Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Dimension reduction with missing response at random

Xu Guo[a], Tao Wang[a], Wangli Xu[c], Lixing Zhu[a,b,*]

[a] Hong Kong Baptist University, Hong Kong

[b] School of Statistics and Mathematics, Yunnan University of Finance and Economics, Kun Ming, China

[c] Center for Applied Statistics, School of Statistics, Renmin University of China, Beijing, China

ARTICLE INFO

Article history:
Received 14 May 2013
Received in revised form 19 July 2013
Accepted 2 August 2013
Available online 24 August 2013

Keywords:
Central subspace
Data-adaptive synthesization
Missing recovery
Missing response at random
Multiple imputation

ABSTRACT

When there are many predictors, how to efficiently impute responses missing at random is an important problem to deal with for regression analysis because this missing mechanism, unlike missing completely at random, is highly related to high-dimensional predictor vectors. In sufficient dimension reduction framework, the fusion-refinement (FR) method in the literature is a promising approach. To make estimation more accurate and efficient, two methods are suggested in this paper. Among them, one method uses the observed data to help on missing data generation, and the other one is an ad hoc approach that mainly reduces the dimension in the nonparametric smoothing in data generation. A data-adaptive synthesization of these two methods is also developed. Simulations are conducted to examine their performance and a HIV clinical trial dataset is analyzed for illustration.

1. Introduction

In practice, due to various reasons such as loss of information caused by uncontrollable factors, unwillingness of some sampled units to provide the desired information and so on, often not all responses are available. The most commonly used method to handle missing response problems simply resorts to the complete-case (CC) analysis by discarding all the incomplete measurements with missing values. However, this practice is undesirable since the resulting estimates are inconsistent unless the missing mechanism is missing completely at random (MCAR); that is, the missingness is independent of all the observed and unobserved variables (Wang and Chen, 2009). A more general missing mechanism is missing at random (MAR) which will be investigated in the present paper. Besides, inference built on the complete case analysis is generally inefficient as it throws away data with missing values. Many efforts have been devoted to address this issue. Generally speaking, there are two ways to handle missing data problems. The first is to impute a plausible value for each missing value and then analyze the data as if they were complete. See, for example, linear regression imputation (Yates, 1933), ratio imputation (Rao, 1996), semiparametric regression imputation (Wang and Rao, 2002), and kernel regression imputation (Cheng, 1994), etc. Rubin (1987) proposed a popular and general multiple imputation (MI) procedure. The other is to use the inverse probability weighted (IPW) approach introduced by Robins et al. (1994); see Zhao et al. (1996), Wang et al. (1997), Robins et al. (1994), Wang et al. (2004), and Guo and Xu (2012). However, existing regression imputation and inverse probability weighted approaches involve high-dimensional smoothing for estimating the completely unknown regression function and selection probability function in nonparametric settings. This difficulty consequently hinders their applications due to the well known curse of dimensionality. One can refer to Little and Rubin (2002) and references therein for a comprehensive review of statistical methods dealing with missing data.

To deal with the dimensionality problem, dimension reduction is necessary for us to efficiently work on regression analysis. Sufficient dimension reduction (SDR) has generated considerable interest in high-dimensional regressions. This

[*] Corresponding author at: Hong Kong Baptist University, Hong Kong. Tel.: +852 34117016.
E-mail address: lzhu@hkbu.edu.hk (L. Zhu).

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.08.001

general methodology aims at dealing with data sparseness in high-dimensional scenarios without parametric model structure. A pioneering research is sliced inverse regression proposed by Li (SIR 1991). Let Y and X be respectively the response and predictor vector. In general, the central subspace (CS, Cook, 1998), denoted by $S_{Y|X}$ in our context, is defined as the subspace S of minimal dimension such that $Y \perp\!\!\!\perp X|P_S X$, where $\perp\!\!\!\perp$ indicates statistical independence and $P_{(\cdot)}$ is a projection operator with respect to the usual inner product. Its dimension $K = \dim(S_{Y|X})$ is often used to refer to structural dimension. We call the vectors forming a basis of $S_{Y|X}$ dimension reduction directions.

Since SIR was proposed, many SDR methods have been developed, including sliced average variance estimation (Cook and Weisberg, 1991), contour regression (Li et al., 2005), directional regression (Li and Wang, 2007), likelihood acquired directions (Cook and Forzani, 2009), discretization–expectation estimation (Zhu et al., 2010a,b) and average partial mean estimation (Zhu et al., 2010a,b) etc. Among these, SIR has been the most popular one in the literature, and there have been many elaborations on the original methodology of SIR such as Zhu and Ng (1995) for its asymptotics.

Recently, in the context of missing predictors, Li and Lu (2008) combined SIR and the augmented inverse probability weighted method for the dimension reduction problem. Zhu et al. (2012) introduced a nonparametric imputation procedure for semiparametric regressions with missing predictors. When the missingness depends on both the completely observed predictors and the response, they still needed a parametric model structure to impute missing values. Ding and Wang (2011) proposed a fusion-refinement (FR) procedure to target the dimension reduction problem with missing response. Let δ be the missingness indicator, which equals 1 if the response Y is observed and 0 otherwise. Following the literatures (Little and Rubin, 2002) in the missing data area, we adopt the commonly used missing mechanism missing at random (MAR) in this paper. To be precise, this means $P(\delta = 1|Y, X) = P(\delta = 1|X) = \pi(X)$ or in other words, $Y \perp\!\!\!\perp \delta|X$. Further $\pi(X)$ is called the selection probability. Assume that $\gamma \in \mathbb{R}^{p\times d}$ is a basis matrix of the central subspace $S_{\delta|X}$, and $\beta \in \mathbb{R}^{p\times K}$ is a basis matrix of the central subspace $S_{Y|X}$. In their fusion stage, they first enlarged the target subspace to be the joint central subspace $S_{(Y,\delta)|X}$. In the second stage of estimation, they used probability mass function (pmf) imputation method and their Theorem 2 in their paper to recover $S_{Y|X}$ that is a subspace of $S_{(\delta,Y)|X}$. The estimate is proved to be consistent. However, their numerical studies indicate that when the angle between the subspace $S_{\delta|X}$ and $S_{Y|X}$ is large, the estimation accuracy is not satisfactory in the sense that $S_{Y|X}$ may not be extracted from $S_{(\delta,Y)|X}$ straightforwardly. Ding and Wang (2011) then suggested an ad hoc way to determine a threshold value of the angle between these two estimated subspaces for the practical use. Their recommendation was based on the simulation results they conducted. It is necessary to have a data-adaptive approach to determine which method should be used for maximizing the estimation accuracy.

From the above observations, in this paper, two alternative approaches are proposed to promote the estimation accuracy. The first is to take care of the inefficiency of nonparametric estimation when the dimension of $S_{(\delta,Y)|X}$ is relatively large. For instance, it is well known that the kernel estimation can have an optimal rate of convergence $O_p(n^{-2/(4+d+K)})$ when the density function of $(\gamma, \beta)^T X$ is two times differentiable. Thus we can see that a more efficient way may be to directly impute Y via the conditional distribution of Y given the projection of X onto $S_{\delta|X}$ such that we can suffer less from the nonparametric estimation with data sparseness in high-dimensional space. This idea should also be consistent with the theme of dimension reduction investigated in this paper. Based on this motivation, a novel two-stage method is proposed. In the first stage, we obtain a basis estimate $\hat{\gamma}$ for $S_{\delta|X}$, and impute Y through the conditional distribution of Y given $\hat{\gamma}^\tau X$. This stage is called Selection Probability Assisted Recovery (SPAR). However, as $S_{\delta|X}$ is not necessarily contained in $S_{Y|X}$, we then use the CC method to assist. Since the estimate deduced from the CC method is consistent, we can develop a Complete Case Assisted Recovery (CCAR). First, obtain a basis estimate $\hat{\beta}$ for $S_{Y|X}$ from the CC analysis and then impute missing responses through the conditional distributions of Y given $\hat{\beta}^\tau X$.

From our comprehensive simulation studies we found that almost uniformly,

- When $S_{Y|X}$ is close to $S_{\delta|X}$, SPAR is better than CCAR, whereas when the angle between these two subspaces is large, CCAR is the winner.

Thus, a natural question is what angle is regarded as either small or large and then we should use either SPAR or CCAR. A straightforward idea is to choose the method which can more efficiently estimate the subspace $S_{Y|X}$. To this end, we propose a data-adaptive method to synthesize both SPAR and CCAR to produce a new estimate. The method can automatically choose one of them in a data-adaptive way to maximize the estimation accuracy. This adaptive approach is realized by the bootstrap method.

Thus, in this paper, we make a comparison between these two methods and the FR procedure through the numerical studies.

The rest of the article is organized as follows. In Section 2 the methods that are based on SPAR and CCAR are introduced. The methodologies are illustrated by adopting the commonly used SIR in Section 3. In Section 4 the data-adaptive approach for SPAR and CCAR is elaborated. Simulation studies are conducted to examine the performance of the methods with a comparison with the FR procedure. In Section 5 the proposed procedures is applied to analyze real data. Conclusions are given in Section 6.

2. Semiparametric dimension reduction assisted recovery

Since the conditional mean imputation, a commonly used imputation method, cannot be applied in our context because it focuses on only one characteristic of the conditional distribution, we then use the multiple imputation (Rubin, 1987). Below