JDS

# Identification and Semiparametric Efficiency Theory of Nonignorable Missing Data With a Shadow Variable

WANG MIAO, Department of Probability and Statistics, Peking University, China
LAN LIU, School of Statistics, University of Minnesota at Twin Cites, USA
YILIN LI, Department of Probability and Statistics, Peking University, China
ERIC J. TCHETGEN TCHETGEN, Department of Statistics, The Wharton School of the University of Pennsylvania, USA
ZHI GENG, School of Mathematics and Statistics, Beijing Technology and Business University, China

**PROBLEM STATEMENT**

Missingness not at random (MNAR) arises in many empirical studies in biomedical, socioeconomic, and epidemiological researches. A fundamental problem of MNAR is the *identification* problem, that is, the parameter of interest may not be uniquely determined with observed data. Besides, statistical inference is challenging under MNAR without identification.

**METHODS**

This paper studies an identification strategy based on a so-called shadow variable. A shadow variable is assumed to be correlated with the outcome, but independent of the missingness process conditional on the outcome and fully observed covariates. A general condition for nonparametric identification of the full data law under MNAR using a valid shadow variable is provided. The corresponding semiparametric efficiency bound for the class of regular and asymptotically linear estimators is established. A doubly robust and locally efficient estimation method is proposed, evaluated on both simulation data, and applied to a real data example about home pricing.

**RESULTS**

The proposed identification condition is satisfied by many commonly-used models; moreover, it is imposed on the complete cases, and therefore has testable implications with observed data only. The closed form for the efficient influence function is obtained, which motivates a doubly robust and locally efficient estimator. The estimator remains consistent even if certain working model is misspecified and attains the semiparametric efficiency bound if all working models are correct.

**SIGNIFICANCE**

The paper describes the largest class of nonparametric models that are identifiable by the shadow variable approach, and establishes the semiparametric theory for this model. A novel doubly robust and locally efficient approach for the analysis of nonignorable missing data with a shadow variable is provided.

Research Article

Code and data links: https://github.com/H4O2/MNAR-shadow-variable

Keywords: Doubly robust estimation, efficient influence function, identification, missing not at random, shadow variable

Authors' addresses: Wang Miao, mwfy@pku.edu.cn, Department of Probability and Statistics, Peking University, China; Lan Liu, liux3771@umn.edu, School of Statistics, University of Minnesota at Twin Cites, USA; Yilin Li, yilinli@pku.edu.cn, Department of Probability and Statistics, Peking University, China; Eric J. Tchetgen Tchetgen, ett@wharton.upenn.edu, Department of Statistics, The Wharton School of the University of Pennsylvania, USA; Zhi Geng, zhigeng@btbu.edu.cn, School of Mathematics and Statistics, Beijing Technology and Business University, China.