# Bayesian Simultaneous Edit and Imputation for Multivariate Categorical Data

Daniel Manrique-Vallier and Jerome P. Reiter*

June 24, 2015

## Abstract

In categorical data, it is typically the case that some combinations of variables are theoretically impossible, such as a three year old child who is married or a man who is pregnant. In practice, however, reported values often include such structural zeros due to, for example, respondent mistakes or data processing errors. To purge data of such errors, many statistical organizations use a process known as edit-imputation. The basic idea is first to select reported values to change according to some heuristic or loss function, and second to replace those values with plausible imputations. This two-stage process typically does not fully utilize information in the data when determining locations of errors, nor does it appropriately reflect uncertainty resulting from the edits and imputations. We present an approach that integrates editing and imputation for categorical microdata with structural zeros. We rely on a Bayesian hierarchical model that includes (i) a Dirichlet process mixture of multinomial distributions as the model for the underlying true values of the data, with support restricted to the set of theoretically possible combinations, (ii) a model for latent indicators of the values that are in error, and (iii) a model for the reported responses for values in error. We illustrate this integrated approach using simulation studies with data from the 2000 U. S. census, and compare it to a two-stage edit-imputation routine. Supplementary material is available online.

Key words: Dirichlet process, error, Fellegi-Holt, latent, structural zero.

*Daniel Manrique-Vallier is Assistant Professor of Statistics, Indiana University, Bloomington, IN 47408 (e-mail dmanriqu@indiana.edu); and Jerome P. Reiter is Mrs. Alexander Hehmeyer Professor of Statistical Science, Duke University, Durham, NC 27708-0251, (e-mail: jerry@stat.duke.edu). This research was supported by a grant from the National Science Foundation (SES-11-31897).

# 1 Introduction

In surveys with multiple categorical data items, the reported data frequently include erroneous values, i.e., combinations of answers that are theoretically impossible or inconsistent across items. These could result from respondent error; for example, a parent accidentally checks boxes indicating that his child is five years old and has attained a university degree, or a person selects "male" for sex and "yes" to answer whether or not the person ever had a hysterectomy. They also could result from data processing errors (Groves and Lyberg, 2010; Biemer, 2010). Regardless of the source, left uncorrected, erroneous values can diminish the quality of analysis and interpretation of the data (Fuller, 1987; Groves, 2004; Durrant and Skinner, 2006).

Recognizing this, many data stewards such as national statistics institutes (NSIs) purge data of overtly erroneous values in a process known as edit-imputation. Edit-imputation is particularly salient for organizations sharing data with the public: releasing files with erroneous values could cause the public to lose confidence in the quality of the data and in the organization more broadly. In fact, many data stewards invest quite substantial resources in edit-imputation processes, exceeding 30% of survey budgets in some NSIs (Granquist and Kovar, 1997; Norberg, 2009).

It is generally impractical and cost-prohibitive to re-contact all respondents with potentially erroneous values. Thus, many data stewards rely in part on automated methods of edit-imputation, in which algorithms select and "correct" erroneous values with minimal human intervention. These automatic editing systems typically run in two steps, an error localization step in which some set of each record's values is determined to be in error, and an imputation step in which those values are replaced with plausibly correct values (De Waal et al., 2011). Most editing systems use variants of the error localization suggested by Fellegi and Holt (1976): for any record with impossible reported values, change the minimum number of fields (variables) needed to make that record a theoretically possible observation (e.g., see Winkler, 1995; Winkler and Petkunas, 1997). The subsequent imputation step usually is a variant of single imputation generated from a hot deck or parametric model (e.g., Winkler, 2003, 2008).

Kim et al. (forthcoming) point out that Fellegi and Holt approaches to edit-imputation—henceforth abbreviated as F-H approaches—have two key drawbacks. First, they do not fully utilize the information in the data in selecting the fields to impute. To illustrate, consider an example where the variables include sex, hysterectomy status, and age in years. When a record is reported as a male with a hysterectomy who is 20 years old, it seems more plausible to change status to no hysterectomy than to change sex to female, because hysterectomies are relatively uncommon among 20 year old women (Merrill, 2008). The minimum number of fields criterion results in changing one of sex or hysterectomy status. The data steward might select among these two solutions based on some heuristic, e.g., change the variable that is more likely to have errors according to experience in other contexts. Second, the organization generally cannot be certain that a F-H (or any) error localization has identified the exact locations of errors. This uncertainty is ignored by specifying a single error localization; hence, analyses of data corrected by F-H approaches underestimate uncertainty (Kim et al., forthcoming). For example, there are 20 year old women with hysterectomies, and inferences should reflect that possibility (as well as other feasible variations, including changing multiple variables) as increased uncertainty.

In this article, we propose an integrated approach to edit-imputation for categorical data that addresses these two shortcomings of the F-H paradigm. Our approach relies on a Bayesian hierarchical model that includes (i) a Dirichlet process mixture of multinomial distributions, also known as a latent class model, for the underlying true values of the data with support that excludes the set of structural zero cells, (ii) a model for latent indicators of the values that are in error, and (iii) a model for the reported responses for values in error. A similar strategy was used by Kim et al. (forthcoming) for data with only continuous variables, in which true values are required to satisfy pre-specified linear inequalities and equalities on the relationships among the variables. By fully integrating the editing and imputation steps, we encode a probability distribution on the error locations that is informed by the observed relationships among the variables in the data, and we fully incorporate uncertainty in the process of selecting and replacing erroneous values. The MCMC algorithm for estimating the model generates datasets without structural zeros as by-products. These can be disseminated as public use files and analyzed using multiple imputation techniques (Rubin, 1987).

The remainder of the article is organized as follows. In Section 2, we present the Bayesian hierarchical model for edit-imputation, which we call the EI-DPM model. We first describe the model for measurement error, which comprises sub-models for the error locations and the reported values. We then describe the Dirichlet process mixture model for the underlying true data. In Section 3, we present a MCMC algorithm for sampling from the EI-DPM model that guarantees all edit constraints are satisfied. In Section 4, we report results of simulation studies based on a subset of data from the 2000 U.S. census. Here, we compare the accuracy of inferences from the EI-DPM to those from a F-H edit-imputation approach. In Section 5, we conclude with a discussion of future research directions.

## 2 The EI-DPM Model

Suppose we have a sample of $n$ subjects measured on $J$ categorical variables. We associate each subject $i \in \{1, \ldots, n\}$ with a true response vector, $\mathbf{x}_i = (x_{i1}, \ldots, x_{iJ})$, with categorical entries, $x_{ij} \in \mathcal{C}_j = \{1, \ldots, L_j\}$. True responses, $\mathbf{x}_i$, correspond to the data we would observe if they were perfectly recorded. Let $\mathcal{X} = (\mathbf{x}_i, \ldots, \mathbf{x}_n)$ be the vector of all true responses. We do not observe $\mathcal{X}$; rather, we observe the reported data $\mathcal{Y} = (\mathbf{y}_1, ..., \mathbf{y}_n)$. Here, each $\mathbf{y}_i \in \mathcal{C} = \mathcal{C}_1 \times ... \times \mathcal{C}_J$ is a potentially contaminated version of its corresponding $\mathbf{x}_i$. We assume that observable data are stochastically generated conditional on the true responses through a measurement process, with density $\mathcal{M}(\mathbf{y}|\mathbf{x}_i, \theta_y)$ for $\mathbf{y} \in \mathcal{C}$. If this process is such that $\mathcal{M}(\mathbf{y}|\mathbf{x}_i, \theta_y) = \delta_{\mathbf{x}_i}(\mathbf{y})$, we call it a *perfect* measurement process. Whenever it is the case that $\mathbf{y}_i \neq \mathbf{x}_i$ we say that the $i$-th individual's observed data *contains errors.*

True responses are subject to a set of edit rules that enumerate impossible responses. Formally, edit rules are a subset $S \subsetneq \mathcal{C}$ for which we know *a priori* that $\Pr(\mathbf{x}_i \in S) = 0$; that is, they are a set of structural zeros as defined by Bishop et al. (1975). Unlike true responses, reported data potentially can take any possible value in $\mathcal{C}$, i.e., $\Pr(\mathbf{y}_i \in S) > 0$. Therefore, whenever $\mathbf{y}_i \in S$ we know for sure that $\mathbf{y}_i$ contains errors. We call errors that result in a direct violation of edit rules *detectable.* Conversely, whenever $\mathbf{x}_i \neq \mathbf{y}_i$ but $\mathbf{y}_i \notin S$, we call the errors *undetectable.* We note that the existence of errors within a record is detectable, not necessarily the location of the errors. For example, when a reported record is a male with a hysterectomy we can be sure that there is an error, but we cannot know—at least not from the record only—whether the error is the reported sex, the reported hysterectomy, or both.

We assume that the true response vectors are iid samples from a common data generating process with density $f_x(\mathbf{x}|\theta_x, S)$ for $\mathbf{x} \in \mathcal{C}$. Here the dependence on $S$ stresses the fact that the support of the distribution must be truncated to the set $\mathcal{C} \setminus S$ to avoid inconsistent values. Under this setup, the objective is to use the contaminated data, $\mathcal{Y}$, to estimate the joint distribution of the true responses.

## 2.1 Measurement model

When we consider $\mathcal{X}$ as given, the measurement model $\mathcal{M}(\mathbf{y}|\mathbf{x}_i, \theta_y)$ encodes assumptions about how the observable responses are generated from the true responses. It is useful to specify the measurement model in two parts. The first part, the error location model, specifies which among the $n \times J$ entries in $\mathcal{X}$ are reported incorrectly. For $i = 1, \dots, n$, let $\mathbf{E}_i = (E_{i1}, ..., E_{iJ})$, where $E_{ij} = 1$ when there is an error in the $(i, j)$-th location, and $E_{ij} = 0$ otherwise. By definition, $E_{ij} = 1$ if and only if $x_{ij} \neq y_{ij}$. Generically, we write the distribution of the error location model for all $n$ records as $p(\mathcal{E} \mid \mathcal{X}, \theta_y)$. The second part, the *reporting model*, specifies which values are observed for fields in error. Generically, we write the distribution of this model as $p(\mathcal{Y} \mid \mathcal{X}, \mathcal{E}, \theta_y)$. Thus, we can write $\mathcal{M}(\mathbf{y}|\mathbf{x}_i, \theta_y)$ as $p(\mathcal{Y} \mid \mathcal{X}, \mathcal{E}, \theta_y)p(\mathcal{E} \mid \mathcal{X}, \theta_y)$.

The error localization and reporting models can be adapted to reflect *a priori* assumptions about the measurement error process. For example, one can allow the likelihood of error for variable $j$ to depend on the level of the true response by using a regression of $E_{ij}$ on $x_{ij}$. More simply, one could favor certain combinations of $\mathbf{E}_i$ over others. For example, one could assume $Pr(E_{ij} = 1 \mid \mathbf{x}_i, \theta_y) = \epsilon_j$, allowing differential probability for changing some variables over others. This is sensible when some variables are more reliably reported than others. Alternatively, one might assume that errors are generated completely at random, so that all $E_{ij}$ have independent Bernoulli distributions with a common error rate $\epsilon$. This location model defines a perfect measurement model when $\epsilon = 0$.

For the reporting model, it is computationally convenient to assume that, conditional on the existence of errors, substitutions are independent. Similar independence assumptions are made, for example, in record linkage contexts (e.g. Fellegi and Sunter, 1969; Steorts et al., 2014). Generally, we write the model as

$$y_{ij}|x_{ij} = l, E_{ij} = e \sim \begin{cases} \delta_l & \text{if } e = 0 \\ \text{Discrete}(\{1, ..., L_j\}, \{q_{jl}(1), q_{jl}(2), ..., q_{jl}(L_j)\}) & \text{if } e = 1. \end{cases} \tag{1}$$

Here $q_{jl}(l^*)$ is the probability of reporting level $l^* \in \{1, ..., L_j\}$ for variable $j$ when the actual value is $l \in \{1, ..., L_j\}$. We consider $\sum_{l^*=1}^{L_j} q_{jl}(l^*) = 1$ and $q_{jl}(l) = 0$. Interpreted as a generative process, this model simply states that whenever the $j$-th variable is erroneous, the reported data will be sampled from all possible values except for the correct one. Absent any information about the propensities of reporting errors, one can assume a uniform substitution process, making $q_{jl}(l^*) \propto 1$ for all $l^* \neq x$ and $q_{jl}(l) = 0$. This leads to the probabilities

$$q_{jl}(l^*) = \begin{cases} q_j & \text{if } l^* \neq l \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where $q_j = 1/(L_j - 1)$.

We complete the specification of the measurement error model with prior distributions for $\theta_y$. For example, for a Bernoulli error location model, one can use

$$\epsilon \sim Beta(a_\epsilon, b_\epsilon). \tag{3}$$

This prior specification has the advantage of being conjugate to the error location model and of having a flexible and natural interpretation. Interpreting the prior expected value $a_\epsilon/(a_\epsilon + b_\epsilon)$ as the "prior error rate", and $a_\epsilon + b_\epsilon$ as the "prior sample number of responses" (see e.g. Gelman et al., 2013, p35), we can encode beliefs about the quality of the observed data. With $\epsilon_j \neq \epsilon$ for all $j$, one instead could use $(a_{\epsilon j}, b_{\epsilon j})$ in (3), selecting values that reflect *a priori* beliefs about the error rate of each variable.

A key feature of this formulation is that we do not assume that the error generation depends in any way on the edit rules $S$. In particular, we do not assume $\mathbf{y}_i \notin S$ implies that $\mathbf{y}_i = \mathbf{x}_i$. From a generative perspective this feature seems a sensible modeling choice—the generation of errors need not be contingent on our ability to detect them. However this is a departure from the F-H approaches applied in most NSIs. F-H approaches are based on the principle of minimizing the number of changes required for the observed data to satisfy all the edits. This implies that whenever $\mathbf{y}_i \neq \mathbf{x}_i$ but $\mathbf{y}_i \notin S$, i.e., the record has errors but satisfies all edits, a F-H approach prescribes no changes to that record. Thus, F-H essentially treats undetectable errors as nonexistent; in other words, it assumes a measurement model that only can generate detectable errors.

Although perhaps unrealistic, measurement models that generate only detectable errors are implicitly used by many NSIs. Typically these organizations are reluctant to change reported data values as a matter of principle. To adhere to this logic, that is, fixing only records with detectable errors, we can adjust the measurement model in the EI-DPM by setting $\mathbf{E}_i = (0, ..., 0)$ for all records with $\mathbf{y}_i \notin S$. This effectively forces the model to assume $\mathbf{x}_i = \mathbf{y}_i$ for all records without detectable errors.

Alternatively, we can enforce changing a small number of fields through the prior distribution on the error rates. Specifying small probabilities of errors with high certainty—for example, using small $a_\epsilon/(a_\epsilon + b_\epsilon)$ with large $a_\epsilon + b_\epsilon$ in (3)—implies that errors are rare by default. This results in posterior distributions where individual records are considered erroneous only when strongly suggested by the data, which is most likely because of a direct violation of the edit rules. This specification can be considered as a model-based analogue of the F-H principle.

## 2.2 True response model

The true responses generation model, $f_x(\mathbf{x}|\theta_x, S)$, in principle can be any multivariate discrete distribution, as long as it is supported only in the set $\mathcal{C} \setminus S$. In practice we desire $f_x(\mathbf{x}|\theta_x, S)$ to be rich enough to capture the relevant multivariate structure in $\mathcal{X}$. One such model is the truncated Bayesian non-parametric latent class model (TNPLCM), introduced by Manrique-Vallier and Reiter (2014a), which we now review.

Latent class models (LCM. Goodman, 1974; Lazarsfeld and Henry, 1968) are widely used for representing discrete multivariate distributions. Putting aside for the moment the complications posed by the structural zeros, the LCM is the finite mixture of $K$ products of multinomial distributions,

$$p(\mathbf{x}|\boldsymbol{\lambda},\boldsymbol{\pi}) = f^{LCM}(\mathbf{x}|\boldsymbol{\lambda},\boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} \lambda_{jk}[x_j], \tag{4}$$

where $\boldsymbol{\pi} = (\pi_1, ..., \pi_K)$ are mixture component probabilities such that $\sum_{k=1}^{K} \pi_k = 1$, and $\boldsymbol{\lambda} = (\lambda_{jk}[l])$ are $J \times K$ sets of multinomial probabilities with $\sum_{l=1}^{L_j} \lambda_{jk}[l] = 1$ for $j = 1, ..., J$ and $k = 1, ..., K$. For large enough $K$, this model can represent any probability distribution over $\mathcal{C}$ arbitrarily well (Dunson and Xing, 2009).

The mixture in (4) can be represented as a two-step generative process for $\mathbf{x}_i$. Let each individual belong to one latent class, $z_i \in (1, \ldots, K)$. For any $i = 1, \ldots, n$, we can write (4) as

$$x_{ij}|z_i \overset{indep}{\sim} \text{Discrete}\left(\{1, ..., L_j\}, (\lambda_{1z_i}[1], ..., \lambda_{Jz_i}[L_j])\right) \quad \text{for } j = 1, ..., J \tag{5}$$

$$z_i \overset{iid}{\sim} \text{Discrete}\left(\{1, ..., K\}, (\pi_1, ..., \pi_K)\right). \tag{6}$$

This representation facilitates estimation via Gibbs sampling (Ishwaran and James, 2001; Si and Reiter, 2013; White and Murphy, 2014).

Dunson and Xing (2009) suggest letting $K = \infty$ and using an infinite stick-breaking process (Sethuraman, 1994) for the prior distribution on $\boldsymbol{\pi}$. This prior distribution has full support on the distribution of probabilities of $\mathbf{x}$, and it does not restrict dependence structures *a priori* (Dunson and Xing, 2009). For an almost-sure approximation (Ishwaran and Zarepour, 2002) that is computationally more convenient, one can set $K$ to a large but finite integer (Ishwaran and James, 2001; Si and Reiter, 2013). In this representation, we let $\pi_k = V_k \prod_{h=1}^{k-1}(1 - V_h)$, where $V_K = 1$ and $V_1, \ldots, V_{K-1} \overset{iid}{\sim} \text{Beta}(1, \alpha)$. The prior specification can be completed with diffuse priors $\alpha \sim \text{Gamma}(0.25, 0.25)$ and $\lambda_{jk}[\cdot] \overset{iid}{\sim} \text{Dirichlet}_{L_j}(1, ..., 1)$ for $j = 1, ..., J$ and $k = 1, ..., K$. See Dunson and Xing (2009) and Si and Reiter (2013) for additional discussion of the prior distributions.

The LCM does not automatically assign zero probability to combinations $\mathbf{x} \in S$, that is, potential outcomes known to have probability zero. To enforce $Pr(\mathbf{x} \in \mathcal{S}) = 0$ when estimating cell probabilities, Manrique-Vallier and Reiter (2014a) introduced the truncated LCM,

$$p(\mathbf{x}|\boldsymbol{\lambda}, \boldsymbol{\pi}, S) = f^{TLCM}(\mathbf{x}|\boldsymbol{\lambda}, \boldsymbol{\pi}) \propto I\{\mathbf{x} \notin S\} \sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} \lambda_{jk}[x_j]. \tag{7}$$

Manrique-Vallier and Reiter (2014a) use the $K$-dimensional stick-breaking process as the prior distribution of $\boldsymbol{\pi}$. To estimate model parameters, Manrique-Vallier and Reiter (2014a) rely on a sample augmentation strategy. They consider $\mathcal{X} = \{\mathbf{x}_i \notin S\}$ to be a subset of a hypothetical sample $\mathcal{X}^*$ of $N \geq n$ records, directly generated from (5)-(6) without caring about the truncation of the support. Let $\mathcal{X}^0$ and $\mathcal{Z}^0$ be the responses and latent class labels corresponding to those samples in $\mathcal{X}^*$ that did fall into the set $S$. Manrique-Vallier and Reiter (2014a) show that, by using the improper prior distribution $p(N) \propto 1/N$ (Meng and Zaslavsky, 2002), the marginal posterior distribution of parameters $(\boldsymbol{\pi}, \boldsymbol{\lambda}, \alpha)$ after integrating out $(N, \mathcal{X}^0, \mathcal{Z}, \mathcal{Z}^0)$ matches that based on the truncated representation in (7).

### 2.3 Edit-imputation model used in illustrations

In the empirical illustration, we use a measurement model with common $\epsilon$ and a uniform substitution process. Putting this together with a TNPLCM as the true response model, we have

$$y_{ij}|x_{ij}, E_{ij} \sim \begin{cases} \delta_{x_{ij}} & \text{if } E_{ij} = 0 \\ \text{Uniform}(\{1, ..., L_j\} \setminus \{x_{ij}\}) & \text{if } E_{ij} = 1 \end{cases} \tag{8}$$

$$E_{ij}|\epsilon \overset{iid}{\sim} \text{Bernoulli}(\epsilon) \tag{9}$$

$$p(\mathbf{x}_i|\boldsymbol{\lambda}, \boldsymbol{\pi}, S) \propto I\{\mathbf{x}_i \notin S\} \sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} \lambda_{jk}[x_{ij}] \tag{10}$$

$$\pi_k = V_k \prod_{h<k} (1 - V_h) \text{ for } k = 1, ..., K \tag{11}$$

$$V_k|\alpha \overset{iid}{\sim} \text{Beta}(1, \alpha) \text{ for } k < K, \text{ and } V_K = 1 \tag{12}$$

$$\lambda_{jk}[\cdot] \overset{iid}{\sim} \text{Dirichlet}_{L_j}(1, ..., 1) \tag{13}$$

$$\alpha \sim \text{Gamma}(0.25, 0.25) \tag{14}$$

$$\epsilon \sim \text{Beta}(a_\epsilon, b_\epsilon). \tag{15}$$

Here $i$ ranges from 1 to $n$, and $j$ from 1 to $J$.

## 3 MCMC Estimation

To estimate the model in (8)–(15), we use the Gibbs sampler outlined in Section 3.2. This sampler utilizes the fact that editing rules, i.e., representations of structural zeros, often can be expressed as the union of non-overlapping *table slices.* A table slice is a subset of $\mathcal{C}$ defined by fixing a subset of the coordinates of the potential responses—for example, the set $\{\mathbf{x} \in \mathcal{C} : x_1 = 1, x_3 = 2\}$. Edit rules frequently are formulated as a collection of combinations of levels of a few variables (often just two at a time) that are deemed impossible. For example, a rule forbidding records where "sex=male" and "hysterectomy=yes" defines a table slice. More complex rules can be decomposed into simple slice definitions. For example, a rule specifying that people younger than 14 years old cannot be married can be translated as the union of all the table slices formed by keeping the variable for marital status fixed at "married," and making the age variable take all the discrete levels corresponding to ages less than 14. Expressing $S$ as table slices can facilitate significant computational gains in the Gibbs sampler for the truncated LCM (Manrique-Vallier and Reiter, 2014b,a). Hence, we begin by defining a notation for table slices corresponding to edit rules.

### 3.1 Representing Edit Rules as Table Slices

Following the notation in Manrique-Vallier and Reiter (2014a), we define *slice definitions* as vectors $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_j)$ where, for $j = 1, \ldots, J$, $\mu_j = x_j$ whenever $x_j$ is fixed at some level and $\mu_j = *$ otherwise, where $*$ is special notation for a placeholder. For example, assuming $J = 5$, the slice definition for the example set $\{\mathbf{x} \in \mathcal{C} : x_1 = 1, x_3 = 2\}$ corresponds to $\boldsymbol{\mu} = (1, *, 2, *, *)$. Let $\mathcal{C}^* = \prod_{j=1}^{J} \{1, 2, ..., L_j, *\}$ be the space of all possible slice definitions $\boldsymbol{\mu}$.

We use slice definitions to denote table slices through the mapping,

$$\overline{\boldsymbol{\mu}} = \{(x_1, ..., x_J) \in \mathcal{C} : x_j = \mu_j \text{ for all } \mu_j \neq *\}, \tag{16}$$

for any $\boldsymbol{\mu} = (\mu_1, ..., \mu_j) \in \mathcal{C}^*$. We call $\overline{\boldsymbol{\mu}}$ the table slice defined by $\boldsymbol{\mu}$. Note that while the slice definition $\boldsymbol{\mu}$ is an element of the set $\mathcal{C}^*$, $\overline{\boldsymbol{\mu}}$ is a subset of $\mathcal{C}$.

The slice definition notation is useful because it allows us to define formal operations in the space $\mathcal{C}^*$ that map directly into set operations in $\mathcal{C}$. In particular, we define the *intersection of slice definitions* $\boldsymbol{\mu}^A$ *and* $\boldsymbol{\mu}^B$ as the slice definition vector $\text{int}(\boldsymbol{\mu}^A, \boldsymbol{\mu}^B) = (\gamma_1, ..., \gamma_J) \in \mathcal{C}^*$ such that, for any $j = 1, ..., J$,

$$\gamma_j = \begin{cases} \mu_j^A & \text{if } (\mu_j^A = \mu_j^B) \text{ or } (\mu_j^A \neq * \text{ and } \mu_j^B = *) \\ \mu_j^B & \text{if } \mu_j^A = * \text{ and } \mu^B \neq *. \end{cases} \tag{17}$$

From this definition it is easy to verify that

$$\overline{\text{int}(\boldsymbol{\mu}^A, \boldsymbol{\mu}^B)} = \overline{\boldsymbol{\mu}^A} \cap \overline{\boldsymbol{\mu}^B}. \tag{18}$$

Definition (17) does not consider cases where $\mu_j^A \neq *$, $\mu_j^B \neq *$ and $\mu_j^A \neq \mu_j^B$ for one or more $j$.

These cases correspond to table slices with an empty intersection. Whenever this is the case we leave $\text{int}(\boldsymbol{\mu}^A, \boldsymbol{\mu}^B)$ undefined and let $\overline{\text{int}(\boldsymbol{\mu}^A, \boldsymbol{\mu}^B)} = \emptyset$.

In what follows we assume that the collection of edit rules is characterized by a collection of $C$ disjoint slice definitions, so that $S = \bigcup_{c=1}^{C} \overline{\boldsymbol{\mu}_c}$. This typically requires a pre-processing step of the original collection of edits. Manrique-Vallier and Reiter (2014a) propose a simple orthogonalization algorithm based on a repeated application of the $\text{int}(\cdot, \cdot)$ operator that could be used for this purpose. We describe the algorithm in the online supplement.

## 3.2 Gibbs sampler

We use a Gibbs sampler to estimate the posterior distribution of $(\mathcal{X}, \mathcal{E}, \mathcal{Z}, \mathcal{Z}^0, \mathcal{X}^0, \boldsymbol{\pi}, \boldsymbol{\lambda}, \alpha, \epsilon, N)$ for the model in Section 2.3. Given $(\mathcal{Z}, \mathcal{Z}^0, \mathcal{X}^0, \boldsymbol{\pi}, \boldsymbol{\lambda}, \alpha, N)$, we update $(\mathbf{x}, \mathbf{E}, \epsilon)$ using the steps in Section 3.2.1. Given a set of true responses $\mathcal{X}$, i.e., after imputing "corrected" values for fields deemed to be erroneous, we update $(\mathcal{Z}, \mathcal{Z}^0, \mathcal{X}^0, \boldsymbol{\pi}, \boldsymbol{\lambda}, \alpha, N)$ using the sampling strategy in Manrique-Vallier and Reiter (2014a). These steps are shown in Section 3.2.2.

After running the Gibbs sampler to convergence, analysts can obtain posterior inferences from relevant parameters. Alternatively, analysts can treat the samples of $\mathcal{X}$ generated by this algorithm as multiple instances of "corrected" datasets for use in multiple imputation inference (Rubin, 1987). To do so, analysts select a modest number, say $M$, of datasets sufficiently spaced so that they are approximately independent.

### 3.2.1 Sampling $(\mathcal{X}, \mathbf{E}, \epsilon)$

For the model with different error rates for each variable, the full conditional distribution of each $\epsilon_j$ is

$$\epsilon_j | ... \sim \text{Beta}(a_{\epsilon j} + s_j, b_{\epsilon j} + N - s_j) \quad \text{for } j = 1, ..., J \tag{19}$$

where $s_j = \sum_{i=1}^{N} I(E_{ij} = 1)$ and $(a_{\epsilon j}, b_{\epsilon j})$ are specific hyperparameters for variable $j$. For the model with $\epsilon_j = \epsilon$ and $(a_{\epsilon j}, b_{\epsilon j}) = (a_\epsilon, b_\epsilon)$ for all $j$, this expression simplifies to

$$\epsilon|... \sim \text{Beta}(a_\epsilon + s, b_\epsilon + N \times J - s) \tag{20}$$

where $s = \sum_j s_j$.

Sampling from $(\mathbf{E}_i, \mathbf{x}_i)$ is more involved. Since the vector $\mathbf{E}_i$ is completely determined by $\mathbf{x}_i$ and $\mathbf{y}_i$, we cannot form Gibbs steps for sampling $\mathbf{E}_i$ independently of $\mathbf{x}_i$ using the full conditionals $p(\mathbf{x}_i|...)$ and $p(\mathbf{E}_i|...)$. Instead, we sample directly from $p(\mathbf{x}_i, \mathbf{E}_i|...)$ using the conditional factorization

$$p(\mathbf{x}_i, \mathbf{E}_i|...) = p(\mathbf{x}_i|..., -\{\mathbf{E}_i\}) \times p(\mathbf{E}_i|...),$$

where $p(\mathbf{x}_i|..., -\{\mathbf{E}_i\})$ denotes the pmf of $\mathbf{x}_i$, conditional on all the parameters and data except for $\mathbf{E}_i$. Allowing different $\epsilon_j$ to present these pmfs in most general form, we have

$$p(\mathbf{x}_i|..., -\{\mathbf{E}_i\}) \propto I(\mathbf{x}_i \notin S) \prod_{j=1}^{J} \lambda_{jk}[x_{ij}] (\epsilon_j q_j)^{I(x_{ij} \neq y_{ij})} (1 - \epsilon_j)^{I(x_{ij} = y_{ij})} \tag{21}$$

$$p(\mathbf{E}_i|...) = \prod_{j=1}^{J} I(x_{ij} \neq y_{ij}). \tag{22}$$

Sampling from (21) is difficult because of the factor $I(\mathbf{x}_i \notin S)$, which induces dependency among the coordinates of $\mathbf{x}_i$. A simple solution is to use a rejection sampling scheme, sampling repeatedly from (21) without considering the truncation until getting a value $\mathbf{x}_i \notin S$. This method works well when the rejection probability is small. When this is not the case (e.g., under severe prior misspecification for $\epsilon$; see discussion at the end), we can use a conditional sampling strategy that exploits the special structure of $S$, which is a disjoint union of table slices. Noting that

$$p(x_{i1}, ..., x_{iJ}|..., -\{\mathbf{E}_i\}) = \prod_{m=1}^{J} p(x_{im}|..., -\{\mathbf{E}_i, x_{i(m+1)}, x_{i(m+2)}, ..., x_{iJ}\}),$$

we sample the coordinates of $\mathbf{x}_i$ one by one using their partial conditional distributions. For this we rely on the following result, proved in Appendix A.

**Theorem 1.** *Let the region of structural zeros, $S$, be defined by a disjoint collection of table slices, i.e.*

$$S = \bigcup_{c=1}^{C} \overline{\boldsymbol{\mu}_c},$$

*with $\overline{\boldsymbol{\mu}_c} \cap \overline{\boldsymbol{\mu}_{c'}} = \emptyset$ for $c \neq c'$. Then, the partial conditional distribution,*

$$\begin{aligned} &p(x_{im}|..., -\{\mathbf{E}_i, x_{i(m+1)}, x_{i(m+2)}, ..., x_{iJ}\}) \\ &\quad \propto a_m(x_{im}) \left[ \prod_{j=m+1}^{J} b_j - \sum_{c=1}^{C} \left( \prod_{j=m+1}^{J} a_j(\rho_j^{(c,m)})^{I(\rho_j^{(c,m)} \neq *)} b_j^{I(\rho_j^{(c,m)} = *)} \right) \right], \end{aligned} \tag{23}$$

*where $a_j(x) = (\epsilon_j q_j)^{I(x \neq y_{ij})}(1-\epsilon_j)^{I(x = y_{ij})} \lambda_{j z_i}(x)$ for $x \in \{1, ..., L_j\}$, $b_j = \sum_{x=1}^{L_j} a_j(x)$, and*

$$(\rho_1^{(c,m)}, ..., \rho_J^{(c,m)}) = int \left[ (x_{i1}, ..., x_{im}, \underbrace{*, ..., *}_{(J-m)\ times}), \boldsymbol{\mu}_c \right].$$

Applying this theorem, we sample from $(\mathbf{x}_i, \mathbf{E}_i)$ as follows. For $m = 1, \ldots, J$. sample

$$x_{im} \sim \text{Discrete}_{1:L_j}(p_1, ..., p_{L_j})$$

where $p_h = \Pr(x_{im} = h | \ldots, -\{\mathbf{E}_i, x_{i(m+1)}, x_{i(m+2)}, \ldots, x_{ij}\})$. We compute $p_1, \ldots, p_{L_j}$ using result (23) from Theorem 1. We make $e_{ij} = I(x_{ij} \neq y_{ij})$ for $j = 1...J$.

This conditional sampling strategy is guaranteed to work; however, computing probabilities using the formula in (23) can be computationally expensive. As a compromise, we suggest and use a hybrid strategy. We start with the rejection sampling scheme, trying to get a proposal accepted until a maximum number of trials (we used a cutoff of 500 attempts in our calculations in the next section). After that threshold is reached, we default to the conditional sampling method. In our experience, this method is reasonably fast and robust, taking advantage of the low computational cost of sampling directly from the LCM when possible and turning to a more sophisticated sampler when not.

### 3.2.2 Sampling $(\mathcal{Z}, \mathcal{Z}^0, \mathcal{X}^0, \boldsymbol{\pi}, \boldsymbol{\lambda}, \alpha, N)$

We sample $(\mathcal{Z}, \mathcal{Z}^0, \mathcal{X}^0, \boldsymbol{\pi}, \boldsymbol{\lambda}, \alpha, N)$ using an adapted version of the seven steps outlined in Manrique-Vallier and Reiter (2014a).

1. For $i = 1, \ldots, n$, sample $z_i \sim \text{Discrete}_{1:K}(p_1, \ldots, p_k)$, with $p_k \propto \pi_k \prod_{j=1}^{J} \lambda_{jk}[x_{ij}]$.

2. For $j = 1, \ldots, J$ and $k = 1, \ldots, K$, sample $\lambda_{jk[\cdot]} \sim \text{Dirichlet}\left(\xi_{jk1}, \ldots, \xi_{jkL_j}\right)$, with $\xi_{jkl} = 1 + \sum_{i=1}^{n} 1\{x_{ij} = l, z_i = k\} + \sum_{i=1}^{n_0} 1\{x_{ij}^0 = l, z_i^0 = k\}$.

3. For $k = 1, ..., K-1$ sample $V_K \sim \text{Beta}(1 + \nu_k, \alpha + \sum_{h=k+1}^{K} \nu_h)$, for $\nu_k = \sum_{i=1}^{n} 1\{z_i = k\} + \sum_{i=1}^{n_0} 1\{z_i^0 = k\}$. Let $V_K = 1$ and make $\pi_k = V_k \prod_{h<k}(1 - V_k)$ for all $k = 1, ..., K$.

4. For $c = 1, \ldots, C$, compute $\omega_c = \Pr(\mathbf{x} \in \overline{\boldsymbol{\mu}}_c | \boldsymbol{\lambda}, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k \prod_{\mu_{cj} \neq *} \lambda_{jk}[\mu_{cj}]$.

5. Sample $(n_1, \ldots, n_C) \sim NM(n, \omega_1, \ldots, \omega_C)$, and let $n_0 = \sum_{c=1}^{C} n_c$.

6. Let $\kappa \leftarrow 1$. Repeat the following for each $c = 1, \ldots, C$.

   (a) Compute the normalized vector $(p_1, \ldots, p_K)$, where $p_k \propto \pi_k \prod_{j:\mu_{cj} \neq *} \lambda_{jk}[\mu_{cj}]$.

   (b) Repeat the following three steps $n_c$ times:

      i. Sample $z_\kappa^0 \sim \text{Discrete}_{1:K}(p_1, \ldots, p_k)$,

      ii. For $j = 1, \ldots, J$ sample

$$x_{\kappa j}^0 \sim \begin{cases} \text{Discrete}_{1:L_j}(\lambda_{jz_\kappa^0}[1], \ldots, \lambda_{jz_\kappa^0}[L_j]) & \text{if } \mu_{cj} = * \\ \delta_{\mu_{cj}} & \text{if } \mu_{cj} \neq * \end{cases}$$

      where $\delta_{\mu_{cj}}$ is a point mass distribution at $\mu_{cj}$,

      iii. Let $\kappa \leftarrow \kappa + 1$.

7. Sample $\alpha \sim \text{Gamma}\left(a - 1 + K, b - \log \pi_K\right)$.

# 4 Empirical Study

We now empirically illustrate the performance of the EI-DPM for edit-imputation and compare it to a F-H approach to edit-imputation. To do so, we use a subset of the 5% public use microdata file for the 2000 U. S. Decennial Census for the state of New York (PUMS; Ruggles et al., 2010). These data also were used by Manrique-Vallier and Reiter (2014a,b). The data comprise $H = 953,076$ subjects and $J = 10$ categorical variables: ownership of dwelling (3 levels), mortgage status (4 levels), age (9 levels), sex (2 levels), marital status (6 levels), single race identification (5 levels), educational attainment (11 levels), employment status (4 levels), work disability status (3 levels), and veteran status (3 levels). This results in a contingency table with 2,566,080 cells. The data documentation indicates 60 pair-wise combinations of variable levels that are deemed impossible; we take these to form a set of edit rules. For example, for any individual $i$, the true response to the variable OWNERSHIP (ownership of dwelling) cannot take the value "Rented" at the same time as the variable MORTGAGE (mortgage status) takes the value "No, owned free and clear." After translating these rules into table slice definitions and applying the algorithm in Manrique-Vallier and Reiter (2014a), we end up with 567 non-overlapping slice definitions. This represents a substantially simplified characterization of the edit rules, as 2,317,030 of the cells correspond to impossible responses.

We consider the $H$ records as a population, from which we take 500 independent, random samples of $n = 1000$ individuals. Since public use files released by the U. S. Bureau of the Census do not contain errors, we contaminate each of the 500 samples using the independent errors and uniform substitution model, with error rate $\epsilon = 0.4$. Thus, in each of the $10 \times 1000$ entries in each test dataset, approximately $4,000$ have been replaced by a random value different from the actual one. With this error rate, we expect 994.0 records per sample—essentially all of them—to have at least one error. However, not all these errors are detectable. In fact, for a given contaminated sub-sample only approximately 78% of the records with errors actually contain one or more violations of the edit rules, meaning that about 22% of the records contain undetectable errors. We note that 40% is a very large fraction of errors; our objective is to put the EI-DPM method through a challenging stress test.

For each sample, we use the EI-DPM model to generate 50 multiply imputed datasets. We use $\epsilon \sim Beta(1,1)$, expressing complete ignorance about the nature of the error rate. We discuss the effects of this and other choices of prior distribution for $\epsilon$ later in this section. We use multiple imputation combining rules (Rubin, 1987) to estimate all $1,824$ three-way marginal proportions that are estimable from the 500 samples. We also create 50 multiply imputed datasets using the F-H paradigm. Specifically, in each of the 500 samples, we independently employ the R package "EditRules" (De Jonge and Van der Loo, 2012) to select a single set of error-localizations that minimizes the number of changes needed to force each record to satisfy all edits. Since the error-localization solution need not be unique, in case of ties we use the default behavior of the package, which is choosing one solution randomly. Once we select the location of the errors, we blank and multiply-impute the selected values. To make comparisons between F-H and EI-DPM as fair as possible, we generate imputations using the model in Manrique-Vallier and Reiter (2014b), which is similar to the sampler in Section 3.2.

Figure 1 illustrates the effects of the contamination procedure on the quality of inferences, before any edit-imputation. Here we contrast the population values and empirical frequencies of each of the $1,824$ three-way marginal probabilities over the 500 replications, using the samples before and after contamination. Unsurprisingly, the uncontaminated frequencies lie almost perfectly on the main diagonal, and the frequencies from the contaminated data are extremely biased, almost consistently underestimating the population values.

Figure 2 displays the 1,824 3-way margins estimated from the 50 multiply edited-imputed datasets generated by EI-DPM and by the application of the F-H method. Comparing to the estimates obtained from the raw contaminated data in Figure 1, the EI-DPM edit-imputation procedure (Figure 2, left panel) produces remarkably accurate estimates of the target quantities. The F-H edit-imputation approach (Figure 2, right panel) is not as accurate. Even after the error-localization and multiple imputation steps, estimates obtained through the F-H approach exhibit notable bias—in fact, they are similar to those obtained directly from the contaminated data (Figure 1). Figure 3 presents the mean squared error (MSE) for these two groups of estimates. The MSE of F-H estimates tend to be around 20 times larger than those from EI-DPM.

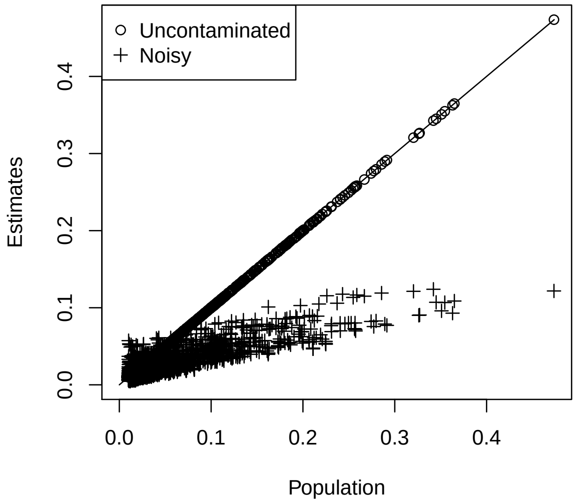

Figure 1: Average over 500 replications of estimates of 3-way margin proportions obtained from 500 contaminated samples ($\epsilon = 0.4$) without edit imputation, versus their actual population values. For clarity we show estimates for quantities whose populational value is larger than 0.01. Estimates from the contaminated samples are extremely biased.

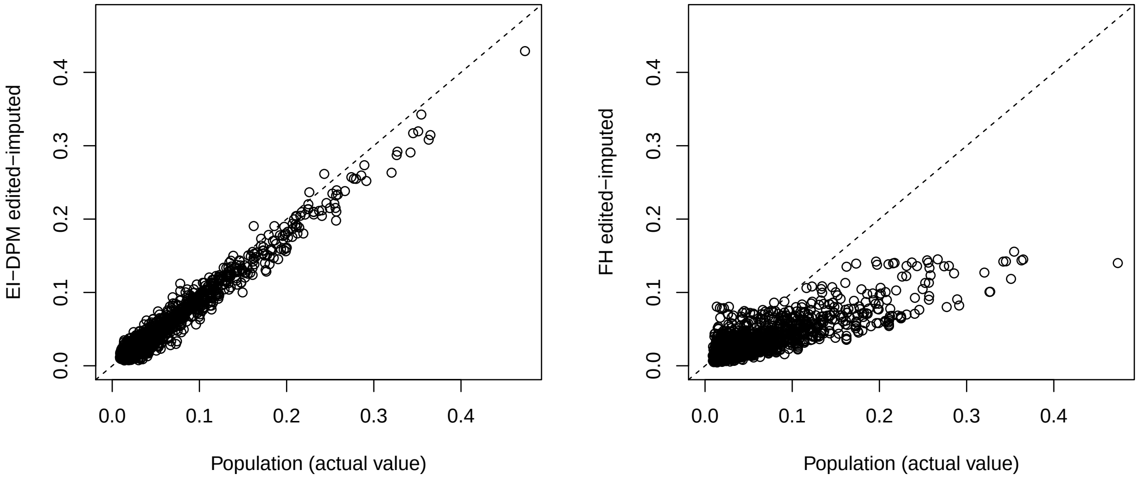

Figure 2: Average over 500 replications of the multiple imputation estimates of 3-way margin proportions versus their actual population values, for faulty data with $\epsilon = 0.4$. EI-DPM in left panel. F-H in right panel.

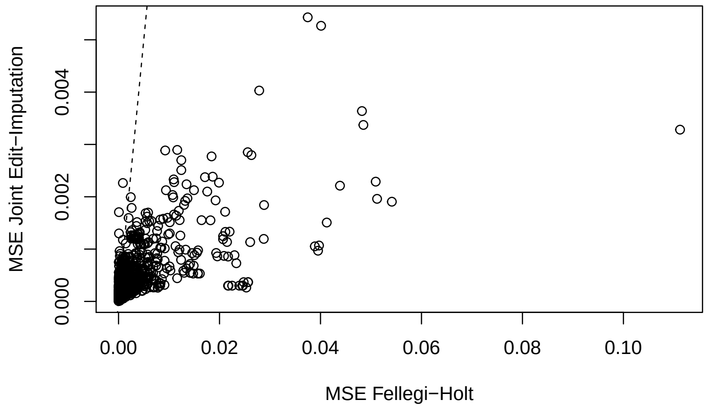

Figure 3: Empirical (over 500 replications) mean squared error of estimates of 1,824 3-way marginal proportions using EI-DPM edit imputation vs. using F-H

Figure 4 displays the empirical coverage rates of 95% confidence intervals for the 1,824 3-way margin test population parameters, using EI-DPM as well as the F-H approach. These intervals are based on the methods of Rubin (1987). We also display the empirical coverage of 95% confidence intervals obtained from multiply imputing the faulty values using the true error locations as generated during the data contamination procedure. We use these values—in principle unattainable from the edited-imputed datasets—as a gold-standard reference to calibrate the coverage rates. Most of the intervals from EI-DPM have at least 80% coverage rates. Typically, the undercoverage results from small absolute biases with small standard errors. In contrast, only 21% of the intervals from the F-H procedure have at least 80% coverage rates, and 30% of them are exactly zero.

We also run the EI-DPM using the prior distribution, $\epsilon \sim \text{Beta}(1, 10^5)$. This expresses a strong (unwarranted in this illustration) belief in the quality of the data, giving essentially a weight equivalent to ten times the data to the prior specification in the posterior inference on $\epsilon$. As a result, we expect to reduce the probability of detecting errors that are not evident, like those that do not result in violation of edit rules. Using one run with $\epsilon = 0.4$ as an example, we found that the mean posterior probability of detecting at least one error in records with inconsistencies is exactly 100%, whereas in faulty records without edit violations it drops down to 5.6%. As expected, the

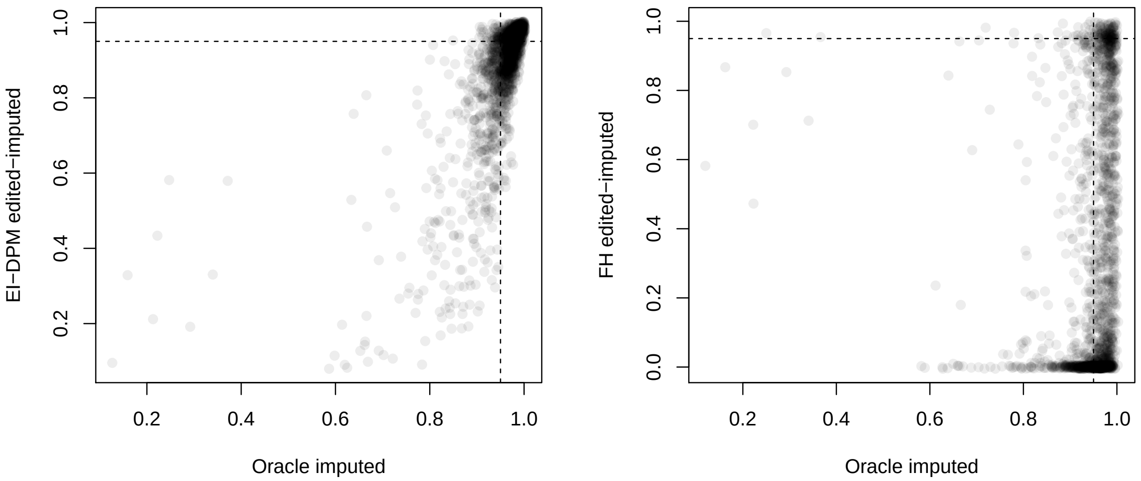

Figure 4: Empirical coverage over 500 replications of multiple-imputation 95% intervals for 1,824 3-way margin proportions obtained from contaminated samples with $\epsilon = 0.4$. Left panel: EI-DPM edited-imputed vs "Oracle" imputed samples. Right panel: F-H vs "Oracle" imputed. Discontinuous lines mark nominal 95% coverage levels. Unif(-0.005, 0.005) noise added for clarity.

failure to catch errors induces large biases in the estimates (left panel Figure 5). We note, however, that these biases are noticeably lower than those obtained with the F-H approach (right panel Figure 5). Stronger priors (e.g. $a_\epsilon = 1$ and $b_\epsilon = 10^6$) result in behaviors even closer to the F-H results, but still with better predictive performance; see the online supplement.

We also examine the method of setting $E_{ij} = 0$ for all $j$ for records $i$ without detectable errors. This prevents the edit-imputation engine from editing records without detectable errors, akin to the F-H principle. Figure 6 displays the result of a repeated sampling experiment in which we randomly contaminate 500 subsamples with rate $\epsilon = 0.4$, but where we leave records that would result in undetectable errors untouched; that is, we reset $\mathbf{y}_i = \mathbf{x}_i$ for these records. The EI-DPM method produces accurate estimates of the target quantities, whereas the F-H method produces highly biased results.

Finally, as a check on whether similar patterns hold at lower rates of error, we repeat the simulation using an error rate of $\epsilon = 10\%$ in place of $\epsilon = 40\%$. As seen in Figure 7, edit-imputation by the EI-DPM offers more accurate estimates than edit-imputation by the F-H method, although

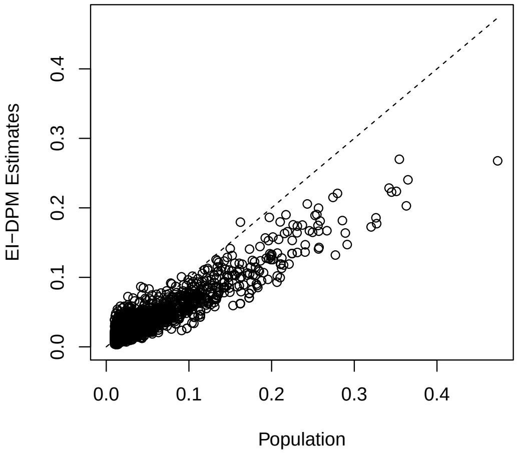

Figure 5: Multiple edit-imputation estimates of 3-way margin proportions versus their actual population values for a single contaminated sample with $\epsilon = 0.4$. The left corresponds to the EI-DPM method with a strong $\epsilon \sim \text{Beta}(1, 10^5)$ prior; the right panel to the F-H method.

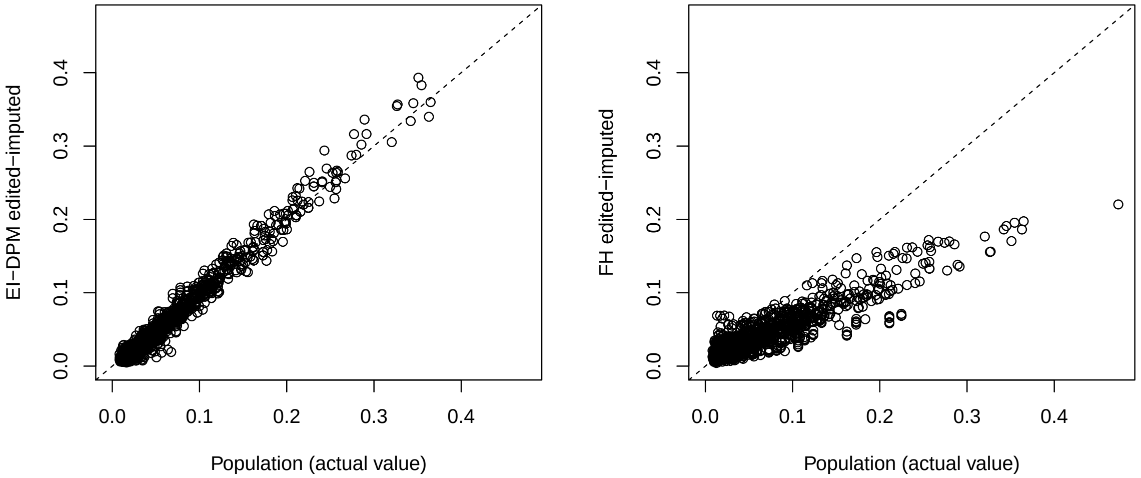

Figure 6: Average over 500 replications of the multiple edit-imputation estimates of 3-way margin proportions versus their actual population values for model and faulty data ($\epsilon = 0.4$) without undetectable errors. EI-DPM method on the left. F-H on the right.

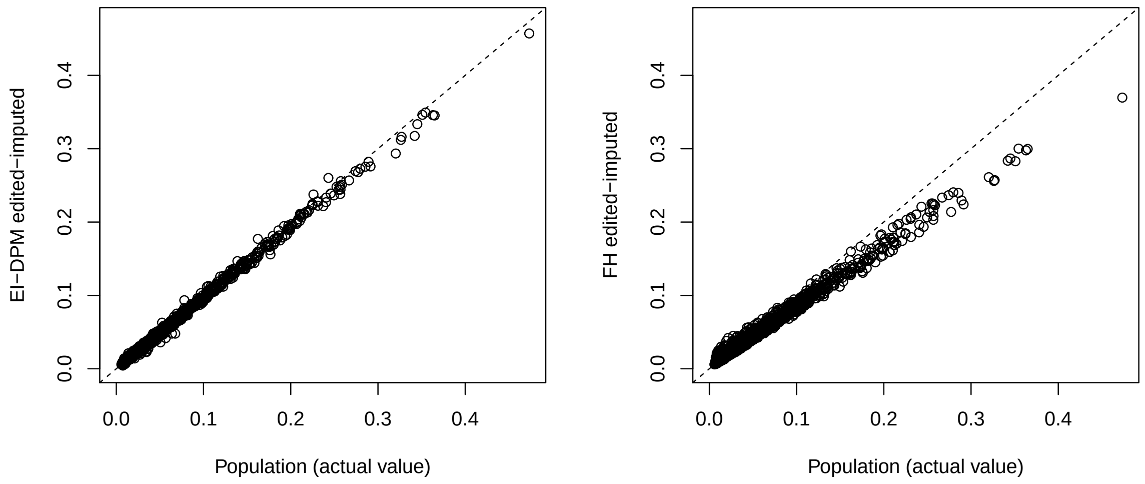

Figure 7: Average over 500 replications of the multiple imputation estimates of 3-way margin proportions versus their actual population values, for faulty data with low error rate $\epsilon = 0.1$. EI-DPM in left panel. F-H in right panel.

as expected the differences are less pronounced due to fewer imputations. See the online supplement for additional results and details on this and other experiements.

## 5 Concluding Remarks

As discussed by Kim et al. (forthcoming), agencies using F-H approaches could force higher probability of editing certain fields based on combinations of variables from other fields. For example, for a reported male with a hysterectomy, the agency could decide to change hysterectomy status from yes to no when the person reports female and 20 years old, whereas change sex from male to female when the person reports having a hysterectomy and 60 years old. Such heuristics could get cumbersome when based on multivariate relationships involving many variables. The EI-DPM automatically lets the data identify unusual combinations based on relationships among all variables, thereby potentially leveraging important associations that were not anticipated by the agency (Kim et al., forthcoming). Unlike F-H approaches with such heuristics, the EI-DPM recognizes the uncertainty in the fields to be edited. The 20 year old person still could be a woman who has undergone a hysterectomy, or a 20 year old man who has not.

The EI-DPM approach can handle missing values simultaneously with erroneous values. One sets $E_{ij} = 1$ for all fields with missing values, forcing $x_{ij}$ to be imputed. We note that this presumes values are missing at random, as is typical in applications of edit-imputation. It also presumes that the same error location and true response models describe the records with errors and records with missing data.

The EI-DPM method can be computationally efficient. In most of our examples with sample sizes $n = 1000$, producing the 50 multiply edited-imputed datasets from a single faulty dataset took only approximately 140 seconds, using personal desktop computers. Computation times are strongly dependent on the particular application. Naturally, longer computing times are required for larger sample sizes (e.g., it took 641 seconds to create 50 completed datasets with a sample of $n = 5000$). Longer computation times also are required when the model uses strong prior specifications for the error rates that do not accord with the true values, for example, using $\epsilon \sim Beta(1, 10^5)$ when in fact $\epsilon = 0.4$. This results from the difficulty of imputing values within the feasible region $\mathcal{C} \setminus S$ when the error rate is estimated with a severe bias. We also note that the proposed sampling algorithm offers ample opportunities for optimization through parallelization. In particular, the many imputation steps (Section 3.2.1, and steps 1 and 6 in Section 3.2.2) can be easily split among different processors or computers. Our current implementation, available upon request, is single threaded C++ with an interface in R. We are currently testing an R package implementing the model in Section 2.3 that will be made available on CRAN.

Independent Bernoulli error location models, potentially with variable-specific error rates, are simple specifications useful for many applied settings. However, sometimes more complex models are appropriate. For example, errors may be correlated within records in situations where individuals tend to have different propensity to commit errors. We conjecture that edit-imputation can be improved in such cases by using mixture models for the vector of error locations. We believe that developing and evaluating versions of the EI-DPM is a topic worthy of future research.

## Supplementary Materials

The supplementary materials include the description of the algorithm for transforming a collection of table slice definitions into a collection of non-overlapping definitions. It also include results from additional simulations.

## References

Biemer, P. P. (2010), "Total Survey Error: Design, Implementation, and Evaluation," *Public Opinion Quarterly*, 74, 817–848.

Bishop, Y., Fienberg, S., and Holland, P. (1975), *Discrete Multivariate Analysis: Theory and Practice*, Cambridge, MA: MIT Press, reprinted in 2007 by Springer-Verlag, New York.

De Jonge, E. and Van der Loo, M. (2012), "editrules: R package for parsing and manipulating of edit rules and error localization," *R package version*, 2.

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

Dunson, D. and Xing, C. (2009), "Nonparametric Bayes modeling of multivariate categorical data," *Journal of the American Statistical Association*, 104, 1042–1051.

Durrant, G. B. and Skinner, C. (2006), "Using missing data methods to correct for measurement error in a distribution function," *Survey Methodology*, 32, 25–36.

Fellegi, I. P. and Holt, D. (1976), "A systematic approach to automatic edit and imputation," *Journal of the American Statistical association*, 71, 17–35.

Fellegi, I. P. and Sunter, A. B. (1969), "A Theory for Record Linkage," *Journal of the American Statistical Association*, 64, 1183–1210.

Fuller, W. A. (1987), *Measurement Error Models*, New York: John Wiley & Sons.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013), *Bayesian data analysis*, CRC press, 3rd ed.

Goodman, L. A. (1974), "Exploratory latent structure analysis using both identifiable and unidentifiable models," *Biometrika*, 61, 215–231.

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

Groves, R. M. (2004), "Measurement Error in Surveys," in *Measurement Errors in Surveys*, eds. Biemer, P. P., Groves, R. M., Lyberg, L. E., Mathiowetz, N. A., and Sudman, S., Hoboken, NJ: John Wiley & Sons, 2nd ed., pp. 1–25.

Groves, R. M. and Lyberg, L. (2010), "Total Survey Error: Past, Present, and Future," *Public Opinion Quarterly*, 74, 849–879.

Ishwaran, H. and James, L. F. (2001), "Gibbs sampling for stick-breaking priors," *Journal of the American Statistical Association*, 96, 161–173.

Ishwaran, H. and Zarepour, M. (2002), "Exact and approximate sum representations for the Dirichlet process," *Canadian Journal of Statistics*, 30, 269–283.

Kim, H. J., Cox, L. H., Karr, A. F., Reiter, J. P., and Wang, Q. (forthcoming), "Simultaneous editing and imputation for continuous data," *Journal of the American Statistical Association.*

Lazarsfeld, P. and Henry, N. (1968), *Latent structure analysis*, Boston: Houghton Mifflin Co.

Manrique-Vallier, D. and Reiter, J. P. (2014a), "Bayesian Estimation of Discrete Multivariate Latent Structure Models with Structural Zeros," *Journal of Computational and Graphical Statistics*, 23, 1061–1079.

— (2014b), "Bayesian Multiple Imputation for Large-Scale Categorical Data with Structural Zeros," *Survey Methodology*, 40, 125–134.

Meng, X. L. and Zaslavsky, A. M. (2002), "Single observation unbiased priors," *The Annals of Statistics*, 30, 1345–1375.

Merrill, R. A. (2008), "Hysterectomy surveillance in the United States, 1997 through 2005," *Medical Science Monitor*, 14, 24–31.

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

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

Ruggles, S., Alexander, T., Genadek, K., Goeken, R., Schroeder, M. B., and Sobek, M. (2010), "Integrated Public Use Microdata Series: Version 5.0 [Machine-readable database]," University of Minnesota, Minneapolis. http://usa.ipums.org.

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.

Steorts, R., Hall, R., and Fienberg, S. (2014), "SMERED: A Bayesian Approach to Graphical Record Linkage and De-duplication." in *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics*, JMLR: W&CP, Reykjavik, Iceland, vol. 33, pp. 922–930.

White, A. and Murphy, T. B. (2014), "BayesLCA: An R Package for Bayesian Latent Class Analysis," *Journal of Statistical Software*, 61.

Winkler, W. (1995), "Editing Discrete Data," in *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 108–113.

— (2008), "General Methods and Algorithms for Imputing Discrete Data under a Variety of Constraints," Tech. Rep. 2008-08, U.S. Bureau of the Census, research Report Series.

Winkler, W. and Petkunas, T. F. (1997), "The DISCRETE edit system," in *Statistical Data Editing*, eds. Kovar, J. and Granquist, L., U.N. Economic Commission for Europe, vol. 2, pp. 52–62.

Winkler, W. E. (2003), "A contingency-table model for imputing data satisfying analytic constraints," in *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

# A Appendix - Proof of Theorem 1

*Proof.* To compute the required pmf we need to marginalize $(x_{i1}, ..., x_{iJ})$ in Equation (21). Let

$$\Omega_m = (x_{i1}, ..., x_{im}, \underbrace{*, ..., *}_{J-m \text{ times}}).$$

Then,

$$\begin{aligned}
p &= p(x_{im}| ... - \{\mathbf{E}_i, x_{i(m+1)}, x_{i(m+2)}, ..., x_{iJ}\}) \\
&\propto \sum_{\mathbf{x}\in\overline{\Omega}_m} I(\mathbf{x} \notin S) \prod_{j=1}^{J} (\epsilon_j q_j)^{I(x_j \neq y_{ij})} (1-\epsilon_j)^{I(x_j = y_{ij})} \lambda_{jk}[x_j] \\
&= \sum_{\mathbf{x}\in\overline{\Omega}_m} I(\mathbf{x} \notin S) \prod_{j=1}^{J} a_j(x_j)
\end{aligned}$$

where $\mathbf{x} = (x_1, ..., x_J)$. Since by hypothesis $S = \bigcup_{c=1}^{C} \overline{\boldsymbol{\mu}_c}$ and the $\boldsymbol{\mu}_c$s are disjoint, we have that

$$\begin{aligned}
I(\mathbf{x} \notin S) &= 1 - I(\mathbf{x} \in S) \\
&= 1 - \sum_{c=1}^{C} I(\mathbf{x} \in \overline{\boldsymbol{\mu}_c}).
\end{aligned}$$

Therefore,

$$\begin{aligned}
p &\propto \sum_{\mathbf{x}\in\overline{\Omega_m}} \left(1 - \sum_{c=1}^{C} I(\mathbf{x} \in \overline{\boldsymbol{\mu}_c})\right) \prod_{j=1}^{J} a_j(x_j) \\
&= \sum_{\mathbf{x}\in\overline{\Omega_m}} \left(\prod_{j=1}^{J} a_j(x_j) - \sum_{c=1}^{C} I(\mathbf{x} \in \overline{\boldsymbol{\mu}_c}) \prod_{j=1}^{J} a_j(x_j)\right) \\
&= \sum_{\mathbf{x}\in\overline{\Omega_m}} \prod_{j=1}^{J} a_j(x_j) - \sum_{c=1}^{C} \sum_{\mathbf{x}\in\overline{\Omega_m}\cap\overline{\boldsymbol{\mu}_c}} \prod_{j=1}^{J} a_j(x_j). \qquad (24)
\end{aligned}$$

Define

$$\boldsymbol{\rho}^{(c,m)} = \left(\rho_1^{(c,m)}, ..., \rho_J^{(c,m)}\right) = \mathrm{int}(\Omega_m, \boldsymbol{\mu}_c). \qquad (25)$$

Then, by properties of the int$(\cdot,\cdot)$ operation, we have that

$$\overline{\Omega_m} \cap \overline{\boldsymbol{\mu}_c} = \overline{\boldsymbol{\rho}^{(c,m)}}.$$

Therefore, substituting in (24),

$$\begin{aligned}
p &\propto \sum_{\mathbf{x}\in\overline{\Omega_m}} \prod_{j=1}^{J} a_j(x_j) - \sum_{c=1}^{C} \sum_{\mathbf{x}\in\overline{\boldsymbol{\rho}^{(c,m)}}} \prod_{j=1}^{J} a_j(x_j) \\
&= \prod_{j=1}^{m} a_j(x_{ij}) \left( \sum_{\mathbf{x}\in\overline{\Omega_m}} \prod_{j=m+1}^{J} a_j(x_j) - \sum_{c=1}^{C} \sum_{\mathbf{x}\in\overline{\boldsymbol{\rho}^{(c,m)}}} \prod_{j=m+1}^{J} a_j(x_j) \right) \\
&\propto a_m(x_{im}) \left( \sum_{\mathbf{x}\in\overline{\Omega_m}} \prod_{j=m+1}^{J} a_j(x_j) - \sum_{c=1}^{C} \sum_{\mathbf{x}\in\overline{\boldsymbol{\rho}^{(c,m)}}} \prod_{j=m+1}^{J} a_j(x_j) \right) \\
&\propto a_m(x_{im}) \left( \prod_{j=m+1}^{J} \sum_{x=1}^{L_j} a_j(x) - \sum_{c=1}^{C} \prod_{\{j:j>m,\rho_j^{(c,m)}\neq *\}} a_j(\rho_j^{(c,m)}) \prod_{\{j:j>m,\rho_j^{(c,m)}=*\}} \sum_{x=1}^{L_j} a_j(x) \right) \\
&= a_m(x_{im}) \left( \prod_{j=m+1}^{J} b_j - \sum_{c=1}^{C} \prod_{\{j:j>m,\rho_j^{(c,m)}\neq *\}} a_j(\rho_j^{(c,m)}) \prod_{\{j:j>m,\rho_j^{(c,m)}=*\}} b_j \right) \\
&= a_m(x_{im}) \left( \prod_{j=m+1}^{J} b_j - \sum_{c=1}^{C} \prod_{j=m+1}^{J} a_j(\rho_j^{(c,m)})^{I(\rho_j^{(c,m)}\neq *)} b_j^{I(\rho_j^{(c,m)}=*)} \right)
\end{aligned}$$

completing the proof.

□