Imputing Missing Data in Time Diary Study Based on A Markov Chain Regression Model

Jianxin Shi and Norman Nie

Stanford Institute for the Quantitative Study of Society
Encina Hall West Room 104
Stanford University Stanford, CA 94305-6048
Phone: (650) 723-7242, Fax: (650) 723-7351
email: jianxins@stanford.edu, nhnie@stanford.edu

Abstract

The Stanford Institute for the Quantitative Study of Society (SIQSS) has conducted a time diary study for Internet use since 2000. This study splits 'yesterday' into six time blocks, and randomly draws one hour from each block to ask respondents about their detailed activities. Imputation methods are then used to estimate the whole day activities that will be used for statistical analysis. In this paper, we propose a new imputation method based on a Markov Chain Regression Model (MCRM) for the SIQSS study design. MRCM characterizes the data dependence structure by a Markov Chain and allows the transition probabilities to depend on demographic variables. We demonstrate by theoretical analysis and simulation studies that MCRM can produce consistent estimation of population mean, population variance and univariate regression equations. On the other hand, we show that naive imputation methods without correctly modelling dependence structure could produce incorrect estimation of the population variance and regression equation. MCRM can also be extended to impute two activities simultaneously to model the relationship of two imputed activities.

1 Introduction

Since 2000, the Stanford Institute for the Quantitative Study of Society (SIQSS) has conducted a time diary study for Internet usage using a nationally representative survey panel. Unlike traditional time diaries based on 24-hour design, this study splits 'yesterday' into six blocks (night, early morning, late morning, afternoon, early evening and late evening), and randomly draws one hour (defined as an interval in the paper) from each block to ask respondents about their detailed activities. Demographic background factors, such as education, marry status, age, household composition, are also collected for each respondent to allow for further multivariate analysis. The time spent on an activity in each surveyed hour may take seven possible values from zero to six: zero if no such activity, one if the time is between zero and ten minutes, and so on. Figure I illustrates the sampling scheme. For more details, please refer to Nie and Erbring (2002), Nie and Hillygus (2002).

The key advantage of the six-hour design over usual 24-hour design is that, it allows the respondents to precisely recall their activities in much greater details without being exhausted. Hence, this design provides data of high quality to examine the relationships between Internet use and time spent in other social activities, and the relationship among the time spent on different Internet activities.

While data is collected for only six of the twenty-four hours for the respondents, in practice one is often interested in obtaining the information for the whole day. This can be accomplished by imputation methods. For simplicity, we use $X_{n,t}$ to denote the time spent on a given activity in tth interval for subject n, and define $Y_n = \sum_{t=1}^{24} X_{n,t}$ to be the total time spent on the activity in the day.

The simplest imputation method is to estimate the activity in each missing interval as the observed value in the same block. The implicit assumption is that the intervals in one time block are fully correlated. We call this naive method block imputation or BI. Another extreme is to assume that $X_{n,t}$ are independent across time. Hence, each $X_{n,t}$ is modelled as a function of t and some demographic variables $\mathbf{Z}_n$: $X_{n,t} = \alpha + \beta' Z_n + \lambda(t)$. Models are fitted using the data in the surveyed intervals. The unobserved activities are imputed using the estimated marginal model. Because we consider only the marginal distribution without considering the correlation structure, we call this strategy marginal imputation or MI.

Obviously, both strategies incorrectly model the dependence structure of $\{X_{n,t}, t = 1, \cdots, 24\}$. As a result, they give incorrect estimates of population variance $var(Y_n)$, which can be problematic for further hypothesis testing, for example, of the mean activity levels between two populations. Moreover, both strategies estimate incorrect regression models.

The purpose of this paper is (1) to develop an appropriate probability framework to estimate the joint distribution of the missing intervals conditioning on observations, and (2) to provide a consistent estimate of the population variance, the coefficients and the associated t-test statistics for a univariate regression analysis when imputed measures are involved.

The paper is organized as follows. In section 2, we describe three problems related with data imputation that will be addressed in the paper. In section 3, we develop a Markov Chain Regression Model (MCRM) to characterize the dependence structure of the activities in twenty four intervals, and then develop methods for estimating the population variance and regression equation based on the estimated MCRM. In section 4, we demonstrate by simulations the validity of the model estimation procedure and the ability to correctly estimate the population variance and regression equations. We conclude by discussing the underlying model assumptions and possible extensions in Section 5.

2 The statement of problem

Throughout this paper, we consider one activity. Let $X_{n,t}$ denote the observation in the tth hour for respondent n, where $1 \leq n \leq N$, $1 \leq t \leq T = 24$ and $X_{n,t} \in \{0, 1, \cdots, 6\}$. $X_{n,t} = 0$ if he spends no time on the activity in the tth hour. $X_{n,t} = k$ if he spends $(10 \times (k-1), 10 \times k]$ minutes. For the nth respondent, six hours $(t_1^n, \cdots, t_6^n)$ are randomly generated from six blocks respectively and the activities are reported for these selected hours. The observation for the nth respondent is $\{t_1^n, \cdots, t_6^n, X_{n,t_1^n}, \cdots, X_{n,t_6^n}, \mathbf{Z}_n\}$, where $\mathbf{Z}_n = (z_{n,1}, \cdots, z_{n,m})$ denotes m demographic variables, such as education level, marry status, etc. Our main interest is to develop methods to impute $Y_n = \sum_{t=1}^{T} X_{n,t}$, the total time spent on the activity yesterday, and to provide valid inferences when Y_n is involved in statistical analysis. Particularly, three problems will be discussed in the paper.

The first problem is to estimate the population variance $var(Y_n)$. This is critical when we detect the difference between two years or two populations. Statistics theory indicates that BI tends to overestimate $var(Y_n)$, hence the power of detecting difference will be reduced substantially. On the other hand, MI tends to underestimate $var(Y_n)$, hence it detects spurious difference with inflated probability than expected.

The second problem concerns the univariate regression analysis with Y_n as the explanation variable. Suppose W_n is generated according to a linear model: $W_n = \alpha + \beta Y_n + \varepsilon_n$, where $\varepsilon \sim N(0, \sigma_0^2)$. Given the incomplete data $\{(X_{n,t_1^n}, \cdots, X_{n,t_6^n}, W_n), n = 1 \cdots, N\}$, we need to estimate β and compute the associated t-statistics.

The third problem also concerns the univariate regression analysis but with Y as the response variable. Let G_n be the explanation variable. Suppose $X_{n,t} = \alpha_t + \beta_t G_n + \varepsilon_{n,t}$ for $t = 1, \cdots, T$, where $\varepsilon_n = (\varepsilon_{n,1}, \cdots, \varepsilon_{n,T})$ follows a multivariate Gaussian distribution. It follows that $Y_n = \sum_{t=1}^T \alpha_t + (\sum_{t=1}^T \beta_t) G_n + \sum_{t=1}^T \varepsilon_{n,t}$. We need to estimate $\beta = \sum_{t=1}^T \beta_t$ and compute the t-statistics from the incomplete observations.

3 Methods

3.1 Model specification

For respondent n, the observed data is $\{t_1^n, \cdots, t_6^n, X_{n,t_1^n}, \cdots, X_{n,t_6^n}, \mathbf{Z}_n\}$. We model $X_{n,t}$ as an ordinal variable. We assume that the current value of $X_{n,t}$ depends only on the value one step ago, i.e.

$$P\{X_{n,t}|X_{n,t-1}, \cdots, X_{n,0}\} = P\{X_{n,t}|X_{n,t-1}\}. \tag{1}$$

Under this assumption, $\{X_{n,t}, t = 1, \cdots, T\}$ forms a Markov chain of the first order with finite states $\{0, 1, \cdots, K = 6\}$. The chain is characterized by the initial distribution and the transition probability matrices. To make the model flexible, we model these probabilities as functions of demographic variables $\mathbf{Z}_n$ and time t.

Define $P_{ij}^n(t) = P\{X_{n,t} \le j | X_{n,t-1} = i\}$ to be the accumulated probability at state j for $i \in \{0, \cdots, K\}$ and $j \in \{0, \cdots, K-1\}$. To allow the transition probability to depend on the demographic variables $\mathbf{Z}_n$ and time t, we parameterize $P_{ij}^n(t)$ as

$$P_{ij}^n(t) = \frac{\exp\left\{\sum_{k=0}^j \mu_{ik} + \lambda_t + \mathbf{Z}_n'\beta\right\}}{1 + \exp\left\{\sum_{k=0}^j \mu_{ik} + \lambda_t + \mathbf{Z}_n'\beta\right\}}. \tag{2}$$

Here, $\beta_i \in R, \mu_{i0} \in R$, but $\mu_{ij} \ge 0$ for $j > 0$ to guarantee $P_{ij}^n(t) \ge P_{i(j-1)}^n(t)$. Under this parametrization, we have

$$\log \frac{P(X_{n,t} > j-1 | X_{n,t-1} = i)}{P(X_{n,t} \le j-1 | X_{n,t-1} = i)} = \sum_{k=0}^{j} \mu_{ik} + \lambda_t + \mathbf{Z}_n'\beta. \tag{3}$$

The transition probability $a_{ij}(t) = P\{X_{n,t} = j | X_{n,t-1} = i\}$ from time $t-1$ to t is then

$$a_{ij}^n(t) = \begin{cases} P_{i0}^n(t) & \text{if } j = 0; \\ P_{ij}^n(t) - P_{i(j-1)}^n(t) & \text{if } j \in \{1, \cdots, K-1\} \\ 1 - P_{i(K-1)}^n(t) & \text{if } j = K. \end{cases}$$

We use A_t^n to denote the transition probability matrix with $[A_t^n]_{ij} = a_{ij}^n(t)$.

Similarly, we define initial probability distribution $q_j^n = P\{X_{n,0} = j\}$ to be

$$q_j^n = \begin{cases} Q_0^n & \text{if } j = 0; \\ Q_j^n - Q_{j-1}^n & \text{if } j \in \{1, \cdots, K-1\} \\ 1 - Q_{K-1}^n & \text{if } j = K, \end{cases}$$

where

$$Q_j^n = P\{X_{n,0} \leq j\} = \frac{\exp\left\{\sum_{k=0}^{j} \nu_k + \mathbf{Z}_n'\gamma\right\}}{1 + \exp\left\{\sum_{k=0}^{j} \nu_k + \mathbf{Z}_n'\gamma\right\}} \tag{4}$$

with $\nu_0 \in R, \gamma_k \in R$ and $\nu_j > 0$ for $j > 0$. Define $\mathbf{Q}^n = (q_0^n, \cdots, q_K^n)$.

Whenever the parameters $(\mu_{ij}, \nu_j, \beta, \gamma, \lambda_t)$ are estimated, we can compute the initial probability vector $\mathbf{Q}^n$ and the transition matrices $(A_1^n, \cdots, A_{T-1}^n)$, hence fully characterize the joint distribution of $\{X_{n,t}\}$.

3.2 Likelihood function

When data are observed at all twenty four hours, the likelihood for the nth respondent is:

$$L_n = P\{X_{n,1} = x_1, \cdots, X_{n,T} = x_T\} = q_{x_1}^n a_{x_1,x_2}^n \cdots a_{x_{(T-1)},x_T}^n$$

from the definition of the first order Markov Chain. For our sampling design with missing values, the available observations form a reduced Markov chain. See Figure 2 for illustration. The transition probability matrices of the reduced Markov chain can be expressed as the product of the transition probability matrices of the original Markov chain. Formally, let $B^n(t_1, t_2)$ denote the transition probability matrix from t_1 to t_2 for subject n, then

$$B^n(t_1, t_2) = A_{t_1}^n \cdots A_{t_2-1}^n. \tag{5}$$

Let $b_{ij}^n(t_1, t_2) = [B^n(t_1, t_2)]_{ij} = P\{X_{n,t_2} = j | X_{n,t_1} = i\}$. Then, the likelihood for observation $\{t_1^n, \cdots, t_s^n, X_{n,t_1^n} = x_1, \cdots, X_{n,t_s^n} = x_s\}$ is given by:

$$L_n = \Big\{ \sum_{k=0}^{K} q_k^n b_{kx_1}(1, t_1^n) \Big\} b_{x_1,x_2}^n(t_1^n, t_2^n) \cdots b_{x_{s-1},x_s}^n(t_{s-1}^n, t_s^n), \tag{6}$$

which is a function of $(\mu_{ij}, \nu_j, \beta, \gamma, \lambda_t)$.

3.3 Model estimation

The model parameters are estimated by maximizing the log likelihood function. Let m denote the number of covariates. Our model has $70 + 2m$ parameters: 42 μ_{ij}, 22 λ_t, 6 ν_j, β and γ.

Here, we have set $\lambda_{23} = 0$ to make the parameters identifiable. The constraints are: $\mu_{ij} > 0$ for $j > 0$, $\nu_j > 0$ for $j > 0$. These positive constraints are eliminated by replacing μ_{ij}, ν_j with $e^{\mu_{ij}}$, e^{ν_j}. Under such parametrization, estimating the model parameters reduces to solve an unconstrained nonlinear optimization problem.

Because of the complex model structure and large number of parameters, it's tedious analytically and expensive computationally to calculate Hessian matrix. Hence, we use BFGS quasi-Newton method. The idea of the method is to approximate the Hessian matrix using the information of previous gradient vectors. The resulting matrix is positively definite, hence, it moves toward the maximum of the log likelihood function in each iteration. Specially, we use a limited memory BFGS (Zhu, Byrd and Nocedal, 1997) routine to search for the maximizer of the log likelihood function. The routine can quickly approach the maximizer. But in our applications, it may stop without meeting the stopping criterion that the norm of the gradient vector is sufficiently small. We take the output from the limited memory BFGS as input and restart the BFGS Quasi-Newton's iteration procedure combined with Armijo-Goldstein linear search. Numerical experiments demonstrate that this strategy can find the maximizer satisfying the stopping criterion quickly. Appendix I gives the detail of computing the gradient vector.

3.4 Data imputation based on MCRM

The goal of this section is to develop methods based on the estimated Markov Chain to address the three problems stated in Section 2.

Estimate population variance $var(Y)$: In Algorithm 3.2, we propose to use multiple imputation procedure (Rubin, 1976, Little and Rubin, 1987) to estimate $var(Y)$ based on a Gibbs sampling algorithm that is described in Algorithm 3.1. In the algorithms, we use $\mathbf{X}_n^+$ to denote the observed data, $\mathbf{X}_n^-$ to denote the unobserved data and $\{s_1, \cdots, s_{T-}\}$ to denote the hours without been surveyed. The idea of multiple imputation is that we draw the missing data $\mathbf{X}_n^-$ conditioning on the observed data $\mathbf{X}_n^+$ multiple times and average the variance estimations from all imputations. The Gibbs sampling algorithm is used to sample $\mathbf{X}_n^-$ conditioning on $\mathbf{X}_n^+$ because the conditional probability is very complicated. For general theory and applications of Gibbs sampling, please refer to Liu (2001). Under mild conditions, we show in Appendix II that this procedure produces a consistent estimation, *i.e.* $\lim_{N,R\to\infty} \widehat{var}_{N,R}(Y) = var(Y)$ *a.s.*.

Estimate linear regression: We first consider estimating $W_n = \alpha + \beta Y_n + \varepsilon_n$ where Y_n serves as the explanation variable. In Appendix III, we prove that a consistent estimation of β

Algorithm 3.1: Simulate $Y_n \sim P(Y_n|\mathbf{X}_n^+)$ using Gipps sampling

---

Compute initial probability vector $\mathbf{Q}^n$ based on MRCM;

Compute transition probability matrix A_t^n for $t = 1, \cdots, T$ based on MRCM;

For each $t \in \{s_1, \cdots, s_{T\_}\}$, set $x_t = \arg\max_k P\{X_{n,t} = k|\mathbf{X}_n^+\}$ as initial states;

For $r = 1 : 2000$

    Randomly choose t from $\{s_1, \cdots, s_{T\_}\}$,

    Draw $X_{n,t} \sim P\{X_{n,t} = k|X_{n,1}, \cdots, X_{n,t-1}, X_{n,t+1}, \cdots, X_{n,T}\}$.

End

Compute $Y_n = \sum_{t=1}^T X_{nt}$.

---

is obtained by imputing Y_n using conditional expectation imputation:

$$\hat{Y}_n = \sum_{t=1}^{T} E(X_{n,t}|\mathbf{X}_n^+). \tag{7}$$

It can be shown that, the t-statistic for testing $H_0 : \beta = 0$ in a univariate linear regression is given by

$$t = \frac{\hat{\beta}}{\sqrt{var(W_k)/var(\hat{Y}_k) - 1}}.$$

Because $\hat{\beta} \to \beta$ and $var(\hat{Y}_k) \leq var(Y_k)$, the t-statistic is asymptotically underestimated. We can replace $var(\hat{Y}_k)$ with $\widehat{var}_{N,I}(Y_k)$ produced from Algorithm 3.2 to derive the correct t-statistic:

$$t = \frac{\hat{\beta}}{\sqrt{var(W_k)/\widehat{var}_{N,I}(Y_k) - 1}}. \tag{8}$$

We then consider estimating $\beta = \sum_{t=1}^T \beta_t$ in linear equation $Y_n = \sum_{t=1}^T \alpha_t + (\sum_{t=1}^T \beta_t)G_t + \sum_{t=1}^T \varepsilon_{n,t}$ (problem 3 in Section 2). Using similar technique to Appendix III (details omitted), we can prove that a consistent estimation of β is obtained using conditional expectation imputation (7). Similarly, the following formula produces a correct value of t-statistic:

$$t = \frac{\hat{\beta}}{\sqrt{\widehat{var}_{N,I}(Y_k)/var(G_k) - 1}}. \tag{9}$$

Algorithm 3.2 Estimate $var(Y)$ using multiple imputation.

For $r = 1 : R$

  For $n = 1 : N$

    Draw $\mathbf{X}_n^- \sim P(\mathbf{X}_n^-|\mathbf{X}_n^+)$ using Gipps sampling (Algorithm 3.1)

    Compute $Y_n^r = \mathbf{X}_n^- + \mathbf{X}_n^+$

  End

  Compute $\hat{\sigma}_r^2 = \sum_{n=1}^N (Y_{n,r} - \bar{Y}^r)^2/N$ with $\bar{Y}^r = \sum_{n=1}^N Y_{n,r}/N$

End

Estimate $var(Y)$ as $\widehat{var}_{N,R}(Y) = \sum_{i=1}^R \hat{\sigma}_r^2/R$.

---

4 Simulation study

4.1 Program

The algorithms were implemented in a C program MCRM 1.0 running under Linux platform.

4.2 Validity of the model estimation procedure

We ran two simulations to validate our model estimation procedure.

Simulation I: We simulate 5000 respondents, each of which has 10 demographic variables generated as independent, normally distributed numbers. $\mu_{ij} = 0.5$ for $i = 0, \cdots, 6, j = 1, \cdots, 5$, $\mu_{i0} = -2$, $\nu_0 = 2$, $\beta_j = \gamma_j = 0.5$, $\lambda_t = 0$. For the nth subject, we compute the initial probability vector $\mathbf{Q}^n = (q_0^n, \cdots, q_6^n)$ and draw $X_1^n = x_1$ according to $\mathbf{Q}^n$. Then, we compute A_1^n and draw $X_2^n = x_2$ according to $P\{X_2^n = x_2|X_1^n = x_1\} = [A_1^n]_{x_1,x_2}$. Similarly, we compute $A_2^n, \cdots, A_{23}^n$ and sequentially simulate $X_2^n, \cdots, X_{23}^n$ according to A_k^n conditioning on the previous step. After simulating the full data for the nth subject, we randomly select one hour in each of the six blocks and mask all other hours.

Simulation II: To test the algorithm with a more realistic parameter specification, we first fit our model with TV time as the target variable using 2006 survey data with 5126 subjects to obtain estimate $(\mu_{ij}, \nu_j, \beta, \gamma, \lambda_t)$. We include 20 dummy variables into analysis formed from the demographic variable such as education, marry status, race, etc. We use these covariates and the estimated parameters to simulate 5216 subjects and mask the observations according to the missing pattern in the real data. Hence, the simulated data has a realistic parameter specification and identical missing pattern to our 2006 data set. We then re-estimate these

model parameters to get $(\mu'_{ij}, \nu'_j, \beta', \gamma', \lambda'_t)$ and compare them with $(\mu_{ij}, \nu_j, \beta, \gamma, \lambda_t)$.

For both simulations, we set 0.1 as the starting values for all parameters. Both models were estimated within 20 minutes on a Linux server powered by an Intel 5160 CPU. We plot the estimated values against the true values on a scatter plot. The parameters are correctly estimated if the points lie on or close to the line $y = x$.

Results: All parameters were accurately estimated for simulation I (the left panel in Figure 3). In simulation II, all parameters $(\mu_{ij}, \beta, \lambda_t)$ related with the transition probabilities were correctly estimated, but the parameters (ν_j, γ) related with the initial distribution were poorly estimated. This is expected considering the fact that the target variable other than sleeping is exactly zero because most of the respondents are sleeping from 12:00PM to 6:00AM. For our data imputation methods based on a Marlov Chain, these parameters have effect only on a few hours prior to the first observation, so it won't degrade our whole procedure.

4.3 Data imputation and statistical inferences

In this section, we use simulations to demonstrate the improvement of our new method MCRM over BI and MI. We will compare the performance in terms of the prediction error, the estimation of population variance and a univariate linear regression.

Simulation III: Model settings are identical to simulation II. We compare the prediction error rate and the estimation of population variance of these methods. We use $X_{n,t}$ to denote the true value and $\hat{X}_{n,t}$ to denote the imputed value. Let $Y_n = \sum_{t=1}^{T} X_{n,t}$ and $\hat{Y}_n = \sum_{t=1}^{T} \hat{X}_{n,t}$. We define two criteria

$$C_1 = \frac{1}{18N} \sum_{n=1}^{N} \sum_{t=1}^{T} (\hat{X}_{n,t} - X_{n,t})^2 \quad \text{and} \quad C_2 = \frac{1}{N} \sum_{n=1}^{N} (\hat{Y}_n - Y_n)^2.$$

Hence, C_1 is the average error square for each hour and each person, C_2 is the average error square for one day and each person.

The simulation results in Table 1 show that our MCRM procedure based on (7) has the lowest prediction error (both C_1 and C_2) compared to BI and MI. In addition, all methods correctly estimated the population mean, but only MCRM can correctly estimate the population variance.

Simulation IV: Model settings are identical to simulation II. In addition, we generate $W_n = \alpha + \beta Y_n + \varepsilon_n$ with $\alpha = \beta = 1$ and ε_n i.i.d. $N(0, 200^2)$. Based on the complete data of 5216 samples, $\hat{\beta}$ is 1.02 and the t-statistic is 44.2. We then ran linear regression using W_n as the response variable and $\hat{Y}_n$ as the explanation variable. Here, $\hat{Y}_n$ is imputed using three methods: BI, MI and MCRM. We compare the estimated regression coefficients and the t-statistics t_β.

Results are presented in table 2. The $\hat{\beta}$ and associated t-statistic t_β produced from our MCRM are very close to those produced from the full data. BI tends to underestimates β while MI tends to overestimates β. In addition, both BI and MI tend to underestimate the t_β.

5 Conclusion and discussion

In this paper, we develop a method for imputing missing data in SIQSS time diary study. We also develop methods for estimating population variance and univariate linear regression. It's worthwhile to point out that, any general imputation method without considering the special data structure may create more problems than it solves, e.g. biasing the variance estimation, regression analysis and hypothesis testing. A well designed imputation procedure requires (1) to specify a probability model on the complete data with appropriate marginal distribution and dependence structure and (2) to develop imputation method based on the probability model. Our MCRM is such a good candidate for SIQSS time diary study because (1) it explicitly models the ordinal target variable to ensure an appropriate marginal distribution, (2) it uses the first order Markov Chain to model the dependence of the 24 hour data, (3) it allows the initial probability and transition probability of the Markov Chain to depend on demographic variables to reflect individual's effect. As the result, it produces consistent estimation for variance, regression equations in our statistical analysis of time data.

Markov Chain models have been widely used in longitudinal data analysis to explore the the relationship between the outcome variable and explanation variables. For example, Zeger, Liang and Self (1985) proposed a first order Markov Chain model for binary response by modelling the marginal distribution as a logistic function of covariates. Cox (1970) proposed to model the transition probability for binary responses. Our MCRM procedure adopts Cox's approach to model the transitional probability but for ordinal responses. Moreover, this model allows us to compute the exact likelihood function even when majority (75%) of the data are missing.

As Zeger, Liang and Self (1985) pointed out, the first order Markov Chain model might not be adequate if the actual data have dependence structure of longer distance. In principal, a higher order Markov Chain Regression Model can be developed similarly for imputation purpose. However, the computational burden for fitting even a second order Markov model would be prohibitive when a lot of missing values are present. On the other hand, modelling longer distance dependence provides little additional information for such a sparse observation pattern. Hence, our MCRM of the first order is a tradeoff between model complexity and efficiency.

In this paper, we considered imputing one target variable at a time. MCRM can also be extended to impute multiple target variables jointly. For example, if email and TV time are jointly modelled, then $X_{n,t}$ may take $(6+1) \times 7/2 = 21$ states, each of which represents one combination of email time and TV time with the restriction that the summation doesn't exceed 6. The joint imputation would be particularly useful when exploring the relationship between the two imputed activities.

Acknowledgement: The authors thank Xiaobing He for preparing the 2006 Survey data.

References

[1] Little, R.J.A. and Rubin, D.B. (1987), Statistical Analysis with Missing Data. *J. Wiley Sons, New York.*

[2] Rubin, D.B. (1976) Inference and missing data. *Biometrika, 63, 581-592.*

[3] Byrd, R.H., Lu, P. and Nocedal., J. (1995) A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM Journal on Scientific and Statistical Computing , 16, 5, 1190-1208.*

[4] Zhu, C., Byrd, R.H. and Nocedal, J. (1997) L-BFGS-B: Algorithm 778: L-BFGS-B, FORTRAN routines for large scale bound constrained optimization. *ACM Transactions on Mathematical Software, Vol 23, Num. 4, 550 - 560.*

[5] Liu, J.S. (2001) Monte Carlo Strategies in Scientific Computing, *Springer, 2001.*

[6] Nie, N.H. and Erbring L. (2002) Internet and Society: A Preliminary Report. *IT and Society 1(1): 275-283.*

[7] Nie, N.H. and Hillygus S. (2002) The Impact of Internet Use on Sociability: Time-Diary Findings. *IT and Society 1(1):1-20.*

[8] Cox, D.R. (1970) The analysis of binary data. *London: Muthuen.*

[9] Zeger, S.L., Liang, K.Y. and Self, S.G. (1985) The analysis of binary longitudinal data with time independent covariates. *Biometrika 1985 72(1):31-38;*

Appendix I: Compute the gradient vector

According to (6), the overall log likelihood function $l = \sum_{n=1}^{N} \log L_n$ can be expressed as the summation of

$$\log P\{X_{n,t_2} = j | X_{n,t_1} = i\} = \log b_{ij}^n(t_1, t_2)$$

and

$$\log P\{X_{n,t} = i\} = \log \Big\{ \sum_{k=0}^{K} P\{X_{n,1} = k\} P\{X_{n,t} = i | X_{n,1} = k\} \Big\} = \log \Big\{ \sum_{k=0}^{K} p_k^n b_{ki}(1, t) \Big\}.$$

Hence, to compute the gradient vector, we only need to compute

$$\frac{\partial b_{ij}^n(t_1, t_2)}{\partial \mu_{kl}}, \frac{\partial b_{ij}^n(t_1, t_2)}{\partial \lambda_t}, \frac{\partial b_{ij}^n(t_1, t_2)}{\partial \beta_k} \tag{10}$$

and

$$\frac{\partial p_i^n}{\partial \nu_k}, \frac{\partial p_i^n}{\partial \gamma_k}. \tag{11}$$

Computing (7) is tedious but straightforward according to the definition of p_i^n. Calculation of (6) is based on the fact:

$$\begin{aligned} \frac{\partial B^n(t_1, t_2)}{\partial y} &= \frac{\partial \{A_{t_1}^n A_{t_1+1}^n \cdots A_{t_2-1}^n\}}{\partial y} \\ &= \frac{\partial A_{t_1}^n}{\partial y} A_{t_1+1}^n \cdots A_{t_2-1}^n + \cdots + A_{t_1}^n A_{t_1+1}^n \cdots \frac{\partial A_{t_2-1}^n}{\partial y}, \end{aligned}$$

where $[\partial A_t^n / \partial y]_{ij} = \partial a_{ij}^n(t) / \partial y$.

Now we consider the computational cost for computing the derivative. When $t_2 - t_1 = 0$ (no gap), we need to compute the derivative once, the algorithm complexity is $O(1)$. When $t_2 - t_1 = d$, we need to compute derivatives for $2K + (d-3)K^2$ times. Therefore, the algorithm is significantly slow for a long gap.

Appendix II

Suppose $\mathbf{X}_n = (X_{n,1}, \cdots, X_{n,T}) \sim f(x_1, \cdots, x_T)$. Let $Y_n = \sum_{t=1}^{T} X_{n,t}$. Denote $\mathbf{X}^+$ as observed information. For each $n = 1, \cdots, N$, draw R i.i.d. samples $Y_n^1, Y_n^2, \cdots, Y_n^R \sim P(Y | \mathbf{X}_n^+)$ that has mean $\mu(\mathbf{X}_n^+)$ and variance $\lambda^2(\mathbf{X}_n^+)$. Define

$$\sigma_r^2 = \frac{1}{N} \sum_{n=1}^{N} (Y_n^r - \overline{Y}^r)^2, \quad \text{where} \quad \overline{Y}^r = \frac{1}{N} \sum_{n=1} Y_n^r.$$

Let $\widehat{var}_{N,R}(Y) = \sum_{r=1}^{R} \sigma_r^2 / R$. Then $\lim_{N,R \to \infty} \widehat{var}_{N,R}(Y) = var(Y)$.

Proof: It's easy to show that

$$\widehat{var}_{N,I}(Y) = \frac{1}{R}\sum_{r=1}^{R}\sigma_r^2 = \frac{1}{N}\sum_{n=1}^{N}\left(\frac{1}{R}\sum_{r=1}^{R}(Y_n^r)^2\right) - \frac{1}{N^2}\sum_{n_1,n_2}\left(\frac{1}{R}\sum_{r}Y_{n_1}^r Y_{n_2}^r\right). \tag{12}$$

Hence,

$$\begin{aligned}
\lim_{N,R\to+\infty}\widehat{var}_{N,I}(Y) &= \lim_{N\to+\infty}\frac{1}{N}\sum_{n=1}^{N}\left(\lim_{R\to+\infty}\frac{1}{R}\sum_{r=1}^{R}(Y_n^r)^2\right) - \lim_{N\to+\infty}\frac{1}{N^2}\sum_{n_1,n_2}\left(\lim_{R\to+\infty}\frac{1}{R}\sum_{r}Y_{n_1}^r Y_{n_2}^r\right)\\
&= \lim_{N\to+\infty}\frac{1}{N}\sum_{n=1}^{N}E(Y_n^r)^2 - \lim_{N\to+\infty}\frac{1}{N^2}\sum_{n_1,n_2}EY_{n_1}^r Y_{n_2}^r\\
&= \lim_{N\to+\infty}\frac{1}{N}\sum_{n=1}^{N}\left(\mu^2(\mathbf{X}_n^+)+\lambda^2(\mathbf{X}_n^+)\right) - \lim_{N\to+\infty}\left(\frac{1}{N}\sum_{n}\mu(\mathbf{X}_n^+)\right)^2\\
&= E\mu^2(\mathbf{X}^+) + E\lambda^2(\mathbf{X}^+) - \left(E\mu(\mathbf{X}^+)\right)^2\\
&= var(\mu(\mathbf{X}^+)) + E\lambda^2(\mathbf{X}^+).
\end{aligned}$$

Because $\mu(\mathbf{X}^+) = E(Y|\mathbf{X}^+)$ and $\lambda^2(\mathbf{X}^+) = var(Y|\mathbf{X}^+)$ by definition, we have

$$\lim_{N,R\to+\infty}\widehat{var}_{N,I}(Y) = var(E(Y|\mathbf{X}^+)) + E[var(Y|\mathbf{X}^+)] = var(Y)$$

according to the well known formula $var(y) = var(E(y|x)) + E(var(y|x))$.

Appendix III

Suppose $\mathbf{X}_n = (X_{n,1},\cdots,X_{n,T}) \sim f(x_1,\cdots,x_T)$. Let $Y_n = \sum_{t=1}^{T}X_{n,t}$ and $\hat{Y}_n = E(Y_n|\mathbf{X}_n^+) = \sum_{t=1}^{T}E(X_{n,t}|\mathbf{X}_n^+)$, where $\mathbf{X}_n^+$ is the observed information for the nth sample. Suppose $W_n = \alpha + \beta Y_n + \varepsilon_n$, where $\varepsilon \sim (0,\sigma_0^2)$. Let $\hat{\beta}$ be the estimated coefficient β in a regression analysis using $\hat{Y}_n$. Then, $\hat{\beta} \to \beta$ *a.s.* as N goes to infinity.

Proof: According to the standard results of univariate linear regression,

$$\hat{\beta} = \frac{\sum_{k=1}^{N}(W_k-\bar{Z})(\hat{Y}_k-\bar{Y})}{\sum_{k=1}^{N}(\hat{Y}_k-\bar{Y})^2} = \frac{\beta\sum_{k=1}^{N}(Y_k-\bar{Y})(\hat{Y}_k-\bar{Y}) + \sum_{k=1}^{N}\varepsilon_k(\hat{Y}_k-\bar{Y})}{\sum_{k=1}^{N}(\hat{Y}_k-\bar{Y})^2}.$$

As N goes to infinity, $\hat{\beta}$ converges to

$$\beta\frac{cov(Y_k,\hat{Y}_k)}{var(Y_k)}.$$

To prove $\hat{\beta} \to \beta$ *a.s.*, it suffices to prove $cov(Y_k, \hat{Y}_k) = var(\hat{Y}_k)$, which follows from the properties of conditional expectation:

$$\begin{aligned} cov(Y_k, \hat{Y}_k) &= cov(Y_k, E(Y_k|\mathbf{X}_k^+)) \\ &= E[Y_k E(Y_k|\mathbf{X}_k^+)] - EY_k E[E(Y_k|\mathbf{X}_k^+)] \\ &= E\Big\{E[Y_k E(Y_k|\mathbf{X}_k^+)]|\mathbf{X}_k^+\Big\} - (EY_k)^2 \\ &= E[E(Y_k|\mathbf{X}_k^+)]^2 - E^2[E(Y_k|\mathbf{X}_k^+)] \\ &= var[E(Y_k|\mathbf{X}_k^+)] \\ &= var\hat{Y}_k. \end{aligned}$$

This completes the proof.

Table 1: Comparison of prediction error, population variance estimations.

	complete data	BI	MI	MCRM
C_1	–	345.5	225.5	**187.9**
C_2	–	11334	6609	**4760**
mean(Y)	152.4	151.7	151.7	151.6
$var(Y)$	15375	26891	5364	**15160**

Table 2: Comparison of univariate regression analysis

	complete data	BI	MI	MCRM
$\hat{\beta}$	1.02	0.58	1.31	**1.00**
t_β	45.06	31.6	32.1	**44.2**

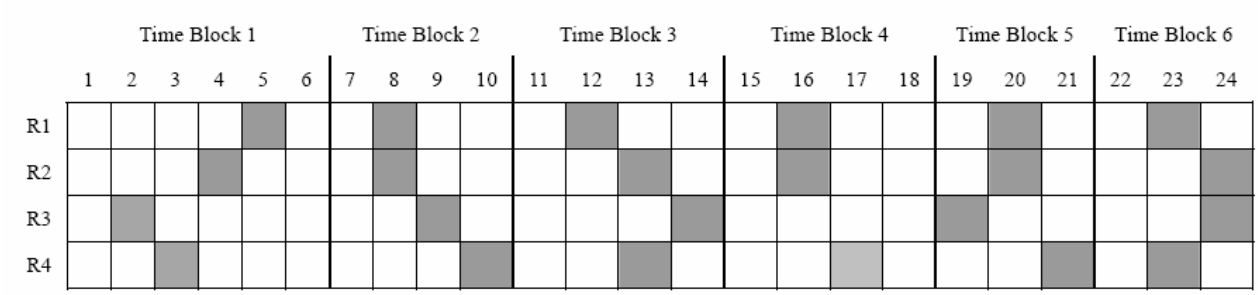

Figure 1: Illustration of the sampling design of SIQSS for four respondents. The whole day is split into six blocks. The first block contains six hours from 12:00AM and to 6:00AM. The last two time blocks contain three hours. Other three bocks contain four hours. Each hour will be referred as an "interval" in the paper. For each respondent, one and only one interval is randomly selected from each block. The survey is then done in the selected six intervals. Different respondents may be surveyed in different combinations of intervals.

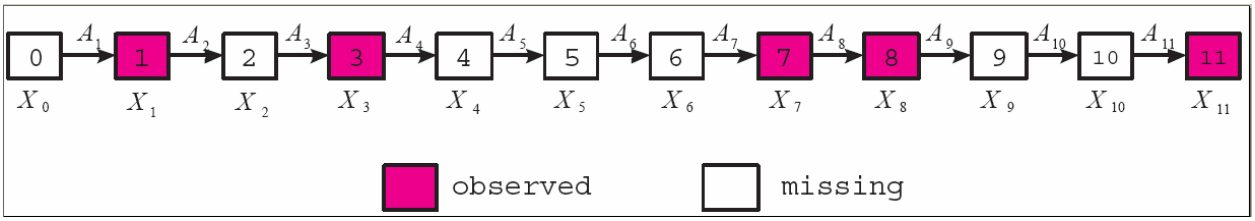

Figure 2: In this example, $\{X_0,\cdots,X_{11}\}$ is a Markov Chain. The observation $\{X_1,X_3,X_7,X_8,X_{11}\}$ forms a reduced Markov Chain. The transition probability matrices are $B_{1,3}=A_{1,2}A_{2,3}$, $B_{3,7}=A_{3,4}A_{4,5}A_{5,6}A_{6,7}$, $B_{7,8}=A_{7,8}$, $B_{8,11}=A_{8,9}A_{9,10}A_{10,11}$, where $A_{i,i+1}$ is the transition matrix of the original Markov Chain chain.

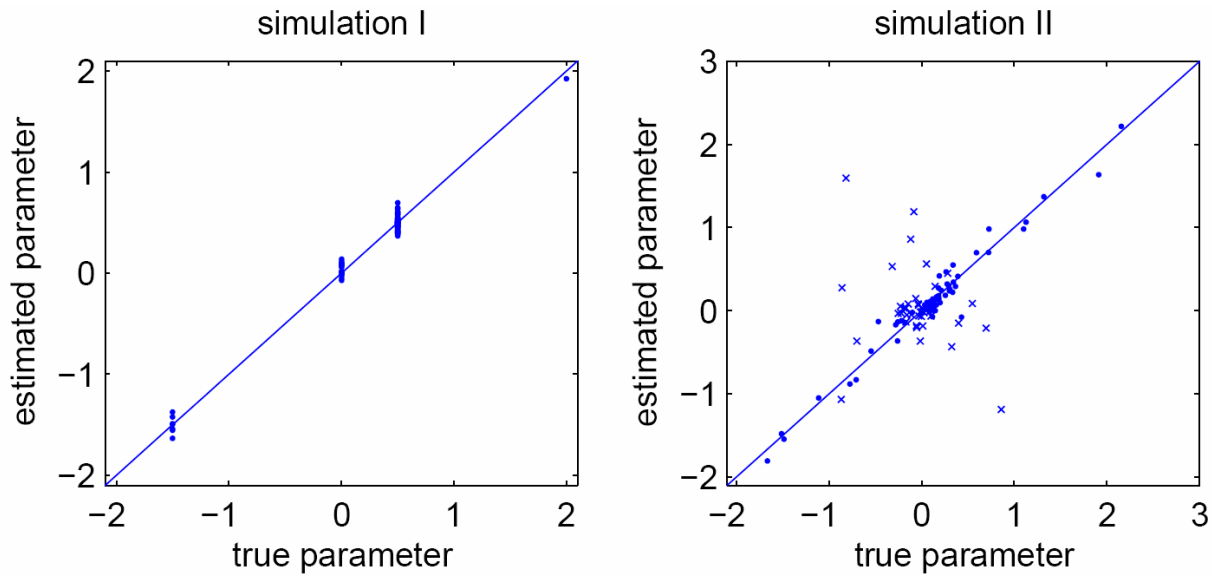

Figure 3: Each point represents a pair of true parameter and estimated parameter. One parameter is accurately estimated if the point is close to the $y = x$ line. In simulation I, parameters are accurately estimated. In simulation II, all 64 parameters related with transition probabilities (labeled as •) are accurately estimated but some of the parameters related with initial probabilities (labeled as ×) are poorly estimated.