



## Rejoinder: statistical inference for non-ignorable missing-data problems: a selective review

Niansheng Tang & Yuanyuan Ju

To cite this article: Niansheng Tang & Yuanyuan Ju (2018) Rejoinder: statistical inference for non-ignorable missing-data problems: a selective review, *Statistical Theory and Related Fields*, 2:2, 146-149, DOI: [10.1080/24754269.2018.1532736](https://doi.org/10.1080/24754269.2018.1532736)

To link to this article: <https://doi.org/10.1080/24754269.2018.1532736>



Published online: 24 Oct 2018.



Submit your article to this journal [↗](#)



Article views: 66



View related articles [↗](#)



View Crossmark data [↗](#)

## Rejoinder: statistical inference for non-ignorable missing-data problems: a selective review

Niansheng Tang and Yuanyuan Ju

Key Lab of Statistical Modeling and Data Analysis of Yunnan Province, Yunnan University, Kunming, People's Republic of China

**ARTICLE HISTORY** Received 2 October 2018; Accepted 3 October 2018

We thank five discussants for their thoughtful comments. All have made significant contributions to the general theme raised in our paper. We will try our best to answer each of points that five discussers have made.

### 1. Response to Dr. Fang and Ni

We agree with the assessment of Fang and Ni that our work mainly focuses on low-dimensional data analysis and feature screening with missing data has not been well addressed. Fang and Ni presented several literatures on feature screening with response missing at random, and noted that there are two possible research topics for feature screening with missing data such as screening with non-ignorable missing response that is rather challenging, and screening with missing covariates.

For feature screening with missing categorical response  $Y$  at random that has  $R$  ( $R > 2$ ) classes  $\{y_1, y_2, \dots, y_R\}$ , if we define  $Y^* = Y$  if  $Y$  is observed and  $Y^* = R + 1$  if  $Y$  is missing, and  $\tau_r(X, Y) = E_X\{F_r(X)\} - 1/2$  that is a dependence measure between  $I(Y = y_r)$  for given  $r = 1, \dots, R$  and continuous covariate  $X$ , where  $F_r(x) = \Pr(X \leq x | Y = y_r)$ , thus one can pursue feature screening by investigating the relationship between  $\tau_r(X, Y)$  and  $\tau_r(X, Y^*)$ . Also, similar to Cui, Li, and Zhong (2015), when there is missing data, one can consider the following index  $MV(X, Y) = E_X[\text{var}_Y\{F_r(X)\}]$ . When categorical response  $Y$  is missing at random, one can consider the relationship between  $MV(X, Y)$  and  $MV(X, Y^*)$ . When continuous response  $Y$  is missing at random, a quantile association-based index for measuring the dependence between  $Y$  and  $X$  can be developed to identify the important covariates. If continuous response  $Y$  is missing at random and  $X$  is continuous, one can use the following index

$$\omega = E[\delta F(X)F(Y) + (1 - \delta)F(X)E\{F(Y)|X, \delta = 1\}] - 1/4$$

to measure the dependence between  $X$  and  $Y$ , where  $F(x) = P(X \leq x)$  and  $F(y) = P(Y \leq y)$  are the

distribution functions of  $X$  and  $Y$ , respectively, and  $\delta$  is the missing data indicator for  $Y$ , i.e.  $\delta = 1$  if  $Y$  is observed and  $\delta = 0$  if  $Y$  is missing.

For feature screening with missing covariates, Fang and Ni only considered a special case: response  $Y$  is binary and covariates are categorical, and proposed considering the relationship between  $IV(X^*, Y)$  and  $IV(X, Y)$ , where the screening index 'information value' is defined as

$$IV(X, Y) = \sum_{j=1}^J \{P(X = j | Y = 2) - P(X = j | Y = 1)\} \times \log \frac{P(X = j | Y = 2)}{P(X = j | Y = 1)}$$

in which  $X$  is a categorical covariate with values  $1, 2, \dots$ , and  $X^* = X$  if the missing data indicator  $\delta_X = 1$  (i.e.  $X$  is observed) and  $X^* = J + 1$  if  $\delta_X = 0$  (i.e.  $X$  is missing). In addition, they also presented an available case (AC) method and a two-step screening procedure. These methods are useful for screening important features under the considered case. However, when response  $Y$  is not binary but categorical or continuous, or covariate  $X$  is continuous, it is rather challenging to develop a new feature screening procedure with missing data.

On the other hand, their method may cause several problems in practical applications. First, the first step of their two-step screening procedure may lead to a biased estimator. Second, under some strong conditions, a relatively high proportion of missing data and unbalanced categorical data, it is rather difficult to guarantee the accuracy of estimation in the second step. To wit, it follows from  $P(U_k = j, Y = r) = \sum_{\mathbf{v}} P(\mathbf{V}^{S_k} = \mathbf{v}, Y = r)P(U_k = j | \mathbf{V}^{S_k} = \mathbf{v}, Y = r, \delta_k = 1)$  that the condition probability  $P(U_k = j | \mathbf{V}^{S_k} = \mathbf{v}, Y = r, \delta_k = 1)$  is strong condition, where  $\mathbf{U} = (U_1, \dots, U_p)$  and  $\mathbf{V} =$

$(V_1, \dots, V_q)$  correspond to covariates with and without missing values, respectively,  $\delta_k$  is the missing data indicator for  $Y$ ,  $S_k$  is a small subset of  $\{1, \dots, q\}$  and  $\mathbf{V}^{S_k} = \{V_l : l \in S_k\}$ .

## 2. Response to Dr. Wang

Longitudinal data are commonly encountered in clinical trials, medicine and social sciences. In longitudinal studies, dropout data invariably occur, that is, some participants might dropout of the study or be lost to follow-up due to some reasons, which leads to the loss of their outcome measurements. In these cases, there are two types of dropout mechanisms: ignorable dropout (i.e. the probability that a participant dropouts the study only depends on the observed data) and non-ignorable dropout (i.e. the probability that a participant dropouts the study depends the missing values possibly except for the observed data). Generally, there are two types of dropout patterns: monotone and non-monotone dropout data. Non-monotone dropout data occur when study participants intermittently miss scheduled visits, while monotone dropout data can be from discontinued participation, loss to follow-up, and mortality.

Longitudinal data analysis with dropout data is not a new topic, and has been studied by many authors. For example, see Hedeker and Gibbons (2006) and Tseng, Elashoff, Li, and Li (2016). Recently, Wang, Qi, and Shao (2018) discussed longitudinal data analysis with non-ignorable dropout by incorporating the idea of instrument variables and estimating equations and a parametric dropout propensity score, presented a two-step generalised method of moments (GMMs) to estimate unknown parameters in the considered parametric propensity score, and investigated asymptotic properties of the proposed GMM estimator. However, they did not consider dropout patterns, which may lead to new approaches for handling dropout data.

On the other hand, considering dropout patterns may improve the efficiency of GMM estimators given in Wang et al. (2018) and solve the optimal estimation issue you concerned. Also, to address the optimal estimation issue, one can make use of all possible information included in responses and covariates and dropout study in constructing estimating equations. Wang et al. (2018) considered a parametric propensity score model for each of non-ignorable dropout data  $y_t$  by incorporating history response data  $\{y_1, \dots, y_{t-1}\}$  and time-independent covariates  $U$ . However, the considered parametric propensity score model has the following several problems. First, there are too many parameters, which may lead to the identifiable issue even if Wang et al. (2018) considered instrument variables, when the sample size is small and  $T$  is relatively large. Second, it may lead to the well-known ill-posed issue when  $T$  is growing with the sample size  $n$ , i.e. a

slow convergence rate in evaluating GMM estimates of parameters. Third, some of interactions of covariates may have a large effect on participant's dropout. Forth, it is impossible to test the plausibility of the posited parametric propensity score model. To address the aforementioned issues, one may consider a sequence of one-dimensional parametric propensity score models as done in Equation (19) of our paper, which can be used to construct estimating equations and select the important variables that have a large effect on participant's dropout via penalty method.

Although a semiparametric model can be adopted for dropout data analysis, there is the well-known 'curse of dimensionality' when  $T$  is large. In this case, one may consider an additive model for  $y_1, \dots, y_{t-1}$ .

## 3. Response to Dr. Zhao

We agree with your valuable assessment that one intrinsic complication of missing data analysis is that it is quite difficult to verify its underlying truth in practical applications because of missing values involved, and it is interesting to develop versatile statistical procedures that are robustness to the misspecification of dropout mechanism model. In this discussion, you considered two types of missingness data mechanism models: conditional independence mechanism and statistical chromatography mechanism, and then discussed model identification, point estimation, hypothesis testing and high-dimensional variable selection. Conditional independence assumption for dropout data utilises the concept of non-response instrument, the resultant statistical inference on model parameter of interest can be carried out without the need of estimating missingness data mechanism, hence it has received a lot of attention in recent years. For example, see Zhao and Shao (2015) and Zhao and Shao (2017). But this method only utilises the observed data information when covariates serve as non-response instrument variables. Moreover, when the considered model is rather complicated, for example, our considered exponential family nonlinear structure equation models, Zhao and Shao (2015) involved intractable high-dimensional integrals. Under non-ignorable dropout assumption, a possible improvement for Zhao and Shao's (2015) method is to estimate unknown parameters  $\theta$  in  $p(Y|\mathbf{X}) = p(Y|\mathbf{X}, \theta)$  via the conditional likelihood of  $p(\mathbf{X}|Y, \delta = 0)$ , where  $\delta$  is the missing data indicator for response variable  $Y$ , i.e.  $\delta = 1$  if  $Y$  is observed and 0 otherwise.

Statistical chromatography model for dropout data:  $\Pr(\delta = 1|\mathbf{X}, Y) = s(\mathbf{X})t(y)$  is an unspecified missing data mechanism, and corresponds to various missing data mechanisms by taking different forms of  $s(\mathbf{X})$  and  $t(y)$ . Based on statistical chromatography model, using the idea of the conditional likelihood and 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) developed an approach to estimate model parameters in  $\theta$ . But Liang and Qin (2000) only utilises the fully observed data, which indicates that their method may lead to biased estimator when missing proportion is high and may be not robust to the misspecification of the considered model. More importantly, their method requires specifying the distribution of response variable, which may limit its applications. In particular, when the considered model is quite complicated, for example, generalised linear model with non-canonical link function, or involves latent variables, such as our considered exponential family nonlinear structural equation models, this method requires handling high-dimensional integrals.

For generalised linear model with canonical link function, i.e.  $p(y|x, \theta) = \exp[\phi^{-1}\{\eta - b(\eta)\} + c(y; \phi)]$  with  $\eta = \alpha + \beta^T x$ , one can obtain the estimation of  $\gamma = \phi^{-1}\beta$  by minimising the objective function:  $\sum_{1 \leq i \leq j \leq m} \log[1 + \exp\{-(y_i - y_j)(x_i - x_j)^T \gamma\}]$ , where the first  $m$  subjects are fully observed without of generality, but it is impossible to obtain the estimation of dispersion parameter  $\phi$  and intercept parameter  $\alpha$  via the method of Liang and Qin (2000). It is interesting to develop an efficient approach to estimate all the parameters in  $\beta$ ,  $\phi$  and  $\alpha$  based on statistical chromatography mechanism.

#### 4. Response to Dr. Morikawa and Professor Kim

We agree with Dr. Morikawa and Professor Kim's comment on the semiparametric estimation of mean functionals, which is a good supplement to our presented estimation procedure for mean functionals in the presence of non-ignorable missing data. Dr. Morikawa and Professor Kim proposed a new approach to restore Wilks' phenomenon in empirical likelihood inference with non-ignorable missing values. A remarkable feature of Morikawa and Kim's method is that the resulting profile empirical log-likelihood ratio statistic can be directly used to construct confidence interval of parameter of interest. However, Morikawa and Kim's empirical likelihood approach has a very limited application scope because it only works for response mean or a single parameter. Let  $\theta \in \mathcal{R}^p$  be a vector of parameters of interest, and  $\theta_0$  be the true value of  $\theta$ . Here  $\theta_0$  is uniquely defined via generalised estimating equations of the form  $E\{\psi(X, Y, \theta_0)\} = 0$ , where  $\psi(X, Y, \theta_0)$  is a vector of  $r$  ( $\geq p$ ) functions, variables  $X$  and  $Y$  follow some unknown joint distribution  $F(X, Y, \theta)$ , and  $E\{\cdot\}$  represents the expectation taken with respect to  $F(X, Y, \theta)$ . Generalised estimating equations encompass a large class of statistical models. For example,  $\psi(X, Y, \theta) = Y - \theta$  for mean functionals of responses, which is studied by Kim and Yu (2001). Let  $\delta$  be the missing data indicator, taking 1 if  $Y$  is observed

and 0 if  $Y$  is missing. It is assumed that  $X = (U^T, Z^T)$  are fully observed, and  $Z$  is an instrument variable and missing data mechanism is specified by a semiparametric propensity score function  $\Pr(\delta = 1|X, Y) = \Pr(\delta = 1|U, Y) = \pi(U, Y, \gamma) = \pi(\gamma)$ . Define  $\hat{\xi}_i(\theta, \gamma) = \delta_i \psi(X_i, Y_i; \theta) + (1 - \delta_i) \hat{m}_\psi^0(X_i; \gamma)$  and  $\hat{\xi}_i^a(\theta, \gamma) = \delta_i \psi(X_i, Y_i; \theta) / \pi_i(\gamma) + \{1 - \delta_i / \pi_i(\gamma)\} \hat{m}_\psi^0(X_i; \gamma)$ , where  $\hat{m}_\psi^0(X_i; \gamma)$  is some consistent estimator of  $m_\psi^0(X_i) = E\{\psi(X_i, Y_i; \theta) | X_i, \delta_i = 0\}$ . Following the idea of Morikawa and Kim and using the idea of Shao and Wang (2016) introduced in the beginning of Section 3 as the calibration conditions, we define the following profile empirical log-likelihood ratio function for  $(\theta, \gamma)$ :

$$\ell(\theta, \gamma) = \sup \left\{ \sum_{i=1}^n \log(np_i) \mid p_i \geq 0, \sum_{i=1}^n p_i = 1, \right. \\ \left. \sum_{i=1}^n \hat{\xi}_i(\theta, \gamma) = 0, \sum_{i=1}^n \hat{\xi}_i^a(\theta, \gamma) = 0, \right. \\ \left. \sum_{i=1}^n p_i M_l(Y_i, U_i, \delta_i, \hat{g}_\gamma, \gamma) = 0 \text{ for } l = 1, \dots, L \right\}.$$

An empirical log-likelihood ratio statistic for testing hypothesis  $H_0 : \theta = \theta_0$  is defined as

$$R_n(\theta_0) = 2 \left\{ \sup_{\theta, \gamma} \ell(\theta, \gamma) - \sup_{\gamma} \ell(\theta_0, \gamma) \right\}.$$

Under some regularity conditions, it is easily shown that  $R_n(\theta_0)$  asymptotically follows the chi-squared distribution with  $p$  degrees of freedom, which is a natural extension of the Wilks' theorem to a general parameter case.

Lazar (2003) pointed out that empirical likelihood can also be used in a posterior inference in place of the parametric likelihood function in Bayes' theorem. Zhang and Tang (2017) extended Bayesian empirical likelihood to quantile structural equation models. Given empirical likelihood function  $\ell(\theta, \gamma)$  and specify a prior  $\xi(\theta, \gamma)$  for  $(\theta, \gamma)$ , we obtain the quasi-posterior density

$$p(\theta, \gamma | \mathbf{O}) = C_0 \exp [\log\{\xi(\theta, \gamma)\} + \ell(\theta, \gamma)],$$

where  $\mathbf{O} = \{(X_i, Y_i, \delta_i), i = 1, \dots, n\}$  and  $C_0$  is a normalising constant such that  $\iint p(\theta, \gamma | \mathbf{O}) d\theta d\gamma = 1$ . Since it is quite easy to calculate the value of empirical log-likelihood function  $\ell(\theta, \gamma)$  given  $\theta$  and  $\gamma$ , the implement of Metropolis–Hastings algorithm is feasible for sampling observations required in making Bayesian inference on  $\theta$  from the posterior  $p(\theta, \gamma | \mathbf{O})$ . Bayesian empirical likelihood approach is a more flexible and effective tool in that it not only can calculate point estimates and confidence intervals but also allows incorporation of prior information, and can circumvent the inherent 'curse of dimensionality' in evaluating empirical likelihood estimators.

## 5. Response to Professor Shao

Professor Shao explores general assumptions for semiparametric non-ignorable propensity model and further leaves an open question on how to perform the nonparametric regression analysis in non-ignorable missing data problems when the semiparametric non-ignorable propensity is of a general form. The key to perform nonparametric regression estimation for non-ignorable missing data is to find a kernel type estimator of the nonparametric part in a semiparametric non-ignorable propensity. Motivated by the comments of Shao, we consider the following more general definition of semiparametric non-ignorable propensity.

Let  $O(U, Y, \gamma_0) = 1/\pi(U, Y, \gamma_0) - 1$ . It can be shown that

$$E\{\delta O(U, Y, \gamma_0) \mid U\} = E\{1 - \delta \mid U\}. \quad (1)$$

Assume that  $O(U, Y, \gamma) = h\{g(U), q_\gamma(Y)\}$ ,  $h(\cdot, \cdot)$  is an arbitrary user-specified function. Such semiparametric conditional odds model also defines a semiparametric non-ignorable propensity  $\pi(U, Y, \gamma_0) = 1/[1 + h\{g(U), q_{\gamma_0}(Y)\}]$ . It follows from (1) that the nonparametric function  $g(U)$  can be profiled using the kernel regression method when the function  $h(a, b)$  is decomposable and has of the forms:  $h(a, b) = \varphi(a)b, \varphi(a)/b, \varphi(a) + b$  and  $\varphi(a) - b$ , among others. Here  $\varphi(a)$  is an arbitrary known function. When  $h(a, b)$  has of a general (implicit) form, developing a kernel regression estimation of semiparametric non-ignorable propensity is challenging, and thus remains an open problem as pointed out by Shao. Therefore, it deserves further research.

## Acknowledgments

The authors are grateful to the Editor for organising the discussion of our paper and the discussants Dr. Fang Fang, Dr. Lyu Ni, Dr. Lei Wang, Dr. Jiwei Zhao, Dr. Kosuke Morikawa, Professor Jae Kwang Kim and Professor Jun Shao, seven leading figures on statistical inference for missing data, for their stimulating comments and insightful contributions to this work.

## Disclosure statement

No potential conflict of interest was reported by the authors.

## Funding

This work was supported by the grants from the Key Projects of the National Natural Science Foundation of China [grant number 11731101] and the National Natural Science Foundation of China [grant number 11671349].

## References

- Cui, H., Li, R., & Zhong, W. (2015). Model-free feature screening for ultrahigh dimensional discriminant analysis. *Journal of the American Statistical Association*, 110, 630–641.
- Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis*. Hoboken, NJ: John Wiley & Sons.
- Kim, J. K., & Yu, C. L. (2011). A semiparametric estimation of mean functionals with nonignorable missing data. *Journal of the American Statistical Association*, 106, 157–165.
- Lazar, N. A. (2003). Bayesian empirical likelihood. *Biometrika*, 90, 319–326.
- Liang, K. Y., & Qin, J. (2000). Regression analysis under non-standard situations: A pairwise pseudolikelihood approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62, 773–786.
- Shao, J., & Wang, L. (2016). Semiparametric inverse propensity weighting for nonignorable missing data. *Biometrika*, 103, 175–187.
- Tseng, C. H., Elashoff, R., Li, N., & Li, G. (2016). Longitudinal data analysis with non-ignorable missing data. *Statistical Methods in Medical Research*, 25, 205–220.
- Wang, L., Qi, C., & Shao, J. (2018). Model-assisted regression estimators for longitudinal data with nonignorable dropout. *International Statistical Review*, to appear.
- Zhang, Y. Q., & Tang, N. S. (2017). Bayesian empirical likelihood estimation of quantile structural equation models. *Journal of Systems Science and Complexity*, 30, 122–138.
- Zhao, J., & Shao, J. (2016). Semiparametric pseudo-likelihoods in generalized linear models with nonignorable missing data. *Journal of the American Statistical Association*, 110, 1577–1590.
- Zhao, J. W., & Shao, J. (2017). Approximate conditional likelihood for generalized linear models with general missing data mechanism. *Journal of Systems Science and Complexity*, 30, 139–153.