Statistica Sinica **33** (2023), 1-22
doi:https://doi.org/10.5705/ss.202021.0070

SEMIPARAMETRIC ESTIMATION OF PROBABILISTIC INDEX MODELS: EFFICIENCY AND BIAS

Karel Vermeulen, Jan De Neve, Gustavo Amorim,
Olivier Thas and Stijn Vansteeland

*Ghent University, Vanderbilt University, Hasselt University,
London School of Hygiene and Tropical Medicine*

Abstract: Many well-known rank tests can be viewed as score tests under probabilistic index models (PIMs), that is, regression models for the conditional probability that the outcome of one randomly chosen subject exceeds the outcome of another independently chosen subject. PIMs provide a natural regression framework for nonparametric rank tests. In addition, PIMs supplement these tests with effect sizes and ease the development of more flexible tests, such as tests that allow for covariate adjustment. Inferences for PIMs are currently based on an estimator, referred to as the standard estimator, that is derived heuristically. By appealing to semiparametric theory and a Hoeffding decomposition, we rigorously derive the class of all consistent and asymptotically normal estimators for the parameters indexing a PIM. We identify the (locally) semiparametric efficient estimator in this class, and derive a second estimator with a smaller second-order finite-sample bias. The properties of the estimators are evaluated theoretically and empirically. The heuristic standard estimator turns out to be the preferred estimator in practice, because it is computationally superior to both the efficient and the bias-reduced estimators, and only suffers from a minor loss in efficiency. We also propose a partition strategy to further improve the computational performance of the standard estimator.

Key words and phrases: Cross correlation, influence function, second-order bias, semiparametric estimation, U-process.

1. Introduction

Probabilistic index models (PIMs, Thas et al. (2012)) form a class of semiparametric models for the probability that the outcome of one randomly chosen subject exceeds the outcome of another independently chosen subject, as a function of covariates. Let $\{\mathbf{Z}_i^T = (Y_i, \mathbf{X}_i^T) : i = 1, \ldots, n\}$ denote a sample of n independent and identically distributed (i.i.d.) random vectors, where Y_i denotes

---

Corresponding author: Jan De Neve, Department of Data Analysis, Ghent University, Ghent, Belgium. E-mail: Jan.DeNeve@UGent.be.