General Methods and Algorithms for Modeling and Imputing Discrete Data under a Variety of Constraints

William E. Winkler[1], william.e.winkler@census.gov 081001 1p
U.S. Census Bureau, Statistical Research Division, Washington, DC 20233-9100

Abstract

Loglinear modeling methods have become quite straightforward to apply to discrete data **X**. The models for missing data involve minor extensions of hot-deck methods (Little and Rubin 2002). Edits are structural zeros that forbid certain patterns. Winkler (2003) provided the theory for connecting edit with imputation. In this paper, we give methods and algorithms for modeling/edit/imputation under linear and convex constraints. The methods can be used for statistical matching (D'Orazio, Di Zio, and Scanu 2006), edit/imputation in which models are also controlled for external constraints such as in benchmark data, and for creating synthetic data with significantly reduced risk of re-identification (Winkler 2007).

Keywords: Data Quality, Loglinear Model Fit, Missing Data, Convex Constraints

1. Introduction

A primary purpose of collecting data is to produce a data file that can be used for two or more analytic purposes. The methods for 'cleaning up' the data include edit/imputation for removing 'implausible combinations of values of fields' and for imputing values that are consistent with underlying uses of the data. The intention of the 'clean-up' procedures is to produce data that are 'fit for use.' In an ideal world, a final data file that has gone through extensive 'clean-up' could be used for several (or all) sets of analyses.

The intent of various clean-up procedures that statistical agencies apply is to replace missing data with plausible values and to change mutually contradictory values to values that are consistent with edit constraints. Edit constraints are conditions that eliminate certain situations such as a child of fifteen or less that is married. Fellegi and Holt (1976) provided a theory of editing that tied certain edit restraints with methods of imputation such as the widely used hot-deck. The advantage of their methods is that edit restraints were contained in easily maintained tables and that imputation (when properly applied) was assured of imputing values that satisfied edits.

Until a few years ago, most editing work involved developing combinatorial optimization algorithms. The algorithms were used in determining the minimum number of fields to impute for missing data or to change for data that contradicted edit restraints. A key difficulty was determining all of the edit restraints that could be logically derived from the original set of edit restraints. Prior to Fellegi and Holt (1976), as a record r_0 was changed to satisfy certain edit constraints that it had failed, the newly changed record r_1 would fail edits that r_0 had not previously failed. Winkler (1997) introduced new set covering and integer programming methods that, in large situations, speeded up computation by as much as a factor of 100 over the previous fastest methods (Garfinkel, Kunnathur, and Liepins 1986).

Winkler (2003) provided theory connecting editing with modern imputation as given by Little and Rubin (2002). In the simplest situations, the methods for imputation involve model building based on straightforward extensions of the missing-at-random assumption that is used in hot-deck imputation. The edit restraints give the structure into which the model building is performed using generalized iterative fitting procedures that have the capability of automatically checking the logical consistency of the system.

The generalization of the modeling/edit/imputation methods allows for convex constraints that can improve the quality of the imputations or control the overall model to benchmark constraints. The convex constraints have been used in record linkage (Winkler 1993) to improve models based on a priori knowledge, in statistical matching to control to benchmarks characteristics (D'Orazio et al. 2006), or in microdata confidentiality to place upper and lower bounds on certain sets of cells to maintain analytic properties while significantly reducing re-identification risk (Winkler 2007). The core set of computational algorithms in these different areas is nearly identical.

This paper provides a summary of the theory and an outline of new generalized modeling/edit/imputation methods that can be easily applied to a large variety of discrete data files quite rapidly. The set covering algorithms (Winkler 1997, Winkler and Chen 2002) provide a file that creates a structure that is used during the main model-building phase. In this paper we do not cover the specifics of finding all of the logically derived edits that, in turn, yield all of the structural zeros needed for the modeling. We assume that the structural zeros are given. The edit part of the software marks certain unacceptable values in combinations of variables as structural zeros.

The model building phase uses a generalized iterative fitting procedure (Winkler 1993) that is similar to the MCECM algorithm of Meng and Rubin (1993) but allows for convex constraints and the structural zeros (Winkler 1990). With the built model (essentially the limited solution under the best-fitting interaction constraints), imputation is straightforward. The values to be imputed are those that are either originally missing or marked as needing to be replaced because certain edits were failed. Cells in the model are matched on the non-missing values in a record and then sampled via a probability- proportional-to-size (*pps*) procedure. The missing values in the record are substituted from the values associated with the sampled cell. Because the model is logically consistent, all imputations are guaranteed to satisfy edits and to preserve the joint distributions of the model.

With hot-deck, there are typically no donors when matching on a moderate number of non-missing values (say 9 of 12 variables are non-missing). To find donors, the number of variables on which matching is performed is typically reduced (*collapsed*) until a potential donor record is found. These hot-deck collapsing rules for finding donors severely distort joint distributions. It is well-known, that even without collapsing, hot-deck will distort joint distributions (e.g. Rao 1997). With the probability structure of the model, variance estimation is straightforward. Estimation of imputation variance is difficult with hot-deck. None of the more than 40 Census Bureau surveys that collect discrete, demographic data and use hot-deck attempt to find the best variables on which to collapse. None attempt estimation of hot-deck imputation variance.

What is remarkable about these new modeling/edit/imputation methods is the ease to which they can be applied to produce good fitting models, high quality joint distributions, and logically consistent data. For smaller discrete-data situations, the methods can be applied in a few days; with larger data situations, several weeks (Winkler and Petkunas 1997; Kovar and Winkler 1996). This is in marked contrast to hot-deck based edit/imputation in which each system is built from scratch, little or no formal modeling is performed to get minimally suitable hot-deck matching rules, hundreds or thousands of if-then-else rules must be checked and often contain errors (both in logic and in programming code), joint distributions are severely compromised, and variance estimation is not possible. Earlier editing methods only assured that imputed values satisfied edits. They did not assure that imputed values preserved joint distributions.

The outline of this paper is as follows. Following this introduction, we provide background on standard loglinear modeling, model building for imputation as in Little and Rubin (2002), the iterative fitting extensions for edit and linear constraints, and the substantial ease with which production systems can be built. In the third section, we provide the main summation of the model building methods that lead to the generalized software. All of the basic methods are known (Winkler 1990, 1993, 2003; Little and Rubin 2002; D'Orazio et al. 2006) but have not been put together in one coherent framework. The methods even extend from general linear constraints to general convex constraints (Winkler 1990, 1993; D'Orazio et al. 2006). The fourth section is discussion, and we end with concluding remarks.

2. Background

In this section we provide background on classical edit/imputation that uses hot-deck and provide a description of how hot-deck was assumed to work by practitioners. As far as we know, there has never been a rigorous development that may justify some of the assumed properties of hot-deck. We also provide background methods of creating loglinear models **Y** (Bishop, Fienberg and Holland 1975) that are straightforward to apply to general discrete data, background on general methods of imputation and editing for missing data under linear constraints that extend the basic methods and can also be straightforward to apply, and an elementary review of the EM algorithm. The application of the general methods and software is straightforward. The application can be done without any modifications that are specific to a particular data file or analytic use.

2.1 Classical data collection, edit rules, and hot-deck imputation

The intent of classical data collection and clean-up was to provide a data file that was free of logical errors and missing data. For a statistical agency, a survey form might be filled out by an interviewer during a face-to-face interview with the respondent. The 'experienced' interviewer would often be able to 'correct' contradictory data or 'replace' missing data during the interview. At a later time analysts might make further 'corrections' prior to the data being placed in computer files. The purpose was to produce a 'complete' (i.e., no missing values) data file that had no contradictory values in some variables. The final 'cleaned' file would be suitable for various statistical analyses. In particular, the statistical file would allow determination of the proportion of specific values of the multiple variables (i.e., joint inclusion probabilities).

In a naïve viewpoint dealing with edits is straightforward. If a child of less than sixteen years old is given a marital status of 'married', then either the age associated with the child might be changed (i.e., to older than 16) or the marital status might be changed to 'single'. The difficulty consistently arose that, as a (computerized) record r_0 was changed to a different record r_1 by changing values in fields in which edits failed, then the new record r_1 would fail other edits that the original record r_0 had not failed.

Fellegi and Holt (1976) were the first to provide an overall model to assure that a changed record r1 would not fail edits. Their theory required the computation of all implicit edits that could be logically derived from an originally specified set of 'explicit' edits. If the implicit edits were available, then it was always possible to change an edit-failing record r_0 to an edit passing record r_1. The availability of 'implicit' edits makes it quite straightforward and fast to determine the minimum number of fields to change in an edit-failing record r_0 to obtain an edit-passing record r_1. Further, Fellegi and Holt indicated how hot-deck might be used to provide the values for filling in missing values or replacing contradictory values. As we shall see, hot-deck is not generally suitable for filling in missing values in a manner that yields records that satisfy edits and preserve joint distributions.

The early set-covering algorithms necessary for the computation of 'implicit' edits required extremely large amounts of computer time (Garfinkel, Kunnathur, and Liepins 1986). A later algorithm (Winkler 1997), while as much as 100 times as fast, is not completely theoretically valid but works in most situations where skip patterns are not present in the survey form. Due to hardware-speed increases, the latter algorithm should work well in most day-to-day survey situations. Boskovitz (2008) recently provided a complete theoretical development, however software based on her algorithms has not yet been written and will likely be 10 times as slow due to the significantly greater amount of information that must be accounted for at different levels of the computational algorithms.

The intent of filling-in missing or contradictory values in edit-failing records r_0 is to obtain a records r_1 that can be used in computing the joint probabilities in a principled manner. The difficulty that had been observed by many individuals is that a well-implemented hot-deck does not preserve joint probabilities. Indeed, Rao (1997) provided a theoretical characterization of why hot-deck fails even in two-dimensional situations. The failure occurs even in 'nice' situations where individuals had previously assumed that hot-deck would work well.

In a real-world survey situation, subject matter 'experts' may develop hundreds or thousands of if-then-else rules that are used for the editing and hot-deck. Because it is exceptionally difficult to develop the logic for such rules, most edit/imputation systems do not assure that records satisfy edits or preserve joint inclusion probabilities. Further, such systems are exceptionally difficult to implement because of (1) logic errors in specifications, (2) errors in computer code, and (3) no effective modeling of hot-deck matching rules. As we will demonstrate, it is effectively impossible with the methods (classical if-then-else and hot-deck) that many agencies use to develop edit/imputation systems that preserve either joint probabilities or that satisfy edit restraints. This is true even in the situations when Fellegi-Holt methods are used for the editing and hot-deck is used for imputation.

An edit/imputation system that effectively uses the edit ideas of Fellegi and Holt (1976) and modern imputation ideas (such as in Little and Rubin 2002) has distinct advantages. First, it is far easier to implement (as we will demonstrate). Edit rules are in easily modified tables, and the logical consistency of the entire system is tested automatically according the mathematics of the Fellegi-Holt model and additional requirements on the preservation of joint inclusion probabilities (Winkler 2003). Second, the optimization that determines the minimum number of fields to change or replace in an edit-failing record is in a fixed mathematical routine that does not need to change. Third, imputation is determined from a model (limiting distribution). Most modeling is very straightforward. It is based on variants of loglinear modeling and extensions of missing data methods that is contained in easily applied, extremely fast computational

algorithms (Winkler 2006, 2007, 2008b). The methods create records that *always* satisfy edits and preserve joint inclusion probabilities.

2.2 How classical hot-deck is assumed to work

In this subsection we provide an explanation of some of the (possibly) subtle issues that significantly degrade the overall analytic characteristics of realistic data files (8 or more variables) that are subjected to *well-implemented* hot-deck. The reason that the issues may be subtle is that in many situations with hot-deck, the probabilistic model is not written down and the effects of the statistical evaluations (say logistic or ordinary regression) on hot-deck collapsing rules for matching are not evaluated. We will describe why it is effectively impossible in many practical survey situations to do the empirical testing and develop program logic necessary for a well-implemented hot-deck. Prior to this we provide some notation and background that will allow us to describe why hot-deck breaks down in terms of the basic modeling frameworks of Little and Rubin (2002) and Winkler (2003).

We assume $\mathbf{X} = X_i = (x_{ij})$, $1 \le i \le N$, $1 \le j \le M$ is a representation of the survey data with N rows (records) and M columns (variables). Record x_i has values x_{ij}, $1 \le j \le M$. The j^{th} variables X_j takes values x_{jk}, $1 \le k \le n_j$. The total number of patterns is $npat = n_1 \times \ldots \times n_M$. In most realistic survey situations (8 or more variables), the number of possible patterns *npat* is far greater than N (i.e., $N << npat$). Under classical hot-deck assumptions (that are essentially universally used in statistical agencies), the typical assumption is that we will be able to match a record $r_0 = (x_{01}, x_{02}, \ldots., x_{0M})$ having missing values of certain variables against a large number of donor records that have no missing variables and that agree with record r_0 on the non-missing values. If record r_0 has eight variables with the last three variables having missing values, then the intent of hot-deck (after it is implemented over an entire file) is to create a set of records that preserve the original probability structure of a hypothetical file **X** having no missing values.

We start with record $r_0 = (x_{01}, x_{02}, \ldots, x_{05}, b, b, b)$ where *b* represents a missing value for x_{06}, x_{07}, and x_{08}. Under the hot-deck assumptions, our matching would allow use to effectively draw from the distribution of $P(X_6, X_7, X_8 \mid X_1=x_{01}, \ldots, X_5=x_{05})$. In practice with real-world data, we typically have zero donors (rather than an exceptionally large number that would be needed to preserve joint distributions). Statistical agencies typically use ad hoc collapsing in which they attempt to match on a subset of the values $x_{01}, x_{02}, \ldots, x_{05}$. For instance, there may be a matching hierarchy in which the first match attempt is on x_{01}, x_{02}, x_{03}. If a donor record is not found matching may be done on x_{01} and x_{02}. If no donor is found, then matching might be done on only x01 where it might be possible to always find a donor.

If we are able to match on x_{01}, x_{02} and x_{03}, we obtain a record $r_d = (x_{d1}, \ldots. x_{d8})$ that yields a hot-deck completed record $r_{0c} = (x_{01}, \ldots, x_{05}, x_{d6}, x_{d7}, x_{d8})$. There is no assurance that the substituted values will preserve joint distributions or create a record that satisfies edits. Indeed, elementary empirical work with exceptionally simple simulated data (that should preserve joint distributions under the hot-deck assumption) also demonstrate that joint distributions are not preserved. Although the elementary work uses data situations that are much nicer than many real-world situations, it still fails to yield hot-deck imputations that preserve joint distributions. To preserve joint distributions, it might be necessary to create some type of basic model for collapsing. A simplistic approach might be to use logistic regression to find what subsets of $x_{01}, \ldots, x_{05}$ are the best predictors of the remaining variables and choose the collapsing hierarchy based on a very large set of logistic regressions.

Even after such work (that is very specific to an individual data set), it is not clear why the joint distributions would be preserved. It would be much better to have a general modeling framework (possibly an extension of Little and Rubin (2002), chapter 13) and software that would work for arbitrary discrete data under mild assumptions. One mild assumption is the *missing-at-random* assumption (Little and Rubin 2002) that is effectively the hot-deck assumption in a framework in which it is possible to preserve joint inclusion probabilities. An effective model might be multinomial (or multinomial with weak Dirichlet prior) that all non-structural-zero cells are given a non-zero (but possibly very close to zero) values. In this situation (p_i), $1 \le i \le npat$ are the probabilities of the multinomial with the individual cells, and we have a suitable probability structure. With this extended hot-deck (effectively Little-Rubin ideas), we match against cells that agree with the non-missing part of a record r0 and choose one cell (donor pattern or record) with probability proportional to size of the cell probability.

2.2 Loglinear modeling

Standard references for loglinear modeling and categorical data analysis are Agresti (2007) and Bishop, Fienberg, and Holland (1975). If $\mathbf{X_i}$ is a table of cell counts from a discrete population, then we are interested in finding a more parsimonious representation $\hat{X}_i$ of the table. If *k*-dimensional table $\mathbf{X_i}$ has $n = n_1 \times n_2 \times \ldots \times n_k$ cells, then we fit models via an iterative proportional fitting procedure (IPF) in which we fit to specific observed margins in cyclic order. The IPF procedure is assured to increase likelihood and typically increases to a maximum likelihood estimate under a multinomial or Poisson model. We will assume multinomial.

In the 3-dimensional situation, we have a general index $\mathbf{i} = (i_1, i_2, i_3)$. The original table has $n_1 \times n_2 \times n_3$ - 1degrees of freedom. The different fitted models will have fewer degrees of freedom. If we fit an independence model, then we successively fit to the observed margins for each single variable and repeat until convergence. If we fit using an all 2-way interaction model, we fit successively to the set of n_1 x n_2 margins determined by X_1 and X_2, then the $n_1 \times n_3$ margins, then $n_2 \times n_3$ and repeat in a cyclic manner. If we use a multinomial distribution (or multinomial distribution with dirichlet prior), we have the log likelihood

$$\sum_{i \le npat} x_i \ln(\hat{p}_i^t) \tag{1}$$

where the summation is over the set of cells, x_i is the count in cell i, and $\hat{p}_i^t$ is the probability estimate from the multinomial distribution at the t^{th} stage of the fitting. We use the standard convention that 0 ln 0 = 0. If we have complete data in **X**, then the log likelihood is known to increase to the maximum likelihood

$$\sum_{i \le npat} x_i \ln(x_i / N) \cdot \tag{2}$$

Commercial software and the software written for this paper have straightforward ways of specifying the interactions. The software user puts in an epsilon that controls the desired accuracy of the final fitted model and an upper bound on the number of iterative-fitting cycles. The software typically monitors either the maximum deviation of the fitted model $\hat{X}$ from the original data **X** or the Kullback-Liebler distance. The maximum deviation is given by

$$\max_{k \le npat} |(\hat{X}_k - X_k)/tot(X)| \tag{3}$$

where *tot(X)* is the sum of the entries in **X,** $\mathbf{X_k}$ is the proportion on the total count tot(X) in cell *k*, and $\hat{X}_k$ is the proportion of the modeled count in cell *k*. The Kullback-Liebler distance is given by

$$\sum_k X_k \ln(X_k / \hat{X}_k)/tot(X). \tag{4}$$

If we multiply the Kullback-Liebler distance by 2 *tot(X),* then we get the difference between the maximum likelihood and the likelihood from the fitted $\hat{X}_k$.

The approximate Chi-square value corresponding to the Kullback-Liebler distance is given by

$$\sum_k 2X_k \ln(X_k / \hat{X}_k). \tag{5}$$

The Kullback-Liebler distance tends to be more conservative than the maximum deviation.

2.4 Missing Data Imputation and Editing

In this section we describe methods of creating a loglinear model **Y** (Bishop, Fienberg and Holland 1975) that is straightforward to extend to models that take account of *edit restraints* (Fellegi and Holt 1976). If **X** is original data and $\mathbf{X_i}$ is a specific cell, then $\mathbf{X_i}$ is a structural zero or cell forbidden by an edit if its count must be zero. For instance, in most Western societies and others, a fifteen year old child is not allowed to be married. To implement a set of edits that are typically pre-specified by subject matter experts, we enumerate them and then use algorithms (Winkler 1997) to determine all implicit edits. For now, we do not need to

know details. We merely need to know that there are straightforward methods for enumerating all implicit edits (structural zeros). Although the current enumeration methods are still ad hoc (Winkler 1997, Winkler and Chen 2002), the methods will quickly enumerate the implicit edits (structural zeros) for most survey situations in which the survey form does not contain skip patterns.

The theoretical justification connecting editing (structural zeros) with imputation (e.g., Little and Rubin 2002) or loglinear modeling is given in Winkler (2003, 1990). The main difference between the EM algorithm of Little and Rubin (2002) and that of Winkler (2003) is that the more general EM algorithm accounts for the structural zeros (and convex constraints that are described in the next section). If the constraints are consistent with the model assumptions and the data, then the fitting procedure is guaranteed to converge to a solution (the likelihood increases) that is dependent on the starting point. We may need to use a set of starting points to insure the robustness of the limiting model that we select. Little and Rubin (2002) typically use a starting point based on the complete data. A complete-data starting point is not always available because the complete data may not be considered to be sufficiently reliable, the sample size may be too small, or an insufficient number of cells have nonzero values.

The models **M** (or $\hat{X}_i$ **)** for the original data X are created under a *missing-at-random (mar)* assumption that is also assumed by hot-deck imputation methods. The basic methods involve imputation only using the EM or various generalizations of EM (Meng and Rubin 1993, Winkler 1993) that have been applied for modeling in the context of statistical matching (D'Orazio, Di Zio, and Scanu 2006) or edit/imputation (Winkler 2003) under either the linear constraints of loglinear modeling or more general convex constraints (Winkler 1990, 1993).

Although we can use a model **M** (e.g. multinomial or multinomial with Dirichlet prior), it is much more convenient to use the fitted data $\hat{X}_i$ directly. If we have a record *r* that contains missing data, then we merely match against the records in $\hat{X}_i$ that agree on the non-missing values in *r* and sample the set of matched records probability proportional to size. The replacement values needed in *r* are substituted from the sampled record r_s to obtain a completed record r_c. This procedure preserves the joint probabilities from the original data **X** in a principled manner.

Hot-deck procedures (when exceptionally well implemented) do not preserve joint probabilities (e.g., Rao 1997, Di Zio et al. 2004). An additional difficulty with hot-deck is the modeling (often using logistic or ordinary regression) needed to create the hot-deck matching rules. If a record represents 12 variables of which three have missing values, then collapsing is often performed so that matching is performed on considerably less than nine variables. The ad hoc and difficult nature of the collapsing rules can cause further distortions in the joint probability distributions and may even distort univariate distributions. With the general procedures of Little and Rubin or of this paper, estimation of the variance component due to imputation is straightforward.

2.5 Review of the EM Algorithm

The EM algorithm was first described generally in Dempster, Laird, and Rubin (1977). A general MCECM (multi-cycle expectation conditional maximization) for linear constraints was described by Meng and Rubin (1993). Winkler (1993) provided a general MCECM algorithm that was extended to convex constraints. It was used in an application to record linkage in which a priori bounds were placed on certain combinations of cells.

We begin by describing the general EM for linear constraints. If f(Y | Θ) is the density associated with data Y given a parametric form represented by Θ, then

$$f(Y|\Theta) = f(Y_{obs}, Y_{mis}|\Theta) = f(Y_{obs}|\Theta) f(Y_{mis}|Y_{obs}, \Theta), \quad (6)$$

where Y_{mis} represents the missing data in *Y* and Y_{obs} represents the observed data. The missing data Y_{mis} can be as a result of actual missing data or of data that are 'blanked' due to a record failing an edit. Taking natural logarithms, the likelihood can be written as

$$\ell(\Theta|Y_{obs}) = \ell(\Theta|Y) - \ln f(Y_{mis}|Y_{obs}, \Theta), \quad (7)$$

where $\ell(\Theta \mid Y_{obs})$ is the observed data likelihood, $\ell(\Theta \mid Y)$ is the complete data likelihood that is quite straightforward to maximize in many situations, and the last term in equation (7) is the missing part of the complete data likelihood.

If we take the expectation of both sides of equation (7) over the missing data Y_{mis} given the observed data Y_{obs} and the current estimate of Θ, say $\Theta^{(t)}$, then we obtain

$$\ell(\Theta \mid Y_{obs}) = Q(\Theta \mid \Theta^{(t)}) - H(\Theta \mid \Theta^{(t)}), \tag{8}$$

where

$$Q(\Theta \mid \Theta^{(t)}) = \sum_{mis} \ell(\Theta \mid Y_{obs}, Y_{mis})\, f(Y_{mis} \mid Y_{obs}, \Theta^{(t)}) \tag{9}$$

and

$$H(\Theta \mid \Theta^{(t)}) = \sum_{mis} \ln f(Y_{mis} \mid Y_{obs}, \Theta)\, f(Y_{mis} \mid Y_{obs}, \Theta^{(t)}) \tag{10}$$

and the summations are over the missing data cells. By Jensen's inequality (Rao 1972, p. 47), we have

$$H(\Theta \mid \Theta^{(t)}) \le H(\Theta^{(t)} \mid \Theta^{(t)}). \tag{11}$$

The appendices give details of several EM algorithms that include convex constraints for controlling the lower and upper bounds of margins, cell probabilities, and combinations of cells to external constraints. The generality of Jensen's inequality (11) gives great flexibility during the E-step. Under very general conditions, I-Projections under convex constraints correspond to maximum likelihood estimates obtained in the M-step (Dykstra and Lemke 1988, El Barmi and Dykstra 1998). Classic iterative fitting algorithms find maximum likelihood estimates via a procedure that progressively finds I-Projections under linear constraints (Bishop, Fienberg, and Holland 1975). Because we are working with a set of probabilities over a set of cells, we can often place upper bounds on certain complementary probabilities that correspond to lower bounds on probabilities (Winkler 1993). This gives great flexibility in assuring that the models represented by the final estimate of Θ approximately satisfy a set of marginal constraints from benchmarks and constraints due to the original data.

3. Illustrating Examples

We consider two variants of a data set and several extensions and variants of the basic EM modeling methods of Little and Rubin (2002, Chapter 13). We compare the imputation methods with hot-deck. Where needed, we also give variants of the EM fitting methods with more realistic data and with convex constraints.

3.1 Example using an Expanded Data of D'Orazio, Di Zio, and Scanu

In this subsection, we use a slightly modified version of the data of D'Orazio et al. (2006). Although the data were used in an application of statistical matching, we use it for an application of edit/imputation that is a straightforward extension of the ideas of Little and Rubin. Our version of the data expands to a fourth variable and one additional edit constraint. The expanded data make it much more straightforward to demonstrate the severe limitations of hot-deck imputation.

To make some of the ideas more clear, we use a modified version of the data of D'Orazio, Di Zio, and Scanu (2006). The data is a sample of records from a large Italian survey for which only three fields were considered. The original fields are AGE, PRO (profession), and EDU (education). An additional field EX is used to illustrate edit constraints and how hot-deck fails to impute records satisfying edit restraints.

Table 1. Response Categories for Fields

Fields	Transformed Response Categories
Age (AGE)	"0"=15-17 years old; "1"=18-22; "2"=23-64; "3"=65+;
Profession (PRO)	"0"=Manager; "1"=Clerk; "2"=Worker
Education (EDU)	"0"=None or compulsory school; "1"=Vocational school; "2"=Secondary school; "3"=Degree
Extra (EX)	"0"=first value; "1"=second value

There are 96 (=4x3x4x2) data patterns of which many are structural zeros. For instance, a person of age 15-17 cannot have a college degree. The sample size is 2313 that we give as a table of counts and then as a table of probabilities. In Tables 2 and 3, we use a lexicographic ordering in which (0,0,0,0)=0, (0.0,0,1)=1, …, (3,2,3,1)=95. We obtain this with the mapping (a1, a2, a3, a4)=a1*24+a2*8+a3*2+a4*1. The first row of the table is cells 0-7; the second row is cells 8-15, and so on. We use 'z' to represent a structural zero that always have probability zero of having a positive value.

The counts in Table 2 can be representative of counts that we might obtain by taking a relatively small proportion of records from a larger population file. If the counts in Table 2 are based on a 1% sampling rate, then most cells having original population counts of 40 or less will not be in the sample (i.e., they are sampling zeros). *Sampling zeros* are cells having count of zero in the sample where the same cells in the original population can have counts greater than zero. A sampling zero is a cell (pattern based on the values that the specific variables may assume) that can plausibly occur with a probability greater than zero in a (much) larger population. To assure that cells corresponding with sampling zeros can possibly be associated with non-zero probability estimates, we often use a Dirichlet prior (say, 0.1) with the multinomial distribution.

Table 2. Population Counts from Sample File

z	z	z	z	z	z	z	z
z	z	z	z	z	z	z	z
15	z	0	0	z	z	z	z
z	z	z	z	z	z	z	z
z	z	0	0	6	0	z	z
27	z	5	0	9	0	z	z
z	z	z	z	95	47	145	75
z	z	86	37	420	133	63	24
759	z	62	28	100	43	0	0
z	z	z	z	0	0	0	0
z	z	0	0	0	0	0	0
12	z	0	0	0	0	0	0

Table 3. Probabilities for Cells from Sample File

0.00000	*0.00000*	*0.00000*	*0.00000*	*0.00000*	*0.00000*	*0.00000*	*0.00000*
0.00000	*0.00000*	*0.00000*	*0.00000*	*0.00000*	*0.00000*	*0.00000*	*0.00000*
0.00678	*0.00000*	0.00000	0.00000	*0.00000*	*0.00000*	*0.00000*	*0.00000*
0.00000	*0.00000*	*0.00000*	*0.00000*	*0.00000*	*0.00000*	*0.00000*	*0.00000*
0.00000	*0.00000*	0.00000	0.00000	0.00271	0.00000	*0.00000*	*0.00000*
0.01220	0.00000	0.00226	0.00000	0.00407	0.00000	*0.00000*	*0.00000*
0.00000	*0.00000*	*0.00000*	*0.00000*	0.04293	0.02124	0.06552	0.03389
0.00000	*0.00000*	0.03886	0.01672	0.18979	0.06010	0.02847	0.01085
0.34297	0.00000	0.02802	0.01265	0.04519	0.01943	0.00000	0.00000
0.00000	*0.00000*	*0.00000*	*0.00000*	0.00000	0.00000	0.00000	0.00000
0.00000	*0.00000*	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
0.00542	*0.00000*	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000

Italics represent structural zeros. Cell probabilities are independently rounded to 5 decimal places.

The empirical data is useful due to having only four variables and 96 cells. Because there are 50 structural zeros, we do not have much flexibility in the fitting procedures. With larger, more realistic situations the much smaller proportion of structural zeros will make the fitting much easier. The structural zeros are induced by certain marginal constraints that are the edit rules specifying that certain combinations of values of certain variables are forbidden. For instance, the marginal (edit) restraint (0, 0, ., .) indicates that any cell for which there is a value of 0 for the first variable and value of 0 for the second variable corresponds to structural zeros.

Table 4. Edits

(0, 0, ., .)	(0, 1, ., .)	(0, ., 2, .)	(0, ., 3, .)
(., 1, 0, .)	(1, ., 3, .)	(., 0, 0, .)	(., 0, 1, .)
(., ., 0, 1)			

If we fit a three-way interaction model, we obtain the probabilities given in Table 5 that agrees quite closely with the true underlying probabilities in Table 3. We build the model with the complete data records only.

Table 5. Estimated Probabilities for Cells Using All 3-way Interactions

0.00000	*0.00000*	*0.00000*	*0.00000*	*0.00000*	*0.00000*	*0.00000*	*0.00000*
0.00000	*0.00000*	*0.00000*	*0.00000*	*0.00000*	*0.00000*	*0.00000*	*0.00000*
0.00678	*0.00000*	0.00002	0.00002	*0.00000*	*0.00000*	*0.00000*	*0.00000*
0.00000	*0.00000*	*0.00000*	*0.00000*	*0.00000*	*0.00000*	*0.00000*	*0.00000*
0.00000	*0.00000*	0.00045	0.00037	0.00267	0.00182	*0.00000*	*0.00000*
0.01220	0.00000	0.00221	0.00229	0.00472	0.00399	*0.00000*	*0.00000*
0.00000	*0.00000*	*0.00000*	*0.00000*	0.04339	0.02078	0.06506	0.03435
0.00000	*0.00000*	0.03973	0.01584	0.18850	0.06140	0.02893	0.01039
0.34297	*0.00000*	0.02719	0.01349	0.04598	0.01863	0.00006	0.00003
0.00000	*0.00000*	*0.00000*	*0.00000*	0.00000	0.00000	0.00004	0.00005
0.00000	*0.00000*	0.00000	0.00000	0.00001	0.00000	0.00001	0.00001
0.00542	*0.00000*	0.00005	0.00005	0.00006	0.00005	0.00000	0.00000

Italics represent structural zeros. Cell probabilities are independently rounded to 5 decimal places.

With the example data, it is straightforward to demonstrate how the hot-deck can break down. Assume we have record r_0 = (3, 2, ., 1) where the value of the third variable is missing. If we attempt to match on the values (3, 2, ., 1), there are no records in the set of complete data records having those values. If we attempt to match on values (3, ., ., 1) only, there are also no complete data records that we can use as donors. If we match on the value of 3 for the first variable, we can obtain the complete data record r_d = (3, 2, 0, 0) that we use to create the completed data record r_{oc}=(3, 2, 0, 1) by substituting that value of the third variable from record r_d. The record r_{oc}=(3, 2, 0, 1) fails edits.

The situation of hot-deck causing edit failures never occurs with imputation using the probabilities of Table 5 (or an alternate table that would model by including records with missing data). First, we always have matching cells. In the case of matching on (3, 2, ., 1), we obtain cells 91=(3, 2, 1, 1), 93=(3, 2, 2, 1), and 95=(3, 2, 3, 1) that have estimated probabilities 0.00005, 0.00005, and 0.00000 (rounded for 5 decimal places), respectively. When we sample these matching cells with probability proportional to size, we sample cell (3, 2, 1, 1) approximately 50% of the time and cell (3, 2, 2, 1) approximately 50% of the time. The *pps* sampling procedure assures that joint inclusion probabilities are maintained. It is well known that hot-deck does not maintain joint inclusion probabilities and, in certain rare situations, may not maintain single-variable inclusion probabilities. Our model and match-sample procedures assure that edit constraints are satisfied.

Summarizing Comments

The procedure in which we create models (sets of probabilities) as in Table 5 assures that the imputation procedure maintains joint inclusion probabilities and creates records that *always* satisfy edit constraints.

As noted by Winkler and Petkunas (1997), it is quite straightforward to put edits into tables of a generalized system that implements the ideas of Fellegi and Holt. They were able to put the edits into tables and build the edit software in less than a half day for each of two small surveys. Their system assured that imputed missing or 'replaced' values always created records that satisfied all edits. By 'replaced' we mean certain values of variables that are set to missing due to failing edits prior to being imputed. Furthermore, Fellegi-Holt systems can check with logical consistency of the entire edit system without training/test data. The actual source code never needs to be changed because the edits are contained in easily maintained tables.

3.2 2nd Example using an Expanded Data of D'Orazio, Di Zio, and Scanu

The intent in this section is to provide an expanded example that extends the ideas in sections 3.1. The extension also generalizes the methods of Little and Rubin (2002, section 13.4) to situations where we additionally use convex constraints. The computational algorithms are based on the EMH algorithm (Winkler 1993). We use the best fitting (all 3-way interaction) model that was used in the fitting of section 3.1. We induce (mild) non-ignorable nonresponse and then use convex constraints to 'adjust' the fitted data to better correspond to underlying true population proportions. If we impute from the fitted model, then we better preserve the underlying true population proportions. This situation possibly corresponds to the situation where subject matter experts might look at certain margins to determine how closely the margins correspond to alternative estimates from external data.

We use a second expanded set of data that are similar to the data of D'Orazio et al. (2006). The counts in Table 6a represent the subtable where EX=1, the counts in Table 6b represent the subtable where EX=0, and the counts in Table 6c represent the subtable where EX=missing. The methods for this example have some similarity to the methods of section 3.1. We created the *non-ignorable* missing data in the following manner. We first took a 50% sample of the records in Tables 6a and 6b and then dropped 50% of the records in the sample that had EX=1. This means that the probability of nonresponse corresponds to whether EX=0 or EX=1. The fully observed complete data in Tables 6a and 6b are available cases that we assume correspond to the true population proportions.

Table 6a. Expanded Data (EX = 1 subtable)

		Education (EDU)			
Age (AGE)	Profession (PRO)	None or	Vocational	Secondary	Degree
		(a) completely classified			
15-17	Manager	z	z	z	z
	Clerk	z	z	z	z
	Worker	z	0	z	z
18-22	Manager	z	z	z	z
	Clerk	z	1	0	z
	Worker	z	0	0	z
23-64	Manager	z	z	32	105
	Clerk	z	23	138	32
	Worker	z	10	38	0
65+	Manager	z	z	0	0
	Clerk	z	0	0	0
	Worker	z	0	0	0

Table 6b. Expanded Data (EX = 0 subtable)

		Education (EDU)			
Age (AGE)	Profession (PRO)	None or	Vocational	Secondary	Degree
		(a) completely classified			
15-17	Manager	z	z	z	z
	Clerk	z	z	z	z
	Worker	15	0	z	z
18-22	Manager	z	z	z	z
	Clerk	z	0	5	z
	Worker	27	5	9	z
23-64	Manager	z	z	110	115
	Clerk	z	100	415	55
	Worker	759	80	105	0
65+	Manager	z	z	0	0
	Clerk	z	0	0	0
	Worker	12	0	0	0

Table 6c. Expanded Data (EX = missing subtable)

		Education (EDU)			
Age (AGE)	Profession (PRO)	None or	Vocational	Secondary	Degree
		(a) completely classified			
15-17	Manager	z	z	z	z
	Clerk	z	z	z	z
	Worker	7	0	z	z
18-22	Manager	z	z	z	z
	Clerk	z	0	3	z
	Worker	13	2	4	z
23-64	Manager	z	z	63	83
	Clerk	z	56	242	30
	Worker	380	42	61	0
65+	Manager	z	z	0	0
	Clerk	z	0	0	0
	Worker	6	0	0	0

Table 7. Comparison of Marginal Probabilities
Available Case, LR model, Model with
Convex Constraint on 4th variable EX

Variable/Value	AVC	LR	Convex
AGE: 0	0.00685	0.00691	0.00689
AGE: 1	0.02145	0.02168	0.02162
AGE: 2	0.96623	0.96576	0.96586
AGE: 3	0.00548	0.00566	0.00564
PRO: 0	0.16522	0.15960	0.16023
PRO: 1	0.35098	0.34559	0.34614
PRO: 2	0.48380	0.49482	0.49364
EDU: 0	0.37106	0.38297	0.38175
EDU: 1	0.09950	0.09991	0.09987
EDU: 2	0.38932	0.38517	0.38573
EDU: 3	0.14012	0.13195	0.13265
EX: 0	0.82702	0.83158	0.82800
EX: 1	0.17298	0.16842	0.17200

Independent rounding to 5 decimal places.

We fit using the all 3-way interaction model with or without convex constraints to obtain tables similar to Table 5 (that we do not show). In examining the margins given in Table 7, we see that the margin associated with EX=1 under the basic model (without convex constraints) has probability 0.16842 that is lower than the available case probability of 0.17298 that we are assuming is the truth (because of the manner in which we created the data). If we put a convex constraint of $P(EX=0) \leq 0.82800$ and repeat the fitting, we get the marginal estimate P(EX=1)=0.17200 that is closer to 0.17298 than the estimate of 0.16842 that we obtained without the convex constraint. We further observe that the marginal probabilities associated with the first three variables change very little. The marginal probabilities in Table 7 do not necessarily add to 1.0 due to independent rounding. The underlying individual cell probabilities always add to 1.0 (to at least eight decimal places).

4. Discussion

Although we cannot typically use the convex constraints to overcome non-ignorable nonresponse when the deviations from the missing-at-random assumption are substantial, the convex constraints are useful when the deviations are not severe. As we were able to repeat the types of results of Table 7 with a number of other data sets, we believe the methods that use convex constraints are a useful tool in creating models for imputation that better correspond to data from a variety of sources.

Prior to using the convex constraints, individuals should attempt to determine whether originally observed data (corresponding to the data of Table 6) is a representative subset of a population where the marginal estimates from external sources might be assumed to be valid.

Subject-matter experts will typically have much more confidence in the microdata produced by an edit/imputation system (Fellegi-Holt based or otherwise) if certain marginal estimates from the microdata do not deviate too severely from corresponding estimates from known external sources. This is particularly true if marginal estimates in the microdata correspond to certain estimates in certain income categories available from other sources.

The present version of the software is suitably fast for iterative fitting under linear constraints on tables having 50 million cells. The general software for fitting under a combination of linear and convex constraints of this paper is equally fast. The current version of the software for fitting under a combination of linear and general convex constraints (Winkler 2007) is one tenth as fast.

The methods of this paper and associated software (Winkler 2008b) provide very fast, general methods for cleaning up data in preparation for data mining (Winkler 2008a).

5. Concluding Remarks

This paper provides a method of modeling in the presence of structural zeros (known as edit constraints). With the final models, it is straightforward to impute for missing data by sampling probability-proportion-to-size in the limiting model among model rows that agree on the non-missing values of the records. Unlike hot-deck imputation, joint distributions are preserved, edit constraints are satisfied automatically, and imputation-variance estimation is straightforward.

1/ This report is released to inform interested parties of (ongoing) research and to encourage discussion (of work in progress). Any views expressed on (statistical, methodological, technical, or operational) issues are those of the author(s) and not necessarily those of the U.S. Census Bureau. The author thanks Dr. Maria Garcia for a number of helpful comments that help improve the exposition.

References

Agresti, A. (2007), *An Introduction to Categorical Data Analysis (2nd Edition)*, New York: J. Wiley.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W., (1975), *Discrete Multivariate Analysis*, Cambridge, MA: MIT Press.

Boskovitz, A. (2008), "Data Editing and Logic: The covering set method from the perspective of logic," CS Ph.D. dissertation, Australia National University, http://thesis.anu.edu.au/public/adt-ANU20080314.163155/index.html .

Csiszár, I. (1975), "I-divergence geometry of probability distributions and minimization problems, *Ann. Prob*. 3, 146-159.

D'Orazio, M., Di Zio, M., and Scanu, M. (2006), "Statistical Matching for Categorical Data: Displaying Uncertainty and Using Logical Constraints," *Journal of Official Statistics*, 22 (1), 137-157.

Di Zio, M., Scanu, M., Coppola, L., Luzi, O., and Ponti, A. (2004), "Bayesian Networks for Imputation," *Journal of the Royal Statistical Society, A*, 167 (2), 309-322.

Dykstra, R. L., and Lemke, J. H. (1988), "Duality of I-Projections and Maximum Likelihood Estimates under Cone Constraints," *Journal of the American Statistical Association*, 83 (402), 446-454.

El Barmi, H., and Dykstra, R. (1998), "Maximum Likelihood Estimates via Duality for Log-convex Models when Cell Probabilities are Subject to Convex Constraints," *Annals of Statistics*, 26 (5), 1878-1893.

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

Garfinkel, R. S., Kunnathur, A. S., and Liepins, G. E., (1986), "Optimal Imputation of Erroneous Data: Categorical Data, General Edits," *Operations Research*, 34, 744-751.

Herzog, T. N., Scheuren, F., and Winkler, W.E., (2007), *Data Quality and Record Linkage Techniques*, New York, N. Y.: Springer.

Kovar, J. G., and Winkler, W. E. (1996), "Editing Economic Data," *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 81-87 (also available as Statistical Research Division Report rr00/04 at http://www.census.gov/srd/www/byyear.html).

Little, R. J. A. and Rubin, D. B. (2002), *Statistic Analysis with Missing Data (2nd Edition)*, John Wiley: New York, N.Y.

Meng, X.-L., and Rubin, D. B. (1993), "Maximum Likelihood via the ECM Algorithm: A General Framework," *Biometrika*, 80, 267-78.

Rao, J. N. K. (1997), "Developments in Sample Survey Theory: An Appraisal," *The Canadian Journal of Statistics, La Revue Canadienne de Statistique*, 25 (1), 1-21.

Rubin, D. B. (1983), "Iteratively Reweighted Least Squares," *Encyclopedia of Statistical Sciences*, 4, 272-275.

Winkler, W. E. (1990), "On Dykstra's Iterative Fitting Procedure," *Annals of Probability*, 18, 1410-1415.

Winkler, W. E. (1993), "Improved Decision Rules in the Fellegi-Sunter Model of Record Linkage," P*roceedings of the Section on Survey Research Methods*, *American Statistical Association*, 274-279 (also http://www.census.gov/srd/papers/pdf/rr93-12.pdf).

Winkler, W.E. (1997), "Set-Covering and Editing Discrete Data," *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 564-569, also http://www.census.gov/srd/papers/pdf/rr9801.pdf .

Winkler, W. E. and Chen, B.-C. (2002), "Extending the Fellegi-Holt Model of Statistical Data Editing," (also http://www.census.gov/srd/papers/pdf/rrs2002-02.pdf).

Winkler, W. E. (2003), "A Contingency Table Model for Imputing Data Satisfying Analytic Constraints," *American Statistical Association, Proc. Survey Research Methods Section*, CD-ROM, also http://www.census.gov/srd/papers/pdf/rrs2003-07.pdf .

Winkler, W. E. (2006), "Statistical Matching Software for Discrete Data," computer software and documentation.

Winkler, W.E. (2007), "Analytically Valid Discrete Microdata and Re-identification," available at http://www.census.gov/srd/papers/pdf/rrs2007-19.pdf .

Winkler, W. E. (2008a), "Data Quality in Data Warehouses," in (J. Wang, ed.) *Encyclopedia of Data Warehousing and Data Mining (2nd Edition)*, in press.

Winkler, W. E. (2008b), "Generalized Modeling/Edit/Imputation Software for Discrete Data," computer software and documentation.

Winkler, W. E., and Petkunas, T. (1997), "The DISCRETE Edit System," in (J. Kovar and L. Granquist, eds.) *Statistical Data Editing, Volume II*, U.N. Economic Commission for Europe, 56-62 (also http://www.census.gov/srd/papers/pdf/rr96-3.pdf).

Appendix. Fitting under general convex constraints

We emphasize that the results of this appendix are only valid for probabilities on a finite set of points **X** having size *n*. For convenience, we assume that the underlying distribution is multinomial (or Dirichlet-multinomial where a Dirichlet prior so that sampling-zero cells can have probability greater than zero.

A.1 Basic I-Projections

Let $J \subset I$ be a subset of the set of indices associated with the set of probabilities P. Let $b > 0$ be a positive constant. Let $r \in P$ be an arbitrary probability. We note that r can be obtained by taking any finite measure on the index set I and renormalizing it to be a probability. Then an I-Projection of the form (El Barmi and Dykstra 1998)

$$\inf_{p \in P,\ \Sigma_{j \in J} p_j \le b} \Sigma_{i \in I} p_i \ln(p_i / r_i) \qquad \text{(A.1)}$$

is given by

$$\begin{cases} r_i & \text{if } \Sigma_{j \in J} r_j \le b, & \text{(A.2a)} \\ r_i \, b / \Sigma_{j \in J} r_j & \text{if } i \in J \text{ and } \Sigma_{j \in J} r_j > b, & \text{(A.2b)} \\ r_i \, (1-b) / \Sigma_{j \in J^c} r_j & \text{if } i \in J^c \text{ and } \Sigma_{j \in J} r_j > b. & \text{(A.2c)} \end{cases}$$

Equation (A.2a) is the I-projection in the situation when probability r already satisfies the convex constraint. Such I-Projections in the above form or in somewhat more general form ($\Sigma_{j \in J} c_j p_j \le b$ for positive constants c_j that add to 1) have been used in record linkage applications in Winkler (1993). Although such I-Projections will not generally increase likelihood, they do in the M-step of the EM algorithm of this paper (see next appendix). Alternative characterizations based on different types of duality allow general computation of I-Projections that increase the likelihood (Dykstra and Lemke 1988, El Barmi and Dykstra 1998).

A.2 A generalized EM algorithm under convex constraints

As with the EM under linear constraints (e.g., Little and Rubin 2002), the EM under convex constraints is intended to replace a (much) more complicated direct computational algorithm with a series of straightforward iterative computational steps. As with other work (Dykstra and Lemke 1988), our constraints need to be both mutually compatible with each other and with the observed data.

To begin, we give additional background and notation that gives insight into the theory and the computational algorithms. Assume that the set of points **X** is from n variables $X_1, \ldots, X_n$ that take $k_1, \ldots, k_n$ values respectively. Assume that the total count of records is N that is divided into N_c, the set of complete records having no missing data, and N_i, the set of incomplete records in which some variables having missing values. Then the product *npat* = $k_1 \ldots k_n$ is the number of patterns of the form $\{X_1=j_1, \ldots, X_l=j_l, \ldots, X_n=j_n\}$. If record r $=\{X_1=j_1, \ldots, X_{m1}= b\ , \ldots X_{m2}=b, X_n=j_n\}$has missing values for variables X_{m1} and X_{m2} , then the count (or probability) associated should be dispersed to $j_{m1} \times j_{m2}$ cells in which the missing values have been filled with either expected values based on the current estimates of the parameters or values based on a particular convex constraint. In this situation, we say that we have d*ispersed* the observed count to the $j_{m1} \times j_{m2}$ virtual cells associated with the missing values. If there are n_z structural zeros within the $j_{m1} \times j_{m2}$ cells, then we only disperse to the $j_{m1} \times j_{m2} - n_z$ cells that are allowed to be have nonzero counts.

We let c_i be the count associated with cell $i \in I$ and p_i (= c_i/N) the corresponding probability. The counts (or equivalently probabilities) can be obtained adding the integer counts from complete data records and the real number counts from the dispersal of the missing data records to the virtual cells.

To facilitate our general EM under convex constraints, we begin by describing the EM for the standard loglinear constraints for the types of discrete-data situations of this paper. Let

$$f(Y|\Theta) = \prod_{i=1}^{N_c} f(y_i|\Theta) \prod_{j}^{N_i} f(y_j|\Theta) =$$

$$\prod_{i=1}^{N_c} p_{\iota(i)} \prod_{n(j)}^{N_i} p_{m(j)} = \prod_{i=1}^{N_c} p_{\iota(i)} \prod_{\iota(j)}^{N_i} \prod_{g=1}^{\lambda(j)} p_{\iota(j,g)}^{z_{jg}} = \qquad \text{(A.3)}$$

$$\prod_{i=1}^{N_c} f(y_i|\Theta) \prod_{j=1}^{N_i} \prod_{g=1}^{\lambda(j)} f(y_j|\Theta) f(z_{jg}|y_j,\Theta)$$

represent the likelihood for the N_c complete data records and the N_i incomplete data records. Here $p_{t(i)}$ is the specific cell (pattern) associated with the i^{th} complete data record, $p_{m(j)}$ is the marginal probability associated with the j^{th} record that has missing data, $p_{s(j,g)}$ is the probability of one of the $\lambda(j)$ virtual records associated with the observed j^{th} incomplete record, and z_{ig} is an unobserved indicator of whether the j^{th} incomplete record is from the g^{th} virtual cell associated with it (i.e., z_{ig} = 1 is the j^{th} record is from associated virtual cell s(j, g); 0 otherwise). The third term in equation (A.3) is the likelihood from the observed data and the fourth term is the likelihood that includes the missing data mechanism given by the z-terms. We observe that the third term is typically difficult to maximize directly. If we knew the unknown z_{ig}, then the fourth term is often straightforward to maximize.

We observe that the p-probabilities (p_i, $i \le npat$) are the parameters that we wish to maximize. Our EM-procedure takes an initial guess ($p_i^{(0)}$, $i \le npat$) with t=0 and updates successively to ($p_i^{(t)}$, $i \le npat$) t=1, 2, …. We take natural logarithms of both sides of equation (A.3) and move terms across the equality sign to obtain

$$l(\Theta|Y_{obs}) = l(\Theta|Y) - \ln f(Y_{mis}|Y_{obs},\Theta). \qquad \text{(A.4)}$$

Here $f(Y_{mis}|Y_{obs},\Theta)$ is the density of z_{ig} given particular observed y_j and a current estimate of Θ. If we take expectations of the right hand side of (A.4) with respect to the current estimate $f(Y_{mis}|Y_{obs},\Theta^{(t)})$ where $\Theta^{(t)}$ is the current estimate of Θ, we obtain

$$l(\Theta|Y_{obs}) = Q(\Theta|\Theta^{(t)}) - H(\Theta|\Theta^{(t)}) \qquad \text{(A.5)}$$

where

$$Q(\Theta|\Theta^{(t)}) = \sum_{i=1}^{N} y_i \ln p_{s(i)} + \sum_{j=1}^{N}\sum_{g=1}^{\lambda(j)} E(z_{jg}|Y_{obs},\Theta^{(t)}) \ln p_{s(i,j)}. \qquad \text{(A.6)}$$

Here $E(z_{jg}|Y_{obs},\Theta^{(t)})$ is the expected value of z_{jg} with respect to the current estimate of the multinomial parameter $\Theta^{(t)} = (p_c^{(t)})$ where $p_c^{(t)}$ is the estimate for cell $c \in I$. In equation (A.6), Θ refers to the p-probabilities. It is straightforward to choose a value of Θ that maximizes or increases the complete data likelihood in equation (A.6) using standard loglinear methods (Bishop et al. 1975, Agresti 2007). Each

$$E(z_{jg}|Y_{obs},\Theta^{(t)}) = p_{s(j,g)}^{(t)} / \sum_{g=1}^{\lambda(j)} p_{s(j,g)}^{(t)} \text{ for } j=1,\ldots,N_i,\ g=1,\ \ldots,\ \lambda(j).$$

Each

$$H(\Theta|\Theta^{(t)}) = \sum_{mis} \ln f(Y_{mis}|Y_{obs},\Theta^{(t)})\, f(Y_{mis}|Y_{obs},\Theta^{(t)}) = \qquad \text{(A.7)}$$
$$\sum_{c\in I} p_c \ln(p^{(ty)}) \le \sum_{c\in I} p_c^{(t)} \ln p_c^{(t)}$$

with the inequality a particular case of Jensen's inequality. (p_i), $i \le npat$, is an arbitrary probability so that the summations in (A.7) are well defined.

To better describe the generalizations from linear constraints to convex constraints, we slightly digress to the (Multi-Cycle) Expectation Conditional Maximization (MCECM) of Meng and Rubin (1993). Meng and Rubin observed, under the linear constraints of their paper, that one could first fill-in missing z_{jg} values and

then go through one cycle of iterative proportional fitting (IPF) or one could repeatedly fill-in expected values and do individual (conditional) maximizations according to a single marginal restraint of IPF. Further, we can observe that the expectation in equation (A.5) can be performed with respect to any probability measure $p^{t,i}$ (current estimate) that is particular to the observed data, the underlying probability model, and a particular marginal constraint. Here $p^{t,i}$ is the probability that we use to fill-in expected values prior to performing a maximization (or computation that only increases likelihood) that is consistent with a given constraint.

With standard MCECM under linear constraints, we easily (and implicitly) get the expected values in the correct form. We could, however, choose the probabilities $p^{t,i}$ to yield expectations for which a maximization (or parameter value that increases the likelihood) with respect to a given convex constraint is straightforward. We observe that we can choose any probability $p^{t,i}$ that differs from $p^{t,ip}$ and the H-values in equation (A.7) would decrease. Here ip refers to the restraint situation immediately preceding the situation with i. We note that each $p^{t,i}$ is a probability that must be compatible with a particular marginal or convex constraint, the observed data, and the particular model (multinomial or Dirichlet-multinomial). We first fill-in 'expected' values according to $p^{t,i}$ and then perform a step that increases the likelihood for the 'completed' likelihood. If $p^{t,i}$ is compatible with the appropriate convex constraint and expectations are taken with respect to $p^{t,i}$, then the maximization (via an I-Projection) with respect to the same convex constraint will increase the likelihood.

To be more specific, we give more details of the specific MCECM procedure under convex constraints. First, initially assume that we are able to find a suitable probability measure r at any given stage (specific iteration step within a cycle) that will yield observed plus expected data that satisfies the desired convex constraint. It is possible to monitor the computation to determine whether this is possible at any given stage. If we are able to do this, then were obtain an equation of the form

$$\sum_{i \le npat} q_i \ln(p_i) \tag{A.8}$$

where (p_i), $i \le npat$, is the probability that we wish to maximize with respect to the data (q_i), $i \le npat$, and the data (q_i), $i \le npat$, satisfy the particular convex constraint at the particular stage of the computation. In comparison to equation (A.6), each

$$q_i = 1_{\{i\}} y_i + \sum_{j \in V(i)} E_r(z_{jh} | Y_{obs}, \Theta^{(t)}) \tag{A.9}$$

where V(i) is the set of indices of virtual cells that are associated with observed cell i, 1_i is an indicator for the observed data y_i because there may be no observed data for cell i in some situations, and z_{jh} is the indicator from the appropriate virtual cell. The collected of appropriate terms in (A.8) is straightforward as it is with usual EM procedure (Little and Rubin 2002, section 13.4). If we use (A.2) to obtain the I-projection (s_i), $i \le npat$, then by Csiszár's inequality we have

$$I(q||p) \ge I(q||s) + I(s||p) \tag{A.10}$$

where $I(q||p) = \sum_i q_i \ln(q_i / p_i)$ and (q_i), $i \le npat$, could be any probability measure satisfying the convex constraint. Because the Kullback-Leibler Information number I(||) is positive, if we drop the I(s || p) and combine terms from (A.10) we obtain that the likelihood increases

$$\sum_i q_i \ln(s_i) \ge \sum_i q_i \ln(p_i) \cdot \tag{A.11}$$

We must be able to choose a probability measure so that (q_i), $i \le npat$, is obtained via (A.9) satisfies the desired convex constraint. It is straightforward to put in computer code that monitors that each (q_i), $i \le npat$, satisfies the convex constraint.

The above algorithm is essentially the same as the EMH algorithm (Winkler 1993) that was used by D'Orazio et al. (2006). In a particular sense, the EMH algorithm generalizes Dykstra and Lemke (1988) and El Barmi and Dykstra (1998) were proved using Fenchel Duality and a generalization that does not require a specific type of duality, respectively. The particular sense is in some situations where we can artificially induce nonresponse (say with a latent variable) and we have convex constraints that are consistent with the observed data and the model assumptions. The authors of the alternate algorithms in the no-missing data situations needed to assume that the space of possible solutions contained the constant probabilities that we do not need to assume in this paper. In some situations, we can also change the no-missing-data situation to the missing-data situation using the methods of Rubin (1983). We note that the non-missing-data situations typically require very specific analytic work and computations and will likely be much faster than the methods of this paper. The methods of this paper, while possibly much slower, allow identical sets of computation (i.e., one set of generalized software) for a very wide range of real data situations.