Statistical Theory and Related Fields

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

Variable screening with missing covariates: a discussion of 'statistical inference for nonignorable missing data problems: a selective review' by Niansheng Tang and Yuanyuan Ju

Fang Fang & Lyu Ni

To cite this article: Fang Fang & Lyu Ni (2018) Variable screening with missing covariates: a discussion of 'statistical inference for nonignorable missing data problems: a selective review' by Niansheng Tang and Yuanyuan Ju, Statistical Theory and Related Fields, 2:2, 134-136, DOI: 10.1080/24754269.2018.1522574

To link to this article: https://doi.org/10.1080/24754269.2018.1522574

Published online: 22 Sep 2018.

STATISTICAL THEORY AND RELATED FIELDS
2018, VOL. 2, NO. 2, 134–136
https://doi.org/10.1080/24754269.2018.1522574

SHORT COMMUNICATION

Variable screening with missing covariates: a discussion of 'statistical inference for nonignorable missing data problems: a selective review' by Niansheng Tang and Yuanyuan Ju

Fang Fang* and Lyu Ni*

School of Statistics, East China Normal University

ABSTRACT
Feature screening with missing data is a critical problem but has not been well addressed in the literature. In this discussion we propose a new screening index based on "information value" and apply it to feature screening with missing covariates.

ARTICLE HISTORY
Received 20 August 2018
Accepted 9 September 2018

KEYWORDS
Feature screening; missing at random; missing covariates

We thank Tang and Ju for their extensively review for the methods dealing with a challenging statistical problem: missing data. The methods discussed in the paper mainly focus on low dimensional data. In the discussion part, the paper mentioned feature screening with missing data, which is a critical research topic but has not been well addressed in the literature.

Several works have been done to handle feature screening with response missing at random. For example, (Lai, Liu, Liu, & Wan, 2017) used inverse probability weighting method to recover the screening indexes when missing data exist. Wang and Li (2018) proposed a missing indicator imputation screening procedure by noting the fact that the set of the active covariates for the response is a subset of the active covariates for the product of the response and missingness indicator. There are two possible directions to further discuss the feature screening methods with missing data. First is to consider screening with nonignorable missing response, which could be quite challenging. Second is to consider screening with missing covariates.

Missing covariate data commonly exist in such health and biomedical related studies as clinical trials, observational data, environmental studies, and health surveys. How to conduct feature screening when some covariates are missing is an interesting problem. While it could be difficult to solve this problem in general, there are special cases in which screening with missing covariates is possible. Here we discuss one special case: the response Y is binary and all the covariates are categorical. If there is no missing data, the PC-SIS in (Huang, Li, & Wang, 2014), IG-SIS in Fang (2016) and APC-SIS in Ni, Fang, and Wan (2017) all have sure screening property (Fan & Lv, 2008). Other than these methods, we propose a new screening index "information value", which is defined as

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

where Y is the binary response with values 1 or 2 and X is a categorical covariate with values $1,2,\ldots,J$. It is easy to see that $IV(X,Y)=0$ if and only if X and Y are statistically independent. If we select the covariates with the largest d estimated IV values as the active covariates, it is not hard to show that this screening procedure has sure screening property.

If X has missing data, then $IV(X,Y)$ can not be estimated directly. Let δ_X be the missingness indicator: $\delta_X=1$ is X is observed and $\delta_X=0$ if X is missing. Define a new categorical covariate as

$$X^*=\begin{cases} X & \text{if } \delta_X=1 \\ J+1 & \text{otherwise}\end{cases}$$

We may want to see what is the relationship between $IV(X^*,Y)$ and $IV(X,Y)$. Actually we have the following two conclusions:

(1) If $P(\delta_X=1|X,Y)=P(\delta_X=1)$, then $IV(X^*,Y)=P(\delta_X=1)IV(X,Y)$.
(2) If $P(\delta_X=1|X,Y)=P(\delta_X=1|X)$, then $IV(X^*,Y)\le IV(X,Y)$.

The first conclusion tells us that if X is missing completed at random, we can use $\widehat{IV}(X^*,Y)/\widehat{P}(\delta_X=1)$ to recover $IV(X,Y)$. The second conclusion tells us that

CONTACT Fang Fang ffang@sfs.ecnu.edu.cn School of Statistics, East China Normal University
*Present address: Key Laboratory of Advanced Theory and Application in Statistics and Data Science (East China Normal University), Ministry of Education.

when the missing probability of X only depends on X itself, $IV(X^*, Y)$ will always underestimate $IV(X, Y)$. So $IV(X^*, Y)$ is not likely to mistakenly select inactive covariates. However, it may miss some active covariates.

Other than considering $IV(X^*, Y)$, we may also consider the commonly used AC (available case) method. That is, we only use the non-missing data of X to estimate $IV(X, Y)$. Denote

$$
\begin{aligned}
& IV_{\{\delta_X=1\}}(X, Y) \\
&= \sum_{j=1}^{J} \{P(X=j|Y=2, \delta_X=1) \\
&- P(X=j|Y=1, \delta_X=1)\} \\
&\log \frac{P(X=j|Y=2, \delta_X=1)}{P(X=j|Y=1, \delta_X=1)}
\end{aligned}
$$

as the AC analog of $IV(X, Y)$. In what situation we can use $IV_{\{\delta_X=1\}}(X, Y)$ to recover $IV(X, Y)$? Consider two covariates X_1 and X_2, where X_1 has missing data and X_2 is always observed. Under the following two conditions:

(C1) $P(\delta_{X_1}=1|X_1, X_2, Y) = P(\delta_{X_1}=1|X_2, Y)$,
(C2) $P(X_1=j_1, X_2=j_2|Y) = P(X_1=j_1|Y)$ $P(X_2=j_2|Y)$,

we have $IV_{\{\delta_X=1\}}(X, Y) = IV(X, Y)$. Condition (C1) means X_1 is missing at random. Condition (C2) means X_1 and X_2 are conditionally (on Y) independent, which is similar to the condition required by naive bayes. Missing at random may be a reasonable assumption in many situations. But conditional independence usually does not hold. However, this AC method still works well in several simulations conducted by us even (C2) does not hold. Just like naive bayes works well in many situations even the conditional independence condition is violated. Here we only discuss two covariates X_1 and X_2, but all the conditions and conclusions can be extended to two groups of covariates, in which one group has missing data and the other group is always observed.

Finally we propose a method which is more applicable than the two methods discussed above based on $IV(X^*, Y)$ or $IV_{\{\delta_X=1\}}(X, Y)$. Denote $\mathbf{U} = (U_1, \ldots, U_p)$ as the covariates with missing data and $\mathbf{V} = (V_1, \ldots, V_q)$ as the covariates without missing data. For each missing covariate U_k, the missing indicator is denoted as δ_k, $k = 1, \ldots, p$. We assume that

$$P(\delta_k = 1|Y, \mathbf{U}, \mathbf{V}) = P(\delta_k = 1|Y, \mathbf{V}^{\mathcal{S}_k}),$$

where $\mathcal{S}_k$ is a small subset of $\{1, \ldots, q\}$ and $\mathbf{V}^{\mathcal{S}_k} = \{V_l : l \in \mathcal{S}_k\}$, i.e. U_k is missing at random and the missing probability only depends on Y and a small subset of covariates that are always observed. Then

$$
\begin{aligned}
& P(U_k = j, Y = r) \\
&= \sum_{\boldsymbol{v}} P(U_k = j, \mathbf{V}^{\mathcal{S}_k} = \boldsymbol{v}, Y = r) \\
&= \sum_{\boldsymbol{v}} P(\mathbf{V}^{\mathcal{S}_k} = \boldsymbol{v}, Y = r) P(U_k = j|\mathbf{V}^{\mathcal{S}_k} = \boldsymbol{v}, Y = r) \\
&= \sum_{\boldsymbol{v}} P(\mathbf{V}^{\mathcal{S}_k} = \boldsymbol{v}, Y = r) \\
& P(U_k = j|\mathbf{V}^{\mathcal{S}_k} = \boldsymbol{v}, Y = r, \delta_k = 1) \qquad (2)
\end{aligned}
$$

can be easily estimated if $\mathcal{S}_k$ is known, where $r = 1$ or 2, and the summation is over all possible values of $\mathbf{V}^{\mathcal{S}_k}$. Then further we can estimate $IV(U_k, Y)$. We propose a two-step screening procedure as follows:

Step 1: Apply APC-SIS or IG-SIS on data $\{\delta_k, \mathbf{V}\}$ to get $\hat{\mathcal{S}}_k$.

Step 2: Estimate $P(U_k = j, Y = r)$ based on $\hat{\mathcal{S}}_k$ and (2). Further estimate $IV(U_k, Y)$ based on (1). $IV(V_l, Y)$ can be estimated regularly since V_l is fully observed. Then we can select the covariates with the largest d estimated IV values.

Under some regularity conditions, this screening procedure has sure screening property.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Fang Fang is an associate professor at Key Laboratory of Advanced Theory and Application in Statistics and Data Science (East China Normal University), Ministry of Education, and School of Statistics, East China Normal University. He is also an Associate Editor of *Journal of Nonparametric Statistics*. His research interests mainly focus on missing data, model averaging and statistical learning.

Lyu Ni is a Ph.D. candidate of statistics at the School of Statistics, East China Normal University. Her research areas are feature screening in high-dimensional data and missing data analysis.

References

Fan, J., & Lv, J. (2008). Sure independent screening for ultrahigh dimensional feature space (with discussion). *Journal of Royal Statistical Society, Series B*, *70*, 849–911.

Fang, L., & Ni, F. (2016). Entropy-based model-free feature screening for ultrahigh-dimensional multiclass classification. *Journal of Nonparametric Statistics*, *28*, 515–530.

Huang, D. Y., Li, R. Z., & Wang, H. S. (2014). Feature screening for ultrahigh dimensional categorical data with applications. *Journal of Business & Economic Statistics*, *32*, 237–244.

Lai, P., Liu, Y. M., Liu, Z., & Wan, Y. (2017). Model free feature screening for ultrahigh dimensional data with responses missing at random. *Computational Statistics & Data Analysis*, *105*, 201–216.

Ni, L., Fang, F., & Wan, F. J. (2017). Adjusted pearson chi-square feature screening for multi-classification with ultra-high dimensional data. *Metrika*, *80*, 805–828.

Wang, Q. H., & Li, Y. J. (2018). How to make model-free feature screening approaches for full data applicable to the case of missing response?. *Scandinavian Journal of Statistics*, *45*, 324–346.