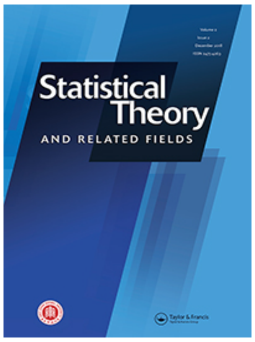

Statistical Theory and Related Fields

ISSN: 2475-4269 (Print) 2475-4277 (Online) Journal homepage: https://www.tandfonline.com/loi/tstf20

Statistical methods without estimating the missingness mechanism: a discussion of ‘statistical inference for nonignorable missing data problems: a selective review’ by Niansheng Tang and Yuanyuan Ju

Jiwei Zhao

STATISTICAL THEORY AND RELATED FIELDS
2018, VOL. 2, NO. 2, 143–145
https://doi.org/10.1080/24754269.2018.1522576

Taylor & Francis
Taylor & Francis Group

SHORT COMMUNICATION

Statistical methods without estimating the missingness mechanism: a discussion of 'statistical inference for nonignorable missing data problems: a selective review' by Niansheng Tang and Yuanyuan Ju

Jiwei Zhao

Department of Biostatistics, State University of New York at Buffalo, Buffalo, NY, USA

ARTICLE HISTORY Received 4 September 2018; Accepted 9 September 2018

First of all, I wholeheartedly congratulate Tang and Ju (referred to as TJ hereafter) on a well-written comprehensive review paper that surveys cutting-edge statistical theory and methodology relevant to estimation, influence analysis and model selection in regression models with missing data.

TJ begins their presentation from the missing data mechanism, a fundamental concept in the missing data literature (Kim and Shao, 2013; Little and Rubin, 2002; Molenberghs et al., 2014; Tsiatis, 2006). In their Section 2, TJ presents a detailed explanation of this definition and underlines its importance to developing downstream statistical methodology. To facilitate this discussion, I adopt the same notation as follows. Consider a regression model where Y is a response variable and $\mathbf{X}$ is a p-dimensional explanatory variable, and $\{(\mathbf{x}_i, y_i) : i = 1, \ldots, n\}$ are n independent and identically distributed realisations of $(\mathbf{X}, Y)$. Assume $\mathbf{X}$ is always fully observed but Y is subject to missingness. Let δ be the missing data indicator for Y, that is, $\delta = 0$ if Y is missing, and $\delta = 1$ otherwise. Then the missing data mechanism is the conditional distribution of δ given $\mathbf{X}$ and Y, i.e.

$$\pi(\mathbf{x}, y) = \mathrm{pr}(\delta = 1 \mid \mathbf{x}, y). \tag{1}$$

One intrinsic complication of the missing data mechanism is that, only except for a few scenarios (d'Haultfoeuille, 2010; Little, 1988), its underlying truth is difficult to verify. The reason due to its plausible dependence on Y, an incompletely observed variable. This issue pronounces more clearly when one moves forward to real application, where the investigators would be more satisfied if a statistical method could make the assumption of the mechanism less stringently so that it is able to be flexibly applied to various scenarios.

My discussion, motivated by the need of developing versatile statistical procedures that would provide robust protection to certain mechanism misspecification, showcases the up-to-date statistical treatments where the mechanism model assumption is only imposed at a minimum level. The discussion concentrates on brief introduction of two types of these assumptions and spans diverse statistical topics including model identification, point estimation, hypothesis testing and high dimensional variable selection.

One distinct feature of the methods in this discussion is that the mechanism model would be treated as a nuisance, hence all the methods could be carried out without the need of estimating the mechanism.

1. Mechanism based on conditional independence

The instrumental variable is a well-studied method in econometrics, epidemiology and related disciplines. The key step of applying this method is certain requirement about the conditional independence among variables. Zhao and Shao (2015) proposed to take advantage of the nonresponse instrument $\mathbf{Z}$, a component of $\mathbf{X}$, to analyse missing data, especially nonignorable missing data. The concept of nonresponse instrument shares the similar spirit to the instrumental variable. To be more specific, Zhao and Shao (2015) assumed that

$$\mathrm{pr}(\delta = 1 \mid \mathbf{x}, y) = \mathrm{pr}(\delta = 1 \mid \mathbf{u}, y), \tag{2}$$

where $\mathbf{x} = (\mathbf{u}^{\mathrm{T}}, \mathbf{z}^{\mathrm{T}})^{\mathrm{T}}$. Some further requirement, e.g. $p(y \mid \mathbf{x}) \neq p(y \mid \mathbf{u})$, is also needed for model identification purpose.

When $\mathbf{X}$ by itself serves as the nonresponse instrument, Tang, Little, and Raghunathan (2003) studied this special situation and proposed to estimate the unknown

CONTACT Jiwei Zhao ✉ zhaoj@buffalo.edu 🖃 Department of Biostatistics, State University of New York at Buffalo, Buffalo, NY, USA

parameter $\boldsymbol{\theta}$ in $p(y \mid \mathbf{x}) = p(y \mid \mathbf{x}; \boldsymbol{\theta})$ through the conditional likelihood of $p(\mathbf{x} \mid y, \delta = 1)$:

$$\prod_{i=1}^{n} \left\{ \frac{p(y_i \mid \mathbf{x}_i; \boldsymbol{\theta})}{\int p(y_i \mid \mathbf{x}; \boldsymbol{\theta}) g(\mathbf{x}) d\mathbf{x}} \right\}^{\delta_i = 1},$$

where $g(\mathbf{x})$ represents the unspecified probability density function of $\mathbf{X}$. Then the objective becomes to a semiparametric function:

$$l(\boldsymbol{\theta}, g) = \sum_{i=1}^{n} r_i \left\{ \log p(y_i \mid \mathbf{x}_i; \boldsymbol{\theta}) - \log \int p(y_i \mid \mathbf{x}; \boldsymbol{\theta}) g(\mathbf{x}) \mathrm{d}\mathbf{x} \right\}.$$

To solve for $\boldsymbol{\theta}$, an estimator of $g(\mathbf{x})$ is needed. Three straightforward $g(\mathbf{x})$ estimators could be considered: the true $g(\mathbf{x})$; a parametric $g(\mathbf{x}) = g(\mathbf{x}; \alpha)$ with α estimated as $\widehat{\alpha}$ through full data likelihood method; a nonparametric $g(\mathbf{x})$ with its cumulative distribution function estimated by its empirical version. These three alternatives lead to three different pseudolikelihood estimators of $\boldsymbol{\theta}$: $\widehat{\boldsymbol{\theta}}_{\mathrm{PL0}}$, $\widehat{\boldsymbol{\theta}}_{\mathrm{PL1}}$ and $\widehat{\boldsymbol{\theta}}_{\mathrm{PL2}}$. At first sight, one would believe that $\widehat{\boldsymbol{\theta}}_{\mathrm{PL0}}$ is superior to the other two in terms of estimation efficiency. However, Tang et al. (2003) showed that $\widehat{\boldsymbol{\theta}}_{\mathrm{PL0}}$ is always less efficient than $\widehat{\boldsymbol{\theta}}_{\mathrm{PL1}}$. In a recent paper, Zhao and Ma (2018) further proved that $\widehat{\boldsymbol{\theta}}_{\mathrm{PL1}}$ is always less efficient than $\widehat{\boldsymbol{\theta}}_{\mathrm{PL2}}$ and there is no other method which could lead to a more efficient estimator than $\widehat{\boldsymbol{\theta}}_{\mathrm{PL2}}$, hence $\widehat{\boldsymbol{\theta}}_{\mathrm{PL2}}$ is optimal.

Other work along this line includes Miao and Tchetgen (2016) exploring different types of doubly robust estimators and Fang, Zhao, and Shao (2018) extending the idea to missing covariate and proposing some imputation approach based on estimating equations.

2. Mechanism based on statistical chromatography

The other unspecified missing data mechanism investigated in the literature is to assume a decomposable model

$$\mathrm{pr}(\delta = 1 \mid \mathbf{x}, y) = s(\mathbf{x}) t(y), \tag{3}$$

where $s(\cdot)$ and $t(\cdot)$ are two unspecified functions. It is clear that, MCAR ($s = t =$ constant) and MAR ($t =$ constant) are special cases of this assumption. When $s =$ constant, it becomes the case discussed in Section 1 where $\mathbf{X}$ on its own serves as the nonresponse instrument.

A pivotal observation following (3) is that, $p(y \mid \mathbf{x})$ and $p(y \mid \mathbf{x}, \delta = 1)$ could be bridged as

$$p(y \mid \mathbf{x}, \delta = 1) = \frac{\mathrm{pr}(\delta = 1 \mid \mathbf{x}, y)}{\mathrm{pr}(\delta = 1 \mid \mathbf{x})} p(y \mid \mathbf{x}).$$

Note that $\mathrm{pr}(\delta = 1 \mid \mathbf{x}, y) / \mathrm{pr}(\delta = 1 \mid \mathbf{x})$ preserves to be a function of $\mathbf{x}$-only multiples a function of y-only. Using the idea of the conditional likelihood (Kalbfleisch, 1978), decomposing the observed y_i's as its rank statistic and order statistic, considering the likelihood conditional on the order statistic, Liang and Qin (2000) proposed the following objective function to estimating $\boldsymbol{\theta}$:

$$\prod_{1 \le i < j \le m} \frac{p(y_i \mid \mathbf{x}_i; \boldsymbol{\theta}) p(y_j \mid \mathbf{x}_j; \boldsymbol{\theta})}{p(y_i \mid \mathbf{x}_i; \boldsymbol{\theta}) p(y_j \mid \mathbf{x}_j; \boldsymbol{\theta}) + p(y_i \mid \mathbf{x}_j; \boldsymbol{\theta}) p(y_j \mid \mathbf{x}_i; \boldsymbol{\theta})}, \tag{4}$$

where the first m subjects are fully observed without the loss of generality.

The key here is that we model the data at a more refined granularity of rank and order statistics, so that sophisticated conditioning arguments could be applied to separate the parameter of interest $\boldsymbol{\theta}$ and other nuisance components. Hence we call this procedure statistical chromatography.

We elaborate under the generalised linear model framework where

$$p(y \mid \mathbf{x}; \boldsymbol{\theta}) = \exp\left[\phi^{-1}\{y\eta - b(\eta)\} + c(y; \phi)\right]$$

with link function structure $g(\mu(\eta)) = \alpha + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$. With canonical link, to maximise (4) is equivalent to minimising

$$\sum_{1 \le i < j \le m} \log\left[1 + \exp\{-(y_i - y_j)(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}} \boldsymbol{\gamma}\}\right],$$

where $\boldsymbol{\gamma} = \phi^{-1}\boldsymbol{\beta}$, $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}^{\mathrm{T}}, \phi)^{\mathrm{T}}$. Hence to compensate for missing data, we could only estimate $\boldsymbol{\gamma}$ as opposed to the whole unknown parameter $\boldsymbol{\theta}$. Although only $\boldsymbol{\gamma}$ is estimable, the hypothesis testing $\boldsymbol{\beta} = \mathbf{0}$ versus $\boldsymbol{\beta} \neq \mathbf{0}$ could still be carried out since the null hypothesis $\boldsymbol{\beta} = 0$ is equivalent to $\boldsymbol{\gamma} = 0$. The detailed Wald type test statistic needs the asymptotic distribution of the estimator of $\boldsymbol{\gamma}$ under this scheme (Zhao and Shao, 2017). With noncanonical link, Zhao and Shao (2017) showed that, interestingly, the whole unknown parameter $\boldsymbol{\theta}$ is estimable under some situations.

Finally I would like to point out a regularisation approach for high-dimensional variable selection with missing data using this approach. The essential idea is to identify 'important' variables through whether the corresponding estimator $\widehat{\gamma_j}$ equals zero or not. The penalised likelihood function is

$$\sum_{1 \le i < j \le m} \log\left[1 + \exp\{-(y_i - y_j)(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}} \boldsymbol{\gamma}\}\right] + \sum_{j=1}^{p} p_\lambda(|\gamma_j|),$$

where $p_\lambda(\cdot)$ could be any penalty function, and $\lambda \ge 0$ is the tuning parameter. Zhao et al. (2018) proved that the validity of the selection consistency allows p to grow at a rate exponentially fast with n as $\log p = o(n^{1-4\kappa}/(\log n)^2)$ with $0 < \kappa < 1/4$. In penalised likelihood approach for variable selection, the determination of the tuning parameter is also critical. Zhao and Yang (2017) further studied some stability enhanced tuning parameter selection methods following this approach.

Disclosure statement

No potential conflict of interest was reported by the author.

Funding

This work was supported by National Center for Advancing Translational Sciences [UL1TR001412].

Notes on contributor

Jiwei Zhao is Assistant Professor in Department of Biostatistics at the State University of New York at Buffalo. He mainly works on statistical problems motivated from various disciplines such as mental health, orthopaedics and sports medicine, women's health, aging research and the use of electronic medical records. He is generally interested in nonignorable missing data, semiparametric theory, nonregular likelihoods and semisupervised learning.

References

d'Haultfoeuille, X. (2010). A new instrumental method for dealing with endogenous selection. *Journal of Econometrics*, *154*, 1–15.

Fang, F., Zhao, J., & Shao, J. (2018). Imputation-based adjusted score equations in generalized linear models with nonignorable missing covariate values. *Statistica Sinica*, *28*.

Kalbfleisch, J. D. (1978). Likelihood methods and nonparametric tests. *Journal of the American Statistical Association*, *73*, 167–170.

Kim, J. K., & Shao, J. (2013). *Statistical Methods for Handling Incomplete Data*. Boca Raton, FL: Chapman & Hall/CRC.

Liang, K.-Y., & Qin, J. (2000). Regression analysis under nonstandard situations: a pairwise pseudolikelihood approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *62*, 773–786.

Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, *83*, 1198–1202.

Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd). Hoboken, NJ: Wiley.

Miao, W., & Tchetgen Tchetgen, E. J. (2016). On varieties of doubly robust estimators under missingness not at random with a shadow variable. *Biometrika*, *103*, 475–482.

Molenberghs, G., Fitzmaurice, G., Kenward, M. G., Tsiatis, A. A., & Verbeke, G. (2014). *Handbook of Missing Data Methodology*. Boca Raton, FL: Chapman & Hall/CRC Press.

Tang, G., Little, R. J., & Raghunathan, T. E. (2003). Analysis of multivariate missing data with nonignorable nonresponse. *Biometrika*, *90*, 747–764.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. New York, NY: Springer.

Zhao, J., & Ma, Y. (2018). Optimal pseudolikelihood estimation in the analysis of multivariate missing data with nonignorable nonresponse. *Biometrika*, *105*, 479–486.

Zhao, J., & Shao, J. (2015). Semiparametric pseudo-likelihoods in generalized linear models with nonignorable missing data. *Journal of the American Statistical Association*, *110*, 1577–1590.

Zhao, J., & Shao, J. (2017). Approximate conditional likelihood for generalized linear models with general missing data mechanism. *Journal of Systems Science and Complexity*, *30*, 139–153.

Zhao, J., & Yang, Y. (2017). Tuning parameter selection in the LASSO with unspecified propensity. In D.-G. Chen, Z. Jin, G. Li, Y. Li, A. Liu, & Y. Zhao (Eds.), *New Advances in Statistics and Data Science* (pp. 109–125). New York, NY: Springer.

Zhao, J., Yang, Y., & Ning, Y. (2018). Penalized pairwise pseudo likelihood for variable selection with nonignorable missing data. *Statistica Sinica*, *28*.