Statistica Sinica **30** (2020), 1419-1436
doi:https://doi.org/10.5705/ss.202017.0366

NETWORK IMPUTATION FOR A SPATIAL AUTOREGRESSION MODEL WITH INCOMPLETE DATA

Zhimeng Sun and Hansheng Wang

Central University of Finance and Economics and Peking University

Abstract: Numerous imputation methods have been developed for missing data. However, these methods apply mainly to independent data, and the assumption of independence disregards connections of units through social relationships (e.g., friendship, follower−followee relationship). In fact, observed responses from connected friends should provide valuable information for missing responses. This motivates us to conduct an imputation by borrowing information from connected friends using a network structure. With the missing−at−random assumption and using observed information only, we propose a partial likelihood approach and develop the corresponding maximum partial likelihood estimator (MPLE). The estimator's consistency and asymptotic normality are established. Using the MPLE, we then develop a novel regression imputation method. The method utilizes both auxiliary information and connected complete units (i.e., network information); using the imputed data, we can compute the sample mean of the responses. We show this method to be consistent and asymptotically normal. Compared with the imputation method using auxiliary information only (i.e., ignoring network information), the proposed estimator is statistically more efficient. Extensive simulation studies are conducted to demonstrate the finite−sample performance of the proposed method. We then analyze a real example about QQ in mainland China.

Key words and phrases: Incomplete data, network imputation, QQ, spatial autoregression.

1. Introduction

Researchers encounter missing data when sampled units fail to provide values for the main variable (Kalton and Kasprzyk (1986)). There are typically three missing data mechanisms: missing completely at random (MCR), missing at random (MAR), and nonignorable missing (NM); see Rubin (1987). In the MCR case, a statistical analysis based on complete units remains asymptotically valid, although its statistical efficiency might be suboptimal. However, in either an MAR or an NM situation, a statistical analysis based on complete units may suffer from significant bias (Shao and Wang (2002)). To address this problem, various imputation methods have been developed and are widely accepted in

practice (Little and Rubin (2002); Schafer (1997)). The most straightforward imputation method is the hot-deck imputation (Rao and Shao (1992)), but its simplicity means it ignores valuable information provided by auxiliary variables. Hence, it can not correct for an estimation bias when the missing data mechanism is not MCR. Moreover, even for MCR, the resulting estimates are statistically inefficient.

As a result, the regression imputation method has become a popular choice (Shao and Wang (2002)). This method builds a regression relationship between auxiliary information and the missing value, and then predicts the missing value accordingly. Parametric regression imputation methods have been studied thoroughly (Srivastava and Cater (1986); Shao and Wang (2002)), whereas semiparametric methods of this type are more recent, having been developed in the past decade. For example, Wang, Linton and Hardle (2004) considered partial linear models for imputations. Liang, Wang and Carroll (2007) studied the measurement errors in covariates. Zhao and Tang (2016) considered an imputation-based statistical inference method for partially linear quantile regression models with missing responses. With the MAR assumption, both methods can provide asymptotically unbiased estimates.

Similar methods have also been developed for NM. For example, Alho (1990) used a logistic regression model to describe the conditional response probability, which led to the maximum conditional likelihood estimation approach. The method requires one or more callbacks to nonrespondents. By assuming a parametric model for the response mechanism and a nonparametric model for the data distribution, Qin, Leung and Shao (2002) proposed a semiparametric likelihood estimation procedure to handle the nonignorable nonresponse problem. Assuming that the missing data mechanism is covariate-dependent, and that the propensity function can be properly specified, Qin, Shao and Zhang (2008) developed a regression imputation procedure that is efficient and robust against a regression model misspecification. Wang, Shao and Kim (2014) proposed an instrumental variable approach for identification and estimation in the case of a nonignorable nonresponse. Wang, Ding and Geng (2016) demonstrated the identifiability of the normal distribution under a monotone missing−data mechanism. They then extended this to normal mixture and t−mixture models with a nonmonotone missing−data mechanism.

Although the aforementioned imputation methods are helpful, they apply mainly to independent data. The assumption of independence basically indicates that different units live in isolated social environments, and thus do not

affect one another. This is obviously incorrect. In fact, most often, the sampled units live in the same social environment and are connected through various social relationships (e.g., friendship, follower−followee relationship). See, for example, Scott (1992), Wasserman and Faust (1994), Cohendet et al. (1998), and LeSage and Pace (2009) for several interesting discussions. Thus, we have a rather complex social network. Previous studies fail to account for this, mainly owing to a lack of network structure data, that is, data on inter-node social relationships. However, following the rapid growth of various social networks (e.g., Facebook, Twitter, QQ, Weibo, and WeChat), data on network structures are becoming increasingly available. Intuitively, socially connected units should be statistically correlated. Thus, observed responses from connected friends should provide valuable information on missing responses. This immediately leads to an extremely valuable prospect for imputation.

Classical imputation methods (as reviewed above) impute a unit's missing value based on its own characteristics. In contrast, with network structure information, we should be able to provide a more accurate imputation by borrowing information from connected friends. Specifically, for each sampled unit, we assume an interested response and a set of auxiliary variables. We assume a linear regression relationship between the response and the auxiliary variables (Srivastava and Cater (1986); Shao and Wang (2002)). Furthermore, we assume that sample units form a complicated social network. Thus, the residuals of units are dependent, and their dependence should be related to the network structure. To model such a dependence relationship, we adopt a spatial autoregression (SAR) model (Bronnenberg and Mahajan (2001); Lee, Liu and Lin (2010); Huang et al. (2016); Zhou et al. (2017)). This is one of the most typical models used for network dependence. Finally, we assume that auxiliary information is fully observed. Furthermore, conditional on the auxiliary information, the response is MAR.

With the MAR assumption, and using observed information only, we propose a partial likelihood estimation and develop the corresponding maximum partial likelihood estimator (MPLE). The estimator's consistency and asymptotic normality are established. Using the MPLE, we develop a novel regression imputation method. The proposed method utilizes both auxiliary information (always observed) and connected complete units (i.e., network information). Using the imputed data, we can compute the sample mean of the responses (both observed and imputed responses), which we show is consistent and asymptotically normal. Compared with the imputation method with auxiliary information only

(i.e., ignoring network information), the proposed estimator is statistically more efficient. Simulation studies are presented to demonstrate the finite−sample performance of the proposed method. Lastly, a real−data example about college student QQ users is discussed.

The rest of this paper is organized as follows. Section 2 presents the proposed methodology. We conduct numerical studies based on both simulated and real data sets in Section 3. Finally, Section 4 concludes the paper. We present the proofs of the theorems in the online Supplementary Material.

2. Methodology

2.1. Full data likelihood

Let $Y_i \in \mathbb{R}^1$ $(1 \le i \le N)$ be the response collected from the ith subject and $X_i = (X_{i1}, \ldots, X_{ip})^\top \in \mathbb{R}^p$ be the associated p-dimensional covariate. We model their regression relationship based on the following standard linear regression model:

$$Y_i = X_i^\top \beta + v_i, \tag{2.1}$$

where v_i is the residual and $\beta = (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p$ is the unknown regression coefficient. We define $\mathbb{Y} = (Y_1, \ldots, Y_N)^\top \in \mathbb{R}^N$ as the response vector, $\mathbb{X} = (X_1, \ldots, X_N)^\top \in \mathbb{R}^{N\times p}$ as the design matrix, and $\mathbb{V} = (v_1, \ldots, v_N)^\top \in \mathbb{R}^N$ as the residual vector. Define the theoretical R-Squared as $R^2 = \{1 - \mathrm{var}(\nu_i)/\mathrm{var}(Y_i)\} \times 100\%$.

We further assume that different subjects are connected through a network, which has an adjacency matrix given by $A = (a_{i_1 i_2}) \in \mathbb{R}^{N\times N}$, where $a_{i_1 i_2} = 1$ if there exists a relationship from i_1 to i_2 (e.g., user i_1 follows user i_2 on Twitter), and $a_{i_1 i_2} = 0$ otherwise. Obviously, the connected users are likely to be correlated with one another in terms of residuals. We model the correlation structure using the following popular spatial network regression model (Anselin (1988); Lee, Liu and Lin (2010)):

$$\mathbb{V} = \rho W \mathbb{V} + \mathcal{E}, \tag{2.2}$$

where ρ is the spatial autocorrelation coefficient, with $|\rho| < 1$. Recall that $\mathbb{V}$ is the noise vector in (2.1). In contrast, $\mathcal{E} = (\varepsilon_1, \ldots, \varepsilon_N)^\top \in \mathbb{R}^N$ is the noise vector in (2.2). To differentiate between the model errors, we refer to $\mathbb{V}$ as the residual vector and to $\mathcal{E}$ as the innovation vector. We assume that the components of the innovation vector $\mathcal{E}$ are mutually independent normal random variables with mean zero and variance σ^2. Furthermore, $W = (w_{i_1 i_2}) \in \mathbb{R}^{N\times N}$ is the so-called

spatial weighting matrix. Depending on the application, the definitions may vary. However, one popular definition can be given as $w_{i_1 i_2} = a_{i_1 i_2}/n_{i_1}$ and $n_{i_1} = \sum_{i_2} a_{i_1 i_2}$ (Anselin (1988)).

Note that (2.2) implies that each v_i consists of two parts. The first is the average value of its connected friends, multiplied by a coefficient ρ. We refer to this as the network effect. The second is a white noise $\mathcal{E}$. The advantage of the model is that it allows for network dependence, making it possible to use information from the node's connected friends. The disadvantage of the model is that its computation cost is significant, especially when the network is large. This is mainly because we need to compute the determinant of a high-dimensional matrix; see, for example, the log-likelihood function in (2.7).

From models (2.1) and (2.2), we know that $\mathbb{Y} - \mathbb{X}\beta = \mathbb{V} = (I - \rho W)^{-1}\mathcal{E}$, where $I \in \mathbb{R}^{N\times N}$ is an identity matrix. We immediately see that $\mathbb{V}$ follows a multivariate normal distribution with mean zero and covariance $\Sigma = \sigma^2(I - \rho W)^{-1}(I - \rho W^\top)^{-1}$. Thus, its log-likelihood function is given by

$$\ell_f^*(\rho, \sigma^2, \beta) = \frac{1}{2}\log\left|(I - \rho W^\top)(I - \rho W)\right| - \frac{N}{2}\log\sigma^2 - \frac{N}{2}\log(2\pi) + \frac{1}{2\sigma^2}(\mathbb{Y} - \mathbb{X}\beta)^\top(I - \rho W^\top)(I - \rho W)(\mathbb{Y} - \mathbb{X}\beta). \tag{2.3}$$

With fixed ρ, we can estimate β using an ideal estimator $\hat{\beta}_f = (\breve{\mathbb{X}}^\top\breve{\mathbb{X}})^{-1}(\breve{\mathbb{X}}^\top\breve{\mathbb{Y}})$, where $\breve{\mathbb{X}} = (I - \rho W)\mathbb{X}$ and $\breve{\mathbb{Y}} = (I - \rho W)\mathbb{Y}$. We consider $\hat{\beta}_f$ an ideal estimator because its computation involves an unknown parameter ρ. By replacing the unknown parameter β in (2.3) with $\hat{\beta}_f$, we arrive at the following profiled full data likelihood function: $\ell_f(\rho, \sigma^2) = \ell_f^*(\rho, \sigma^2, \hat{\beta}_f)$. By optimizing $\ell_f(\rho, \sigma^2)$ with respect to σ^2, we obtain an analytic estimator of σ^2 as $\hat{\sigma}_f^2 = (\mathbb{Y} - \mathbb{X}\hat{\beta}_f)^\top(I - \rho W^\top)(I - \rho W)(\mathbb{Y} - \mathbb{X}\hat{\beta}_f)/N$. Then, we obtain the profiled maximum likelihood estimator (MLE) of ρ as $\hat{\rho}_f = \arg\max_\rho \ell_f^*(\rho, \hat{\sigma}_f^2, \hat{\beta}_f)$. Here, the subscript f indicates that we obtain the MLE by considering the full data likelihood; that is, the response vector $\mathbb{Y}$ is fully observed. Under appropriate regularity conditions, Lee (2004) studied the asymptotic distribution of $(\hat{\beta}_f, \hat{\rho}_f, \hat{\sigma}_f^2)$ in a similar manner.

2.2. Incomplete data likelihood

In practice, the response vector $\mathbb{Y}$ is very often not completely observed. For example, consider a social network, comprising QQ (*www.qq.com*) users from the same university. Let Y_i be the self-reported natural age from the

ith user. Because many QQ users consider age a private matter and, thus, refuse to report Y_i publicly, we cannot observe a considerable portion of $\mathbb{Y}$. Consequently, we decompose $\mathbb{Y}$ into two parts. Without loss of generality, we write $\mathbb{Y} = (\mathbb{Y}_1^\top, \mathbb{Y}_2^\top)^\top$, where $\mathbb{Y}_1 = (Y_1, \ldots, Y_n)^\top \in \mathbb{R}^n$ is the observed response vector, and $\mathbb{Y}_2 = (Y_{n+1}, \ldots, Y_N)^\top \in \mathbb{R}^{N-n}$ is the unobserved vector. We can decompose other notation accordingly, yielding $\mathbb{V} = (\mathbb{V}_1^\top, \mathbb{V}_2^\top)^\top$, $\mathcal{E} = (\mathcal{E}_1^\top, \mathcal{E}_2^\top)^\top \in \mathbb{R}^N$, and $\mathbb{X} = (\mathbb{X}_1^\top, \mathbb{X}_2^\top)^\top$. Then, the adjacency matrix A and weighting matrix W can be partitioned as $A = (A_{11}, A_{12}; A_{21}, A_{22})$ and $W = (W_{11}, W_{12}; W_{21}, W_{22})$, respectively.

Before continuing, we need to define the missing data mechanism of $\mathbb{Y}_2$. First, for an arbitrary subject i, the binary indicator δ_i takes the value one if Y_i is observed, and zero otherwise. For consistency with the notation defined in the previous subsections, we see immediately that the subjects are appropriately ordered such that $\delta_i = 1$ for $1 \leq i \leq n$, and $\delta_i = 0$ for $n < i \leq N$. In a network context, it is likely that the missingness of a response is affected by both the individual and his/her connected friends. Thus, it is realistic to assume that the conditional missingness probability depends on both the covariate $\mathbb{X}$ and the response $\mathbb{Y}$. Here, $\mathbb{Y}$ is the entire response vector. It collects responses from both the target node and its connected friends, regardless of whether or not $\mathbb{Y}$ is observed. Under a regression setup, this is equivalent to assuming that the missingness probability depends on both $\mathbb{X}$ and $\mathbb{V}$. This leads to the following assumption:

$$P(\delta_i = 1|\mathbb{X}, \mathbb{Y}, A) = P(\delta_i = 1|\mathbb{X}, \mathbb{X}\beta + \mathbb{V}) = P(\delta_i = 1|\mathbb{X}, \mathbb{V}). \tag{2.4}$$

By assuming that the probability function is smooth in $\mathbb{V}$, we can conduct a Taylor−type expansion about $\mathbb{V}$ at the point $\mathbb{V} = 0$. This leads to

$$P(\delta_i = 1|\mathbb{X}, \mathbb{V}) = P(\delta_i = 1|\mathbb{X}, \mathbb{V})|_{\mathbb{V}=0} + \frac{dP(\delta_i = 1|\mathbb{X}, \mathbb{V})}{d\mathbb{V}}|_{\mathbb{V}=0} \cdot \mathbb{V} + o(\mathbb{V}). \tag{2.5}$$

Note that $\mathbb{V}$ is not fully observed in our data set. As a result, we ignore the higher−order terms involving $\mathbb{V}$. Consequently, only the intercept term is kept, and serves as the first−order approximation to the true missingness probability. This approximation leads to

$$P(\delta_i = 1|\mathbb{X}, \mathbb{Y}, A) \approx P(\delta_i = 1|X_i). \tag{2.6}$$

Interestingly, this first−order approximation is free of the residual vector $\mathbb{V}$, and

thus is free of the response vector $\mathbb{Y}$. This is an MAR assumption (Rubin (1987)), which means the classical MAR assumption can be viewed as a first−order approximation of the true missingness data mechanism. Although this is a practical approximation, our simulation studies show that it is helpful.

Because $\mathbb{Y}_2$ is not observed, the estimator based on full data is no longer computable. Thus, we can use the observed data $\mathbb{Y}_1$ only. Because $\mathbb{V}$ and, thus, $\mathbb{Y}$ are jointly normal, the marginal distribution of $\mathbb{Y}_1$ is also normal, with mean $\mathbb{X}_1\beta$ and residual $\mathcal{E}_1$. Thus, to derive the marginal likelihood of $\mathbb{Y}_1$, we have only to specify the marginal covariance matrix of $\mathcal{E}_1$. We then consider the partition of the whole covariance Σ, which leads to $\Sigma = (\Sigma_{11}, \Sigma_{12}; \Sigma_{21}, \Sigma_{22})$. Here, Σ_{11} is the marginal covariance of $\mathcal{E}_1$. Recall that $\Sigma^{-1} = \sigma^{-2}(I-\rho W^\top)(I-\rho W)$. Then, write $(I-\rho W^\top)(I-\rho W)$ by $\Upsilon(\rho) = \Big[\Upsilon_{11}(\rho), \Upsilon_{12}(\rho); \Upsilon_{21}(\rho), \Upsilon_{22}(\rho)\Big]$, with

$$\begin{aligned}
\Upsilon_{11}(\rho) &= I_{11} - \rho(W_{11}^\top + W_{11}) + \rho^2(W_{11}^\top W_{11} + W_{21}^\top W_{21}),\\
\Upsilon_{12}(\rho) &= -\rho(W_{12} + W_{21}^\top) + \rho^2(W_{11}^\top W_{12} + W_{21}^\top W_{22}),\\
\Upsilon_{21}(\rho) &= -\rho(W_{12}^\top + W_{21}) + \rho^2(W_{12}^\top W_{11} + W_{22}^\top W_{21}),\\
\Upsilon_{22}(\rho) &= I_{22} - \rho(W_{22}^\top + W_{22}) + \rho^2(W_{12}^\top W_{12} + W_{22}^\top W_{22}).
\end{aligned}$$

We now have $\Sigma_{11} = \sigma^2\Big[\Upsilon_{11}(\rho) - \Upsilon_{12}(\rho)\Upsilon_{22}^{-1}(\rho)\Upsilon_{21}(\rho)\Big]^{-1}$. This leads to the following log partial likelihood based on incomplete data:

$$\begin{aligned}
\ell(\beta, \rho, \sigma^2) = &\frac{1}{2}\log|\Omega(\rho)| - \frac{n}{2}\log\sigma^2\\
&- \frac{1}{2\sigma^2}(\mathbb{Y}_1 - \mathbb{X}_1\beta)^\top\Omega(\rho)(\mathbb{Y}_1 - \mathbb{X}_1\beta) - \frac{n}{2}\log(2\pi), \qquad (2.7)
\end{aligned}$$

where $\Omega(\rho) = \Upsilon_{11} - \Upsilon_{12}\Upsilon_{22}^{-1}\Upsilon_{21}$. Here, by adopting the matrix $\Omega(\rho)$, we extend the full data likelihood (2.3) to the incomplete data likelihood (2.7).

To estimate the parameter $\theta = (\beta^\top, \rho, \sigma^2)^\top$, we can adopt a profiled estimation approach. We first fix ρ and σ^2 and optimize (2.7) with respect to β to obtain an estimator of β: $\hat\beta(\rho) = \Big[\mathbb{X}_1^\top\Omega(\rho)\mathbb{X}_1\Big]^{-1}\mathbb{X}_1^\top\Omega(\rho)\mathbb{Y}_1$. Then, by replacing β in (2.7) with $\hat\beta(\rho)$, we obtain the profiled objective function for (σ^2, ρ),

$$\begin{aligned}
\ell(\hat\beta, \sigma^2, \rho) = &-\frac{n}{2}\log(2\pi) + \frac{1}{2}\log|\Omega(\rho)| - \frac{n}{2}\log\sigma^2\\
&- \frac{1}{2\sigma^2}(\mathbb{Y}_1 - \mathbb{X}_1\hat\beta(\rho))^\top\Omega(\rho)(\mathbb{Y}_1 - \mathbb{X}_1\hat\beta(\rho)).
\end{aligned}$$

Next, we optimize $\ell(\hat{\beta}, \sigma^2, \rho)$ with respect to σ^2. This leads to the following estimator of σ^2: $\hat{\sigma}^2(\rho) = n^{-1}(\mathbb{Y}_1 - \mathbb{X}_1\hat{\beta})^\top \Omega(\rho)(\mathbb{Y}_1 - \mathbb{X}_1\hat{\beta})$. By applying $\hat{\sigma}^2(\rho)$ and $\hat{\beta}(\rho)$ back to (2.7), we obtain the profiled objective function $\ell(\rho) = \ell(\rho, \hat{\sigma}^2, \hat{\beta})$. This leads to the final estimator, $\hat{\rho} = \text{argmax}\ell(\rho)$. Note that ρ is a scaler in $(-1, 1)$, we can thus compute $\hat{\rho}$ using a grid search, which is computationally stable. We apply $\hat{\rho}$ back to the formula of $\hat{\beta}(\rho)$ and $\hat{\sigma}^2(\rho)$ to obtain the estimators for β and σ^2 as $\hat{\beta} = \hat{\beta}(\hat{\rho})$ and $\hat{\sigma}^2 = \hat{\sigma}^2(\hat{\rho})$, respectively.

2.3. Network-based imputation method

We next examine how to impute the missing responses in $\mathbb{Y}_2$, with the help of $\hat{\theta} = (\hat{\beta}^\top, \hat{\rho}, \hat{\sigma}^2)^\top$. From model (2.1), we have $\mathbb{Y}_2 = \mathbb{X}_2\beta + \mathbb{V}_2$, where $\mathbb{X}_2$ is observed, and we can consistently estimate β using $\hat{\beta}$ based on observed complete data. Thus, imputing $\mathbb{V}_2$ based on the information given by $\mathbb{X}_1$, $\mathbb{Y}_1$, and W is the key step to considering the network information. On the other hand, $\mathbb{V}_1$ and $\mathbb{V}_2$ are correlated, and their correlation structure is fully determined by the network weighting matrix W. This motivates us to investigate the conditional distribution of $\mathbb{V}_2$, given $\mathbb{V}_1$. Because $\mathbb{V} = (\mathbb{V}_1^\top, \mathbb{V}_2^\top)^\top$ is jointly normal, we can obtain the conditional distribution $\mathbb{V}_2|\mathbb{V}_1$, as stated in the following proposition.

Proposition 1. *Assuming model* (2.2)*, the conditional distribution of* $\mathbb{V}_2$*, given* $\mathbb{V}_1$*, is multivariate normal, with mean* $\Sigma_{21}\Sigma_{11}^{-1}\mathbb{V}_1$ *and covariance* $\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$.

From proposition 1, we know that $\mathrm{E}(\mathbb{V}_2|\mathbb{V}_1) = \Sigma_{21}\Sigma_{11}^{-1}\mathbb{V}_1$, where we can accurately approximate $\mathbb{V}_1$ using $\hat{\mathbb{V}}_1 = \mathbb{Y}_1 - \mathbb{X}_1^\top\hat{\beta}$. This leads to the imputed response $\widehat{\mathbb{Y}}_2 = (\hat{Y}_{n+l}, l = 1, \ldots, N-n)$, with $\hat{Y}_{n+l} = X_{n+l}^\top\hat{\beta} + \hat{V}_{n+l}$, where $\hat{V}_{n+l} = \Sigma_{21,l}(\hat{\rho})\Sigma_{11}^{-1}(\hat{\rho})\hat{\mathbb{V}}_1 = -\Upsilon_{22,l}^{-1}(\hat{\rho})\Upsilon_{21}(\hat{\rho})\hat{\mathbb{V}}_1$, and $\Sigma_{21,l}^{-1}(\hat{\rho})$ and $\Upsilon_{22,l}^{-1}(\hat{\rho})$ are the lth rows of $\Sigma_{21}^{-1}(\hat{\rho})$ and $\Upsilon_{22}^{-1}(\hat{\rho})$, respectively. Using the imputed $\mathbb{Y}_2$, we estimate the mean $E(Y_i) = \mu$ by

$$\hat{\mu}_N = N^{-1}\left(\sum_{i=1}^{n} Y_i + \sum_{i=n+1}^{N} \hat{Y}_i\right),$$

where we use the subscript N in $\hat{\mu}_N$ to emphasize that this estimator includes network information (i.e., W). Thus, we refer to $\widehat{\mathbb{Y}}_2$ as a network-based (NB) imputed response. We also impute $\mathbb{Y}_2$ using $\widetilde{\mathbb{Y}}_2 = \mathbb{X}_2\tilde{\beta}$, where $\tilde{\beta} = (\mathbb{X}_1^\top\mathbb{X})^{-1}(\mathbb{X}_1^\top\mathbb{Y}_1)$. Clearly, $\tilde{\beta}$ is a standard ordinary least squares estimate, computed with complete units only and ignoring all network information. For convenience, we refer to $\widetilde{\mathbb{Y}}_2$ as a regression-based (RB) imputed response. As we show subsequently, the NB estimate $\widehat{\mathbb{Y}}_2$ is much more accurate than the RB estimate $\widetilde{\mathbb{Y}}_2$; see Theorem 2 in

the next subsection and Section 3 for numerical evidence.

2.4. Technical conditions

We next consider the theoretical properties of the proposed estimators, including the asymptotic distribution of various estimators. To this end, we write

$$\begin{aligned}\dot{\Omega}(\rho) =& \frac{d\Omega(\rho)}{d\rho} \\ =& \Upsilon_{12}(\rho)\Upsilon_{22}^{-1}(\rho)\Big[2\rho(W_{12}^{\top}W_{12}+W_{22}^{\top}W_{22})-(W_{22}+W_{22}^{\top})\Big]\Upsilon_{22}^{-1}(\rho)\Upsilon_{21}(\rho) \\ &+\Big[W_{12}+W_{21}^{\top}-2\rho(W_{11}^{\top}W_{12}+W_{21}^{\top}W_{22})\Big]\Upsilon_{22}^{-1}(\rho)\Upsilon_{21}(\rho) \\ &+\Upsilon_{12}(\rho)\Upsilon_{22}^{-1}(\rho)\Big[W_{12}^{\top}+W_{21}-2\rho(W_{12}^{\top}W_{11}+W_{22}^{\top}W_{21})\Big] \\ &-(W_{11}^{\top}+W_{11})+2\rho(W_{11}^{\top}W_{11}+W_{21}^{\top}W_{21}).\end{aligned}$$

For an arbitrary matrix H, we denote the product HH by H^2. We then have the following technical conditions:

(C1) $\Omega(\rho)$ is a positive−definite matrix, for any $\rho > 0$. There exists another positive−definite matrix Λ_{11} and a finite positive constant Λ_{22}, such that $(n\sigma^2)^{-1}\mathbb{X}_1^{\top}\Omega(\rho)\mathbb{X}_1 \to \Lambda_{11}$ and $n^{-1}\mathrm{tr}\Big[\{\Omega^{-1}(\rho)\dot{\Omega}(\rho)\}^2\Big] \to \Lambda_{22}$.

(C2) Constants $c_{\min}$ and $c_{\max}$ exist, such that $\lim_{n\to\infty} 2n^{-1}tr\Big[\{B(\rho)\Omega^{-1}(\rho)\}^2\Big] > c_{\min}$ and $\lambda_{\max}\Big\{B(\rho)\Omega^{-1}(\rho)\Big\} \le c_{\max}$, for $B = \dot{\Omega}(\rho)$ and $\ddot{\Omega}(\rho)$.

(C3) A constant $0 < r < 1$ exists, such that $\lim_{n\to\infty} n/N = r$.

Both (C1) and (C2) are essentially moment conditions involving the design matrix $\mathbb{X}_1$, network structure W, and unknown parameter ρ. However, we argue that both (C1) and (C2) are fairly reasonable conditions. To gain an intuitive understanding, we conduct Taylor's expansion for $\Omega(\rho)$, $\dot{\Omega}(\rho)$, and $\ddot{\Omega}(\rho)$ around ρ. In many reported real applications, the empirically estimated ρ is small. This enables us to approximate $\Omega(\rho)$ using $\Omega(0)$. That is, $\Omega(\rho) \approx \Omega(0)$. Similar approximations hold for $\dot{\Omega}(\rho)$ and $\ddot{\Omega}(\rho)$. However, we do not use this simple approximation to develop the asymptotic theory of the proposed method, but instead use it only to gain a quick understanding of the technical conditions (C1) and (C2).

First, we consider (C1). Because $\Omega(\rho) \approx \Omega(0) = I_{11}$, we have $(n\sigma^2)^{-}\mathbb{X}_1^{\top}\Omega(\rho)$ $\mathbb{X}_1 \approx (n\sigma^2)^{-1}\mathbb{X}_1^{\top}\mathbb{X}_1$. Thus, the first condition in (C1) is equivalent to $(n\sigma^2)^{-1}\mathbb{X}_1^{\top}$

$\mathbb{X}_1 \to \Lambda_{11}$, approximately, which is just a law of large numbers (LLN)-type assumption. This condition can be satisfied easily if the predictors from different nodes are independent or are weakly dependent. As with $\rho \approx 0$, we can verify that $\dot{\Omega}(\rho) \approx \dot{\Omega}(0) = -W_{11}^\top - W_{11}$. Because $\Omega^{-1}(\rho) \approx \Omega^{-1}(0) = I_{11}$, the second condition in (C1) is approximately equivalent to $n^{-1}tr(W_{11}^\top + W_{11})^2 \to \Lambda_{22}$. When W_{11} is symmetrical, this condition is approximately equivalent to $n^{-1}4\sum_{i_1=1}^n \sum_{i_2=1}^n w_{i_1 i_2}^2 \to \Lambda_{22}$. This is reasonable, because $w_{i_1 i_2}$, for $i_2 = 1, \ldots, n$, are all positive and sum to one. We can use similar arguments for the conditions in (C2). To summarize, although both (C1) and (C2) seem quite complicated in form, their conditions are fairly reasonable.

2.5. Theoretical properties

Using the technical conditions given in the previous subsection, we investigate the asymptotic properties of the proposed estimators.

Theorem 1. *Assuming models* (2.1) *and* (2.2) *and conditions* (C1)–(C2)*, we have* (1) $(\hat{\beta}, \hat{\rho}, \hat{\sigma}^2)$ *is a consistent estimator of* (β, ρ, σ^2)*, and* $\hat{\beta}$ *and* $(\hat{\rho}, \hat{\sigma}^2)$ *are asymptotically independent; and* (2) $\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \Lambda_{11}^{-1})$*,* $\sqrt{n}(\hat{\rho} - \rho, \hat{\sigma}^2 - \sigma^2)^\top \xrightarrow{d} N(0, \Lambda_2^{-1})$*, where* $\xrightarrow{d}$ *represents convergence in the distribution. In addition,* $\Lambda_2 = (\Lambda_{22}, \Lambda_{23}; \Lambda_{32}, \Lambda_{33})$*,* $\Lambda_{22} = \lim_{n\to\infty}(2n)^{-1}tr[(\Omega^{-1}(\rho)\dot{\Omega}(\rho))^2]$*, and* $\Lambda_{23} = \Lambda_{32} = \lim_{n\to\infty} -(2n\sigma^4)^{-1}E[(\mathbb{Y}_1 - \mathbb{X}_1\beta)^\top \dot{\Omega}(\rho)(\mathbb{Y}_1 - \mathbb{X}_1\beta)], \Lambda_{33} = (2\sigma^4)^{-1}$.

From Theorem 1, we know that the proposed estimators are all $\sqrt{n}$-consistent and asymptotically normal. Note that $\hat{\beta}$ and $(\hat{\rho}, \hat{\sigma}^2)$ are asymptotically independent. This is as expected, because the information for $\hat{\beta}$ mainly comes from the mean, whereas that for $(\hat{\rho}, \hat{\sigma}^2)$ comes from the covariance.

Remark 1. By Little and Rubin (2002) and under the MAR assumption, we expect the parameter to be estimated consistently from observed data, in theory. Theorem 1 formally confirms this expectation. In addition, Theorem 1 provides an analytically tractable formula for the asymptotic variance, which makes the corresponding variance estimation easy.

Theorem 2. *Assuming models* (2.1) *and* (2.2)*, we know that, for* $l = 1, \ldots, N - n$*,* $\hat{Y}_{n+l} - Y_{n+l} \xrightarrow{d} N(0, \Phi)$*, where* $\Phi = EV_{n+l}^2 - E[E(V_{n+l}|\mathbb{V}_1)]^2$.

Note that if we ignore the network information by taking $\rho = 0$, and then estimate Y_{n+l} using $\tilde{Y}_{n+l} = X_{n+l}^\top \tilde{\beta}$, we can verify that $\tilde{Y}_{n+l} - Y_{n+l} \to_d = -V_{n+l}$, with a variance of EV_{n+l}^2. By Theorem 2, we know that the NB imputation is likely to

be more efficient than its RB counterpart if $\mathrm{E}V_{n+l}^2 - \Phi = \mathrm{E}\{\mathrm{E}(V_{n+l}|\mathbb{V}_1)\}^2 > 0$. However, if $\rho = 0$, we then know $\mathrm{E}(V_{n+l}|\mathbb{V}_1) = 0$. In this case, $\mathrm{E}V_{n+l}^2 = \Phi$. Thus, the efficiencies of both methods become identical. This suggests that the superiority of the NB imputation is highly dependent on ρ.

Theorem 3. *Assuming models* (2.1) *and* (2.2), *we have* $\sqrt{n}(\hat{\mu}_N - \mu) \xrightarrow{d} N(0, \phi)$, *with* $\phi = Var(X_1\beta) + \lim_{n\to\infty}(r/n)\Big[1_n^\top - 1_{N-n}^\top \Upsilon_{22}^{-1}(\rho)\Upsilon_{21}(\rho) + B\Big]\Sigma_{11}\Big[1_n^\top - 1_{N-n}^\top \Upsilon_{22}^{-1}(\rho)\Upsilon_{21}(\rho) + B\Big]^\top$ *and* $B = \lim_{n\to\infty} 1_{N-n}^\top \Big[\mathbb{X}_2 + \Upsilon_{22}^{-1}(\rho)\Upsilon_{21}(\rho)\mathbb{X}_1\Big] \cdot \Big[\mathbb{X}_1^\top \Omega(\rho)\mathbb{X}_1\Big]^{-1} \mathbb{X}_1^\top \Omega(\rho)$.

3. Numerical Studies

3.1. simulation studies

To demonstrate the finite−sample performance of the proposed method, we present a number of simulation studies. Specifically, we set $N = 500$ and fix $p = 2$. For each i, we set $X_{i1} = 1$ and simulate X_{i2} from a standard normal distribution. The corresponding regression parameters are given by $\beta = (1, 1)^\top \in \mathbb{R}^2$. Furthermore, we consider $\rho = 0$, 0.3, and 0.5. For each ρ, we generate ε_i from a standard independent normal distribution. Finally, we generate the adjacency matrix A, as follows. We first generate N independent random variables from $N(3, 1)$, denoted by E_i, with $1 \le i \le N$. Next, for every node pair (i, j), with $i \neq j$, we define $a_{ij} = 1$ if $|i - j| \le E_i$, and zero otherwise. Further, we define $a_{ij} = 0$ whenever $i = j$. This leads to the adjacency matrix A and its row-normalized weighting matrix W. We then generate $\mathbb{Y}$ according to (2.1) and (2.2).

We consider two cases of the missing data mechanism: the MAR and the NM. In the MAR case, for every node i, we set its response as missing with probability $\exp(\gamma + X_{i1} + X_{i2})/\{1 + \exp(\gamma + X_{i1} + X_{i2})\}$, where γ is a tuning parameter controlling the level of missing data. We consider three different values of γ: $\gamma = -2$, -1, and 0.5. In our model, the three values yield missing rates of approximately 20%, 50%, and 80%, respectively, on average. For each γ specification, the experiment is randomly replicated $M = 1,000$ times.

Let $n^{(r)}$ be the number of complete units generated in the rth simulation replication, and their average be given by $\bar{n} = M^{-1}\sum_r n^{(r)}$. For each parameter, for example, β_1, let $\hat{\beta}_1^{(r)}$ be the corresponding estimator obtained in the rth simulation replication. Then, we estimate the true variance (VAR) as $\mathrm{VAR} = M^{-1}\sum_r(\hat{\beta}_1^{(r)} - \bar{\beta}_1)^2$, with $\bar{\beta}_1 = M^{-1}\sum_r \hat{\beta}_1^{(r)}$. From Theorem 1,

we analytically state the asymptotic variance of $\hat{\beta}_1$. This enables us to provide an estimator for the true VAR by replacing the unknown quantities in $\Lambda_{11}^{-1}(2,2)$ with their estimates. We denote the estimate by $\widehat{\mathrm{VAR}}^{(r)}$, with a mean of $\widehat{\mathrm{VAR}} = M^{-1}\sum_r \widehat{\mathrm{VAR}}^{(r)}$. Next, we construct a 95% confidence interval for β_1 as $\mathrm{CI}^{(r)} = (\hat{\beta}_1^{(r)} - \widehat{\mathrm{SE}}^{(r)} z_{1-\alpha/2}, \hat{\beta}_1^{(r)} + \widehat{\mathrm{SE}}^{(r)} z_{1-\alpha/2})$, where $\widehat{\mathrm{SE}}^{(r)} = \{\widehat{\mathrm{VAR}}^{(r)}\}^{1/2}$, and z_α represents the αth lower quantile of a standard normal distribution. We then evaluate their coverage probability (CP) as $\mathrm{CP} = M^{-1}\sum_r I(\beta_1 \in \mathrm{CI}^{(r)})$. Other estimates (i.e., $\hat{\beta}_0$,$\hat{\rho}$, $\hat{\sigma}^2$, and $\hat{\mu}$) are summarized similarly. We also evaluate the forecasting error (FE) of $\widehat{\mathbb{Y}}_2$ by $\mathrm{FE}_1 = M^{-1}\sum \|\widehat{\mathbb{Y}}_2^{(r)} - \mathbb{Y}_2^{(r)}\|^2/(N - n^{(r)})\}$, where $\mathbb{Y}_2^{(r)}$ is the missing response vector generated in the rth replication. Then, we can evaluate $\widetilde{\mathbb{Y}}_2$ in a similar manner, as $\mathrm{FE}_2 = M^{-1}\sum \|\widetilde{\mathbb{Y}}_2^{(r)} - \mathbb{Y}_2^{(r)}\|^2/(N - n^{(r)})\}$. The relative improvement margin is then given by $\mathrm{RIM} = (1 - \mathrm{FE}_1/\mathrm{FE}_2) \times 100\%$.

From Table 1, in the MAR case, we find that as the sample size of the complete units (i.e., n) increases, the performance of all estimators improves, with VAR steadily approaching zero. This suggests that the proposed estimators are consistent. Moreover, we find that the VAR estimate (i.e., $\widehat{\mathrm{VAR}}$) approximates VAR relatively well. This confirms that our asymptotic results given in Theorems 1 and 3 should be correct. This is further confirmed by the reported CP values, which are fairly close to the nominal level of 95%.

From Table 2, we find that the performance of the NB estimator depends on the parameter ρ. Specifically, for $\rho = 0$, the performance of the NB estimator and the RB estimator are nearly identical, with the RB estimator performing slightly better (probably owing to its simplicity). However, for $\rho = 0.3$ and 0.5, which indicate the existence of network dependence, the forecasting accuracy of the NB estimator is considerably better than that of the RB estimator. Additionally, the forecasting accuracy improves as ρ increases. The RIM could be as large as 22.84%. This corroborates our theoretical findings in Theorem 2 quite well.

To assess the effect of the MAR assumption, we conduct a simulation study with an NM missing data mechanism. In this case, for every node i, we set its response as missing with probability $\exp(\gamma + X_{i1} + X_{i2} + 0.1 \cdot V_i)/\{1 + \exp(\gamma + X_{i1} + X_{i2} + 0.1 \cdot V_i)\}$. Accordingly, neither the NB nor the RB estimators is still consistent. As a result, the empirical performance deteriorates, as shown in Table 3. The reported CP values for the intercept term β_0 could be far below the nominal level 95%. This is particularly true when $R^2 = 10\%$. However, it seems that the results improve as R^2 increases. In this case, the residual $\mathbb{V}$ becomes smaller, enabling the first−order approximation (2.4), (2.5), and (2.6) to work

better. This corroborates our theoretical finding.

3.2. Real−data example

To demonstrate our method, we present an interesting real−data example. The data set is a sampled subnetwork of QQ (www.qq.com), which is perhaps the largest social network instant messaging (IM) software in mainland China, with more than 800 million users. We obtain the data set using a convenient snowball−type sampling method on a university campus. The objective is to demonstrate QQ network dependence among college students. We start with eight convenient QQ users, who are college students. We next collect their QQ friends. We collect only those QQ users whose self-reported age is missing or between 18 and 25 years old in order to ensure the sampled QQ users are college students, or at least of a similar age. This forms a sample of $N = 396$ QQ users. Nothe that the sampling method means that we expect the sample to be biased if the target is the whole QQ population. However, if we consider the college student users in the intended university campus as the population, we conjecture that the bias could be considerably reduced.

For each QQ user, we take the user's natural age as the response. QQ users have the right to decide whether to disclose their age. Some might refuse to do so owing to privacy concerns. This leads to a considerable portion of users showing no age. In our data set, the number of complete units is $n = 332$; this leaves $N-n = 64$ units with missing responses, accounting for about $(N-n)/N = 16\%$ of the total sample size.

For two arbitrary QQ users (i and j), we define $a_{ij} = a_{ji} = 1$ if they are friends, and zero otherwise. This leads to the adjacency matrix A and its row-normalized weighting matrix W. Similarly, for each QQ user i, we collect the following covariate data: $X_1 = 1$ (intercept); X_2 (gender); X_3 (QQ age, i.e. how long the user has been using QQ; this is different to the user's natural age); X_4 (QQ grade; this is a comprehensive measure for the user's QQ age and activeness), X_5 (total number of photos posted in the user's QQ space; QQ space is a Facebook-type personal QQ homepage); X_6 (total number of comments posted in the user's QQ space); X_7 (total number of articles posted in the user's QQ space); and X_8 (total number of messages left by the user's QQ friends in the QQ space). The observed age ranges from 18 to 25, with mean 23.24, median 23, and standard deviation 1.20. Before conducting a formal analysis, we standardize all quantitative covariates so that they have mean zero and variance one.

We then apply the proposed method to the data set to obtain the estimated

Table 1. Detailed Results from 1,000 Simulation Replications.

	$\rho = 0$				$\rho = 0.3$				$\rho = 0.5$			
	$\bar{n}$	VAR	$\widehat{\text{VAR}}$	CP(%)	$\bar{n}$	VAR	$\widehat{\text{VAR}}$	CP(%)	$\bar{n}$	VAR	$\widehat{\text{VAR}}$	CP(%)
	103	0.01496	0.01558	94.2	103	0.01751	0.01780	94.1	103	0.02299	0.02317	94.5
β_0	250	0.00520	0.00500	94.6	250	0.00746	0.00688	93.8	250	0.01169	0.01089	93.9
	389	0.00278	0.00282	94.8	389	0.00496	0.00472	94.3	389	0.00903	0.00862	94.3
	103	0.01223	0.01153	94.0	103	0.01300	0.01199	93.6	103	0.01453	0.01320	92.7
β_1	250	0.00495	0.00481	94.5	250	0.00502	0.00488	94.5	250	0.00516	0.00502	94.2
	389	0.00313	0.00290	94.3	389	0.00312	0.00288	94.1	389	0.00307	0.00285	94.7
	103	0.01863	0.09080	93.1	103	0.03382	0.05702	93.8	103	0.02433	0.02543	94.5
ρ	250	0.00413	0.01711	96.5	250	0.00915	0.00968	96.1	250	0.00483	0.00478	95.4
	389	0.00218	0.00733	96.7	389	0.00488	0.00464	94.6	389	0.00277	0.00271	95.4
	103	0.02163	0.02019	91.2	103	0.02536	0.02451	90.5	103	0.03278	0.03187	90.8
σ^2	250	0.00777	0.00789	92.9	250	0.00839	0.00858	93.2	250	0.00937	0.00955	93.2
	389	0.00511	0.00507	93.7	389	0.00519	0.00524	93.9	389	0.00535	0.00550	94.8

Table 2. Comparison of Prediction Errors for Different ρ.

	$\bar{n}$	PE_1	PE_2	RIM (%)
$\rho = 0$	103	1.03334	1.02520	-0.79383
	250	1.01052	1.00705	-0.34444
	389	1.00213	0.99975	-0.23781
$\rho = 0.3$	103	1.08412	1.09729	1.20035
	250	1.02930	1.07923	4.62681
	389	0.99572	1.07061	6.99538
$\rho = 0.5$	103	1.20045	1.30298	7.86864
	250	1.05946	1.28261	17.39817
	389	0.98113	1.27153	22.83846

Table 3. Detailed Simulation Results for NM with $\rho = 0.3$.

		VAR	$\widehat{\text{VAR}}$	CP
$R^2 = 10\%$	β_0	0.043285	0.041870	0.82
	β_1	0.027782	0.025787	0.86
	ρ	0.005208	0.004879	0.95
	σ^2	0.429914	0.416175	0.92
$R^2 = 20\%$	β_0	0.019356	0.018806	0.89
	β_1	0.012360	0.011533	0.91
	ρ	0.005285	0.004843	0.94
	σ^2	0.087596	0.083071	0.94
$R^2 = 50\%$	β_0	0.004895	0.004732	0.93
	β_1	0.003110	0.002893	0.93
	ρ	0.005431	0.004823	0.93
	σ^2	0.005392	0.005229	0.94

Table 4. Results: Analysis of the QQ Data Set.

Variable	Name	Estimate	SE	p-Value
X_1	Intercept	23.226	0.1448	**0.000**
X_2	Gender	-0.04	0.1239	0.773
X_3	QQ age	0.27	0.0801	**0.001**
X_4	QQ Grade	0.06	0.0709	0.396
X_5	Photos	0.10	0.0683	0.152
X_6	Comments	-0.20	0.0677	**0.003**
X_7	Articles	-0.01	0.0591	0.910
X_8	Messages	-0.13	0.0633	**0.038**

results shown in Table 4. From Table 4, we find four estimates to be statistically significant at the 5% level: X_1 (the intercept), X_3 (QQ age), X_6 (total number of comments) and X_8 (total number of messages). In addition, we find that ρ

Table 5. Results: Analysis of the QQ Data Set from Complete Case.

Variable	Name	Estimate	SE	p-Value
X_1	Intercept	23.30	0.0889	**0.000**
X_2	Gender	-0.05	0.1340	0.714
X_3	QQ age	0.35	0.0817	**0.000**
X_4	QQ Grade	0.08	0.0739	0.264
X_5	Photos	0.17	0.0726	**0.022**
X_6	Comments	-0.23	0.0720	**0.002**
X_7	Articles	-0.01	0.0626	0.887
X_8	Messages	-0.23	0.0664	**0.001**

is estimated to be 0.59 with $\widehat{\text{SE}}$=0.10. The resulting p-value is less than 1%, suggesting that even after controlling for the effects of the aforementioned covariates, a positive correlation still exists between QQ friends in terms of their natural age. We also include a complete case analysis, for comparison purpose. It shows that the corresponding p-value of the model is nearly zero, which indicates the significance of the model. Table 5 reports the regression coefficient estimation from the complete−case analysis. The resulting root mean prediction error is 1.12.

Then, we can impute the missing responses of the 64 incomplete units using both the NB and RB methods. Because the missing responses of a real data set are not observed, we cannot accurately evaluate the forecasting error of the NB and RB methods. To overcome this difficulty, we focus on the subnetwork generated by the complete units. This leads to a subnetwork size of 332, with both responses and covariates observed. Next, let i be an arbitrarily selected node from this subnetwork. We then randomly set Y_i as missing, with probability $\exp(1+W_X)/(1+\exp(1+W_X))$, with $W_X = 0.5\cdot X_{i1}+0.1\cdot X_{i2}+0.2\cdot X_{i3}+0.3\cdot X_{i4}+0.4\cdot X_{i5}+0.3\cdot X_{i6}+0.2\cdot X_{i7}+0.1\cdot X_{i8}$. This leads to approximately 20% incomplete units. Then, we impute the missing responses using either the NB or the RB method, and compare the imputed values against the true responses, similarly to the simulation study. The experiment is replicated 1,000 times. The average root mean prediction error (ARMPE) is computed. The ARMPE value for NB is 1.10, and for RB is 1.16, with a RIM of 10.09%.

4. Conclusion

In this work, we develop a network-based imputation method to analyze missing data in a network. The proposed method explores information from

both regression and network dependence points of view. This yields a consistent estimator for $E(Y)$. However, it can not be used to estimate higher-order statistics, such as $E(Y^2)$. Thus, the proposed imputation method can be adapted only to estimate $E(Y)$, and should be extended accordingly. However, estimating higher−order moments for network data with SAR-type dependence is an open question, which we leave to future work.

Supplementary Material

The Supplementary Material contains detailed proofs of Theorem 1, Theorem 2 and Theorem 3.

Acknowledgements

Zhimeng Sun's(sunzhimeng99@126.com) research is supported by the National Natural Science Foundation of China (No.11871488,11301561), National Statistical Science Research Project(No.2018LZ28), Youth Talent Development Support Program (No.QYP1810), Social Survey and Research Database Construction Project, Double First−class Construction Project at the Central University of Finance and Economics. Hansheng Wang's(hansheng@pku.edu.cn) research is supported in part by the National Natural Science Foundation of China (NO.11131002, 11271032), the Business Intelligence Research Center at Peking University, and the Center for Statistical Science at Peking University.

References

Alho (1990). Adjusting for nonresponse bias using logistic regression. *Biometrika* **77**, 617–624.

Anselin, L. (1988). Spatial econometrics: methods and models. *Kluwer Academic* **20**, 284–299.

Bronnenberg, B. J. and Mahajan, V. (2001). Unobserved retailer behavior in multimarket data: Joint spatial dependence in marketing shares and promotion variables. *Marketing Science* **20**, 284–299.

Cohendet, P., Llerena, P., Stahn, H. and Umbhauer, G. (1998). *The Economics of Networks: Interactions and Behaviors*. Springer, New York.

Huang, D., Yin, J., Shi, T. and Wang, H. S. (2016). A statistical model for social network labeling. *Journal of Business and Economic Statistics* **34**, 368–374.

Kalton, G. and Kasprzyk, D. (1986). The treatment of missing data. *Survey Methodology* **12**, 1–16.

Lee (2004). Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. *The Econometrics Journal* **72**, 1899–1925.

Lee, L. F., Liu, X. D. and Lin, X. (2010). Specification and estimation of social interaction models with network structures. *The Econometrics Journal* **13**, 145–176.

LeSage, J. and Pace, R. K. (2009). *Introduction to Spatial Econometrics*. Chapman & Hall, New

York.

Liang, H., Wang, S. and Carroll, R. J. (2007). Partially linear models with missing response variables and error-prone covariates. *Biometrika* **94**, 185–198.

Little, R. J. and Rubin, D. B. (2002). *Statistical Analysis With Missing Data*. 2nd Edition. John Wiley and Sons, Inc.

Qin, J., Leung, D. and Shao, J. (2002). Estimation with survey data under non-ignorable nonresponse or informative sampling. *Journal of the American Statistical Association* **97**, 813–817.

Qin, J., Shao, J. and Zhang, B. (2008). Efficient and doubly robust imputation for covariate-dependent missing response. *Journal of the American Statistical Association* **103**, 797–810.

Rao, J. N. K. and Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika* **79**, 811–822.

Rubin, D. B. (1987) *Multiple Imputation for Nonrespondents in Surveys*. Wiley, New York.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman and Hall, London.

Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice and Visualization*. John Wiely, New York.

Shao, J. and Wang, H. (2002). Sample correlation coefficients based on survey data under regression imputation. *Journal of the American Statistical Association* **97**, 544–552.

Srivastava, M. S. and Cater, E. M. (1986). The maximum likelihood method for nonresponse in sample surveys. *Survey Methodology* **12**, 61–72.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge University Press.

Wang, M., Deng, P. and Geng, Z. (2016). Identifiability of normal and normal mixture models with nonignorable missing data. *Journal of the American Statistical Association* **111**, 1673–1683.

Wang, Q. H., Linton, O. and Hardle, W. (2004). Semiparametric regression analysis with missing response at random. *Journal of the American Statistical Association* **99**, 334–345.

Wang, S., Shao, J. and Kim, J. K. (2014). An instrumental variable approach for identification and estimation with nonignorable nonresponse. *Statistica Sinica* **24**, 1097–1116.

Zhao, P. X. and Tang, X. R. (2016). Imputation based statistical inference for partially linear quantile regression models with missing responses. *Metrika* **79**, 991–1009.

Zhou, J., Tu, Y. D., Chen, Y. X. and Wang, H. S. (2017). Estimating spatial autocorrelation with sampled network data. *Journal of Business and Economic Statistics* **35**, 130–138.

School of Statistics and Mathematics, Central University of Finance and Economics, Beijing, P. R. China 100081.

E-mail: sunzhimeng99@126.com

Guanghua School of Management, Peking University, Beijing, P. R. China 100871.Guanghua School of Management, Peking University, Beijing, P. R. China 100871.

E-mail: hansheng@pku.edu.cn

(Received October 2016; accepted September 2018)