Discussion of "Probabilistic Index Models," by Thas, O., de Neve, J., Lieven, C. and Ottoy., J.-P.

Citation for published version:
Inácio, V, de Carvalho, M & Turkman, MA 2012, 'Discussion of "Probabilistic Index Models," by Thas, O., de Neve, J., Lieven, C. and Ottoy., J.-P.' Journal of the Royal Statistical Society: Series B, vol. 74, pp. 659-661. DOI: 10.1111/j.1467-9868.2011.01020.x

Digital Object Identifier (DOI):
10.1111/j.1467-9868.2011.01020.x

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Journal of the Royal Statistical Society: Series B

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

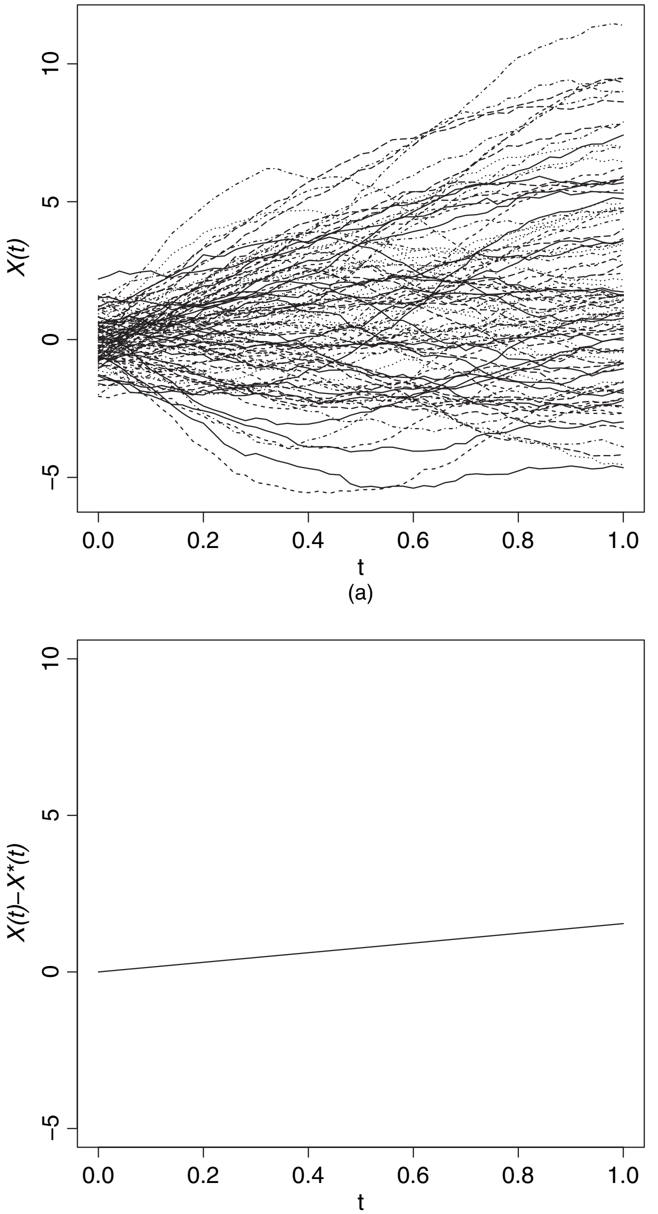

Fig. 9. (a) 100 simulated predictor trajectories and (b) hypothetical difference curve $X(t) - X^*(t)$

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

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