Supplementary Materials: Appendices

Multiple Imputation of Missing or Faulty Values Under Linear Constraints

Hang J. Kim, Jerome P. Reiter, Quanli Wang, Lawrence H. Cox, and Alan F. Karr
Duke University and National Institute of Statistical Sciences

A. SENSITIVITY ANALYSIS OF PRIORS

To check the impact of prior settings on the MCMC fitting and multiple imputation inferences, we perform several simulation studies using the hyperparameter settings in Table A.1. For each setting, we run 20 independent chains, each with 10,000 iterations. From each chain, we store every 1,000th iterate, resulting in $m = 10$ completed datasets. We estimate the regression coefficients for the model in (17) in the main text using the usual multiple imputation point estimators (Rubin 1987). As shown in Figure A.1, the multiple imputation inferences are insensitive to the choice of these prior distributions.

Table A.1: Different hyperparameter settings for sensitivity analysis. The middle row of each study (bold character) shows the default setting used in Section 3 of the main text.

Setting	a_α	b_α	a_ϕ	b_ϕ	h	Setting	a_α	b_α	a_ϕ	b_ϕ	h
Study1	1.0	.25	.25	.25	1.0	Study2	.05	.05	.25	.25	1.0
	.50	.25	.25	.25	1.0		.10	.10	.25	.25	1.0
	.25	**.25**	**.25**	**.25**	**1.0**		**.25**	**.25**	**.25**	**.25**	**1.0**
	.25	.50	.25	.25	1.0		1.0	1.0	.25	.25	1.0
	.25	1.0	.25	.25	1.0		2.0	2.0	.25	.25	1.0
Study3	.25	.25	1.0	.25	1.0	Study4	.25	.25	.05	.05	1.0
	.25	.25	.50	.25	1.0		.25	.25	.10	.10	1.0
	.25	**.25**	**.25**	**.25**	**1.0**		**.25**	**.25**	**.25**	**.25**	**1.0**
	.25	.25	.25	.50	1.0		.25	.25	1.0	1.0	1.0
	.25	.25	.25	1.0	1.0		.25	.25	2.0	2.0	1.0
Study5	.25	.25	.25	.25	0.1						
	.25	.25	.25	.25	0.2						
	.25	.25	.25	.25	0.5						
	.25	**.25**	**.25**	**.25**	**1.0**						
	.25	.25	.25	.25	2.0						
	.25	.25	.25	.25	3.0						
	.25	.25	.25	.25	5.0						

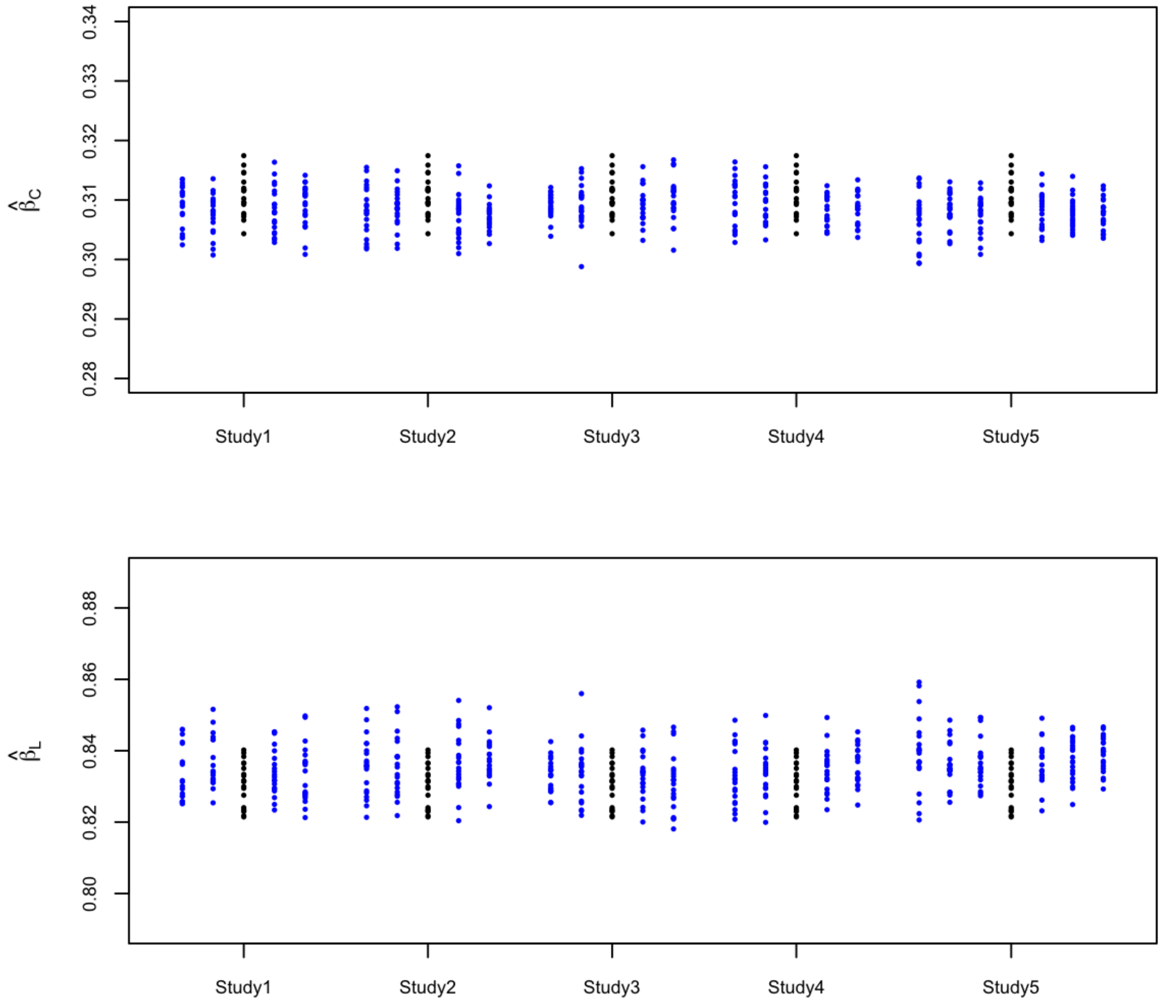

Figure A.1: Estimated regression coefficients of labor in the regression in Section 4 based on the CDPMMN multiple imputation under the prior specifications in Table A.1. In each setting, results are based on 20 independent replications, and the middle column includes the estimates under the prior specification recommended in Section 3 in the main text. Inference is not sensitive to these prior specifications.

B. TRANSFORMING THE VARIABLES AND THE LINEAR CONSTRAINTS

Let x_{ij} be the value of jth variable of ith subject on the original scale. In the analyses in Section 4 of the main text, we take $\log(x_{ij})$ and standardize the logged values using observed data, so that we model

$$y_{ij} = \frac{\log x_{ij} - \tilde{x}_j}{\tilde{s}_j},$$

where $\tilde{x}_j = \sum_{i=1}^n \delta_{ij} \log x_{ij} / \sum_{i=1}^n \delta_{ij}$, $\tilde{s}_j^2 = \sum_{i=1}^n \delta_{ij}(\log x_{ij} - \tilde{x}_j)^2 / (\sum_{i=1}^n \delta_{ij} - 1)$, and $\delta_{ij} = 1$ when x_{ij} is observed and $\delta_{ij} = 0$ otherwise. The motivation of the standardization, i.e., using y_{ij} instead of x_{ij}, is to improve the MCMC fitting, since fewer components are needed to approximate the distribution of the logged and standardized variables, and to simplify specification of hyperparameters in the prior distributions. To be thorough, we repeated the simulation study in Section 4.1 using the unnormalized values $\log x_{ij}$ and the priors described in Section 3.2. The simulation results with the unnormalized values were nearly identical to those with the normalized values y_{ij} shown in Figures 2–5.

We now must account for the transformation in the system of linear inequalities. Suppose that on the original scale, the range restrictions for any x_{ij} are given by

$$L_j \leq x_{ij} \leq U_j, \tag{B.1}$$

where U_j and L_j are agency-fixed upper and lower limits, respectively. Further, for any (x_{ij}, x_{ik}), the ratio edits are given by

$$L_{jk} \leq x_{ij}/x_{ik} \leq U_{jk}, \tag{B.2}$$

where again U_{jk} and L_{jk} are agency-fixed upper and lower limits, respectively.

For ratio constraints involving values x_{ij} and x_{ik}, where $j \neq k$, we rewrite (B.2) as

$$\frac{\log L_{jk} - (\tilde{x}_j - \tilde{x}_k)}{\tilde{s}_j \tilde{s}_k} \leq \frac{y_{ij}}{\tilde{s}_k} - \frac{y_{ik}}{\tilde{s}_j} \leq \frac{\log U_{jk} - (\tilde{x}_j - \tilde{x}_k)}{\tilde{s}_j \tilde{s}_k}.$$

The new ratio constraint limits, (L^*_{jk}, U^*_{jk}) for y_{ij} and y_{ik} can be expressed as

$$L^*_{jk} \leq y_{ij} - c_{jk} y_{ik} \leq U^*_{jk} \tag{B.3}$$

where $c_{kj} = \tilde{s}_k/\tilde{s}_j$, $L^*_{jk} = (\log L_{jk} - \tilde{x}_j + \tilde{x}_k)/\tilde{s}_j$ and $U^*_{jk} = (\log U_{jk} - \tilde{x}_j + \tilde{x}_k)/\tilde{s}_j$. The new range limits, (L^*_j, U^*_j), for y_{ij} are

$$L^*_j \leq y_{ij} \leq U^*_j \tag{B.4}$$

where $L^*_j = (\log L_j - \tilde{x}_j)/\tilde{s}_j$ and $U^*_j = (\log U_j - \tilde{x}_j)/\tilde{s}_j$.

C. FINDING THE FEASIBLE REGION $\mathcal{A}_i$

In this section, we present an illustrative example of the matrix-based approach to finding feasible regions in systems of linear inequalities. For simplicity and to show the main ideas, suppose that $\boldsymbol{y}_i = (y_{i1}, y_{i2}, y_{i3})'$. The range restrictions and ratio inequalities from (B.3) and

(B.4) can be written as $A\boldsymbol{y}_i \le \boldsymbol{b}$, where

$$A = \begin{pmatrix} 1 & -c_{12} & 0 \\ 1 & 0 & -c_{13} \\ 0 & 1 & -c_{23} \\ -1 & c_{12} & 0 \\ -1 & 0 & c_{13} \\ 0 & -1 & c_{23} \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{b} = \begin{pmatrix} U_{12}^* \\ U_{13}^* \\ U_{23}^* \\ -L_{12}^* \\ -L_{13}^* \\ -L_{23}^* \\ U_1^* \\ U_2^* \\ U_3^* \\ -L_1^* \\ -L_2^* \\ -L_3^* \end{pmatrix}.$$

Suppose that y_{i1} and y_{i3} are missing/blanked but y_{i2} is observed. We need to find the feasible range for (y_{i1}, y_{i3}) given y_{i2}. Let A_{13} be the sub-matrix of A associated with y_{i1} and y_{i3}, comprising the first and third columns of A whose entries are not zeros, and, let $\boldsymbol{b}_{13}$ and $\boldsymbol{a}_2$ be the corresponding vectors from $\boldsymbol{b}$ and the second column of A. We have

$$A_{13} = \begin{pmatrix} 1 & 0 \\ 1 & -c_{13} \\ 0 & -c_{23} \\ -1 & 0 \\ -1 & c_{13} \\ 0 & c_{23} \\ 1 & 0 \\ 0 & 1 \\ -1 & 0 \\ 0 & -1 \end{pmatrix}, \quad \boldsymbol{b}_{13} = \begin{pmatrix} U_{12}^* \\ U_{13}^* \\ U_{23}^* \\ -L_{12}^* \\ -L_{13}^* \\ -L_{23}^* \\ U_1^* \\ U_3^* \\ -L_1^* \\ -L_3^* \end{pmatrix}, \quad \text{and} \quad \boldsymbol{a}_2 = \begin{pmatrix} -c_{12} \\ 0 \\ 1 \\ c_{12} \\ 0 \\ -1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

The feasible region for the missing (y_{i1}, y_{i3}) is $\mathcal{A}_i = \{(y_{i1}, y_{i3})' : A_{13}(y_{i1}, y_{i3})' \le \boldsymbol{b}_{13} - \boldsymbol{a}_2 y_{i2}\}$. The vector and matrices are defined similarly to compute general feasible regions.

Table D.1: Summaries of simulated datasets under MAR assumptions

		RVA	CAP	SL	USL	RMU
Parameters	$\tau_{0,j}$	-5.0	2.0	2.0	-4.0	-4.0
	$\tau_{S,j}$	0.1	-0.2	-0.2	0.1	0.1
	$\tau_{U,j}$	0.2	-0.2	-0.2	0.2	0.2
Missingness rate		12.8%	13.4%	13.4%	27.9%	27.9%

D. SIMULATION STUDY UNDER MISSING AT RANDOM

In this section, we apply the CDPMMN method in a simulation with a missing at random (MAR) mechanism. We modify the repeated simulation in Section 4.2 of the main text as follows. Assume that the variables SW and USW are fully observed, and the remaining variables are subject to item nonresponse. For each remaining variable, the missing data model follows the logistic regression,

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \tau_{0,j} + \tau_{S,j}\log \mathrm{SW}_i + \tau_{U,j}\log \mathrm{USW}_i,\ \ i = 1,\ldots,n.$$

Here, p_{ij} is the probability that variable j of record i is missing, where $j \in \{$RVA,CAP,SL, USL,RMU$\}$.

For each of 500 replications of the simulation, we first randomly select 1,000 records from the the 6,607 plants in the 1991 data. For each i in the sample, for each variable we introduce missing values with probability p_{ij}. Table D.1 displays the parameters of the logistic regression model and the average missingness rate per variable in the 500 simulations.

We create $m = 10$ completed data sets of the missing items using the CDPMMN model. We estimate the regression coefficients described in Section 4.2 using the usual multiple

Table D.2: Properties of point estimators across the 500 simulations for the original data, the CDPMMN multiple imputation, and the complete cases analysis of the simulated data generated under MAR assumption.

	β_0			β_C			β_L		
	Bias	*TSE*	*TSD*	*Bias*	*TSE*	*TSD*	*Bias*	*TSE*	*TSD*
Original data	-.007	4.03	–	.001	0.21	–	-.001	0.60	–
CDPMMN	.005	5.48	1.69	-.003	0.26	0.07	.003	0.78	0.22
Complete cases	.129	24.14	20.80	-.014	0.78	0.56	-.013	2.05	1.36

imputation point estimators, as well as the point estimates based on only the complete cases. Table D.2 displays the results of the simulation study. The CDPMMN is effective in this MAR scenario, whereas the complete cases analysis results in biased estimation.

References

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