first group and the last $n_2=n-n_1$ to the second. Let $X_i=0$ if $1\leqslant i\leqslant n_1$ and $X_i=1$ if $n_1+1\leqslant i\leqslant n$. Consider the PIM with identity link,

$$P(Y<Y^*|X,X^*) = \tfrac{1}{2} + (X^*-X)\beta \qquad \text{for } (X,X^*)\in\mathcal{X}=\{(X,X^*): X<X^*\}. \tag{14}$$

The offset $\beta_0=\frac{1}{2}$ is not strictly necessary, because the scope of the model does not include $X=X^*$. However, by having it in the model, the traditional two-sample null hypothesis becomes equivalent to $\beta=0$, which is the default null hypothesis in most statistical software. The order relation restriction in $\mathcal{X}$ implies that only $X=0$ and $X^*=1$ are allowed so that the model can be reformulated in a more convenient form. We use the notation $Y^{(1)}$ and $Y^{(2)}$ to denote two independent observations from the first ($X^{(1)}=0$) and the second ($X^{(2)}=1$) group respectively. The model is now reformulated as

$$P(Y^{(1)}<Y^{(2)}|X^{(1)}=0, X^{(2)}=1) = \tfrac{1}{2} + (X^{(2)}-X^{(1)})\beta = \tfrac{1}{2}+\beta. \tag{15}$$

The estimating equation (8) becomes

$$\sum_{i=1}^{n_1}\sum_{j=n_1+1}^{n} \frac{I(Y_i<Y_j)-\frac{1}{2}-\beta}{(\frac{1}{2}+\beta)(\frac{1}{2}-\beta)} = 0.$$

Therefore $\hat{\beta} = (n_1 n_2)^{-1}\Sigma_{i=1}^{n_1}\Sigma_{j=n_1+1}^{n} I(Y_i<Y_j) - \frac{1}{2}$. With $\mathrm{MW} = \Sigma_{i=1}^{n_1}\Sigma_{j=n_1+1}^{n} I(Y_i<Y_j)$ denoting the Mann–Whitney test statistic, we see immediately that $\hat{\beta} = \mathrm{MW}/n_1 n_2 - \frac{1}{2}$. The

traditional Mann–Whitney test, however, is usually based on the standardized test statistic $T_{\mathrm{MW}} = (\mathrm{MW} - n_1 n_2/2)/\sigma_0$, where σ_0 is the standard deviation of MW under the two-sample null hypothesis $H_0 : F_1 = F_2$, with F_1 and F_2 the distribution functions of $Y^{(1)}$ and $Y^{(2)}$ respectively. Under this restrictive null hypothesis $\sigma_0^2 = n_1 n_2 (n+1)/12$. Using a variance which is obtained under the null hypothesis is related to score tests, whereas using a variance estimator that is more generally consistent is related to the Wald test. The advantage of using a more generally consistent variance estimator is that the test may then also be used for testing the null hypothesis $H_0 : P(Y^{(1)} < Y^{(2)}) = \frac{1}{2}$ *versus* $H_0 : P(Y^{(1)} < Y^{(2)}) \neq \frac{1}{2}$ (i.e. $H_0 : \beta = 0$ *versus* $H_1 : \beta \neq 0$). Such a variance estimator was proposed by Fligner and Policello (1981) and, using the equality $\hat{\beta} = \mathrm{MW}/n_1 n_2 - \frac{1}{2}$, their results give immediately a variance estimator for $\hat{\beta}$ which can be written

$$\hat{\sigma}^2_{\hat{\beta}} = (n_1 n_2)^{-1} \{(n_1 - 1)\hat{\phi}_1^2 + (n_2 - 1)\hat{\phi}_2^2 + (\tfrac{1}{2} + \hat{\beta})(\tfrac{1}{2} - \hat{\beta})\}, \tag{16}$$

with

$$\hat{\phi}_1^2 = \frac{1}{n_1 n_2 (n_1 - 1)} \sum_{j=n_1+1}^{n} \sum_{i=1}^{n_1} \sum_{\substack{i'=1 \\ i' \neq i}}^{n_1} I(Y_i < Y_j) I(Y_{i'} < Y_j) - \left(\frac{1}{2} + \hat{\beta}\right)^2$$

and

$$\hat{\phi}_2^2 = \frac{1}{n_1 n_2 (n_2 - 1)} \sum_{i=1}^{n_1} \sum_{j=n_1+1}^{n} \sum_{\substack{j'=n_1+1 \\ j' \neq j}}^{n} I(Y_i < Y_j) I(Y_i < Y_{j'}) - \left(\frac{1}{2} + \hat{\beta}\right)^2.$$

It can be easily shown that the sandwich variance estimator of lemma 2 gives exactly the same expression. The PIM and the inference based on the estimating equations thus include the Wald-type WMW test of Fligner and Policello (1981). We refer to chapter 9 of Thas (2009) for more information about the use of the WMW test in a semiparametric setting.

We started this section by assuming that the response Y is continuous, resulting in a simplification of $P(Y \preccurlyeq Y^*)$ and $I(Y_i \preccurlyeq Y_j)$. However, when the continuity assumption on Y is dropped and ties are allowed, the relationship with the WMW test statistics still holds, but with midranks instead of ranks.

For the K-sample problem, the PIM can be similarly parameterized so that each parameter, $\hat{\beta}_{kl}$, say, corresponds to $\mathrm{MW}_{kl}/n_k n_l - \frac{1}{2}$, with MW_{kl} the Mann–Whitney test statistic for comparing groups k and l and n_k (or n_l) the sample size of group k (or l), $k < l$ and $k, l = 1, \ldots, K$. The equivalence between a PIM with this parameterization and the Kruskal–Wallis test is based on an equivalent representation of the Kruskal–Wallis statistic in terms of Mann–Whitney statistics; see Fligner (1985) for more details.

4.4. Rank regression and Hodges–Lehmann estimators

For the class of linear models of Section 4.1 the parameters can be estimated by means of several methods. With no full parametric assumption on the error distribution, least squares is probably the most popular method. However, least squares suffers from the drawback that it is very 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 excellent reviews. Although rank regression parameter estimation can be defined in a general way, we shall formulate it here only with the Wilcoxon scores. The parameters of the linear regression model are estimated by minimizing

$$\sum_{i=1}^{n} \left[\frac{R\{Y_i - (\mu + \alpha X_i)\}}{n+1} - \frac{1}{2} \right] \{Y_i - (\mu + \alpha X_i)\}, \tag{17}$$

where $R\{Y_i-(\mu+\alpha X_i)\}$ denotes the rank of the residual $Y_i-(\mu+\alpha X_i)$ among the n residuals. The estimate of α is thus obtained by solving the estimating equation (based on the partial derivative of expression (17))

$$\sum_{i=1}^{n} X_i\left[\frac{R\{Y_i-(\mu+\alpha X_i)\}}{n+1}-\frac{1}{2}\right]=0. \tag{18}$$

In what follows we shall replace the denominator $n+1$ with n (asymptotically equivalent). The relationship with the estimating equation (8) of the PIM parameters becomes more transparent when the rank in equation (18) is replaced by an expression involving the indicator function. We assume that there are no ties in the residuals. Equation (18) may then be written as

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

This can be simplified to

$$\sum_{i=1}^{n}\sum_{j=1}^{n} X_i\left[I\{Y_j\leqslant Y_i-\alpha(X_i-X_j)\}-\frac{1}{2}\right]=0. \tag{19}$$

Consider now the estimating equation (8) of a PIM with identity link, $Z_{ij}=X_i-X_j$ and with a very simple index function $A(Z_{ij},\beta)=Z_{ij}$,

$$\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i-X_j)\{I(Y_i\preccurlyeq Y_j)-\beta(X_i-X_j)-\tfrac{1}{2}\}=0. \tag{20}$$

Assuming that there are no ties, straightforward algebra shows that the left-hand side of equation (20) equals

$$2\sum_{i=1}^{n}\sum_{j=1}^{n} X_i\{I(Y_i\leqslant Y_j)-\beta(X_i-X_j)-\tfrac{1}{2}\}. \tag{21}$$

By comparing the two estimating equations (19) and (20) with the left-hand side of the latter replaced by expression (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. Thus, in rank regression the parameter α is estimated as $\hat{\alpha}$ so that, after subtracting $\hat{\alpha}X_i$ from the responses Y_i, the estimated PI equals $\frac{1}{2}$. The estimator of β in the PIM makes on average, for each X_i-X_j, the estimated PI deviate from $\frac{1}{2}$ by $\hat{\beta}(X_i-X_j)$. 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 asymptotically in the estimating equation (19) 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 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. This setting corresponds to the two-sample problem of Section 4.3. The estimating equation (18) becomes

$$\sum_{i=1}^{n_1}\left[\frac{R\{Y_i-(\mu+\alpha)\}}{n+1}-\frac{1}{2}\right]=0,$$

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

$$\sum_{i=1}^{n_1} \sum_{j=n_1+1}^{n} \{I(Y_i \leqslant Y_j) - \beta - \tfrac{1}{2}\} = 0,$$

which is $\hat{\beta} = (1/n_1 n_2) \sum_{i=1}^{n_1} \sum_{j=n_1+1}^{n} I(Y_i \leqslant Y_j) - \frac{1}{2}$, as in Section 4.3.

5. 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 data-generating models. Moreover, in the class of data-generating distributions of the PIMs there may be a complicated relationship between the parameters of both models. Here, we have chosen to generate data with a normal linear regression model, an exponential generalized linear model and multinomial regression model. For the first two models the relationship with the PIM is provided in Section 4, and for the last more details will be given later. 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 give only 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. Moreover, we 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.

All computations have been performed with the R software (R Development Core Team, 2010) and all PIMs are defined for the lexicographical order relation because they all satisfy the antisymmetry condition; see Section 3 for more information.

5.1. Checking asymptotic properties of the estimators

The theoretical properties of the estimators of Section 3 are evaluated in a simulation study. Since a PIM does not represent a unique data-generating model we simulate data from two models for which we have established a relationship with the PIMs: a normal linear model and an exponential model.

5.1.1. Normal linear model

We consider the model

$$Y_i = \alpha X_i + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{22}$$

where $\varepsilon_i | X_i$ are IID $N\{0, \sigma_\varepsilon^2(X_i)\}$. 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 $u = 10$. The parameter α equals 1 or 10. Table 1 presents the results for a constant standard deviation, i.e. $\sigma_\varepsilon(X) = \sigma$, with $\sigma = 1$ or $\sigma = 5$. The corresponding PIM is given by

$$\Phi^{-1}\{P(Y \preccurlyeq Y^* | X, X^*)\} = \beta(X^* - X),$$

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

Table 1. Simulation results for the normal linear homoscedastic model, based on 1000 Monte Carlo runs†

α	u	σ	β	$Av(\hat{\beta})$	$var(\hat{\beta})$	$Av(\hat{S}_{\hat{\beta}})$	EC	$Av(\bar{\beta})$	$var(\bar{\beta})$	$Av(\tilde{\beta})$	$var(\tilde{\beta})$
$n=25$											
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

†β is the true parameter, Av($\hat{\beta}$) the average of the β-estimates according to the semiparametric PIM theory, var($\hat{\beta}$) the sample variance of the simulated $\hat{\beta}$, 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, Av($\bar{\beta}$) the average of the least squares estimates, var($\bar{\beta}$) the sample variance of the simulated $\bar{\beta}$, Av($\tilde{\beta}$) the average of the maximum likelihood estimates and var($\tilde{\beta}$) the sample variance of the simulated $\tilde{\beta}$.

estimator of Section 3 is denoted by $\hat{\beta}$, and it is further referred to as the PIM estimator. Table 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. Table 1 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 model (22) by means of least squares and maximum likelihood. In the normal linear regression model least squares and maximum likelihood 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 difference estimators of β, particularly in small samples.

From Table 1 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 close to their nominal level for sample sizes of 50 and more. The simulation study also reveals that the sample distribution of $\hat{\beta}$ is close to normal (the results are not shown). As expected the parametric estimators are more efficient, but, when α or the range of X increases, the difference in efficiency decreases.

Table 2 shows the results of simulations of heteroscedastic data with $\sigma_\varepsilon(X) = \sigma\sqrt{X}$, where $\sigma = 1$ or $\sigma = 5$. The corresponding PIM is given by

Table 2. Simulation results for the normal linear heteroscedastic model, based on 1000 Monte Carlo runs†

α	u	σ	β	$Av(\hat{\beta})$	$var(\hat{\beta})$	$Av(\hat{S}_{\hat{\beta}})$	EC	$Av(\bar{\beta})$	$var(\bar{\beta})$	$Av(\tilde{\beta})$	$var(\tilde{\beta})$
$n=25$											
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

†β is the true parameter, Av($\hat{\beta}$) the average of the β-estimates according to the semiparametric PIM theory, var($\hat{\beta}$) the sample variance of the simulated $\hat{\beta}$, 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, Av($\bar{\beta}$) the average of the least squares estimates, var($\bar{\beta}$) the sample variance of the simulated $\bar{\beta}$, Av($\tilde{\beta}$) the average of the maximum likelihood estimates and var($\tilde{\beta}$) the sample variance of the simulated $\tilde{\beta}$.

$$\Phi^{-1}\{P(Y \preccurlyeq Y^*|X, X^*)\} = \beta \frac{X^* - X}{\sqrt{(X^* + X)}},$$

where $\beta = \alpha/\sigma$. Similar conclusions hold to those for the homoscedastic case. Surprisingly the semiparametric PIM estimator is more efficient than least squares and maximum likelihood when $\alpha = 10$, $u = 1$ and $\sigma = 1$. This observation does not contradict the theory, which only assures the efficiency of the maximum likelihood estimator in an asymptotic sense.

5.1.2. Exponential model

Let $Y_i|X_i$ be IID Exponential$\{\gamma(X_i)\}$ with

$$\gamma(X_i) = \exp(\alpha X_i), \qquad i = 1, \ldots, n. \tag{23}$$

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 $u = 10$ and α takes the value 0.1 or -2. The corresponding PIM is

$$\text{logit}\{P(Y \preccurlyeq Y^*|X, X^*)\} = \beta(X - X^*), \tag{24}$$

where $\beta = \alpha$. Table 3 gives the results when model (24) is analysed with the semiparametric PIM theory, resulting in $\hat{\beta}$. As a result of the identity $\beta = \alpha$, the parameter β can also be estimated

Table 3. Simulation results for the exponential model, based on 1000 Monte Carlo runs†

α	u	σ	β	$Av(\hat{\beta})$	$var(\hat{\beta})$	$Av(\hat{S}_{\hat{\beta}})$	EC	$Av(\bar{\beta})$	$var(\bar{\beta})$	$Av(\tilde{\beta})$	$var(\tilde{\beta})$
$n=25$											
−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

†β is the true parameter, Av($\hat{\beta}$) the average of the β-estimates by using the semiparametric PIM theory, var($\hat{\beta}$) the sample variance of the simulated $\hat{\beta}$, Av($\hat{S}_{\hat{\beta}}$) the average of the sandwich variance estimates by using the semiparametric PIM theory, EC the empirical coverage of a 95% confidence interval for β, Av($\bar{\beta}$) the average of the semiparametric proportional hazards estimates, var($\bar{\beta}$) the sample variance of the simulated β, Av($\tilde{\beta}$) the average of the maximum likelihood estimates and var($\tilde{\beta}$) the sample variance of the simulated $\tilde{\beta}$.

on the basis of 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 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. For large ranges of X the efficiency of the PIM estimator is close to the efficiency of the semiparametric proportional hazards estimator.

5.2. Power

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

$$g\{P(Y \preccurlyeq Y^* | X_1, X_1^*, X_2, X_2^*)\} = \beta_1(X_1^* - X_1) + \beta_2(X_2^* - X_2), \tag{25}$$

where X_1 and X_1^* are 0–1 dummy variables 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 response is better than the response of an untreated patient is 50%. To our knowledge hardly any statistical tests have been described in the literature for this problem. In Section 1 we have discussed the most important competitors. In this simulation study we have opted for the test of Brumback *et al.* (2006). 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,

$$g\{P(Y \preccurlyeq Y^* | X_1 < X_1^*, X_2, X_2^*)\} = \delta_1 + \delta_2(X_2^* - X_2), \tag{26}$$

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.

Three simulation scenarios are described next. For each scenario a parametric or a semiparametric test is included as a competitor test.

(a) $Y_i|X_i$ are IID $N(\alpha_1 X_{1i} + \alpha_2 X_{2i}, 1)$. The data are analysed by least squares in a marginal linear model with conditional mean $E(Y|X_1, X_2) = \gamma_1 X_1 + \gamma_2 X_2$, by the PIM (25) with probit link function and by Brumback's bootstrap test based on equation (26) with probit link. The least squares 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.

(b) $Y_i|X_i$ are IID Exponential$\{\exp(\alpha_1 X_{1i} + \alpha_2 X_{2i})\}$. The data are analysed by partial likelihood in a proportional hazards model with hazard function $\lambda(X) = \exp(\gamma_1 X_1 + \gamma_2 X_2)$, by the PIM (25) with logit link and by Brumback's bootstrap test based on equation (26) with logit link. The powers with the partial likelihood method may be considered as corresponding to a semiparametric competitor of the PIM, although the proportional hazard model does not coincide with the class of PIMs: they express different restrictions on the distribution $f_{Y|X}$.

(c) $Z_i|X_i$ are IID Logistic$(\alpha_1 X_{1i} + \alpha_2 X_{2i}, 1)$, for which the latent response variable Z_i is discretized into four ordered categories as described in section 6.2 of Agresti (2007). The resulting ordinal response is denoted by Y_i. The data are analysed by maximum likelihood in the proportional odds model $\text{logit}\{P(Y \leqslant j|X_1, X_2)\} = \mu_j + \gamma_1 X_1 + \gamma_2 X_2$, by the PIM (25) with logit link and by Brumback's bootstrap test based on equation (26) with logit link. The R package `MASS` (Venables and Ripley, 2002) is used to fit the proportional odds model.

The following design is considered. The covariate X_1 is a 0–1 balanced dummy variable, X_2 is equally spaced over $[0.1, 10]$ and α_1 takes the values 0, 0.5 and 1 whereas 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 4 shows the empirical powers based on 1000

Table 4. Empirical powers based on 1000 Monte Carlo runs for the three data-generating models†

α_1	*Powers for the following data-generating models:*								
	1			*2*			*3*		
	PIM	*LS*	*BT*	*PIM*	*PL*	*BT*	*PIM*	*ML*	*BT*
n = 20									
0.0	7.6	9.5	0.0	9.7	4.3	0.0	10.8	4.5	2.2
0.5	15.0	27.3	0.0	22.7	16.2	0.0	14.1	7.3	2.8
1.0	45.9	72.3	0.2	42.3	44.4	0.0	25.3	16.8	4.8
n = 50									
0.0	5.7	6.4	2.0	8.1	6.4	3.3	7.7	5.1	4.9
0.5	35.3	50.6	24.4	30.1	38.4	17.5	18.3	15.6	12.9
1.0	89.5	97.5	78.7	76.0	89.2	57.6	39.7	37.5	33.4
n = 200									
0.0	4.7	5.3	4.2	4.8	4.7	4.1	7.1	6.1	6.4
0.5	93.4	98.0	91.0	77.1	93.3	75.3	36.8	37.5	35.6
1.0	100.0	100.0	100.0	100.0	100.0	100.0	88.5	88.8	87.4

†For each scenario the power of PIM is compared with a traditional regression technique: least squares, LS, partial likelihood, PL, maximum likelihood, ML, or the bootstrap, BT.

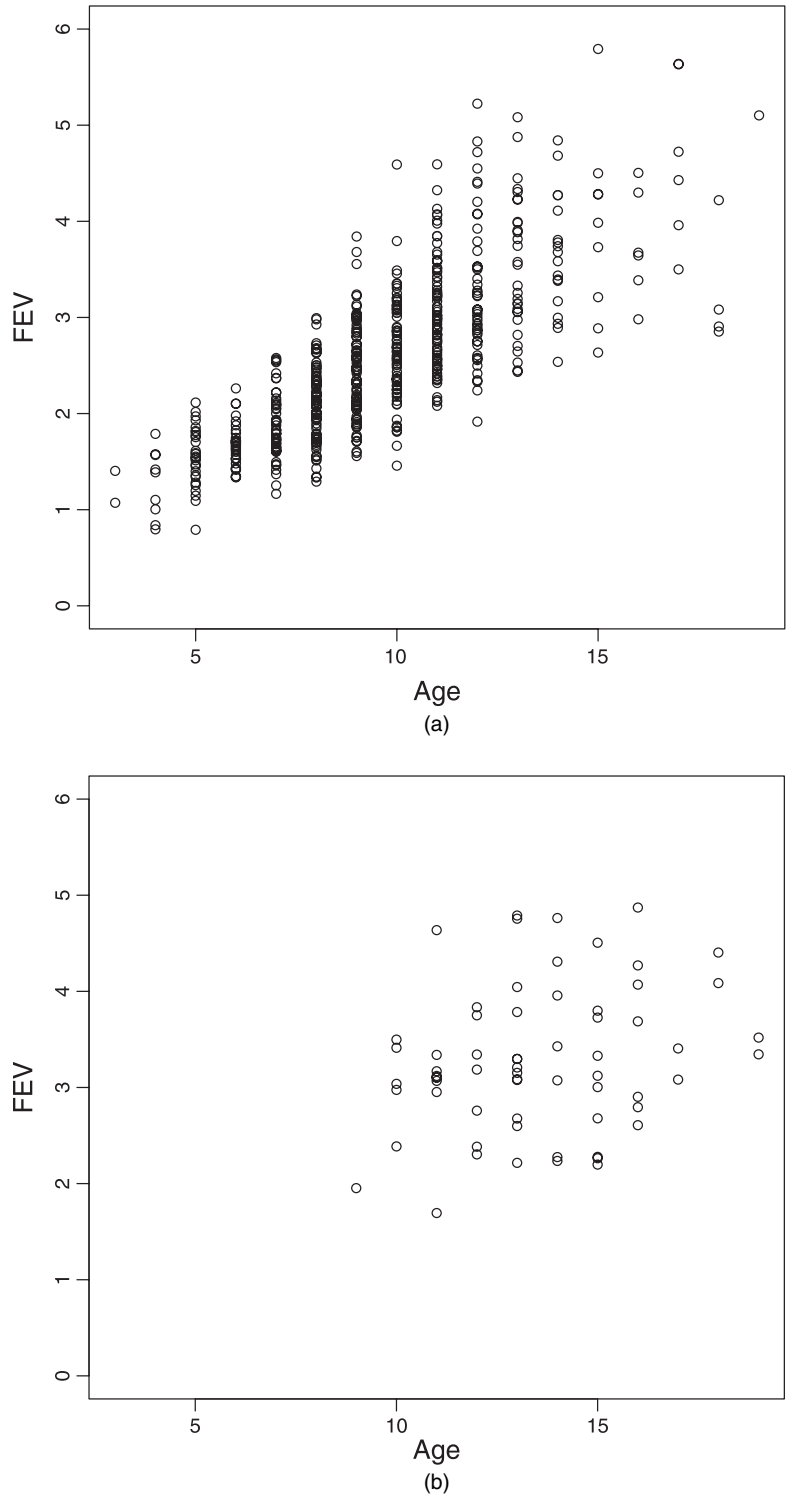

Fig. 2. FEV as a function AGE for (a) smokers and (b) non-smokers

Monte Carlo simulation runs. For a small sample size ($n = 20$) Brumback's test shows complete breakdown by showing virtually no power, and the tests based on the PIM are liberal. The tests based on least squares are also liberal under model 1. When $n = 50$ all tests have sizes that are not too far from the nominal level of 5%, but the PIM-based tests are often still slightly liberal and Brumback's test is often still conservative (although not for model 3). When $n = 200$ all tests are nearly unbiased. The powers of the tests in the PIM framework are generally larger than those of Brumback's test. The test based on least squares (model 1), partial likelihood (model 2) and maximum likelihood (model 3) are slightly more powerful, as expected.

6. Examples

To illustrate the interpretation of the PIM we present several examples. In Section 6.1 we present the data analysis for a continuous response and two predictors showing interaction. The example of Section 6.2 has an ordinal response variable and two predictors with no interaction. An example data set with a continuous heteroscedastic response variable and one single continuous regressor is presented in Section 6.3. All PIMs are defined for the lexicographical order relation and they all satisfy the antisymmetry condition; see Section 3 for more information. For notational convenience we 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 Development Core Team, 2010).

6.1. Childhood respiratory disease study

The 'Childhood respiratory disease study' is a longitudinal study following pulmonary function in children. We consider only the part of this study that was provided by Rosner (1999). The response variable is the forced expiratory volume FEV, which is an index of pulmonary function measured as the volume of air expelled after 1 s of constant effort. Along with FEV (litres), the covariates 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 years. See Rosner (1999), page 41, for more information. The primary focus is on the analysis of the effect of smoking status on pulmonary function. Fig. 2 displays FEV as a function of the AGE and SMOKING status; note that all very young children are non-smokers. The WMW test is a natural choice. However, it is believed that age may be a potential confounder, and thus the effect of smoking on FEV should be adjusted for age. This is illustrated in Fig. 3, which shows density estimates of the FEV distributions for several combinations of smoking status and age. Fig. 3 also suggests an interaction between age and smoking status. It is also of interest to quantify the effect of age.

For comparison we first analyse the data with a linear regression model with mean

$$E(\text{FEV}) = \alpha_0 + \alpha_1 \,\text{AGE} + \alpha_2 \,\text{SMOKE} + \alpha_3 \,\text{AGE} * \text{SMOKE}. \tag{27}$$

Table 5 gives the model fit with ordinary least squares. Since the residual plot (which is not shown) indicates non-constant variance of the error, we also analyse the data by using weighted least squares (see Table 5). The weights were obtained by fitting the absolute residuals of ordinary least squares in a linear regression model with the fitted values of ordinary least squares as the regressor.

With weighted least squares the effect of smoking on the mean level of FEV, while controlling for age, is estimated as $1.84 - 0.15\,\text{AGE}$. If we consider, for example, the age categories 12, 13, 14 and 15 years from Fig. 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 $[-0.19, 0.21]$,

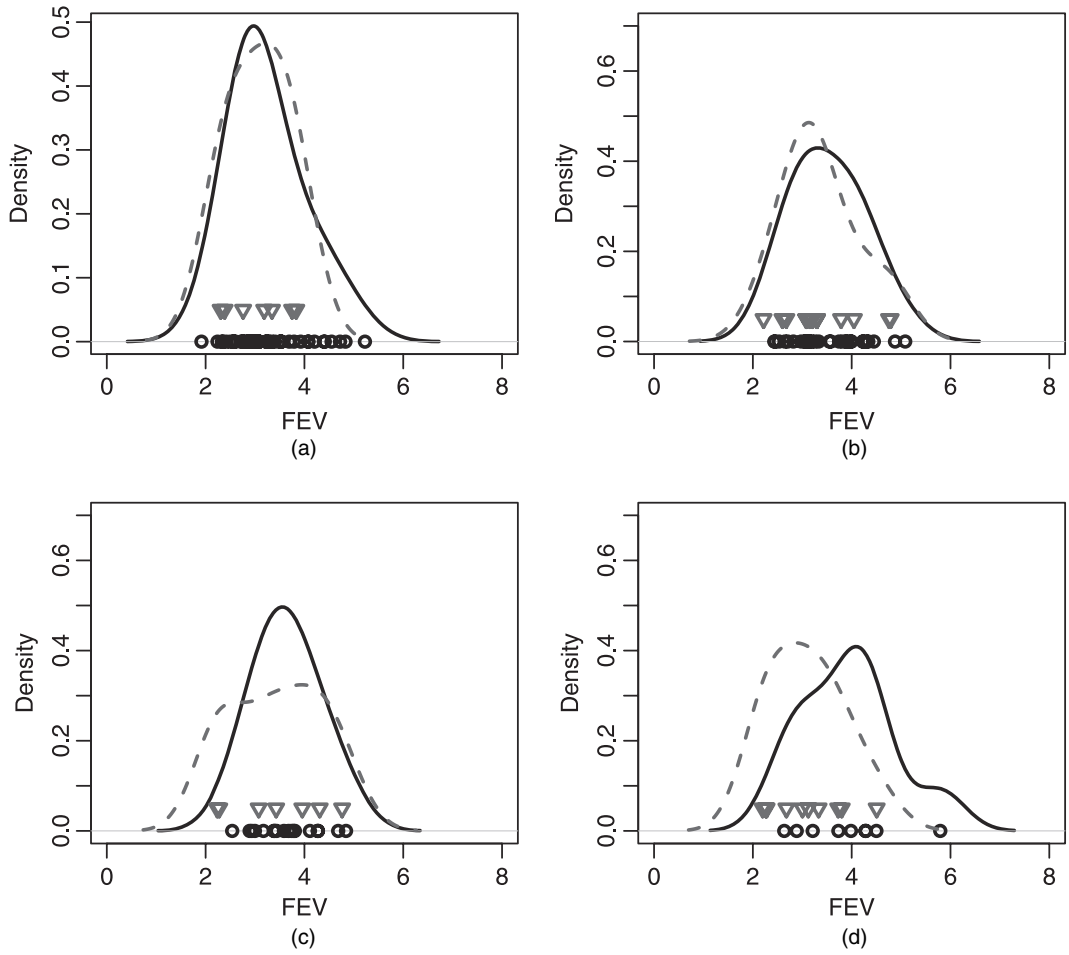

Fig. 3. Kernel density estimates of the FEV-distributions for smokers (∇) and non-smokers (○) of age (a) 12 years, (b) 13 years, (c) 14 years and (d) 15 years: the densities are estimated by using a Gaussian kernel with a bandwidth of 0.5; beneath each kernel density plot is a rug plot to identify better the individual sample observations that are used for the density estimation

$[-0.33, 0.05]$, $[-0.49, -0.09]$ and $[-0.68, -0.21]$. Thus for the ages of 14 and 15 years the mean FEV of non-smokers is significantly larger. When the smoking status is fixed, the mean FEV is estimated to change by $0.24 - 0.15\,\mathrm{SMOKE}$ when age increases by 1 year. For non-smokers this effect is thus 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 $[0.009, 0.156]$. Fig. 3 suggests that, while controlling for age, smoking not only affects the mean. The effect of smoking is also visible in higher order moments. The PI is well suited to quantify effects that do not act on one single moment of the response distribution.

We consider the PI model with interaction:

$$\begin{aligned}\mathrm{logit}\{P(\mathrm{FEV} \preccurlyeq \mathrm{FEV}^*)\} &= \beta_1(\mathrm{AGE}^* - \mathrm{AGE}) + \beta_2(\mathrm{SMOKE}^* - \mathrm{SMOKE}) \\ &\quad + \beta_3(\mathrm{AGE}^* * \mathrm{SMOKE}^* - \mathrm{AGE} * \mathrm{SMOKE}). \qquad (28)\end{aligned}$$

The model has no intercept, because, when $\mathrm{AGE}^* = \mathrm{AGE}$ and $\mathrm{SMOKE}^* = \mathrm{SMOKE}$, the model must give $P(\mathrm{FEV} \preccurlyeq \mathrm{FEV}^*) = \mathrm{expit}(0) = \frac{1}{2}$. The parameter estimates are presented in Table 5.

Table 5. Results of the ordinary least squares and weighted least squares fits of model (27) and the results of the fit of the PIM (28)

Parameter	*Estimate*	*Standard error*	*p-value*
Linear regression model ordinary least squares			
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
Linear regression model weighted least squares			
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

For a fixed age, the probability of having a smaller FEV, as a non-smoker as compared with a smoker, is estimated as $\text{expit}(\hat{\beta}_2 + \hat{\beta}_3\,\text{AGE}) = \text{expit}(5.31 - 0.46\,\text{AGE})$. This illustrates that the effect of smoking on the PI depends on age. For the age categories 12, 13, 14 and 15 years from Fig. 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.

In contrast, if the smoking status is fixed, the probability of having a larger FEV when age increases by 1 year is estimated as $\text{expit}(\hat{\beta}_1 + \hat{\beta}_3\,\text{SMOKE}) = \text{expit}(0.61 - 0.46\,\text{SMOKE})$. Thus for non-smokers this probability is estimated by $\text{expit}(0.61) = 65\%$ whereas for smokers this drops to $\text{expit}(0.15) = 54\%$. 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 response and the covariates. As for any other regression model, it is important to assess the validity of the model for a given data set. For this purpose we propose a simple graphical diagnostic tool which is based on a lack-of-fit method for logistic regression models (Hosmer and Lemeshow, 1980; Lemeshow and Hosmer, 1982; Hosmer *et al.*, 1988). When the model fits the data well, we expect that the predicted probabilities are close to the observed (empirical) probabilities. Thus a plot of the former *versus* the latter could serve for graphical model fit assessment. Hosmer and Lemeshow (1980) proposed to calculate the empirical probabilities within groups of observations. In particular, observations with similar predicted probabilities are grouped by partitioning the [0, 1] interval of the predicted probabilities on the basis of their deciles. For each interval the average predicted probability and the empirical probability are calculated. Fig. 4 shows the diagnostic plot; it suggests that the PIM fits the data well. As the pseudo-observations are not mutually independent, the distribution theory of the Hosmer–Lemeshow goodness-of-fit test does not directly apply to our setting.

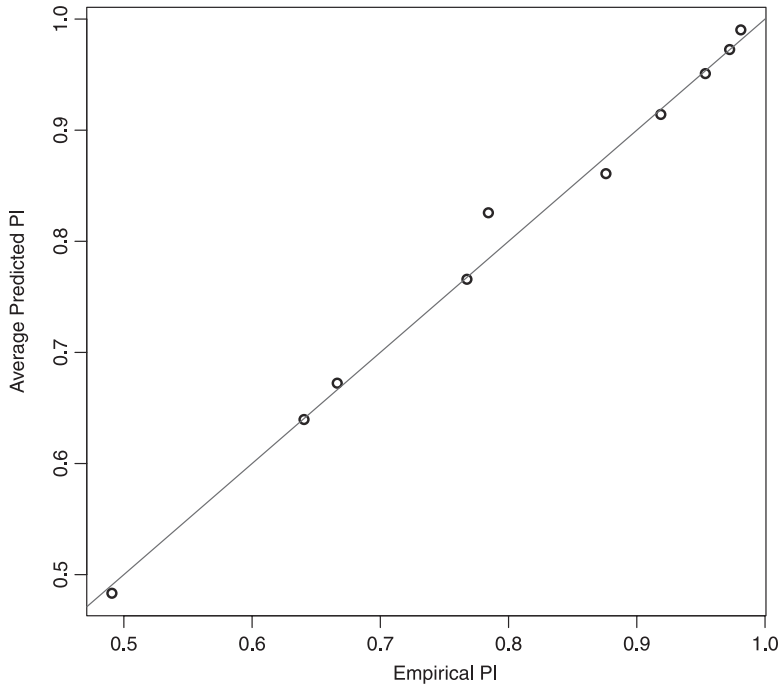

Fig. 4. Diagnostic plot for the respiratory disease data: the plot shows the average predicted PI according to the fitted PIM *versus* the empirical PI; the grouping is based on the deciles of the predicted PI

6.2. Mental health study

The 'Mental health study' is a study of mental health for a random sample of 40 adult residents of Alachua County, Florida. See Agresti (2007), page 185, for more information. The response variable 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 socio-economic status SES are also reported. SES is a binary variable coded as 0 (low socio-economic status) and 1 (high socio-economic status). LI is a composite measure that quantifies the severity and the number of important life events such as birth of a child, death in the family and divorce. One of the objectives of the study is to assess whether SES has an effect on MI. As the average MI-score has no clear interpretation, Agresti (2007) analysed the data with a cumulative logistic regression model. Here we analyse the data in terms of the PI. As it is believed that LI may be a potential confounder, we propose to analyse the mental health data with the PIM

$$\text{logit}\{P(\text{MI} \preccurlyeq \text{MI}^*)\} = \beta_1(\text{SES}^* - \text{SES}) + \beta_2(\text{LI}^* - \text{LI}). \tag{29}$$

The parameter estimates are presented in Table 6. The diagnostic plot for model (29) is shown in Fig. 5 (see Section 6.1 for details on the construction). The graph demonstrates that the PIM fits the data quite well. For comparison Table 6 also contains the maximum likelihood parameter estimates of the cumulative logit model,

$$\text{logit}\{P(\text{MI} \leqslant j)\} = \mu_j + \alpha_1 \text{SES} + \alpha_2 \text{LI}, \qquad j = 1, 2, 3, \tag{30}$$

for which the parameter estimates are obtained by using the `MASS` R package (Venables and Ripley, 2002).

Table 6. Results of the fits of the PIM (29) and the cumulative logit model (30)

Parameter	*Estimate*	*Standard error*	*p-value*
PIM (29)			
SES (β_1)	−0.74	0.34	0.03
LI (β_2)	0.20	0.07	0.006
Cumulative logit model (30)			
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

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. With $\hat{\beta}_1 = -0.74$ we conclude that, of people with equal LI, someone with a high socio-economic status has an estimated probability of $\text{expit}(-0.74) = 32\%$ to have a larger MI-score than someone with a low socio-economic status and a 95% confidence interval is given by [20%, 48%]. People with a low socio-economic status are thus more likely to be mentally impaired than others with a high socio-economic status, while all having the same LI. The effect of LI on MI can be estimated by the probability $\text{expit}(\hat{\beta}_2)$. In particular, among people with the same SES, those with an LI of 1 unit smaller than the LI of another group of people have a smaller MI-score with estimated probability $\text{expit}(0.2) = 55\%$, with a 95% confidence interval of [51%, 59%]. Thus, the larger the LI, the more likely someone is to be mentally impaired.

The cumulative logit model (30) gives no significant effect of SES at the 5% level of significance ($p = 0.07$). Similarly to PIM analysis, there is a significant effect of the life index on the cumulative logit ($p = 0.008$): if LI increases by 1 unit, the odds that the mental impairment score is not larger than a particular level decreases by an estimated factor $\exp(-0.32) = 0.73$ with a 95% confidence interval of [0.56, 0.91]. Although the conclusions based on the PIM and the cumulative logit model agree quite well, there is thus a difference in interpretation. The cumulative logit model (30) can be further extended so that the covariate effect on the odds ratios for the events $\text{MI} \leqslant j$ depends on the level j. Since this more complex model does not fit significantly better (the results are not shown; $p = 0.68$), we keep the model with the proportional odds assumption. The PIM (29) can also be extended so that the effects of SES and LI on the PI depend not only on the differences $\text{SES}^* - \text{SES}$ and $\text{LI}^* - \text{LI}$, but also on the covariates themselves. For example,

$$\text{logit}\{P(\text{MI} \preccurlyeq \text{MI}^*)\} = \beta_1(\text{SES}^* - \text{SES}) + \beta_2(\text{LI}^* - \text{LI}) + \beta_3\,\text{SES} + \beta_4\,\text{LI}, \tag{31}$$

which is well defined for the strict lexicographical order restriction $\text{SES} < \text{SES}^*$, or $\text{SES} = \text{SES}^*$ and $\text{LI} < \text{LI}^*$. However, this more complex model did not fit significantly better (the results are not shown; $p = 0.77$). Note that the addition of $\beta_3\,\text{SES}$ and $\beta_4\,\text{LI}$ in model (31) is another way of introducing an interaction effect.

Although both models may have their merits and shortcomings we believe that the PIM has the advantage of quantifying the effects of the covariates on the response distribution more directly. More specifically the PIM (29) models the log-odds,

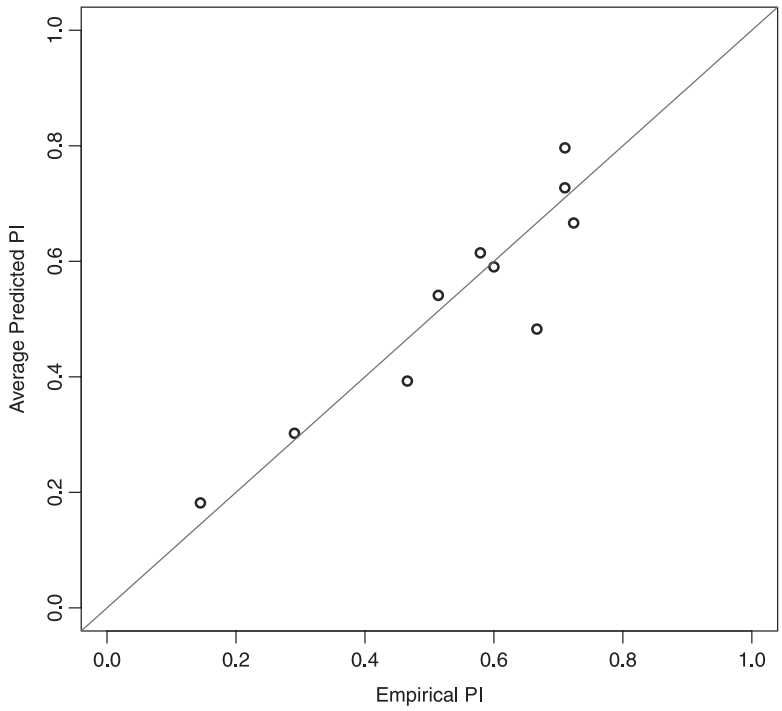

Fig. 5. Diagnostic plot for the mental health data: the plot shows the average predicted PI according to the fitted PIM *versus* the empirical PI; the grouping is based on the deciles of the predicted PI

$$\log\{\text{odds}(\text{MI} \preccurlyeq \text{MI}^* | \text{SES}, \text{SES}^*, \text{LI}, \text{LI}^*)\},$$

whereas the proportional odds model (30) models the log-odds-ratio,

$$\log\left\{\frac{\text{odds}(\text{MI} \leqslant j | \text{SES}, \text{LI})}{\text{odds}(\text{MI}^* \leqslant j | \text{SES}^*, \text{LI}^*)}\right\}.$$

6.3. Food expenditure data set

The food expenditure data set contains data on food expenditure FE (in Belgian francs), and 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 that is spent on food falls with increasing income, even if actual expenditure on food rises. The data were also used in Koenker (2005) for the illustration of quantile regression; the data are also available in the `quantreg` R package (Koenker, 2011). Fig. 6(a) plots FE *versus* HI as well as showing a fitted linear model based on weighted least squares. The weights were obtained by fitting the squared residuals of a classical least squares fit in a linear regression model with the fitted values of the least squares fit as the regressor. As a result of the increasing variability in food expenditure as household income increases, we analyse the data with the PIM

$$\text{logit}\{P(\text{FE} \preccurlyeq \text{FE}^*)\} = \beta \frac{\text{HI}^* - \text{HI}}{\sqrt{(\text{HI}^* + \text{HI})}}, \tag{32}$$

in which the denominator $\sqrt{(\text{HI}^* + \text{HI})}$ is suggested by the arguments that were given in Section 4.1. The estimated slope is $\hat{\beta} = 0.39$ ($p < 0.001$ and 95% confidence interval $[0.34, 0.44]$). This

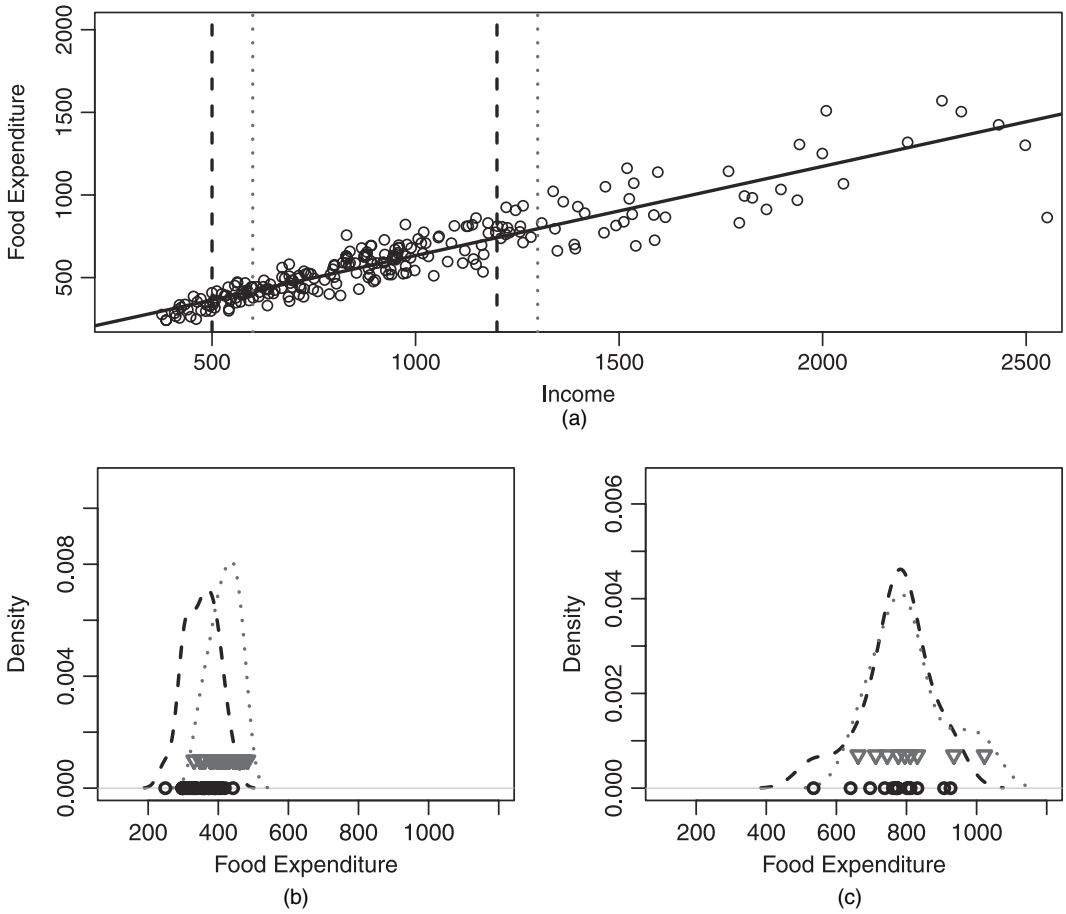

Fig. 6. (a) Scatter plot of the food expenditure data with a fitted linear regression line, and non-parametric Gaussian kernel smoother density estimates with bandwidths (b) 20 and (c) 50 of the food expenditure for household incomes (b) 500 (○) and 600 (∇) and (c) 1200 (○) and 1300 BEF (∇): beneath each kernel density plot is a rug plot to identify better the individual sample observations that are used for the density estimation; the notation $P\{Y(500) < Y(600)\}$ and $P\{Y(1200) < Y(1300)\}$ is used as a compact notation for the PI

analysis supports Engel's hypothesis. Indeed if the household income is 500 BEF then the probability of larger food expenditure with a household income of 600 BEF is estimated as 76% with a 95% confidence interval of [74%, 79%]. When we compare households with 1200 and 1300 BEF this estimated probability drops to 69% with a 95% confidence interval of [66%, 71%]. This is an example of the negative effect modification of the increasing error variance (see Section 4.1). Figs 6(b) and 6(c) illustrate this phenomenon. As the data set contains no two households with exactly the same income, an observation with income u is assigned to income v if $|u - v| < 50$ BEF.

The diagnostic plot is presented in Fig. 7; it shows convincingly a very good fit.

7. Conclusion

We have introduced a general class of semiparametric models for the PI. The models apply to continuous and ordinal response variables. The parameters of the PIM have direct

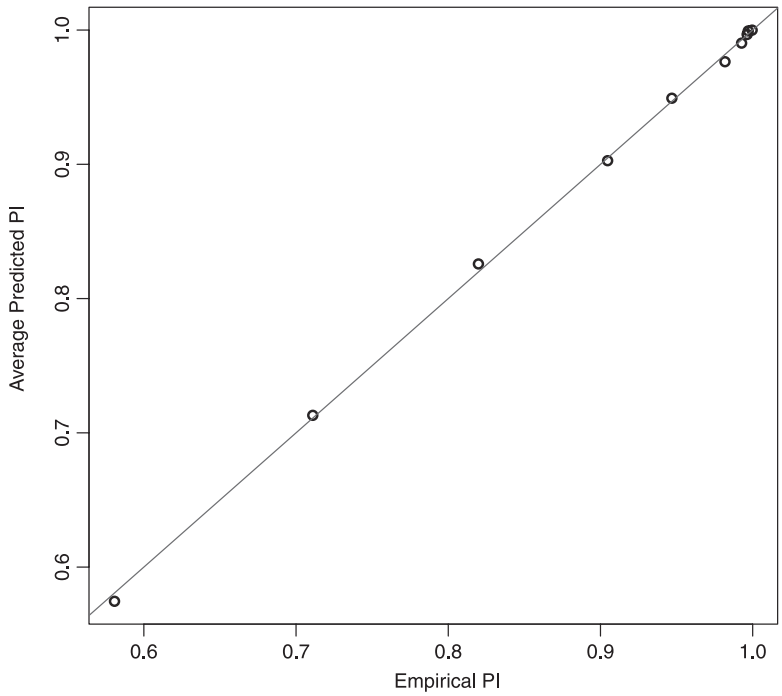

Fig. 7. Diagnostic plot for the food expenditure data: the plot shows the average predicted PI according to the fitted PIM *versus* the empirical PI; the grouping is based on the deciles of the predicted PI

and informative interpretations that have been illustrated on four data sets. The PIM framework may be considered as a generalization of the area under the curve regression models of Dodd and Pepe (2003) and of the related covariate-corrected WMW test of Brumback *et al.* (2006). It extends these methods by providing a more flexible model formulation that

(a) not only applies to the comparison of response variables for two treatment groups,
(b) is not restricted to continuous responses and
(c) includes a consistent estimator of the covariance matrix of the parameter estimators without relying on the bootstrap method.

The asymptotic theory that we have presented is based on the work of Lumley and 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 future research we plan to improve the methods further by the construction of efficient score functions. The results of our simulation study demonstrate that the theoretical properties of the parameter and variance estimators apply well to moderately sized samples, and that the powers of our tests are quite good.

The semiparametric PIMs are flexible, but, as for all regression models, they impose some restrictions on the conditional distribution of the response variable. Therefore we have proposed a simple graphical diagnostic tool that is based on the ideas of Hosmer and Lemeshow (1980). The development of more formal lack-of-fit tests for the PIMs may be an interesting direction for future research. In particular, we believe that the ideas of Deschepper *et al.* (2006) and Hart

(1997) may be helpful. Another restriction in the present definition of the PIMs is the linearity in the predictor, which in our current model is formulated in the spirit of generalized linear models (McCullagh and Nelder, l989). Future extensions may involve non-parametric regression terms which may be estimated by means of kernel smoothers, splines or any other type of non-parametric estimator, eventually resulting in PIMs that resemble generalized additive models (Hastie and Tibshirani, 1990) for the PI.

In the present paper PIMs are only defined for use with mutually independent observations. Extensions to clustered and longitudinal data would also be very useful. This may involve the introduction of random-effect terms in the linear predictor, or it may be accomplished through extensions of the estimating equations.

Finally, we want to stress that PIMs are not to be considered as a competitor of other classes of statistical models. We rather think that the PIM framework is a valuable addition to the statisticians' toolbox which may be used whenever the PI is chosen as a meaningful scale for the formulation of the research question.

Acknowledgements

This research was supported by Interuniversity Attraction Pole research network grant P6/03 of the Belgian Government (Belgian science policy). The authors also thank Stijn Vansteelandt for interesting discussions, and the referees for very constructive comments.

References

Acion, L., Peterson, J., Temple, S. and Arndt, S. (2006) Probabilistic index: an intuitive non-parametric approach to measuring the size of treatment effects. *Statist. Med.*, **25**, 591–602.

Agresti, A. (2007) *An Introduction to Categorical Data Analysis*. Hoboken: Wiley.

Beck, A., Steer, R. and Garbin, M. (1988) Psychometric properties of the beck depression inventory: twenty-five years of evaluation. *Clin. Psychol. Rev.*, **8**, 77–100.

Beyerlein, A., Fahrmeir, L., Mansmann, U. and Toschke, A. (2008) Alternative regression models to assess increase in childhood BMI. *BMC Med. Res. Methodol.*, **8**, no. 1.

Browne, R. (2010) The t-test p value and its relationship to the effect size and $P(X > Y)$. *Am. Statistn*, **64**, 30–33.

Brumback, L., Pepe, M. and Alonzo, T. (2006) Using the ROC curve for gauging treatment effect in clinical trials. *Statist. Med.*, **25**, 575–590.

Chamberlain, G. (1987) Asymptotic efficiency in estimation with conditional moment restrictions. *J. Econmetr.*, **34**, 305–334.

Cox, D. R. (1972) Regression models and life-tables (with discussion). *J. R. Statist. Soc.* B, **34**, 187–220.

Deschepper, E., Thas, O. and Ottoy, J. (2006) Regional residual plots for assessing the fit of linear regression models. *Data Anal. Computnl Statist.*, **50**, 1995–2013.

Dodd, L. and Pepe, M. (2003) Semi-parametric regression for the area under the receiver operating characteristics curve. *J. Am. Statist. Ass.*, **98**, 409–417.

Enis, P. and Geisser, S. (1971) Estimation of the probability that $Y < X$. *J. Am. Statist. Ass.*, **66**, 162–168.

Fishburn, P. C. (1974) Lexicographic orders, utilities and decision rules: a survey. *Mangmnt Sci.*, **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. *J. Am. Statist. Ass.*, **76**, 162–168.

Hart, J. (1997) *Nonparametric Smoothing and Lack-of-fit Tests*. Berlin: Springer.

Hastie, T. and Tibshirani, R. (1990) *Generalized Additive Models*. New York: Chapman and Hall.

Hodges, J. and Lehmann, E. (1963) Estimation of location based on ranks. *Ann. Math. Statist.*, **34**, 598–611.

Højsgaard, S., Halekoh, U. and Yan, J. (2005) The R package geepack for generalized estimating equations. *J. Statist. Softwr.*, **15**, no. 2, 1–11.

Holt, J. and Prentice, R. (1974) Survival analysis in twin studies and matched pair experiments. *Biometrika*, **61**, 17–30.

Hosmer, D. and Lemeshow, S. (1980) A goodness-of-fit test for the multiple logistic regression model. *Communs Statist. Theor. Meth.*, **10**, 1043–1069.

Hosmer, D., Lemeshow, S. and Klar, J. (1988) Goodness-of-fit testing for multiple logistic regression analysis when the estimated probabilities are small. *Biometr. J.*, **30**, 1–14.
Kalbfleisch, J. and Prentice, R. (1973) Marginal likelihoods based on Cox's regression and life model. *Biometrika*, **60**, 267–278.
Koenker, R. (2005) *Quantile Regression*. Cambridge: Cambridge University Press.
Koenker, R. (2011) quantreg: quantile regression. *R Package Version 4.54*.
Kotz, S., Lumelskii, Y. and Pensky, M. (2003) *The Stress–Strength Model and Its Generalizations: Theory and Applications*. Singapore: World Scientific Publishing.
Laine, C. and Davidoff, F. (1996) Patient-centered medicine: a professional evolution. *J. Am. Med. Ass.*, **275**, 152–156.
Lemeshow, S. and Hosmer, D. (1982) A review of goodness-of-fit statistics for use in the development of logistic regression models. *Am. J. Epidem.*, **115**, 92–106.
Liang, K. and Zeger, S. (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.
Liu, I. and Agresti, A. (2005) The analysis of ordered categorical data: an overview and a survey of recent developments. *Test*, **14**, 1–73.
Lumley, T. and Hamblett, N. (2003) Asymptotics for marginal generalized linear models with sparse correlations. *Technical Report 207*. University of Washington, Seattle.
McCullagh, P. (1980) Regression models for ordinal data (with discussion). *J. R. Statist. Soc.* B, **42**, 109–142.
McCullagh, P. and Nelder, J. (1989) *Generalized Linear Models*, 2nd edn. London: Chapman and Hall.
McKean, J. (2004) Robust analysis of linear models. *Statist. Sci.*, **19**, 562–570.
McKean, J., Terpstra, J. and Kloke, J. (2009) Computational rank-based statistics. *Wiley Interdisc. Rev. Computnl Statist.*, **1**, 132–140.
Myles, P., Troedel, S., Boquest, M. and Reeves, M. (1999) The pain visual analog scale: is it linear or nonlinear? *Anesth Analg.*, **89**, 1517–1520.
Newey, W. (1988) Adaptive estimation of regression models via moment restrictions. *J. Econmetr.*, **38**, 301–339.
Pepe, M. (2003) *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford: Oxford University Press.
R Development Core Team (2010) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
Rosner, B. (1999) *Fundamentals of Biostatistics*. Pacific Grove: Duxbury.
Thas, O. (2009) *Comparing Distributions*. New York: Springer.
Therneau, T. and Lumley, T. (2010) survival: survival analysis, including penalised likelihood. *R Package Version 2.36-2*.
Tian, L. (2008) Confidence intervals for $P(Y1 > Y2)$ with normal outcomes in linear models. *Statist. Med.*, **27**, 4221–4237.
Tsiatis, A. (2006) *Semiparametric Theory and Missing Data*. New York: Springer.
Turk, D., Rudy, T. and Sorkin, B. (1993) Neglected topics in chronic pain treatment outcome studies: determination of success. *Pain*, **53**, 3–16.
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. *J. Clin. Psychpharm.*, **28**, 147–155.
Venables, W. N. and Ripley, B. D. (2002) *Modern Applied Statistics with S*, 4th edn. New York: Springer.
Wallerstein, S. (1984) *Scaling Clinical Pain and Pain Relief*. New York: Elsevier.
Zeger, S. and Liang, K. (1986) Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, **42**, 121–130.
Zhou, W. (2008) Statistical inference for $P(X < Y)$. *Statist. Med.*, **27**, 257–279.

Discussion on the paper by Thas, De Neve, Clement and Ottoy

Thomas Alexander Gerds (*University of Copenhagen*)

I am pleased to welcome this paper to the Society. At a first glance the probabilistic index model (PIM) is the instrument that has always been missing in my toolbox: a multiple-regression model which generalizes the Wilcoxon rank sum test. If it is as indicated, we can now leave the multiple (normal) linear regression model and use a PIM as a robust alternative. I believe that this class of models will have a significant influence on applied statistical work. Let me outline two arguments.

(a) PIMs will be used by young statisticians. Let us think about a young statistician as someone working with a *default* toolbox equipped with *default* tools composed according to their type and place of education. Such a person aims to apply the *correct* tool to a given problem and may believe that it is *wrong*, say, to apply a t-test when the outcome is a discrete variable. Hence, a young statistician applies a PIM when the task is to do multiple-regression analysis of apparently not normally distributed outcomes. However, a PIM certainly cannot solve all the problems of

experienced statisticians who know how to apply the *wrong* tools and still arrive at sound conclusions.

(b) PIMs will find their way into medical statistics. The generic argument of Thas and his colleagues is that their models provide effect measures which have an intuitive interpretation. They also make an important connection to area under the curve regression. The relationship can be extended to the concordance index which generalizes the area under the curve. The concordance index is widely used to assess the discrimination ability of risk prediction models (Harrell *et al.*, 1996). It is usually defined as the probability that the risk predicted for person i is greater than that for person j given that the event occurs earlier for person i, i.e.

$$C = P(R_i \succcurlyeq R_j | T_i < T_j).$$

From the PIM perspective a more natural formulation is to condition the order of the outcome on the order of the predicted risks:

$$C = P(T_i \preccurlyeq T_j | R_i > R_j).$$

This latter formulation is appealing since one does not condition on the future. It can be noted that if both the predicted risks and the event times are continuous variables then $P(T_i < T_j) = P(R_i > R_j) = 0.5$ and hence the two formulations are equivalent. A potentially interesting new application of PIMs is to test whether a biomarker X improves the predictive ability of the prediction model with a suitably formulated 'concordance index model', e.g.

$$P(T_i \preccurlyeq T_j | R_i, R_j, X_i, X_j) = g^{-1}(\beta \mathbf{1}\{R_i > R_j\} + \gamma \mathbf{1}\{X_i - X_j\}).$$

Thas and his colleagues discover a fascinating relationship between the PIM and the Cox model in Section 4.2 but they do not deal with censored data. In survival analysis we observe only $\min(T_i, C_i)$ where C_i is the censoring time. A necessary first step is to truncate the pseudovalue at a time t where the probability of being uncensored is positive:

$$I_{ij}(t) = \mathbf{1}\{T_i \preccurlyeq T_j, T_i < t\},$$
$$P(C_i > t) > 0.$$

Still, the value of the pseudovalue is unknown for pairs where $T_i > C_i$. To deal with censored data one possibility would be to apply inverse probability weighting. Here I propose a different approach. The idea is to construct a pseudovalue for the pseudovalue following Andersen *et al.* (2003). Apart from correction terms it is given as.

$$\tilde{I}_{ij}(t) = n^2 \int_0^t \{1 - \hat{F}(s)\} \, \mathrm{d}\hat{F}(s) - (n-1)^2 \int_0^t \{1 - \hat{F}^{(j)}(s)\} \, \mathrm{d}\hat{F}^{(j)}(s)$$

where $\hat{F}$ is the Kaplan–Meier estimate calculated with all the data and $\hat{F}^{(j)}$ is the Kaplan–Meier estimate when the data from the jth patient have been removed. I conjecture that if censoring is independent of the covariates and the event times then one can argue by using a second-order von Mises expansion of the Kaplan–Meier estimator as in Graw *et al.* (2009) to show that

$$E\{\tilde{I}_{ij}(t) | X_i, X_j\} = E\{I_{ij}(t) | X_i, X_j\}.$$

Under the usual regularity conditions on the link function, estimating equations based on the pseudo-pseudovalue will be asymptotically consistent. Note that in uncensored data the pseudo-pseudovalues are equal to the pseudovalues. I close my discussion with a couple of remarks.

(i) Thas and his colleagues treat ties rigorously throughout their paper. A potentially important further distinction is between ties that occur due to observational imprecision and real ties where the underlying characteristics are equal for some individuals.

(ii) In Section 6, pseudocalibration plots are used to assess goodness of fit. Thas and his colleagues note a problem with the arbitrary grouping that is inherent in these plots and in the Hosmer–Lemeshow test. To avoid arbitrary grouping one could use non-parametric smoothing (Le Cessie and Van Houwelingen, 1991), or measure the calibration by the expected value of a strictly proper scoring rule (Gneiting and Raftery, 2007). For example, one could measure calibration by the average mean-squared pseudoerror

$$\frac{1}{n}\frac{1}{n}\sum_i\sum_j[\mathbf{1}\{Y_i \leqslant Y_j\} - \mathrm{PIM}(X_i, X_j, \beta)]^2.$$

The PIM should score below the benchmark of 25% which is obtained when 50% chance is predicted for the event $\{Y_i \leqslant Y_j\}$ independently of i and j.

(iii) To improve the interpretation of PIMs further one could introduce an offset into the probabilistic index: $P(Y^* \leqslant Y - \varepsilon)$. Then the regression parameters in a suitably defined PIM would express the effects of predictor variables on the probability that the outcome will be reduced by at least ε, which could be a clinically meaningful change.

In summary, I think that Thas and his colleagues have provided us with a new hammer for the *default* toolbox. It gives me great pleasure to propose the vote of thanks.

Stephen Senn (*Centre de Recherche Public de la Santé, Strassen*)

We need many ways of looking at data and a technical exploration of an alternative approach to modelling with a chance for discussion should always be welcome to this Society. As such it is a pleasure to second the vote of thanks for this interesting paper. It is, however, the tradition of this Society for seconders to be critical and although I think that it will be good for the applied statistician to know that these techniques exist and have been developed in detail I also think that will *usually* be wise for the statistician not to use them (Senn, 2011).

Before explaining why, I draw attention to some connections. The authors use the term *probabilistic index* (PI). They refer to the fact that *individual exceedence probability* has been used before but do not give the reference, which I now provide (Senn, 1997). Recently Buyse (2010) has proposed a multivariate version called the *proportion in favour of treatment*. More important, however, is that, in the context of longitudinal data and factorial experiments, there is an extensive treatment of *relative treatment effects* using *normalized empirical distribution functions* in the beautiful book by Brunner *et al.* (2001). I urge the authors to study this as I believe that they will find many interesting connections to their work. The indicator function that is defined at the beginning of Section 3.1 is essentially, of course, the Heaviside function $H(d), d = Y^* - Y$, and this raises the possibility of a close connection to the very extensive theory of counting processes applied to survival analysis. Of course the authors themselves develop a connection in Section 4.2 and, indeed, Kalbfleisch and Prentice (1980) to whom they refer used $H(d)$.

I illustrate my reasons for distrusting the PI as a measure of effect by looking at the first example. First, note that violation of the linearity assumption for this example is a red herring. The technique proposed does *nothing* to deal with this. If the conditional distribution of the response depends on the dose in the way implied, then there is no universal effect of a 5-g change whether measured by the PI or more conventionally. Furthermore, the reference to the ordinal nature of the Beck depression inventory is also misleading. This is a sum of 21 items and since change from baseline is used a 21st linear operation has been added unnecessarily to the 20 already performed in its construction. If the Beck depression inventory change score is truly ordinal it only is so because it is interval. Furthermore, the use by the authors of change scores immediately raises a worrying issue. Suppose that we take the simplest case where we assume that although baseline is related to outcome it is unrelated to dose (if this is not so then *any* modelling approach will have to tread delicately). If this is so then it makes no difference (in expectation) to the conditional estimate of the 'effect' of dose by using conventional least squares whether we use raw outcomes only, or differences from baselines only or condition on the baselines by using them as a covariate. However, even in the best behaved of cases the PI would be quite different since it is, essentially, a signal-to-noise ratio: the degree of overlap depends not just on the signal but also on the noise.

The authors seem to see this as a good thing. I cannot agree. Consider a placebo-controlled trial of an angiotensin-converting enzyme inhibitor in hypertension with diastolic blood pressure as the outcome measures. Any one of the following will change the value of the PI even if the effect as conventionally measured is stable across patients: narrowing or broadening the inclusion criteria; taking more precise measurements; using the average of a number of measurements; using the difference from baseline; stratifying. Can this be a good thing? Can physicians let alone patients interpret the resulting PI? Consider the authors' first example. It is certainly a challenge to explain what this 70.2% is. Is it the probability that a randomly chosen patient will improve his or her value if given 5 g extra?: no. Is it the probability that such a patient will benefit from taking 5 g extra?: no. Is it an inherent property of the treatment?: no. It is a combined property of the treatment, the variability of the way that we measure it and the variability of the patients we happen to have recruited into a study for which we did not use random sampling. I shall repeat what I have said in discussion of such measures previously: only those who misunderstand them will find them simple (Senn, 2006).

Of course, all statisticians use measures that are not collapsible: odds ratios and hazard ratios (Ford *et al.*, 1995; Gail *et al.*, 1984; Robinson and Jewell, 1991) are cases in point. However, one justification for using such measures is that they should be interpreted ultimately in terms of predictions (Lee and Nelder, 2004; Senn, 2004). Whatever problems such measures have the PI will have much worse.

However, now that my grumble is over, I return to my opening comments. I found this paper interesting and stimulating, as I am sure will do many who read it and as did all who heard it delivered. In encouraging us to think again it increases our understanding of what we are doing in statistical modelling. I applaud the authors' final remarks that the approach to analysis is not a superior substitute for conventional approaches but a possible alternative or perhaps supplement on occasion. If this is so one cannot but welcome this exposition and exploration and I am very pleased to second the vote of thanks.

The vote of thanks was passed by acclamation.

Ingrid Van Keilegom (*Université catholique de Louvain*)

I first congratulate the authors for this interesting paper and important contribution to the area of semiparametric regression modelling. The model proposed has many links with other known models in the literature and has the advantage of allowing the response to be discrete and even ordinal.

As the authors point out, it is not clear for the moment whether their estimation procedure is efficient. To shed some light on the estimation of the model and on its semiparametric efficiency bound, which is at the same time an important and a very difficult problem, I concentrate here on the case where Y is continuous and rewrite the model as

$$h(Y) = g^{-1}(Z^{\mathrm{T}}\beta) + \varepsilon, \tag{33}$$

where $h(y) = P(Y^* \geqslant y|X^*)$, $E(\varepsilon|Z) = 0$ and $\mathrm{var}(\varepsilon|Z) = \sigma^2(Z)$. Indeed, we can write $P(Y \leqslant Y^*|X, X^*) = E_Y[P(Y \leqslant Y^*|X, X^*, Y)|X, X^*] = E[h(Y)|X, X^*]$. This shows that the probabilistic index model (PIM) is a special case of a transformation model. The transformation methodology has been quite successful and a large literature exists on this subject for parametric models; see for example Carroll and Ruppert (1988) among many others. To estimate β, we can now proceed as follows. Define

$$m(Z, Y, \beta, h, \sigma^2) = \sigma^{-2}(Z)\{h(Y) - g^{-1}(Z^{\mathrm{T}}\beta)\}\frac{\partial g^{-1}(Z^{\mathrm{T}}\beta)}{\partial \beta}.$$

Then, $E[m(Z, Y, \beta, h, \sigma^2)] = 0$. To estimate β, first replace h and σ^2 by non-parametric estimators (say $\hat{h}$ and $\hat{\sigma}^2$), and then define the estimator $\hat{\beta}$ by solving the system of equations

$$n^{-1}\sum_{i=1}^{n} m(Z_i, Y_i, \beta, \hat{h}, \hat{\sigma}^2) = 0$$

with respect to β. The asymptotic normality of $\hat{\beta}$ can be obtained from Chen *et al.* (2003), who developed primitive conditions for the asymptotic normality of any semiparametric Z-estimator.

A second way to look at the PIM is by rewriting the model as

$$\phi\{S(\cdot|X^*)|X\} = g^{-1}(Z^{\mathrm{T}}\beta), \tag{34}$$

where $S(y|X^*) = P(Y^* \geqslant y|X^*)$ and $\phi\{S(\cdot|X^*)|X\} = E[S(Y|X^*)|X, X^*]$. By writing the PIM in this way, it becomes a special case of the model that was studied by Grigoletto and Akritas (1999), except that the function ϕ depends on X here.

Since models (33) and (34) are well known and have been well studied in the literature, they can be helpful in determining the semiparametric efficiency bounds of the PIM. However, the estimation of these models builds almost inevitably on the estimation of conditional functions (h and σ^2 for model (33), and ϕ for model (34)), which can be a difficult task involving for example the delicate choice of smoothing parameters, whereas the estimation method that is proposed by the authors does not rely on any smoothing methods.

Lori E. Dodd (*National Institute of Allergy and Infectious Diseases, Bethesda*) (© US Government)

The probabilistic index (PI) arises naturally where relative orderings of outcomes from pairs of observations can be assigned. PI-like indices have been used extensively in psychophysics, in which subjective readers may be unable to assign scores directly but can rank pairs of images with respect to 'signal' or 'noise' in what are referred to as two-alternative forced choice experiments (Green and Swets, 1966). In clinical trials, the PI has been proposed as a clinically intuitive way of combining multiple outcomes

Table 7. Coefficient estimates under different orderings

	SES $\leqslant$ SES*	SES* $\leqslant$ SES	No ordering
$(\hat{\beta}_1, \hat{\beta}_2 - \hat{\beta}_1)$	(−0.04, 0.60)	(−0.53, 1.09)	(0.70, 0.89)

(Follmann, 2002). For example, a rule-based method to combine the outcomes of death and hospitalization might proceed as follows.

(a) Death is the worst outcome; earlier death worse than later.
(b) Among survivors, hospitalization for disease is the worst outcome, with early hospitalization worse than later.

Patient outcomes can be naturally ordered and covariate effects evaluated by using 'pairwise ordering regression' (POR) (Follmann, 2002), which is similar to probabilistic index models (PIMs).

Thas and his colleagues nicely demonstrate the relationship between PIMs and standard linear regression, proportional hazards (PH) and rank regression models. For PHs, the PIM covariate effects provide an alternative interpretation that may be easier to explain to collaborators. Follmann (2002) showed a similar connection between logistic PIMs and PH models in POR, but POR allows for censoring. Alternative models of the PIM can be obtained by considering the PIM as an expected placement value—i.e. the expected 'place' in the conditional survivor distribution function, $E_{Y*}[S_{Y|X}(Y^*)]$ (Pepe and Cai, 2004; Cai and Dodd, 2008), This interpretation suggests alternative estimating equations that may be more efficient than those developed by Thas and his colleagues.

In general, the choice of and effect of lexicographic ordering on covariates and model coherence is not self-evident. For binary covariates, the use of a strict lexicographical ordering $X < X^*$ results in the Wald-type Mann–Whitney statistic. This is one case in which the lexicographical ordering is clear. However, more guidance and intuition about this in the general setting would be helpful. Equation (31) presents a model for which the authors impose the lexicographical ordering SES $\leqslant$ SES*. Consider a simpler model (and its complement):

$$\text{logit}\{P(\text{MI} \preccurlyeq \text{MI}^*)\} = \beta_1(\text{SES}^* - \text{SES}) + \beta_2 \text{SES}, \tag{35}$$

$$\text{logit}\{P(\text{MI}^* \preccurlyeq \text{MI})\} = \beta_1'(\text{SES} - \text{SES}^*) + \beta_2' \text{SES}^*. \tag{36}$$

It was not clear whether the authors would expect a lack of coherency, in the sense that $m(\text{SES}, \text{SES}^*) \neq 1 - m(\text{SES}^*, \text{SES})$, for $(\text{SES}, \text{SES}^*) = (1, 0)$. However, the fitted PIs for this case demonstrate coherency for all $(\text{SES}, \text{SES}^*)$. Now, consider two different orderings SES* $\leqslant$ SES and no ordering. Within a given ordering, coherency holds, as can be seen in Table 7. Results are presented in terms of β_1 and $\beta_2 - \beta_1$ because they describe the effect on SES and SES* respectively.

The estimates in Table 7 imply different relationships between the covariates and the PI. What is the preferred ordering? It may be SES $\leqslant$ SES* because the PI is defined as $P(\text{MI} \leqslant \text{MI}^*)$ but the motivation should be made more explicit.

I conclude with two final cautionary notes about PIs. First, it is well known that, when receiver operating characteristic curves cross, conclusions about covariate effects on the PI become more complex, as this phenomenon can mask true covariate relationships on $S_{Y|X}$. Graphical procedures displaying receiver operating characteristic curve regression models may provide a complementary tool for diagnosing this phenomenon (Pepe, 2000). Additionally, Hand (1992) cautioned against using the PI for causal effects and provided examples for which the PI would lead to the incorrect conclusion about which of two treatments is better.

Wicher Bergsma (*London School of Economics and Political Science*) **and Marcel Croon, Jacques A. Hagenaars and Andries van der Ark** (*Tilburg University*)

We would like to point out the relationship to Bergsma *et al.* (2009), where probabilistic index models (PIMs) were introduced under the name of *Bradley–Terry-type models*, and full maximum likelihood for fitting and testing with categorical variables was used. Below we also point out possible interpretational problems with certain PIMs, and how to avoid these.

We begin by giving a justification for the use of the probabilistic index. Consider a set of ordinal random variables $\{Y_i, i \in \mathcal{I}\}$ (not necessarily independently or identically distributed). Being ordinal, the Y_i are only

meaningful comparatively, i.e. an individual Y_i has no meaning. However, a set of meaningful sufficient statistics is

$$\{\text{sgn}(Y_i - Y_j) | i \neq j\}.$$

This suggests the use of

$$L_{ij} = E[\text{sgn}(Y_i - Y_j)] = P(Y_i > Y_j) - P(Y_i < Y_j)$$

which is related to the probabilistic index via

$$\text{PI}_{ij} = (1 - L_{ij})/2.$$

Linking to the notation of Thas and his colleagues, write $Y_i = (Y|\mathbf{X} = i)$ so that

$$\text{PI}_{ij} = P(Y < Y^*|\mathbf{X} = i, \mathbf{X}^* = j).$$

We see that models based on the L_{ij} or the PI_{ij} are truly ordinal, in contrast with, for example McCullagh's logistic models and normal threshold models, which assume that ordinal data are realizations of some underlying interval level variable.

It might be tempting to interpret $L_{ij} > 0$ as '$Y_i > Y_j$'. However, a problem is that it is possible that

$$L_{ij} > 0, \quad L_{jk} > 0 \text{ and } L_{ik} > 0$$

so $Y_i > Y_j$, $Y_j > Y_k$ and $Y_k > Y_i$, i.e. the inequality relation is intransitive. For PIM (31) in the paper an intransitive solution arises if $\beta_1 = \beta_2 = \beta_4 = 0$, $\beta_3 > 0$, $\text{SES}_i = \text{SES}_j = \text{SES}_k > 0$, in which case $\text{MI}_i < \text{MI}_j$, $\text{MI}_{jk} < \text{MI}_k$ and $\text{MI}_k < \text{MI}_i$.

Ideally, we would like to be able to interpret L_{ij} as a difference in location of Y_i and Y_j. However, this is not possible in general, since we may have

$$L_{ij} + L_{jk} \neq L_{ik}.$$

However, if the Bradley–Terry-type model

$$L_{ij} = \lambda_i - \lambda_j \tag{37}$$

holds, then

$$L_{ij} + L_{jk} = L_{ik}$$

and the λs can be interpreted as ordinal location parameters for the Ys. A regression model for the ordinal locations λ_i can then be formulated as

$$\lambda_i = \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta}. \tag{38}$$

More generally than model (37) for a link g, we can consider

$$g(L_{ij}) = \lambda_i - \lambda_j. \tag{39}$$

Substitution of equation (38) into equation (39) yields

$$g(L_{ij}) = (\mathbf{X}_i - \mathbf{X}_j)^{\mathrm{T}} \boldsymbol{\beta}$$

which is a subclass of the PIMs that were considered by Thas and his colleagues. Note that, assuming that equation (39) holds, our formulation (38) is easy to interpret and falls within the classical regression framework.

Bergsma *et al.* (2009) considered a very broad class of models, which includes PIMs, and derived multinomial 'maximum' likelihood equations. These equations apply to PIMs for the case that the response variable is categorical. However, the Lagrangian algorithm that was described there (and implemented in Bergsma and Van der Ark (2009)) appears to suffer from numerical problems when covariates are continuous. We wonder how a full likelihood method could be implemented for the continuous case.

Stijn Vansteelandt (*Ghent University and London School of Hygiene and Tropical Medicine*)

I thank the authors for an interesting and stimulating paper. When interest lies in the effect of treatment A (1, treatment; 0, no treatment) on outcome Y, covariate-adjusted probabilistic indices have been suggested to avoid attenuation of the estimated treatment effect (Brumback *et al.*, 2005), to boost its precision

(Brumback *et al.*, 2005) or to adjust for confounding by Thas and his colleagues. I shall reflect on these various suggestions.

I remind the reader that covariate adjustment is a subtle consideration, even when the treatment is randomly assigned. The unadjusted analysis then targets the marginal probabilistic index $P\{Y(0) \preccurlyeq Y^*(1)\}$, which expresses how likely it is, if one picks two random individuals and randomly chooses to treat one (in which case we observe $Y^*(1)$) but not the other (in which case we observe $Y(0)$), for the untreated individual to score lower. The adjusted analysis targets the same comparison, but for two random individuals with the same covariate value L. As one adjusts for increasingly more baseline predictors of the outcome, the covariate-adjusted probabilistic index will tend to move increasingly further from 0.5, and to come increasingly closer to the within-subject comparison $P\{Y(0) \preccurlyeq Y(1)\}$, which expresses how likely it is that a random individual would score lower if untreated than if treated. Although such within-subject comparison may be the statistician's ultimate dream, interpretation as such is always hindered by the fact that one will never know how close the approximation is.

Since covariate adjustment thus changes the treatment effect estimand, it cannot be used to boost its precision and in fact would often inflate its standard error (Robinson and Jewell, 1991). Rather than letting the choice of covariate set change the interpretation, I shall here focus on the marginal probabilistic index $P\{Y(0) \preccurlyeq Y^*(1)\}$. Adjustment for confounding may now alternatively happen by calculating

$$\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} I(A_i = 0, A_j = 1)\, I(Y_i \preccurlyeq Y_j)/\{P(A_i = 0|L_i)\, P(A_j = 1|L_j)\}}{\sum_{i=1}^{n}\sum_{j=1}^{n} I(A_i = 0, A_j = 1)/\{P(A_i = 0|L_i)\, P(A_j = 1|L_j)\}}. \tag{40}$$

This has the advantage that it relies instead on a model for the propensity score $P(A|L)$, which is arguably easier to specify than the dependence of the probabilistic index on the covariate values of two individuals. Using semiparametric efficiency theory under the model defined by the sole restrictions of a propensity score model, more efficient estimators of $P\{Y(0) \preccurlyeq Y^*(1)\}$ can be constructed. Application of a semiparametric efficient estimator would guarantee that the adoption of auxiliary covariate information boosts precision. When the exposure is randomly assigned, the resulting inference would be (asymptotically) distribution free, because of the estimator's reliance on the known randomization probabilities $P(A|L)$ (see for example expression (40) and Zhang *et al.* (2008)), in contrast with inference under (covariate-adjusted) probabilistic index models which does not exploit that knowledge.

The following contributions were received in writing after the meeting.

David Draper (*University of California, Santa Cruz*)

The authors have offered an interesting semiparametric approach to regression modelling based on their probabilistic index (PI). However, I do not see that this technique offers significant gains when compared with existing Bayesian non-parametric fitting methods. Consider, for instance, the authors' example examining the relationship between Beck depression inventory (BDI) improvement and dose of quetiapine in Section 1, which is illustrated in the paper's Fig. 1(a). The authors point out correctly that their PI analysis is superior to a naive linear regression, in two ways: their approach attempts to capture non-linearity in a particular way, and it also attempts to respond to the evident heteroscedasticity. However, Fig. 8 presents the results from fitting a *treed Gaussian process* model (Gramacy and Lee, 2008) to this data set, using the freeware R function `btgp` 'straight out of the box', with no special tuning or other user intervention. This is a Bayesian non-parametric technique that finds an optimal partition of the X-space, for fitting Gaussian process regression models to the separate regions identified by the partition. The treed Gaussian process analysis automatically adapts to the heteroscedasticity and non-linearity in this data set, and in so doing it reveals an important scientific finding that was not discovered with the authors' PI approach: the improvement in BDI is constant in the quetiapine dose up to about 19 g, above which it is approximately linear with a slope of about 0.5 BDI points per gram.

In obtaining these results, I did not instruct `btgp` to find a specified number of partition sets, as defined by the dose variable, or where to locate the 'change-points'; the algorithm correctly deduced that the optimal number of separate Gaussian process models to fit in this case is 2. I say 'correctly' and 'optimal' because—in an analysis not presented here in more detail, because of space limitations—I generated 100 simulated data sets, each with 49 observations, matching the structure of the BDI improvement by dose data set (with one change-point randomly located between 16 and 22 g, a constant relationship to the left of the change-point at a value varying randomly from 2 to 8 BDI improvement points, a linear relationship

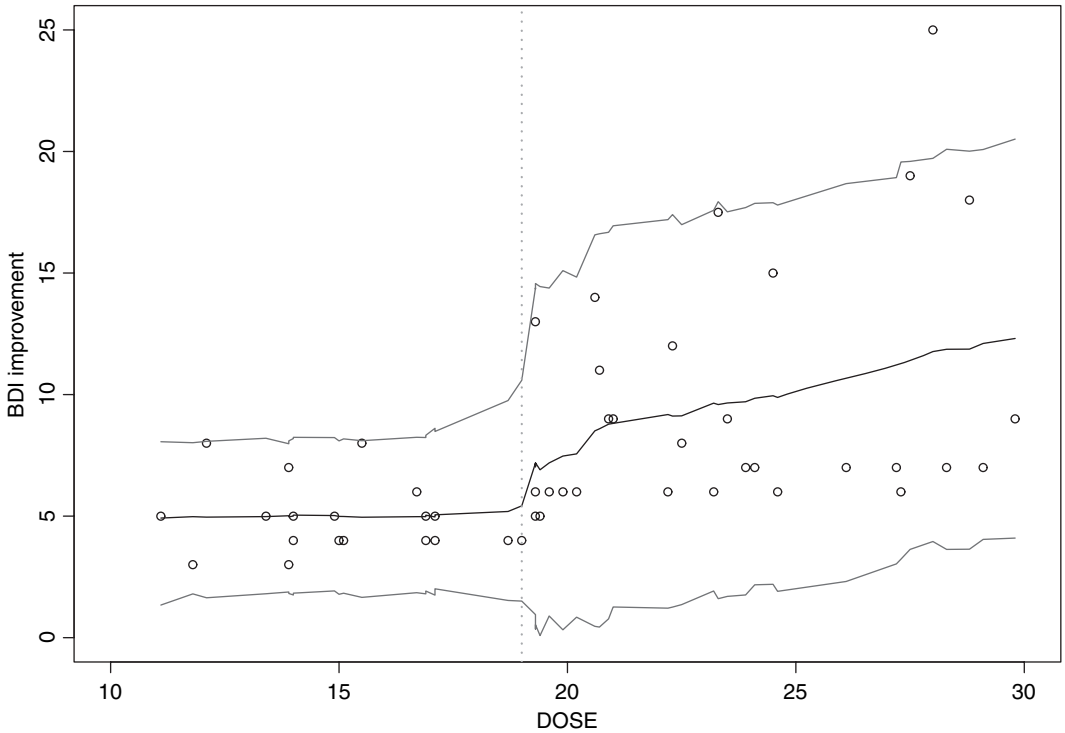

Fig. 8. Bayesian treed Gaussian process, fit to the relationship between DOSE and BDI improvement in Fig. 1: ———, estimated underlying mean function; ═══, 90% uncertainty bands; ⋮, optimal partition

to the right of the break point with a slope varying randomly from 0.25 to 0.75, and heteroscedasticity values chosen randomly from ranges similar to those in the observed data set), and `btgp` identified the correct structure in 93 of these 100 replications.

Michael P. Fay (*National Institute of Allergy and Infectious Diseases, Bethesda*) (© US Government)
Although Thas and his colleagues discussed the k-sample case, it is helpful to compare the probabilistic index models (PIMs) and linear models in the simple three-sample case to point out some non-intuitive behaviour of the PIMs. Let $Y^{(a)}$ be a random response from group a, and the associated covariate be $X^{(a)}$, a 3×1 vector with the ath element equal to 1 and the others 0. The associated linear model has $E(Y|X) = X^{\mathrm{T}}\mu$, where $\mu = (\mu_1, \mu_2, \mu_3)$, and the model imposes no additional structure on the means. For comparing groups a and b in the linear model, we use the difference $E(Y^{(b)} - Y^{(a)}) = \mu_b - \mu_a$. So knowing μ allows us to obtain any pairwise comparison between the groups.

Now consider a PIM for the three-sample case. Let $P_{ab} = P(Y^{(a)} \preccurlyeq Y^{(b)})$ and let $\beta_{ab} = P_{ab} - \frac{1}{2}$. Suppose that our model of P_{ab} is $m(X^{(a)}, X^{(b)}; \beta) = \frac{1}{2} + (X^{(b)} - X^{(a)})\beta = \frac{1}{2} + \beta_b - \beta_a$, where $\beta = (\beta_1, \beta_2, \beta_3)$. For this model, the comparison between group a and b gives $\beta_{ab} = \beta_b - \beta_a$. As with the linear model, knowing β we can model all three pairwise comparisons, β_{12}, β_{23} and β_{13}, and if we know two of the three pairwise comparisons we can obtain the third. Further, since there are only three unique pairs for comparisons, it would appear that three parameters would not impose any additional structure. This is not true, since there are distributions for which the PIM model above does not fit the data.

Consider three discrete distributions, each with three possible values which occur with equal probability. Here are the possible values: group 1 (1,5,9), group 2 (2,6,7) and group 3 (3,4,8). Then $P(Y^{(1)} \preccurlyeq Y^{(2)}) = P(Y^{(2)} \preccurlyeq Y^{(3)}) = P(Y^{(3)} \preccurlyeq Y^{(1)}) = 5/9$ and $\beta_{12} = \beta_{23} = \beta_{31} = 1/18$. This is an example of the intransitivity of the PI (see for example Brown and Hettmansperger (2006)). If we try to fit the model $m(X^{(a)}, X^{(b)}; \beta) = \frac{1}{2} + \beta_b - \beta_a$ to this scenario, then there are no values of the parameter vector β such that $\beta_2 - \beta_1 = \beta_3 - \beta_2 = \beta_1 - \beta_3 = 1/18$. So with these three distributions our model is misspecified.

Asymptotically, with equal numbers in the three groups, it seems that the estimates of β_1, β_2 and β_3 would all approach 0. For large samples, would we erroneously conclude that $P_{12} \approx \frac{1}{2}$ in the presence of the third group, but conclude that $P_{12} \approx 5/9$ if we did not observe the third group? Perhaps diagnostic plots are important even in very simple cases where we might not use them in linear models.

Dean Follmann (*National Institutes of Allergy and Infectious Diseases, Bethesda*) (© US Government)
I very much liked this paper. It gave a thoughtful development of a flexible probabilistic index model (PIM) approach, explored connections with other methods, had nice theoretical results and gave three substantial examples. I also am hopeful that this approach becomes part of an applied statistician's toolbox because I think that there are settings where it will be the perfect choice.

In this comment I wanted to expand on an aspect of this approach that I became painfully aware of when working on a similar method (Follmann, 2002). Under a simple version of a PIM, one postulates that the probability that outcome i is better than j is given by a logistic regression with intercept 0 and covariate $X_i - X_j$. I applied this pairwise logistic approach (PLA) to a clinical trial by using standard software and waited for the result. After a while, I quit waiting as I realized what the hitherto esoteric expression $O(n^2)$ (the order of the number of terms in the PLA likelihood) truly meant for a data set with $n = 4228$. And, even if I were patient, I would have had to wait even longer for the covariance estimate based on $O(n^3)$ operations. Being impatient and needing an example, I decided to analyse a subgroup of 645 diabetics to illustrate the method. Unfortunately, this is not a universal solution to the problem of large n.

If we assume a proportional hazards (PH) model for the outcomes, then the pairwise logistic regression model obtains. The PLA does not imply a PH model, and thus the PH model requires a stronger assumption. But there are tempting reasons to make this assumption. First, we can just run Cox regression software on the data. Under no censoring this should involve $O(n)$ terms for the partial likelihood. Another reason is that, under the PH model, partial likelihood gives more efficient estimates than from the PLA. To crystallize these points, I conducted one small simulation in *R*, for the two-group setting with $n = 20$ and then $n = 200$ per group, $X = 0$ or $X = 1$ the group indicator, exponentially distributed outcomes and no censoring. On the basis of 1000 replications, the ratio of mean-squared errors for the pairwise to partial likelihoods was 1.66 ($n = 20$) and 1.31 ($n = 200$) whereas the ratio of computation times was about 14.3 ($n = 20$) and 1.15×10^4 ($n = 200$). The PH assumption has real advantages and it is not exactly clear what additional flexibility the weaker assumption of the PLA buys us. And the PH model still allows us the nice PIM interpretation of our parameters.

Vanda Inácio (*Lisbon University*), **Miguel de Carvalho** (*Ecole Polytechnique Fédérale de Lausanne and Universidade Nova de Lisboa*) **and Antónia Amaral Turkman** (*Lisbon University*)
We congratulate the authors for this stimulating paper. In the space available, we concentrate on the relationship of the probabilistic index model (PIM) to the normal linear regression model, and its possible extension for the case of functional predictors. Consider the normal functional linear model with functional predictor and scalar response

$$Y = \int_T \alpha(t)\, X(t) + \varepsilon = \langle \alpha, X \rangle + \varepsilon, \qquad \varepsilon \sim N(0, \sigma), \tag{41}$$

where the predictor X and the functional parameter α are square integrable over a compact T. Similarly to what has been shown by the authors, we have

$$P(Y < Y^* | X, X^*) = \Phi\left(\frac{\langle \alpha, X - X^* \rangle}{\sigma\sqrt{2}} \right) = \langle \beta, X - X^* \rangle, \tag{42}$$

where $X - X^* = X(t) - X^*(t)$, for $t \in T$, with $\beta = \alpha/\sigma\sqrt{2}$ being a functional parameter in this context. To estimate the functional PIM in equation (42) we only need to estimate α and σ. Cardot *et al.* (1999) proposed to estimate α on the basis of functional principal components, using the estimator

$$\hat{\alpha} = \sum_{j=1}^{K} \frac{\Delta_n \hat{v}_j}{\hat{\lambda}_j} \hat{v}_j.$$

Here Δ_n is the empirical cross-covariance operator and $\hat{v}_1, \ldots, \hat{v}_K$ are the eigenfunctions associated with the K largest eigenvalues $\hat{\lambda}_1, \ldots, \hat{\lambda}_K$ of the empirical covariance operator of the sample $X_1, \ldots, X_n$. For further details see Cardot *et al.* (1999). Estimation of the PIM in equation (42) is completed after obtaining

$$\hat{\sigma} = \left\{ \frac{\sum_{i=1}^{n} (Y_i - \langle \hat{\alpha}, X_i \rangle)^2}{n - K - 1} \right\}^{1/2}.$$

Recently, Inácio *et al.* (2012) have extended receiver operating characteristic curve regression methodology to the functional context. They investigated how the accuracy of gamma glutamyl transferase, as a diagnostic test to detect metabolic syndrome, is affected by the nocturnal arterial oxygen saturation, which was measured densely over the patient's sleep. It would be interesting to study this relationship by

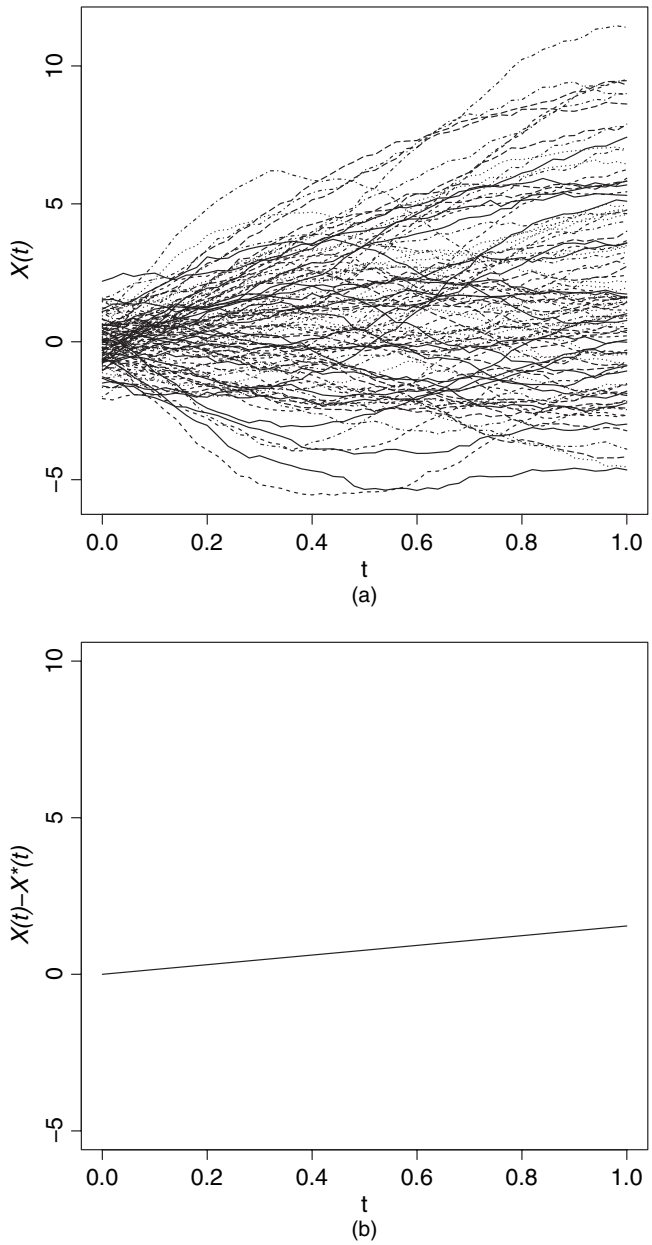

Fig. 9. (a) 100 simulated predictor trajectories and (b) hypothetical difference curve $X(t) - X^*(t)$

means of a (functional) PIM. For example, it would be interesting to use an estimate of the probability in model (42) as an index to compare the gamma glutamyl transferase values of someone with a 'high' curve of arterial oxygen saturation against someone with a 'low' curve of arterial oxygen saturation.

We illustrate our thoughts by means of a numerical experiment, where we simulated 100 independent data sets (sample size 100) according to model (41); Fig. 9(a) gives an idea of the shape of the predictor curves $X(t)$, whereas Fig. 9(b) represents a hypothetical difference curve $X(t) - X^*(t)$. The true probability in model (42) under our simulated scenario is 0.710 and its average estimate (2.5%, 97.5% simulation quantiles) is 0.712 (0.677,0.746).

Tom King and Lara E. Harris (*Southampton University*)

For subjective listening ratings, Wolfe and Firth (2002) showed the need for modelling personal response scales. The ABX listening test remains a popular approach for subjective listening experiments for this reason and other bias problems (Zielinski *et al.*, 2008). This is a type of two-alternative forced choice test that was mentioned by Dodd such that listeners are presented with two excerpts A and B and asked to identify which is X. In a more general version, listeners are asked to identify which of A and B are most similar to X, repeating these tests for multiple iterations of A and B from a finite list of excerpts.

Standard approaches to analysing results test null hypotheses of no audible difference by using exact binomial probabilities (Leventhal, 1986). These also allow for an estimate of the proportion of correct identifications to be made (Burstein, 1989), assuming equal allocation of forced 'don't knows'. Multiple comparisons mean losing power without borrowing strength by using covariates. A density could be estimated by using more advanced methods to estimate a ranking but this would be opaque to many working in audio. Non-parametric methods might be able to test for a preferred ranking but would not afford much insight into the relative support for different rankings, or the influence of covariates.

The probabilistic index model should be ideal for this type of data. The question in this instance is to test preference of bass reproduction through digital simulation of a number of loudspeaker designs. The probabilistic index model should be able to incorporate all the relevant covariates and to estimate specific preferences and to estimate design preferences as well as identifying preference variation. More details are given in Harris *et al.* (2012).

A. J. Lawrance (*University of Warwick, Coventry*)

I enjoyed this paper at the meeting but, in spite of the attractive presentation and a little reflection afterwards, I still have a few points of query. As a person without previous knowledge of the area, it is still not clear to me why a probabilistic index model (PIM) is in general a natural non-parametric regression way to go which stands on its own two feet. I do understand that quite a few well-known methods can be cast in the PIM way and be extended via a PIM, but this does not make it natural. The topic is regression so one would expect to see a connection to the conditional distribution of response given covariates, even if not fully specified. It seems very curious that this appears to be absent, at least on the surface, and even more so that the PIM focuses on the difference distribution of two independent response variables. That seems a very awkward way to relate to the conditional regression distribution. Nor do I know what information is being neglected by a PIM by this formulation. The lack of a connection to the conditional distribution would appear to be the reason why no sort of likelihood is available. Finally, to ride my graphical hobby horse, can I plead for common scales in comparative graphs such as in Fig. 3 and between Figs 6(b) and 6(c)? Discussion at the meeting illustrated high regard for the work and I quite expect the authors to be able to answer all my main points satisfactorily, and I look forward to the revelations.

Chenlei Leng (*National University of Singapore*) **and Guang Cheng** (*Purdue University, West Lafayette*)

We congratulate the authors on developing an interesting class of semiparametric models, i.e. probabilistic index models (PIMs), that directly relates the probabilistic index to the covariates. The construction of a PIM is well motivated by the ordinal response variable. We shall comment on the semiparametric efficiency issues.

Given the pseudo-observations $\{I_{ij} \equiv I(Y_i \preccurlyeq Y_j), \mathbf{Z}_{ij}\}_{(i,j)\in\mathcal{I}_n}$, the PIM is essentially a special case of the semiparametric conditional moment model. The authors thus propose to estimate $\boldsymbol{\beta}$ on the basis of the quasi-likelihood estimating equation (8) in the presence of the nuisance function f_{xy}. For the longitudinal data modelled in the marginal generalized estimating equation framework, i.e. $E(Y_{ij}|\mathbf{X}_{ij}) = g^{-1}(\mathbf{X}_{ij}'\boldsymbol{\beta})$ for $i = 1, \ldots, n$ and $j = 1, \ldots, m_i$, it is not difficult to derive the efficient score function of $\boldsymbol{\beta}$ as

$$\tilde{l}_\beta = \sum_{i=1}^{n} \left(\frac{\partial g^{-1}(\mathbf{X}_i\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right)' \boldsymbol{\Sigma}_i^{-1} \{\mathbf{Y}_i - g^{-1}(\mathbf{X}_i\beta)\}, \tag{43}$$

where $g^{-1}(\mathbf{X}_i\boldsymbol{\beta}) = (g^{-1}(\mathbf{X}'_{i1}\boldsymbol{\beta})), \ldots, g^{-1}(\mathbf{X}'_{im_i}\boldsymbol{\beta}))'$, $\boldsymbol{\Sigma}_i = \mathrm{var}(\mathbf{Y}_i|\mathbf{X}_i)$ and $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{im_i})'$, by only assuming the conditional moment restrictions and bounded m_is. The semiparametric efficiency bound trivially follows from expression (43). However, efficiency bound calculation in this paper is non-trivial owing to the more complicated dependence structure, i.e. sparsely correlated data. We suggest that the authors modify equation (8) as an (approximate) efficient score function, according to Hansen (1985), who considered the efficiency bound under weakly dependent data, which may be solved to obtain the efficient estimate.

We next discuss the efficient estimation of the PIM based on the fact that the PIM can be induced from some marginal regression model with full parametric likelihood, i.e. f_{XY} is known. Here, we focus on a prototypical example (given independent, identically distributed observations):

$$\Lambda(Y) = \mathbf{X}'\boldsymbol{\beta} + \varepsilon, \tag{44}$$

where Λ is an unknown increasing function and ε follows a known distribution F_ε. In this case, we can choose an appropriate error distribution such that $P\{\varepsilon - \varepsilon^* < (\mathbf{X}^* - \mathbf{X})'\boldsymbol{\beta}\} = m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta})$ for the $m(\cdot)$ in the PIM of interest. Linear transformation models (44) have a long history. Bickel and Ritov (1997) proposed an efficient estimation of $\boldsymbol{\beta}$ based on rank statistic methods. Cheng *et al.* (1995) proposed a class of estimating equations for $\boldsymbol{\beta}$ under possibly right-censored observations. Moreover, Han (1987) even allowed F_ε to be unknown and gave the maximum rank correlation estimate

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \sum_{i<j} I(Y_i < Y_j)\, I(\mathbf{X}'_i\boldsymbol{\beta} < \mathbf{X}'_j\boldsymbol{\beta}) \tag{45}$$

that is shown to be asymptotically normal with root n rate. It would be quite interesting to compare the asymptotic variances of the above estimators with that of the PIM estimate theoretically and empirically.

Thomas Lumley (*University of Auckland*)

The authors use the results of Lumley and Hamblett (2003) in their proofs, I believe from my suggestion when I visited Ghent. Since that time, I have found out that related central limit theorem results were proved much earlier in the probability literature where the concept that we called 'sparse correlation' is described in terms of 'graph-structured dependence'. In particular, Baldi and Rinott (1989) gave a bound for the departure from normality of a sum of random variables in terms of the maximal degree of the dependence graph.

Jorge Mateu (*University Jaume I, Castellón*) **and Carlos Diaz-Avalos** (*National University of Mexico, Mexico City*)

The authors present in a clear manner the definition and theoretical issues related to probability index models, and how they can be used to model the effect of covariates, with emphasis on the cases of categorical non-ordered covariates. We were pleased to read the clear review of the subject that they gave and the examples shown in the paper. These are enlightening and motivate the reader to follow the subject further.

We believe that the methods shown in the paper are applicable in the area of spatial analysis. The advent of geographical information systems now makes information about spatial covariates easily available, and for spatial variables of interest, say Z, models of the type $E[Z(\mathbf{u})] = X\beta$ where $\mathbf{u} \in \mathcal{D}$ are becoming common. Testing $P(Z < Z^*|X, X^*)$ by using probability index models is an attractive choice if one is interested in deciding whether, at some set of points $\mathbf{u} \in \mathcal{D}$, a spatial random variable $Z(\mathbf{u})$ is below a prescribed threshold value Z^* representing an upper limit for water quality, for instance. $Z(\mathbf{u})$ may represent a random field, a Markov random field or the intensity function of a spatial point process. Another application could be in testing the significance of spatial covariates. This is an issue of interest in several fields, such as plant ecology. To our knowledge, in the field of spatial point process modelling little has been done regarding significance tests for covariates included in the parametric models for the intensity function. Few references (Rathbun *et al.*, 2004; Waagepetersen, 2007) have considered such problems from the fully parametric point of view but rely on their significance tests in confidence intervals resulting from asymptotic assumptions that may not be realistic in applications, so the power of the tests may be overestimated.

Spatial observations are usually dependent, and the probabilistic index model as presented in the paper is not directly applicable. However, the spatial association may be incorporated either by adding a spatial term in the linear predictor, i.e. defining the pseudo-observations as

$$I(\mathbf{u}) = I\{Z(\mathbf{u}) < Z^*(\mathbf{u})\} = g^{-1}\{(X - X^*)\beta + W(\mathbf{u})\},$$

or by directly extending the sparse correlation structure of the pseudo-observations (Section 3.2) for the case of a Markov random-field approach.

Joseph W. McKean (*Western Michigan University, Kalamazoo*)

Thas and his colleagues have presented an interesting procedure for semiparametric models. Their probability index model (PIM) relates the simple Wilcoxon–Mann–Whitney probability $P(Y < Y^*)$ to a linear function of predictors through a link function. Although, as the authors caution, it is not necessarily a competitor to robust procedures for linear or specified non-linear models, it seems useful for a wide variety of semiparametric models. I confine my remarks to rank-based estimation and a few remarks on pseudonorms.

The authors compare their PIM procedure with several procedures, including rank-based (rank regression) procedures for linear models. These estimates for Wilcoxon scores are obtained by minimizing the dispersion function given in expression (17). This is equivalent to minimizing a pseudo-norm of the residuals as discussed in section 5 of McKean and Schrader (1980); see, also, chapter 3 of Hettmansperger and McKean (2011). Abebe *et al.* (2010) extended these rank-based estimates to a general estimating equation model which was discussed in Liang and Zeger (1986); see, also, section 5.5 of Hettmansperger and McKean (2011) for a sketch of this development. On the basis of their asymptotic theory, as well as empirical studies, these rank-based generalized estimating equation estimates are robust and highly efficient. An appropriate choice of weights results in estimates that are robust in factor space. Also, the theory holds for general scores, so optimal procedures for skewed as well as symmetric error distributions are feasible. Although the asymptotic theory assumes continuous responses, the estimates can be obtained for discrete responses. Hence, a comparison of these rank-based generalized estimating equation estimates with the authors' PIM estimates over continuous and discrete response models should prove interesting.

With regard to the authors' discussion on page 635, for the Wilcoxon scores, the pseudonorm of expression (17) can be written to a constant multiplier as

$$\sum_{i,j} |Y_i - Y_j - (X_i - X_j)\alpha|;$$

see, for instance, section 2.2.2 of Hettmansperger and McKean (2011). Hence, in addition to an asymptotic equivalence between these objective functions, as the authors point out, the equivalence holds for finite n. Note, further, that least squares estimation can be obtained, except for the intercept, by minimizing the squared pseudonorm

$$\sum_{i,j} \{Y_i - Y_j - (X_i - X_j)\alpha\}^2.$$

So least squares estimates are invariant to observations with the same vector of covariates, similar to rank-based and PIM estimates.

I thank the authors for their presentation of the PIM procedure. I look forward to applying it to data sets on which I am consulting and to comparing it with other procedures.

Hannu Oja (*University of Tampere*)

I congratulate the authors for an interesting and inspiring piece of work. It is always good to have new and different tools for statistical inference. What I consider important for further analysis and development of the approach are as follows.

(a) The dependence between Y and $\mathbf{X}$ is described through a function $H : \mathbb{R}^d \times \mathbb{R}^d \to [0, 1]$ such that

$$H_{Y|\mathbf{X}}(\mathbf{X}_1, \mathbf{X}_2) = P(Y_1 < Y_2|\mathbf{X_1}, \mathbf{X}_2) + \tfrac{1}{2} P(Y_1 < Y_2|\mathbf{X}_1, \mathbf{X}_2).$$

It is remarkable that $H_{Y|\mathbf{X}}(\mathbf{X}_1, \mathbf{X}_2) = H_{g(Y)|\mathbf{X}}(\mathbf{X}_1, \mathbf{X}_2)$ for all strictly increasing functions g, and therefore the tests and estimates for unknown $H_{Y|\mathbf{X}}$ should depend on $Y_1, \ldots, Y_n$ only through their ranks $R_1, \ldots, R_n$. To find a realistic parametric model for $H_{Y|\mathbf{X}}(\mathbf{X}_1, \mathbf{X}_2)$ in a practical data analysis is a demanding task indeed.

(b) A natural next step could be to consider triples $(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), (\mathbf{X}_3, Y_3)$ instead of pairs and to define

$$\begin{aligned} H_{Y|\mathbf{X}}(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3) = {} & P(Y_1 < Y_2 < Y_3|\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3) + \tfrac{1}{2} P(Y_1 < Y_2 = Y_3|\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3) \\ & + \tfrac{1}{2} P(Y_1 = Y_2 < Y_3|\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3) + \tfrac{1}{6} P(Y_1 = Y_2 = Y_3|\mathbf{X}_1, \mathbf{X}_2\mathbf{X}_3), \end{aligned}$$

and so on. Finally, the partial likelihood function that is used for Cox's proportional hazard model is, in the continuous case, the probability

$$P(Y_{S_1} < \ldots < Y_{S_n} | \mathbf{X}_1, \ldots, \mathbf{X}_n)$$

where $(S_1, \ldots, S_n)$ are observed inverse ranks, i.e. $R_{S_i} = i$, $i = 1, \ldots, n$.

(c) The estimating equations in expression (8) use variances of I_{ij} but ignore the non-zero covariances between I_{ij} and $I_{i'j'}$. More efficient estimates could be obtained if the whole variance–covariance matrix was used to give weights for $I_{ij} - g^{-1}(\mathbf{Z}_{ij}^{\mathrm{T}}\beta)$. This is what is planned for future research.

(d) I am a little worried about how you define the true β-parameter β_0. In my mind, the true population value to be estimated should depend only on the conditional distribution $f_{Y|\mathbf{X}}$ or the joint distribution $f_{Y,\mathbf{X}}$ or $H_{Y|\mathbf{X}}(\mathbf{X}_1, \mathbf{X}_2)$. It should not be a function of the sequence of design values $(\mathbf{X}_n)$.

I hope that the authors will be interested to develop their approach further.

Emanuel Parzen and Subhadeep Mukhopadhyay (*Texas A&M University, College Station*)
We are inspired by this outstanding paper about the probabilistic index (PI) to discuss an extension, the comparison mid-probability index (CMPI). Our research (extending research by Parzen (1979, 1994, 2004) on non-parametric quantile data modelling) is currently developing (Mukhopadhyay *et al.*, 2011; Parzen and Mukhopadhyay, 2012) comprehensive approaches to the classification–dependence problem: observe continuous or discrete variables (Y dimension 1; X dimension p); model the conditional distribution of Y given X, the dependence between Y and X, and influential subsets of X.

To unify continuous and discrete cases, define the mid-distribution function $F^{\mathrm{mid}}(y; Y) = \Pr(Y < y) + 0.5 \Pr(Y = y)$. Define $\mathrm{CMPI}(Y^*, Y|X) = \mathbb{E}[F^{\mathrm{mid}}(Y^*; Y)|X]$ where Y^* and Y are independent and identically distributed. The authors' PI compares conditional distributions $Y|X$ and $Y^*|X^*$. Our index compares the conditional distribution $Y^*|X^*$ with the unconditional distribution Y. When Y is continuous and X is binary, the CMPI estimates the Wilcoxon statistic; when Y is binary and X continuous, the CMPI estimates $\Pr(Y = 1|X)$.

Step 1: construct (from sample distributions) marginally orthonormal score functions $S_j(Y)$, $S_k(X)$ and $S_1(Y) = \{F^{\mathrm{mid}}(Y; Y) - 0.5\}/\sigma_{\mathrm{mid}}$, with σ^2_{mid} the variance of $F^{\mathrm{mid}}(Y)$. Construct $S_j(Y)$, $j > 1$, by the Gram–Schmidt method from powers of $S_1(Y)$, and discrete Legendre polynomials. For vectors X, k integers k', construct $S_k(X)$ as the product of $S_{k'}(X')$ of each component X' of X.
Step 2: compute score co-moments $\mathrm{LP}(j, k; Y, X) = \mathbb{E}[S_j(Y) S_k(X)]$.
Step 3: compute the non-parametric estimator

$$\mathrm{CMPI}(Y^*, Y|X) = \sum_k S_k(X) E[S_k(X) F^{\mathrm{mid}}(Y^*; Y)].$$

Plot it on a scatter plot $(F^{\mathrm{mid}}(X; X), F^{\mathrm{mid}}(Y; Y))$. Measure the dependence (mutual information) of Y and X non-parametrically by the sum of squares of $\mathrm{LP}(j, k; Y, X)$.
Step 4: the parametric logistic regression model for the CMPI regresses on influential $S_k(X)$ identified from the largest co-moments $\mathrm{LP}(j, k; Y, X)$; choose sufficient statistics before parameters.
Step 5: the copula density function of (Y, X) is non-parametrically estimated by maximum entropy (exponential model) density estimation. The copula density is interpreted as the joint density of $(F^{\mathrm{mid}}(Y; Y), F^{\mathrm{mid}}(X, X))$; $F^{\mathrm{mid}}(X; X)$ has components $F^{\mathrm{mid}}(X'; X')$ marginal mid-distributions.

The authors deserve our appreciation for a path breaking and inspiring paper. Our comments aim to outline additional tools for statisticians' toolbox of modern applied statistics, looking at data as well as modelling them.

Details and graphs can be obtained from `www.stat.tamu.edu/~deep/discussionPIM.pdf`.

Emilio Porcu (*Universidad de Valparaiso*) **and Alessandro Zini** (*University of Milano Bicocca*)
We congratulate the authors for their nice paper: we have the following comments.

(a) The semiparametric class of models proposed enables us to understand better the statistical meaning of both the parameters of a wide range of classic models or class of them (generalized linear models and generalized additive models) and the relationships with both the applied estimators (Mann–Whitney and Wilcoxon–Mann–Whitney) and estimating techniques (quasi-likelihood).

(b) The class of probabilistic index models seems not to be nested with respect to several wide classes of models.

(c) In some situations, it contains models taking into account the heteroscedasticity of the data, in spite of other traditional models.

(d) In other contexts, with respect to a competitor, the probabilistic index model proposed appears equivalent, but the convergence to asymptotic behaviour is faster.
(e) Referring to points (b) and (d), the authors propose, also, a simple new graphical tool, to evaluate the specification or misspecification of a candidate model.
(f) In our opinion, the efficiency of estimators, instead of consistency established by the authors, is a more minor question than specification or misspecification of models, even in small samples.

The authors may want to consider the following points.

(a) A general and crucial point is that of the choice, in the specification of a model, of the discrete (natural) scale with which some phenomenon is subjectively measured: the authors assume, for simplicity, scales on integers, subjectively chosen both by researchers and patients, implicitly claiming 'granularity'. But, who or what guarantees equidistance about the subjective choices? From our perspective, this matter should be taken into account (endogenously) by the model. We sketch here two potential ways.
    (i) When comparing the term $\beta(X^* - X)$ in function m, the following choices may be taken into account:

$$\beta\{(X^*)^\gamma - X^\gamma\},$$

and

$$\beta(X^* - X)^\gamma,$$

for $\gamma \in \mathbf{R}$ where the former choice underlies some Box–Cox transformation. The latter choice, which is coherent with a future point of research in the authors' conclusions about non-linearity with respect to modelling dependence from covariates, may be an interesting alternative for specific problems. Both alternatives pose the problem of equidistance in categories.
    (ii) For a discussion about the choice of the scale, a useful reference may be Zini (2008), where the implications about ordering are discussed, though in the authors' specific context.

Mark A. van de Wiel (*VU University, Amsterdam*)

I congratulate the authors for an excellent paper on this exciting and potentially very useful class of regression models. The authors show the wide applicability of probabilistic index models (PIMs) in various examp1es. Below I address a few issues.

First of all, a philosophical one: PIMs are definitely useful for ordinal responses, in particular because the ordering is then the only meaningful property of the response. To some extent this also holds for (medical) survival data, at least in many settings where modelling of absolute survival is hopeless. However, I believe that the use of PIMs for well-characterized continuous responses is limited. It seems to me that we should use the 'richness of the continuity' for the response and not only its ordering. Of course, this may lead to more complex models (e.g. including heteroscedasticity), but these should give more insight on how the covariate impacts the response than does a PIM.

The 'competition' with parametric models becomes even more important when considering a paired or clustered setting (which the authors briefly mention in Section 7). In an unpaired setting, the power of the PIM-based test relative to that of parametric counterparts is relatively good, because all binomial $\binom{n}{2}$ pairs are used in the PIM statistic. A well-known example is the high asymptotic power of the two-sample Wilcoxon test with respect to a two-sample t-test. However, this relative power drops dramatically in a paired setting when only the ordering within pairs (or clusters) can be used, unless additional distributional assumptions are made.

My final concern is the relatively bad control of the type I error for the asymptotic test in the case of small to moderate sample sizes (Table 4). It should be possible to obtain better small sample results, even when multiple covariates are present. I understand the authors' wish to avoid the bootstrap, but it would have been nice to have these results for their setting. When concentrating on one β, it seems that these could be obtained by assuming asymptotic (joint) normality for the other parameters under the assumption that $\beta = 0$, which defines a sampling model, and then compare $\hat{\beta}$ with its bootstrap counterparts. Alternatively, approximations that use higher moments, such as Edgeworth expansions or saddle point approximations, could be explored.

Wang Zhou (*National University of Singapore*)

It was my pleasure to read this important and interesting paper that proposes probabilistic index models. I shall make two comments.

My first comment is about the inference of the parameter β. In theorem 1, the authors derive the asymptotic normality for their estimators β_n, which satisfies equation (8). However, normal approximations are often too rough to be useful in practice for small to moderate sample sizes. To improve the inference, one may consider using some other techniques. We propose to use the empirical likelihood method.

To make the idea simple, we assume only that the function m is antisymmetric about 1. So equation (8) becomes

$$\mathbf{U}_n(\beta) = \sum_{1 \leqslant i < j \leqslant n} h(\mathbf{X}_i, \mathbf{X}_j) = \mathbf{0},$$

where $h(\mathbf{X}_i, \mathbf{X}_j) = \mathbf{A}(\mathbf{Z}_{ij}; \beta)\{I_{ij} - g^{-1}(\mathbf{Z}_{ij}^{\mathrm{T}}\beta)\} + \mathbf{A}(\mathbf{Z}_{ji}; \beta)\{I_{ji} - g^{-1}(\mathbf{Z}_{ji}^{\mathrm{T}}\beta)\}$. This is a U-structured estimation equation. So we can use the jackknife empirical likelihood (see Jing *et al.* (2009)) for inference on β. To be more specific, let

$$T_n = \binom{n}{2}^{-1} \mathbf{U}_n(\beta),$$

$T_{n-1}^{(-i)} = T(\mathbf{X}_1, \ldots, \mathbf{X}_{i-1}, \mathbf{X}_{i+1}, \ldots, \mathbf{X}_n; \beta)$, the statistic computed on the orginal data set with the ith observation removed. The jackknife pseudovalues

$$\hat{V}_i(\beta) = nT_n - (n-1)T_{n-1}^{(-i)}, \qquad i = 1, \ldots, n,$$

can be shown to be asymptotically independent under mild conditions. Since $T_n = n^{-1}\Sigma_{i=1}^n \hat{V}_i(\beta)$, a standard empirical likelihood ratio can then be constructed on $\hat{V}_i$ as follows:

$$\mathcal{R}(\beta) = \max\left\{ \prod_{i=1}^n n p_i : \sum_{i=1}^n p_i = 1, \sum_{i=1}^n p_i \hat{V}_i(\beta) = 0 \right\}.$$

One can prove that $-2\log\{\mathcal{R}(\beta_0)\} \to^{\mathrm{d}} \chi_p^2$ as $n \to \infty$ under mild conditions, where p is the dimension of β.

My second comment is about Section 4.3, the two-sample problem. At the beginning of Section 4.3, the authors assume that (Y_i, X_i), $i = 1, \ldots, n$, are independent and identically distributed. So n_1 and n_2 should be random. Their MW is different from the classical Mann–Whitney test statistic in which the two sample sizes n_1 and n_2 are fixed.

The **authors** replied later, in writing, as follows.

First we thank the discussants for reading our paper and for taking time to prepare interesting and very insightful comments. After having prepared for writing this rejoinder, we are more than ever aware that the probabilistic index model (PIM) method can be looked at from so many angles that it will take us, and hopefully also many other researchers, quite some time to disentangle its colourful set of flavours. For brevity we cannot respond to all the comments in detail.

We have organized this rejoinder as follows. Instead of replying to each discussant separately, we have tried to arrange our answers by topic. As not all issues could be grouped, at the end we shall briefly give some feedback to particular questions or problems.

Efficiency

Several discussants make suggestions for improving the efficiency of the parameter estimators. Ingrid Van Keilegom proposes to embed a PIM in the transformation model, because efficient estimators in such models have been described. In particular, she defines her model (33) with $h(y) = P(Y^* \geqslant y|\mathbf{X}^*) = 1 - F_{Y|X}(y; \mathbf{X}^*)$. Within the transformation model framework, a PIM is presented as $E_{Y|X}\{h(Y)|\mathbf{X}, \mathbf{X}^*\} = m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta})$. Note that this construction resembles the PIM formulation using expected placement values, as suggested by Lori Dodd. Efficient estimators can be obtained when $h(\cdot)$ is known, but under certain conditions $h(\cdot)$ may be replaced by a consistent estimator. Ingrid Van Keilegom recognizes that this may not be simple in our setting. At this point we refer to Cai and Dodd (2008), who, in developing regression methods for the partial area under the receiver operating characteristic curve came across a similar problem: the conditional distribution function $F_{Y|X}(y; \mathbf{X}^*)$ must be replaced by a consistent estimator in the estimating equation. Although this is feasible under additional smoothness conditions, we deliberately did not want to proceed along this path initially, because we fear that sparseness in the covariate space may obstruct its use in real data settings. In the next section we briefly describe a simpler version of a PIM, for

which the estimation of the nuisance function $h(\cdot)$ becomes easier without having to introduce stringent smoothness conditions.

Marginal probabilistic index model

For brevity, we limit the discussion here to continuous responses in the absence of ties. Consider the model

$$P(Y \preccurlyeq Y^*|\mathbf{X}) = g^{-1}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}), \tag{46}$$

in which the probability refers to the conditional distribution of Y given $\mathbf{X}$ and the marginal distribution of Y^*, i.e. Y^* has distribution function $F_Y(y) = E_X\{F_{Y|X}(y|\mathbf{X})\}$. We refer to model (46) as a marginal PIM. The transformation model with $h(y) = P(Y^* \geqslant y) = 1 - F_Y(y)$ now reduces to the marginal PIM. As before, $h(\cdot)$ is unknown, but now its estimation is straightforward without requiring many additional assumptions. In particular, $\hat{h}(y) = 1 - \hat{F}_Y(y)$ in which $\hat{F}_Y$ is the empirical distribution function of the response variable. It would be interesting to study this model further in the transformation model setting to find semiparametric efficient estimators. Later in our rejoinder we come back to the interpretation of this model, and its relationship to the Kruskal–Wallis rank test. Finally, we also note that this marginal PIM resembles very closely the *comparison mid-probability index* of Emanuel Parzen and Subhadeep Mukhopadhyay. At this point we would like to take the opportunity to thank them for their very stimulating contribution (which has been made available on their Web site) and their deep insights into the non-parametric modelling of the comparison mid-probability index.

We are happy that Stephen Senn reminds us about Brunner *et al.* (2001) which describes rank methods for analysing longitudinal and factorial experiments. The book inspired us in the early days of our PIM research, and we should indeed have referred to this very nice work in the paper. In the simple setting of the K-sample problem, Brunner *et al.* (2001) defined their *relative effect* as $p_i = P(Y_i \leqslant Y^*)$, where Y_i has distribution function F_i and Y^* has the marginal distribution function

$$F_Y(y) = \frac{1}{K}\sum_{i=1}^{K} F_i(y)$$

(assuming equal sample sizes). The relative effect is thus directly related to the marginal PIM of the previous paragraph. Brunner and colleagues further extended their methods to factorial and longitudinal studies. Their focus is on hypothesis testing in which non-parametric estimators of the relative effects play a central role.

K-sample problem and transitivity

We now continue with the K-sample setting to shed some light on the issues that are related to intransitivity raised by Michael Fay and by Wicher Bergsma and colleagues. As in the previous section we use Y_j to denote a response variable with distribution function F_j ($j = 1, \ldots, K$). Consider a PIM with identity link so that ($j < k$)

$$P(Y_j \preccurlyeq Y_k) = \tfrac{1}{2} + \beta_k - \beta_j, \tag{47}$$

with the constraint $\Sigma_{j=1}^{K}\beta_j = 0$. This corresponds to a dummy coding of $\mathbf{X}_j$ and $\mathbf{X}_k$ and with $Z_{jk}^{\mathrm{T}}\boldsymbol{\beta} = (\mathbf{X}_k - \mathbf{X}_j)^{\mathrm{T}}\boldsymbol{\beta}$ as used for most examples in the paper. Let $P_{jk} = P(Y_j \leqslant Y_k) - \frac{1}{2}$. Then, for all $j < k < l$,

$$P_{jk} + P_{kl} = P_{jl}, \tag{48}$$

which expresses the same type of restriction as the one that holds for the L_{jk} of Bergma and colleagues in their parameterization of the Bradley–Terry model. This restriction implies a kind of transitivity. Model (47) can be extended to ($j < k$)

$$P(Y_j \preccurlyeq Y_k) = \tfrac{1}{2} + \beta_k - \beta_l + \beta_{jk}, \tag{49}$$

with the constraints $\Sigma_{j=1}^{K}\beta_j = 0$ and $\Sigma_{j<k}\beta_{jk} = 0$ for all $k = 2, \ldots, K$. With this model equation (48) no longer holds and transitivity is no longer guaranteed. Hypothesis tests for testing that all β_{jk} are 0 may be used for testing for transitivity. Whereas transitivity is often a desired property, or at least it is a convenient characteristic that, for example, always holds in location–shift models, there are settings in which it is not guaranteed or in which detecting intransitivity is even of interest. Several discussants (Lori Dodd, Wicher Bergsma, Tom King and Dean Follmann) refer to examples of studies in which the response data come immediately in the form of pairwise orderings (pseudo-observations) that do not necessarily satisfy transitivity. This suggests that PIMs may also be useful for this type of application. It is also worth noting that

some study designs ensure that the pairwise orderings (pseudo-observations) are mutually independent so that sparse correlation is no longer an issue.

In Section 4.3 of our paper we demonstrated how the Wilcoxon–Mann–Whitney statistic is related to the PIM parameter estimators in the two-sample problem. We now briefly show how the transitive PIM (47) and the marginal PIM relate to the Kruskal–Wallis statistic. First note that the transitive PIM (47) has $K-1$ independent parameters, and the intransitive PIM (49) has $\frac{1}{2}K(K-1)$ independent parameters. For both PIMs the marginal PIM becomes

$$P(Y \preccurlyeq Y_j) = \frac{1}{K}\sum_{k=1}^{K} P(Y_k \preccurlyeq Y_j) = \tfrac{1}{2} + \beta_j.$$

This demonstrates that the marginal PIM cannot distinguish between the two PIMs. We further expect that the Kruskal–Wallis test statistic is asymptotically equivalent to (up to a proportionality factor) $\Sigma_{j=1}^{K}\hat{\beta}_j^2$, with $\hat{\beta}_j$ the estimator of β_j in equation (47) (equivalent to the non-parametric estimator of $P(Y \preccurlyeq Y_j) - \frac{1}{2}$).

Model (47) is also used by Michael Fay for illustrating that a PIM sometimes may be misspecified. We agree with him, but we remark that his example only demonstrates that sometimes transitivity does not hold. The saturated model (49) will fit his data. Goodness-of-fit methods may also be used for assessing the quality of the fit. Recently we (De Neve *et al.*, 2012) have developed a new method for assessing the fit of a PIM.

Model formulation

In connection with the usefulness of goodness-of-fit methods, we also refer to the proposal of Emilio Porcu and Alessandro Zini. They suggest considering terms $\beta\{(X^*)^\gamma - X^\gamma\}$ or $\beta(X^* - X)^\gamma$ in the PIM. Again a model assessment will be required to evaluate the adequacy of the model.

Tony Lawrance wonders about the unconventional way in which PIMs refer to the conditional response distribution and about the fact that PIMs require two independent response variables. Equation (3) in the paper shows the most explicit connection between a PIM and the conditional response distributions. One could also think about defining classical linear regression models in terms of the conditional distribution of the difference $Y - Y^*$. In particular, consider

$$E(Y^* - Y|\mathbf{X}, \mathbf{X}^*) = (\mathbf{X}^* - \mathbf{X})^{\mathrm{T}}\boldsymbol{\alpha}.$$

With this model specification, the natural least squares criterion would be the squared pseudonorm that is provided by Joe McKean in his contribution. Joe McKean gives also the corresponding L_1-pseudo-norm, and he concludes that this demonstrates that least squares, rank-based and PIM estimates share the property that observations with the same covariate patterns do not contribute to the estimate.

The relationship between area under the curve regression and PIMs has been explained in the paper. Thomas Gerds suggests the use of PIMs to model the concordance index, which generalizes the area under the curve for assessing the discrimination ability of prediction models. His *concordance index model* is basically a PIM and thus the PIM machinery may be used, for example, to test whether a biomarker further improves the predictive ability of a prediction model. Another formulation of a concordance index model may be

$$P(T_i \preccurlyeq T_j|R_i, R_j, X_i, X_j) = g^{-1}\{\beta_1 I(X_i = X_j) + \beta_2 I(X_i \neq X_j)\},$$

only defined for $\chi = \{(R_i, X_j, R_j, X_j)|R_i > R_j\}$. In this way

$$g^{-1}(\beta_1) = P(T_i \preccurlyeq T_j|R_i > R_j, X_i = X_j),$$
$$g^{-1}(\beta_2) = P(T_i \preccurlyeq T_j|R_i > R_j, X_i \neq X_j),$$

and the null hypothesis of interest is $H_0: \beta_1 = \beta_2$.

Finally, we remark that Lori Dodd's models (35) and (36) with her lexicographical orderings SES $\leqslant$ SES* and SES* $\leqslant$ SES do not agree with our model formulation of equation (31). We explicitly restricted the PIM to a strict lexicographical ordering SES $<$ SES* so that we do not run into the problem that she encountered.

Computation

We agree with Dean Follmann, who has experience with modelling pseudo-observations (Follmann, 2002), that the estimation procedure may be computationally demanding. We have some experience with large data sets for which we implemented an approximation that seems to work well. The procedure goes as

follows. First we randomly split the data set of size n into s disjoint subsamples of size $m = n/s$ ($i = 1, \ldots, s$). With each subsample we fit a PIM, resulting in the parameter estimates $\hat{\boldsymbol{\beta}}_i$ and covariance matrix estimates $\hat{\boldsymbol{\Sigma}}_i$ ($i = 1, \ldots, s$). Finally we combine the estimates into

$$\tilde{\boldsymbol{\beta}} = \frac{1}{s} \sum_{i=1}^{s} \hat{\boldsymbol{\beta}}_i$$

and

$$\tilde{\boldsymbol{\Sigma}} = \frac{1}{s^2} \sum_{i=1}^{s} \hat{\boldsymbol{\Sigma}}_i.$$

We can demonstrate that these estimators are asymptotically equivalent to the original estimators (with fixed $s < \infty$). The order of the computation time for estimating $\boldsymbol{\beta}$ reduces from $O(n^2)$ to $O(n^2/s)$, i.e. a reduction by a factor s.

The method that is described in the previous paragraph may also turn out to be useful when further research focuses on using computationally intensive methods for inference in PIMs. For example, Mark van de Wiel and Wang Zhou suggest adopting a bootstrap or an empirical likelihood procedure.

Estimating equations

Hannu Oja suggests adopting estimation equations to account for the pairwise non-zero covariances between the pseudo-observations I_{ij} and I_{kl}. A possible way forward is to use pseudolikelihood by constructing the product of all bivariate distributions of two pseudo-observations. We are currently exploring this path for PIMs for clustered data (a collaboration with Stijn Vansteelandt and Fanghong Zhang from Ghent University).

Inspired by the relationship between the PIM and Cox proportional hazard models, Thomas Gerds suggests further developing the PIM framework by allowing for censored data. He proposes two strategies. The first involves inverse probability weighting and the second makes use of pseudo-pseudovalues. Here we mention only that inverse probability weighting has already been described by Cheng *et al.* (1995) in a class of semiparametric linear transformation models that generalize Cox proportional hazard models that make use of estimating equations similar to ours. In passing we note that the discussants Chenlei Leng and Guang Cheng made this observation too. Given the importance of missing and censored data we would very much welcome further research along the lines suggested by Thomas Gerds.

Starting from the same linear transformation model as Cheng *et al.* (1995) do, Chenglei Leng and Guang Cheng suggest going even one step further by not having to specify the distribution function of the additive error term in the transformation model; this would result in a maximum rank correlation estimator as in Han (1987). We have two remarks. First, with only a single covariate X, there will be no unique maximum rank correlation estimator, because the order relation restriction on the covariates will make $I(X_i\beta < X_j\beta) = 1$ for all positive β and $I(X_i\beta < X_j\beta) = 0$ for all negative β (or the other way around). Perhaps his problem disappears when $\mathbf{X}$ contains multiple regressors. Finally, on using the relationship between the error distribution and the link function, maximum rank correlation estimates may also be advertised as an appropriate method for situations in which the link function is left unspecified.

Stijn Vansteelandt argues that covariate adjustment may have several disadvantages when the primary focus is on the probabilistic index as the effect size of a treatment. For example, the interpretation of the treatment effect changes with covariates selected in the model, and the variance of the treatment effect parameter estimator may be inflated by adding covariates. As an alternative solution, he proposes to adjust for confounding by changing the estimating equation of the marginal probabilistic index by incorporating the propensity score. This approach seems to have many advantages for comparative studies, and we sincerely hope that this method will be further developed.

Unstructured responses

Wang Zhou argues that our two-sample setting (Section 4.3) is different from the classical setting in the sense that we allow the sample sizes n_1 and n_2 to be random. We understand the misunderstanding. We should actually have added that the X_i are subject to the restriction $\Sigma_{i=1}^{n} X_i = n_2$.

Jorge Mateu and Carlos Diaz-Avalos, and Vanda Inácio and colleagues suggest extensions that are related to functional data analysis. We welcome their suggestions and we encourage further research in this important area. We also thank Thomas Lumley for pointing us to a reference that describes an asymptotic theory that sheds a different light on the concept of sparse correlation. In the interest of further extending and generalizing the PIM method to more complicated data structures, we believe that it will be necessary to find a more general asymptotic theory that can deal with the type of weak dependences that we encounter.

Throughout the paper we have stressed several times that we consider the PIM to be a valuable additional tool in the statistician's toolbox. When, however, the scientific focus is on the mean response, other regression techniques are favourable. This is, for example, illustrated by David Draper, who reanalysed the Beck depression inventory data with a *treed Gaussian process* model. Another method for an informative analysis of this data set is semiparametric quantile regression (Koenker, 2005).

Finally we turn to Stephen Senn. We understand his concerns related to the use of the probabilistic index as an effect size measure. His criticism applies, however, to most of the methods that focus on the probabilistic index. We hope that further research can solve the issues that he raises.

References in the discussion

Abebe, A., McKean, J. W. and Kloke, J. D. (2010) Iterated reweighted rank-based estimates for GEE models. *Technical Report*. Western Michigan University, Kalamazoo.

Andersen, P., Klein, J. and Rosthøj, S. (2003) Generalised linear models for correlated pseudo-observations, with applications to multi-state models. *Biometrika*, **90**, 15–27.

Baldi, P. and Rinott, Y. (1989) On Normal approximations in terms of dependency graphs. *Ann. Probab.*, **17**, 1646–1650.

Bergsma, W., Croon, M. and Hagenaars, J. A. (2009) *Marginal Models for Dependent, Clustered and Longitudinal Categorical Data*. New York: Springer.

Bergsma, W. and Van der Ark, A. (2009) CMM: categorical marginal models. *R Package*.

Bickel, P. J. and Ritov, Y. (1997) LAN for ranks in transformation models. In *Festschrift for Lucien Le Cam* (eds D. Pollard, E. Torgersen and G. Yang). New York: Springer.

Brown, B. M. and Hettmansperger, T. P. (2002) Kruskal-Wallis, multiple comparisons and Efron dice. *Aust. New Zeal. J. Statist.*, **44**, 427–438.

Brumback, L., Pepe, M. and Alonzo, T. (2006) Using the ROC curve for gauging treatment effect in clinical trials. *Statist. Med.*, **25**, 575–590.

Brunner, E., Domhof, S. and Langer, F. (2001) *Nonparametric Analysis of Longitudinal Data in Factorial Experiments*. New York: Wiley.

Burstein, H. (1989) Transformed binomial confidence limits for listening tests. *J. Audio Engng Soc.*, **37**, 363.

Buyse, M. (2010) Generalized pairwise comparisons of prioritized outcomes in the two-sample problem. *Statist. Med.*, **29**, 3245–3257.

Cai, T. and Dodd, L. (2008) Regression analysis for the partial area under the ROC curve. *Statist. Sin.*, **18**, 817–836.

Cardot, H., Ferraty, F. and Sarda, P. (1999) Functional linear model. *Statist. Probab. Lett.*, **45**, 11–22.

Carroll, R. J. and Ruppert, D. (1988) *Transformation and Weighting in Regression*. New York: Chapman and Hall.

Chen, X., Linton, O. and Van Keilegom, I. (2003) Estimation of semiparametric models when the criterion function is not smooth. *Econometrica*, **71**, 1591–1608.

Cheng, S. C., Wei, L. J. and Ying, Z. (1995) Analysis of transformation models with censored data. *Biometrika*, **92**, 835–845.

De Neve, J., Thas, O. and Ottoy, J. P. (2012) Goodness-of-fit methods for probabilistic index models. *Communs Statist. Theor. Meth.*, to be published.

Follmann, D. A. (2002) Regression analysis based on pairwise ordering of patients' clinical histories. *Statist. Med.*, **21**, 3353–3367.

Ford, I., Norrie, J. and Ahmadi, S. (1995) Model inconsistency, illustrated by the Cox proportional hazards model. *Statist. Med.*, **14**, 735–746.

Gail, M. H., Wiand, S. and Piantadosi, S. (1984) Biased estimates of treatment effects in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika*, **71**, 431–444.

Gneiting, T. and Raftery, A. (2007) Strictly proper scoring rules, prediction, and estimation. *J. Am. Statist. Ass.*, **102**, 359–378.

Gramacy, R. B. and Lee, H. K. H. (2008) Bayesian treed Gaussian process models, with an application to computer modeling. *J. Am. Statist. Ass.*, **103**, 1119–1130.

Graw, F., Gerds, T. A. and Schumacher, M. (2009) On pseudo-values for regression analysis in competing risks models. *Liftim. Data Anal.*, **15**, 241–255.

Green, D. M. and Swets, J. A. (1966) *Signal Detection Theory and Psychophysics*. New York: Wiley.

Grigoletto, M. and Akritas, M. G. (1999) Analysis of covariance with incomplete data via semiparametric model transformations. *Biometrics*, **55**, 1177–1187.

Han, A. K. (1987) Non-parametric analysis of a generalized regression model. *J. Econmetr.*, **35**, 303–316.

Hand, D. J. (1992) On comparing two treatments. *Am. Statistn*, **46**, 190–192.

Hansen, L. P. (1985) A method for calculating bounds on the asymptotic covariance matrices of generalized method of moments estimators. *J. Econmetr.*, **30**, 203–238.

Harrell, F. E., Lee, K. L. and Mark, D. B. (1996) Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statist. Med.*, **15**, 361–387.

Harris, L. E., Holland, K. R. and King, T. (2012) Use of a two-alternative-forced-choice method in listening experiments: procedure and analysis. Submitted to *J. Acoust. Soc. Am*.
Hettmansperger, T. P. and McKean, J. W. (2011) *Robust Nonparametric Statistical Methods*, 2nd edn. Boca Raton: Chapman and Hall.
Inácio, V., González-Manteiga, W., Febrero-Bande, M., Gude, F., Alonzo, T. A. and Cadarso-Suárez, C. (2012) Extending induced ROC methodology to the functional context. *Biostatistics*, to be published, doi 10.1093/biostatistics/kxs007.
Jing, B., Yuan, J. and Zhou, W. (2009) Jackknife empirical likelihood. *J. Am. Statist. Ass.*, **104**, 1224–1232.
Kalbfleisch, J. D. and Prentice, R. L. (1980) *The Statistical Analysis of Failure Time Data*. New York: Wiley.
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.
Lee, Y. and Nelder, J. A. (2004) Conditional and marginal models: another view. *Statist. Sci.*, **19**, 219–228.
Leventhal, L. (1986) Type 1 and Type 2 errors in the statistical analysis of listening tests. *J. Audio Engng Soc.*, **34**, 437–453.
Liang, K. and Zeger, S. (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.
Lumley, T. and Hamblett, N. (2003) Asymptotics for marginal generalized linear models with sparse correlations. *Technical Report 207*. University of Washington, Seattle.
McKean, J. W. and Schrader, R. M. (1980) The geometry of robust procedures in linear models. *J. R. Statist. Soc.* B, **42**, 366–371.
Mukhopadhyay, S., Parzen, E. and Lahiri, S. N. (2011) A unifying quantile based framework for classification: Bayes rule, copula, comparison density and the fundamental question. *Preprint*. Texas A&M University, College Station.
Parzen, E. (1979) Nonparametric statistical data modeling. *J. Am. Statist. Ass.*, **74**, 105–131.
Parzen, E. (1994) From comparison density to two sample data analysis. In *The Frontiers of Statistical Modeling: an Informational Approach* (ed. H. Bozdogan). Amsterdam: Kluwer.
Parzen, E. (2004) Quantile probability and statistical data modeling. *Statist. Sci.*, **19**, 652–662.
Parzen, E. and Mukhopadhyay, S. (2012) Modeling, dependence, classification, united statistical science, many cultures. *Preprint arXiv:1204.4699*.
Pepe. M. S. (2000) An interpretation for the ROC curve and inference using GLM procedures. *Biometrics*, **52**, 352–359.
Pepe, M. S. and Cai, T. (2004) The analysis of placement values for evaluating discriminatory measures. *Biometrics*, **60**, 528–535.
Rathbun, S. L., Shiffman, S. and Gwaltney, C. J. (2007) Modelling the effects of partially observed covariates on Poisson process intensity. *Biometrika*, **94**, 153–165.
Robinson, L. D. and Jewell, N. P. (1991) Some surprising results about covariate adjustment in logistic regression models. *Int. Statist. Rev.*, **58**, 227–240.
Senn, S. J. (1997) Testing for individual and population equivalence based on the proportion of similar responses. *Statist. Med.*, **16**, 1303–1306.
Senn, S. J. (2004) Conditional and marginal models: another view—comments and rejoinders. *Statist. Sci.*, **19**, 228–238.
Senn, S. J. (2006) Probabilistic index: an intuitive non-parametric approach to measuring the size of the treatment effects by L. Acion, J. J. Peterson, S. Temple and S. Arndt. *Statist. Med.*, **25**, 3944–3948.
Senn, S. (2011) U is for Unease: reasons to mistrust overlap measures in clinical trials. *Statist. Biopharm. Res.*, **3**, 302–309.
Waagepetersen, R. (2007) Estimating functions for inhomogeneous spatial point processes with incomplete covariate data. *Biometrika*, **95**, 351–363.
Wolfe, R. and Firth, D. (2002) Modelling subjective use of an ordinal response scale in a many period crossover experiment. *Statist. Appl.*, **51**, 245–255.
Zhang, M., Tsiatis, A. A. and Davidian, M. (2008) Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics*, **64**, 707–715.
Zielinski, S., Rumsey, F. and Bech, S. (2008) On some biases encountered in modern audio quality listening test—a review. *J. Audio Engng Soc.*, **56**, 427–451.
Zini, A. (2008) A note on the bipolar mean: is it a single mean? *Statist. Appl.*, **6**, no. 1.