Multiple Imputation of Missing or Faulty Values Under Linear Constraints

Hang J. Kim

Duke University and National Institute of Statistical Sciences, Durham, NC 27708 (*hangkim@niss.org*)

Jerome P. Reiter

Department of Statistical Science, Duke University, Durham, NC 27708 (*jerry@stat.duke.edu*)

Quanli Wang

Department of Statistical Science, Duke University, Durham, NC 27708 (*quanli@stat.duke.edu*)

Lawrence H. Cox

National Institute of Statistical Sciences, Research Triangle Park, NC 27709 (*cox@niss.org*)

Alan F. Karr

National Institute of Statistical Sciences, Research Triangle Park, NC 27709 (*karr@niss.org*)

ABSTRACT

Many statistical agencies, survey organizations, and research centers collect data that suffer from item nonresponse and erroneous or inconsistent values. These data may be required to satisfy linear constraints, e.g., bounds on individual variables and inequalities for ratios or sums of variables. Often these constraints are designed to identify faulty values, which then are blanked and imputed. The data also may exhibit complex distributional features, including nonlinear relationships and highly non-normal distributions. We present a fully Bayesian, joint model for modeling or imputing data with missing/blanked values under linear constraints that (i) automatically incorporates the constraints in inferences and imputations, and (ii) uses a flexible Dirichlet process mixture of multivariate normal distributions to reflect complex distributional features. Our strategy for estimation is to augment the observed data with draws from a hypothetical population in which the constraints are not present, thereby taking advantage of computationally expedient methods for fitting mixture models. Missing/blanked items are sampled from their posterior distribution using the Hit-and-Run sampler, which guarantees that all imputations satisfy the constraints. We illustrate the approach using manufacturing data from Colombia, examining the potential to preserve joint distributions and a regression from the plant productivity literature. Supplementary materials are available online.

KEY WORDS: Edit; Hit-and-Run; Mixture; Survey; Truncation.

1. INTRODUCTION

Most economic data sets suffer from missing data. For example, among the plants surveyed in the 2007 U. S. Census of Manufactures and across all 6-digit NAICS industries, 27% of plants are missing total value of shipments, 32% of plants are missing book values of assets, and 42% of plants are missing cost of materials. As is well-known (Little and Rubin 2002), using only the complete cases (all variables are observed) or available cases (all variables for the particular analysis are observed) can cause problems for statistical inferences, even when variables are missing at random (Rubin 1976). By discarding cases with partially observed data, both approaches sacrifice information that could be used to increase precision. Further, using only available cases complicates model comparisons, since different models could be estimated on different sets of cases; standard model comparison strategies do not account for such disparities. For data collected with complex survey designs, using only available cases complicates survey-weighted inference, since the original weights generally are no longer meaningful for the available sample.

An alternative to complete/available cases is to fill in the missing items with multiple imputations (Rubin 1987). The basic idea is to simulate values for the missing items by sampling repeatedly from predictive distributions. This creates $m > 1$ completed data sets that can be analyzed or, as relevant for many data producers, disseminated to the public. When the imputation models meet certain conditions (Rubin 1987, Chapter 4), analysts of the m completed data sets can make valid inferences using complete-data statistical methods and software. Specifically, the analyst computes point and variance estimates of interest with each data set and combines these estimates using simple formulas developed by Rubin (1987). These formulas serve to propagate the uncertainty introduced by missing data and imputation through the analyst's inferences. See Reiter and Raghunathan (2007) for a review of multiple imputation.

In many settings, imputations of numerical variables must respect linear constraints. These constraints arise particularly in the context of editing faulty data, i.e., values that are not plausible (e.g., a plant with one million employees, or a large plant with an average salary per employee of $1 million) or that are inconsistent with other values on the data file (e.g., an establishment reporting an entry year of 2012 that also reports non-zero employment in earlier years). For example, the U. S. Census Bureau requires numerical variables to satisfy an extensive set of ratio constraints intended to catch errors in survey reporting or data capture in the Census of Manufactures, the Annual Survey of Manufactures, and the Services Sectors Censuses (SSC) in the Economic Census (Winkler and Draper 1996; Draper and Winkler 1997; Thompson et al. 2001). Examples of non-U. S. economic data products subject to editing include the Survey of Average Weekly Earnings of the Australian Bureau of Statistics (Lawrence and McDavitt 1994), the Structural Business Survey of Statistics Netherlands (Scholtus and Goksen 2012), and the Monthly Inquiry into the Distribution and Services Sector of the U. K. Office for National Statistics (Hedlin 2003).

Despite the prevalence and prominence of such contexts, there has been little work on imputation of numerical variables under constraints (De Waal et al. 2011). Most statistical agencies use hot deck imputation: for each record with missing values, find a record with complete data that is similar on all observed values, and use that record's values as imputations. However, many hot deck approaches do not guarantee that all constraints are satisfied and, as with hot deck imputation in general, can fail to describe multivariate relationships and typically result in underestimation of uncertainty (Little and Rubin 2002). Tempelman (2007) proposes to impute via truncated multivariate normal distributions that assign zero support to multivariate regions not satisfying the constraints. This approach presumes relationships in the data are well-described by a single multivariate normal distribution, which for skewed economic variables is not likely in practice even after transformations (which themselves can be tricky to implement because the constraints may be difficult to express on the transformed scale). Raghunathan et al. (2001) and Tempelman (2007) propose to use sequential regression imputation, also called multiple imputation by chained equations (Van Buuren and Oudshoorn 1999), with truncated conditional models. The basic idea is to impute each variable y_j with missing data from its regression on some function of all other variables $\{y_s : s \neq j\}$, enforcing zero support for the interval values of y_j not satisfying the constraints. While more flexible than multivariate normal imputation, this technique faces several challenges in practice. Relations among the variables may be interactive and nonlinear, and identifying these complexities in each conditional model can be a laborious task with no guarantee of success. Furthermore, the set of specified conditional distributions may not correspond to a coherent joint distribution and thus is subject to odd theoretical behaviors. For example, the order in which variables are placed in the sequence could impact the imputations (Baccini et al. 2010).

In this article, we propose a fully Bayesian, flexible joint modeling approach for multiple imputation of missing or faulty data subject to linear constraints. To do so, we use a Dirichlet process mixture of multivariate normal distributions as the base imputation engine, allowing the data to inform the appropriate number of mixture components. The mixture model allows for flexible joint modeling, as it can reflect complex distributional and dependence structures automatically (MacEachern and Müller 1998), and it is readily fit with MCMC techniques (Ishwaran and James 2001). We restrict the support of the mixture model only to regions that satisfy the constraints. To sample from this restricted mixture model, we utilize the Hit-and-Run sampler (Boneh and Golan 1979; Smith 1980), thereby guaranteeing that the imputations satisfy all constraints. We illustrate the constrained imputation procedure with data from the Colombian Annual Manufacturing Survey, estimating descriptive statistics and coefficients from a regression used in the literature on plant productivity.

The remainder of the article is organized as follows. In Section 2, we review automatic editing and imputation processes, which serves as a motivating setting for our work. In

Section 3, we present the constrained Dirichlet process mixture of multivariate normals (CDPMMN) multiple imputation engine, including the MCMC algorithm for generating imputations. In Section 4, we apply the CDPMMN to the Colombian manufacturing data. In Section 5, we conclude with a brief discussion of future research directions.

2. BACKGROUND ON DATA EDITING

Although the CDPMMN model is of interest as a general estimation and imputation model, it is especially relevant for national statistical agencies and survey organizations—henceforth all called agencies—seeking to disseminate high-quality data to the public. These agencies typically dedicate significant resources to imputing missing data and correcting faulty data before dissemination. For example, Granquist and Kovar (1997) estimated that national statistical agencies spend 40% of the total budget for business surveys (20% for household surveys) on edit and imputation processes, and in an internal study of 62 products at Statistics Sweden, Norberg (2009) reports that the agency allocated 32.6% of total costs in business surveys to editing processes. Improving these processes serves as a primary motivation for the CDPMMN imputation engine, as we now describe.

Since the 1960s, national statistical agencies have leveraged the expertise of subject matter analysts with automated methods for editing and imputing numerical data. The subject matter experts create certain consistent rules, called edit rules or edits, that describe feasible regions of data values. The automated methods identify and replace values that fail the edit rules.

For numerical data, typical edit rules include *range restrictions*, e.g., marginal lower and upper bounds on a single variable, *ratio edits*, e.g., lower and upper bounds on a ratio of two variables, and *balance edits*, e.g., two or more items adding to a total item. In this article, we consider range restrictions and ratio edits, but not balance edits. Formally, let x_{ij} be

Table 1: Exemplary ratio edits in the SSC. SLS = sales/receipts, APR = annual payroll, OPX = annual operating expenses, GOP = purchases, EMP = employment, QPR = first quarter payroll, BIN = beginning inventory, and ENV = ending inventory.

Ratio	Lower bound	Upper bound	Industry avg.	No. of records	No. of fail	Perc. of fail
SLS / APR	1.0	152.8	21.8	4,785	223	4.7 %
SLS / OPX	1.0	57.5	9.4	3,963	224	5.7 %
SLS / GOP	0.5	2.2	1.1	4,097	335	8.2 %
SLS / BIN	0.1	358.0	16.4	3,685	188	5.1 %
APR / EMP	1.4	106.0	29.2	4,804	55	1.1 %
APR / QPR	2.6	7.6	4.3	4,565	87	1.9 %
OPX / APR	1.0	9.4	2.6	3,877	333	8.6 %
ENV / BIN	0.3	2.1	0.9	3,681	148	4.0 %

the value of the jth variable for the ith subject, where $i = 1, \ldots, n$ and $j = 1, \ldots, p$. For variables with range restrictions, we have $L_j \leq x_{ij} \leq U_j$, where U_j and L_j are agency-fixed upper and lower limits, respectively. For any two variables (x_{ij}, x_{ik}) subject to ratio edits, we have $L_{jk} \leq x_{ij}/x_{ik} \leq U_{jk}$, where again U_{jk} and L_{jk} are agency-fixed upper and lower limits, respectively. An exemplary system of ratio edits is displayed in Table 1. These rules, defined by the Census Bureau for data collected in the SSC, were used to test editing routines in 1997 Economic Census (Thompson et al. 2001).

The set of edit rules defines a feasible region (convex polytope) comprising all potential records passing the edits. A potential data record $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})'$ satisfies the edits if and only if the constraints $A\boldsymbol{x}_i \leq \boldsymbol{b}$ for some matrix A and vector $\boldsymbol{b}$ are satisfied. Any record that fails the edits is replaced (imputed) by a record that passes the edits.

Finding records that fail rules is straightforward, at least conceptually, with automated routines and modern computing. However, identifying the faulty values within any record (the error localization procedure) and correcting those erroneous values (the imputation procedure) are far more complicated tasks. Among national statistical agencies, the most commonly used error localization procedure is that of Fellegi and Holt (1976), who develop a set covering algorithm that identifies the minimum number of values to change in the data to satisfy all edit rules. Fellegi and Holt (1976) error localization algorithms are the basis for many automated edit systems, including the Census Bureau's SPEER (Draper and Winkler 1997), Statistics Canada's GEIS (Whitridge and Kovar 1990), Spanish National Statistical Institute's DIA (Garcia-Rubio and Villan 1990), Statistics Netherlands' CherryPi (De Waal 2000), and Istituto Nazionaledi Statistica's SCIA (Manzari 2004). We do not discuss error localization steps further, as our concern is with the imputation procedure; that is, we assume that the erroneous fields of survey data that violate edit rules have been detected and blanked by the agency, and we focus on imputing the blanked values.

3. MODEL AND ESTIMATION

We now present the CDPMMN imputation engine. We begin in Section 3.1 by reviewing Dirichlet mixtures of multivariate normal distributions without any truncation or missing data. We adapt the model in Section 3.2 to handle truncation and present an MCMC algorithm to estimate the model, again assuming no missing data. We modify this algorithm in Section 3.3. to account for missing data, making use of a Hit-and-Run algorithm in the MCMC steps. We discuss several implementation issues in Section 3.4.

3.1 Dirichlet Process Mixtures of Normals Without Truncation

Let $Y_n = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ comprise n complete observations not subject to edit constraints, where each $\boldsymbol{y}_i$ is a p-dimensional vector. To facilitate modeling, we assume that the analyst standardizes each variable before modeling to have observed mean equal to zero and observed standard deviation equal to one; see Appendix B in the Supplementary Material for details of the standardization. We suppose that each individual i belongs to exactly one of $K < \infty$ latent mixture components. For $i = 1, \ldots, n$, let $z_i \in \{1, \ldots, K\}$ indicate the component of individual i, and let $\pi_k = \Pr(z_i = k)$. We assume that $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ is the same for all individuals. Within any component k, we suppose that the p variables follow a component-specific multivariate normal distribution with mean $\boldsymbol{\mu}_k$ and variance Σ_k. Let $\Theta = (\mu, \Sigma, \boldsymbol{\pi})$, where $\mu = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K)$ and $\Sigma = (\Sigma_1, \ldots, \Sigma_K)$. Mathematically, the finite mixture model can be expressed as

$$\begin{aligned} \boldsymbol{y}_i \mid z_i, \mu, \Sigma &\sim \mathrm{N}(\boldsymbol{y}_i \mid \boldsymbol{\mu}_{z_i}, \Sigma_{z_i}) && (1) \\ z_i \mid \boldsymbol{\pi} &\sim \mathrm{Multinomial}(\pi_1, \ldots, \pi_K). && (2) \end{aligned}$$

Marginalizing over z_i, this model is equivalent to

$$p(\boldsymbol{y}_i \mid \Theta) = \sum_{k=1}^{K} \pi_k \mathrm{N}(\boldsymbol{y}_i \mid \boldsymbol{\mu}_k, \Sigma_k). \qquad (3)$$

As a mixture of multivariate normal distributions, the model is flexible enough to capture distributional features like skewness and non-linear relationships that a single multivariate normal distribution would fail to encode.

To complete the Bayesian specification, we require prior distributions for Θ. As suggested by Lavine and West (1992), for $(\boldsymbol{\mu}_k, \Sigma_k)$ we use

$$\begin{aligned} \boldsymbol{\mu}_k \mid \Sigma_k &\sim \mathrm{N}(\boldsymbol{\mu}_0, h^{-1}\Sigma_k) && (4) \\ \Sigma_k &\sim \mathrm{InverseWishart}(f, \Phi), && (5) \end{aligned}$$

where f is the degrees of freedom, $\Phi = diag(\phi_1, \ldots, \phi_p)$, and

$$\phi_j \sim \mathrm{Gamma}(a_\phi, b_\phi) \qquad (6)$$

with mean a_ϕ / b_ϕ for $j = 1, \ldots, p$. These conditionally conjugate prior distributions simplify the MCMC computations. We defer discussion of these hyperparameter values until Section 3.2, when we present the constrained model.

For the component weights, $\boldsymbol{\pi}$, we use the stick-breaking representation of a truncated Dirichlet process (Sethuraman 1994; Ishwaran and James 2001), shown in (7) – (8) below.

$$\begin{aligned} \pi_k &= v_k \prod_{g<k} (1 - v_g) \text{ for } k = 1, \ldots, K & (7) \\ v_k &\sim \text{Beta}(1, \alpha) \text{ for } k = 1, \ldots, K-1; \ v_K = 1 & (8) \\ \alpha &\sim \text{Gamma}(a_\alpha, b_\alpha). & (9) \end{aligned}$$

Here, (a_α, b_α) are analyst-supplied constants. Following Dunson and Xing (2009), we recommend setting $(a_\alpha = .25, b_\alpha = .25)$, which represents a small prior sample size and hence vague specification for Gamma distributions. This ensures that the information from the data dominates the posterior distribution (Escobar and West 1995). The specification of prior distributions in (7) – (9) encourages π_k to decrease stochastically with k. When α is very small, most of the probability in $\boldsymbol{\pi}$ is allocated to the first few components, thus reducing the risks of over-fitting the data as well as increasing computational efficiency. We note that Dirichlet process distributions are widely used for mixture models in economic analyses (e.g., Hirano 2002; Gilbride and Lenk 2010).

As a prior distribution for $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$, one also could use a symmetric Dirichlet distribution, $\boldsymbol{\pi} \sim \text{Dir}(a/K, \ldots, a/K)$, for example with $a = 1$. We expect that imputations based on the truncated Dirichlet process prior distribution and this symmetric Dirichlet distribution generally will not differ noticeably. We use the truncated Dirichlet process prior distribution primarily because (i) it has worked well in other applications of mixture models for multiple imputation of missing data (Si and Reiter 2013; Manrique-Vallier and Reiter forthcoming) and (ii) in our experience, it facilitates reasonably fast MCMC convergence.

With specified hyperparameters, the model can be estimated using a Gibbs sampler, as each full conditional distribution is in closed form; see Ishwaran and James (2001).

3.2 Dirichlet Process Mixtures of Normals With Truncation

The model in Section 3.1 has unrestricted support and hence is inappropriate for imputation under linear constraints. Instead, when $\boldsymbol{y}_i$ is restricted to lie in some feasible region $\mathcal{A}$, we need to replace (3) with

$$p(\boldsymbol{y}_i \mid \Theta, \mathcal{A}) \;=\; \frac{1}{h(\mathcal{A}, \Theta)} \sum_{k=1}^{K} \pi_k \mathrm{N}(\boldsymbol{y}_i \mid \boldsymbol{\mu}_k, \Sigma_k) I(\boldsymbol{y}_i \in \mathcal{A}), \quad i = 1, \ldots, n, \tag{10}$$

where $I(\cdot) = 1$ when the condition inside the parentheses is true, and $I(\cdot) = 0$ otherwise. Here, $h(\mathcal{A}, \Theta)$ is the normalizing constant such that

$$h(\mathcal{A}, \Theta) \;=\; \int_{\{\boldsymbol{y}: \boldsymbol{y} \in \mathcal{A}\}} \sum_{k=1}^{K} \pi_k \mathrm{N}(\boldsymbol{y} \mid \boldsymbol{\mu}_k, \Sigma_k) d\boldsymbol{y}. \tag{11}$$

Equivalently, using the representation conditional on z_i, we need to replace (1) with

$$p(\boldsymbol{y}_i \mid z_i, \mu, \Sigma, \mathcal{A}) \;=\; \frac{1}{h(\mathcal{A}, \boldsymbol{\mu}_{z_i}, \Sigma_{z_i})} \mathrm{N}(\boldsymbol{y}_i \mid \boldsymbol{\mu}_{z_i}, \Sigma_{z_i}) I(\boldsymbol{y}_i \in \mathcal{A}) \tag{12}$$

where

$$h(\mathcal{A}, \boldsymbol{\mu}_k, \Sigma_k) = \int_{\{\boldsymbol{y}: \boldsymbol{y} \in \mathcal{A}\}} \mathrm{N}(\boldsymbol{y} \mid \boldsymbol{\mu}_k, \Sigma_k) d\boldsymbol{y}. \tag{13}$$

We leave all other features of the model as described in (2) and (4) – (9).

Unfortunately, due to the truncation, full conditional distributions depending on the likelihood do not have conjugate forms, thus complicating the MCMC. To avoid computation with $h(\mathcal{A}, \Theta)$, we use a data augmentation technique developed by O'Malley and Zaslavsky (2008) in which we (i) conceive of the observed values as a sample from a larger, hypothetical sample not subject to the constraints, (ii) construct the hypothetical sample by augmenting the observed data with values from outside of $\mathcal{A}$, and (iii) use the augmented data to estimate parameters using the usual, unconstrained Gibbs sampler. With appropriate prior distributions, the resulting parameter estimates correspond to those from the CDPMMN model.

Specifically, we assume that there exists a hypothetical sample of unknown size $N > n$ records, $Y_N = \{Y_n, Y_{N-n}\}$, where $Y_n = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ includes the n observed values and $Y_{N-n} = \{\boldsymbol{y}_{n+1}, \ldots, \boldsymbol{y}_N\}$ includes $N - n$ augmented values, such that

$$p(Y_N \mid \Theta) \; = \; \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k \mathrm{N}(\boldsymbol{y}_i \mid \boldsymbol{\mu}_k, \Sigma_k). \tag{14}$$

We consider the observed data $Y_n = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ to be a sample from Y_N with probability $h(\mathcal{A}, \Theta)$. Thus, for known N, the number of cases in $\mathcal{A}$ has distribution

$$n \mid N, \Theta, \mathcal{A} \; \sim \; \mathrm{Binomial}(N, h(\mathcal{A}, \Theta)). \tag{15}$$

We use the prior distribution suggested by Meng and Zaslavsky (2002) and O'Malley and Zaslavsky (2008), $p(N) \propto 1/N$ for $N > n$, so that

$$N - n \mid n, \Theta, \mathcal{A} \; \sim \; \mathrm{NegativeBinomial}\left(n, 1 - h(\mathcal{A}, \Theta)\right). \tag{16}$$

With this construction, we can estimate Θ without computing $h(\mathcal{A}, \Theta)$ using an MCMC algorithm. Let $Z_N = \{Z_n, Z_{N-n}\}$ be the set of membership indicators corresponding to Y_N, where $Z_n = \{z_1, \ldots, z_n\}$ and $Z_{N-n} = \{z_{n+1}, \ldots, z_N\}$. For each group $k = 1, \ldots, K$, let $N_k = \sum_{i=1}^{N} I(z_i = k)$ be the number of cases in the augmented sample in group k; let $\bar{\boldsymbol{y}}_k = \sum_{\{i: z_i = k\}} \boldsymbol{y}_i / N_k$; and, let $S_k = \sum_{\{i: z_i = k\}} (\boldsymbol{y}_i - \bar{\boldsymbol{y}}_k)(\boldsymbol{y}_i - \bar{\boldsymbol{y}}_k)'$. After initialization, the MCMC algorithm proceeds with the following Gibbs steps.

S1. For $k = 1, \ldots, K$, sample values of $(\boldsymbol{\mu}_k, \Sigma_k)$ from the full conditionals,

$$\boldsymbol{\mu}_k \mid \Sigma_k, Y_N, Z_N \sim \mathrm{N}\left(\boldsymbol{\mu}_k^*, \frac{1}{h}\Sigma_k\right), \quad \Sigma_k \mid Y_N, Z_N \sim \text{InverseWishart}(f_k, \Phi_k)$$

where $\boldsymbol{\mu}_k^* = (N_k \bar{\boldsymbol{y}}_k + h\boldsymbol{\mu}_0)/(N_k + h)$, $f_k = f + N_k$, $\Phi_k = \Phi + S_k + (\boldsymbol{\mu}_k^* - \boldsymbol{\mu}_0)(\boldsymbol{\mu}_k^* - \boldsymbol{\mu}_0)'/(1/N_k + 1/h)$.

S2. For $k = 1, \ldots, K-1$, sample values of v_k from the full conditional,

$$v_k \mid Z_N, \alpha \ \sim \ \text{Beta}\left(1 + N_k, \alpha + \sum_{g>k} N_g\right).$$

Set $v_K = 1$. For $k = 1, \ldots, K$, let $\pi_k = v_k \prod_{g<k}(1 - v_g)$ as in (7).

S3. For $j = 1, \ldots, p$, sample each ϕ_l from the full conditional,

$$\phi_l \mid \Sigma \ \sim \ \text{Gamma}\left(a_\phi + \frac{K(p+1)}{2}, b_\phi + \frac{1}{2}\sum_{k=1}^{K} \Sigma_{k(r,r)}^{-1}\right)$$

where $\Sigma_{k(r,r)}^{-1}$ is the rth diagonal element of Σ_k^{-1}.

S4. Given $\boldsymbol{\pi}$, sample α from the full conditional,

$$\alpha \mid \boldsymbol{\pi} \ \sim \ \text{Gamma}\left(a_\alpha + K - 1, b_\alpha - \log \pi_K\right).$$

S5. For $i = 1, \ldots, n$, sample each z_i from the full conditional,

$$z_i \mid \boldsymbol{y}_i, \Theta \ \sim \ \text{Multinomial}(\pi_{i1}^*, \ldots, \pi_{iK}^*),$$

where $\pi_{ik}^* = \pi_k \mathrm{N}(\boldsymbol{y}_i \mid \boldsymbol{\mu}_k, \Sigma_k) / \left\{\sum_{g=1}^{K} \pi_g \mathrm{N}(\boldsymbol{y}_i \mid \boldsymbol{\mu}_g, \Sigma_g)\right\}$.

S6. Sample (N, Z_{N-n}, Y_{N-n}) jointly from their full conditional distribution following the approach suggested by O'Malley and Zaslavsky (2008). Specifically, based on the result in (16), we draw each $(z_i, \boldsymbol{y}_i) \in (Z_{N-n}, Y_{N-n})$ from a negative binomial data-generating process. To begin, set $c_{in} = c_{out} = 0$. Then, perform the following steps.

S6.1. Draw $z^* \sim \text{Multinomial}(\pi_1, \ldots, \pi_K)$.

S6.2. Draw $\boldsymbol{y}^* \mid z^* \sim \text{N}(\boldsymbol{\mu}_{z^*}, \Sigma_{z^*})$.

S6.3. If $\boldsymbol{y}^* \in \mathcal{A}$, set $c_{in} = c_{in} + 1$.

If $\boldsymbol{y}^* \in \mathcal{A}^c$, set $c_{out} = c_{out} + 1$, $\boldsymbol{y}_{n+c_{out}} = \boldsymbol{y}^*$, and $z_{n+c_{out}} = z^*$.

S6.4. Repeat S6.1 through S6.3 until $c_{in} = n$.

S6.5. Let $N = n + c_{out}$.

For the prior distributions in (4) – (6), we recommend using a prior mean of $\boldsymbol{\mu}_0 = \mathbf{0}$ since each variable is mean-centered, using $f = p + 1$ degrees of freedom to ensure a proper distribution without overly constraining Σ, and setting $h = 1$ mostly for convenience. We recommend setting (a_ϕ, b_ϕ) to be modest but not too small, so as to allow substantial prior mass at modest-sized variances. Following Dellaportas and Papageorgiou (2006), we use $a_\phi = b_\phi = .25$. In the Colombia data illustration, results were insensitive to other sensible choices of hyperparameter values; details of the sensitivity analysis are in Appendix A, available as on-line Supplementary Material.

3.3 Accounting for Missing Data: Hit-and-Run Algorithm

Having developed an MCMC algorithm for fitting the CDPMMN model without any missing values, we now extend to include imputation of missing data or, equivalently, imputation of blanked values due to edit rules. Here, we assume the missing data mechanism is ignorable (Rubin 1976). Without loss of generality, suppose that the first $s \leq n$ records in Y_n have some missing values. Let $Y_s = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s\}$ be these first s records, and $Y_{n-s} = \{\boldsymbol{y}_{s+1}, \ldots, \boldsymbol{y}_n\}$ be the set of fully observed records. For each $\boldsymbol{y}_i \in Y_s$, let $\boldsymbol{y}_i = (\boldsymbol{y}_{i,0}, \boldsymbol{y}_{i,1})$, where $\boldsymbol{y}_{i,0}$ comprises the missing values and $\boldsymbol{y}_{i,1}$ comprises the observed values. Finally, let $Y_{s,0} = \{\boldsymbol{y}_{1,0}, \ldots, \boldsymbol{y}_{s,0}\}$, and let $Y_{s,1} = \{\boldsymbol{y}_{1,1}, \ldots, \boldsymbol{y}_{s,1}\}$.

Formally, we seek to estimate the posterior distribution, $f(Y_{s,0}, Y_{N-n}, Z_N, \Theta, \Phi, \alpha, N \mid Y_{s,1}, Y_{n-s}, \mathcal{A})$. We do so in two steps, namely (i) given a draw of $Y_{s,0}$ satisfying $\mathcal{A}$, draw values of $(Y_{N-n}, Z_N, \Theta, \Phi, \alpha, N)$ using S1 – S6 from Section 3.2, and (ii) given a draw of $(Y_{N-n}, Z_N, \Theta, \Phi, \alpha, N)$, draw imputations for $Y_{s,0}$ satisfying $\mathcal{A}$ using a Metropolis version of the Hit-and-Run (H&R) sampler (Chen and Schmeiser 1993). Thus, we only need to add an imputation step to the MCMC algorithm from Section 3.2, as we now describe.

We begin by finding the region of feasible imputations for each $\boldsymbol{y}_{i,0}$, which we write as $\mathcal{A}_i = \{\boldsymbol{y}_{i,0}; \boldsymbol{y}_i = (\boldsymbol{y}_{i,0}, \boldsymbol{y}_{i,1}) \in \mathcal{A}\}$. For systems of linear constraints like those in Table 1 and typically employed in edit-imputation contexts, $\mathcal{A}_i$ can be defined using matrix algebra operations; see Appendix C, Supplementary Material, for an illustrative example. For more complex linear constraints, feasible regions can be found by linear programming and related optimization techniques.

The H&R sampler proceeds as follows. At any iteration t of the MCMC sampler, we presume the current values of each $\boldsymbol{y}_{i,0}$, say $\boldsymbol{y}_{i,0}^{(t)}$, where $i = 1, \ldots, s$, satisfy the constraints; see Section 3.4 for a suggestion to initialize the chain at feasible values. The basic idea of the H&R sampler is to pick a random direction in R^{p_i}, where p_i is the number of variables in $\boldsymbol{y}_{i,0}$; follow that direction starting from $\boldsymbol{y}_{i,0}^{(t)}$ until hitting the boundary of $\mathcal{A}_i$, say at some point $\boldsymbol{b}_{i,0}^{(t)}$; and, sample a new point along the line segment with end points $(\boldsymbol{y}_{i,0}^{(t)}, \boldsymbol{b}_{i,0}^{(t)})$. For convex $\mathcal{A}_i$, the selected point is guaranteed to be inside the feasible region. Formally, we implement this process via the Metropolis step in S7 below.

S7. For $i = 1, \ldots, s$, update each current value $\boldsymbol{y}_{i,0}^{(t)}$ with a Metropolis accept/reject step as follows.

S7.1. Propose a direction $\boldsymbol{d}^* \sim r(\cdot)$ where $r(\cdot)$ is a uniform distribution on the surface of the unit sphere in $\mathbb{R}^p$.

S7.2. Find the set of candidate proposal distances, which we write as $\Omega(\boldsymbol{d}^*, \boldsymbol{y}_{i,0}^{(t)}) = \{\lambda \in \mathbb{R} : \boldsymbol{y}_{i,0}^{(t)} + \lambda \boldsymbol{d}^* \in \mathcal{A}_i\}$.

S7.3. Draw a signed distance $\lambda^* \in \Omega(\boldsymbol{d}^*, \boldsymbol{y}_{i,0}^{(t)})$ from a uniform distribution with the density $1/\mathcal{M}\left\{\Omega(\boldsymbol{d}^*, \boldsymbol{y}_{i,0}^{(t)})\right\}$ where $\mathcal{M}$ is a Lebesgue measure.

S7.4. Accept or reject the proposal $\boldsymbol{y}_{i,0}^q = \boldsymbol{y}_{i,0}^{(t)} + \lambda^* \boldsymbol{d}^*$ with the acceptance probability ρ_i, where

$$\rho_i = \min\left\{1, \frac{p(\boldsymbol{y}_{i,0}^q \mid z_i, \mu, \Sigma, \mathcal{A}_i)}{p(\boldsymbol{y}_{i,0}^{(t)} \mid z_i, \mu, \Sigma, \mathcal{A}_i)}\right\}.$$

Note that the acceptance probability can be calculated by using the fact that the conditional distribution of $\boldsymbol{y}_{i,0}$ is proportional to that of $\boldsymbol{y}_i$, i.e., $p(\boldsymbol{y}_{i,0} \mid z_i, \mu, \Sigma, \mathcal{A}_i) \propto p(\boldsymbol{y}_i \mid z_i, \mu, \Sigma, \mathcal{A})$ as in (12). After canceling common terms, the acceptance probability simplifies to $\rho_i = \min\left\{1, \mathrm{N}(\boldsymbol{y}_i^q \mid \boldsymbol{\mu}_{z_i}, \Sigma_{z_i})/\mathrm{N}(\boldsymbol{y}_i^{(t)} \mid \boldsymbol{\mu}_{z_i}, \Sigma_{z_i})\right\}$ where $\boldsymbol{y}_i^q = (\boldsymbol{y}_{i,0}^q, \boldsymbol{y}_{i,1})$.

Because the H&R sampler randomly moves in any direction, it can cover multivariate spaces more efficiently than typical Gibbs samplers that move one direction in each conditional step. The efficiency gain can be substantial when the sample space is a convex polytope and when variables are highly correlated or sharply constrained (Chen and Schmeiser 1993; Berger 1993). Lovász and Vempala (2006) prove that the H&R sampler mixes fast, and that mixing time does not depend on the choice of starting points. Hence, for imputing multivariate data subject to the constraints, the H&R sampler offers a potentially significant computational advantage over typical Gibbs steps.

To obtain m completed data sets for use in multiple imputation, one selects m of the sampled $Y_{s,0}$ after MCMC convergence. These data sets should be spaced sufficiently to be approximately independent (given the observed data). This involves thinning the MCMC samples so that the autocorrelations among parameters are close to zero.

3.4 Implementation Issues

In this section, we discuss several features of implementing the CDPMMN imputation engine, beginning with the choice of K. Setting K too small can result in insufficient flexibility to capture complex features of the distributions, but setting K too large is computationally inefficient. We recommend initially setting K to a reasonably large value, say $K = 50$. Analysts can examine the posterior distribution of the number of unique values of z_i across MCMC iterates to diagnose if K is large enough. Significant posterior mass at a number of classes equal to K suggests that the truncation limit should be increased. We note that one has to assume a finite K because of the truncation; otherwise, the model can generate arbitrary components with nearly all mass outside $\mathcal{A}$.

One also has to select hyperparameter values, as we discussed in Section 3.2. Of note here is the value of h in (4), which determines the spread of mean vectors. For example, small values of h generally imply that, *a priori*, $\boldsymbol{\mu}_k$ can be far from $\boldsymbol{\mu}_0$. Empirically, our experience is that small values of h can generate large values of N, the number of augmented values, which can slow the MCMC algorithm. We recommend avoiding very small h, say by setting $h > .1$, to reduce this computational burden. As noted in Section 3.2 and the sensitivity analysis in Appendix A, our application results are robust to changes of h and other hyperparameters.

To initialize the MCMC chains when using the H&R sampler, we need to choose a set of points all within the polytope defined by $\mathcal{A}$. To do so, we find the extreme points of the polytope along each dimension by solving linear programs (Tervonen et al. 2013), and set the starting point of the H&R chain as the arithmetic mean of the extreme points. We run this procedure for finding starting values using the *R*-package "hitandrun" (available at http://cran.r-project.org/web/packages/hitandrun/index.html). In our empirical application, other options of finding starting points within the feasible region, including randomly weighted extreme points or vertex representation, do not change the final results.

With MCMC algorithms it is essential to examine convergence diagnostics. Due to the complexity of the models and many parameters, as well as the truncation and missing values, monitoring convergence of chains is not straightforward. We suggest that MCMC diagnostics focus on the draws of N, α, the ordered samples of $\boldsymbol{\pi}$, and, when feasible, values of the $\boldsymbol{\pi}$-weighted averages of $\boldsymbol{\mu}_k$ and Σ_k, e.g., $\sum_k \pi_k \boldsymbol{\mu}_k$, rather than specific component parameters, $\{\boldsymbol{\mu}_k, \Sigma_k\}$. Specific component parameters are subject to label switching among the mixture components, which complicates interpretation of the components and MCMC diagnostics; we note that label switching does not affect the multiple imputations.

4. EMPIRICAL EXAMPLE: COLOMBIAN MANUFACTURING DATA

We illustrate the CDPMMN with multiple imputation and analysis of plant-level data from the Colombian Annual Manufacturing Survey from 1977 to 1991 (except for 1981, for which we have no data). Similar data have been used in analyses of plant-level productivity (e.g., Fernandes 2007; Petrin and White 2011). Following Petrin and White (2011), we use the seven variables in Table 2. We remove a small number of plants with missing item data or nonpositive values, so as to make a clean file on which to evaluate the imputations. The resulting annual sample sizes range from 5,770 to 6,873.

To illustrate the performance of the CDPMMN imputation engine, we introduce range restrictions on each variable (see Table 2) and ratio edits on each pair of variables (see

Table 2: Variables in the Colombian Annual Manufacturing Survey, with range restrictions that we introduce for illustration.

Variable	Definition	Range
RVA	real valued-added	$0.2 \leq \text{RVA} \leq 6,600,000$
CAP	capital in real terms	$0.07 \leq \text{CAP} \leq 9,000,000$
SL	skilled labor	$0.99 \leq \text{SL} \leq 2,800$
USL	unskilled labor	$0.99 \leq \text{USL} \leq 6,700$
RMU	real material used in products	$0.03 \leq \text{RMU} \leq 9,500,000$
SW	wages paid to skill labor	$3.3 \leq \text{SW} \leq 11,000,000$
USW	wages paid to unskilled labor	$3.5 \leq \text{USW} \leq 16,000,000$

Table 3: Introduced ratio edits in Colombian Manufacturing Survey.

Ratio	Minimum	Maximum	Ratio	Minimum	Maximum
RVA/CAP	7.0e-05	6,100	SL/USL	3.0e-03	270
RVA/SL	7.2e-02	500,000	SL/RMU	8.0e-06	1,350
RVA/USL	4.2e-02	1,000,000	SL/SW	4.0e-05	1.2
RVA/RMU	6.0e-05	1,300,000	SL/USW	4.0e-06	5.0
RVA/SW	5.9e-05	1,000	USL/RMU	1.0e-06	13,000
RVA/USW	5.3e-05	3,333	USL/SW	1.0e-06	2.5
CAP/SL	2.4e-02	100,000	USL/USW	1.0e-06	1.5
CAP/USL	1.5e-02	333,333	RMU/SW	2.2e-07	1,250
CAP/RMU	4.0e-05	1,111,111	RMU/USW	7.1e-08	2,500
CAP/SW	1.5e-05	1,250	SW/USW	4.0e-03	850
CAP/USW	1.7e-05	250			

Table 3) like those used by government agencies in economic data. Limits are wide enough that all existing records satisfy the constraints, thus offering a set of true values to use as a gold standard. However, the limits are close enough to several observed values that the constraints matter, in the sense that estimation and imputation under models that ignore the truncation would have nonnegligible support in the infeasible region. Figure 1 displays a typical example of the linear constraints for two variables, log(USW) and log(CAP).

We then randomly introduce missing values in the observed data, and use the CDPMMN imputation engine to implement multiple imputation of missing values subject to linear

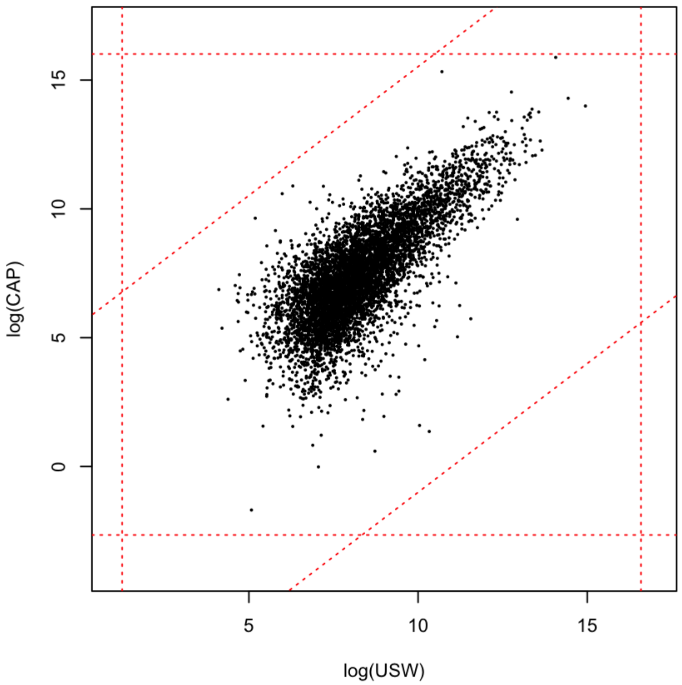

Figure 1: The linear constraints for two variables, log(USW) and log(CAP), for 1982 data

constraints. We do so in two empirical studies: a set of analyses designed to compare one-off inferences from the imputed and original data (before introduction of missing values), and a repeated sampling experiment designed to illustrate efficiency gains in using the CDPMMN over complete-case methods.

In both studies, we make inferences for unknown parameters according to the multiple imputation inferences of Rubin (1987), which we review briefly here. Suppose that the analyst seeks to estimate some quantity Q, such as a regression coefficient or population mean. Let $\{D^{(1)}, \ldots, D^{(m)}\}$ be m completed data sets drawn from the CDPMMN imputation engine. Given completed data, the analyst estimates Q with some point estimator q and estimates its variance with u. For $l = 1, \ldots, m$, let $q^{(l)}$ and $u^{(l)}$ be, respectively, the value of q and u computed with $D^{(l)}$. The analyst uses $\bar{q}_m = \frac{1}{m} \sum_{l=1}^{m} q^{(l)}$ to estimate Q and $T_m = \bar{u}_m + (1 + 1/m) b_m$ to estimate its variance, where $\bar{u}_m = \sum_{l=1}^{m} u^{(l)}/m$ and $b_m = \sum_{l=1}^{m}(q^{(l)} -$

$\bar{q}_m)^2/(m-1)$. For large samples, inferences for Q are obtained from the t-distribution, $(\bar{q}_m - Q) \sim t_{\nu_m}(0, T_m)$, where the degrees of freedom is $\nu_m = (m-1)\left[1 + \bar{u}_m / \{(1+1/m)b_m\}\right]^2$. Tests of significance for multicomponent null hypotheses are derived by Li et al. (1991), Meng and Rubin (1992), and Reiter (2007).

4.1 Comparisons of CDPMMN-Imputed and Original Data

For each year $t = 1977, \ldots, 1991$ other than 1981, let $D_{orig,t}$ be the original data before introduction of missing values. For each year, we generate a set of missing values $Y_{s,t}$ by randomly blanking data from $D_{orig,t}$ as follows: 1,500 records have one randomly selected missing item, 1,000 records have two randomly selected missing items, and 500 records have three randomly selected missing items. One can interpret these as missing data or as blanked fields after an error localization procedure. Missingness rates per variable range from 10.2% to 11.5%. This represents a missing completely at random mechanism (Rubin 1976), so that complete-case analyses based only on records with no missing data, i.e., $Y_{n-s,t}$, would be unbiased but inefficient, sacrificing roughly half of the observations in any one year. We note that these rates of missing data are larger than typical fractions of records that fail edit rules, so as to offer a strong but still realistic challenge to the CDPMMN imputation engine.

We implement the CDPMMN imputation engine using the model and prior distributions described in Section 3 to create $m = 10$ completed data sets, each satisfying all constraints. To facilitate model estimation, we work with the standardized natural logarithms of all variables (take logs first, then standardize using observed data).

After imputing on the transformed scale, we transform back to the original scale before making inferences. Of course, we also transform the limits of the linear inequalities to be consistent with the use of standardized logarithms; see Appendix B, available as on-line Supplementary Material. We run the MCMC algorithm of Section 3 for 10,000 iterations

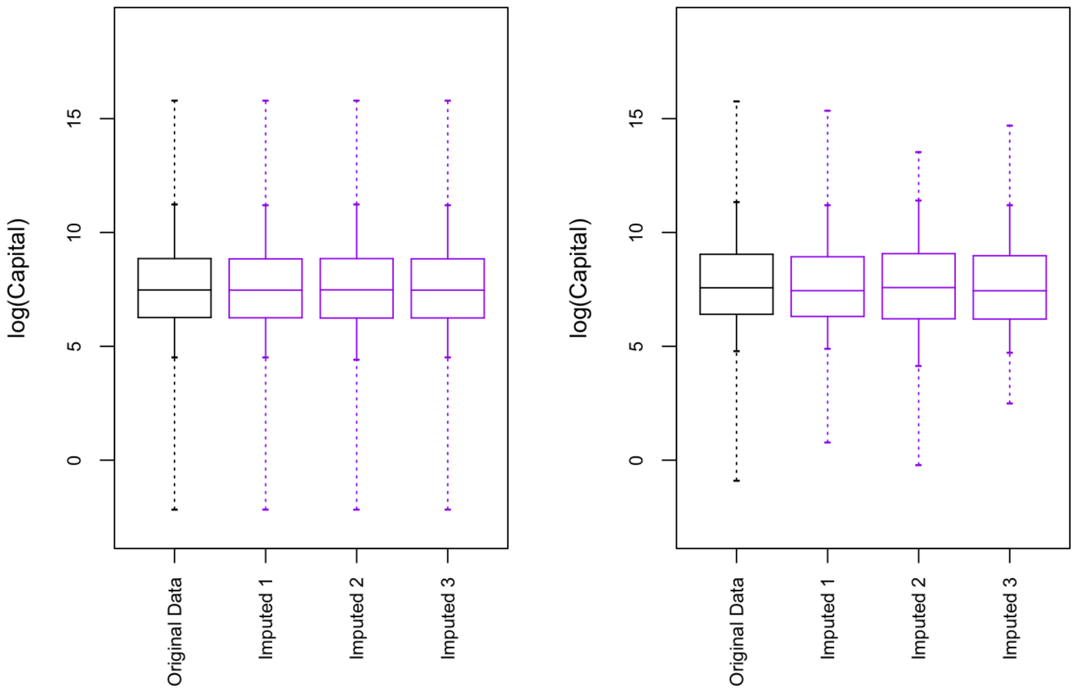

Figure 2: Left panel: Distributions of log(CAP) from original data and three completed data sets for all 6,607 records in the 1991 data. Right panel: Distributions of log(CAP) for 714 blanked values in the original data and the imputed values in three completed data sets. Boxes include 25th, 50th, and 75th percentiles. Dotted lines stretch from minimum value to 5th percentile, and from 95th percentile to maximum value. The marginal distributions in the imputed data sets are similar to those in the original data.

after a burn-in period, and store every 1,000th iterate to obtain the $m = 10$ completed data sets. We use $K = 40$ components, which we judged to be sufficient based on the posterior distributions of the numbers of unique z_i among the n cases.

We begin by focusing on results from the 1991 survey; results from other survey years are similar. Figure 2 summarizes the marginal distributions of log(CAP) in the original and completed data sets. For the full distribution based on all 6,607 cases, the distributions are nearly indistinguishable, implying that the CDPMMN completed data approximate the marginal distributions of the original data well. The distribution of the 714 values of log(CAP) in D_{orig} before blanking is also similar to those in the three completed data sets.

Figure 3 displays a scatter plot of log(USW) versus log(CAP)—at least one of these

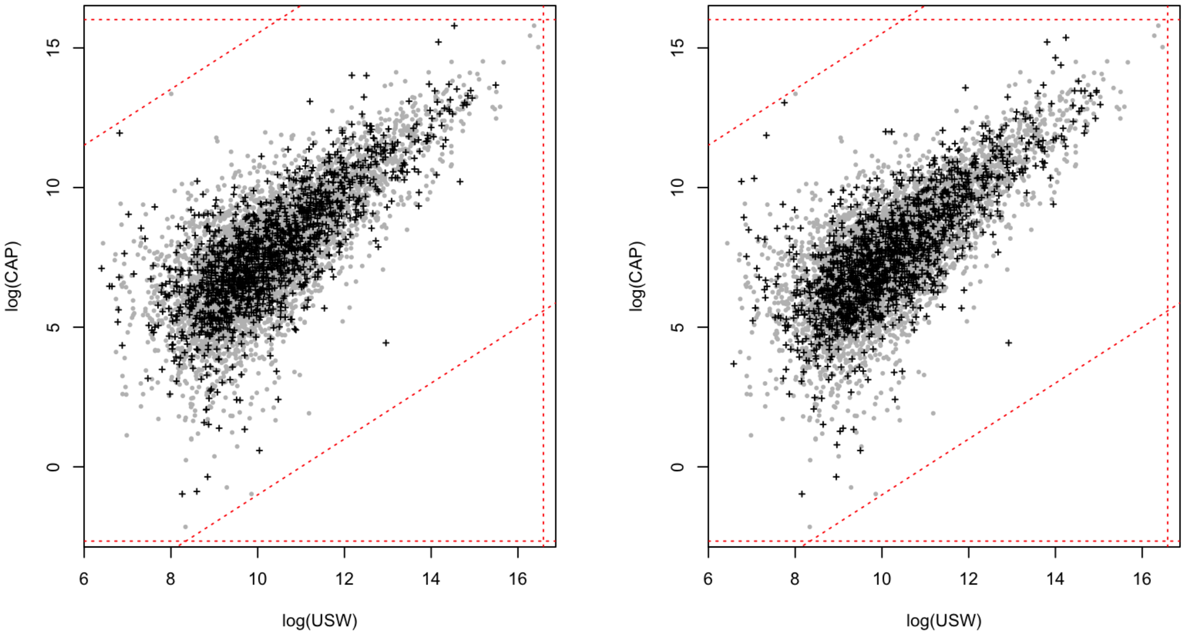

Figure 3: Plots of log(USW) versus log(CAP) for 1991 data. The left plot displays the original data and the right plot displays one completed data set. The gray blobs represent records whose values are the same in both plots. The crosses represent the records subject to blanking. The dotted lines show the range restrictions and ratio edits. The distributions are similar.

variables is missing for 20.3% of cases—for D_{orig} and one of the completed data sets; results for other $D^{(l)}$ are similar. The overall bivariate patterns are very similar, suggesting once again the high quality of the imputations. We note that the shape of the joint density implies that a model based on a single multivariate normal distribution, even absent truncation, would not be appropriate for these data.

Figure 4 displays scatter plots of pairwise correlations across all variables in the original and three completed data sets. For the full distribution based on all 6,607 cases, the correlations are very similar. Correlations based on only the 3,000 cases with at least one missing item also are reasonably close to those based on the original data.

We next examine inferences for coefficients in the regression used by Petrin and White

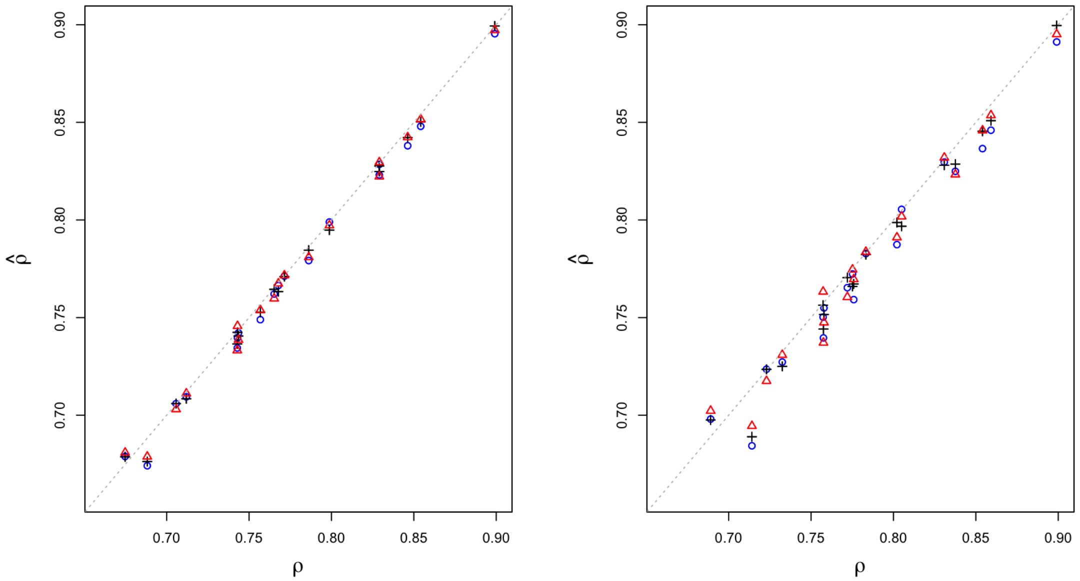

Figure 4: Plots of correlations among all seven variables for 1991 data. The horizontal coordinates are the correlations in the original data, and the vertical coordinates are the correlations in three completed data sets. Circles represent values for $D^{(1)}$, crosses represent values for $D^{(2)}$, and triangles represent values for $D^{(3)}$. The left plot uses all 6,607 cases and the right plot uses only the 3,000 cases with at least one missing item.

(2011) to analyze plant productivity, namely

$$\log(\mathrm{RVA}_i) = \beta_0 + \beta_C \log(\mathrm{CAP}_i) + \beta_L \log(\mathrm{LAB}_i) + \varepsilon_i, \quad \varepsilon_i \sim \mathrm{N}(0, \sigma^2), \tag{17}$$

where $\mathrm{LAB}_i = \mathrm{SL}_i + \mathrm{USL}_i$. We estimate regressions independently in each year. Figure 5 displays OLS 95% confidence intervals from $D_{orig,t}$ and from the CDPMMN multiply-imputed data sets in each year. For comparison, it also displays intervals based on only the complete cases. The intervals for β based on the CDPMMN completed data sets are similar to those based on D_{orig}, with somewhat wider lengths due to the missing values. For comparison, the complete-case results typically have even wider standard errors.

4.2 Repeated Simulation

The results in Figure 5 suggest that the CDPMMN multiple imputation offers more efficient inferences than the complete cases analysis. To verify this further, we perform a repeated sampling study. We assume that the 6,607 plants in the 1991 data comprise a population. We then randomly sample 500 independent realizations of D_{orig} from this population, each comprising 1,000 records. For each sampled D_{orig}, we introduce missing values by blanking one value for 200 randomly selected records, two values for 200 randomly sampled records, and three values for 100 randomly sampled records. We create $m = 10$ completed data sets using the CDPMMN imputation engine using the same approach as in Section 4.1, and estimate the regression coefficients in (17) using the multiple imputation point estimator. We also compute the point estimates based on D_{orig} and only the complete cases.

To evaluate the point estimators, for each method we compute three quantities for each coefficient in $\boldsymbol{\beta} = (\beta_0, \beta_C, \beta_L) = (\beta_0, \beta_1, \beta_2)$. The first quantity is the simulated bias, $Bias_j = \sum_{r=1}^{500} \hat{\beta}_{r,j}/500 - \beta_{pop,j}$, $j = 0, 1, 2$, where $\hat{\beta}_{r,j}$ is a method-specific point estimate of β_j in replication r and $\beta_{pop,j}$ is the value of β_j based on all 6,607 cases. The second quantity is the total squared error, $TSE_j = \sum_{r=1}^{500}(\hat{\beta}_{r,j} - \beta_{pop,j})^2$. The third quantity is the total squared distance between the point estimates based on the imputed (or only complete) data and the original data, $TSD_j = \sum_{r=1}^{500}(\hat{\beta}_{r,j} - \hat{\beta}_{r,j(orig)})^2$.

Table 4 displays the results of the simulation study. All methods are approximately unbiased, with the complete-cases analysis being far less efficient than the CDPMMN multiple imputation analysis. The CDPMMN multiple imputation closely follows the analysis based on the original data.

Since the repeated simulation results are based on a missing completely at random (MCAR) mechanism, we also perform a repeated simulation study with a missing at random (MAR) mechanism; see Appendix D in the Supplementary Materials for details. The

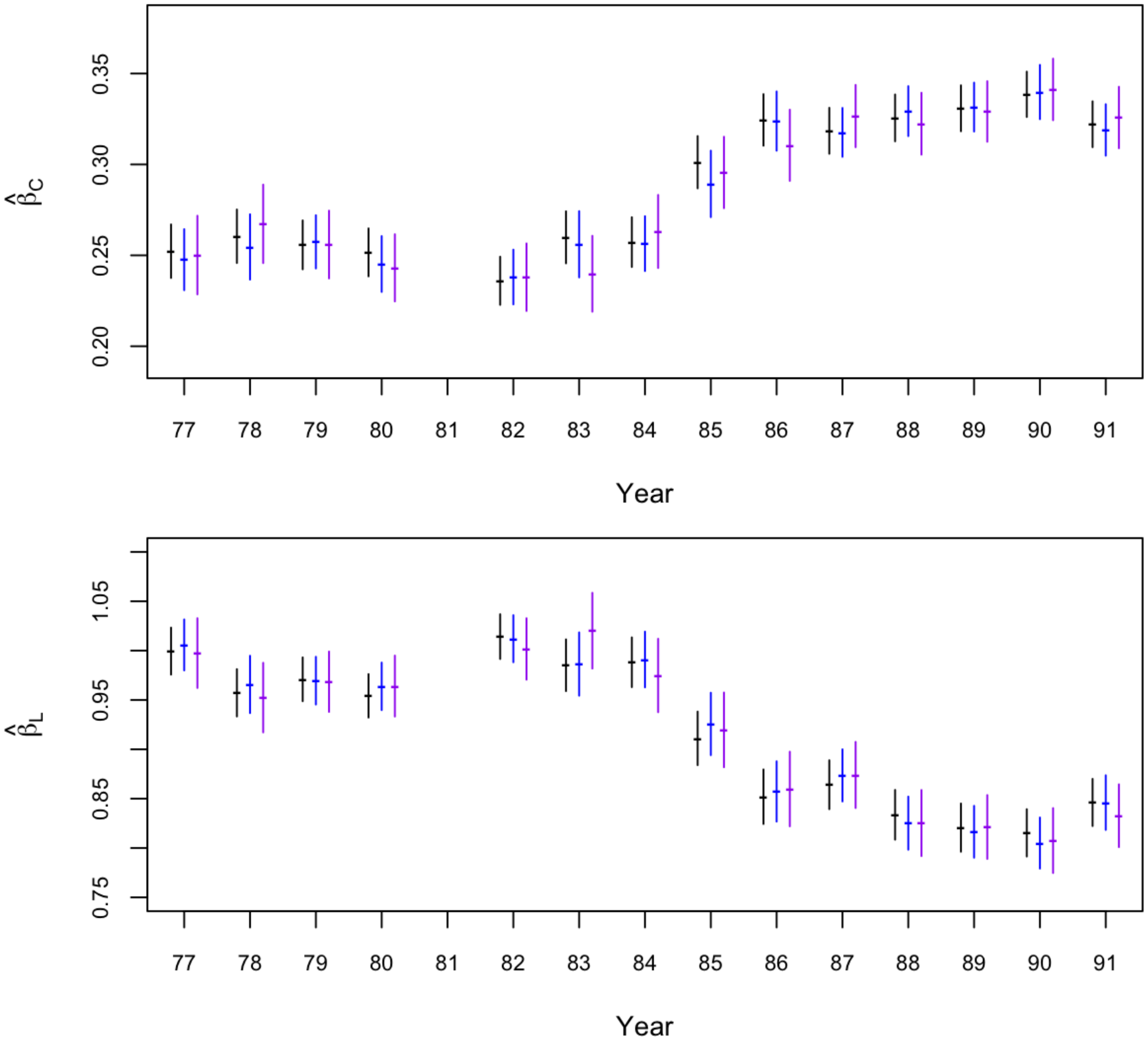

Figure 5: 95% confidence intervals for β_C and β_L from original data (first displayed interval), CDPMMN multiply-imputed data (second displayed interval), and the complete cases (third displayed interval). Using CDPMMN results in similar intervals as the original data, and shorter intervals than the complete cases.

Table 4: Properties of point estimators across the 500 simulations for the original data, the CDPMMN multiple imputation, and the complete cases analysis.

	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	-.002	5.19	1.31	-.001	0.27	0.07	.001	0.78	0.20
Complete-cases	-.006	8.55	4.71	.000	0.48	0.26	.002	1.38	0.75

performance of the CDPMMN imputations are akin to those in Table 4: point estimates are approximately unbiased with roughly a 30% increase in variance due to missing data. The complete-cases analysis results in biased estimates.

For both the MCAR and MAR scenarios, we also examine results for multiple imputation by chained equations using the MICE software (Van Buuren and Groothuis-Oudshoorn 2011) in R. We use the default implementation, which is based on predictive mean matching using linear models. In any completed dataset, there are usually a handful of imputed values that result in violations of the inequality constraints. Agencies seeking to release data clean of violations would have to change these values manually, whereas the CDPMMN imputation automatically respects such constraints. Nonetheless, we found that point estimators from the predictive mean matching MICE and the CDPMMN imputation performed quite similarly on our three evaluation measures, both in the MCAR and MAR scenarios. Of course, the performance of any method depends on the features of the data. Determining general conditions when the CDPMMN outperforms MICE, and vice versa (after manually fixing violations), is a topic worthy of further study.

5. CONCLUDING REMARKS

The empirical analyses of the Colombia manufacturing data suggest that the CDPMMN offers a flexible engine for generating coherent imputations guaranteed to respect linear inequalities. We expect the CDPMMN to be computationally feasible with efficient parallel computing when the number of variables is modest, say on the order of 40 to 50. For larger numbers of variables, analysts may need to use techniques other than the CDPMMN, for example models based on conditional independence assumptions such as Bayesian factor models (Aguilar and West 2000). We note that the general strategy of data augmentation combined with a Hit-and-Run sampler can be applied to any Bayesian multivariate model. In contrast, we do not view the number of records as a practically limiting factor, because the computations can be easily parallelized or implemented with GPU computing (Suchard et al. 2010). If computationally necessary, and potentially for improved accuracy, analysts can split the data into subsets of rows, e.g., by industry classifications, and estimate the model independently across subsets.

As with any imputation model specification, it is prudent to examine the fit of the CDPMMN model for the particular context. As described in Gelman et al. (2005), one can compare the marginal distributions of the observed and imputed values as a "sanity check." Highly dissimilar distributions can suggest the model does not describe the data well. One also can compute posterior predictive checks (e.g., He et al. 2010; Burgette and Reiter 2010; Si and Reiter 2013). As described by Si and Reiter (2013), the basic idea is to use the imputation model to generate not only Y_s but an entirely new full dataset, i.e., create a completed dataset $D^{(l)} = (Y_s^{(l)}, Y_{n-s})$ and a replicated dataset $R^{(l)}$ in which both Y_s and Y_{n-s} are simulated from the imputation model. After generating many pairs $(D^{(l)}, R^{(l)})$, one compares each $R^{(l)}$ with its corresponding $D^{(l)}$ on statistics of interest, such as regression coefficients and tail area probabilities. When the statistics are dissimilar, the diagnostic indicates that the imputation model does not generate replicated data that look like the completed data, so that it may not be generating plausible imputations for the missing data. When the statistics are not dissimilar, the diagnostic does not indicate evidence of imputation model inadequacy (with respect to that statistic).

Although the CDPMMN is intended primarily for continuous data, similar data augmentation strategies can be applied in other contexts of imputation under constraints. Manrique-Vallier and Reiter (forthcoming) use a truncated Dirichlet process mixture of multinomial distributions for imputation of missing unordered categorical data when the data include structural zeros, i.e., certain combinations are constrained to have probability zero (e.g., pregnant men). For mixed categorical and continuous data with no missing categorical variables, one can apply the CDPMMN imputation model separately within cells defined by the implied contingency table, provided sample sizes within each cell are adequate. We are unaware of methods for fitting joint mixture models for mixed data when both the categorical and continuous variables are subject to constraints.

In addition to extending these ideas to mixed data settings, there are several key areas in imputation under constraints that need future research. Some variables may need to satisfy linear equalities, for example logical sums. We did not account for these types of constraints here, although we anticipate that the hit-and-run sampler can be modified to do so. In edit-imputation settings, it is not clear that the Fellegi and Holt (1976) paradigm for error localization is optimal or advantageous in all settings. Error localization methods based on measurement error models derived from empirical evidence, combined with a fully coherent joint model for the imputation step, could result in higher quality edited/imputed data and subsequent analyses.

ACKNOWLEDGMENTS

This research was supported by the National Science Foundation (SES-11-31897). We thank the editors and two anonymous referees for useful and constructive comments.

SUPPLEMENTARY MATERIALS

Further explanations of results: File consisting of (i) results of simulations that illustrate the insensitivity of multiple imputation inferences to specifications of the prior distributions for the CDPMMN imputation engine, (ii) the expression for the limits of range restrictions and ratio edits when using standardized logarithms in the imputation models, (iii) an illustration of how to find boundaries of feasible region when constraints are represented by range restrictions and ratio edits, and (iv) the simulation study under the MAR assumption. (PDF file)

REFERENCES

Aguilar, O., and West, M. (2000), "Bayesian Dynamic Factor Models and Portfolio Allocation," *Journal of Business & Economic Statistics*, 18, 338–357.

Baccini, M., Cook, S., Frangakis, C. E., Li, F., Mealli, F., Rubin, D. B., and Zell, E. R. (2010), "Multiple Imputation in the Anthrax Vaccine Research Program," *Chance*, 23, 16–23.

Berger, J. O. (1993), "The Present and Future of Bayesian Multivariate Analysis," in *Multivariate Analysis: Future Directions*, ed. C. R. Rao, Amsterdam: North-Holland.

Boneh, A., and Golan, A. (1979), "Constraints' Redundancy and Feasible Region Boundedness by Random Feasible Point Generator," unpublished paper presented at Third European Congress on Operations Research, EURO III, Amsterdam, Netherlands.

Burgette, L. F., and Reiter, J. P. (2010), "Multiple Imputation for Missing Data via Sequential Regression Trees," *American Journal of Epidemiology*, 172, 1070–1076.

Chen, M.-H., and Schmeiser, B. (1993), "Performance of the Gibbs, Hit-and-Run, and Metropolis Samplers," *Journal of Computational and Graphical Statistics*, 2, 251–272.

De Waal, T. (2000), "A Brief Overview of Imputation Methods Applied at Statistics Netherlands," *Netherlands Official Statistics*, 15, 23–27.

De Waal, T., Pannekoek, J., and Scholtus, S. (2011), *Handbook of Statistical Data Editing and Imputation*, Hoboken, NJ: Wiley.

Dellaportas, P., and Papageorgiou, I. (2006), "Multivariate Mixtures of Normals With Unknown Number of Components," *Statistics and Computing*, 16, 57–68.

Draper, L. R., and Winkler, W. E. (1997), "Balancing and Ratio Editing With the New Speer System," Research Report RR97/05, Statistical Research Division, US Bureau of the Census, Washington, DC.

Dunson, D. B., and Xing, C. (2009), "Nonparametric Bayes Modeling of Multivariate Categorical Data," *Journal of the American Statistical Association*, 104, 1042–1051.

Escobar, M. D., and West, M. (1995), "Estimating Normal Means With a Dirichlet Process Prior," *Journal of the American Statistical Association*, 89, 268–277.

Fellegi, I. P., and Holt, D. (1976), "A Systematic Approach to Automatic Edit and Imputation," *Journal of the American Statistical Association*, 71, 17–35.

Fernandes, A. M. (2007), "Trade Policy, Trade Volumes and Plant-Level Productivity in Colombian Manufacturing Industries," *Journal of International Economics*, 71, 52–71.

Garcia-Rubio, E., and Villan, I. (1990), "DIA system: Software for the Automatic Editing of Qualitative Data," in *Proceedings of 1990 Annual Research Conference, US Bureau of the Census*, Arlington, VA, pp. 525–537.

Gelman, A., Van Mechelen, I., Verbeke, G., Heitjan, D. F., and Meulders, M. (2005), "Multiple Imputation for Model Checking: Completed-Data Plots with Missing and Latent Data," *Biometrics*, 61, 74–85.

Gilbride, T. J., and Lenk, P. J. (2010), "Posterior Predictive Model Checking: An Application to Multivariate Normal Heterogeneity," *Journal of Marketing Research*, 47, 896–909.

Granquist, L., and Kovar, J. G. (1997), "Editing of Survey Data: How Much Is Enough?" in *Survey Measurement and Process Quality*, eds. L. Lyberg, P. Biemer, M. Collins, E. De Leeuw, C. Dipp, N. Schwarz, and D. Trewin, New York: Wiley, pp. 415–435.

He, Y., Zaslavsky, A., Landrum, M., Harrington, D., and Catalano, P. (2010), "Multiple Imputation in a Large-Scale Complex Survey: A Practical Guide," *Statistical Methods in Medical Research*, 19, 653–670.

Hedlin, D. (2003), "Score Functions to Reduce Business Survey Editing at the U.K. Office for National Statistics," *Journal of Official Statistics*, 19, 177–199.

Hirano, K. (2002), "Semiparametric Bayesian Inference in Autoregressive Panel Data Models," *Econometrica*, 70, 781–799.

Ishwaran, H., and James, L. F. (2001), "Gibbs Sampling Methods for Stick-Breaking Priors," *Journal of the American Statistical Association*, 96, 161–173.

Lavine, M., and West, M. (1992), "A Bayesian Method for Classification and Discrimination," *Canadian Journal of Statistics*, 20, 451–461.

Lawrence, D., and McDavitt, C. (1994), "Significance Editing in the Australian Survey of Average Weekly Earnings," *Journal of Official Statistics*, 10, 437–447.

Li, K. H., Raghunathan, T. E., and Rubin, D. B. (1991), "Large Sample Significance Levels From Multiply Imputed Data Using Moment-Based Statistics and an F Reference Distribution," *Journal of the American Statistical Association*, 86, 1065–1073.

Little, R. J. A., and Rubin, D. B. (2002), *Statistical Analysis With Missing Data* (2nd ed.), Hoboken, NJ: Wiley.

Lovász, L., and Vempala, S. (2006), "Hit-And-Run from a Corner," *SIAM Journal on Computing*, 35, 985–1005.

MacEachern, S. N., and Müller, P. (1998), "Estimating Mixture of Dirichlet Process Models," *Journal of Computational and Graphical Statistics*, 7, 223–238.

Manrique-Vallier, D., and Reiter, J. P. (forthcoming), "Bayesian Estimation of Discrete Multivariate Latent Structure Models with Structural Zeros," *Journal of Computational and Graphical Statistics*.

Manzari, A. (2004), "Combining Editing and Imputation Methods: An Experimental Application on Population Census Data," *Journal of the Royal Statistical Society, Sereis A*, 167, 295–307.

Meng, X.-L., and Rubin, D. B. (1992), "Performing Likelihood Ratio Tests With Multiply-Imputed Data Sets," *Biometrika*, 79, 103–111.

Meng, X.-L., and Zaslavsky, A. M. (2002), "Single Observation Unbiased Priors," *Annals of Statistics*, 30, 1345–1375.

Norberg, A. (2009), "Editing at Statistics Sweden – Yesterday, Today and Tomorrow," in *Modernisation of Statistics Production 2009*, Sockholm, Sweden.

O'Malley, A. J., and Zaslavsky, A. M. (2008), "Domain-Level Covariance Analysis for Multilevel Survey Data With Structured Nonresponse," *Journal of the American Statistical Association*, 103, 1405–1418.

Petrin, A., and White, T. K. (2011), "The Impact of Plant-Level Resource Reallocations and Technical Progress on U.S. Macroeconomic Growth," *Review of Economic Dynamics*, 14, 3–26.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., and Solenberger, P. (2001), "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models," *Survey Methodology*, 27, 85–95.

Reiter, J. P. (2007), "Small-Sample Degrees of Freedom for Multi-Component Significance Tests With Multiple Imputation for Missing Data," *Biometrika*, 94, 502–508.

Reiter, J. P., and Raghunathan, T. E. (2007), "The Multiple Adaptations of Multiple Imputation," *Journal of the American Statistical Association*, 102, 1462–1471.

Rubin, D. B. (1976), "Inference and Missing Data," *Biometrika*, 63, 581–592.

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

Scholtus, S., and Goksen, S. (2012), "Automatic Editing With Hard and Soft Edits," Discussion Paper 201225, Statistics Netherlands.

Sethuraman, J. (1994), "A Constructive Definition of Dirichlet Priors," *Statistica Sinica*, 4, 639–650.

Si, Y., and Reiter, J. P. (2013), "Nonparametric Bayesian Multiple Imputation for Incomplete Categorical Variables in Large-Scale Assessment Surveys," *Journal of Educational and Behavioral Statistics*, 38, 499–521.

Smith, R. (1980), “A Monte Carlo Procedure for the Random Generation of Feasible Solutions to Mathematical Programming Problems,” in *Bulletin of the TIMS/ORSA Joint National Meeting*, Washington, DC.

Suchard, M. A., Wang, Q., Chan, C., Frelinger, J., Cron, A., and West, M. (2010), “Understanding GPU Programming for Statistical Computation: Studies in Massively Parallel Massive Mixtures,” *Journal of Computational and Graphical Statistics*, 19, 419–438.

Tempelman, C. (2007), “Imputation of Restricted Data,” Ph. D. dissertation, University of Groningen.

Tervonen, T., van Valkenhoef, G., Baştürk, N., and Postmus, D. (2013), “Hit-And-Run Enables Efficient Weight Generation for Simulation-Based Multiple Criteria Decision Analysis,” *European Journal of Operational Research*, 224, 552–559.

Thompson, K. J., Sausman, K., Walkup, M., Dahl, S., King, C., and Adeshiyan, S. A. (2001), “Developing Ratio Edits and Imputation Parameters for the Services Sector Censuses Plain Vanilla Ratio Edit Module Test,” Economic Statistical Methods Report ESM-0101, US Bureau of the Census, Wahsington, DC.

Van Buuren, S., and Groothuis-Oudshoorn, K. (2011), “mice: Multivariate Imputation by Chained Equations in R,” *Journal of Statistical Software*, 45.

Van Buuren, S., and Oudshoorn, K. (1999), “Flexible Multivariate Imputation by MICE,” Technical Report PG/VGZ/99.054, TNO Prevention and Health, Leiden, Netherlands.

Whitridge, P., and Kovar, J. G. (1990), “Applications of the Generalized Edit and Imputation System at Statistics Canada,” in *American Statistical Association Proceedings of the Survey Research Method Section*, pp. 105–110.

Winkler, W. E., and Draper, L. R. (1996), "Application of the SPEER Edit System," Research Report RR96/02, Statistical Research Division, US Bureau of the Census, Washington, DC.

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						

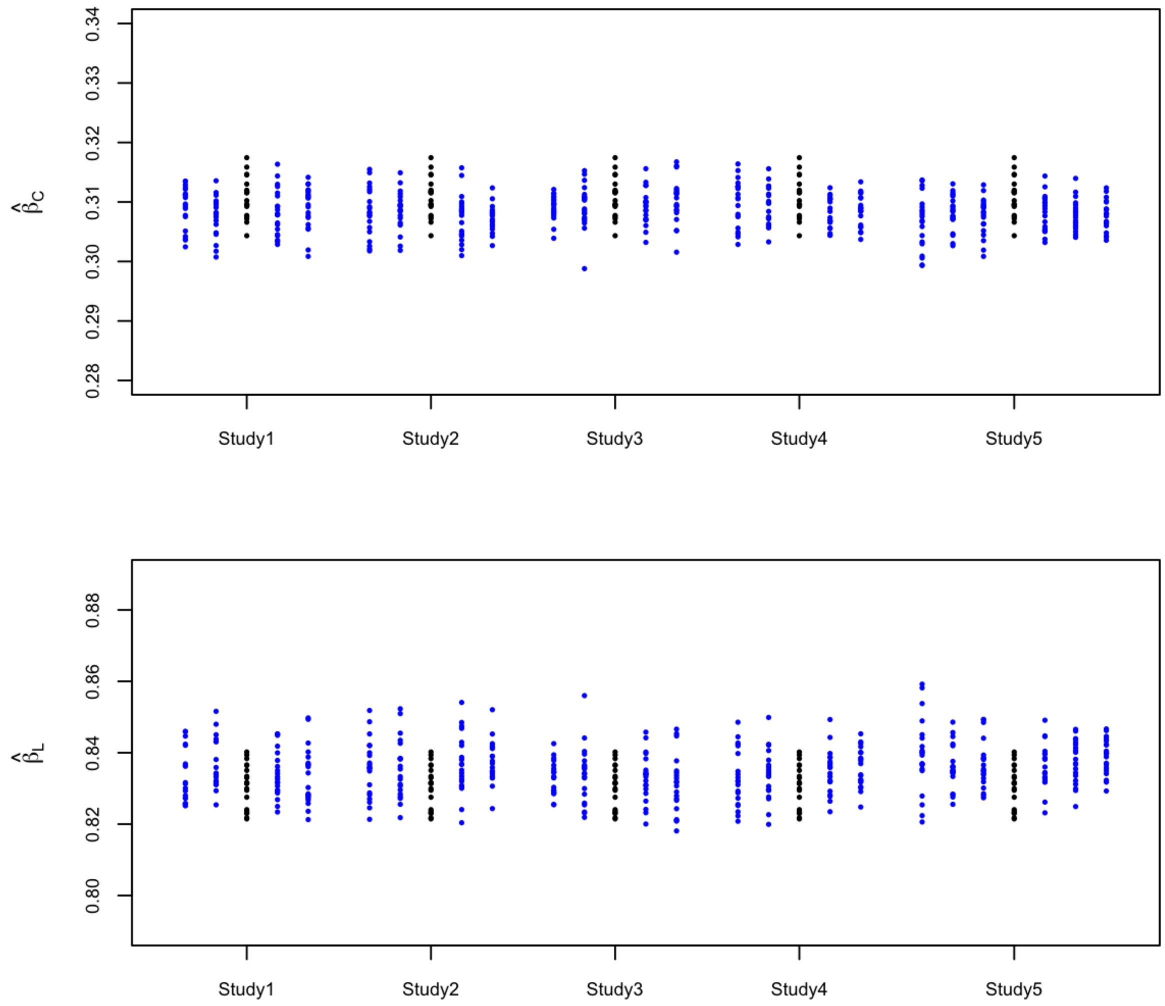

Figure A.1: Estimated regression coefficients of labor in the regression in Section 4 based on the CDPMMN multiple imputation under the prior specifications in Table A.1. In each setting, results are based on 20 independent replications, and the middle column includes the estimates under the prior specification 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 $A\boldsymbol{y}_i \leq \boldsymbol{b}$, 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 $\mathcal{A}_i = \{(y_{i1}, y_{i3})' : A_{13}(y_{i1}, y_{i3})' \leq \boldsymbol{b}_{13} - \boldsymbol{a}_2 y_{i2}\}$. 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 $j \in \{\text{RVA,CAP,SL, USL,RMU}\}$.

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.