Methods for Imputing Missing Values and Synthesizing Confidential Values for Continuous and Magnitude Data

by

Lan Wei

Department of Statistical Science
Duke University

Date: ______________________

Approved:

___________________________
Jerome P. Reiter, Supervisor

___________________________
Surya Tapas Tokdar

___________________________
Mine Çetinkaya-Rundel

___________________________
Ashwin Machanavajjhala

Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistical Science in the Graduate School of Duke University

2016
Abstract

Methods for Imputing Missing Values and Synthesizing Confidential Values for Continuous and Magnitude Data

by

Lan Wei

Department of Statistical Science
Duke University

Date: ____________________
Approved:

________________________
Jerome P. Reiter, Supervisor

________________________
Surya Tapas Tokdar

________________________
Mine Çetinkaya-Rundel

________________________
Ashwin Machanavajjhala

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistical Science in the Graduate School of Duke University 2016
Abstract

Continuous variable is one of the major data types collected by the survey organizations. It can be incomplete such that the data collectors need to fill in the missingness. Or, it can contain sensitive information which needs to be aggregated in order not to be re-identified. In this thesis, I represent novel methods of multiple imputation (MI) that can be applied to impute missing values and synthesize confidential values for continuous and magnitude data.

The first method is for limiting the disclosure risk of the continuous microdata whose marginal sums are fixed. The motivation comes from the magnitude tables of non-negative integer values in economic surveys. I present approaches based on a mixture of Poisson distributions to describe the multivariate distribution so that the marginals of the synthetic data are guaranteed to sum to the original totals. At the same time, I present methods for assessing disclosure risks in releasing such synthetic magnitude microdata. The illustration on a survey of manufacturing establishments shows that the disclosure risks are low while the information loss is acceptable.

The second method is for releasing synthetic continuous microdata by interval-protected MI methods. Typically, MI fits a synthesis model directly on the confidential values and then generates multiple synthetic datasets from the model. Thus, its disclosure risk can be high especially when the original data contain extreme values. From a new perspective, I present MI approaches conditioned on the protective intervals. The basic idea is to estimate the parameters of the synthesis model
from these intervals and/or restrict the synthetic values from truncated distributions. The results of simple simulation studies are encouraging, which suggests the potential of interval-protected MI in limiting the posterior disclosure risk for continuous microdata.

The third method is for imputing missing values in continuous and categorical variables. It is extended from a hierarchically coupled mixture model with local dependence. However, the new method separates the variables into non-focused (e.g., almost-fully-observed) and focused (e.g., missing-a-lot) ones. The sub-model structure of focused variables is more complex than that of non-focused ones with their cluster indicators being linked together by a tensor factorization. In addition, the focused continuous variables depend locally on non-focused values. This model property suggests that moving the strongly associated non-focused variables to the side of focused ones can help to improve estimation accuracy, which is examined by several simulation studies. And this method is applied to data from the American Community Survey.
Contents

Abstract iv
List of Tables x
List of Figures xii
Acknowledgements xxiii

1 Introduction 1
  1.1 Multiple Imputation for Missing Data 3
    1.1.1 Missing Data Mechanisms 4
    1.1.2 Multiple Imputation 6
  1.2 Multiple Imputation for Disclosure Control 7
  1.3 Dirichlet Process Mixture Modelling 8
    1.3.1 Finite Mixture Modelling 8
    1.3.2 Dirichlet Process Mixtures 9
    1.3.3 Dirichlet Process Mixtures of Product Multinomial Distributions 10
    1.3.4 Hierarchically Coupled Mixture Model with Local Dependency 11

2 Releasing Synthetic Magnitude Microdata Constrained to Fixed Marginal Totals 14
  2.1 Introduction 14
  2.2 Mixtures of Poisson Distributions 15
  2.3 Synthesizers Based on Mixtures of Poissons 17
    2.3.1 The Mixture of Poisson Synthesizer 17

vi
List of Tables

2.1 Synthetic and original values of the two largest values of skw for each synthesizer. Synthetic data results are averages and standard deviations across the $L$ datasets. .............................................. 31

2.2 Posterior probability that the largest value of skl lies in the target range of (378, 496) for different synthesizers under Scenario III. ........... 34

3.1 Methods of releasing the synthetic data $Y$ for the true data $X$. Here, $D = \{d_i = (a_i, b_i) : i = 1, \ldots, n\}$ is the collection of protection intervals for each observation $X_i = x_i$. And $W = \{w_i : i = 1, \ldots, n\}$ is the collection of indicators such that $w_i = 1$ if $a_i \leq x_i \leq b_i$ and $w_i = 0$ otherwise. .................................................. 43

3.2 The posterior disclosure risk metrics of learning $X_{[1]} = 6$ for $g = 0.1$. .... 55

3.3 The posterior disclosure risk metrics of knowing $X_{[1]} = 6$ for $g = 0.5$. .... 55

3.4 Methods of releasing the synthetic data $Y = (Y_1, Y_2)$ for the true data $X = (X_1, X_2)$. Here, $D_1 = \{d_{1,i} = (a_{1,i}, b_{1,i}) : i = 1, \ldots, n\}$ is the collection of protection intervals for each observation $X_{1,i} = x_{1,i}$. And $W_1 = \{w_{1,i} : i = 1, \ldots, n\}$ is the collection of indicators such that $w_{1,i} = 1$ if $a_{1,i} \leq x_{1,i} \leq b_{1,i}$ and $w_{1,i} = 0$ otherwise. Similarly, $D_2 = \{d_{2,i} = (a_{2,i}, b_{2,i}) : i = 1, \ldots, n\}$ is the collection of protection intervals for each observation $X_{2,i} = x_{2,i}$. And $W_2 = \{w_{2,i} : i = 1, \ldots, n\}$ is the collection of indicators such that $w_{2,i} = 1$ if $a_{2,i} \leq x_{2,i} \leq b_{2,i}$ and $w_{2,i} = 0$ otherwise. .............................................. 56

3.5 The means of the estimated regression coefficients from the synthetic data for each method compared to the coefficient on the real data (i.e., $\beta = 1.1$) together with the standard errors. ................................. 62

3.6 The posterior disclosure risk metrics of knowing $X_{[2]} = 6.7$ for $g_1 = g_2 = 0.1$. ................................................................. 63

3.7 The posterior risk metrics of knowing $X_{2,[1]} = 6.7$ for $g_1 = g_2 = 0.5$. 64
3.8 The posterior disclosure risk metrics of knowing $X_{[1]} = 6$ when $(g^{(a)} = 5, g^{(b)} = 25)$. ................................................................. 66

3.9 The posterior disclosure risk metrics of knowing $X_{[1]} = 6$ when $(g^{(a)} = 15, g^{(b)} = 25)$. ................................................................. 66

3.10 The means of the estimated regression coefficients from the synthetic data for each method compared to the coefficient on the real data (i.e., $\beta = 1.1$) together with the standard errors. ................................. 67

3.11 The posterior risk metrics of knowing $X_{2,[1]} = 6.7$ for $g_1^{(a)} = g_2^{(a)} = 5$ and $g_1^{(b)} = g_2^{(b)} = 25$. ................................................................. 68

3.12 The posterior risk metrics of knowing $X_{2,[1]} = 6.7$ for $g_1^{(a)} = g_2^{(a)} = 15$ and $g_1^{(b)} = g_2^{(b)} = 25$. ................................................................. 69

4.1 High-level description of scenarios and quantities of interest in the simulation studies. In scenario "$X^{(F)}$ and $X^{(NF)}$", we are interested in the accuracies of the bivariate cell frequencies between $X^{(F)}$ and the associated $X^{(NF)}$. In scenario "$X^{(F)}$ and $Y^{(NF)}$", we are interested in the accuracies of the dependencies between $X^{(F)}$ and the associated $Y^{(NF)}$. In scenario "$X^{(F)}$ and $X^{(NF)}$", we are interested in the accuracies of the dependencies between $Y^{(F)}$ and the associated $X^{(NF)}$. In scenario "$Y^{(F)}$ and $Y^{(NF)}$", we are interested in the accuracies of the correlation coefficients between $Y^{(F)}$ and the associated $Y^{(NF)}$. .................................................. 98

4.2 The variables selected from ACS in the year of 2010. The first four variables are for each household. The other variables are for individual household head. Cts is short for continuous variable and Cat is short for categorical variables. For a categorical variable, # Levels is its number of factorial levels. PROPTX99 is a categorical variable with large number of levels. It is treated as a continuous variable only when we represent the related results. ................................................................. 108

4.3 The relative mutual information shared among focus and non-focus variables in the study of HMI. INCTOT and OCCSCORE are the top two $N^{\mathcal{F}}$ variables which shares a relatively equal and high amount of mutual information with EDUC and PROPTX99. ................................................................. 109

4.4 The relative mutual information shared among focus and non-focus variables in the study of LMI. PROPTX99 is the non-focus variable which shares a relatively high amount of mutual information with all the focus ones. ................................................................. 124
List of Figures

2.1 Scatter plots of the observed values in the Columbia manufacturing survey data. .......................................................... 30

2.2 Boxplots of $skl, unsk, skw, unskw$. In each plot, the columns are arranged from the left to the right in the order of original data, the MP, the MM, the TCMM with $q = .997, .995, .993, \text{ and } .991$. ....... 36

2.3 Density plots of the synthetic pair-wise correlations for different synthesizers. Each vertical line is the value of the correlation from the original data. ................................................................. 37

2.4 Density plots of the synthetic estimates of $(\beta_1, \ldots, \beta_6)$ for different synthesizers. Each vertical line is the value of the coefficient from the original data. ................................................................. 38

2.5 The prior and the disclosure risk of the top largest $skl$ under the MP synthesizer for the strong assumption. ......................... 39

2.6 The prior and the disclosure risk of the top largest $skl$ under the MP synthesizer for the strong assumption. ......................... 39

2.7 The prior and the posterior risk for the record with the largest value of $skl$ for the TCMM synthesizer under Scenario I. ............. 40

3.1 Results for univariate additive noise simulations. The histograms of the true data $X$ versus 10 sets of synthesis $Y$. ......................... 53

3.2 The posterior predictive densities of guessing $X_{11} = 6$ for $g = 0.1$. When $D$ is known, the plots of Method 3 and 4 are deleted because their posterior disclosure risks are uniformly distributed according to (3.3) and (3.4). ................................................................. 71
3.3 The posterior predictive densities of guessing $X_{[1]} = 6$ for $g = 0.5$. When $D$ is known, the plots of Method 3 and 4 are deleted because their posterior disclosure risks are uniformly distributed according to (3.3) and (3.4). ................................................................. 72

3.4 Results for bivariate additive noise simulations. The histograms of the true data $X_2$ versus 10 sets of synthesis $Y_2$ when $g_1 = g_2 = 0.1$. ........... 73

3.5 Results for bivariate additive noise simulations. The histograms of the true data $X_2$ versus 10 sets of synthesis $Y_2$ when $g_1 = g_2 = 0.5$. .......... 74

3.6 The posterior predictive densities of guessing $X_{2,[1]} = 6.3$ for $g_1 = g_2 = 0.1$ when $D_1$ is known in Method 2 and $(D_1, D_2)$ are known in Method 3. When $(D_1, D_2)$ are known in Method 4, 5, and 6, the plots are deleted because their posterior disclosure risks are uniformly distributed according to Section 3.3.1. ................................. 75

3.7 The posterior predictive densities of guessing $X_{2,[1]} = 6.7$ for $g_1 = g_2 = 0.1$ when $(D_1, D_2)$ are unknown. ............................. 76

3.8 The posterior predictive densities of guessing $X_{2,[1]} = 6.7$ for $g_1 = g_2 = 0.5$ when $D_1$ is known in Method 2 and $(D_1, D_2)$ are known in Method 3. When $(D_1, D_2)$ are known in Method 4, 5, and 6, the plots are deleted because their posterior disclosure risks are uniformly distributed according to Section 3.3.1. ................................. 77

3.9 The posterior predictive densities of guessing $X_{2,[1]} = 6.7$ for $g_1 = g_2 = 0.1$ when $(D_1, D_2)$ are unknown. ............................. 78

3.10 Results for univariate multiplicative noise simulations. The histograms of the true data $X$ versus 10 sets of synthesis $Y$. ....................... 79

3.11 The posterior predictive densities of guessing $X_{[1]} = 6$ for $(g^{(a)} = 5, g^{(b)} = 25)$. When $D$ is known, the plots of Method 3 and 4 are deleted because the posterior disclosure risks are uniformly distributed according to (3.3) and (3.4). ................................. 80

3.12 The posterior predictive densities of guessing $X_{[1]} = 6$ for $(g^{(a)} = 15, g^{(b)} = 25)$. When $D$ is known, the plots of Method 3 and 4 are deleted because their posterior disclosure risks are uniformly distributed according to (3.3) and (3.4). ................................. 81

3.13 Results for bivariate multiplicative noise simulations. The histograms of the true data $X_2$ versus 10 sets of synthesis $Y_2$ when $g_1^{(a)} = g_2^{(a)} = 5$ and $g_1^{(b)} = g_2^{(b)} = 25$. ................................. 82
3.14 Results for bivariate multiplicative noise simulations. The histograms of the true data $X_2$ versus 10 sets of synthesis $Y_2$ when $g_1^{(a)} = g_2^{(a)} = 15$ and $g_1^{(b)} = g_2^{(b)} = 25$. 

3.15 The posterior predictive densities of guessing $X_{2,1} = 6.7$ for $g_1^{(a)} = g_2^{(a)} = 5$ and $g_1^{(b)} = g_2^{(b)} = 25$ when $D_1$ is known in Method 2 and $(D_1, D_2)$ are known in Method 3. When $(D_1, D_2)$ are known in Method 4, 5, and 6, the plots are deleted because their posterior disclosure risks are uniformly distributed according to Section 3.3.1.

3.16 The posterior predictive densities of guessing $X_{2,1} = 6.7$ for $g_1^{(a)} = g_2^{(a)} = 5$ and $g_1^{(b)} = g_2^{(b)} = 25$ when $(D_1, D_2)$ are unknown.

3.17 The posterior predictive densities of guessing $X_{2,1} = 6.7$ for $g_1^{(a)} = g_2^{(a)} = 15$ and $g_1^{(b)} = g_2^{(b)} = 25$ when $D_1$ is known in Method 2 and $(D_1, D_2)$ are known in Method 3. When $(D_1, D_2)$ are known in Method 4, 5, and 6, the plots are deleted because their posterior disclosure risks are uniformly distributed according to Section 3.3.1.

3.18 The posterior predictive densities of guessing $X_{2,1} = 6.7$ for $g_1^{(a)} = g_2^{(a)} = 15$ and $g_1^{(b)} = g_2^{(b)} = 25$ when $(D_1, D_2)$ are unknown.

4.1 Graphical model representation of HCMM-FNF. $X^{(F)}$, $Y^{(F)}$, $X^{(NF)}$, and $Y^{(NF)}$ are the observed categorical and continuous variables. $H^{(F)}$ and $H^{(NF)}$ are the mixture components of $F$ and $NF$ variables, respectively. $Z$ is the mixture component for the whole record.

4.2 Scenario “$X^{(F)}$ and $X^{(NF)}$”. All four focus categorical variables $X_{1:4}^{(F)}$ depend on the first four non-focus categorical variables $X_{1:4}^{(NF)}$ by (4.14). Under the strategy of Move, we move $X_{1:4}^{(NF)}$ to the side of $X_{1:4}^{(F)}$.

4.3 Scenario “$X^{(F)}$ and $Y^{(NF)}$”. All four focus categorical variables $X_{1:4}^{(F)}$ depend on the first four non-focus continuous variables $Y_{1:4}^{(NF)}$ by (4.15). Under the strategy of Move, we bring $Y_{1:4}^{(NF)}$ to the side of $Y^{(F)}$.

4.4 Scenario “$Y^{(F)}$ and $X^{(NF)}$”. All four focus continuous variables $Y_{1:4}^{(F)}$ depend on the first four non-focus categorical variables $X_{1:4}^{(NF)}$ by (4.16), (4.17) or (4.18). Under the strategy of Move, we move $X_{1:4}^{(NF)}$ to the side of $X^{(F)}$. 

xiv
4.5 Scenario “\(Y^{(F)}\) and \(Y^{(NF)}\)”. All four focus continuous variables \(Y_{1:4}^{(F)}\) depend on the first four non-focus continuous variables \(Y_{1:4}^{(NF)}\) by (4.19), (4.20), or (4.21). Under the strategy of \(Move\), we move \(Y_{1:4}^{(NF)}\) to the side of \(Y_{1:4}^{(F)}\) .......................... 101

4.6 The comparison of the bivariate cell frequencies between \(Stay\) and \(Move\) in “\(X^{(F)}\) and \(X^{(NF)}\)”. The left plot is from Metric I. A triangle symbol is a difference when \(Move\) versus \(Stay\) has a p-value below 0.01. The right one is from Metric II, where the median is 0 for \(Stay\) and 0 for \(Move\). The bottom plot zooms in the right one by focusing on the percentage change between \(-5\) and \(5\), where the median is still 0 for \(Stay\) and 0 for \(Move\) .......................... 112

4.7 The comparison of the logistic regression coefficients between \(Stay\) and \(Move\) in “\(X^{(F)}\) and \(Y^{(NF)}\)” with the associations described by (4.16). The left plot is from Metric I. A triangle symbol is a difference when \(Move\) versus \(Stay\) has a p-value below 0.01. The right plot is from Metric II, in which the median is \(-44.79\) for \(Stay\) and \(-9.93\) for \(Move\) .......................... 113

4.8 The comparison of the logistic regression coefficients between \(Stay\) and \(Move\) in “\(Y^{(F)}\) and \(X^{(NF)}\)” with the associations described by (4.17). The left plot is from Metric I. The right one is from Metric II, where the median is \(-0.0884\) for \(Stay\) and \(-0.1048\) for \(Move\) .......................... 114

4.9 The comparison of the logistic regression coefficients between \(Stay\) and \(Move\) in “\(Y^{(F)}\) and \(X^{(NF)}\)” with the associations described by (4.18). The left plot is from Metric I. The right one is from Metric II, where the median is \(-2.9172\) for \(Stay\) and \(-3.101\) for \(Move\) .......................... 114

4.10 The comparison of the correlation coefficients between \(Stay\) and \(Move\) in “\(Y^{(F)}\) and \(X^{(NF)}\)” with the associations described by (4.19). The left plot is from Metric I. The right one is from Metric II, where the median is \(-12.47\) for \(Stay\) and \(-4.26\) for \(Move\) .......................... 115

4.11 The comparison of the correlation coefficients between \(Stay\) and \(Move\) in “\(Y^{(F)}\) and \(Y^{(NF)}\)” with the associations described by (4.20). The left plot is from Metric I. The right one is from Metric II, where the median is \(-36\) for \(Stay\) and \(-30\) for \(Move\) .......................... 115
4.13 The comparison of the correlation coefficients between Stay and Move in \( Y^{(F)} \) and \( Y^{(NF)} \) with the associations described by (4.21). The left plot is from Metric I. The right one is from Metric II, where the median is \(-25\) for Stay and \(-28\) for Move.

4.14 The variable assignment of focus and non-focus variables for ACS 2010 in the state of North Carolina in the study of HMI. INCTOT and OCCSCORE are the variables which will be moved to the side of \( Y^{(F)} \) under the strategy of move.

4.15 The marginal probability distributions of PROPTX99 in the study of HMI. In each plot, the black solid line with gray area is from the truth with its 95% bootstrap confidence interval. The other lines are from the completed datasets after imputation.

4.16 The marginal boxplots of EDUC in the study of HMI. Each panel corresponds to a factor level in EDUC. And each boxplot is a summary of all 10 imputed counts in that factor level from a particular imputation model. Here, H, M, S, T stand for HCMM-LD, Move, Stay, and Truth, respectively.

4.17 The contour plots from the kernel density of log(1+INCTOT) (standardized) and PROTPX99 for the missing observations in the study of HMI. For each imputation model, the contour plots of all its imputed kernel densities are identically and independently drawn and look similar, so that we randomly select one imputation for the purpose of display.

4.18 The contour plots from the kernel density of OCCSCORE (standardized) and PROTPX99 for the missing observations in the study of HMI. For each imputation model, the contour plots of all its imputed kernel densities are identically and independently drawn and look similar, so that we randomly select one imputation for the purpose of display.

4.19 The means of the standardized log(1+INCTOT) by EDUC in the study of HMI. H, M, and S stand for HCMM-LD, HCMM-FNF with Stay, and HCMM-FNF with Move. For each imputation model, the error bars are obtained by applying the combination rule on all its imputations.
4.20 The means of the standardized OCCSCORE by EDUC in the study of HMI. T, H, M, and S stand for True, HCMM-LD, HCMM-FNF with Stay, and HCMM-FNF with Move. For each imputation model, the error bars are obtained by applying the combination rule on all its imputations. .................................................. 121

4.21 The means of PROPTX99 by EDUC in the study of HMI. T, H, M, and S stand for True, HCMM-LD, HCMM-FNF with Stay, and HCMM-FNF with Move. For each imputation model, the error bars are obtained by applying the combination rule on all its imputations. 122

4.22 The variable assignment of focus and non-focus variables for ACS 2010 in the state of North Carolina in the study of LMI. PROPTX99 is the variable which will be moved to the side of $X^{(F)}$ under the strategy of move. .................................................. 123

4.23 The marginal probability distributions of log(1+INCTOT) (standardized) in the study of LMI. In each plot, the black solid line with gray area is from the truth with its 95% bootstrap confidence interval. The other lines are from the completed datasets after imputation. . . . . 125

4.24 The marginal probability distributions of OCCSCORE (standardized) in the study of LMI. In each plot, the black solid line with gray area is from the truth with its 95% bootstrap confidence interval. The other lines are from the completed datasets after imputation. . . . . . 126

4.25 The marginal boxplots of EDUC in the study of LMI. Each panel corresponds to a factor level in EDUC. And each boxplot is a summary of all 10 imputed counts in that factor level from a particular imputation model. Here, H, M, S, T stand for HCMM-LD, Move, Stay, and Truth, respectively. .................................................. 130

4.26 The marginal boxplots of DIFFSENS in the study of LMI. Each panel corresponds to a factor level in DIFFSENS. And each boxplot is a summary of all 10 imputed counts in that factor level from a particular imputation model. Here, H, M, S, T stand for HCMM-LD, Move, Stay, and Truth, respectively. .................................................. 131

4.27 The contour plots of the kernel density for log(1+INCTOT) and OCCSCORE (standardized) for the missing observations in the study of LMI. For each imputation model, the contour plots of all its imputed kernel densities are identically and independently drawn and look similar, so that we randomly select one imputation for the purpose of display. .................................................. 132
4.28 The contour plots of the kernel density for PROPTX99 and the standardized log(1+INCTOT) for the missing observations in the study of *LMI*. For each imputation model, the contour plots of all its imputed kernel densities are identically and independently drawn and look similar, so that we randomly select one imputation for the purpose of display. .................................................. 133

4.29 The contour plots of the kernel density for PROPTX99 and OCCSCORE (standardized) for the missing observations in the study of *LMI*. For each imputation model, the contour plots of all its imputed kernel densities are identically and independently drawn and look similar, so that we randomly select one imputation for the purpose of display. .................................................. 134

4.30 The means of the standardized log(1+INCTOT) by EDUC in the study of *LMI*. Here, H, M, and S stand for *HCMM-LD, Stay,* and *Move*. For each imputation model, the error bars are obtained by applying the combination rule on all its imputations. ................................. 135

4.31 The means of the standardized OCCSCORE by EDUC in the study of *LMI*. T, H, M, and S stand for *True, HCMM-LD, Stay,* and *Move*. For each imputation model, the error bars are obtained by applying the combination rule on all its imputations. ................................. 135

4.32 The means of PROPTX99 by EDUC in the study of *LMI*. T, H, M, and S stand for *True, HCMM-LD, Stay,* and *Move*. For each imputation model, the error bars are obtained by applying the combination rule on all its imputations. ................................. 136

4.33 The means of the standardized log(1+INCTOT) by DIFFSENS in the study of *LMI*. T, H, M, and S stand for *True, HCMM-LD, Stay,* and *Move*. For each imputation model, the error bars are obtained by applying the combination rule on all its imputations. ................................. 136

4.34 The means of the standardized OCCSCORE by DIFFSENS in the study of *LMI*. T, H, M, and S stand for *True, HCMM-LD, Stay,* and *Move*. For each imputation model, the error bars are obtained by applying the combination rule on all its imputations. ................................. 137

4.35 The means of PROPTX99 by DIFFSENS in the study of *LMI*. T, H, M, and S stand for *True, HCMM-LD, Stay,* and *Move*. For each imputation model, the error bars are obtained by applying the combination rule on all its imputations. ................................. 137
B.1 The marginal probability distributions of PROPTX99 in the study of *HMI*. In each plot, the black solid line with gray area is from the truth with its 95% bootstrap confidence interval. The other lines are from the completed datasets after imputation. There is no significant difference between this and Figure 4.15. . . . . . . . . . . . . . . . . . 146

B.2 The marginal boxplots of EDUC in the study of *HMI*. Each panel corresponds to a factor level in EDUC. And each boxplot is a summary of all 10 imputed counts in that factor level from a particular imputation model. Here, H, M, S, T stand for *HCMM-LD*, *Move*, *Stay*, and *Truth*, respectively. The difference between this and Figure 4.16 is small, which does not influence our main conclusion in Section 4.4. 147

B.3 The contour plots from the kernel density of log(1+INCTOT) (standardized) and PROTPX99 for the missing observations in the study of *HMI*. For each imputation model, the contour plots of all its imputed kernel densities are identically and independently drawn and look similar, so that we randomly select one imputation for the purpose of display. The only difference between this and Figure 4.17 seems to be from the model *HCMM-LD*. However, this does not affect our main conclusion in Section 4.4. . . . . . . . . . . . . . . . . . 148

B.4 The contour plots from the kernel density of OCCSCORE (standardized) and PROTPX99 for the missing observations in the study of HMI. For each imputation model, the contour plots of all its imputed kernel densities are identically and independently drawn and look similar, so that we randomly select one imputation for the purpose of display. There is no significant difference between this and Figure 4.18. . . . . 149

B.5 The means of the standardized log(1+INCTOT) by EDUC in the study of *HMI*. T, H, M, and S stand for *True*, *HCMM-LD*, *HCMM-FNF* with *Stay*, and *HCMM-FNF* with *Move*. For each imputation model, the error bars are obtained by applying the combination rule on all its imputations. There is no significant difference between this and Figure 4.19. . . . . . . . . . . . . . . . . . 150

B.6 The means of the standardized OCCSCORE by EDUC in the study of *HMI*. T, H, M, and S stand for *True*, *HCMM-LD*, *HCMM-FNF* with *Stay*, and *HCMM-FNF* with *Move*. For each imputation model, the error bars are obtained by applying the combination rule on all its imputations. The difference between this and Figure 4.20 is small, which does not influence our main conclusion in Section 4.4. . . . . 151
B.7 The means of PROPTX99 by EDUC in the study of HMI. T, H, M, and S stand for True, HCMM-LD, HCMM-FNF with Stay, and HCMM-FNF with Move. For each imputation model, the error bars are obtained by applying the combination rule on all its imputations. The difference between this and Figure 4.21 is small, which does not affect our main conclusion in Section 4.4.

B.8 The marginal probability distributions of log(1+INCTOT) (standardized) in the study of LMI. In each plot, the black solid line with gray area is from the truth with its 95% bootstrap confidence interval. The other lines are from the completed datasets after imputation. There is no significant difference between this and Figure 4.23.

B.9 The marginal probability distributions of OCCSCORE (standardized) in the study of LMI. In each plot, the black solid line with gray area is from the truth with its 95% bootstrap confidence interval. The other lines are from the completed datasets after imputation. There is no significant difference between this and Figure 4.24.

B.10 The marginal boxplots of EDUC in the study of LMI. Each panel corresponds to a factor level in EDUC. And each boxplot is a summary of all 10 imputed counts in that factor level from a particular imputation model. Here, H, M, S, T stand for HCMM-LD, Move, Stay, and Truth, respectively. There is some difference between this and Figure 4.25. However, such a difference does not affect the main conclusion we get in Section 4.4.

B.11 The marginal boxplots of DIFFSENS in the study of LMI. Each panel corresponds to a factor level in DIFFSENS. And each boxplot is a summary of all 10 imputed counts in that factor level from a particular imputation model. Here, H, M, S, T stand for HCMM-LD, Move, Stay, and Truth, respectively. There is no significant difference between this and Figure 4.26.

B.12 The contour plots of the kernel density for log(1+INCTOT) and OCCSCORE (standardized) for the missing observations in the study of LMI. For each imputation model, the contour plots of all its imputed kernel densities are identically and independently drawn and look similar, so that we randomly select one imputation for the purpose of display. There is no big difference between this and Figure 4.27.

xx
B.13 The contour plots of the kernel density for PROPTX99 and the standardized log(1+INCTOT) for the missing observations in the study of LMI. For each imputation model, the contour plots of all its imputed kernel densities are identically and independently drawn and look similar, so that we randomly select one imputation for the purpose of display. There is no significant difference between this and Figure 4.28.

B.14 The contour plots of the kernel density for PROPTX99 and OCCSCORE (standardized) for the missing observations in the study of LMI. For each imputation model, the contour plots of all its imputed kernel densities are identically and independently drawn and look similar, so that we randomly select one imputation for the purpose of display. There is no significant difference between this and Figure 4.29.

B.15 The means of the standardized log(1+INCTOT) by EDUC in the study of LMI. T, H, M, and S stand for True, HCMM-LD, Stay, and Move. For each imputation model, the error bars are obtained by applying the combination rule on all its imputations. There is some difference between this and Figure 4.30. However, such a difference does not influence the main conclusion in Section 4.4.

B.16 The means of the standardized OCCSCORE by EDUC in the study of LMI. T, H, M, and S stand for True, HCMM-LD, Stay, and Move. For each imputation model, the error bars are obtained by applying the combination rule on all its imputations. The difference between this and Figure 4.31 is small, which does not affect the main conclusion in Section 4.4.

B.17 The means of PROPTX99 by EDUC in the study of LMI. T, H, M, and S stand for True, HCMM-LD, Stay, and Move. For each imputation model, the error bars are obtained by applying the combination rule on all its imputations. There is some difference between this and Figure 4.32. However, such a difference does not influence the main conclusion we get in Section 4.4.

B.18 The means of the standardized log(1+INCTOT by DIFFSENS in the study of LMI. T, H, M, and S stand for True, HCMM-LD, Stay, and Move. For each imputation model, the error bars are obtained by applying the combination rule on all its imputations. There is some difference between this and Figure 4.33. However, such a difference does not affect the main conclusion we get in Section 4.4.
B.19 The means of the standardized OCCSCORE by DIFFSENS in the study of LMI. T, H, M, and S stand for True, HCMM-LD, Stay, and Move. For each imputation model, the error bars are obtained by applying the combination rule on all its imputations. There is some difference between this and Figure 4.34. However, such a difference does not influence the main conclusion we get in Section 4.4.

B.20 The means of PROPTX99 by DIFFSENS in the study of LMI. T, H, M, and S stand for True, HCMM-LD, Stay, and Move. For each imputation model, the error bars are obtained by applying the combination rule on all its imputations. There is some difference between this and Figure 4.35. However, such a difference does not affect the main conclusion we get in Section 4.4.
Acknowledgements

First, I would like to express my sincere gratitude towards my adviser, Professor Jerome P. Reiter. During my entire PhD study, he always provides me with invaluable mentoring which I am greatly thankful. Without his great patience in guidance, this thesis would not have been possible.

At the same time, I would like to thank all the faculty members in the department. Their passion in research and excellence in teaching helps me to increase my knowledge in Bayesian statistics. In particular, I would like to thank Professor Mike West for his encouragement during my PhD study. I would like to thank Professor Mine Çetinkaya-Rundel, Professor Ashwin Machanavajjhala, and Professor Surya Tapas Tokdar. I am very grateful for their precious time in serving on my PhD thesis committees.

I would also like to thank my colleagues in Professor Reiter’s group. Especially, I would like to thank Jared Murray for sharing me his research work. I would like to thank Maria De Yoreo for the discussion and input. And I would like to thank Quanli Wang for giving me advice on my coding. In addition, I would like to thank all my friends at Duke for making my PhD study more enjoyable.

My PhD work was supported by the National Science Foundation through grant NSF-SES-11-31897, which is greatly acknowledged.

Last but not least, I would like to express my deepest gratitude to my family for their great support and unconditional love.
Survey organizations sharing data with the public typically have to deal with two key issues, namely imputing missing values and protecting data subjects’ confidentiality. One approach that can do both is to use multiple imputations (MI, Rubin 1987). For missing data, MI fills in each missing value with $m \geq 2$ sensible draws from a joint probability distribution, resulting in $m$ completed data sets. The variability in these $m$ completed data sets reflect the uncertainty due to nonresponse and the uncertainty in predicting the missing values. For protecting a survey with confidential information, MI creates multiple synthetic datasets by replacing confidential values with draws from a model. The goal is to ensure that the disclosure risks are low while the information loss is acceptable.

In this thesis, I present novel methods for MI that can be applied to data confidentiality and missing data problems. The thesis is organized in four chapters. In Chapter 2, I present approaches to generating synthetic microdata for multivariate data that take on non-negative integer values, such as magnitude data in economic surveys. The basic idea is to estimate a mixture of Poisson distributions to describe the multivariate distribution, and release draws from the posterior predictive distri-
bution of the model. I develop approaches that guarantee the synthetic data sum to marginal totals computed from the original data, as well approaches that do not enforce this equality. For both cases, I present methods for assessing disclosure risks inherent in releasing synthetic magnitude microdata. I illustrate the methodology using economic data from a survey of manufacturing establishments.

In Chapter 3, I introduce a method of generating synthetic continuous data conditioned on protective intervals rather than the confidential values. The basic idea is to estimate synthesis models with the model parameters drawn from the protective intervals, with the goal that intruders learn no more than the intervals from the synthetic data. I illustrate this approach with simple simulation studies. The results suggest that this is a promising direction for releasing synthetic data with low posterior disclosure risk and acceptable data utility.

In Chapter 4, I present an approach for multiple imputation of missing data for continuous and categorical variables with focused and non-focused clustering. The basic idea is to separate variables into non-focused (e.g., almost-fully-observed) and focused (e.g., missing-a-lot) ones, and use two different sub-models for each. For the focused variables, we use a complicated sub-model that can capture complex distributions. For the non-focused variables, we use a relatively simple model structure. The model allows focused variables to depend locally on non-focused ones, as well as encodes dependence by linking two sets of cluster indicators. I propose a strategy of moving non-focused variables to the side of focused ones to improve estimation accuracy. I examine this strategy with several simulation studies. I apply the model to data from the American Community Survey. I also compare it against a hierarchically coupled mixture model with local dependence (Murray and Reiter, 2016).

In the remainder of this chapter, I provide necessary background material on MI for missing data in Section 1.1, on MI for disclosure control in Section 1.2, and on
Dirichlet process mixture modelling in Section 1.3.

1.1 Multiple Imputation for Missing Data

There are basically two types of nonresponse — unit and item. A survey has unit nonresponse when some of its entire records are missing. A survey suffers from item nonresponse when the values of the records are partially observed. Both of these kinds of missingness can have effects on the joint distribution. Consequently, missing values can lead to biased statistical inferences (Little and Rubin, 2002).

There are several strategies for dealing with missing data. In general, they fall into two groups. The strategies in the first group tend to draw the conclusions based on the observed data only. A most common and simple solution is casewise deletion, which removes any records with missing values. The estimates from casewise deletion can be biased when the missing values are systematically different from the observed one (Little and Rubin, 2002). Another solution is to weight the observed data according to the probability of being sampled in a given population. Then, the weighted observed data are used to compensate the missingness. Compared with casewise deletion, the last method can generate less biased estimates. However, it still sacrifices partially observed information.

The strategies in the second group tend to estimate the joint distribution of the completed data and impute the missing values. Imputation can be either single or multiple. A single imputation (SI) fills in one value for each missing item. A naive SI, for example, is to substitute the mean value for all missingness. It tends to underestimate the variance. A more complicated SI, for instance, can be a “hot-deck” method. It is applied in Current Population Survey to impute each missing item with one corresponding “similar” value (Kalton and Kasprzyk, 1986). A “similar” value is found by matching the missingness with the doner records. Even though SI imputes the missing values, the joint distribution of its completed data can be
biased. Additionally, it underestimates imputation variation.

1.1.1 Missing Data Mechanisms

There are various reasons for a survey to have nonresponse items or units. Basically, all the reasons can be categorized into three types of missing data mechanisms. Each mechanism has its own unique assumption about why the observations are missing. This leads to different data missing approaches. Thus, it is important to consider missing data mechanism before choosing missing data approaches. In the following, I briefly describe the concepts of missing data mechanisms (Rubin, 1976; Raghunathan, 2004; Little and Rubin, 2002; Paiva, 2014).

First, I define some relevant notations. Suppose data \( Y = (Y_{mis}, Y_{obs}) \) have the dimension \( n \times p \) with \( n \) observations and \( p \) variables. \( Y_{mis} \) is the missing data and \( Y_{obs} \) is the observed data. Let \( I \) be the matrix of nonresponse indicators with the same dimension of \( Y \), where \( I_{ij} = 1 \) when \( Y_{ij} \) is missing and \( I_{ij} = 0 \) otherwise. The matrix \( I \) is random because one cannot know the value of each \( I_{ij} \) until the data \( Y \) are collected. At the same time, I assume \( Y \) and \( I \) can be jointly captured by a model with parameters \( \theta = (\theta^{(Y)}, \theta^{(I)}) \). Here, \( \theta^{(Y)} \) is the parameter for the full data model \( Pr(Y|\theta^{(Y)}) \) and \( \theta^{(I)} \) is the parameter for the missing mechanism model \( Pr(I|Y, \theta^{(I)}) \).

When there is no relationship between the nonresponse indicators and any values (observed or missing), this is termed as missing completely at random (MCAR). Statistically, MCAR can be described as

\[
Pr(I|Y, \theta) = Pr(I|\theta^{(I)}),
\]

that is, the nonresponse indicators are independent of \( Y \) and \( \theta^{(Y)} \). This also implies that the analysis of \( Y_{obs} \) can give valid inferences about \( Pr(\theta^{(Y)}|Y) \) although there will generally be some loss of information. An example of MCAR is that values in
data file are missing due to accidental entry errors.

When there is a systematic relationship between the propensity of nonresponse indicators and the observed data, this is named as missing at random (MAR). Statistically, MAR can be represented as

\[ Pr(I|Y, \theta) = Pr(I|Y_{obs}, \theta^{(I)}), \]  

(1.2)

that is, the nonresponse indicators do depend on the observed data but not on the unobserved data (and data model parameter \( \theta^{(Y)} \)). This is equivalent to saying that any units with the same values of \( Y_{obs} \) follow the same statistical distribution, \( Pr(Y|\theta^{(Y)}) \). An example of MAR is that a patient’s body weight is measured only when his blood pressure is high, and his blood pressure is always measured. From (1.1) and (1.2), we can see that MAR is a weaker assumption than MCAR.

When the nonresponse indicators depend on the unobserved \( Y_{mis} \), we call this missing not at random (MNAR). That is,

\[ Pr(I|Y, \theta) = Pr(I|Y_{mis}, Y_{obs}, \theta^{(Y)}, \theta^{(I)}). \]  

(1.3)

It implies that modelling \( Pr(Y|\theta^{(Y)}) \) alone is insufficient. Therefore, a joint model of \( Y \) and \( I \) needs to be specified to include the missing data mechanism (Daniels and Hogan, 2000; Little and Rubin, 2002). An example for MNAR is that a subject refuses to reveal his salary because it is large.

Among the three data missing mechanisms, MAR is the most common one used in practice. However, distinguishing MAR from MNAR is practically difficult and mathematically impossible (Molenberghs et al., 2008). This is because the only way to do so is by analyzing the reason for missingness, which often is not possible. For example, a data collector needs to design a follow-up survey for non-respondents with a few key questions. From how the designed questions are answered, he may infer the missing data mechanism for the unobserved items. Since conducting addi-
tional surveys required more human and monetary resources, this method is typically unavailable.

1.1.2 Multiple Imputation

In MI, we need to combine analyses of multiple completed data drawn from a Bayesian posterior distribution (Rubin, 1987). This is called repeated-imputation inferences. We consider two versions, one for missing data and the other for data confidentiality (see Section 1.2).

For missing data, let $Q$ be a scalar quantity of interest in the survey. Denote $m$ as the number of completed data. Let $\hat{Q}_i$ be the statistical estimate of $Q$ from the $i^{th}$ completed data. Let $U_i$ be the estimated variance of each $\hat{Q}_i$. These estimates are combined as follows.

$$\bar{Q}_m = \frac{1}{m} \sum_{l=1}^{m} \hat{Q}_l/m$$  \hspace{1cm} (1.4)$$

$$\bar{U}_m = \frac{1}{m} \sum_{l=1}^{m} U_i/m$$  \hspace{1cm} (1.5)$$

$$B_m = \sum_{l=1}^{m} (\hat{Q}_l - \bar{Q}_m)^t (\hat{Q}_l - \bar{Q}_m)/(m - 1)$$  \hspace{1cm} (1.6)$$

$$T_m = (1 + m^{-1})B_m + \bar{U}_m.$$  \hspace{1cm} (1.7)$$

Here, $\bar{Q}_m$ is the point estimate of $Q$ from $m$ completed data and $T_m$ is the variance of $\bar{Q}_m$.

In theory, $(Q - \bar{Q}_x) \sim \mathcal{N}(0, T_x)$ for $m \to \infty$ and large $n$. In practice, $(Q - \bar{Q}_m)$ follows a Student’s $t$ distribution with the degree of freedom $\nu_m = (m - 1)(1 + \bar{U}_m/((1+m^{-1})B_m))^2$ for small $n$. This approximation is often sufficient for $m$ between 5 and 10. Barnard and Rubin (1999) modified the degree of freedom for even smaller $m$. For a vector $Q$, this combining rule can be further generalized according to Rubin (1987, Chapter 3), or Schafer (1997, Chapter 4), or Reiter and Raghunathan (2007).
1.2 Multiple Imputation for Disclosure Control

Surveys with microdata tend to contain sensitive information for each survey respondent. When sharing them, the data collectors and relevant users are required by laws to prevent identification of data subjects. Typical ways to disseminate disclosure-treated versions of confidential data include cell suppression, data masking, and data swapping (Willenborg and De Waal, 2012). These methods can decrease the utility in the released data in that they can easily alter the joint distribution. To avoid such drawbacks, Rubin (1993) proposed to apply MI for creating multiple, synthetic data sets for public release. Several studies have successfully applied MI for confidential data sets (Rubin, 1993; Little, 1993; Raghunathan et al., 2003; Reiter, 2003, 2004; Reiter and Raghunathan, 2007; Wang and Reiter, 2012; Hu et al., 2014; Paiva et al., 2014).

For the problem of data confidentiality, Raghunathan et al. (2003) proposed a variation of the combining rule for full synthesis (also see Reiter and Raghunathan (2007)). They proved that in fully synthetic samples, (1.7) needs to be changed to

\[ T_m = (1 + m^{-1})B_m - \bar{U}_m. \]  

(1.8)

When working with fully synthetic populations, the \( m \) repeated estimates \( \hat{Q}_i^{(p)} \) can be combined by (1.4). However, each variance \( U_i^{(p)} = 0 \). This is because each synthesis is an entire imputed population. Then, the total variance is \( T_m = (1 + m^{-1})B_m \) with \( B_m \) given in (1.6). Another variation is the combining rule for partially synthetic data proposed by (Reiter, 2003).
1.3 Dirichlet Process Mixture Modelling

1.3.1 Finite Mixture Modelling

The realization of MI is based on the joint distribution of the complete data. However, in practice such a distribution can have multiple modes, various distributional shapes and complex interactions. Thus, modelling it is challenging. One solution is to apply a finite mixture of standard distributions with various locations and scales, as I now review.

Suppose a data point $X_i$ has the dimension $p \geq 2$, for $i = 1, \ldots, n$ individuals. Denote $Z_i$ as its mixture component index. Then, the probability function of $X_i$ can be modelled as:

$$Pr(X_i = x_i) = \sum_{k=1}^{K} Pr(Z_i = k)Pr(x_i|\theta_k, Z_i = k). \quad (1.9)$$

Here, $K$ is a finite number of mixtures, and $Pr(x_i|\theta_k, Z_i = k)$ is the probability of $X_i = x_i$ in the $k^{th}$ mixture component with model parameter $\theta_k$.

Although based on standard distributions, (1.9) can pose complex computational challenges (Diebolt and Robert, 1994). Therefore, estimating its parameters relies on computationally intensive methods such as the Expectation- Maximization algorithm (EM, Dempster et al. 1977) or Markov chain Monte Carlo method (MCMC, Metropolis et al. 1953; Hastings 1970; Geman and Geman 1984; Tanner and Wong 1987; Gelfand and Smith 1990; Geyer 1992; Tierney 1994). The EM, as an optimization approach, addresses the problem by solving the likelihood equation for the mixture of distributions. The Bayesian approach using MCMC transforms the complicated model structure into a set of simple conditional models using latent variables.

In Chapter 2, a finite mixture of product Poisson distributions is proposed for count data with marginal sum constrains.
A typical question for a finite mixture modelling is how to choose the value of $K$ in (1.9). There exist different answers (Gershman and Blei, 2012). For example, one can fit the data with different values for $K$, and select the optimal $K$ by a model comparison method (Claeskens and Hjort, 2008). In a nonparametric Bayesian approach, one fits a single model that can adapt the model complexity to the data (Hjort et al., 2010; Gershman and Blei, 2012). Moreover, the complexity of a nonparametric Bayesian model grows automatically as more data are observed.

In nonparametric Bayesian statistics, a common way to extend a finite mixture model to an infinite one is to use the Dirichlet Process (DP, Ferguson 1973). A DP is a distribution over some probability measurable space $\theta$. That is, a random distribution $G$ is Dirichlet process distributed with base distribution $G_0$ and concentration parameter $\alpha$. This can be written as $G \sim \text{DP}(\alpha, G_0)$ if $(G(A_1), \ldots, G(A_p)) \sim \text{Dir}(\alpha G_0(A_1), \ldots, \alpha G_0(A_p))$ for any finite measurable partition $A_1, \ldots, A_p$ of $\theta$ (Teh, 2011).

For any measurable set $A \subset \theta$, we have $E[G(A)] = G_0(A)$. Thus, $G_0$ is the mean of the DP. The variance of $V[G(A)]$ is equal to $G_0(A)(1 - G_0(A))/(\alpha + 1)$. That is, the larger the concentration parameter $\alpha$, the smaller $V[G(A)]$ will be. In other words, the DP will concentrate more of its mass around the mean for large $\alpha$.

When $K$ goes to infinity, a DP mixture model for (1.9) can be described as

$$X_i = x_i|\theta_i \sim Pr(x_i|\theta_i) $$ (1.10)

$$\theta_i|G \sim G(\theta) $$ (1.11)

$$G \sim \text{DP}(\alpha, G_0). $$ (1.12)
We can write these in a stick-breaking construction (Sethuraman, 1994) as

\[ X_i | Z_i \sim Pr(X_i | \theta_{Z_i}) \]  
\[ \theta \sim G_0 \]  
\[ Pr(Z_i = k) = \pi_k \]  
\[ \pi_k = \lambda_k \prod_{l<k} (1 - \lambda_l) \text{ with } \pi_k = \lambda_1 \]  
\[ \lambda_l \overset{i.i.d.}{\sim} B(1, \alpha), \]

in which \( k = 1, \ldots, \infty \).

If we integrate out parameters \( \theta \) and mixing proportions \( \pi \), at most \( n \) components will be associated with data. Usually the number of the occupied mixture components is much less than \( n \). Thus, often we can approximate the infinite DP with a finite version, provided \( K \) is large (Ishwaran and James, 2001).

1.3.3 Dirichlet Process Mixtures of Product Multinomial Distributions

Dunson and Xing (2009) proposed the Dirichlet Process Mixtures of Product Multinomial (DPMPM) model for unordered categorical data. It is a nonparametric Bayesian approach with the prior on the entire space of distribution for multivariate nominal data. Thus, the model allows dependencies in unordered categorical variables a priori.

Si and Reiter (2013) applied DPMPM model to impute missingness in the Trends in International Mathematics and Science Study. Hu et al. (2014) proposed a modified version of DPMPM to synthesize the nested household data in the American Community Survey. Murray and Reiter (2016) embedded DPMPM into their model for more complicated continuous and categorical data types.

In Chapter 4, I extend the model from Murray and Reiter (2016). Since DPMPM is involved, I briefly describe its model structure. Suppose we have unordered cate-
girical data $X$ with $n$ observations and $p$ variables. Then, we have

$$Pr(X_i = x_i | Z_i = k,-) = \prod_{j=1}^{p} \psi_{kz_{ij}}^{(j)},$$

(1.18)

where $Z_i$ is the latent mixture component index for $X_i$ with $z = 1, \ldots, \infty$. The probability of $Z_i = k$ is assigned as $\pi_k$, which can be drawn from a stick-breaking representation (Sethuraman, 1994) of the Dirichlet process (Ferguson, 1973; Blackwell and MacQueen, 1973). We have

$$\pi_k = V_k \prod_{g < k} (1 - V_g)$$

(1.19)

$$V_g \sim \mathcal{B}(1, \alpha)$$

(1.20)

where $\alpha$ is the precision parameter for the Dirichlet process of $\pi_k$. When $\alpha$ is small, the prior of $\pi_k$ favors a sparse representation with most of the weight on the first few components. In a fully Bayesian setting, the hyper prior for $\alpha$ can be taken as $\alpha \sim \mathcal{G}(a_\alpha, b_\alpha)$. Constant $a_\alpha$ and $b_\alpha$ can be chosen so that they allow the data to inform about an appropriate degree of sparsity. In practice, when the data size is big, the influences from $a_\alpha$ and $b_\alpha$ can be small (Murray and Reiter, 2016).

In (1.18), $\psi_{z_{ij}}^{(j)}$ is the probability of $X_{ij} = x_{ij}$ in the latent class $z$. Suppose the $j^{th}$ variable in $X$ has the levels from $j_1$ to $j_{d_j}$. Then, $\psi_{zj} = (\psi_{zj_1}, \ldots, \psi_{zj_{d_j}})'$ is a vector of probabilities for $X_j$ being assigned into any one of these levels. We use a conjugate prior for $\psi_{zj}$, $\text{Dir}(c_{j_1}, \ldots, c_{j_{d_j}})$ with $c_j = (c_{j_1}, \ldots, c_{j_{d_j}})'$ being either $(1, \ldots, 1)'$ or $(1/d_j, \ldots, 1/d_j)'$ for noninformative Bayesian priors.

1.3.4 Hierarchically Coupled Mixture Model with Local Dependency

Murray and Reiter (2016) developed the hierarchically coupled mixture model with local dependence (HCMM-LD) to evaluate the redesign for the Survey of Income and Program Participation (SIPP) by MI. SIPP has numerous continuous and categorical
variables with complicated distributions. For example, the mean and variance of annual income can vary by different levels of educational attainment. Within each level, the distribution of annual income can be heavily skewed. Thus, SIPP contains complex dependencies with nonstandard distributions.

Multivariate modelling methods such as a general location model (Olkin and Tate, 1961; Little and Schluchter, 1985; Schafer, 1997) or a “chained equations” approach (Van Buuren and Oudshoorn, 1999; Raghunathan et al., 2001) can fail mainly due to the possibility of mis-specifying the key interactions. In addition, Dunson and Bhattacharya (2010) found that the number of clusters with fully factorized kernels in a Dirichlet process mixture must grow to capture the true dependence in the joint distribution.

Both of these two issues motivate Murray and Reiter (2016) to propose HCMM-LD. Its main feature is the property of local dependence. This gives one freedom to specify the link function between the continuous and categorical variables within each local cluster. At the same time, this prevents the number of mixture components from growing rapidly. A related idea of preventing the proliferation of mixture components can be found in the research papers of Banerjee et al. (2013).

In the following, I briefly summarize the HCMM-LD model. Denote $Y$ as the continuous variable and $X$ as the categorical features. The dimension of $Y$ is $n \times q$ and the dimension of $X$ is $n \times p$, with $n$ being the number of observations. Let $D(\cdot)$ be a function which takes the main effect of the input variables. The data model in HCMM-LD is

$$
(Y_i | X_i = x_i, H_i^{(Y)} = r, -) \sim \mathcal{N}(D(x_i)B_r, \Sigma_r)
$$

(1.21)

$$
Pr(X_i = x_i | H_i^{(X)} = s, -) = \prod_{j=1}^{p} \psi_{s_{x_{ij}}(j)},
$$

(1.22)

where $H_i^{(Y)}$ and $H_i^{(X)}$ are the mixture component indices for $Y_i$ and $X_i$. Here, $B_r$ and
\( \Sigma_r \) are the regression coefficients and covariance matrix for \( (Y_i | X_i = x_i, H_i^{(Y)} = r, -) \) within the latent cluster \( r \) in \( H_i^{(Y)} \), and \( \psi_{sx_{ij}}^{(j)} \) is the probability of the \( j^{th} \) variable in \( X_i \) being \( x_{ij} \) within the latent cluster \( s \) in \( H_i^{(X)} \).

The hierarchical priors of the mixture component indices are

\[
Pr(H_i^{(X)} = s, H_i^{(Y)} = r | Z_i = z) = \phi_{zr}^{(x)} \phi_{sr}^{(y)} \tag{1.23}
\]

\[
Pr(Z_i = z) = \lambda_z, \tag{1.24}
\]

where \( \phi_{zr}^{(x)} = (\phi_{z1}^{(x)}, \ldots, \phi_{zk(X)}^{(x)})' \) and \( \phi_{zr}^{(y)} = (\phi_{z1}^{(y)}, \ldots, \phi_{zk(Y)}^{(y)})' \) are probability vectors assigned independent truncated stick breaking priors (Ishwaran and James, 2001).

From (1.22) to (1.24), we see that HCMM-LD is a mixture of mixture distributions. That is, it hierarchically combines a mixture of multivariate normals for \( Y \) and a DPMPM model for \( X \) through the local linkage function \( D(\cdot) \). The virtue of local dependence of this model can be seen through the conditional probability of \( Y_i \) given \( X_i \):

\[
Pr(Y_i = y_i | X_i = x_i) = \sum_{r=1}^{k^{(Y)}} \frac{w_r(x_i)}{\sum_{i=1}^{k^{(Y)}} w_i(x_i)} N(y_i; D(x_i)B_r, \Sigma_r) \tag{1.25}
\]

with the weight \( w_r(x_i) \) being \( \sum_{j=1}^{k^{(Y)}} \lambda_z \phi_{zr}^{(y)} \sum_{s=1}^{k^{(X)}} \phi_{zs}^{(x)} \prod_{j=1}^{p} \psi_{sx_{ij}}^{(j)} \). Therefore, HCMM-LD is able to capture the interactions between \( Y_i \) and \( X_i \) that are not coded in \( D(X_i) \).

In Chapter 4, I extend this HCMM-LD model with focused and non-focused clustering. The major difference is that HCMM-LD is a joint modelling method that does not distinguish focused (i.e., missing-a-lot) variables from non-focused (i.e., almost-fully-observed) ones. On the contrary, the new model in Chapter 4 is a locally conditional model that tries to take the advantage of non-focused variables to help focused characteristics. The new model sacrifices model complexity for non-focused variables, in favor of more accurate model fitting for focused ones.
2

Releasing Synthetic Magnitude Microdata Constrained to Fixed Marginal Totals

This chapter very closely follows the paper of Wei and Reiter (2016).

2.1 Introduction

Many statistical agencies collect data that take on non-negative integer values, such as magnitudes in economic data. Agencies typically disseminate tabular summaries of such data, reporting marginal totals and suppressing or otherwise blurring internal cells of the table (Willenborg and De Waal, 2012; Federal Committee on Statistical Methodology, 2005). They generally do not provide unrestricted, public access to microdata files, that is, data on individual records, because of concerns that doing so could result in too high risks of disclosures of confidential information.

In this article, we present an approach to releasing magnitude microdata based synthetic data, i.e., data simulated from statistical models estimated with the confidential data (Rubin, 1993; Little, 1993; Reiter and Raghunathan, 2007). Synthetic data approaches have been used to create public use files for several high profile
products in the United States, including the Survey of Income and Program Participation (Abowd et al., 2006), the American Community Survey group quarters data (Hawala, 2008), the OnTheMap application (Machanavajjhala et al., 2008), and the Longitudinal Business Database (Kinney et al., 2011, 2014). They also have been used in other countries, notably in Germany (Drechsler et al., 2008a,b).

As the synthesis model, we use a finite mixture of Poisson distributions. This multivariate model can capture complex associations among the variables without a priori assuming particular correlation structure, as well as respect the integer nature of the data. As we show, the model can be adapted to generate synthetic values that sum to fixed marginal totals. Finally, the model facilitates approaches to protect outlying observations, as one can manipulate mixture component memberships to effect a form of multivariate microaggregation of records.

The remainder of this article is organized as follows. In Section 2.2, we introduce the mixture of Poisson distributions model and a Bayesian algorithm for estimating the posterior distribution of its parameters. In Section 2.3, we describe several methods for using this model to generate synthetic microdata. In Section 2.4, we present methods for assessing the disclosure risks inherent in releasing the synthetic microdata. In Section 2.5, we illustrate the methods using data from a manufacturing survey from the country Colombia. Finally, in Section 2.6, we conclude with directions for future research.

2.2 Mixtures of Poisson Distributions

For $i = 1,\ldots,n$ and $j = 1,\ldots,p$, let $y_{ij}$ be the value of the $j$th variable for the $i$th individual. Let $Y_i = (y_{i1},\ldots,y_{ip})$, and let $\mathcal{Y} = \{Y_i : i = 1,\ldots,n\}$. We assume that each individual belongs to one of $K$ latent classes, also known as mixture components; we discuss the choice of $K$ later. For $i = 1,\ldots,n$, let $z_i$ indicate the component of individual $i$. We assume that $Pr(z_i = k) = \pi_k$ for $k = 1,\ldots,K$; that
is, component probabilities are the same for all individuals. Within each component \( k \), we assume that each variable \( j \) has a Poisson distribution with rate parameter \( \theta_{kj} \). Let \( \theta_k = (\theta_{k1}, \ldots, \theta_{kp}) \), and let \( \Theta = \{\theta_k : k = 1, \cdots, K\} \). We assume that variables are independent within components.

Putting this together, we have the mixture of Poisson distributions model,

\[
Y_i \mid z_i, \Theta \sim \prod_{j=1}^{p} \text{Pois}(y_{ij} \mid \theta_{zij})
\]

\[
z_i \mid \pi \sim \text{Mult}(1, \pi),
\]

where \( \pi = (\pi_1, \ldots, \pi_K) \) and “\( \text{Mult}(1, \pi) \)” indicates a multinomial distribution with one trial and probabilities \( \pi \). Averaging over \( z_i \), the model can be expressed as

\[
\Pr(Y_i \mid \Theta, \pi) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{p} \frac{\theta_{y_{ij}}^{y_{ij}} \exp(-\theta_{zij})}{y_{ij}!}.
\]

Probabilities for various marginal distributions can be computed using appropriate subsets of the product in (2.3); for example,

\[
\Pr(y_{ij} \mid \Theta, \pi) = \sum_{k=1}^{K} \pi_k \frac{\theta_{y_{ij}}^{y_{ij}} \exp(-\theta_{zij})}{y_{ij}!}
\]

\[
\Pr(y_{ij}, y_{ij'} \mid \Theta, \pi) = \sum_{k=1}^{K} \pi_k \frac{\theta_{y_{ij}}^{y_{ij}} \exp(-\theta_{zij}) \theta_{y_{ij'}}^{y_{ij'}} \exp(-\theta_{zij'})}{y_{ij}! y_{ij'}!}.
\]

The model favors putting individuals with similar data values in the same component. Because of this, even though variables are independent within components, the mixture model encodes potential dependence when averaged over components. To see this, note from (2.4) and (2.5) that \( \Pr(y_{ij} \mid \Theta, \pi) \Pr(y_{ij'} \mid \Theta, \pi) \neq \Pr(y_{ij}, y_{ij'} \mid \Theta, \pi) \).

Since we observe only \( Y \), we have to estimate \( Z = \{z_i : i = 1, \ldots, n\} \) and the parameters \( (\Theta, \pi) \). To do so, we use a Bayesian model with conjugate prior
distributions for the parameters. We use the prior distributions,
\[
\pi = (\pi_1, \cdots, \pi_K) \sim \text{Dir}(d_1, \cdots, d_K), \quad \text{with } d_k = 1/K \text{ for } k = 1, \cdots, K \tag{2.6}
\]
\[
\log(\theta_k) = (\log(\theta_{k1}), \cdots, \log(\theta_{kp})) \overset{iid}{\sim} \text{N}_p(\cdot|\mu, \Sigma), \quad \text{for } k = 1, \cdots, K \tag{2.7}
\]
where “Dir” indicates a Dirichlet distribution and “N_p” indicates a p-variate normal distribution. We use a standard Jeffreys prior distribution for the hyperparameters \((\mu, \Sigma)|\Sigma|^{-(p+1)/2}\). We estimate the model via Markov Chain Monte Carlo (MCMC) techniques. See the appendix for the full conditional distributions for the sampler.

We set \(K\) large enough to capture complex features of the multivariate distribution, but small enough to facilitate efficient computation. Setting \(K\) too small tends to degrade estimates of the dependence relationships across variables. For highly correlated variables typical in magnitude data, we recommend starting with a modest value of \(K\), say \(K = n/10\), and running the algorithm to convergence. In each iteration of the MCMC (after burn-in), we check the number of components occupied by at least one observation. If the numbers of occupied classes reach \(K\), we increase \(K\) and redo the MCMC sampling.

2.3 Synthesizers Based on Mixtures of Poissons

We now turn the mixture of Poisson distributions model into synthetic data generators. We present generators that ensure the synthetic values sum to the marginal totals in the confidential data, and generators that do not consider these marginal totals.

2.3.1 The Mixture of Poisson Synthesizer

When we do not constrain the synthetic totals to equal the totals in the confidential data, we can generate synthetic data directly from the model in Section 2.2. To do so, we sample a draw of \((\pi, \Theta)\) from the posterior distribution. Using the sampled \(\pi\),
we draw \( n \) new component indicators, \( z_1, \ldots, z_n \), using (2.2). For each \( z_i \), we use the corresponding drawn value of \( \theta_{z_i} \) in (2.1) to generate a synthetic \( \tilde{Y}_i = (\tilde{y}_{i1}, \ldots, \tilde{y}_{ip}) \). The synthetic dataset is \( \mathcal{Y} = \{ \tilde{Y}_i : i = 1, \ldots, n \} \). We call this the mixture of Poissons (MP) synthesizer.

One can generate \( L \) synthetic datasets, \( \tilde{Y}^{(l)} \) with \( l \in \{1, \ldots, L\} \), as recommended in (Raghunathan et al., 2003; Reiter, 2003), by independently repeating the process of sampling parameters and data \( L \) times. Alternatively, one can use the posterior mode of \( (\pi, \Theta) \) in each replication, as suggested for partial synthesis or census files by (Reiter and Kinney, 2012).

2.3.2 The Mixture of Multinomial Synthesizer

To generate synthetic data with marginal totals matching those in the confidential data, we take advantage of the well-known mathematical relationship between sums of Poisson random variables and the multinomial distribution. As an example of this relationship, let \( (X_1, X_2, X_3) \) be three independently distributed random variables with rate parameter \( \lambda \). Given \( T = X_1 + X_2 + X_3 \), one can show that \( (X_1, X_2, X_3) \mid T \sim \text{Mult}(T, (1/3, 1/3, 1/3)) \).

For any draw of \( Z \), say \( Z^{(l)} \), we can compute and store the totals in each occupied component; that is, for \( k = 1, \ldots, K \) and \( j = 1, \ldots, p \), we compute \( T_{kj}^{(l)} = \sum_{i=1}^n y_{ij}(z_i^{(l)} = k) \), where \( (.) \) equals one when the condition inside the parenthesis is true and equals zero otherwise. We also compute \( n_k^{(l)} = \sum_{i=1}^n (z_i^{(l)} = k) \), the number of cases that belong to component \( k \) according to \( Z^{(l)} \). Clearly, when \( n_k^{(l)} = 1 \) we have \( T_{kj}^{(l)} = y_{ij} \) for the single individual with \( z_i^{(l)} = k \). We call individuals placed in components with \( n_k^{(l)} = 1 \) singletons, and all other individuals nonsingletons. Note that singleton status can change for different draws of \( Z^{(l)} \).

We generate synthetic data as follows.
(1) Using the model in Section 2.3.1, sample values of $\Theta^{(l)}$ and $Z^{(l)}$.

(2) Using $Z^{(l)}$, compute $\{n_k^{(l)}, T_{kj}^{(l)} : k = 1, \ldots, K; j = 1, \ldots, p\}$.

(3) Generate synthetic values for singleton and nonsingleton components —

(3.1) For each component $k$ with $n_k^{(l)} > 1$, sample one draw of variable $j$ from the multinomial distribution with sample size equal to $T_{kj}^{(l)}$ and $n_k^{(l)}$ levels with all probabilities equal to $1/n_k^{(l)}$. This results in a vector of $n_k^{(l)}$ values of $\tilde{y}_{ij}^{(l)}$. The support of each $\tilde{y}_{ij}^{(l)}$ is $\{0, \ldots, T_{kj}^{(l)}\}$, with $\sum_{i:z_i^{(l)}=k} \tilde{y}_{ij}^{(l)} = T_{kj}^{(l)}$.

We repeat the process for each $j = 1, \ldots, p$ by sampling from independent multinomial distributions within each component.

(3.2) When $Z^{(l)}$ includes singletons, we create a new component $K + 1$ that includes the data from all singletons. Let $n_{K+1}^{(l)} = \sum_{k=1}^{K} n_k^{(l)} = 1$ and $T_{(K+1)j}^{(l)} = \sum_{i=1}^{n} y_{ij}(n_{zi} = 1)$ be, respectively, the number of cases and marginal total of variable $j$ in this new component. For each $z_i = k$ such that $n_k^{(l)} = 1$, we collect its value of $\theta_{zi,j}^{(l)}$, resulting in a $n_{K+1}^{(l)}$ dimensional vector that we write as $(\lambda_{1j}^{(l)}, \ldots, \lambda_{n_{K+1}j}^{(l)})$. Finally, we sample one draw of variable $j$ from the multinomial distribution with sample size equal to $T_{(K+1)j}^{(l)}$ and $n_{K+1}^{(l)}$ levels with probabilities

$$\lambda^{(l)} = (\lambda_{1j}^{(l)}/\sum_{m=1}^{n_{K+1}^{(l)}} \lambda_{mj}^{(l)}, \ldots, \lambda_{n_{K+1}j}^{(l)} / \sum_{m=1}^{n_{K+1}^{(l)}} \lambda_{mj}^{(l)}) \quad (2.8)$$

This results in a vector of $n_{K+1}^{(l)}$ values of $\tilde{y}_{ij}^{(l)}$. The support of each $\tilde{y}_{ij}^{(l)}$ is $\{0, \ldots, T_{(K+1)j}^{(l)}\}$ with $\sum_{i:n_{zi}^{(l)}=1} y_{ij} = T_{(K+1)j}^{(l)}$.

Since the marginal totals in the synthetic and confidential data match in each component, the overall marginal totals also match. We call this the mixture of multinomials
To see why we use Step 3.2 for singletons, consider what happens if instead we generate synthetic values for singleton cases using the method in Step 3.1. The synthesizer would generate synthetic values for these individuals exactly equal to their original values, which likely represents an unacceptable disclosure risk. Step 3.2 introduces uncertainty in the synthesized values for singletons by forcing $n_{k+1}^{(l)} > 1$. We note that one could adapt the MM synthesizer for other aggregations, for example combine all $k$ with $n_{k}^{(l)} \leq 3$. Generally, this reduces the quality of the synthetic data but also could reduce disclosure risks. The idea of collapsing clusters motivates our third synthesis procedure.

We also considered a synthesizer that moves any individual $i$ that is a singleton into the $z_i$ with the second largest posterior probability for that $i$. This approach also can be used when the data include only one singleton component. In our empirical investigations, this approach had less attractive risk-utility profiles than the MM synthesizer; hence, we do not discuss it further here.

### 2.3.3 The Tail-Collapsed Mixture of Multinomial Synthesizer

In many contexts, agencies consider individuals in the tails of distributions to be at the highest risks of disclosure. Agencies apply heavier disclosure treatment to those individuals. We implement a synthesizer in this spirit, which we call the tail-collapsed mixture of multinomials (TCMM) synthesizer, by combining components that feature individuals with risky (e.g., large) values into a single component. This can be used to collapse tails of individual variables, effectively performing a model-based variation of microaggregation plus noise (Oganian and Karr, 2006). Specifically, we synthesize data as follows.

1. Repeat Step (1) from the MM synthesizer.
(2) Repeat Step (2) from the MM synthesizer.

(2a) For $k = 1, \ldots, K$ and $j = 1, \ldots, p$, compute $\bar{y}_{kj}^{(l)} = \sum_{z_{ik}^{(l)} = k} y_{ij}^{(l)}/n_k^{(l)}$. For $j = 1, \ldots, p$, let $q_j^{(l)}$ be the $q$th quantile of all $\{\bar{y}_{kj}^{(l)} : k = 1, \ldots, K\}$. Let $r_k = 1$ for any $k$ such that $\bar{y}_{kj}^{(l)} > q_j^{(l)}$ for at least one $j$, and let $r_k = 0$ otherwise. We collapse all components with $r_k = 1$ into a new single component, which we denote $K + 2$.

(3) Repeat Step (3) from the MM synthesizer, using Step (3.1) for component $K + 2$ and the appropriate steps for the remaining singleton and nonsingleton components.

The criterion in Step (2a) labels individuals as being at risk, not their specific variables. This is because the model assigns individuals to components, not variables to components. It is not straightforward to assign different variables for the same record to different components. We note that one can adjust any $q_j^{(l)}$ to control the extent of collapsing.

2.4 Disclosure Risk Metrics

When all values are synthesized, it is nonsensical for intruders to match any synthetic dataset to external files, since there is no unique mapping of the rows (individuals) in the synthetic datasets, $\tilde{\mathbf{Y}} = \{\tilde{\mathbf{Y}}^{(l)} : l = 1, \ldots, L\}$, to the rows in the original data $\mathbf{Y}$ (Hu et al., 2014). We therefore consider questions of the form: can intruders accurately infer from $\tilde{\mathbf{Y}}$ the original data values for particular individuals in $\mathbf{Y}$? With magnitude data, such questions usually focus on individuals with large values of confidential variables.

The ability of intruders to guess values accurately depends on the amount of information they possess about $\mathbf{Y}$. Of course, it is impossible to know this. Hence,
we assess disclosure risks under multiple intruder attack scenarios. We consider several “worst case” scenarios in which the intruder knows all but one or two values in $\mathbf{Y}$, and seeks to use $\tilde{\mathbf{Y}}$ to learn about these unknown values. As a variant on these scenarios, we assume the intruder knows all data values for all but one or two records. Finally, we consider a scenario commonly used by official statistics agencies with business establishment data, namely that an individual with the second largest value of a certain variable attempts to use the released data to learn about the individual with the largest value of the variable.

2.4.1 Scenario I — Largest Two Values Are Unknown

Suppose that an intruder does not know the values of $y_{(1)j}$ and $y_{(2)j}$, where the subscripts (1) and (2) indicate the largest and second largest values of variable $j$ in $\mathbf{Y}$. The intruder knows all other data in $\mathbf{Y}$, which we denote as $\mathbf{Y}_{-(1,2)j} = \mathbf{Y} - \{y_{(1)j}, y_{(2)j}\}$. Let $Y_{(1)j}$ and $Y_{(2)j}$ be random variables representing the intruder’s guesses about $y_{(1)j}$ and $y_{(2)j}$. When the total $T_j$ of variable $j$ from $\mathbf{Y}$ is available, the intruder knows the value of $T_{2j} = Y_{(1)j} + Y_{(2)j} = y_{(1)j} + y_{(2)j}$, since $T_{2j}$ is simply $T_j$ minus the sum of $y_{ij}$ for values available in $\mathbf{Y}_{-(1,2)j}$. Let $(\hat{y}_{(1)j}, \hat{y}_{(2)j})$ represent a possible guess at $(Y_{(1)j}, Y_{(2)j})$.

When $T_j$ is available, we assume that the intruder computes the posterior probability,

$$Pr(Y_{(1)j} = \hat{y}_{(1)j}, Y_{(2)j} = \hat{y}_{(2)j} \mid \tilde{\mathbf{Y}}, \mathbf{Y}_{-(1,2)j}, Y_{(1)j} + Y_{(2)j} = T_{2j})$$

$$\propto Pr(\tilde{\mathbf{Y}} \mid \mathbf{Y}_{-(1,2)j}, Y_{(1)j} = \hat{y}_{(1)j}, Y_{(2)j} = \hat{y}_{(2)j})$$

$$Pr(Y_{(1)j} = \hat{y}_{(1)j}, Y_{(2)j} = \hat{y}_{(2)j} \mid \mathbf{Y}_{-(1,2)j}, Y_{(1)j} + Y_{(2)j} = T_{2j}),$$

(2.9)

for all feasible values of $(\hat{y}_{(1)j}, \hat{y}_{(2)j})$. The first probability after the proportionality sign is the likelihood of generating the proposed-to-be-released synthetic data, assuming that the unknown $(Y_{(1)j}, Y_{(2)j}) = (\hat{y}_{(1)j}, \hat{y}_{(2)j})$. The probability is high (low)
for values of \((\hat{y}_{(i)j}, \hat{y}_{(2)j})\) that are (not) likely to have generated the synthetic data. The second probability after the proportionality sign is the intruder’s prior distribution on \((Y_{(1)j}, Y_{(2)j})\). We assume that the intruder knows which synthesizer from Section 2.3 was used to generate \(\tilde{Y}\), but does not know any parameter values used in the generation, including \(q_j\).

When \(T_j\) is not available, we consider a worst case scenario of an intruder who knows all values except one, in particular the largest value. Let \(\mathbf{Y}_{-(1)j} = \mathbf{Y} - y_{(1)j}\). We assume that the intruder seeks to compute the posterior probability,

\[
Pr(Y_{(1)j} = \hat{y}_{(1)j} \mid \tilde{Y}, \mathbf{Y}_{-(1)j}) \propto Pr(\tilde{Y} \mid \mathbf{Y}_{-(1)j}, Y_{(1)j} = \hat{y}_{(1)j}) Pr(Y_{(1)j} = \hat{y}_{(1)j} \mid \mathbf{Y}_{-(1)j}),
\]

for a range of plausible values of \(\hat{y}_{(i)j}\).

For either case, we need to set the prior distribution. A sensible choice for (2.9) is the uniform distribution on all feasible values of \((\hat{y}_{(1)j}, \hat{y}_{(2)j})\). The analogue for (2.10) is a uniform distribution over a reasonable range \(\hat{y}_{(1)j}\), bounded below by \(y_{(2)j}\). These uniform prior distributions allow the probability to be determined entirely from the released synthetic data. An alternative, informative prior distribution uses the model in Section 2.3, estimated with \(\mathbf{Y}_{-(1,2)j}\) or \(\mathbf{Y}_{-(1)j}\), depending on what is assumed known to the intruder. In this case, the probabilities associated with any feasible \((\hat{y}_{(1)j}, \hat{y}_{(2)j})\) are determined from the estimated model parameters. Estimates generated from the informative prior distributions could be close to the true values when variable \(j\) is highly correlated with other variables assumed known to the intruder. Thus, we use the estimates generated from the informative prior distributions as reference points to assess how much additional risk is incurred by releasing the synthetic data in these worst case scenarios.

To approximate the probabilities in (2.9) and (2.10), we use an importance sampling strategy to reduce computation time, following the approach suggested by (Reiter et al., 2014; Hu et al., 2014; Paiva et al., 2014). We present the ap-
proach in detail for (2.9). Suppose that the MCMC sampler results in $B$ draws from the posterior distribution of $(\pi, \Theta, Z)$. Let $\mathcal{P}_b$, where $b = 1, \ldots, B$, represent each draw. We use these draws as the proposal distribution for the importance sampler, which we write as $Pr(\mathcal{P}_b \mid \mathcal{Y})$. Using this proposal, we approximate

$$Pr(\tilde{\mathcal{Y}} \mid \mathcal{Y}_{-(1,2)}, Y_{(1)} = \hat{y}_{(1)}, Y_{(2)} = \hat{y}_{(2)})$$

in (2.9) as

$$Pr(\tilde{\mathcal{Y}} \mid \mathcal{Y}_{-(1,2)}, Y_{(1)} = \hat{y}_{(1)}, Y_{(2)} = \hat{y}_{(2)}) = \frac{\sum_{b=1}^B Pr(\tilde{\mathcal{Y}} \mid \mathcal{P}_b)Pr(\mathcal{P}_b \mid Y_{(1)} = \hat{y}_{(1)}, Y_{(2)} = \hat{y}_{(2)}, \mathcal{Y}_{-(1,2)})}{\sum_{b=1}^B Pr(\mathcal{P}_b \mid Y_{(1)} = \hat{y}_{(1)}, Y_{(2)} = \hat{y}_{(2)}, \mathcal{Y}_{-(1,2)})}$$

(2.11)

$$\approx \frac{\sum_{b=1}^B Pr(\tilde{\mathcal{Y}} \mid \mathcal{P}_b)Pr(Y_{(1)} = \hat{y}_{(1)}, Y_{(2)} = \hat{y}_{(2)} \mid \mathcal{P}_b)Pr(Y_{(1)} = y_{(1)}, Y_{(2)} = y_{(2)} \mid \mathcal{P}_b)}{\sum_{b=1}^B Pr(Y_{(1)} = \hat{y}_{(1)}, Y_{(2)} = \hat{y}_{(2)} \mid \mathcal{P}_b)Pr(Y_{(1)} = y_{(1)}, Y_{(2)} = y_{(2)} \mid \mathcal{P}_b)}$$

(2.12)

Here, $Pr(\tilde{\mathcal{Y}} \mid \mathcal{P}_b) = \prod_{i=1}^L Pr(\tilde{\mathcal{Y}}^{(i)} \mid \mathcal{P}_b)$ is computed based on the synthesizer, using the draws in $\mathcal{P}_b$. For any $(\hat{y}_{(1)}, \hat{y}_{(2)})$, the values of $Pr(Y_{(1)} = \hat{y}_{(1)}, Y_{(2)} = \hat{y}_{(2)} \mid \mathcal{P}_b, \mathcal{Y}_{-(1,2)})$ are computed from the mixture Poisson model. Once we approximate $Pr(\tilde{\mathcal{Y}} \mid \mathcal{Y}_{-(1,2)}, Y_{(1)} = \hat{y}_{(1)}, Y_{(2)} = \hat{y}_{(2)})$, we multiply it by the prior distribution in (2.9) to obtain the value of the kernel at the particular $(\hat{y}_{(1)}, \hat{y}_{(2)})$ under consideration. Repeating for all $(\hat{y}_{(1)}, \hat{y}_{(2)})$ of interest, we obtain the normalizing constant.

Intruders cannot actually use $Pr(\mathcal{P}_b \mid \mathcal{Y})$, since it relies on values that are presumed unknown to the intruder. Hence, the importance sampler tends to offer a conservative approximation to the intruder’s probabilities. The intruder could obtain a proposal distribution using the $\mathcal{P}_b$ estimated from the model fit to $\tilde{\mathcal{Y}}$. This could be unreliable for data generated by the TCMM method, since the tails of the synthetic data could be quite different than those of the original data. Alternatively, the intruder could use $Pr(\mathcal{P}_b \mid \mathcal{Y}_{-(1,2)})$, averaging over the missing values of $(y_{(1)}, y_{(2)})$ in the Gibbs sampler. We use this proposal distribution in the applications of Section 2.5.

When $T_j$ is not available, we use an importance sampling strategy akin to (2.12).
2.4.2 Scenario II — Two Entire Records Are Unknown

As a slight weakening of the information available to the intruder in Scenario I, we now consider intruders who do not know any values for two entire records, say \(y_{(1)}\) and \(y_{(2)}\), and want to learn their values of some variable \(j\). For example, \(y_{(1)}\) and \(y_{(2)}\) might include the data for the individuals with the largest two values of variable \(j\). As in (2.9), when \(T_2 = (T_{21}, \ldots, T_{2p})\) is known to the intruder, we compute

\[
Pr(Y_{(1)j} = \hat{y}_{(1)j}, Y_{(2)j} = \hat{y}_{(2)j} | \tilde{\mathbf{Y}}, \mathbf{Y}_{-(1,2)}, \mathbf{Y}_{(1)} + \mathbf{Y}_{(2)} = T_2)
\]

\[
\propto Pr(\tilde{\mathbf{Y}} | \mathbf{Y}_{-(1,2)}, Y_{(1)j} = \hat{y}_{(1)j}, Y_{(2)j} = \hat{y}_{(2)j}, T_2)
\]

\[
Pr(Y_{(1)j} = \hat{y}_{(1)j}, Y_{(2)j} = \hat{y}_{(2)j} | \mathbf{Y}_{-(1,2)}, \mathbf{Y}_{(1)} + \mathbf{Y}_{(2)} = T_2),
\]

(2.13)

where \(\mathbf{Y}_{-(1,2)} = \mathbf{Y} - \{y_{(1)}, y_{(2)}\}\).

When \(T_2\) is not available, we consider a worst case scenario of an intruder who knows values for all individuals except some \(y_{(1)}\). Let \(\mathbf{Y}_{-(1)} = \mathbf{Y} - y_{(1)}\). We compute the posterior probability,

\[
Pr(Y_{(1)j} = \hat{y}_{(1)j} | \tilde{\mathbf{Y}}, \mathbf{Y}_{-(1)}) \propto Pr(\tilde{\mathbf{Y}} | \mathbf{Y}_{-(1)}, Y_{(1)j} = \hat{y}_{(1)j})Pr(Y_{(1)j} = \hat{y}_{(1)j} | \mathbf{Y}_{-(1)}),
\]

(2.14)

for a range of plausible values of \(\hat{y}_{(1)j}\).

For both cases, we can adapt the importance sampling strategy outlined in Section 2.4.1 to approximate the probabilities. For example, we approximate (2.13) with

\[
Pr(\tilde{\mathbf{Y}} | \mathbf{Y}_{-(1,2)}, Y_{(1)j} = \hat{y}_{(1)j}, Y_{(2)j} = \hat{y}_{(2)j}, Y_{(1)} + Y_{(2)} = T_2)
\]

\[
\approx \frac{\sum_{b=1}^B Pr(\tilde{\mathbf{Y}} | \mathbf{P}_b)Pr(\mathbf{P}_b | Y_{(1)j} = \hat{y}_{(1)j}, Y_{(2)j} = \hat{y}_{(2)j}, \mathbf{Y}_{-(1,2)})/Pr(\mathbf{P}_b | \mathbf{Y})}{\sum_{b=1}^B Pr(Y_{(1)j} = \hat{y}_{(1)j}, Y_{(2)j} = \hat{y}_{(2)j}, \mathbf{Y}_{-(1,2)})/Pr(\mathbf{P}_b | \mathbf{Y})}
\]

\[
\propto \frac{\sum_{b=1}^B Pr(\tilde{\mathbf{Y}} | \mathbf{P}_b)Pr(Y_{(1)j} = \hat{y}_{(1)j}, Y_{(2)j} = \hat{y}_{(2)j} | \mathbf{P}_b)/Pr(Y_{(1)j} = y_{(1)j}, Y_{(2)j} = y_{(2)} | \mathbf{P}_b)}{\sum_{b=1}^B Pr(Y_{(1)j} = \hat{y}_{(1)j}, Y_{(2)j} = \hat{y}_{(2)} | \mathbf{P}_b)/Pr(Y_{(1)j} = y_{(1)j}, Y_{(2)j} = y_{(2)} | \mathbf{P}_b)}.
\]

(2.15)

The key difference between (2.15) and (2.11) is the role of the missing values, \((Y_{(1)j'}, Y_{(2)j'})\) for all \(j' \neq j\). In (2.15), they do not affect \(Pr(\mathbf{P}_b | Y_{(1)j} = \hat{y}_{(1)j}, Y_{(2)j} = \hat{y}_{(2)j})\).
\( \hat{y}_{(2)j} \), so that this probability is determined entirely by \((\hat{y}_{(1)j}, \hat{y}_{(2)j})\). This is not the case in (2.11), where the variables \( j' \neq j \) serve to downweight the probability of components that would generate implausible values of \((y_{(1)j'}, y_{(2)j'})\). In settings with highly correlated variables, Scenario I and Scenario II should result in similar approximations to the posterior probabilities—assuming the same uniform prior distributions in each scenario—for \( y_{(1)j}, y_{(2)j} \), particularly when they are large compared to the rest of the data. In those cases, \( Pr(\mathcal{P}_b \mid Y_{(1)j} = \hat{y}_{(1)j}, Y_{(2)j} = \hat{y}_{(2)j}) \) is very similar regardless of whether or not we use the variables with \( j' \neq j \). Indeed, we find this to be the case in the empirical applications of Section 2.5.

2.4.3 Scenario III — Intruder Knows Only Second Largest

We now consider a scenario where the individual with the second largest value of some variable \( j \) uses \( \tilde{\mathcal{Y}} \) to learn about \( y_{(1)j} \). When \( T_j \) is available, we assume that the intruder seeks to compute

\[
Pr(Y_{(1)j} = \hat{y}_{(1)j} \mid \tilde{\mathcal{Y}}, y_{(2)}, Y_{(1)j} \geq y_{(2)j}, T_j) \\
\times Pr(\tilde{\mathcal{Y}} \mid Y_{(1)j} = \hat{y}_{(1)j}, y_{(2)}, T_j) Pr(Y_{(1)j} = \hat{y}_{(1)j} \mid y_{(2)}, Y_{(1)j} \geq y_{(2)j}, T_j).
\]

(2.16)

On the surface, this computation involves averaging over all unknown values, i.e., \( \mathcal{Y}_{-(1,2)} \) and values of \( y_{(1)} \) other than variable \( j \), which is computationally prohibitive. However, we can avoid this averaging by taking advantage of a common feature of the synthesis procedures: records that are not in the latent class occupied by the target record with \( y_{(1)j} \) are irrelevant for generating the synthetic \( \tilde{y}_{(1)j} \). Hence, we can disregard them when computing (2.16).

To take advantage of this feature, the intruder needs to know the members of the component occupied by the target record across each synthetic dataset. Given only \( y_{(2)} \) this is not possible. However, the intruder can use \( \tilde{\mathcal{Y}} \) to approximate these sets of records. For each synthetic dataset \( \tilde{\mathcal{Y}}^{(l)} \), the intruder first identifies the record with the largest synthetic value, \( \tilde{y}_{(1)j}^{(l)} \). For \( l = 1, \ldots, L \), the intruder then finds the maximum likelihood estimates of the Poisson mixture model fit to \( \tilde{\mathcal{Y}}^{(l)} \), and
determines the set of records belonging to the same component as \( \tilde{y}_{(1)j}^{(l)} \). Let this set be \( S^{(l)} \) comprising \( n_{S}^{(l)} \) records, and let \( T_{S_j}^{(l)} = \sum_{i \in S^{(l)}} \tilde{y}_{ij}^{(l)} \) be the total of the synthetic values in \( S^{(l)} \). Using this simplification, we have

\[
Pr(Y_{(1)j} = \hat{y}_{(1)j}^{(l)} \mid \mathbf{Y}, y_{(2)}, Y_{(1)j} \geq y_{(2)j}, T_j) \\
\propto \left[ \Pi_{i=1}^{L} Pr(\{\tilde{y}_{ij}^{(l)} : i \in S^{(l)}\} \mid Y_{(1)j} = \hat{y}_{(1)j}^{(l)}, y_{(2)}, T_{S_j}^{(l)}) \right] \\
Pr(Y_{(1)j} = \hat{y}_{(1)j} \mid y_{(2)}, Y_{(1)j} \geq y_{(2)j}, T_{S_j}^{(l)}).
\tag{2.17}
\]

For the MM synthesizer, \( Pr(\{\tilde{y}_{ij}^{(l)} : i \in S^{(l)}\} \mid Y_{(1)j} = \hat{y}_{(1)j}^{(l)}, y_{(2)}, T_{S_j}^{(l)}) \) is computed from the relevant multinomial distribution. When the target record is in a nonsingleton component according to the MLE with \( \mathbf{Y}^{(l)} \), we compute the mass associated with \( \{\tilde{y}_{ij}^{(l)} : i \in S^{(l)}\} \) from the multinomial distribution with sample size equal to \( T_{S_j}^{(l)} \) and \( n_{S}^{(l)} \) levels with all probabilities equal to \( 1/n_{S}^{(l)} \). When the target is a singleton according to the MLE with \( \mathbf{Y}^{(l)} \), we mimic the MM synthesizer step and combine it with all other singleton components. Here, we use a multinomial distribution with sample size equal to \( T_{S_j}^{(l)} \) and \( n_{S}^{(l)} \) levels with probabilities of each component determined according to the procedure in Step 3.2, using the result of the model estimated with \( \tilde{\mathbf{Y}}^{(l)} \). For the TCMM synthesizer, when the target record is part of the components that would be collapsed, we redefine \( S \) to include all records in the collapsed component. We use the multinomial distribution with sample size equal to \( T_{S_j}^{(l)} \) and \( n_{S}^{(l)} \) levels with all probabilities equal to \( 1/n_{S}^{(l)} \). Otherwise, we follow the strategies for the MM synthesizer.

When \( n_{S}^{(l)} > 1 \) for every \( l = 1, \ldots, L \), or when the target record is part of the components that would be collapsed by the TCMM synthesizer, one need not compute \( \Pi_{i=1}^{L} Pr(\{\tilde{y}_{ij}^{(l)} \in S^{(l)}\} \mid Y_{(1)j} = \hat{y}_{(1)j}^{(l)}, y_{(2)}, T_{S_j}^{(l)}) \). In these cases, the term is constant for all \( \hat{y}_{(1)j} \), so that it cancels after normalizing (2.17). When \( n_{S}^{(l)} = 1 \) in the MM synthesizer, and additionally when the target record is not collapsed in the TCMM synthesizer, we simplify computation by instead using \( \Pi_{i=1}^{L} Pr(\{\tilde{y}_{ij}^{(l)} \mid Y_{(1)j} = \hat{y}_{(1)j}^{(l)} \cap Y_{(1)j} \geq y_{(2)j}, T_{S_j}^{(l)} \}
\]
\( \hat{y}_{(1)j}, y_{(2)}, T_{S_j}^{(l)} \), thereby avoiding averaging over the unknown true values among all singletons. This simplification does not diminish the accuracy of the estimation noticeably.

To compute \( Pr(Y_{(1)j} = \hat{y}_{(1)j} \mid y_{(2)}, Y_{(1)j} \geq y_{(2)j}, T_{S_j}^{(l)} \)\), we have to assume an intruder’s prior distribution. As described in Section 2.4.1, one choice is the uniform distributions on all \( \hat{y}_{(1)j} > y_{(2)j} \). However, intruders can use the information in \( \tilde{Y} \), and more specifically in the components corresponding to all \( S^{(l)} \), for more accurate prior guesses. Let \( \theta_{(1)j}^{(l)} \), where \( l = 1, \ldots, L \), be the Poisson rate of the component corresponding to \( S^{(l)} \). We approximate the probability as

\[
Pr(Y_{(1)j} = \hat{y}_{(1)j} \mid y_{(2)}, Y_{(1)j} \geq y_{(2)j}, n_{S^{(l)}}) = (1/L) \sum_{l=1}^{L} \text{Pois}(Y_{(1)j} = \hat{y}_{(1)j} \mid \theta_{(1)j}^{(l)}, Y_{(1)j} \geq y_{(2)j}, n_{S^{(l)}}).
\] (2.18)

Here, we do not condition on \( T_{S_j}^{(l)} \), since the mixture Poisson model does not fix totals. We note that \( \Pi_{i=1}^{L} Pr(\{\tilde{y}_{ij}^{(l)} : i \in S^{(l)}\} \mid Y_{(1)j} = \hat{y}_{(1)j}, Y_{(2)}, T_{S_j}^{(l)}) \) accounts for totals, so that values \( \hat{y}_{(1)j} \) that are not plausible (e.g., that exceed some \( T_{S_j}^{(l)} \)) have zero probability. When \( n_{S^{(l)}} = 1 \), we assume that \( \theta_{(1)j}^{(l)} \approx \hat{y}_{(1)j} \). When \( n_{S^{(l)}} > 1 \), as is possible with MM synthesizers and likely with TCMM synthesizers, we use \( \theta_{(1)j}^{(l)} \approx T_{S^{(l)j}}/n_{S^{(l)}} \). The probability can be computed analytically or via simulation, e.g., repeatedly simulate draws from a Poisson distribution with parameter \( \theta_{(1)j}^{(l)} \), and count the percentage of times that the maximum value equals \( \hat{y}_{(1)j} \).

The expression in (2.16) suggests the opportunity for a potentially disclosive attack on data generated from the MP synthesizer. When some \( S^{(l)} \) includes only \( (\tilde{y}_{(1)j}, \tilde{y}_{(2)j}) \)—which can happen since the Poisson mixture model tends to allocate observations with similar values to the same cluster—the intruder simply computes \( T_{S^{(l)}} \) and subtracts \( y_{(2)j} \) to arrive at \( y_{(1)j} \). Because of the potential of this attack, the MP synthesizer may not be sufficiently protective against the Scenario III attack.
When $T_j$ is not available and data are generated from the MP synthesizer, the intruder can use (2.18) directly to estimate the risk. Here, $\hat{\theta}^{(d)}_{(1),j}$ is estimated by fitting the mixture of Poisson model on $\hat{Y}^{(d)}$.

2.5 Empirical Illustration

We now illustrate the synthesizers using data on business establishments. The data come from a survey of business establishments in the country Colombia taken in the year 1977. These data are available for public use, which is unusual for establishment data (and hence why we use them). The data were utilized previously by (Kim et al., 2014) to illustrate multiple imputation techniques in different contexts and with different models than we consider here. We use data comprising $n = 1051$ food manufacturing establishments measured on $p = 4$ magnitude variables: the number of skilled ($\text{skl}$) and unskilled laborers ($\text{unskl}$) and their corresponding wages ($\text{skw}$ and $\text{unskw}$). There are no missing values.

As evident in Figure 2.1, all four variables are right-skewed. Indeed, for each variable, the maximum value is more than 40 times larger than the third quartile. In particular, $\text{skw}$ and $\text{unskw}$ have large values relative to the other values in the dataset. The variables have high, positive correlations.

We estimate the mixture of Poisson distributions model using $K = 100$ components, although far fewer are actually occupied in any MCMC run. We run the MCMC chain for 30,000 iterations. We check convergence of the chain by examining the trace plots, autocorrelations, and mixing properties for $\pi$, $\mu$, and $\Sigma$. We generate $L = 10$ sets of component assignments and parameter values by running the chain another 100 times, using every tenth iteration. We use these sets to generate $L = 10$ synthetic datasets for each synthesizer. For the TCMM, we use $q \in \{.997, .995, .993, .991\}$ for every variable, representing increasing amounts of collapsing.
Figure 2.1: Scatter plots of the observed values in the Columbia manufacturing survey data.

2.5.1 Utility Analysis

We begin with evaluations of the analytic usefulness of the synthetic datasets generated by each method. To do so, we compare representative analyses performed on the synthetic and original datasets. We focus on marginal distributions, correlations, and regression relationships.

Marginal Distributions

Figure 2.2 displays the marginal distributions for the original data and one randomly selected synthetic dataset for each synthesizer. Plots for other synthetic datasets have similar shapes. From a utility perspective, the MP and MM synthesizers are remarkably effective at preserving the marginal distributions. In contrast, by design the TCMM synthesizer shrinks the largest values thereby distorting the tails of the distribution, with greater degree of truncation as $q$ decreases. The distributions of
Table 2.1: Synthetic and original values of the two largest values of \( skw \) for each synthesizer. Synthetic data results are averages and standard deviations across the \( L \) datasets.

<table>
<thead>
<tr>
<th></th>
<th>1st Largest</th>
<th>2nd Largest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>104246</td>
<td>87315</td>
</tr>
<tr>
<td>MP</td>
<td>104336(326)</td>
<td>86844(284)</td>
</tr>
<tr>
<td>MM</td>
<td>104286(247)</td>
<td>86984(355)</td>
</tr>
<tr>
<td>TCMM with ( q = .997 )</td>
<td>66897(145)</td>
<td>66729(93)</td>
</tr>
<tr>
<td>TCMM with ( q = .995 )</td>
<td>51133(149)</td>
<td>50999(63)</td>
</tr>
<tr>
<td>TCMM with ( q = .993 )</td>
<td>49044(137)</td>
<td>48926(62)</td>
</tr>
<tr>
<td>TCMM with ( q = .991 )</td>
<td>39974(113)</td>
<td>398797(98)</td>
</tr>
</tbody>
</table>

the synthesized top two values for \( skw \) are summarized in Table 2.1.

Correlations

For each \( \tilde{Y}^{(l)} \), we compute the six pair-wise correlations for each synthesizer. For each correlation in each scenario, we compute the average and variance across the \( L \) datasets. Figure 2.3 displays density plots of \( t \)-distributions with \( L - 1 \) degrees of freedom, with first and second moments equal to each correlation’s average and standard deviation. These portray values of the correlations that are likely to result from the synthesizer. In general, the correlations from the synthetic data are reasonably close to those from the original data, with differences tending to be .05 or smaller. As expected, the MP and the MM synthesizers respect the correlation structure more faithfully than TCMM does. For the TCMM synthesizer, the relationship between \( q \) and the quality of the correlation estimates is not monotonic. As \( q \) becomes smaller (collapsing more establishments), the correlation between \( skl \) and \( skw \) loses accuracy. However, this is not the case for the estimated correlation between \( skl \) and \( unskw \).

Linear Regression Coefficients

We next estimate a linear regression, predicting the number of unskilled laborers from main effects on the other variables. We first transform each variable using

\[
y_{i,j}^* = \frac{\log(\sqrt{y_{i,j}} + 1) - \text{mean}(\log(\sqrt{y_{i,j}} + 1))}{\text{sd}(\log(\sqrt{y_{i,j}} + 1))}
\]

This transformation makes the distributions...
of the values more symmetric. We estimate the linear regression:

\[ unnkl_i^* = \beta_0 + \beta_1 skl^* + \beta_2 (skl^*)^2 + \beta_3 skw^* + \beta_4 (skw^*)^2 + \beta_5 unnkw^* + \beta_6 (unnkw^*)^2 + \epsilon_i; \quad \epsilon_i \sim N(0, \sigma^2). \]  

(2.19)

In the original data, the \( R^2 \approx 0.89 \). The normality assumption is somewhat dubious here due to the outliers even on the log scale. However, our purpose is to compare model estimates based on the actual and synthetic data, so that the quality of the model fit suffices for our purposes.

For each \( \tilde{Y}^{(l)} \), we estimate \( (\beta_1, \ldots, \beta_6) \) for each synthesizer. For each coefficient in each scenario, we compute the average and variance across the \( L \) datasets. Figure 2.3 displays density plots of \( t \)-distributions with \( L - 1 \) degrees of freedom, with first and second moments equal to each coefficient’s average and standard deviation. Here, the MP and MM synthesizers result in reasonably accurate estimates of the coefficients. However, the TCMM synthesizer results are quite inaccurate for some coefficients, even causing them to switch signs. Apparently, collapsing the tails seriously degrades this regression analysis.

### 2.5.2 Disclosure Risk Assessment

To illustrate the disclosure risk computations, we assume the intruder seeks to use \( \tilde{Y} \) to estimate the largest value of \( skl \) in the original data. This value \( y_{(1)1} = 414 \). We present results for Scenario I and Scenario III for each synthesizer; Scenario II results were similar to those in Scenario I. Throughout, we assume that the intruder knows which synthesizer was applied to create \( \tilde{Y} \). For all risk measures, we use uniform prior distributions for \( Y_{(1)1} \), allowing the synthetic data essentially to determine the posterior risk. We note that the second largest value of \( skl \) is \( y_{(2)1} = 378 \).

After estimating the model on \( Y \), the record with the largest value of \( skl \) is frequently in a singleton component across draws of the MCMC. For the synthetic datasets we generated, it is almost always combined with other singletons in the MM synthesizer. For the TCMM synthesizer, this record is collapsed with other records in the tails for all values of \( q \).
Results for Scenario I

For the MP synthesizer, we assume that the intruder knows $y_{(2)1}$ but does not know $T_1$. Theoretically $Y_{(1)1}$ has no finite upper bound. However, we assume an intruder interested in computing posterior probabilities in the range $[378, 496]$, which we obtain from $[\max(y_{(2)1}, 8y_{(1)1}), 1.2y_{(1)1}]$. Figure 2.5 displays both the prior and posterior disclosure risks for all $Y_{(1)1} \in [378, 496]$. The value $\hat{y}_{(1)1} = 424$ has the maximum probability at .061. The true $\hat{y}_{(1)1} = 414$ has the probability .019, and falls inside 95% predictive risk interval. The prior risk distribution in Figure 2.5 strongly favors values as small as possible. Hence, we conclude from these results that, under the strong assumption, the synthetic data provide a substantial amount of information about the location of $y_{(1)1}$.

For the MM and TCMM synthesizers, we assume that the intruder knows $Y_{(-1,2)j}$ and $T_1$. Since by assumption the intruder knows the third largest value of $skl$, $y_{(3)1}$, the intruder knows that $T_{21}/2 \leq Y_{(1)1} \leq T_{21} - y_{(3)1}$, so that $396 \leq Y_{(1)1} \leq 423$. Hence, simply releasing $T_j$ already substantially constrains $\hat{y}_{(1)1}$ under Scenario I.

Figure 2.6 displays both the prior risks and posterior risks for the MM synthesizer for all $Y_{(1)1} \in [396, 496]$. The value $\hat{y}_{(1)1} = 407$ has the maximum probability at .042. The true $\hat{y}_{(1)1} = 414$ has probability .037. The prior risk distribution favors the largest values of $\hat{y}_{(1)1}$, with the prior risk probability at $\hat{y}_{(1)1} = 414$ nearly 1/4 as large as the probability at $\hat{y}_{(1)1} = 423$. Apparently, the synthetic data dramatically shift the intruder’s risk probabilities. However, it is not obvious that the intruder seeking an accurate estimate of $y_{(1)j}$ is much better off for using $\hat{Y}$ under Scenario 1; indeed, the prior risk at $\hat{y}_{(1)1} = 414$ is also .037.

Figure 2.7 displays the prior risks and the posterior risks under the TCMM synthesizer for all $Y_{(1)1} \in [396, 496]$. Since we compute the prior risk using the informative prior distribution, the prior and posterior risk are exactly the same. Compared with the MM synthesizer, the posterior risk at the true $y_{(1)1} = 414$ is smaller under the TCMM synthesizer. Because of the fixed marginal totals and the intruder’s prior
knowledge, in Scenario I the posterior risk for TCMM does not depend on the size of $q$ (provided that the top two establishments are in the collapsed component).

### 2.5.3 Results for Scenario III

In this scenario, the intruder now knows only that $Y_{(1)1} > y_{(2)1} = 378$. We assume that the intruder is particularly interested in estimating the chance that $378 < Y_{(1)j} \leq 496$. Table 2.2 shows these probabilities under Scenario III for the different synthesizers. The risks under both the MP and the MM synthesizers are high. This is because the largest case is in a singleton component, and the released data provide enough information about the Poisson rate in this component to enable accurate estimation. For the TCMM synthesizers, the synthetic data provide little information about the Poisson rate for the singleton component, because the collapsing has disguised that rate effectively. As such, the risks are low.

### 2.6 Discussion

The empirical illustration suggests several conclusions about the different synthesizers. First, without any collapsing of cases, the mixture of Poissons model can result in accurate estimates of the joint distribution of the variables, resulting in synthetic data with high utility. However, for outliers, draws from the model can be too close to the original values, leading to potentially unacceptable disclosure risks. Because synthetic values are sampled from Poisson distributions within each cluster, in any synthetic dataset the values for some variable $j$ are drawn with standard deviation equal to the square root of the mean of variable $j$. For records in singleton clusters with extreme values of variable $j$, this may not introduce sufficient noise to the original value. As such, we do not recommend the MP synthesizer for data with extreme outliers. In this case, one potential solution is to perturb outlying values.
first so as to protect their confidentiality, then synthesize the resulting data; similar ideas are discussed by (Yang et al., 2014). Second, when agencies want to release marginal totals from the confidential data along with synthetic microdata, the MM synthesizer can provide high utility and ensure that the synthetic and original totals match. Once again, however, by reproducing the tail reasonably well, the MM synthesizer may not offer sufficient protection. Thus, we also do not recommend the MM synthesizer for data with extreme outliers. Third, the TCMM synthesizer can estimate the bulk of the distribution quite well, but by design it sacrifices information about the tails of the distributions. This was the only synthesizer with potentially acceptable disclosure risks in the presence of outliers.
Figure 2.2: Boxplots of $skl$, $unsk$, $skw$, $unskw$. In each plot, the columns are arranged from the left to the right in the order of original data, the MP, the MM, the TCMM with $q = .997$, .995, .993, and .991.
Figure 2.3: Density plots of the synthetic pair-wise correlations for different synthesizers. Each vertical line is the value of the correlation from the original data.
Figure 2.4: Density plots of the synthetic estimates of $(\beta_1, \ldots, \beta_6)$ for different synthesizers. Each vertical line is the value of the coefficient from the original data.
Figure 2.5: The prior and the disclosure risk of the top largest $skl$ under the MP synthesizer for the strong assumption.

Figure 2.6: The prior and the disclosure risk of the top largest $skl$ under the MP synthesizer for the strong assumption.
Figure 2.7: The prior and the posterior risk for the record with the largest value of $skl$ for the TCMM synthesizer under Scenario I.
3.1 Introduction

To protect continuous microdata from being disclosed, data collectors have to distort them. There are several disclosure limitation techniques used for distortion, including noise addition (Kim, 1986; Fuller, 1993), data distortion by probability distribution (Liew et al., 1985), resampling, microaggregation (Defays and Nanopoulos, 1993; Domingo-Ferrer and Mateo-Sanz, 2002), lossy compression (Domingo-Ferrer and Torra, 2001a), rank swapping (Moore Jr, 1996), and multiple imputation (MI) (Rubin, 1987).

MI (Rubin, 1987) is a model-based approach which samples and releases multiple fully synthetic datasets from the model fitted on the confidential data. The virtue of MI is that its syntheses can represent the distribution of the original microdata. It has been applied by statisticians to produce detailed public data without compromising confidentiality (Rubin, 1993; Little, 1993; Raghunathan et al., 2003; Reiter, 2003, 2004; Reiter and Raghunathan, 2007; Wang and Reiter, 2012; Hu et al., 2014; Paiva
et al., 2014). However, a standard MI may not be adequate when the continuous microdata has extreme values, which may influence the released distribution.

In this chapter, we explore some non-standard MI methods for protecting continuous microdata with outliers. A common feature of these non-standard methods is that they use protection intervals to add an extra layer of protection. These protection intervals are determined by data collectors and can vary for different individuals. We conjecture that data collectors sample model parameters of the synthesis model by drawing them conditioned on the protection intervals. Alternatively, they can restrict the released synthetic data by sampling them from truncated distributions.

Although finding the optimal mechanism of generating the protection intervals is not the main focus of this study, it is important in reducing the disclosure risk for a given MI approach. Basically, the locations and the widths of the protection intervals can directly affect the degree of the distortion to the original data. In addition, how the intervals relate to the truth can impact the disclosure risk in a more complicated way. In this study, we briefly investigate the width of the additive and multiplicative intervals both of which cover the original data.

To assess the performance of these methods, risk metrics are needed. Typically for the disclosure control techniques involving the protection intervals, the risk measurement is to check the probability of the original value falls within an interval (Domingo-Ferrer and Torra, 2001b). This interval is centered on the corresponding masked value. The width of the interval is usually estimated according to the rank of the protected variable or to its standard deviation. Obviously, if the interval is not centered, this measurement will be less accurate. We define the disclosure risk using the Bayesian posterior distribution of the true data values (Hu et al., 2014; Reiter et al., 2014; Paiva et al., 2014). This metric allows us to incorporate the intruder’s prior knowledge of the original dataset and the data synthesis procedure.

The remainder of this chapter is structured as follows, Section 3.2 provides the
Table 3.1: Methods of releasing the synthetic data $Y$ for the true data $X$. Here, $D = \{d_i = (a_i, b_i) : i = 1, \ldots, n\}$ is the collection of protection intervals for each observation $X_i = x_i$. And $W = \{w_i : i = 1, \ldots, n\}$ is the collection of indicators such that $w_i = 1$ if $a_i \leq x_i \leq b_i$ and $w_i = 0$ otherwise.

| Draw $Y$ from | Draw $\theta$ from | $f(\theta | X)$ | $f(\theta | D, W)$ |
|---------------|-------------------|----------------|-----------------|
| $f(Y; \theta)$ | Method 1          | Method 3       |
| $f(Y; \theta, D, W)$ | Method 2          | Method 4       |

interval-protected MI methods together with the posterior risk metrics for a single continuous variable. Section 3.3 extends these approaches together with the risk metrics to two continuous variables. Section 3.4 describes a way to generate multiplicative intervals based on a fuzzy noise factor (Abowd et al., 2005). Section 3.5 concludes this study.

3.2 Study with a single variable

Suppose the data comprise a univariate continuous variable $X$ measured on $n$ individuals which can be fitted by a parametric model $f(X; \theta)$. Suppose the data collector specifies a range $(a_i, b_i)$ for each $X_i$. We let $D$ be the set of all the protection intervals $d_i = (a_i, b_i)$ and $W$ be the set of all the protection indicators $w_i$ such that $w_i = 1$ if $a_i \leq X_i \leq b_i$ and $w_i = 0$ otherwise. We define several synthesis techniques that use $D$ and $W$. There are summarized in Table 3.1. For the later simulations, we always allow $w_i = 1$ so that we can focus on the influence of the width of the protection intervals on both data utilities and posterior disclosure risks.

Method 1 is a standard approach to release synthetic data. That is, it estimates the model parameters $\theta$ from $X$. Then, it samples $Y$ based on this model. Method 2 is similar to Method 1 except that it draws each $Y_i$ from the parametric model within the range of $(a_i, b_i)$. Thus, Method 2 might be able to provide some additional protection for $X$ by forcing each synthetic $Y_i$ to lie within its protection interval, although as we shall see this additional protection is not strong when $w_i = 1$. Method
3 adds extra uncertainty for $\theta$ by inferring it from the protection intervals. By doing so, the estimated $\theta$ may not overly disclose a particular $X_i$ beyond what is learned from $(a_i, b_i)$. Then, the synthetic $Y$ released from the parametric model can be less informative about the true $X$. Method 4 restricts the synthetic data by the protection intervals, in addition to sampling $\theta$ conditioned on the intervals.

3.2.1 Posterior disclosure risk

Assume that an intruder wants to make disclosures about some values in a collected data $X = (X^{(nr)}, X^{(r)})$. Here, we let $X^{(nr)}$ be the data known to intruders, and we let $X^{(r)}$ be the data unknown to intruders, which are the targets of the intruder’s attacks. Then, the prior disclosure risk is the intruder’s knowledge of $X^{(r)}$ given $X^{(nr)}$. The posterior disclosure risk is the density of guessing the true $X^{(r)}$ given the intruder’s prior knowledge and the synthetic data $Y$. In the later simulations, we allow $X^{(r)}$ to be a single value and $X^{(nr)}$ include all but the single $X^{(r)}$.

In the following, we introduce the posterior predictive density of each method for the situation when $(D, W)$ are known or unknown, respectively. And we assume the prior disclosure risk is uniformly distributed, which means the intruder does not have any specific preference about how $X^{(r)}$ should look like.
For Method 1, the posterior predictive density of guessing $X^{(r)}$ is

\[
f(X^{(r)}|X^{(nr)}, Y, D, W) = f(X^{(r)}|X^{(nr)}, Y) \\
\propto \int f(Y|X^{(r)}, X^{(nr)}, \theta) f(\theta|X^{(r)}, X^{(nr)}) d\theta \cdot f(X^{(r)}|X^{(nr)}) \\
\propto \int f(Y|X^{(r)}, X^{(nr)}, \theta) f(\theta|X^{(r)}, X^{(nr)}) d\theta \\
\propto \int f(Y; \theta) f(\theta|X^{(r)}, X^{(nr)}) d\theta \\
\propto \sum_{b=1}^{B} f(Y; \theta^{(b)})/B, \quad \text{(3.1)}
\]

in which we assume the prior disclosure risk density $f(X^{(r)}|X^{(nr)})$ is uniformly distributed over all the possible values for $X^{(r)}$. Here, \{\theta^{(b)} : b = 1, \ldots, B\} are independently and identically distributed (i.i.d.) from $f(\theta|X^{(r)}, X^{(nr)})$. Thus, when the guessed value of $X^{(r)}$ is close to the true $X$, the draws of $\theta^{(b)}$ are more similar to those based on the true $X$. This can make the posterior predictive density around the true $X^{(r)}$ large.

For Method 2, the posterior predictive density of guessing $X^{(r)}$ is

\[
f(X^{(r)}|X^{(nr)}, Y, D, W) \propto \int f(Y|X^{(r)}, X^{(nr)}, \theta, W, D) f(\theta|X^{(r)}, X^{(nr)}, D, W) d\theta \\
\propto \int f(Y|X^{(r)}, X^{(nr)}, \theta, D, W) f(\theta|X^{(r)}, X^{(nr)}, D, W) d\theta \\
\propto \int f(Y; \theta, D, W) f(\theta|X^{(r)}, X^{(nr)}) d\theta \\
\propto \sum_{b=1}^{B} f(Y; \hat{\theta}_b, D, W)/B, \quad \text{(3.2)}
\]
where the prior disclosure risk density $f(X^{(r)}|X^{(nr)}, D, W)$ is uniformly distributed on the support of $D^{(r)}$ if $W^{(r)} = 1$ or $(D^{(r)})^c$ if $W^{(r)} = 0$. The model parameters $\{\hat{\theta}_b : b = 1, \ldots, B\}$ are i.i.d. draws from $f(\theta|X^{(r)}, X^{(nr)})$.

For Method 3, the posterior predictive density of guessing $X^{(r)}$ is

$$f(X^{(r)}|X^{(nr)}, Y, D, W) \propto \int f(Y|X^{(r)}, X^{(nr)}, \theta, D, W) f(\theta|X^{(r)}, X^{(nr)}, D, W) d\theta \cdot f(X^{(r)}|X^{(nr)}, D, W)$$

$$\propto \int f(Y|X^{(r)}, X^{(nr)}, \theta, D, W) f(\theta|X^{(r)}, X^{(nr)}, D, W) d\theta$$

$$\propto \int f(Y; \theta) f(\theta|D, W) d\theta$$

$$\propto \sum_{b=1}^{B} f(Y; \hat{\theta}_b)/B,$$  \hfill (3.3)

in which the prior disclosure risk density is the same as that for Method 2. However, $\{\hat{\theta}_b : b = 1, \ldots, B\}$ are i.i.d. draws from $f(\theta|D, W)$. Since changing the guesses of $X^{(r)}$ does not affect how $\hat{\theta}_b$ is drawn when $(D, W)$ are known, (3.3) is a uniform distribution on the same support as its prior disclosure risk density.

For Method 4, the posterior predictive density of guessing $X^{(r)}$ is

$$f(X^{(r)}|X^{(nr)}, Y, D, W) \propto \int f(Y|X^{(r)}, X^{(nr)}, \theta, D, W) f(\theta|X^{(r)}, X^{(nr)}, D, W) d\theta \cdot f(X^{(r)}|X^{(nr)}, D, W)$$

$$\propto \int f(Y|X^{(r)}, X^{(nr)}, \theta, D, W) f(\theta|X^{(r)}, X^{(nr)}, D, W) d\theta$$

$$\propto \int f(Y; \theta, D, W) f(\theta|D, W) d\theta$$

$$\propto \sum_{b=1}^{B} f(Y; \hat{\theta}_b, D, W)/B,$$  \hfill (3.4)

where the prior disclosure risk density is also the same as that in Method 2. Each
\( \hat{\theta}_b \) is independently and identically drawn from \( f(\theta|D,W) \) for \( b = 1, \ldots, B \). Due to the same reason for Method 3, \( f(X^{(r)}|X^{(nr)}, Y, D, W) \) is a uniform distribution on the same support as its prior disclosure risk density.

\( (D,W) \) are unknown to the intruder

The posterior predictive density under Method 1 in (3.1) does not depend on \( D,W \).

Thus, we only introduce the posterior predictive densities for the other three methods in this section. We assume the prior disclosure risk density \( f(X^{(r)}|X^{(nr)}) \) is uniformly distributed over all the possible values of \( X^{(r)} \).

For Method 2, the posterior predictive density of guessing \( X^{(r)} \) is

\[
\begin{align*}
 f(X^{(r)}|X^{(nr)}, Y) & \propto \int_W \int_D \int_\theta f(Y|X^{(r)}, X^{(nr)}, \theta, D, W) f(\theta, D, W|X^{(r)}, X^{(nr)}) d\theta dD dW \\
 & \propto \int_W \int_D \int_\theta f(Y|X^{(r)}, X^{(nr)}, \theta, D, W) f(\theta, D, W|X^{(r)}, X^{(nr)}) d\theta dD dW \\
 & \propto \int_W \int_D \int_\theta f(Y; \theta, D, W) f(\theta, D, W|X^{(r)}, X^{(nr)}) d\theta dD dW \\
 & \propto \sum_{b=1}^B f(Y; \hat{\theta}_b, \hat{D}_b, \hat{W}_b)/B, \quad (3.5)
\end{align*}
\]

in which \( \{(\hat{\theta}_b, \hat{D}_b, \hat{W}_b) : b = 1, \ldots, B \} \) are i.i.d. draws from \( f(\theta, W, D|X^{(r)}, X^{(nr)}) \).

One can sample them by a Bayesian approach. However, the full distribution of \( \theta \) only depends on \( (X^{(r)}, X^{(nr)}) \) because we assume the intruder samples \( \theta \) from \( f(\theta|X) \) in the same way as how the agent does in Method 2. Thus, the main difference between (3.2) and (3.5) is whether the intruder needs to estimate \( (D,W) \) or not.
For Method 3, the posterior predictive density of guessing $X^{(r)}$ is

$$f(X^{(r)}|X^{(nr)}, Y) \propto \int_{W} \int_{D} \int_{\theta} f(Y|X^{(r)}, X^{(nr)}, \theta, D, W) f(\theta, D, W|X^{(r)}, X^{(nr)}) d\theta dDdW$$

$$\cdot f(X^{(r)}|X^{(nr)})$$

$$\propto \int_{W} \int_{D} \int_{\theta} f(Y|X^{(r)}, X^{(nr)}, \theta, D, W) f(\theta, D, W|X^{(r)}, X^{(nr)}) d\theta dDdW$$

$$\propto \int_{W} \int_{D} \int_{\theta} f(Y; \theta) f(\theta, D, W|X^{(r)}, X^{(nr)}) d\theta dDdW$$

$$\propto \sum_{b=1}^{B} f(Y; \hat{\theta}_b)/B,$$

(3.6)

where $\{\hat{\theta}_b := 1, \ldots, B\}$ are i.i.d. draws from $f(\theta|X^{(r)}, X^{(nr)}, D, W)$.

For Method 4, the posterior predictive density of guessing $X^{(r)}$ is

$$f(X^{(r)}|X^{(nr)}, Y) \propto \int_{W} \int_{D} \int_{\theta} f(Y|X^{(r)}, X^{(nr)}, \theta, D, W) f(\theta, D, W|X^{(r)}, X^{(nr)}) d\theta dDdW$$

$$\cdot f(X^{(r)}|X^{(nr)})$$

$$\propto \int_{W} \int_{D} \int_{\theta} f(Y|X^{(r)}, X^{(nr)}, \theta, D, W) f(\theta, D, W|X^{(r)}, X^{(nr)}) d\theta dDdW$$

$$\propto \int_{W} \int_{D} \int_{\theta} f(Y; \theta, D, W) f(\theta, D, W|X^{(r)}, X^{(nr)}) d\theta dDdW$$

$$\propto \sum_{b=1}^{B} f(Y; \hat{\theta}_b, \hat{D}_b, \hat{W}_b)/B,$$

(3.7)

in which $\{ (\hat{\theta}_b, \hat{W}_b, \hat{D}_b) : b = 1, \ldots, B \}$ are i.i.d. draws from $f(\theta, W, D|X^{(r)}, X^{(nr)})$.

**Summary posterior risk metrics**

Based on the posterior predictive density, we derive posterior risk metrics including interval risk, relative risk, posterior mean, and relative MSE (mean squared error).

The interval risk is the probability of the posterior disclosure risk within the range
where $R$ is defined by the agency as a risky range. It is

$$
\int_R f(X^{(r)}) = \hat{x}^{(r)}|X^{(nr)}, W, D, Y) dX^{(r)}.
$$

The relative risk is the ratio of the posterior disclosure risk over the prior disclosure risk within the range $R$. It is

$$
\frac{\int_R f(X^{(r)}) = \hat{x}^{(r)}|X^{(nr)}, W, D, Y) dX^{(r)}}{\int_R f(X^{(r)}) = \tilde{x}^{(r)}|X^{(nr)}, W, D) dX^{(r)}}.
$$

In later simulations, we choose $R$ to be from 95% to 105% of the unknown individuals, although in principle one could use any range.

The posterior mean is

$$
E(\hat{X}^{(r)}) = \int_{X^{(r)}} \hat{x}^{(r)} f(X^{(r)} = \hat{x}^{(r)}|X^{(nr)}, W, D, Y) dX^{(r)},
$$

which is the mean of the posterior predictive density. Values near $X^{(r)}$ are considered risky.

And the relative MSE is:

$$
E((\hat{X}^{(r)} - x^{(r)})^2) / x^{(r)} = \int_{X^{(r)}} (\hat{x}^{(r)} - x^{(r)})^2 f(X^{(r)} = \hat{x}^{(r)}|X^{(nr)}, W, D, Y) dX^{(r)} / x^{(r)}
$$

which is similar to a second moment of the posterior predictive density. Values near zero are considered risky.

When $D$ and $W$ are unknown, we drop them in (3.8) to (3.11).

**3.2.2 Simulation study**

To illustrate these ideas, we use a simple simulation study. We generate the true $X$ from $\mathcal{N}(0, 1)$ with $n = 1,000$ observations. Then, we replace the largest value with $6\sigma = 6$, which represents a large outlier. However, we continue to base the synthetic
data model on normal distributions, treating the manufactured outlier as a regular
data point.

We allow each $X_i$ to be always within the protection intervals and the values of $(a_i, b_i)$ are simply taken as

$$
\begin{align*}
  a_i &= x_i - c_i^{(a)} \\
  b_i &= x_i + c_i^{(b)}
\end{align*}
$$

with

$$
\begin{align*}
  c_i^{(a)} &\sim U(0, g) \\
  c_i^{(b)} &\sim U(0, g).
\end{align*}
$$

One can control the parameter $g$ in (3.12) such that small $g$ allows $(a_i, b_i)$ to be informative about $x_i$. Contrarily, big $g$ makes $(a_i, b_i)$ be vague about it. Because of this, we would like to see how risky each method will be when $g = 0.1$ and $g = 0.5$.

To generate the model-based synthesis, we need to sample the model parameters $\theta = (\mu, \sigma^2)$ from $X$ or $(D, W)$. When we sample them from $X$, we apply the standard Bayesian approach with a Jeffreys prior $\sigma^{-2}$. The full conditional distributions of $(\mu, \sigma^2)$ in Method 1 and 2 are

$$
\begin{align*}
  \pi(\mu|\cdot) &= N(\mu; \sum_{i=1}^{n} x_i/n, \sigma^2/n) \\
  \pi(\sigma^2|\cdot) &= G^{-1}(\sigma^2; n/2, \sum_{i=1}^{n} (x_i - \mu)^2/2).
\end{align*}
$$

When we sample $(\mu, \sigma^2)$ from $(D, W)$ in Method 3 and 4, we apply the approach of data augmentation. To do so, we augment $(D, W)$ by a continuous random variable $\tilde{X}$ sampled from

$$
\begin{align*}
  f(\tilde{x}_i|\mu, \sigma^2, w_i, d_i) &= N(\tilde{x}_i; \mu, \sigma^2) \cdot \mathbb{I}_{(a_i, b_i)}(\tilde{x}_i) \cdot \mathbb{I}_{(a_i, b_i)^c}(\tilde{x}_i).
\end{align*}
$$

This means that if $w_i = 1$, $\tilde{x}_i$ follows $N(\tilde{x}_i; \mu, \sigma^2)$ truncated to the support of $(a_i, b_i)$. Otherwise, $\tilde{x}_i$ follows $N(\tilde{x}_i; \mu, \sigma^2)$ truncated to the support of $(-\infty, a_i) \cup (b_i, \infty)$. We use the method of inverse CDF according to the property of normal distribution,
sampling values $\tilde{x}_i$ for $i = 1, \ldots, n$ from $\tilde{x}_i = \mu + \sigma \Phi^{-1}(v)$ with $v$ as

$$v = \begin{cases} u(\Phi(\frac{b_i - \mu}{\sigma}) - \Phi(\frac{a_i - \mu}{\sigma})), & \text{for } w_i = 1 \\ c_i^{-1}u, & \text{for } w_i = 0 \ And \ u < c_i \Phi(\frac{a_i - \mu}{\sigma}) \\ c_i^{-1}u + \Phi(\frac{b_i - \mu}{\sigma}) - \Phi(\frac{a_i - \mu}{\sigma}), & \text{for } w_i = 0 \ And \ u > c_i \Phi(\frac{a_i - \mu}{\sigma}). \end{cases} \tag{3.16}$$

Here, $u \sim U(0, 1)$ and $c_i^{-1} = 1 - \Phi(\frac{b_i - \mu}{\sigma}) + \Phi(\frac{a_i - \mu}{\sigma})$.

We sample $(\mu, \sigma^2)$ from the full conditional distributions with the Jeffreys prior $\pi(\mu, \sigma^2) = \sigma^{-2}$ by using the augmented data,

$$(\mu|\sigma^2, \tilde{X}, W, D) \sim \mathcal{N}(\sum_{i=1}^n \tilde{x}_i/n, \sigma^2/n) \tag{3.17}$$

$$(\sigma^2|\mu, \tilde{X}, W, D) \sim \mathcal{G}^{-1}(n/2, \sum_{i=1}^n (\tilde{x}_i - \mu)^2/2). \tag{3.18}$$

We assume that the intruder knows every $X_i$ except the largest one $X_{[1]}$. He also knows each $w_i = 1$. The intruder has a knowledge of (3.12) in the sense that he knows how $(a_i, b_i, c_i^{(b)}_i, c_i^{(a)}_i)$ are generated and what the value of $g$ is. Moreover, we assume he knows how we sample the model parameters by the different methods. Therefore, when $D$ is known he can sample $(\mu, \sigma^2)$ according (3.13) to (3.14). When $D$ is known, in Method 2 he has

$$f(\theta, D|X^{(r)}, X^{(nr)}) \propto f(X^{(r)}, X^{(nr)}; \theta, D)\pi(\theta)\pi(D)$$

$$\propto f(X^{(r)}, X^{(nr)}; \theta)\pi(\theta)\pi(D)$$

$$\propto \left\{ \prod_{i=1}^n f(x_i; \mu, \sigma^2) \right\} \{\pi(\mu, \sigma^2)\} \{\prod_{i=1}^n \pi(a_i)\pi(b_i)\}$$

$$\propto \left\{ \prod_{i=1}^n \mathcal{N}(x_i; \mu, \sigma^2) \right\} \{\sigma^{-2}\} \{\prod_{i=1}^n \mathcal{U}(a_i; x_i - g, x_i)\mathcal{U}(b_i; x_i, x_i + g)\}. \tag{3.19}$$

Here, $f(X^{(r)}, X^{(nr)}; \theta, D) = f(X^{(r)}, X^{(nr)}; \theta)$ because the data $(X^{(r)}, X^{(nr)})$ is fitted by the model without being masked just like how the true data $X$ is in Method 2.
Thus, the full conditional distributions of \((\mu, \sigma^2)\) are the same as those in (3.13) and (3.14). And the full conditional distributions of \((a_i, b_i)\) are

\[
\pi(a_i|\cdot) = \mathcal{U}(a_i; x_i - g, x_i) \\
\pi(b_i|\cdot) = \mathcal{U}(b_i; x_i, x_i + g).
\] (3.20) (3.21)

When \((\mu, \sigma^2)\) are drawn from unknown \(D\) in Method 3 and 4, the intruder needs to first estimate \(D\), say by \(\hat{D}\), using (3.20) and (3.21). Then, he augments \((\hat{D}, W)\) by \(\tilde{X}\) as shown in (3.15) and (3.16). Finally, he draws \((\mu, \sigma^2)\) according to (3.17) and (3.18).

In the following, we generate 10 sets of fully synthetic data for each method using a common \(X\). We treat Method 1 as our benchmark because it is the standard approach to generate the synthetic data without any interval protection methods. Comparing others with it can help to measure the pros and cons of the new approaches. For computational convenience, we assume the intruder makes guesses between \(0.75X_{[1]}\) and \(1.25X_{[1]}\) by an increment of 0.01 for Method 1 as well as for the other methods when \(D\) is unknown. When \(D\) is known, we assume the intruder makes guesses within \((a_{X_{[1]}}, b_{X_{[1]}})\) by an increment of 0.01 for all the interval-protected methods. Here, \(a_{X_{[1]}}\) and \(b_{X_{[1]}}\) are the lower and upper bounds for \(X_{[1]}\) implied by \((D, W)\).

Data utilities

Figure 3.1 shows the histograms of the true data \(X\) versus the synthetic data \(Y\) when \(g = 0.1\) and \(g = 0.5\), respectively. The synthetic histograms of Method 1 and 3 are close to the true one, but those of Method 2 and 4 are right on. That’s because preventing \(Y\) from being outside of the protection intervals by either Method 2 or 4 leads to synthetic data very close to the truth. Increasing \(g\) from 0.1 to 0.5 does not seem to affect the synthetic histograms of Method 3 much. However, this does influence the data utilities of Method 2 and 4, as their synthetic histograms become misaligned with the truth.
Figure 3.1: Results for univariate additive noise simulations. The histograms of the true data $X$ versus 10 sets of synthesis $Y$.

Posterior disclosure risks

Figure 3.2 displays the posterior predictive densities of all four methods for $g = 0.1$. Table 3.2 summarizes the posterior disclosure risk metrics. When $D$ is known, all the methods have a high posterior disclosure risk for the true $X_1$. For Method 1, this is expected because it does not have any restrictions on either the model parameters
or the released data. For the interval-protected methods, the intruder can narrow his guesses around the truth with the help of the known and informative intervals. Here, comparing the relative risks is difficult because their denominators (i.e., the prior disclosure risks) are different within \((0.95X_{[1]}, 1.05X_{[1]})\).

When \(D\) is unknown, intuitively, the posterior risk metrics of all the interval-protected methods should not have a major increase or decrease, since \(g\) is small. Although the relative risk of Method 2 and 4 are higher than those when \(D\) is known, their interval risks and posterior means remain almost the same. The reason of having higher relative risk in Method 2 and 4 is that: (1) their interval risks are still equal to 1, since their posterior predictive densities are non-zero only within the estimated range \((\hat{a}_{X_{[1]}}, \hat{b}_{X_{[1]}})\) which is inside \((0.95X_{[1]}, 1.05X_{[1]});\) and (2) their prior risk probability within \((\hat{a}_{X_{[1]}}, \hat{b}_{X_{[1]}})\) becomes smaller, since their prior disclosure risks are now uniformly distributed over a wider range of the possible guesses. Then, the ratios of the posterior predictive probability over the prior disclosure risk probability within \((0.95X_{[1]}, 1.05X_{[1]}):\) can become large.

When \(D\) is unknown, the shape of the posterior predictive density of Method 3 looks similar to that of Method 1. This makes sense, because both of them do not restrict the released synthetic data. However, the interval risk as well as the relative risk of Method 3 are slightly lower than those of Method 1. That’s because the intruder needs to sample the model parameters from the unknown intervals, which can make his guesses less accurate about \((\mu, \sigma^2)\), and hence \(X_{[1]}\).

Figure 3.3 displays the posterior predictive densities of all the four methods for \(g = 0.5\). The corresponding posterior disclosure risk metrics are summarized in Table 3.3. In general, increasing the value of \(g\) decreases the posterior disclosure risks for all the interval-protected methods. This is especially true when \(D\) is unknown.

When \(D\) is unknown, the interval risks as well the relative risks of all the interval-protected methods drop. In addition, their posterior means deviate more from the true \(X_{[1]}\) than those in Table 3.2. In increasing order, the posterior disclosure risks of the methods are: Method 3 < Method 4 < Method 2. This indicates that restricting
Table 3.2: The posterior disclosure risk metrics of learning $X_{[1]} = 6$ for $g = 0.1$.

<table>
<thead>
<tr>
<th>Method</th>
<th>Interval Risk</th>
<th>Relative Risk</th>
<th>Posterior Mean</th>
<th>Relative MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method 1</td>
<td>0.3</td>
<td>1.48</td>
<td>6.08</td>
<td>0.11</td>
</tr>
<tr>
<td>Method 2</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>$1 \times 10^{-3}$</td>
</tr>
<tr>
<td>Method 3</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>$1 \times 10^{-3}$</td>
</tr>
<tr>
<td>Method 4</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>$1 \times 10^{-3}$</td>
</tr>
</tbody>
</table>

When $D$ is unknown

<table>
<thead>
<tr>
<th>Method</th>
<th>Interval Risk</th>
<th>Relative Risk</th>
<th>Posterior Mean</th>
<th>Relative MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method 1</td>
<td>0.3</td>
<td>1.48</td>
<td>6.08</td>
<td>0.11</td>
</tr>
<tr>
<td>Method 2</td>
<td>0.98</td>
<td>4.85</td>
<td>6.07</td>
<td>0.02</td>
</tr>
<tr>
<td>Method 3</td>
<td>0.22</td>
<td>1.1</td>
<td>5.56</td>
<td>0.14</td>
</tr>
<tr>
<td>Method 4</td>
<td>0.96</td>
<td>4.74</td>
<td>5.88</td>
<td>0.03</td>
</tr>
</tbody>
</table>

Table 3.3: The posterior disclosure risk metrics of knowing $X_{[1]} = 6$ for $g = 0.5$.

<table>
<thead>
<tr>
<th>Method</th>
<th>Interval Risk</th>
<th>Relative Risk</th>
<th>Posterior Mean</th>
<th>Relative MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method 1</td>
<td>0.3</td>
<td>1.48</td>
<td>6.08</td>
<td>0.11</td>
</tr>
<tr>
<td>Method 2</td>
<td>1</td>
<td>1</td>
<td>6.01</td>
<td>$2 \times 10^{-3}$</td>
</tr>
<tr>
<td>Method 3</td>
<td>1</td>
<td>1</td>
<td>6.02</td>
<td>$6 \times 10^{-3}$</td>
</tr>
<tr>
<td>Method 4</td>
<td>1</td>
<td>1</td>
<td>6.02</td>
<td>$6 \times 10^{-3}$</td>
</tr>
</tbody>
</table>

When $D$ is unknown

<table>
<thead>
<tr>
<th>Method</th>
<th>Interval Risk</th>
<th>Relative Risk</th>
<th>Posterior Mean</th>
<th>Relative MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method 1</td>
<td>0.3</td>
<td>1.48</td>
<td>6.08</td>
<td>0.11</td>
</tr>
<tr>
<td>Method 2</td>
<td>0.98</td>
<td>4.85</td>
<td>6.07</td>
<td>0.02</td>
</tr>
<tr>
<td>Method 3</td>
<td>0.22</td>
<td>1.1</td>
<td>5.56</td>
<td>0.14</td>
</tr>
<tr>
<td>Method 4</td>
<td>0.96</td>
<td>4.74</td>
<td>5.88</td>
<td>0.03</td>
</tr>
</tbody>
</table>

the model parameters by sampling them from the protection intervals can help to reduce the posterior disclosure risk. And not restricting the released data can further lower this risk. Thus, guessing $X_{[1]}$ by Method 3 is most inaccurate. The reason is that unlike Method 2 the intruder needs to sample the model parameters from the unknown intervals by Method 3. In addition, unlike Method 4 the released $Y$ from Method 3 is not restricted.
Table 3.4: Methods of releasing the synthetic data $Y = (Y_1, Y_2)$ for the true data $X = (X_1, X_2)$. Here, $D_1 = \{d_{1,i} = (a_{1,i}, b_{1,i}) : i = 1, \ldots, n\}$ is the collection of protection intervals for each observation $X_{1,i} = x_{1,i}$. And $W_1 = \{w_{1,i} : i = 1, \ldots, n\}$ is the collection of indicators such that $w_{1,i} = 1$ if $a_{1,i} \leq x_{1,i} \leq b_{1,i}$ and $w_{1,i} = 0$ otherwise. Similarly, $D_2 = \{d_{2,i} = (a_{2,i}, b_{2,i}) : i = 1, \ldots, n\}$ is the collection of protection intervals for each observation $X_{2,i} = x_{2,i}$. And $W_2 = \{w_{2,i} : i = 1, \ldots, n\}$ is the collection of indicators such that $w_{2,i} = 1$ if $a_{2,i} \leq x_{2,i} \leq b_{2,i}$ and $w_{2,i} = 0$ otherwise.

| Draw $(Y_1, Y_2)$ from | Draw $\theta = (\theta_1, \theta_2)$ from | $f(\theta|X_1, X_2)$ | $f(\theta|D_1, W_1, D_2, W_2)$ |
|------------------------|-----------------------------------------------|---------------------|-------------------------------|
| $f(Y_1, Y_2; \theta)$ | $f(Y_1, Y_2; \theta, D_1, W_1)$ | Method 1            | Method 4                      |
| $f(Y_1, Y_2; \theta, D_1, W_1)$ | $f(Y_1, Y_2; \theta, D_1, W_1, D_2, W_2)$ | Method 2            | Method 5                      |
| $f(Y_1, Y_2; \theta, D_1, W_1, D_2, W_2)$ |                                  | Method 3            | Method 6                      |

3.3 Study with two variables

We now turn to a simulation with two variables, so as to illustrate a general path forward for multivariate data. Of course, synthesizing a large number of variables, as well as evaluating disclosure risks, becomes increasingly complicated computationally as we add more variables. Hence, the results here serve as a preliminary study of the potential of the protected-interval approach for data synthesis.

Suppose the true data $X = (X_1, X_2)$ can be fitted by $X_1 \sim f_1(X_1; \theta_1)$ and $(X_2|X_1) \sim f_2(X_2; X_1, \theta_2)$. We let $X_2$ contain the sensitive information. To protect $X_2$, the agent may want to use a protective interval for $X_1$ as well, even if it is deemed not confidential. Otherwise, the intruder may guess the sensitive values in $X_2$ from the distribution of $(X_2|X_1) \sim f_2(X_2; X_1, \theta_2)$. We consider six different ways of generating synthetic data as listed in Table 3.4.

Method 1, 2, 3 draw the model parameters $\theta = (\theta_1, \theta_2)$ from the true data $X$. Among them, Method 1 is a standard approach for data synthesis that does not use protected intervals for either $Y_1$ or $Y_2$. Method 2 uses protected intervals for $Y_1$. However, its released $Y_2$ is sampled from $f_2(X_2; X_1, \theta_2)$ without any protection. Method 3 uses protected intervals for both $Y_1$ and $Y_2$. That is, it draws $Y_1$ from $f_1(X_1; \theta_1, D_1, W_1)$ and $Y_2$ from $f_2(X_2; X_1, \theta_2, D_2, W_2)$. The ways how Method 4, 5, and 6 release the synthesis are similar to those by Method 1, 2, and 3. The only
difference is that Method 4, 5, and 6 draw the model parameters \( \theta \) from the protected intervals \( (D_1, D_2, W_1, W_2) \).

Here is an example of how the agent samples the model parameters \( \theta \) from the true data \( X \). Suppose \( X \) can be fitted by \( X_1 \sim \mathcal{N}(\mu_1, \sigma^2_1) \) and \( (X_2 | X_1) \sim \mathcal{N}(\beta X_1, \sigma^2_1) \). To draw the model parameters \( \theta = (\mu_1, \sigma^2_1, \beta, \sigma^2_2) \) from \( X \), the agent can assume conjugate priors for \( \theta \) as \( \pi(\mu_1, \sigma^2_1) \propto \sigma^{-2} \) and \( \pi(\beta, \sigma^2_2) = \pi(\beta | \sigma^2_2) \pi(\sigma^2_2) = \mathcal{N}(\mu_0, \sigma^2_2 \tau_0^{-1}) \mathcal{G}^{-1}(\nu_0/2, \omega_0 s_0^2/2) \). Then, the full conditional distributions are:

\[
\pi(\mu_1 | -) = \mathcal{N}(\sum_{i=1}^{n} x_{1,i}/n, \sigma^2_1/n) \tag{3.22}
\]

\[
\pi(\sigma^2_1 | -) = \mathcal{G}^{-1}(n/2, \sum_{i=1}^{n} (x_{1,i} - \mu_1)^2/2) \tag{3.23}
\]

\[
\pi(\beta | -) = \mathcal{N}(\mu_n, \sigma^2_2 \tau_n^{-1}) \tag{3.24}
\]

\[
\pi(\sigma^2_2 | -) = \mathcal{G}^{-1}((\nu_0 + n)/2, (\omega_0 s_0^2 + \sum_{i=1}^{n} x_{2,i}^2 + \tau_0 \mu_0^2 - \tau_n \mu_n^2)/2) \tag{3.25}
\]

in which \( \tau_n = \sum_{i=1}^{n} x_{1,i}^2 + \tau_0 \) and \( \mu_n = \tau_n^{-1}(\tau_0 \mu_0 + \sum_{i=1}^{n} x_{1,i} x_{2,i}) \).

If the agent needs to draw the model parameters \( \theta = (\mu_1, \sigma^2_1, \beta, \sigma^2_2) \) from the protection intervals \( (D_1, W_1, D_2, W_2) \), he can apply the approach of data augmentation and then replace the true data \( X \) with the augmented data \( \tilde{X} \) in the above full conditional distributions. The way how to augment the data is similar to the example of univariate normal shown in (3.15) and (3.16). That is, the augmented data \( \tilde{X} \) follows

\[
f(\tilde{x}_{1i} | \mu_1, \sigma^2_1, w_{1i}, d_{1i}) \propto \mathcal{N}(\tilde{x}_{1i}; \mu_1, \sigma^2_1) \cdot \prod_{(a_1,b_{1i})}^{w_{1i}} \cdot \prod_{(a_1,b_{1i})}^{1-w_{1i}} \tag{3.26}
\]

\[
f(\tilde{x}_{2i} | \tilde{x}_{1i}, \beta, \sigma^2_2, w_{2i}, d_{2i}) \propto \mathcal{N}(\tilde{x}_{2i} ; \beta \tilde{x}_{1i}, \sigma^2_2) \cdot \prod_{(a_2,b_{2i})}^{w_{2i}} \cdot \prod_{(a_2,b_{2i})}^{1-w_{2i}} \tag{3.27}
\]

which can be drawn by the method of inverse CDF such that \( \tilde{x}_{1i} = \mu_1 + \sigma_1 \Phi^{-1}(v_1) \)
and $\tilde{x}_{2i} = \beta \tilde{x}_{1i} + \sigma_2 \Phi^{-1}(v2)$ with $v_1$ and $v_2$ as

$$
v_1 = \begin{cases} 
  u_1 \left( \Phi \left( \frac{\tilde{x}_{1i} - \mu_1}{\sigma_1} \right) - \Phi \left( \frac{\tilde{x}_{1i} - \mu_1}{\sigma_1} \right) \right) + \Phi \left( \frac{\tilde{x}_{1i} - \mu_1}{\sigma_1} \right), & \text{for } w_{1i} = 1 \\
  c_{1i}^{-1} u_1, & \text{for } w_{1i} = 0 \& u_1 < c_{1i} \Phi \left( \frac{\tilde{x}_{1i} - \mu_1}{\sigma_1} \right) \\
  c_{1i}^{-1} u_1 + \Phi \left( \frac{\tilde{x}_{1i} - \mu_1}{\sigma_1} \right) - \Phi \left( \frac{\tilde{x}_{1i} - \mu_1}{\sigma_1} \right), & \text{for } w_{1i} = 0 \& u_1 > c_{1i} \Phi \left( \frac{\tilde{x}_{1i} - \mu_1}{\sigma_1} \right), 
\end{cases}
$$

(3.28)

$$
v_2 = \begin{cases} 
  u_2 \left( \Phi \left( \frac{\tilde{x}_{1i} - \beta \tilde{x}_{1i}}{\sigma_2} \right) - \Phi \left( \frac{\tilde{x}_{1i} - \beta \tilde{x}_{1i}}{\sigma_2} \right) \right) + \Phi \left( \frac{\tilde{x}_{1i} - \beta \tilde{x}_{1i}}{\sigma_2} \right), & \text{for } w_{2i} = 1 \\
  c_{2i}^{-1} u_2, & \text{for } w_{2i} = 0 \& u_2 < c_{2i} \Phi \left( \frac{\tilde{x}_{1i} - \beta \tilde{x}_{1i}}{\sigma_2} \right) \\
  c_{2i}^{-1} u_2 + \Phi \left( \frac{\tilde{x}_{1i} - \beta \tilde{x}_{1i}}{\sigma_2} \right) - \Phi \left( \frac{\tilde{x}_{1i} - \beta \tilde{x}_{1i}}{\sigma_2} \right), & \text{for } w_{2i} = 0 \& u_2 > c_{2i} \Phi \left( \frac{\tilde{x}_{1i} - \beta \tilde{x}_{1i}}{\sigma_2} \right). 
\end{cases}
$$

(3.29)

Here, $u_1$ and $u_2$ are i.i.d. draws from $U(0, 1)$. The constants $c_{1i}^{-1} = 1 - \Phi \left( \frac{\tilde{x}_{1i} - \mu_1}{\sigma_1} \right) + \Phi \left( \frac{\tilde{x}_{1i} - \mu_1}{\sigma_1} \right)$ and $c_{2i}^{-1} = 1 - \Phi \left( \frac{\tilde{x}_{1i} - \beta \tilde{x}_{1i}}{\sigma_2} \right) + \Phi \left( \frac{\tilde{x}_{1i} - \beta \tilde{x}_{1i}}{\sigma_2} \right)$.

### 3.3.1 Posterior predictive density

Suppose the risky values in $X$ are $X^{(r)} = (X_1^{(r)}, X_2^{(r)})$ and the non-risky ones are $X^{(nr)} = (X_1^{(nr)}, X_2^{(nr)})$. The way to calculate the posterior predictive density is similar to that of the univariate case. We assume the prior disclosure risk density is uniformly distributed within the possible support. Then, we compute the posterior predictive density according to the situations that $(D_1, D_2, W_1, W_2)$ are known and unknown, respectively.

For example, when $(D_1, D_2, W_1, W_2)$ are known the posterior predictive density
If the agent draws $\theta$ from $X$, to be consistent with how the agent does, we assume the
intruder will sample $\hat{\theta}_b$ from $f(\theta|X_1^{(r)}, X_2^{(r)}, X_1^{(nr)}, X_2^{(nr)}).$ Otherwise, he will sample
it from $f(\theta|D_1, D_2, W_1, W_2)$. Thus, when the protection intervals are known the
posterior predictive densities under Method 4, 5, 6 are uniformly distributed over
all the possible guesses. In addition,

$$f(Y_1, Y_2|\hat{\theta}_b, D_1, D_2, W_1, W_2)$$

$$= \begin{cases} 
  f_1(Y_1|\hat{\theta}_1, b, D_1, W_1)f_2(Y_2|Y_1, \hat{\theta}_2, b) & \text{under Method 1 and 4} \\
  f_1(Y_1|\hat{\theta}_1, b, D_1, W_1)f_2(Y_2|Y_1, \hat{\theta}_2, b) & \text{under Method 2 and 5} \\
  f_1(Y_1|\hat{\theta}_1, b, D_1, W_1)f_2(Y_2|Y_1, \hat{\theta}_2, b, D_2, W_2) & \text{under Method 3 and 6}
\end{cases}$$

(3.31)
Likewise, when \((D_1, D_2, W_1, W_2)\) are unknown the posterior predictive density is:

\[
f(X_1^{(r)}, X_2^{(r)}|X_1^{(nr)}, X_2^{(nr)}; Y_1, Y_2) \approx \prod_{Y_2} \int_{Y_1} f(Y_1, Y_2|X_1^{(r)}, X_2^{(r)}, X_1^{(nr)}, X_2^{(nr)}; \theta, D_1, D_2, W_1, W_2) \cdot f(\theta, D_1, D_2, W_1, W_2) d\theta dD_1 dD_2 dW_1 dW_2 \]

\[
\times \int_{Y_1} f(Y_1, Y_2|X_1^{(r)}, X_2^{(r)}, X_1^{(nr)}, X_2^{(nr)}, D_1, D_2, W_1, W_2, \theta) \cdot f(\theta, D_1, D_2, W_1, W_2|X_1^{(r)}, X_2^{(r)}, X_1^{(nr)}, X_2^{(nr)}) d\theta dD_1 dD_2 dW_1 dW_2 \]

\[
\approx \prod_{Y_2} \int_{Y_1} f(Y_1, Y_2|\hat{\theta}_b, \hat{D}_1, \hat{D}_2, \hat{W}_1, \hat{W}_2) \cdot \frac{1}{B} \sum_{b=1}^B f(Y_1, Y_2|\hat{\theta}_b, \hat{D}_1, \hat{D}_2, \hat{W}_1, \hat{W}_2) \quad (3.32)
\]

### 3.3.2 Simulation study

Similar to the univariate case, we generate the true data \(X_1\) from \(N(\mu_1, \sigma_1^2)\) with \(n = 1,000\). We replace its biggest value by \(6\sigma\); call this point \(X_{1,[1]}\). Then, we sample \(X_2\) from \(N(\beta X_1, \sigma_2^2)\) for all points but the record with \(X_{1,[1]}\), and set \(X_{2,[1]} = \beta X_{1,[1]} + 0.1\). Here, \(\mu_1 = 0, \sigma_1^2 = 1, \beta = 1.1, \sigma_2^2 = 1\). We allow \(w_{1,i} = w_{2,i} = 1\) for all \(i = 1, \ldots, n\).

At the same time, we let \(a_{1,i} = x_{1,i} - c_{1,i}^{(a_1)}, b_{1,i} = x_{1,i} + c_{1,i}^{(b_1)}\), \(a_{2,i} = x_{2,i} - c_{2,i}^{(a_2)}, b_{2,i} = x_{2,i} + c_{2,i}^{(b_2)}\) with \(c_{1,i}^{(a_1)} \sim \mathcal{U}(0, g_1), c_{1,i}^{(b_1)} \sim \mathcal{U}(0, g_1), c_{1,i}^{(a_2)} \sim \mathcal{U}(0, g_2), c_{1,i}^{(b_2)} \sim \mathcal{U}(0, g_2)\). We define \(a_{1,X_{1,[1]}}\) and \(b_{1,X_{1,[1]}}\) to be the lower and upper bounds for \(X_{1,[1]}\). Likewise, \(a_{2,X_{2,[1]}}\) and \(b_{2,X_{2,[1]}}\) are the lower and upper bounds for \(X_{2,[1]}\).

We assume the intruder’s target is \(X_{2,[1]}\). He knows the true \(X\) except for \(X_{[1]} = (X_{1,[1]}, X_{2,[1]})\). He also knows \(w_{1,i} = w_{2,i} = 1\) for all \(i\). In addition, he has the knowledge of how we generate the protection intervals and knows the values of \((g_1, g_2)\).

We assume the intruder searches over the grids with the size of \(0.01 \times 0.01\) in the area of \((0.75X_{1,[1]}, 1.25X_{1,[1]}) \times (0.75X_{2,[1]}, 1.25X_{2,[1]})\) for Method 1, for the other methods when both \(D_1\) and \(D_2\) are unknown. If \(D_1\) is known but \(D_2\) is
not, he will search through the grids with the size of 0.01 \times 0.01 in the area of 
\((a_{1i,1}, b_{1i,1}) \times (a_{2i,2}, b_{2i,2}) = \beta_{a} a_{1i,1}, \beta_{b} b_{1i,1})\) for Method 2. If both \(D_{1}\) and \(D_{2}\) are known, he will search over the grids with the size of 0.01 \times 0.01 in the area of 
\((a_{1i,1}, b_{1i,1}) \times (a_{2i,2}, b_{2i,2})\) for Method 3, 4, 5, and 6.

When \(\theta = (\mu_{1}, \sigma_{1}^{2}, \beta, \sigma_{2}^{2})\) is drawn from the true data \(X = (X_{1}, X_{2})\) (as in Method 1, 2, 3), the full conditional distributions of \(\theta\) are shown in (3.22) \(\sim\) (3.25). When \(\theta\) is drawn from the known protection intervals \((D_{1}, D_{2})\) (as in Method 4, 5, and 6), the intruder only needs to replace the true data \(X\) with the augmented data \(\tilde{X}\) in these full conditional distributions.

When \(D_{1}\) is unknown, its conditional distributions are \(\pi(a_{1i}|-) = U(x_{1i} - g_{1}, x_{1i})\) and \(\pi(b_{1i}|-) = U(x_{1i}, x_{1i} + g)\). Similarly, when \(D_{2}\) is unknown its conditional distributions are \(\pi(a_{2i}|-) = U(x_{2i} - g_{2}, x_{2i})\) and \(\pi(b_{2i}|-) = U(x_{2i}, x_{2i} + g)\). Then, when \(\theta\) is drawn from the unknown protection intervals \((D_{1}, D_{2})\) the intruder can still sample the augmented data by (3.28) and (3.29) by substituting the true protection intervals with the estimated ones.

We generate 10 sets of fully synthetic data for each method. Since Method 1 is the standard approach to release the synthesis which does not involve any protection intervals, we still treat it as a benchmark. Its posterior predictive density remains the same no matter how we make changes of \((g_{1}, g_{2})\).

**Data utilities**

Figure 3.4 and 3.5 display the histograms of the synthetic \(Y_{2}\) versus the true \(X_{2}\), and Table 3.5 shows the statistical summaries of the estimated regression coefficients when \(g_{1} = g_{2} = 0.1\) or \(g_{1} = g_{2} = 0.5\), respectively. By comparison, we can see that the data utilities of Method 3 and 6 which restrict \(Y_{2}\) are more sensitive to the slight changes in \(g_{1}\) and \(g_{2}\). When \(g_{1} = g_{2} = 0.1\), the synthetic histograms from all the interval-protected methods are reasonably similar to those from Method 1, with those from Method 3 and 6 being the most favorable. However, when \(g_{1} = g_{2} = 0.5\) the data utilities of Method 3 and 6 deteriorate more quickly than those of the other
Table 3.5: The means of the estimated regression coefficients from the synthetic data for each method compared to the coefficient on the real data (i.e., $\beta = 1.1$) together with the standard errors.

<table>
<thead>
<tr>
<th>Method</th>
<th>S.E.</th>
<th>Mean</th>
<th>Method</th>
<th>S.E.</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>when $g_1 = g_2 = 0.1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method 1</td>
<td>0.01</td>
<td>1.11</td>
<td>Method 2</td>
<td>0.01</td>
<td>1.1</td>
</tr>
<tr>
<td>Method 3</td>
<td>$1 \times 10^{-3}$</td>
<td>1.1</td>
<td>Method 4</td>
<td>0.01</td>
<td>1.11</td>
</tr>
<tr>
<td>Method 5</td>
<td>0.01</td>
<td>1.1</td>
<td>Method 6</td>
<td>$1 \times 10^{-3}$</td>
<td>1.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>when $g_1 = g_2 = 0.5$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method 1</td>
<td>0.01</td>
<td>1.11</td>
<td>Method 2</td>
<td>0.01</td>
<td>1.1</td>
</tr>
<tr>
<td>Method 3</td>
<td>$1 \times 10^{-3}$</td>
<td>1.1</td>
<td>Method 4</td>
<td>0.01</td>
<td>1.09</td>
</tr>
<tr>
<td>Method 5</td>
<td>0.01</td>
<td>1.09</td>
<td>Method 6</td>
<td>$2 \times 10^{-3}$</td>
<td>1.09</td>
</tr>
</tbody>
</table>

Posterior disclosure risks

Figure 3.6 and 3.7 display the posterior predictive densities of all six synthetic methods for $g_1 = g_2 = 0.1$ when $(D_1, D_2)$ are known or unknown, respectively. Table 3.6 summarizes the posterior disclosure risk metrics. When the protective intervals are known all the methods have a high posterior disclosure risk due to the fact that the intruder can make use of the informative protection intervals to target the true $X_2[1]$. Here, one cannot easily compare the relative risks because their prior disclosure risks within $(0.95X_2[1], 1.05X_2[1])$ are different.

Similar to the univariate case, when $D_1$ and $D_2$ are unknown, the posterior disclosure risk metrics of all the interval-protected methods remain almost the same. However, the interval risk of Method 4 drops from 1 to 0.2 with its relative risk being slightly higher than 1. At the same time, the relative risks of the other interval-protected methods are about four to five times bigger than those in Table 3.6. That’s because their interval risks remain close to 1 while their prior disclosure risks become smaller.

Figure 3.8 and 3.9 display the posterior predictive densities of all six synthesis methods for $g_1 = g_2 = 0.5$ when $(D_1, D_2)$ are known or unknown, respectively. Table 3.7 summarizes the corresponding posterior disclosure risk metrics. Basically,
Table 3.6: The posterior disclosure risk metrics of knowing $X_{[2]} = 6.7$ for $g_1 = g_2 = 0.1$.

<table>
<thead>
<tr>
<th>Interval Risk</th>
<th>Relative Risk</th>
<th>Posterior Mean</th>
<th>Relative MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method 1</td>
<td>0.32</td>
<td>1.6</td>
<td>6.7</td>
</tr>
<tr>
<td>Method 2</td>
<td>1</td>
<td>1</td>
<td>6.69</td>
</tr>
<tr>
<td>Method 3</td>
<td>1</td>
<td>1</td>
<td>6.72</td>
</tr>
<tr>
<td>Method 4</td>
<td>1</td>
<td>1</td>
<td>6.71</td>
</tr>
<tr>
<td>Method 5</td>
<td>1</td>
<td>1</td>
<td>6.71</td>
</tr>
<tr>
<td>Method 6</td>
<td>1</td>
<td>1</td>
<td>6.71</td>
</tr>
</tbody>
</table>

when $(D_1, D_2)$ are unknown

<table>
<thead>
<tr>
<th>Interval Risk</th>
<th>Relative Risk</th>
<th>Posterior Mean</th>
<th>Relative MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method 1</td>
<td>0.32</td>
<td>1.6</td>
<td>6.7</td>
</tr>
<tr>
<td>Method 2</td>
<td>0.86</td>
<td>4.27</td>
<td>6.85</td>
</tr>
<tr>
<td>Method 3</td>
<td>1</td>
<td>4.94</td>
<td>6.73</td>
</tr>
<tr>
<td>Method 4</td>
<td>0.2</td>
<td>0.99</td>
<td>6.25</td>
</tr>
<tr>
<td>Method 5</td>
<td>0.91</td>
<td>4.52</td>
<td>6.39</td>
</tr>
<tr>
<td>Method 6</td>
<td>1</td>
<td>4.94</td>
<td>6.76</td>
</tr>
</tbody>
</table>

increasing $g_1$ and $g_2$ can help to reduce the posterior disclosure risks of all the interval-protected methods. This is especially true for the ones that do not directly restrict $Y_2$ (i.e., Method 2, 4, 5). Among Method 2, 4, 5, the interval risk, the relative risk, and the posterior mean are lowest for Method 4. The reason is that Method 4 samples the model parameters from the unknown intervals and both its $Y_1$ and $Y_2$ are unrestricted. Although Method 5 also samples the model parameters from the unknown intervals, its $Y_1$ is restricted. Thus, its posterior disclosure risk is a little higher than than of Method 4, but still lower than that of Method 2.

3.4 Multiplicative intervals

Additive noise approaches use a constant noise distribution regardless of the values of $X$. In practice, many applications of noise infusion methodology prefer to use greater noise for larger values and smaller noise for smaller values. This often is accomplished in practice using multiplicative noise rather than additive noise. In this section, we repeat the experiments of Section 3.2.2 and 3.3.2 using a variant of
Table 3.7: The posterior risk metrics of knowing $X_{2,[1]} = 6.7$ for $g_1 = g_2 = 0.5$.

when $D_1$ is known in Method 2 & $(D_1, D_2)$ are known in Method 3, 4, 5, and 6

<table>
<thead>
<tr>
<th>Interval Risk</th>
<th>Relative Risk</th>
<th>Posterior Mean</th>
<th>Relative MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method 1</td>
<td>0.32</td>
<td>1.6</td>
<td>6.7</td>
</tr>
<tr>
<td>Method 2</td>
<td>1</td>
<td>1</td>
<td>6.69</td>
</tr>
<tr>
<td>Method 3</td>
<td>1</td>
<td>1</td>
<td>6.62</td>
</tr>
<tr>
<td>Method 4</td>
<td>1</td>
<td>1</td>
<td>6.8</td>
</tr>
<tr>
<td>Method 5</td>
<td>1</td>
<td>1</td>
<td>6.8</td>
</tr>
<tr>
<td>Method 6</td>
<td>1</td>
<td>1</td>
<td>6.8</td>
</tr>
</tbody>
</table>

when $(D_1, D_2)$ are unknown

<table>
<thead>
<tr>
<th>Interval Risk</th>
<th>Relative Risk</th>
<th>Posterior Mean</th>
<th>Relative MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method 1</td>
<td>0.32</td>
<td>1.6</td>
<td>6.7</td>
</tr>
<tr>
<td>Method 2</td>
<td>0.39</td>
<td>1.95</td>
<td>6.36</td>
</tr>
<tr>
<td>Method 3</td>
<td>1</td>
<td>4.94</td>
<td>6.77</td>
</tr>
<tr>
<td>Method 4</td>
<td>0.11</td>
<td>0.55</td>
<td>7.53</td>
</tr>
<tr>
<td>Method 5</td>
<td>0.14</td>
<td>0.73</td>
<td>7.14</td>
</tr>
<tr>
<td>Method 6</td>
<td>1</td>
<td>4.94</td>
<td>6.64</td>
</tr>
</tbody>
</table>

multiplicative noise.

Abowd et al. (2005) use a fuzzy noise factor to protect the Quarterly Workforce Indicators released by the Census Bureau. The model of this noise factor is

$$Pr(\delta_i) = \begin{cases} 
  \frac{(b - \delta)}{(b - a)^2}, & \delta \in [a, b] \\
  \frac{(b + \delta - 2)}{(b - a)^2}, & \delta \in [2 - b, 2 - a] \\
  0, & \text{otherwise,} 
\end{cases}$$

(3.33)

in which $a = 1 + c/100$, $b = 1 + d/100$, and $1 < a < b < 2$. Thus, the fuzzy noise factor is centered at 1 with the minimum distortion of $c$ percent and maximum distortion of $d$ percent ($c$ and $d$ are confidential). Then, a true value $X_i$ is distorted as $X_i^* = X_i\delta_i$.

We modify this idea slightly. We let the upper and lower bound of each protection interval be $a_i = x_i\delta_i^{(a)}$ and $b_i = x_i\delta_i^{(b)}$ with

$$Pr(\delta_i^{(a)}) = 2(c^{(b)} - \delta_i^{(a)})/(c^{(b)} - c^{(a)})^2, \quad \delta_i^{(a)} \in [c^{(a)}, c^{(b)}]$$

$$Pr(\delta_i^{(b)}) = 2(c^{(b)} + \delta_i^{(b)} - 2)/(c^{(b)} - c^{(a)})^2, \quad \delta_i^{(b)} \in [2 - c^{(b)}, 2 - c^{(a)}].$$

(3.34)
with $c^{(a)} = 1 + g^{(a)}/100$ and $c^{(b)} = 1 + g^{(b)}/100$.

The approach in Abowd et al. (2005) directly releases the noisy data distorted by the fuzzy noise factor. Instead, we generate two noisy values for each original one, use these noisy values to construct the protection intervals, and release multiple synthetic values by restricting them within these intervals and/or sampling the model parameters conditioned on them. Since $g^{(a)}$ controls the minimum percentage in distortion, we expect that increasing it can help to reduce the posterior disclosure risk.

3.4.1 Simulation study

We apply (3.34) to construct the protection intervals for the studies with single variable and with two variables. For the study with single variable, we allow $(g^{(a)} = 5, g^{(b)} = 25)$ or $(g^{(a)} = 15, g^{(b)} = 25)$ to protect the same data we have in Section 3.2.2. For the study with two variables, we allow $(g_1^{(a)} = g_2^{(a)} = 15, g_1^{(b)} = g_2^{(b)} = 25)$ or $(g_1^{(a)} = g_2^{(a)} = 15, g_1^{(b)} = g_2^{(b)} = 25)$ to protect the same data we have in Section 3.3.2.

Study with a single variable

Figure 3.10 displays the histograms of the true $X$ versus the synthetic $Y$ of all the four methods when $(g^{(a)} = 5, g^{(b)} = 25)$ and $(g^{(a)} = 15, g^{(b)} = 25)$, respectively. By comparison, we can see that increasing the minimum distortion rate (i.e., $g^{(a)}$) does not seem to affect the data utilities of all the interval-protected methods much. However, increasing $g^{(a)}$ does help to lower their posterior disclosure risks. Basically, their posterior predictive densities are shifted to the left (see Figure 3.11 and 3.12 as well as Table 3.8 and 3.9). For example, the spikes of Method 2 become concentrated around 5.18 instead of 6.05. Those of Method 4 are concentrated around 4.97 instead of 5.88. The peak of Method 3 moves to 4.6 from 5.1.
Table 3.8: The posterior disclosure risk metrics of knowing $X_{[1]} = 6$ when $(g^{(a)} = 5, g^{(b)} = 25)$.

<table>
<thead>
<tr>
<th>Method</th>
<th>Interval Risk</th>
<th>Relative Risk</th>
<th>Posterior Mean</th>
<th>Relative MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method 1</td>
<td>0.3</td>
<td>1.48</td>
<td>6.08</td>
<td>0.11</td>
</tr>
<tr>
<td>Method 2</td>
<td>0.44</td>
<td>1</td>
<td>6.37</td>
<td>0.09</td>
</tr>
<tr>
<td>Method 3</td>
<td>0.44</td>
<td>1</td>
<td>6.37</td>
<td>0.09</td>
</tr>
<tr>
<td>Method 4</td>
<td>0.44</td>
<td>1</td>
<td>6.37</td>
<td>0.09</td>
</tr>
</tbody>
</table>

when $D$ is unknown

<table>
<thead>
<tr>
<th>Method</th>
<th>Interval Risk</th>
<th>Relative Risk</th>
<th>Posterior Mean</th>
<th>Relative MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method 1</td>
<td>0.3</td>
<td>1.48</td>
<td>6.083</td>
<td>0.11</td>
</tr>
<tr>
<td>Method 2</td>
<td>0.53</td>
<td>2.63</td>
<td>6.14</td>
<td>0.05</td>
</tr>
<tr>
<td>Method 3</td>
<td>$1 \times 10^{-3}$</td>
<td>$8 \times 10^{-3}$</td>
<td>5.1</td>
<td>0.15</td>
</tr>
<tr>
<td>Method 4</td>
<td>0.36</td>
<td>1.77</td>
<td>5.83</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Table 3.9: The posterior disclosure risk metrics of knowing $X_{[1]} = 6$ when $(g^{(a)} = 15, g^{(b)} = 25)$.

<table>
<thead>
<tr>
<th>Method</th>
<th>Interval Risk</th>
<th>Relative Risk</th>
<th>Posterior Mean</th>
<th>Relative MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method 1</td>
<td>0.3</td>
<td>1.48</td>
<td>6.08</td>
<td>0.11</td>
</tr>
<tr>
<td>Method 2</td>
<td>0.27</td>
<td>1</td>
<td>6.19</td>
<td>0.11</td>
</tr>
<tr>
<td>Method 3</td>
<td>0.27</td>
<td>1</td>
<td>6.19</td>
<td>0.11</td>
</tr>
<tr>
<td>Method 4</td>
<td>0.27</td>
<td>1</td>
<td>6.19</td>
<td>0.11</td>
</tr>
</tbody>
</table>

when $D$ is unknown

<table>
<thead>
<tr>
<th>Method</th>
<th>Interval Risk</th>
<th>Relative Risk</th>
<th>Posterior Mean</th>
<th>Relative MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method 1</td>
<td>$1 \times 10^{-5}$</td>
<td>$5 \times 10^{-5}$</td>
<td>5.18</td>
<td>0.13</td>
</tr>
<tr>
<td>Method 2</td>
<td>$1 \times 10^{-14}$</td>
<td>$8 \times 10^{-14}$</td>
<td>4.6</td>
<td>0.23</td>
</tr>
<tr>
<td>Method 3</td>
<td>$7 \times 10^{-7}$</td>
<td>$3 \times 10^{-6}$</td>
<td>4.97</td>
<td>0.17</td>
</tr>
</tbody>
</table>

Study with two variables

Figure 3.13 and 3.14 display the histograms of the true $X$ versus the synthetic $Y$, and Table 3.10 shows the statistical summary of the estimated regression coefficient of all the six methods when $(g^{(a)} = g^{(a)} = 5, c^{(b)}_1 = c^{(b)}_2 = 25)$ and $(g^{(a)} = g^{(a)} = 15, c^{(b)}_1 = c^{(b)}_2 = 25)$, respectively. Again, we observe that increasing the minimum distortion
Table 3.10: The means of the estimated regression coefficients from the synthetic data for each method compared to the coefficient on the real data (i.e., $\beta = 1.1$) together with the standard errors.

<table>
<thead>
<tr>
<th>Method</th>
<th>S.E.</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.01</td>
<td>1.11</td>
</tr>
<tr>
<td>2</td>
<td>0.01</td>
<td>1.08</td>
</tr>
<tr>
<td>3</td>
<td>$2 \times 10^{-3}$</td>
<td>1.09</td>
</tr>
<tr>
<td>4</td>
<td>0.02</td>
<td>1.12</td>
</tr>
<tr>
<td>5</td>
<td>0.02</td>
<td>1.08</td>
</tr>
<tr>
<td>6</td>
<td>$3 \times 10^{-3}$</td>
<td>1.09</td>
</tr>
</tbody>
</table>

when $g_1^{(a)} = g_2^{(a)} = 5$ and $g_1^{(b)} = g_2^{(b)} = 25$

<table>
<thead>
<tr>
<th>Method</th>
<th>S.E.</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.01</td>
<td>1.11</td>
</tr>
<tr>
<td>2</td>
<td>0.01</td>
<td>1.08</td>
</tr>
<tr>
<td>3</td>
<td>$2 \times 10^{-3}$</td>
<td>1.09</td>
</tr>
<tr>
<td>4</td>
<td>0.02</td>
<td>1.12</td>
</tr>
<tr>
<td>5</td>
<td>0.02</td>
<td>1.08</td>
</tr>
<tr>
<td>6</td>
<td>$3 \times 10^{-3}$</td>
<td>1.09</td>
</tr>
</tbody>
</table>

when $g_1^{(a)} = g_2^{(a)} = 15$ and $g_1^{(b)} = g_2^{(b)} = 25$

3.5 Conclusion

In this study, we explore the interval-protected MI methods for releasing continuous data. These include synthesizing each data point within its individual protective interval, sampling the model parameters from these intervals, and a combination of these two. To illustrate how to measure the risk, we simulate the scenarios in which an intruder uses his knowledge of the data synthesis procedure as well as the interval generating mechanism to target data points. The results reveal that the degree to which an intruder can learn depends on the nature of the protection intervals, whether the intruder knows the intervals or not, and which data synthesis approaches an agent applies.

Basically, when the protection intervals are very informative about the truth, all the methods have a high posterior disclosure risk for large outliers. This is simply because these intervals basically reveal the actual values though the released data
Table 3.11: The posterior risk metrics of knowing $X_{2[1]} = 6.7$ for $g_1^{(a)} = g_2^{(a)} = 5$ and $g_1^{(b)} = g_2^{(b)} = 25$.

When $D_1$ is known in Method 2 & $(D_1, D_2)$ are known in Method 3, 4, 5, and 6

<table>
<thead>
<tr>
<th>Method</th>
<th>Interval Risk</th>
<th>Relative Risk</th>
<th>Posterior Mean</th>
<th>Relative MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.32</td>
<td>1.6</td>
<td>6.7</td>
<td>0.11</td>
</tr>
<tr>
<td>2</td>
<td>$1 \times 10^{-9}$</td>
<td>$3 \times 10^{-9}$</td>
<td>5.97</td>
<td>0.1</td>
</tr>
<tr>
<td>3</td>
<td>0.83</td>
<td>2.59</td>
<td>6.46</td>
<td>0.05</td>
</tr>
<tr>
<td>4</td>
<td>0.36</td>
<td>1</td>
<td>6.81</td>
<td>0.08</td>
</tr>
<tr>
<td>5</td>
<td>0.36</td>
<td>1</td>
<td>6.81</td>
<td>0.08</td>
</tr>
<tr>
<td>6</td>
<td>0.36</td>
<td>1</td>
<td>6.81</td>
<td>0.08</td>
</tr>
</tbody>
</table>

When $(D_1, D_2)$ are unknown

<table>
<thead>
<tr>
<th>Method</th>
<th>Interval Risk</th>
<th>Relative Risk</th>
<th>Posterior Mean</th>
<th>Relative MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.32</td>
<td>1.6</td>
<td>6.7</td>
<td>0.11</td>
</tr>
<tr>
<td>2</td>
<td>0.25</td>
<td>1.28</td>
<td>6.35</td>
<td>0.05</td>
</tr>
<tr>
<td>3</td>
<td>0.66</td>
<td>3.3</td>
<td>6.59</td>
<td>0.04</td>
</tr>
<tr>
<td>4</td>
<td>$2 \times 10^{-7}$</td>
<td>$1 \times 10^{-6}$</td>
<td>5.19</td>
<td>0.22</td>
</tr>
<tr>
<td>5</td>
<td>$3 \times 10^{-7}$</td>
<td>$1.5 \times 10^{-6}$</td>
<td>5.61</td>
<td>0.16</td>
</tr>
<tr>
<td>6</td>
<td>0.97</td>
<td>4.79</td>
<td>6.47</td>
<td>0.04</td>
</tr>
</tbody>
</table>

and/or the model parameters. However, when the protection intervals become less informative about the truth, the posterior disclosure risks of all the interval-protected methods decrease, and the choice of the data synthesis methods becomes important. This is especially true when the intervals are unknown to the intruder. We find that the methods which restrict the model parameters by drawing them from the unknown intervals can help to reduce the risk more than those which do not restrict the model parameters. We also notice that the methods which do not restrict the released data have lower risks than those which do restrict the released data. For example, in the study with single variable Method 3 has lower risk than Method 4, and Method 4 has lower risk than Method 2. Also in the study with two variables, Method 4 has lower risk than Method 5, and Method 5 has lower risk than Method 2.

For the data utility, generally speaking, the methods which restrict the release data have high data utilities when the protection intervals are informative, but their utilities go down quickly when these intervals become less informative. However, we
Table 3.12: The posterior risk metrics of knowing \( X_{2,[1]} = 6.7 \) for \( g_1^{(a)} = g_2^{(a)} = 15 \) and \( g_1^{(b)} = g_2^{(b)} = 25 \).

When \( D_1 \) is known in Method 2 & \((D_1, D_2)\) are known in Method 3, 4, 5, and 6

<table>
<thead>
<tr>
<th>Method</th>
<th>Interval Risk</th>
<th>Relative Risk</th>
<th>Posterior Mean</th>
<th>Relative MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method 1</td>
<td>0.32</td>
<td>1.6</td>
<td>6.7</td>
<td>0.11</td>
</tr>
<tr>
<td>Method 2</td>
<td>( 2 \times 10^{-16} )</td>
<td>( 8 \times 10^{-16} )</td>
<td>5.7</td>
<td>0.14</td>
</tr>
<tr>
<td>Method 3</td>
<td>0.02</td>
<td>0.05</td>
<td>7.12</td>
<td>0.06</td>
</tr>
<tr>
<td>Method 4</td>
<td>0.32</td>
<td>1</td>
<td>6.73</td>
<td>0.09</td>
</tr>
<tr>
<td>Method 5</td>
<td>0.32</td>
<td>1</td>
<td>6.73</td>
<td>0.09</td>
</tr>
<tr>
<td>Method 6</td>
<td>0.32</td>
<td>1</td>
<td>6.73</td>
<td>0.09</td>
</tr>
</tbody>
</table>

When \((D_1, D_2)\) are unknown

<table>
<thead>
<tr>
<th>Method</th>
<th>Interval Risk</th>
<th>Relative Risk</th>
<th>Posterior Mean</th>
<th>Relative MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method 1</td>
<td>0.32</td>
<td>1.6</td>
<td>6.7</td>
<td>0.11</td>
</tr>
<tr>
<td>Method 2</td>
<td>( 4 \times 10^{-4} )</td>
<td>( 2 \times 10^{-3} )</td>
<td>6.17</td>
<td>0.07</td>
</tr>
<tr>
<td>Method 3</td>
<td>( 1 \times 10^{-3} )</td>
<td>( 4 \times 10^{-3} )</td>
<td>6.27</td>
<td>0.06</td>
</tr>
<tr>
<td>Method 4</td>
<td>0</td>
<td>0</td>
<td>5.02</td>
<td>0.24</td>
</tr>
<tr>
<td>Method 5</td>
<td>( 2 \times 10^{-11} )</td>
<td>( 1 \times 10^{-10} )</td>
<td>5.13</td>
<td>0.23</td>
</tr>
<tr>
<td>Method 6</td>
<td>0.08</td>
<td>0.4</td>
<td>6.16</td>
<td>0.08</td>
</tr>
</tbody>
</table>

do not observe such a phenomenon in the methods which do not restrict the released data. In addition, for the multiplicative intervals increasing the minimum distortion rate does not seem to affect the data utilities of all the interval-protected methods.

This study is a preliminary research for applying the interval-protected MI with the protection intervals to mask continuous microdata. There are many gaps that we need to fill in through the future work. Since we only compare these methods by having all the intervals cover the truth, it will be helpful to know their performance by allowing the original data to be outside of the intervals. In addition, we assume the variables are normally distributed, which is usually not true for the real data. For example, the individual annual incomes tend to have a skewed distribution with a heavy right tail. Thus, it will be necessary to have the data simulated from such densities. Moreover, for many continuous variables we need to determine ways to synthesize and compute risks in an efficient way. Finally, the biggest challenge will be to figure out the mechanism of generating the protection intervals. This is a complicated but worth investigating problem. There will be so many factors one
needs to consider when constructing the protection intervals, including not only the widths and locations of the intervals but also the distribution of the original data.
Figure 3.2: The posterior predictive densities of guessing $X_{[1]} = 6$ for $g = 0.1$. When $D$ is known, the plots of Method 3 and 4 are deleted because their posterior disclosure risks are uniformly distributed according to (3.3) and (3.4).
Figure 3.3: The posterior predictive densities of guessing $X_{[1]} = 6$ for $g = 0.5$. When $D$ is known, the plots of Method 3 and 4 are deleted because their posterior disclosure risks are uniformly distributed according to (3.3) and (3.4).
Figure 3.4: Results for bivariate additive noise simulations. The histograms of the true data $X_2$ versus 10 sets of synthesis $Y_2$ when $g_1 = g_2 = 0.1$. 

(a) Method 1
(b) Method 2
(c) Method 3
(d) Method 4
(e) Method 5
(f) Method 6
Figure 3.5: Results for bivariate additive noise simulations. The histograms of the true data $X_2$ versus 10 sets of synthesis $Y_2$ when $g_1 = g_2 = 0.5$. 
Figure 3.6: The posterior predictive densities of guessing $X_{2,[1]} = 6.3$ for $g_1 = g_2 = 0.1$ when $D_1$ is known in Method 2 and $(D_1, D_2)$ are known in Method 3. When $(D_1, D_2)$ are known in Method 4, 5, and 6, the plots are deleted because their posterior disclosure risks are uniformly distributed according to Section 3.3.1.
Figure 3.7: The posterior predictive densities of guessing $X_{2,[1]} = 6.7$ for $g_1 = g_2 = 0.1$ when $(D_1, D_2)$ are unknown.
Figure 3.8: The posterior predictive densities of guessing $X_{2,[1]} = 6.7$ for $g_1 = g_2 = 0.5$ when $D_1$ is known in Method 2 and $(D_1, D_2)$ are known in Method 3. When $(D_1, D_2)$ are known in Method 4, 5, and 6, the plots are deleted because their posterior disclosure risks are uniformly distributed according to Section 3.3.1.
Figure 3.9: The posterior predictive densities of guessing $X_{2,[1]} = 6.7$ for $g_1 = g_2 = 0.1$ when $(D_1, D_2)$ are unknown.
Figure 3.10: Results for univariate multiplicative noise simulations. The histograms of the true data $X$ versus 10 sets of synthesis $Y$. 

(a) Method 1

(b) Method 2 when $(g^{(a)} = 5, g^{(b)} = 25)$

(c) Method 3 when $(g^{(a)} = 5, g^{(b)} = 25)$

(d) Method 4 when $(g^{(a)} = 5, g^{(b)} = 25)$

(e) Method 2 when $(g^{(a)} = 15, g^{(b)} = 25)$

(f) Method 3 when $(g^{(a)} = 15, g^{(b)} = 25)$

(g) Method 4 when $(g^{(a)} = 15, g^{(b)} = 25)$
Figure 3.11: The posterior predictive densities of guessing $X_{[1]} = 6$ for $(g^{(a)} = 5, g^{(b)} = 25)$. When $D$ is known, the plots of Method 3 and 4 are deleted because the posterior disclosure risks are uniformly distributed according to (3.3) and (3.4).
(a) Method 1

(b) Method 2 when $D$ is known

(c) Method 2 when $D$ is unknown

(d) Method 3 when $D$ is unknown

(e) Method 4 when $D$ is unknown

**Figure 3.12:** The posterior predictive densities of guessing $X_{[1]} = 6$ for $(g^{(a)} = 15, g^{(b)} = 25)$. When $D$ is known, the plots of Method 3 and 4 are deleted because their posterior disclosure risks are uniformly distributed according to (3.3) and (3.4).
Figure 3.13: Results for bivariate multiplicative noise simulations. The histograms of the true data $X_2$ versus 10 sets of synthesis $Y_2$ when $g_1^{(a)} = g_2^{(a)} = 5$ and $g_1^{(b)} = g_2^{(b)} = 25$. 

82
Figure 3.14: Results for bivariate multiplicative noise simulations. The histograms of the true data $X_2$ versus 10 sets of synthesis $Y_2$ when $g_1^{(a)} = g_2^{(a)} = 15$ and $g_1^{(b)} = g_2^{(b)} = 25$. 
Figure 3.15: The posterior predictive densities of guessing $X_{2,[1]} = 6.7$ for $g^{(a)}_1 = g^{(a)}_2 = 5$ and $g^{(b)}_1 = g^{(b)}_2 = 25$ when $D_1$ is known in Method 2 and $(D_1, D_2)$ are known in Method 3. When $(D_1, D_2)$ are known in Method 4, 5, and 6, the plots are deleted because their posterior disclosure risks are uniformly distributed according to Section 3.3.1.
Figure 3.16: The posterior predictive densities of guessing $X_{2,1} = 6.7$ for $g_{1}^{(a)} = g_{2}^{(a)} = 5$ and $g_{1}^{(b)} = g_{2}^{(b)} = 25$ when $(D_1, D_2)$ are unknown.
Figure 3.17: The posterior predictive densities of guessing $X_{2|[1]} = 6.7$ for $g^{(a)}_1 = g^{(a)}_2 = 15$ and $g^{(b)}_1 = g^{(b)}_2 = 25$ when $D_1$ is known in Method 2 and $(D_1, D_2)$ are known in Method 3. When $(D_1, D_2)$ are known in Method 4, 5, and 6, the plots are deleted because their posterior disclosure risks are uniformly distributed according to Section 3.3.1.
Figure 3.18: The posterior predictive densities of guessing $X_{2[1]} = 6.7$ for $g_1^{(a)} = g_2^{(a)} = 15$ and $g_1^{(b)} = g_2^{(b)} = 25$ when $(D_1, D_2)$ are unknown.
4.1 Introduction

Nonparametric Bayesian (NB) mixture modelling is a flexible method to model complicated data (Müller and Quintana, 2004; Dunson and Xing, 2009; Müller and Mitra, 2013; DeYoreo and Kottas, 2014; DeYoreo et al., 2015a), including multiple imputation for missing data. As examples, Si and Reiter (2013) use the Dirichlet process mixture of products of multinomial (DPMPM) model to impute multivariate unordered categorical variables. Hu et al. (2016) use a nested DPMPM to impute unordered categorical survey variables from nested household heads. Manrique-Vallier and Reiter (2013, 2014) use a discrete multivariate truncated latent structure models to impute categorical data with structural zeros. Kim et al. (2014, 2015) use a constrained Dirichlet process mixture of multivariate normals to impute continuous data with linear constraints.

Although NB mixture modelling allows the data to determine the complexity of
the model, it typically involves a strong assumption of local independence. This can have negative effects on estimation (Banerjee et al., 2013; Murray and Reiter, 2016). For example, the NB mixture modelling can demand a large number of mixture components when the joint distribution of categorical variables do not follow an independence structure. When this occurs, the estimated joint distribution may be dominated by a few variables resulting in unreliable estimation.

To avoid assuming local independence, Murray and Reiter (2016) propose the hierarchical coupled mixture model with local independence (HCMM-LD). The model combines two Dirichlet processes (DP), one for multivariate continuous data and the other for multivariate categorical data (Dunson and Xing, 2009), through a hierarchical tensor factorization. The main advantage of this model is that it allows local dependence between continuous and categorical variables. This can alleviate the burden of proliferating mixture components under the strong assumption of local independence. HCMM-LD has been successfully applied to multiply impute the missing values in samples from the Survey of Income and Program Participation (Murray and Reiter, 2016).

Missing information does not tend to be uniformly scattered across all the survey variables. Often missing values are concentrated on a few variables with other variables nearly or completely observed. In this study, we call variables that have a high rate of missingness as focus variables, labeled as $F$. Focus variables also can be the variables that the users are most interested in. We call the remaining variables with few or no missing values as non-focus variables, labeled as $NF$.

When many survey variables can be treated as $NF$ variables, the number of $F$ variables can be significantly smaller than that of $NF$. This type of missingness pattern suggests that there is room to improve the NB imputation approaches for multivariate data. Imputing all the missingness, including those which are trivial to the users, by complex nonparametric Bayesian MI could be inefficient. The situation
could become even worse when the number of almost fully observed $NF$ is big, as NB mixture modelling has the potential to underestimate the distribution of a small number of variables in complicated data (DeYoreo et al., 2015b; Hannah et al., 2011; Wade et al., 2014). When this happens, $NF$ could dominate the entire model structure which can lead to a poor model fitted on $F$.

In this chapter, I present a model called the *Hierarchical Coupled Mixture Model with Focused and Non-Focused Clustering (HCMM-FNF)*. It is an extension model of *HCMM-LD* (Murray and Reiter, 2016). The basic idea is to use one set of cluster indicators for $F$ variables and a second set for $NF$ variables, and connect the two sets using a tensor factorization prior. The major difference between these two models is that *HCMM-LD* uses one model for the continuous and the categorical variables, whereas *HCMM-FNF* uses separate sub-models for $F$ and $NF$ variables.

Additionally, for the *HCMM-FNF* model I describe how to effectively borrow information from $NF$ to further assist in capturing the distribution of $F$. Carefully choosing some of the $NF$ variables to be locally clustered with $F$ variables can help the model better estimate associations among $F$. We call this method the strategy of *Move*. We consider a way of selecting a to-be-moving $NF$ variable based upon the degree of dependence between it and the $F$ variables. We compare the improvement *Move* can provide against the strategy of *Stay*, where we do not move any $NF$ variables to the side of $F$.

The remainder of this chapter is organized as follows. In Section 4.2, we mathematically and graphically present the model with emphasis on the focused and non-focused clustering. Then, we introduce the motivation for the strategy of *Move* from the model structure. In Section 4.3, we describe eight simple scenarios that explore the potential benefits of the strategy of *Move*. The results from the simulation studies show advantages of *Move* over *Stay* for some scenarios. In Section 4.4, we apply *HCMM-FNF* to data from the American Community Survey. We compare the
performance of Move versus Stay against the reference model HCMM-LD. Basically, the results are consistent with the simulation studies. In Section 4.5, we summarize the model properties and the benefit of Move for HCMM-FNF. We also discuss some future research directions.

4.2 Model specification and properties

4.2.1 Model specification

We denote $Y$ as continuous variables and $X$ as categorical variables. We use a superscript $F$ to denote focus variables and the superscript $NF$ to denote non-focus variables. Thus, $Y^{(F)}$, $X^{(F)}$, $Y^{(NF)}$ and $X^{(NF)}$ are the focus continuous, focus categorical, non-focus continuous, and non-focus categorical variables, respectively. We assume there are $n$ observations. For each observation $i$, we have $Y^{(F)}_i = (Y^{(F)}_{i1}, \ldots, Y^{(F)}_{iq(F)})^T$, $X^{(F)}_i = (X^{(F)}_{i1}, \ldots, X^{(F)}_{iq(F)})^T$, $Y^{(NF)}_i = (Y^{(NF)}_{i1}, \ldots, Y^{(NF)}_{iq(NF)})^T$, and $X^{(NF)}_i = (X^{(NF)}_{i1}, \ldots, X^{(NF)}_{iq(NF)})^T$. Let $D_i$ be a regression design matrix containing the main effects of $X^{(F)}_i$, $Y^{(NF)}_i$, and $X^{(NF)}_i$ with the dimension of $\mathcal{R}^{1 \times p^*}$. A similar regression approach is used by Murray and Reiter (2016). The model of HCMM-FNF can be described as follows.
(\(Y_i^{(F)} = y_i^{(F)}\) | \(X_i^{(F)} = x_i^{(F)}\), \(Y_i^{(NF)} = y_i^{(NF)}\), \(X_i^{(NF)} = x_i^{(NF)}\), \(H_i^{(FY)} = a, -\))
\[\sim \mathcal{N}(y_i^{(F)} | D_i B_a^{(F)}, \Sigma_a^{(F)})\]  
(4.1)

\((X_i^{(F)} = x_i^{(F)} | H_i^{(FX)} = b, -) \sim \prod_{j=1}^{p(F)} \psi_{b,x_{ij}}^{(F)}(j)\)  
(4.2)

\((Y_i^{(NF)} = y_i^{(NF)} | H_i^{(NF)} = h, -) \sim \mathcal{N}(y_i^{(NF)} | B_h^{(NF)}, \Sigma_h^{(NF)})\)  
(4.3)

\((X_i^{(NF)} = x_i^{(NF)} | H_i^{(NF)} = h, -) \sim \prod_{j=1}^{p(NF)} \psi_{z,x_{ij}}^{(NF)}(j)\)  
(4.4)

\(Pr(H_i^{(fy)} = a, H_i^{(fx)} = b | Z_i = z) = \phi_{z,a}^{(FY)} \phi_{z,b}^{(FX)}\)  
(4.5)

\(Pr(H_i^{(NF)} = h | Z_i = z) = \phi_{z,h}^{(NF)}\)  
(4.6)

\(Pr(Z_i = z) = \lambda_z\).  
(4.7)

Here, \(H_i^{(FY)}\) is the mixture component index of \(Y_i^{(F)}\) with the support of \(H_i^{(FY)} \in \{1, \cdots, k^{(FY)}\}\). \(H^{(FX)}\) is the mixture component index of \(X_i^{(F)}\) with the support of \(H_i^{(FX)} \in \{1, \cdots, k^{(FX)}\}\). \(H_i^{(NF)}\) is the mixture component index of \(Y_i^{(NF)}\) and \(X_i^{(NF)}\) with the support of \(H_i^{(NF)} \in \{1, \cdots, k^{(NF)}\}\). \(Z_i\) is the mixture component index of \(H_i^{(F)}\) and \(H_i^{(NF)}\) with the support of \(W \in \{1, \cdots, k^{(Z)}\}\). \(B_a^{(F)}\) and \(\Sigma_a^{(F)}\) are the matrix of regression coefficients and the covariance matrix for \(H_i^{(FY)}\) in the latent cluster \(a\). \(\psi_{b,x_{ij}}^{(F)}\) is the probability of \(X_{ij}^{(F)} = x_{ij}^{(F)}\) for \(H_i^{(FX)}\) in the latent cluster \(b\). \(B_h^{(NF)}\) and \(\Sigma_h^{(NF)}\) are the mean vector and the covariance matrix for \(H_i^{(NF)}\) in the latent cluster \(h\). Here, \(\Sigma_h^{(NF)}\) is a diagonal matrix of \(Diag(\eta_{h,1}^{(NF)}, \cdots, \eta_{h,q}^{(NF)})\). Thus, the variables in \(Y_i^{(NF)}\) are conditionally independently. \(\psi_{h,x_{ij}}^{(NF)}\) is the probability of \(X_{ij}^{(NF)} = x_{ij}^{(NF)}\) for \(H_i^{(FX)}\) in the latent cluster \(h\). More details about prior specifications and full conditionals can be found in Appendix B.1.
4.2.2 The property of focused and non-focused clustering

A graphical representation of $HCMM-FNF$ is shown in Figure 4.1. It shows that the model has a hierarchical structure for the NB mixture components. This hierarchical structure gives the model relative freedom to cluster $F$ and $NF$ separately. In the $N\mathcal{F}$ clustering, $X_i^{(NF)}$ and $Y_i^{(NF)}$ are conditionally independent within a given mixture component $H_i^{(NF)}$. We let $X_i^{(NF)}$ follow a $DPMPM$ and $Y_i^{(NF)}$ follow a mixture of multivariate normals with diagonal covariance matrix. On the contrary, in the $F$ clustering $X_i^{(F)}$ and $Y_i^{(F)}$ are dependent within a given mixture component $H_i^{(F)}$. Although $X_i^{(F)}$ still follows a $DPMPM$, $Y_i^{(F)}$ follows a mixture of multivariate normals with its mean regressed on the main effects of $X_i^{(F)}$, $X_i^{(NF)}$, and $Y_i^{(NF)}$ and its covariance matrix being non-diagonal.

![Graphical model representation of HCMM-FNF](image)

**Figure 4.1:** Graphical model representation of $HCMM-FNF$. $X^{(F)}$, $Y^{(F)}$, $X^{(NF)}$, and $Y^{(NF)}$ are the observed categorical and continuous variables. $H^{(F)}$ and $H^{(NF)}$ are the mixture components of $F$ and $N\mathcal{F}$ variables, respectively. $Z$ is the mixture component for the whole record.

Thus, the structures of $N\mathcal{F}$ and $F$ clustering are distinct with the clustering of $F$ being more complicated. This model structure could be preferred for a small number of $F$ variables with a high missing rate in a complex dataset. We show the
joint distribution of HCMM-FNF in (4.8). It is a mixture of $\mathcal{F}$ clustering and $\mathcal{NF}$ clustering. The $\mathcal{F}$ clustering is similar to the joint distribution of HCMM-LD in the sense that it is a product of a mixture of multivariate normals and a $\text{DPMPM}$ model with local dependence. The $\mathcal{NF}$ clustering is the joint distribution of a mixture model with the assumption of local independence in all $\mathcal{NF}$ variables.

$$
Pr(Y_i^{(F)} = y_i^{(F)}, X_i^{(F)} = x_i^{(F)}, Y_i^{(NF)} = y_i^{(NF)}, X_i^{(NF)} = x_i^{(NF)})
$$

$$
= \sum_{z=1}^{k^{(F)}} \sum_{h=1}^{k^{(NF)}} \sum_{a=1}^{k^{(FY)}} \sum_{b=1}^{k^{(FX)}} Pr((y_i^{(F)}|x_i^{(F)}, y_i^{(NF)}, x_i^{(NF)})|H_i^{(FY)}) = a, -) \cdot Pr(x_i^{(F)}|H_i^{(FX)} = b, -) \cdot Pr(y_i^{(NF)}|H^{(NF)} = h, -) \cdot Pr(x_i^{(NF)}|H_i^{(NF)} = h, -) \cdot Pr(H_i^{(FY)} = a, H_i^{(FX)} = b|Z_i = z) \cdot Pr(H_i^{(NF)} = h|Z_i = z) \cdot Pr(Z_i = z)
$$

$$
= \sum_{z=1}^{k^{(F)}} \left\{ \sum_{a=1}^{k^{(FY)}} \left\{ \sum_{b=1}^{k^{(FX)}} \phi_{z,a}^{(FY)} N(Y_i^{(F)}|D_i B_a^{(F)}, \Sigma_a^{(F)}) \cdot \sum_{b=1}^{k^{(FX)}} \phi_{z,b}^{(FX)} \prod_{j=1}^{p^{(F)}} \psi_{b,x_{ij}}^{(j)} \right\} \right\} \sum_{h=1}^{k^{(NF)}} \phi_{z,h}^{(NF)} N(Y_i^{(NF)}|B_h^{(NF)}, \Sigma_h^{(NF)}) \cdot \prod_{j=1}^{p^{(NF)}} \psi_{h,x_{ij}}^{(j)} \lambda_z. \ \ \ \ (4.8)
$$

The conditional distribution of $Y_i^{(F)}$ given the other variables is a mixture of multivariate normals as in (4.9)

$$
Pr(Y_i^{(F)} = y_i^{(F)}|X_i^{(F)} = x_i^{(F)}, Y_i^{(NF)} = y_i^{(NF)}, X_i^{(NF)} = x_i^{(NF)}, \Theta)
$$

$$
= \sum_{z=1}^{k^{(F)}} \sum_{a=1}^{k^{(FY)}} \frac{w_{z,a}(x_i^{(F)}, y_i^{(NF)}, x_i^{(NF)})}{\sum_{z'=1}^{k^{(F)}} \sum_{a'=1}^{k^{(FY)}} w_{z',a'}(x_i^{(F)}, y_i^{(NF)}, x_i^{(NF)})} \cdot N(Y_i^{(F)}|D_i B_a^{(F)}, \Sigma_a^{(F)}), \ \ \ \ (4.9)
$$
in which

\[
\begin{align*}
\quad w_{z,a}(x_i^{(F)}, y_i^{(NF)}, x_i^{(NF)}) \\
= \phi_{z,a}^{(FY)} \lambda_z \cdot \sum_{b=1}^{k^{(FX)}} \phi_{z,b}^{(FX)} \prod_{j=1}^{p^{(F)}} \psi_{b,z_{ij}}^{(j)} \cdot \sum_{h=1}^{k^{(NP)}} \phi_{z,h}^{(NF)} N(y_i^{(NF)} | B_h^{(NF)}, \Sigma_h^{(NF)}) \prod_{j=1}^{p^{(N)}} \psi_{h,x_{ij}^{(NF)}}^{(j)}.
\end{align*}
\]

(4.10)

The component-specific mean, \( D_i B_a^{(F)} \), implies that \( Y_i^{(F)} \) and the other variables are not locally independent within \( H_i^{(FY)} = a \). In addition, the local dependence structure can capture nonlinear effects not encoded in \( D_i \). That’s because

\[
E(Y_i^{(F)} | x_i^{(F)}, Y_i^{(NF)} = y_i^{(NF)}, X_i^{(NF)} = x_i^{(NF)}, \Theta) = \sum_{z=1}^{k^{(Z)}} \sum_{a=1}^{k^{(FY)}} w_{z,a}(x_i^{(F)}, y_i^{(NF)}, x_i^{(NF)}) D_i B_a^{(F)},
\]

(4.11)

with

\[
\sum_{z=1}^{k^{(Z)}} \sum_{a=1}^{k^{(FY)}} w_{z,a}(x_i^{(F)}, y_i^{(NF)}, x_i^{(NF)}) = 1
\]

being a nonlinear function of \((x_i^{(F)}, y_i^{(NF)}, x_i^{(NF)})\).

The joint distribution of \( X_i^{(F)} \) and the \( N \! F \) variables given the mixture component \( Z_i = z \) is

\[
Pr(X_i^{(F)} = x_i^{(F)}, Y_i^{(NF)} = y_i^{(NF)}, X_i^{(NF)} = x_i^{(NF)} | Z_i = z, \Theta) = \left\{ \begin{array}{ll}
\sum_{b=1}^{k^{(FX)}} \phi_{z,b}^{(FX)} \prod_{j=1}^{p^{(F)}} \psi_{b,x_{ij}}^{(j)} & \quad \text{DPMMPM} \\
\sum_{h=1}^{k^{(NP)}} \phi_{z,h}^{(NF)} N(Y_i^{(NF)} | B_h^{(NF)}, \Sigma_h^{(NF)}) \prod_{j=1}^{p^{(N)}} \psi_{h,x_{ij}^{(NF)}}^{(j)} & \quad \text{\( N \! F \) clustering}
\end{array} \right\}.
\]

(4.12)

Clearly, \( X_i^{(F)} \) and \( N \! F \) \(_i\) are locally independent given \( Z_i = z \). However, their associations can be captured by integrating out \( Z_i \). To see this, the conditional distribution
of $X_i^{(F)}$ given the $NF$ variables is

\[
Pr(X_i^{(F)} = x_i^{(F)} | Y_i^{(NF)} = y_i^{(NF)}, X_i^{(NF)} = x_i^{(NF)}, \Theta) = \sum_{z=1}^{k^{(Z)}} \sum_{z'=1}^{k^{(Z)}} w_z(y_i^{(NF)}, x_i^{(NF)}) \left\{ \sum_{b=1}^{k^{(FX)}} \phi_{z,b}^{(FX)} \prod_{j=1}^{p^{(F)}} \psi_{j,bx_{ij}}^{(F)} \right\}_{DPMPM},
\]

in which $w_z(y_i^{(NF)}, x_i^{(NF)}) = \left\{ \sum_{h=1}^{k^{(NF)}} \phi_{z,h}^{(NF)} N(Y_i^{(NF)} \mid B_h^{(NF)}, \Sigma_h^{(NF)}) \prod_{j=1}^{p^{(NF)}} \psi_{j,hx_{ij}}^{(NF)} \right\}_{NF \text{ clustering}} \lambda_z.$

Thus, the influence of the $NF$ variables only appears in the weights.

For simplicity, we suppress the relevant model parameters in $\Theta$ in (4.8), (4.9), (4.12), and (4.13).

4.2.3 The motivation of the strategy of Move

Since $Y^{(NF)}$ are almost fully observed, bringing highly associated $Y^{(NF)}$ to the side of $Y^{(F)}$ may help the model to efficiently borrow useful information for $Y^{(F)}$ within latent class $H^{(FY)}$. This strategy could be especially useful when the associations between $Y^{(F)}$ and $Y^{(NF)}$ are nonlinear. For the same reason, if $Y^{(F)}$ and $X^{(NF)}$ are highly associated, moving $X^{(NF)}$ to the side of $X^{(F)}$ may help the model improve the predictions of the associations between $Y^{(F)}$ and $X^{(NF)}$.

Figure 4.1 does not show any direct connections between $X^{(F)}$ and $\{NF\}$. This implies that $X^{(F)}$ and $\{NF\}$ are locally independent within latent class $Z$. The strong assumption of local independence may cause problems if the true joint probability cannot be approximated by a mixture of local independent distributions. Thus, if some variables in $Y^{(NF)}$ are highly associated with $X^{(F)}$, bringing them to the side of $Y^{(F)}$ can allow $X^{(F)}$ to borrow information more efficiently within the local latent class $Z$. Likewise, if some variables in $X^{(NF)}$ are highly associated with $X^{(F)}$,
moving them to the side of $X^{(F)}$ can better model the associations between them within latent cluster $H^{(FX)}$.

These arguments suggest that the strategy of Move has the potential to improve the associations between $F$ variables with the highly associated $NF$ ones. We now investigate this potential using simulation studies.

4.3 Simulation Studies

From Section 4.2, it seems that moving the highly associated $NF$ variables to the side of $F$ may improve the estimates of the related associations. However, we also know that the model is complicated in the sense that it has focused and non-focused clustering with local dependence between $Y_{i}^{(F)}$ and $NF_f$. Thus, it is challenging to study the strategy of Move for all the highly-associated $NF$ variables simultaneously. Therefore, we design simple simulations that zoom in on situations when one type of $F$ variables is highly associated with the other type of $NF$ variables.

4.3.1 Scenarios

We design eight scenarios in which any of the four associations, i.e., $(X^{(F)}, X^{(NF)})$, $(X^{(F)}, Y^{(NF)})$, $(Y^{(F)}, X^{(NF)})$, and $(Y^{(F)}, Y^{(NF)})$, may be improved by the strategy of Move (see Table 4.1).

Scenario “$X^{(F)}$ and $X^{(NF)}$”

We begin with the scenario in which some $X^{(NF)}$ are highly associated with $X^{(F)}$ as evident in Figure 4.2. Here, we have six binary $X^{(NF)}$ variables generated from a DPMPM model. We have them heavily depend on the first four $X^{(NF)}$ variables according to

$$Pr(X_j^{(F)} = x | X_j^{(NF)} = x) = 0.9,$$  \hspace{1cm} (4.14)
Table 4.1: High-level description of scenarios and quantities of interest in the simulation studies. In scenario “\(X^{(F)}\) and \(X^{(NF)}\)”, we are interested in the accuracies of the bivariate cell frequencies between \(X^{(F)}\) and the associated \(X^{(NF)}\). In scenario “\(X^{(F)}\) and \(Y^{(NF)}\)”, we are interested in the accuracies of the dependencies between \(X^{(F)}\) and the associated \(Y^{(NF)}\). In scenario “\(Y^{(F)}\) and \(X^{(NF)}\)”, we are interested in the accuracies of the dependencies between \(Y^{(F)}\) and the associated \(X^{(NF)}\). In scenario “\(Y^{(F)}\) and \(Y^{(NF)}\)”, we are interested in the accuracies of the correlation coefficients between \(Y^{(F)}\) and the associated \(Y^{(NF)}\).

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Description</th>
<th>Key Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>(X^{(F)}) and (X^{(NF)})</td>
<td>Move (X^{(NF)}) to (X^{(F)})</td>
<td>(Pr(X^{(F)}, X^{(NF)}))</td>
</tr>
<tr>
<td>(X^{(F)}) and (Y^{(NF)})</td>
<td>Move (Y^{(NF)}) to (Y^{(F)})</td>
<td>(Pr(X^{(F)}, Y^{(NF)}))</td>
</tr>
<tr>
<td>(Y^{(F)}) and (X^{(NF)})</td>
<td>Move (X^{(NF)}) to (X^{(F)})</td>
<td>(Pr(Y^{(F)}, X^{(NF)}))</td>
</tr>
<tr>
<td>(Y^{(F)}) and (Y^{(NF)})</td>
<td>Move (Y^{(NF)}) to (Y^{(F)})</td>
<td>(Pr(Y^{(F)}, Y^{(NF)}))</td>
</tr>
</tbody>
</table>

with \(x\) equal to either 1 or 2 for \(j = 1, \ldots, 4\). Thus, \(X^{(NF)}_{1:4}\) are highly associated with \(X^{(F)}_{1:4}\). Under the strategy of Move, we bring the first four non-focus categorical variables to the side of \(X^{(F)}\).

**Figure 4.2:** Scenario “\(X^{(F)}\) and \(X^{(NF)}\)”. All four focus categorical variables \(X^{(F)}_{1:4}\) depend on the first four non-focus categorical variables \(X^{(NF)}_{1:4}\) by (4.14). Under the strategy of Move, we move \(X^{(NF)}_{1:4}\) to the side of \(X^{(F)}_{1:4}\).

**Scenario “\(X^{(F)}\) and \(Y^{(NF)}\)”**

Next, we are interested in some \(Y^{(NF)}\) being highly associated with \(X^{(F)}\) as evident in Figure 4.3. Here, we have six \(Y^{(NF)}\) variables generated from a mixture of normal distributions. Then, we let the binary \(X^{(F)}\) depend on the first four of \(Y^{(NF)}\) such that

\[
\log \left( \frac{Pr(X_j^{(F)} = 2|Y_j^{(NF)} = y_j^{(NF)})}{Pr(X_j^{(F)} = 1|Y_j^{(NF)} = y_j^{(NF)})} \right) = y_j^{(NF)},
\]

(4.15)
with $X_j^{(F)}$ equal to either 1 or 2 for $j = 1, \cdots, 4$. Under Move, we move $Y_{1:4}^{(NF)}$ to the side of $Y^{(F)}$.

**Figure 4.3:** Scenario “$X^{(F)}$ and $Y^{(NF)}$”. All four focus categorical variables $X_j^{(F)}$ depend on the first four non-focus continuous variables $Y_{1:4}^{(NF)}$ by (4.15). Under the strategy of Move, we bring $Y_{1:4}^{(NF)}$ to the side of $Y^{(F)}$.

**Scenario “$Y^{(F)}$ and $X^{(NF)}$”**

In “$Y^{(F)}$ and $X^{(NF)}$”, we have some $X^{(NF)}$ highly associated with $Y^{(F)}$. The six binary $X^{(NF)}$ variables are generated from a DPMPM model. Then, $Y^{(F)}$ depend on the first four of $X^{(NF)}$ according to

$$
(Y_j^{(F)}|X_j^{(NF)} = x_j^{(NF)}) \sim \mathcal{N}(x_j^{(NF)}, 0.005),
$$

with $j = 1, \cdots, 4$.

(4.16) is a case when $X^{(NF)}$ and $Y^{(F)}$ are highly and linearly associated. We are also interested in the situation when they are highly and nonlinearly associated. That’s because the direct link between $X^{(NF)}$ and $Y^{(F)}$ is linear in HCMM-FNF and we would like to test some scenarios that violate this model assumption. In (4.17), we have

$$
(log(Y_j^{(F)})|X_j^{(NF)} = x_j^{(NF)}) \sim \mathcal{N}(x_j^{(NF)}, 0.005),
$$

where the mean of $log(Y^{(F)})$ depends on the first four $X^{(NF)}$.

We also make this nonlinear association more complicated by

$$
log\left(\frac{Pr(X_j^{(NF)} = 2|Y_j^{(F)} = y_j^{(F)})}{Pr(X_j^{(NF)} = 1|Y_j^{(F)} = y_j^{(F)})}\right) = y_j^{(F)},
$$

with $X_j^{(NF)}$ equal to either 1 or 2 for $j = 1, \cdots, 4$. 99
Figure 4.4: Scenario “$Y^{(F)}$ and $X^{(NF)}$”. All four focus continuous variables $Y_{1:4}^{(F)}$ depend on the first four non-focus categorical variables $X_{1:4}^{(NF)}$ by (4.16), (4.17) or (4.18). Under the strategy of Move, we move $X_{1:4}^{(NF)}$ to the side of $X^{(F)}$.

For scenario “$Y^{(F)}$ and $X^{(NF)}$”, we bring the first four $X^{(F)}$ variables to the side of $X^{(F)}$ under the strategy of Move (see Figure 4.4).

Scenario “$Y^{(F)}$ and $Y^{(NF)}$”

In “$Y^{(F)}$ and $Y^{(NF)}$”, $Y^{(NF)}$ variables are highly associated with the first four variables $Y^{(F)}$. Again, $Y^{(NF)}$ variables are generated from a mixture of normal distributions. The associations between them are

$$ (Y_j^{(F)} | Y_j^{(NF)} = y_j^{(NF)}) \sim \mathcal{N}(0.9 \times y_j^{(NF)}, 0.005), $$

with $j = 1, \ldots, 4$.

We would also like to know the effectiveness of Move when $Y^{(F)}$ and $Y^{(NF)}$ are highly and nonlinearly associated. In (4.20), we have

$$ (log(Y_j^{(F)}) | Y_j^{(NF)} = y_j^{(NF)}) \sim \mathcal{N}(0.9 \times y_j^{(NF)}, 0.08), $$

with $j = 1, \ldots, 4$.

In (4.21), we have

$$ (Y_j^{(F)} | Y_j^{(NF)} = y_j^{(NF)}) \sim \mathcal{N}(0.9 \times (y_j^{(NF)} + y_j^{(NF)^2}), 0.08), $$

with $j = 1, \ldots, 4$.

For “$Y^{(F)}$ and $Y^{(NF)}$”, we bring the first four $Y^{(NF)}$ variables to the side of $Y^{(F)}$ under the strategy of Move (see Figure 4.5).
4.3.2 Metrics for Comparison

We have two metrics for the comparisons. One metric is based on the absolute difference between the estimated quantity and the truth. The other relies on the percentage change. Let \( q^{(s)}_{k,j,l} \) be the \( k^{th} \) quantity of interest in the \( j^{th} \) repeated sample for the \( l^{th} \) imputation. The superscript \( (s) \) indicates that this estimate is from the strategy of Stay. Similarly, we have \( q^{(m)}_{k,j,l} \) for the estimate obtained by the strategy of Move. Notations without any superscript, such as \( q_{k,j} \), stand for the quantities from the truth defined as the complete data without any missing values.

**Metric I:** The absolute differences are \( d^{(s)}_{k,j,l} = |q^{(s)}_{k,j,l} - q_{k,j}| \) for Stay and \( d^{(m)}_{k,j,l} = |q^{(m)}_{k,j,l} - q_{k,j}| \) for Move. We take the mean of these quantities over all the imputations, \( d^{(s)}_{k,j} = \frac{1}{L} \sum_{l=1}^{L} d^{(s)}_{k,j,l} \) and \( d^{(m)}_{k,j} = \frac{1}{L} \sum_{l=1}^{L} d^{(m)}_{k,j,l} \). For a fixed quantity of interest \( k \), \( d^{(s)}_{k,j} \) are identically and independently distributed for all repeated studies \( m \). We can do a paired t-test for the hypothesis of \( H_0 : \mu^{(s)}_k = \mu^{(m)}_k \). Here, \( \mu^{(s)}_k \) is the population mean of \( d^{(s)}_{k,j} \) and \( \mu^{(m)}_k \) is the population mean of \( d^{(m)}_{k,j} \). If the p-value of this t-test is below 0.01, we consider the difference between Stay and Move for this particular quantity of interest is statistically significant.

**Metric II:** The percentage changes are \( \Delta d^{(s)}_{k,j,l} = \frac{q^{(s)}_{k,j,l} - q_{k,j}}{q_{k,j}} \times 100\% \) for Stay and \( \Delta d^{(m)}_{k,j,l} = \frac{q^{(m)}_{k,j,l} - q_{k,j}}{q_{k,j}} \times 100\% \) for Move. Due to the fact that the true quantities of interest are usually not in the same range, this metric can help us better understand
the relative difference. Again, we take $\Delta d_k^{(s)} = \frac{1}{JL} \sum_{j=1}^{J} \sum_{l=1}^{L} \Delta d_{k,m,t}^{(s)}$ and $\Delta d_k^{(m)} = \frac{1}{JL} \sum_{j=1}^{J} \sum_{l=1}^{L} \Delta d_{k,m,t}^{(m)}$. We draw side-by-side boxplots for all $\{\Delta d_k^{(s)} : k \in K\}$ and $\{\Delta d_k^{(m)} : k \in K\}$. Here, $k \in K$ means that these quantities of interest belong to the same type $K$. For example, if we draw a side-by-side boxplot of all $\{\Delta d_k^{(s)} : k \in K\}$ and $\{\Delta d_k^{(m)} : k \in K\}$ for $\text{Cor}(Y^{(F)}, Y^{(NF)})$, then $K$ stands for all the possible correlations between $Y^{(F)}$ and $Y^{(NF)}$.

### 4.3.3 Results

We generate 100 independent datasets from each scenario. Each dataset has $n = 1,000$ complete observations, in which we choose 1% of values for non-focus variables and 50% of values for focus variables to be missing completely at random. Then, we fit the $HCMM-FNF$ model with the strategy of $Stay$ and $Move$. We run the MCMC chains for 25,000 iterations as its burn-in. After that, we perform the standard diagnosis of MCMC convergence. Finally, we keep $L = 10$ imputations spaced 100 iterations apart. These simulation studies are conducted in R batch mode on the Duke Shared Cluster Recourse.

In summary, the results from all the simulation studies show that (1) moving highly associated $\mathcal{NF}$ variables to the side of $\mathcal{F}$ can improve inferences for the quantities of interest related with the associated $X^{(F)}$, and (2) no matter whether these associations are linear or nonlinear, the strategy of move does not substantially improve the inferences for those quantities related with the associated $Y^{(F)}$.

**Scenario “$X^{(F)}$ and $X^{(NF)}$”**

Figure 4.6 displays the results from “$X^{(F)}$ and $X^{(NF)}$”. The plot on the left is from Metric I and the one on the right is from Metric II. In the left plot, a symbol below the diagonal means that an absolute difference under $Stay$ is greater than that under $Move$. In the right plot, an imputation model is preferred if the corresponding
boxplot has a median close to zero with a narrow spread. Thus, from the left plot we can see that the cell probabilities from the two-way tables of $X^{(F)}$ and $X^{(NF)}$ can be better estimated under Move. The improved counts are from the two-way tables of $X_j^{(NF)}$ and $X_j^{(F)}$ for $j = 1, \cdots, 4$. At the same time, the two boxplots show that the percentage change under Move has a smaller spread than that under Stay. Therefore, the strategy of Move is preferred in “$X^{(F)}$ and $X^{(NF)}$”.

**Scenario “$X^{(F)}$ and $Y^{(NF)}$”**

In “$X^{(F)}$ and $Y^{(NF)}$”, the quantities of interest are the coefficients of the logistic regression between $X^{(F)}$ and $Y^{(NF)}$. That is, we regress the logit of each $X^{(F)}$ variables upon any one of $Y^{(NF)}$ variables. Figure 4.7 displays that these coefficients can be better estimated by the strategy of Move. The plot on the left shows that the absolute differences under Stay are significantly bigger than those under Move for the associations between $X^{(F)}$ and $Y^{(NF)}$. Again, the improved logistic regression coefficients are mostly from the associations between $X_j^{(F)}$ and $Y_j^{(NF)}$ for $j = 1, \cdots, 4$. The side-by-side boxplot also shows that the median of the percentage change under Move is closer to zero than that under Stay.

**Scenario “$Y^{(F)}$ and $X^{(NF)}$”**

In “$Y^{(F)}$ and $X^{(NF)}$”, we are interested in the associations between $Y^{(F)}$ and $X^{(NF)}$. There are several ways to measure such associations. For example, one can regress a $Y^{(F)}$ on the associated $X^{(NF)}$ and then compare the regression coefficients between Stay and Move. We choose to measure these associations by the logistic regression of $X^{(NF)}$ given $Y^{(F)}$ without the intercept, so that the associations can be reflected by just one regression coefficient. In detail, we compare the coefficients of $\text{logit}(X_j^{(NF)})$ regressed upon $Y_j^{(F)}$ for $j^{(X)} \in \{1, 2, 3, 4\}$ and $j^{(Y)} \in \{1, 2, 3, 4, 5, 6\}$ under Move versus under Stay. All the comparisons from these three scenarios are displayed by Figure
4.8, 4.9, and 4.10. Basically, there are no significant differences between Move and Stay from Metric I. The boxplots for Metric II show that the medians between these two methods are similar, although the spread of Move is smaller than that of Stay in Figure 4.8 and 4.9.

Scenario “$Y^{(F)}$ and $Y^{(NF)}$”

In “$Y^{(F)}$ and $Y^{(NF)}$”, we choose to use the pairwise correlation to measure the associations between $Y^{(F)}$ and $Y^{(NF)}$. All the comparisons from these three scenarios are displayed by Figure 4.11, 4.12, and 4.13. Again, there are no significant differences between Move and Stay from Metric I. And, there are no significant differences between these two methods from the boxplots for Metric II.

### 4.3.4 Discussion

These results can be explained by the model structure. In the scenario of “$X^{(F)}$ and $X^{(NF)}$”, we know that the true data has

$$Pr(X_j^{(F)}, X_j^{(NF)}) \neq Pr(X_j^{(F)})Pr(X_j^{(NF)}),$$

for $j = 1, \cdots, 4$. However, Stay has

$$Pr(X_j^{(F)}, X_j^{(NF)}) = \sum_{z=1}^{k(2)} \prod_{i:Z_i = z} Pr(x_{ij}^{(F)}|Z_i = z)\prod_{i:Z_i = z} Pr(x_{ij}^{(NF)}|Z_i = z)Pr(Z_i = z),$$

with

$$Pr(x_{ij}^{(F)}|Z_i = z) = \sum_{b=1}^{k(FX)} \phi_{z,b}^{(FX)} \psi_{b,x_{ij}^{(F)}}^{(F)(j)}$$

$$Pr(x_{ij}^{(NF)}|Z_i = z) = \sum_{h=1}^{k(NF)} \phi_{z,h}^{(NF)} \psi_{h,x_{ij}^{(NF)}}^{(NF)(j)}$$

(4.24)
Thus, \textit{Stay} enforces $X_j^{(F)}$ and $X_j^{(NF)}$ to be locally independent. Their strong association can only be estimated through the tensor factorization at the layer of $Z_i$. Because of this, \textit{Stay} is an inappropriate model for “$X^{(F)}$ and $X^{(NF)}$”. It tends to consume more mixture components than necessary to approximate the true data structure.

On the contrary, \textit{Move} has

$$Pr(X_j^{(F)}, X_j^{(NF)}) = \sum_{z=1}^{k(Z)} \{ \prod_{i:Z_i=z} Pr(x_{ij}^{(F)}, x_{ij}^{(NF)} | Z_i = z) \} Pr(Z_i = z), \quad (4.25)$$

with $Pr(x_{ij}^{(F)}, x_{ij}^{(NF)} | Z_i = z) = \sum_{b=1}^{k(FX)} \phi_{z,b}^{(FX)}(j) \phi_{b,x_{ij}^{(F)}}(j) \psi_{X_{ij}^{(NF)}}$. It captures this strong dependence locally by the DPMPM in the sub-model of $F$ cluster, which can be more efficient and accurate. Similarly, we have $Pr(X_j^{(F)}, Y_j^{(NF)}) \neq Pr(X_j^{(F)})Pr(Y_j^{(NF)})$ for scenario “$X^{(F)}$ and $Y^{(NF)}$”. However, $X_{ij}^{(F)} \perp Y_{ij}^{(NF)} | Z_i = z$ under the strategy of \textit{Stay}. But, $X_j^{(F)}$ and $Y_j^{(NF)}$ are conditionally dependent under the strategy of \textit{Move}. Thus, \textit{Move} is better than \textit{Stay} for “$X^{(F)}$ and $Y^{(NF)}$”. Therefore, the strategy of \textit{Move} can benefit focus categorical variable $X^{(F)}$ when $X^{(F)}$ and $NF$ variables are highly associated.

For focus continuous variable $Y^{(F)}$, the strategy of \textit{Move} does not bring any significant benefit when the associated $NF$ variable is linearly connected with it. The model already has the mean function of $Y^{(F)}$ regressed upon the main effect of $X^{(F)}$, $X^{(NF)}$, and $Y^{(NF)}$ in each assigned mixture component. Therefore, the model fitting under the strategy of \textit{Stay} and that under the strategy of \textit{Move} is practically the same. The results from “$Y^{(F)}$ and $X^{(NF)}$” and “$Y^{(F)}$ and $Y^{(NF)}$” reflect this. At the same time, if the true dependence between $Y^{(F)}$ and $NF$ is nonlinear, the model performance between \textit{Move} and \textit{Stay} also are practically similar. The reason is that their joint distribution is a mixture distribution either under the strategy of \textit{Stay}
or under the strategy of Move. For example, the joint density of \( Y^{(F)} \) and \( Y^{(NF)} \) is a mixture of bivariate normal distributions within each latent class \( z \) under Stay. Their joint density is also a mixture of bivariate normal distributions within each latent class \( a \) in \( H^{(FY)} \) under Move. Each of these two mixture distributions can represent a variety of shapes that a single distribution cannot easily model. Thus, Move and Stay do not show much difference in scenarios “\( Y^{(F)} \) and \( X^{(NF)} \)” and “\( Y^{(F)} \) and \( Y^{(NF)} \)”.

4.4 Empirical Study

Although \( HCMM-FNF \) is a complex NB mixture modelling with focus and non-focused clustering, it does have the constraints of local independence between \( X^{(F)} \) and the variables in \( NF \). On the other hand, the model structure implies that we can try to overcome this limitation by applying the strategy of Move for those \( NF \) variables that highly associate with \( X^{(F)} \). This suggestion is further supported by the simple scenarios in Section 4.3. In this section we would like to see if similar behavior holds for genuine data.

4.4.1 Illustrative Data: The American Community Survey

The American Community Survey (ACS), an ongoing survey conducted by the U.S. Census Bureau, collects demographic, housing, social, and economic data from sampled households along with information on the people who live in these households. It is a rich and dynamic resource for public policy decision making and analysis. Researchers can access public use files from the Integrated Public Use Microdata Series (IPUMS, \( usa.ipums.org \)).

Relationships among variables in ACS can be complex. For example, the shape of the distributions of total annual income can vary with different educational attainment. It is challenging to capture such relationships with a linear regression
model for gross annual income. Therefore, imputations and models of the ACS can benefit from a flexible modelling technique such as nonparametric Bayesian mixture modelling.

With this dataset, we are interested in knowing the performance of HCMRFNF. To be more specific, we would like to know (1) how the conditional modelling HCMRFNF compares with the joint modelling HCMMLD, and (2) whether or not the strategy of Move can benefit the estimation of the distribution of the focus categorical variables.

When conducting this empirical study, we consider the balance between data size and computational speed. We choose 17 variables from ACS shown in Table 4.2. These variables reflect a household’s economic variable, household head’s demographics, race, health insurance, education, income, occupational standing, migration, employment, and disability. Since IPUMS reprocesses the raw data, the rate of missing values is low for the variables (see Table 4.2). Thus, it is difficult to assign focus and non-focus variables only based on the rate of missingness.

We choose a sub-population that consists of household heads who own their living units and were employed during the year of 2010 in the state of North Carolina. After we remove the nonavailable values, this subset has 19,492 complete cases from which we systematically sample (with the sampling interval of 10) 1,026 household heads as our constructed sample.

4.4.2 Studies

We have two studies — one with high relative mutual information shared between \(X^{(F)}\) and \(\mathcal{N}\mathcal{F}\) and the other with low relative mutual information shared between them. We calculate the relative mutual information to quantify the association between any two variables. The mutual information between any two continuous
Table 4.2: The variables selected from ACS in the year of 2010. The first four variables are for each household. The other variables are for individual household head. Cts is short for continuous variable and Cat is short for categorical variables. For a categorical variable, # Levels is its number of factorial levels. PROPTX99 is a categorical variable with large number of levels. It is treated as a continuous variable only when we represent the related results.

<table>
<thead>
<tr>
<th>Name</th>
<th>Label</th>
<th>Cts./Cat.[#Levels]</th>
<th>NA%</th>
</tr>
</thead>
<tbody>
<tr>
<td>PROPX99</td>
<td>Annual property taxes</td>
<td>Categorical[67]</td>
<td>0</td>
</tr>
<tr>
<td>COSTELEC</td>
<td>Annual electricity cost</td>
<td>Continuous</td>
<td>0</td>
</tr>
<tr>
<td>COSTGAS</td>
<td>Annual gas cost</td>
<td>Continuous</td>
<td>0</td>
</tr>
<tr>
<td>COSTWATR</td>
<td>Annual water cost</td>
<td>Continuous</td>
<td>0</td>
</tr>
<tr>
<td>AGE</td>
<td>Age [of head]</td>
<td>Continuous</td>
<td>0</td>
</tr>
<tr>
<td>SEX</td>
<td>Sex [of head]</td>
<td>Categorical[2]</td>
<td>0</td>
</tr>
<tr>
<td>MARST</td>
<td>Marital status [of head]</td>
<td>Categorical[6]</td>
<td>0</td>
</tr>
<tr>
<td>RACE</td>
<td>Race [of head]</td>
<td>Categorical[7]</td>
<td>0</td>
</tr>
<tr>
<td>HCOVANY</td>
<td>Any health insurance coverage [of head]</td>
<td>Categorical[2]</td>
<td>0</td>
</tr>
<tr>
<td>EDUC</td>
<td>Educational attainment [of head]</td>
<td>Categorical[9]</td>
<td>0.59</td>
</tr>
<tr>
<td>SCHLTYPE</td>
<td>Public or private school [of head]</td>
<td>Categorical[3]</td>
<td>0</td>
</tr>
<tr>
<td>INCTOT</td>
<td>Total personal income [of head]</td>
<td>Continuous</td>
<td>0.12</td>
</tr>
<tr>
<td>OCCSCORE</td>
<td>Occupational income score [of head]</td>
<td>Continuous</td>
<td>0</td>
</tr>
<tr>
<td>PWTYPE</td>
<td>Place of work: metropolitan status [of head]</td>
<td>Categorical[5]</td>
<td>2.00</td>
</tr>
<tr>
<td>MIGRATE1</td>
<td>Migration status, 1 year [of head]</td>
<td>Categorical[4]</td>
<td>0</td>
</tr>
<tr>
<td>DIFFSENS</td>
<td>Vision or hearing difficulty [of head]</td>
<td>Categorical[2]</td>
<td>0</td>
</tr>
</tbody>
</table>

variables is given by

$$I(X, Y) = \int_Y \int_X p(x, y) \log \left( \frac{p(x, y)}{p(x)p(y)} \right) dx dy.$$  \hspace{1cm} (4.26)

The relative mutual information with respect to a variable $X$ is a ratio of $I(X, Y)$ over $I(X, X)$. For categorical variables, we just need to replace the integral with the summation in (4.26).

In each study, we impute the missingness in the constructed sample with three different models — HCMM-LD, HCMM-FNF with Stay, and HCMM-FNF with Move. We use HCMM-LD as a benchmark to quantify the performance of Stay and Move.

**High Mutual Information (HMI) Study**

We begin with the study in which $X^F$ and $N F$ share a high amount of relative mutual information. From all the available categorical variables in Table 4.2, we select EDUC and PROPTX99 as two focus ones (which we assume are missing completely
Table 4.3: The relative mutual information shared among focus and non-focus variables in the study of HMI. INCTOT and OCCSCORE are the top two NF variables which shares a relatively equal and high amount of mutual information with EDUC and PROPTX99.

<table>
<thead>
<tr>
<th></th>
<th>EDUC</th>
<th>PROPTX99</th>
</tr>
</thead>
<tbody>
<tr>
<td>INCTOT</td>
<td>0.12</td>
<td>0.26</td>
</tr>
<tr>
<td>OCCSCORE</td>
<td>0.14</td>
<td>0.22</td>
</tr>
<tr>
<td>COSTELEC</td>
<td>0.03</td>
<td>0.09</td>
</tr>
<tr>
<td>COSTGAS</td>
<td>0.05</td>
<td>0.09</td>
</tr>
<tr>
<td>COSTWATR</td>
<td>0.04</td>
<td>0.11</td>
</tr>
<tr>
<td>AGE</td>
<td>0.02</td>
<td>0.07</td>
</tr>
<tr>
<td>SEX</td>
<td>0.003</td>
<td>0.01</td>
</tr>
<tr>
<td>RACE</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>MARST</td>
<td>0.02</td>
<td>0.04</td>
</tr>
<tr>
<td>SCHLTYPE</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>HCOVNAY</td>
<td>0.02</td>
<td>0.01</td>
</tr>
<tr>
<td>PWTYPE</td>
<td>0.02</td>
<td>0.04</td>
</tr>
<tr>
<td>MIGRATE1</td>
<td>0.005</td>
<td>0.01</td>
</tr>
<tr>
<td>DIFFSENS</td>
<td>0.008</td>
<td>0.007</td>
</tr>
</tbody>
</table>

The values of relative mutual information shared between these two variables and the others are listed in Table 4.3. From this table, we expect that treating INCTOT and OCCSCORE as focus continuous variables under the strategy of Move will improve the estimates of the associations.

The assignment of $F$ and $NF$ is summarized in Figure 4.14. Since we don’t have the true missing rates for all these variables, we assume the missing mechanism is missing completely at random with the missing rate in focus as 50% and non-focus as 1%. This is consistent with the missing rates in our simulation studies.

We analyze the marginal probabilities of PROTPX99 and EDUC, and pay special attention to associations between to-be-moved NF (i.e., INCTOT and OCCSCORE) and $X(r)$ variables.

Figure 4.15 shows the marginal density of PROPTX99. Compared with HCMM-LD, Stay and Move are both better in the sense that HCMM-LD has difficulty in
modelling the values of PROPTX99 beyond 50. The decreasing order of performance is \textit{Move} (0.61) > \textit{Stay} (0.77) > \textit{HCMM-LD} (0.83), where the value in the parenthesis is the average of the absolute difference between the true and the imputed densities,

\begin{equation}
\text{Average of the absolute difference} = \frac{1}{L} \sum_{l=1}^{L} \left\{ \sum_{x} |\hat{f}^{(l)}(x) - f(x)| \right\},
\end{equation}

in which \( \hat{f}^{(l)}(\cdot) \) is the \( l^{th} \) imputed density and \( f(\cdot) \) is the true density.

The marginal cell counts of EDUC are displayed in Figure 4.16. In our constructed sample, EDUC has 9 categories from level 1 (nursery school to grade 4) to level 10 (4 years of college) without level 8 (2 years of college). This marginal distribution is hard to impute accurately because the number of household heads whose education attainment is no higher than level 4 (grade 10) is small. There are no major differences among the three models, although the patterns of the imputation performances from them can vary under each panel.

The contour plots from the kernel density of the standardized log(1+INCTOT) and PROTPX99 for the missing observations are displayed in Figure 4.17. The true density is unimodal, concentrated in the area with PROPTX99 from 5 to 45 and log(1+INCTOT) from -1.5 to 1.2. By comparison, the completed data estimates of the kernel densities under \textit{HCMM-LD} and under \textit{Stay} have a large spread with their contours heavily distorted. The imputed kernel density under \textit{Move} looks most similar to the truth.

Figure 4.18 displays the kernel densities of the standardized OCCSCORE and PROPTX99 for the missing observations. The true density has two big connected modes with one small isolated mode. The reason of having a small mode is because few household heads whose occupational score is around 1 (around 41 on the original scale) need to pay a high amount for their property taxes. Due to the small number of people who are in the isolated mode, it is challenging to capture this density so
that both HCMM-LD and Stay have inadequate fits. By comparison, Move is better than the other two models.

The bivariate associations related with EDUC are displayed in Figure 4.19, 4.20, and 4.21. In Figure 4.19, Move is slightly better in the tenth panel. In Figure 4.20, Move is slightly worse in the first panel and Stay is slightly worse in the tenth panel. In Figure 4.21, Move is slightly worse in the first panel while HCMM-LD is slightly worse in the fourth and tenth panels. From all these panel plots, therefore, it is hard to state that Move is better than the other methods. The reason could be that the highest relative mutual information between EDUC and the others is only about 0.14 (see Table 4.3), which is low. Thus, the imputation performances between Move and Stay are comparable.
Figure 4.6: The comparison of the bivariate cell frequencies between Stay and Move in “X(F) and X(NF)”. The left plot is from Metric I. A triangle symbol is a difference when Move versus Stay has a p-value below 0.01. The right one is from Metric II, where the median is 0 for Stay and 0 for Move. The bottom plot zooms in the right one by focusing on the percentage change between −5 and 5, where the median is still 0 for Stay and 0 for Move.
Figure 4.7: The comparison of the logistic regression coefficients between Stay and Move in \(X^{(F)}\) and \(Y^{(NF)}\). The left plot is from Metric I. A triangle symbol is a difference when Move versus Stay has a p-value below 0.01. The right plot is from Metric II, in which the median is \(-44.79\) for Stay and \(-9.93\) for Move.

Figure 4.8: The comparison of the logistic regression coefficients between Stay and Move in \(Y^{(F)}\) and \(X^{(NF)}\) with the associations described by (4.16). The left plot is from Metric I. The right one is from Metric II, where the median is \(-0.0884\) for Stay and \(-0.1048\) for Move.
Figure 4.9: The comparison of the logistic regression coefficients between Stay and Move in \( Y^{(F)} \) and \( X^{(NF)} \) with the associations described by (4.17). The left plot is from Metric I. The right one is from Metric II, where the median is \(-0.7822\) for Stay and \(-0.7393\) for Move.

Figure 4.10: The comparison of the logistic regression coefficients between Stay and Move in \( Y^{(F)} \) and \( X^{(NF)} \) with the associations described by (4.18). The left plot is from Metric I. The right one is from Metric II, where the median is \(-2.9172\) for Stay and \(-3.101\) for Move.
**Figure 4.11**: The comparison of the correlation coefficients between *Stay* and *Move* in “$Y^F$ and $Y^{(NF)}$” with the associations described by (4.19). The left plot is from Metric I. The right one is from Metric II, where the median is $-12.47$ for *Stay* and $-4.26$ for *Move*.

**Figure 4.12**: The comparison of the correlation coefficients between *Stay* and *Move* in “$Y^F$ and $Y^{(NF)}$” with the associations described by (4.20). The left plot is from Metric I. The right one is from Metric II, where the median is $-36$ for *Stay* and $-30$ for *Move*. 
Figure 4.13: The comparison of the correlation coefficients between \textit{Stay} and \textit{Move} in “$Y^{(F)}$ and $Y^{(NF)}$” with the associations described by (4.21). The left plot is from Metric I. The right one is from Metric II, where the median is $-25$ for \textit{Stay} and $-28$ for \textit{Move}.

Figure 4.14: The variable assignment of focus and non-focus variables for ACS 2010 in the state of North Carolina in the study of \textit{HMI}. $\text{INCTOT}$ and $\text{OCCSCORE}$ are the variables which will be moved to the side of $Y^{(F)}$ under the strategy of move.
Figure 4.15: The marginal probability distributions of PROPTX99 in the study of HMI. In each plot, the black solid line with gray area is from the truth with its 95% bootstrap confidence interval. The other lines are from the completed datasets after imputation.
Figure 4.16: The marginal boxplots of EDUC in the study of HMI. Each panel corresponds to a factor level in EDUC. And each boxplot is a summary of all 10 imputed counts in that factor level from a particular imputation model. Here, H, M, S, T stand for HCMM-LD, Move, Stay, and Truth, respectively.
Figure 4.17: The contour plots from the kernel density of log(1+INCTOT) (standardized) and PROTPX99 for the missing observations in the study of HMI. For each imputation model, the contour plots of all its imputed kernel densities are identically and independently drawn and look similar, so that we randomly select one imputation for the purpose of display.
Figure 4.18: The contour plots from the kernel density of OCCSCORE (standardized) and PROTPX99 for the missing observations in the study of HMI. For each imputation model, the contour plots of all its imputed kernel densities are identically and independently drawn and look similar, so that we randomly select one imputation for the purpose of display.
Figure 4.19: The means of the standardized log(1+INCTOT) by EDUC in the study of HMI. H, M, and S stand for HCMM-LD, HCMM-FNF with Stay, and HCMM-FNF with Move. For each imputation model, the error bars are obtained by applying the combination rule on all its imputations.

Figure 4.20: The means of the standardized OCCSCORE by EDUC in the study of HMI. T, H, M, and S stand for True, HCMM-LD, HCMM-FNF with Stay, and HCMM-FNF with Move. For each imputation model, the error bars are obtained by applying the combination rule on all its imputations.
Figure 4.21: The means of PROPTX99 by EDUC in the study of HMI. T, H, M, and S stand for True, HCMM-LD, HCMM-FNF with Stay, and HCMM-FNF with Move. For each imputation model, the error bars are obtained by applying the combination rule on all its imputations.

Low Mutual Information (LMI) Study

For ACS, learning people’s socioeconomic status is an important analysis goal. A person’s social standing can be measured as a combination of education, income, occupation, and health. Thus, we treat EDUC, INCTOT, OCCSCORE, and DIFFSENS as focus variables and the remaining variables as non-focus ones (see Figure 4.22). We assume focus variables have 50% items missing completely at random. The non-focus variables have a missing rate of only 1%.

Table 4.4 lists all the pairwise relative mutual information values between one focus and one non-focus variable in our constructed population. From this table, we see that PROPTX99 is the top variable that shares a high amount of relative mutual information with all focus ones. However, we do not expect moving PROPTX99 to the side of $X^{(F)}$ will improve the accuracy. From the simulation studies of Section 4.3, we know that the improvement by Move is most evident when to-bemoved $\mathcal{N.F}$
and $X^{(F)}$ are highly associated. However, their relative mutual information is only 0.16. Thus, moving may not make a difference in accuracy. Thus, we name the second study as Low Mutual Information study. Additionally, Move does not benefit focus continuous variables much according to the simulation studies. However, in the study of LMI we would still like to compare the imputation performance of Move and Stay when we assign some not quite related $\mathcal{N} \setminus \mathcal{F}$ variables to the side of $\mathcal{F}$.

The marginal probability densities of the standardized log(1+INCTOT) are displayed in Figure 4.23. The true density has two modes — one big mode around 0 and one small mode around 2. This distribution is heavily skewed to the left, which means some household heads have annual total income far below average. In Figure 4.23, Stay seems to be the best because all of its imputed marginal densities fall into the 95% bootstrap confidence interval of the truth. However, the advantage of Stay over the other two models is not significantly obvious.
Table 4.4: The relative mutual information shared among focus and non-focus variables in the study of LMI. PROPTX99 is the non-focus variable which shares a relatively high amount of mutual information with all the focus ones.

<table>
<thead>
<tr>
<th></th>
<th>EDUC</th>
<th>DIFFSENS</th>
<th>INCTOT</th>
<th>OCCSCORE</th>
</tr>
</thead>
<tbody>
<tr>
<td>PROPTX99</td>
<td>0.16</td>
<td>0.16</td>
<td>0.26</td>
<td>0.22</td>
</tr>
<tr>
<td>COSTELEC</td>
<td>0.03</td>
<td>0.05</td>
<td>0.06</td>
<td>0.05</td>
</tr>
<tr>
<td>COSTGAS</td>
<td>0.05</td>
<td>0.02</td>
<td>0.07</td>
<td>0.06</td>
</tr>
<tr>
<td>COSTWATR</td>
<td>0.04</td>
<td>0.04</td>
<td>0.07</td>
<td>0.07</td>
</tr>
<tr>
<td>AGE</td>
<td>0.02</td>
<td>0.13</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>SEX</td>
<td>0.003</td>
<td>0.0003</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>RACE</td>
<td>0.03</td>
<td>0.02</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>MARST</td>
<td>0.02</td>
<td>0.05</td>
<td>0.02</td>
<td>0.03</td>
</tr>
<tr>
<td>SCHLTYPE</td>
<td>0.01</td>
<td>0.01</td>
<td>0.008</td>
<td>0.01</td>
</tr>
<tr>
<td>HCOVANY</td>
<td>0.02</td>
<td>$1 \times 10^{-5}$</td>
<td>0.03</td>
<td>0.02</td>
</tr>
<tr>
<td>PWTYPE</td>
<td>0.02</td>
<td>0.03</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>MIGRATE1</td>
<td>0.005</td>
<td>0.01</td>
<td>0.01</td>
<td>0.009</td>
</tr>
</tbody>
</table>

The marginal probability densities of the standardized OCCSCORE are displayed in Figure 4.24. Its truth has two big and one small mode. The big modes are around $-1$ and 1. The small mode is around 2.8. Again, this distribution cannot be easily estimated by a standard parametric method, suggesting that NB models can help us recover it. Figure 4.24 shows that (1) HCM-M-LD is slightly better than the other two for OCCSCORE $<-1$; (2) Stay is slightly better than the other two for OCCSCORE $>1.5$; and (3) Move is slightly better than the other two for OCCSCORE between $-0.5$ and $0.5$. Thus, it is difficult to conclude which model is the best for the marginal probability density of OCCSCORE.

Figure 4.25 displays the marginal cell counts for EDUC. We can hardly differentiate the results for all the three models based on all the panel plots in Figure 4.25.

Figure 4.26 displays the marginal cell counts for DIFFSENS, which has only two categories: no vision or hearing difficulty (1), and has vision or hearing difficulty (2). Like EDUC, the marginal distribution of DIFFSENS is difficult to impute due to the small number of household heads in the second category. From Figure 4.26,
we can see that HCMM-LD tends to overestimate the number of healthy people and underestimate that of people with disability. Move is the only model that is consistently close to the truth. The performance of Stay is somewhat between HCMM-LD and Move.

Next, we take a look at the bivariate associations related with focus variables. The kernel densities of the standardized log(1+INCTOT) and OCCSCORE for the missing observations are displayed in Figure 4.27. The true kernel density seems to have two modes. However, none of the completed-data density estimates looks similar to the truth. The kernel densities of the standardized log(1+INCTOT) and
Figure 4.24: The marginal probability distributions of OCCSCORE (standardized) in the study of \textit{LMI}. In each plot, the black solid line with gray area is from the truth with its 95\% bootstrap confidence interval. The other lines are from the completed datasets after imputation.

PROPTX99 for the missing observations are displayed in Figure 4.28. This time all the three imputed densities look similar to the truth with \textit{Stay} being slightly better. The kernel densities of the standardized OCCSCORE and PROPTX99 for the missing observations are displayed in Figure 4.29. The truth has three modes — two are connected and one is isolated. All three models can capture the isolated mode with PROPTX99 around 68. However, none of them accurately imputes the two connected modes with PROPTX99 around 25. From all these bivariate kernel plots, we can see that \textit{Stay} and \textit{Move} are very comparable. Additionally, a joint modelling method \textit{HCMM-LD} shows no advantages over \textit{HCMM-FNF}. 

126
Finally, we examine the bivariate associations related with focus categorical variables. Figure 4.30 displays the means of the standardized log(1+INCTOT) by EDUC. It seems that HCMM-LD is slightly worse in the fourth panel. Figure 4.31 displays the means of the standardized OCCSCORE by EDUC. All three models are quite competitive. Figure 4.32 displays the means of PROPTX99 by EDUC. This time HCMM-LD is slightly worse in the second and fifth panels and Stay is slightly worse in the third, ninth, and tenth panels.

Figure 4.33 displays the means of the standardized log(1+INCTOT) by DIFFSENS. It seems that HCMM-FNF are slightly worse in the first panel. Figure 4.34 displays the means of the standardized OCCSCORE by DIFFSENS. HCMM-LD is slightly worse in the first panel and Stay is slightly worse in the second one. Figure 4.35 displays the means of PROPTX99 by DIFFSENS. This time HCMM-LD is slightly worse in the second panel.

Again, these panel plots do not clearly indicate which model is the best to estimate the associations related with $X^{(F)}$.

4.4.3 Discussion

In general, the results of this empirical study are consistent with the conclusions from the simulation studies of Section 4.3. The strategy of Move can improve the associations between the focus categorical variables and the highly associated non-focus variables. This explains why the quantities of interest associated with INCTOT and OCCSCORE in the study of LMI are not significantly improved under the strategy of Move. By contrast, Move significantly improves the estimates between INCTOT and PROPTX99, or OCCSCORE and PROPTX99 in the study of HMI. The degree of improvement depends on the amount of relative mutual information shared between $X^{(F)}$ and $\mathcal{NF}$. Thus, the advantage of Move is not apparent for the estimates related with EDUC in both HMI and LMI as well as those related with
DIFFSENS in the study of LMI.

We have also tried HMI and LMI on another random sample of 1,026 household heads who own their living units and were employed during the year of 2010 in the state of North Carolina. Basically, we have found that the patterns (see Appendix B.2) from these cases look similar to those presented here.

4.5 Conclusion

HCMM-FNF is motivated by the limitation of NB mixture modelling with local independence in the situation when some variables are almost fully observed and the others are missing a lot. Basically, the model inherits the property of local dependence from HCMM-LD to limit the proliferation of mixture components for the highly associated variables. Additionally, it has the focused and non-focused clustering for $\mathcal{F}$ and $\mathcal{NF}$ survey variable.

In HCMM-FNF, there is a direct link between $Y^{(F)}$ and $\mathcal{NF}$. However, there is no such link between $X^{(F)}$ and $\mathcal{NF}$. This implies that a weakness of HCMM-FNF can exist when $X^{(F)}$ and $\mathcal{NF}$ are highly associated. To avoid such drawbacks, we can move variables from $\mathcal{NF}$ to $\mathcal{F}$. The simulation studies in Section 4.3 show that the associations between $X^{(F)}$ with either $X^{(NF)}$ or $Y^{(NF)}$ can be improved when they highly depend on each other.

To further investigate this finding, we apply Move to a sample from American Community Survey in Section 4.4. The High-Mutual-Information study shows that moving INCTOT and OCCSCORE to the side of $Y^{(F)}$ can increase the imputation performance for estimating the associations between these two variables and PROPTX99. We also compare the imputation quality of Move versus that of Stay when there are no strong associations between $\mathcal{NF}$ and $X^{(F)}$ variables in the Low-Mutual-Information study. This study shows no much difference between these two models.
The empirical studies also show that \textit{HCMM-FNF} and the benchmark model \textit{HCMM-LD} are competitive. This suggests that, for some datasets, \textit{HCMM-LD} could be substituted by \textit{HCMM-FNF} without much loss when the survey variables can be partitioned into $\mathcal{F}$ and $\mathcal{N}\mathcal{F}$. Compared with \textit{HCMM-LD} in this case, the benefit of \textit{HCMM-FNF} is its computational efficiency. The vector $Y \in \mathcal{R}^q$ in \textit{HCMM-LD} can be partitioned into $Y^{(F)} \in \mathcal{R}^{q(F)}$ and $Y^{(NF)} \in \mathcal{R}^{q(NF)}$ with $q^{(F)} + q^{(NF)} = q$. Then, $Y^{(F)}$ and $Y^{(NF)}$ can be drawn from their own individual multivariate normal densities in \textit{HCMM-FNF}. This is more efficient than drawing the entire $Y$ from a multivariate Gaussian in \textit{HCMM-LD}. \textit{HCMM-FNF} further assumes that $Y^{(NF)}$ are locally independent, which can simplify the computation.

Since \textit{HCMM-FNF} is designed for contexts with almost fully observed non-focus variables, the missing rates in $\mathcal{N}\mathcal{F}$ variables need to be low. \textit{HCMM-FNF} also is most appropriate when there are not too many focus variables. Thus, we believe that the performance of \textit{HCMM-FNF} is related with these two factors. Continuing work can involve quantifying the imputation performance of \textit{HCMM-FNF} according to combinations of these conditions.

In addition, some recent advances in NB imputation (Manrique-Vallier and Reiter, 2013, 2014; Kim et al., 2014, 2015; DeYoreo et al., 2015b) shed light on interesting future directions for \textit{HCMM-FNF} in the areas of structural zeros, linear constrains, and various data types. For example, we can modify this model to incorporate ordinal data through the method proposed by DeYoreo et al. (2015b).
Figure 4.25: The marginal boxplots of EDUC in the study of LMI. Each panel corresponds to a factor level in EDUC. And each boxplot is a summary of all 10 imputed counts in that factor level from a particular imputation model. Here, H, M, S, T stand for HCMM-LD, Move, Stay, and Truth, respectively.
Figure 4.26: The marginal boxplots of DIFFSENS in the study of LMI. Each panel corresponds to a factor level in DIFFSENS. And each boxplot is a summary of all 10 imputed counts in that factor level from a particular imputation model. Here, H, M, S, T stand for HCMM-LD, Move, Stay, and Truth, respectively.
Figure 4.27: The contour plots of the kernel density for log(1+INCTOT) and OCCSCORE (standardized) for the missing observations in the study of LMI. For each imputation model, the contour plots of all its imputed kernel densities are identically and independently drawn and look similar, so that we randomly select one imputation for the purpose of display.
Figure 4.28: The contour plots of the kernel density for PROPTX99 and the standardized log(1+INCTOT) for the missing observations in the study of LMI. For each imputation model, the contour plots of all its imputed kernel densities are identically and independently drawn and look similar, so that we randomly select one imputation for the purpose of display.
Figure 4.29: The contour plots of the kernel density for PROPTX99 and OCC-SCORE (standardized) for the missing observations in the study of LMI. For each imputation model, the contour plots of all its imputed kernel densities are identically and independently drawn and look similar, so that we randomly select one imputation for the purpose of display.
Figure 4.30: The means of the standardized log(1+INCTOT) by EDUC in the study of LMI. Here, H, M, and S stand for HCMM-LD, Stay, and Move. For each imputation model, the error bars are obtained by applying the combination rule on all its imputations.

Figure 4.31: The means of the standardized OCCSCORE by EDUC in the study of LMI. T, H, M, and S stand for True, HCMM-LD, Stay, and Move. For each imputation model, the error bars are obtained by applying the combination rule on all its imputations.
Figure 4.32: The means of PROPTX99 by EDUC in the study of LMI. T, H, M, and S stand for True, HCMM-LD, Stay, and Move. For each imputation model, the error bars are obtained by applying the combination rule on all its imputations.

Figure 4.33: The means of the standardized log(1+INCTOT by DIFFSENS in the study of LMI. T, H, M, and S stand for True, HCMM-LD, Stay, and Move. For each imputation model, the error bars are obtained by applying the combination rule on all its imputations.
Figure 4.34: The means of the standardized OCCSCORE by DIFFSENS in the study of LMI. T, H, M, and S stand for True, HCMM-LD, Stay, and Move. For each imputation model, the error bars are obtained by applying the combination rule on all its imputations.

Figure 4.35: The means of PROPTX99 by DIFFSENS in the study of LMI. T, H, M, and S stand for True, HCMM-LD, Stay, and Move. For each imputation model, the error bars are obtained by applying the combination rule on all its imputations.
Appendix A

Releasing Synthetic Magnitude Microdata
Constrained to Fixed Marginal Totals

A.1 Full conditionals for mixture of Poissons model

We estimate the mixture of Poisson distributions using a Metropolis within Gibbs sampler. In the full conditional that follow, we let “IW” indicate the inverse Wishart distribution and “Ga” indicate the Gamma distribution. In the conditioning, we
write a – to indicate all other parameters are fixed at current draws.

\[ z_i \mid \sim \text{Mult}(1, w_i = (w_{i1}, \ldots, w_{iK})) \text{ with } w_{ik} = \frac{\pi_k \prod_{j=1}^p \text{Pois}(y_{ij} | \theta_{kj})}{\sum_{i=1}^K \pi_i \prod_{j=1}^p \text{Pois}(y_{ij} | \theta_{ij})}, \quad \text{(A.1)} \]

\[ \pi \mid \sim \text{Dir}(d_1^*, \ldots, d_K^*) \text{ with } d_k^* = n_k + d_k, \text{ where } n_k = \sum_{i=1}^n (z_i = k). \quad \text{(A.2)} \]

\[ \Sigma \mid \sim \text{IW}(K, \tilde{\Sigma}^{-1}), \]

\[ \tilde{\Sigma} = \sum_{k=1}^K (\log(\theta_k) - \bar{\log}(\theta))(\log(\theta_k) - \bar{\log}(\theta))' \text{ and } \bar{\log}(\theta) = \sum_{k=1}^K \log(\theta_k)/K. \quad \text{(A.3)} \]

\[ \mu \mid \sim N_p(\bar{\log}(\theta), \Sigma/K). \quad \text{(A.4)} \]

\[ \Pr(\theta_k \mid \) \propto \begin{cases} \prod_{j=1}^p \text{Ga}(\sum_{i=1}^n y_{ij}(z_i = k) + 1, n_k)N(\log(\theta_k) | \mu, \Sigma) \prod_{j=1}^p \frac{1}{\theta_{kj}}, & n_k > 0 \\ N(\log(\theta_k) | \mu, \Sigma) \prod_{j=1}^p \frac{1}{\theta_{kj}}, & n_k = 0. \end{cases} \quad \text{(A.5)} \]

To sample each \( \theta_k \), we use a Metropolis step with the acceptance rate,

\[ \alpha = \min\left\{1, \frac{Pr(\theta_k^* \mid \theta_k) \cdot Q(\theta_k^* | \theta_k)}{Pr(\theta_k \mid \theta_k^*) \cdot Q(\theta_k \mid \theta_k^*)} \right\} \]

where \( \theta_k^* \) is the candidate for \( \theta_k \) and \( Q(\theta_k^* | \theta_k) \) is the proposal density,

\[ Q(\theta_k^* | \theta_k) = \begin{cases} \frac{\theta_k^* \text{ ind}}{\theta_{kj}} \sim \text{Ga}(\sum_{i=1}^n y_{ij}(z_i = k) + 1, n_k), & \text{for } j = 1, \ldots, p, \text{ when } n_k > 0 \\ \log(\theta_k^*) \sim N_p(\mu, \Sigma), & \text{when } n_k = 0. \end{cases} \quad \text{(A.6)} \]
Appendix B

Hierarchically Coupled Mixture Model with Focused and Non Focused Clustering for Continuous and Categorical Data

B.1 Model for HCMM-FNF

B.1.1 Priors

The priors for the mixture of product multinomials are
\[
\psi_b^{(F)}(j) \sim \text{Dir}(\gamma_{b,1}, \ldots, \gamma_{b,d_F}), \quad \gamma_b^{(j)} = (1/d_{j}^{(F)}, \ldots, 1/d_{j}^{(F)}) \tag{B.1}
\]
\[
\psi_h^{(NF)}(j) \sim \text{Dir}(\gamma_{h,1}, \ldots, \gamma_{h,d_{NF}}), \quad \gamma_h^{(j)} = (1/d_{j}^{(NF)}, \ldots, 1/d_{j}^{(NF)}). \tag{B.2}
\]

The priors for the mixture of multivariate normals are
\[
\Pr(B_a^{(F)}, \Sigma_a^{(F)}) = \mathcal{N}(B_0^{(F)}, I, T_B^{(F)}) \times \mathcal{W}(\nu^{(F)}, \Sigma^{(F)}) \tag{B.3}
\]
\[
\Pr(B_h^{(NF)}) = \mathcal{N}(B_0^{(NF)}, T_B^{(NF)}) \tag{B.4}
\]
\[
\Pr(\eta_{h,j}^{(NF)}) = \mathcal{G}(\nu^{(NF)}, \eta_j^{(NF)}), \tag{B.5}
\]

where \(T_B^{(F)} = \text{Diag}(\tau_1^{(F)}, \ldots, \tau_{d_F}^{(F)})\) and \(T_B^{(NF)} = \text{Diag}(\tau_1^{(NF)}, \ldots, \tau_{d_{NF}}^{(NF)})\) with
\[
\tau_j^{(F)} \sim \mathcal{G}(\alpha_{\tau^{(F)}}, \beta_{\tau^{(F)}}) \tag{B.6}
\]
\[
\tau_j^{(NF)} \sim \mathcal{G}(\alpha_{\tau^{(NF)}}, \beta_{\tau^{(NF)}}). \tag{B.7}
\]
The hyper priors are

\[
Pr(B_0^{(F)}, \Sigma^{(F)}) = \mathcal{N}(0, I, \sigma_0^{(F)^2} I) \times \mathcal{W}(w^{(F)}, \Sigma^{(F)}) \tag{B.8}
\]

\[
Pr(B_0^{(NF)}) = \mathcal{N}(0, \sigma_0^{(NF)^2} I) \tag{B.9}
\]

\[
Pr(\eta^{(NF)}) = \mathcal{IG}(\nu^{(NF)}, \eta_0^{(NF)}). \tag{B.10}
\]

Here, we take \(v^{(F)} = q^{(F)} + 2\), \(v^{(NF)} = 2\), \(w^{(F)} = q^{(F)} + 1\), \(w^{(NF)} = 1\), \(\Sigma_0^{(F)} = I/(q^{(F)} + 1)\), \(\eta_0^{(NF)} = 1\).

The hierarchical priors for the latent variables follow a truncated version of the stick-breaking construction of Dirichlet Process (Sethuraman, 1994), introduced in Ishwaran and James (2001)

\[
\phi_{z,a}^{(FY)} = V_{z,a}^{(FY)} \prod_{l < a} (1 - V_{z,l}^{(FY)}), V_{z,a}^{(FY)} \overset{iid}{\sim} \mathcal{B}(1, \beta^{(FY)}), V_{z,k}^{(FY)} = 1 \tag{B.11}
\]

\[
\phi_{z,b}^{(FX)} = V_{z,b}^{(FX)} \prod_{l < b} (1 - V_{z,l}^{(FX)}), V_{z,b}^{(FX)} \overset{iid}{\sim} \mathcal{B}(1, \beta^{(FX)}), V_{z,k}^{(FX)} = 1 \tag{B.12}
\]

\[
\phi_{z,h}^{(NFX)} = V_{z,h}^{(NFX)} \prod_{l < h} (1 - V_{z,l}^{(NFX)}), V_{z,h}^{(NFX)} \overset{iid}{\sim} \mathcal{B}(1, \beta^{(NFX)}), V_{z,k}^{(NFX)} = 1 \tag{B.13}
\]

\[
\lambda_z = W_z \prod_{l < z} (1 - W_l), W_z \overset{iid}{\sim} \mathcal{B}(1, \alpha), W_k(z) = 1. \tag{B.14}
\]
B.1.2 Full Conditionals

The full conditionals for the parameters in the stick-breaking construction are

\begin{align*}
Pr(\alpha | \cdot) &= G(a_\alpha + k^{(z)} - 1, b_\alpha - \log(\lambda_{k^{(z)}})) & \quad (B.15) \\
Pr(\beta^{(NF)} | \cdot) &= G(a_{\beta^{(NF)}} + n_{occ}(k^{(NF)} - 1), b_{\beta^{(NF)}} - \sum_{z:Z_{occ}} \log(\phi_{z,k}^{(NF)})) \quad (B.16) \\
Pr(\beta^{(FX)} | \cdot) &= G(a_{\beta^{(FX)}} + n_{occ}(k^{(FX)} - 1), b_{\beta^{(FX)}} - \sum_{z:Z_{occ}} \log(\phi_{z,k}^{(FX)})) \quad (B.17) \\
Pr(\beta^{(FY)} | \cdot) &= G(a_{\beta^{(FY)}} + n_{occ}(k^{(FY)} - 1), b_{\beta^{(FY)}} - \sum_{z:Z_{occ}} \log(\phi_{z,k}^{(FY)})) \quad (B.18)
\end{align*}

\begin{align}
Pr(W_z | \cdot) &= B(1 + n_z^{(z)}, \alpha + \sum_{k=1}^{k(z)} n_k^{(z)}, n_k^{(z)} = \sum_{i=1}^{n} I(Z_i = k), \text{ for } z = 1, \ldots, k(z) - 1 \quad (B.19)
\end{align}

\begin{align}
Pr(V_{i,a}^{(FY)} | \cdot) &= B(1 + n_{i,a}^{(FY)}, \beta^{(FY)} + \sum_{k=a+1}^{k^{(FY)}} n_{z,k}^{(FY)}, n_{z,k}^{(FY)} = \sum_{i=1}^{n} I(Z_i = z, H_i^{(FY)} = k), \quad \text{for } a = 1, \ldots, k^{(FY)} - 1 \quad (B.20)
\end{align}

\begin{align}
Pr(V_{i,b}^{(FX)} | \cdot) &= B(1 + n_{i,b}^{(FX)}, \beta^{(FX)} + \sum_{k=b+1}^{k^{(FX)}} n_{z,k}^{(FX)}, n_{z,k}^{(FX)} = \sum_{i=1}^{n} I(Z_i = g, H_i^{(FX)} = k), \quad \text{for } b = 1, \ldots, k^{(FX)} - 1 \quad (B.21)
\end{align}

\begin{align}
Pr(V_{i,h}^{(FX)} | \cdot) &= B(1 + n_{i,h}^{(NF)}, \beta^{(NF)} + \sum_{k=h+1}^{k^{(NF)}} n_{z,k}^{(NF)}, n_{z,k}^{(NF)} = \sum_{i=1}^{n} I(Z_i = g, H_i^{(NF)} = k), \quad \text{for } h = 1, \ldots, k^{(NF)} - 1. \quad (B.22)
\end{align}

Here, \( n_{occ} \) is the number of occupied clusters in latent class \( Z \), and \( Z_{occ} \) is the set of occupied clusters in latent class \( Z \).

The full conditionals for the parameters in the mixture of product multinomials are

\begin{align}
Pr(\psi_{b}^{(F)(j)} | \cdot) &= D(\gamma_{b,1}^{(j)} + \sum_{i:H_i^{(FX)} = b} I(x_{i,j}^{(F)} = 1), \ldots, \gamma_{b,d_j}^{(j)} + \sum_{i:H_i^{(FX)} = b} I(x_{i,j}^{(F)} = d_j^{(F)})) \quad (B.23) \\
Pr(\psi_{h}^{(NF)(j)} | \cdot) &= D(\gamma_{h,1}^{(j)} + \sum_{i:H_i^{(NF)} = h} I(x_{i,j}^{(NF)} = 1), \ldots, \gamma_{h,d_j^{(NF)}}^{(j)} + \sum_{i:H_i^{(NF)} = h} I(x_{i,j}^{(NF)} = d_j^{(NF)})). \quad (B.24)
\end{align}

The full conditionals for the parameters and hyper-parameters in the mixture of
multivariate normals are

$$Pr(\mathbf{B}_0^{(F)} \mid -) = \mathcal{N}(B_m, I, V), V = (I/\sigma_0^{(F)})^2 + k^{(FY)}T_B^{(F)^{-1}},$$

$$B_m = B_sT_B^{(F)^{-1}}, B_s = \sum_{a=1}^{k^{(FY)}}B_a^{(F)} \quad \text{(B.25)}$$

$$Pr(\Sigma^{(F)} \mid -) = \mathcal{W}(\mathbf{w}^{(F)} + \nu^{(F)}k^{(FY)}, (\Sigma_a^{(F)} - \Sigma_s^{(F)} - \Sigma_s^{(F)} - 1)\Sigma_s^{(F)} - 1 = \sum_{a=1}^{k^{(FY)}}\Sigma_a^{(F)} \quad \text{(B.26)}$$

$$Pr(\tau_j^{(F)} \mid -) = \mathcal{G}(a_{\tau(F)} + k^{(FY)}p^* \ell, b_{\tau(F)} + \sum_{a=1}^{k^{(FY)}}(B_a^{(F)} - B_0^{(F)})/(B_a^{(F)} - B_0^{(F)})/2), \quad \text{for } j = 1, \cdots , q^{(F)}$$

$$Pr(B_{a,j}^{(F)}) = \mathcal{N}(\tau_j^{(F)}B_{0,j}^{(F)} + D_a^{(F)}\tilde{Y}_{a,j}^{(F)}/\tilde{\sigma}_{a,j}^{(F)2}, V), V = (\tau_j^{(F)}I + D_a^{(F)}D_a^{(F)} / \tilde{\sigma}_{a,j}^{(F)2})^{-1}. \quad \text{(B.28)}$$

Here, $B_{a,j}^{(F)}$ is the $j^{th}$ column in $B_a^{(F)}$ and $B_0^{(F)}$ is the $j^{th}$ column in $B_0^{(F)}$. Matrix $D_a$ are the rows in design matrix $D$ whose latent class label is $H_i^{(FY)} = a$. Data $\tilde{Y}_{a,j}^{(F)} = Y_{a,j}^{(F)} - (Y_{a,j}^{(F)} - D_a^{(F)}B_{a,j}^{(F)})\Sigma_{a,j,j}^{(F)}$. Here, $Y_{a}^{(F)}$ is the collection of $Y_{i}^{(F)}$'s whose latent class label $H_i^{(FY)} = a$, $Y_{a,j}^{(F)}$ is the $j^{th}$ column of $Y_{a}^{(F)}$, and $Y_{a,j}^{(F)}$ is $Y_{a}^{(F)}$ without column $j$. The regression coefficient matrix $B_{a,j}^{(F)}$ is $B_a^{(F)}$ without column $j$.

The covariance matrix $\Sigma_{a,j,j}^{(F)}$ is $\Sigma_a^{(F)}$ without row $j$ and column $j$. Finally, $\Sigma_{a,j}^{(F)}$ is $\Sigma_a^{(F)}$ with row $j$ but without column $j$, and $\tilde{\sigma}_{a,j}^{(F)^2} = \Sigma_{a,j,j}^{(F)} - \Sigma_{a,j,j}^{(F)} - \Sigma_{a,j,j}^{(F)} - \Sigma_{a,j,j}^{(F)}$. We have
The full conditionals for the mixture components are

\[ \Pr(\Sigma_a^{(F)} | -) = \mathcal{W}(\nu^{(F)} + \eta_a^{(FY)}, \Sigma^{(F)} + Q_a), Q_a = (Y_a - D_a B_a^{(F)})' (Y_a - D_a B_a^{(F)}) \]  
\[ \Pr(B_0^{(NF)} | -) = \mathcal{N}(B_m, V), V = (I/\sigma_0^{(NF)} + k^{(NF)} T_B^{(NF)-1})^{-1}, \]
\[ B_m = B_s T_B^{(NF)-1}, B_s = \sum_{h=1}^{k^{(NF)}} B_h^{(NF)} \]  
\[ \Pr(\eta_j^{(NF)} | -) = \mathcal{G}(w^{(NF)} + \nu^{(NF)} \kappa^{(NF)}, \eta_0^{(NF)} + \sum_{h=1}^{k^{(NF)}} \sum_{j=1}^{q^{(NF)}} \eta_{h,j}^{(NF)-1}) \]
\[ \Pr(\tau_j^{(NF)} | -) = \mathcal{G}(a_{\tau^{(NF)}} + k^{(Z)}/2, b_{\tau^{(NF)}} + \sum_{z=1}^{k^{(Z)}} (B_z^{(NF)} - B_{0,z}^{(NF)})^2/2), \text{for } j = 1, \cdots, q^{(NF)} \]
\[ \Pr(B_h^{(NF)} | -) = \mathcal{N}((B_0^{(NF)} T_B^{(NF)-1} + 1') Y_h \Sigma_h^{(NF)-1}) V, V = (T_B^{(NF)-1} + n_h^{(NF)} \Sigma_h^{(NF)-1})^{-1} \]
\[ \Pr(\eta_{h,j}^{(NF)} | -) = \mathcal{IG}(\nu^{(NF)} + n_h^{(NF)} / 2, \eta_j^{(NF)} + \sum_{i: H_i^{(NF)} = h} (y_{ij} - B_{h,j})^2/2). \]

The full conditionals for the mixture components are

\[ \Pr(Z_i = z | -) = \mathcal{X} \lambda_z \phi_z^{(FY)} \phi_z^{(FY)} \phi_z^{(FY)} \]
\[ \Pr(H_i^{(FY)} = a | -) = \mathcal{X} \phi_{z,a}^{(FY)} N(y_i^{(FY)} | D_i B_a^{(F)}, \Sigma_a^{(F)}) \]
\[ \Pr(H_i^{(FX)} = b | -) = \mathcal{X} \phi_{z,b}^{(FX)} \prod_{j=1}^{p^{(FX)}} \phi_{b_d^{(F)}(j)}. \]

In the code, we set hyper parameters \( a_\alpha = a_\beta^{(F)} = a_\beta^{(NF)} = a_\beta^{(FY)} = a_{\tau^{(F)}} = a_{\tau^{(NF)}} = 0.5 \) and \( b_\alpha = b_{\beta^{(F)}} = b_{\beta^{(NF)}} = b_{\beta^{(FY)}} = b_{\beta^{(FY)}} = b_{\tau^{(F)}} = b_{\tau^{(NF)}} = 0.5 \).

**B.1.3 Imputation**

The subscripts \( mis \) means the missing values, and \( obs \) means the observed values. We use the following distributions to impute the missing values in each observation.
\[ P_{i}(X^{(F)}_{i, \text{mis}} = c | H^{(FX)}_{i} = b, H^{(FY)}_{i} = a, -) \mathcal{X}_{b,c}^{(F)}(j) N(Y^{(F)}_{i} | D_{i}B_{a}^{(F)}, \Sigma_{a}^{(F)}) \]  
(B.38)

\[ P_{i}(X^{(NF)}_{i, \text{mis}} = c | Z_{i} = z, H^{(FY)}_{i} = a, -) \mathcal{X}_{z,c}^{(NF)}(j) N(Y^{(F)}_{i} | D_{i}B_{a}^{(F)}, \Sigma_{a}^{(F)}) \]  
(B.39)

\[ P_{i}(Y^{(NF)}_{i, \text{mis}} | -) = P_{i}(Y^{(NF)}_{i, \text{mis}} | Y^{(NF)}_{i, \text{obs}}) = \eta(N^{(NF)}, V^{(NF)}) \]  
(B.40)

\[ P_{i}(Y^{(F)}_{i, \text{mis}} | -) = P_{i}(Y^{(F)}_{i, \text{mis}} | Y^{(F)}_{i, \text{obs}}, -) = \eta(M^{(F)}, V^{(F)}) \]  
(B.41)

in which

\[ M^{(NF)} = B_{h, \text{mis}}^{(NF)} - (Y^{(NF)}_{i, \text{obs}} - B_{h, \text{mis}}^{(NF)}) \Sigma_{h, \text{obs,obs}}^{(NF)} \Sigma_{h, \text{obs,obs}}^{(NF)} \]  
(B.42)

\[ V^{(NF)} = \Sigma_{h, \text{mis,mis}}^{(NF)} - \Sigma_{h, \text{mis,obs}}^{(NF)} \Sigma_{h, \text{mis,obs}}^{(NF)} \Sigma_{h, \text{mis,obs}}^{(NF)} \]  
(B.43)

\[ M^{(F)} = (D_{i}B_{a}^{(F)})_{\text{mis}} - (Y^{(F)}_{i, \text{obs}} - (D_{i}B_{a}^{(F)})_{\text{obs}}) \Sigma_{a, \text{obs,obs}}^{(F)} \Sigma_{a, \text{obs,obs}}^{(F)} \]  
(B.44)

\[ V^{(F)} = \Sigma_{a, \text{mis,mis}}^{(F)} - \Sigma_{a, \text{mis,obs}}^{(F)} \Sigma_{a, \text{mis,obs}}^{(F)} \Sigma_{a, \text{mis,obs}}^{(F)} \]  
(B.45)

### B.2 Results of HMI and LMI

Here, we present the results of HMI and LMI with another random sample of 1,026 observation who own their living units and were employed during the year of 2010 in the state of North Carolina.

#### B.2.1 High Mutual Information (HMI) Study
Figure B.1: The marginal probability distributions of PROPTX99 in the study of HMI. In each plot, the black solid line with gray area is from the truth with its 95% bootstrap confidence interval. The other lines are from the completed datasets after imputation. There is no significant difference between this and Figure 4.15.
Figure B.2: The marginal boxplots of EDUC in the study of HMI. Each panel corresponds to a factor level in EDUC. And each boxplot is a summary of all 10 imputed counts in that factor level from a particular imputation model. Here, H, M, S, T stand for HCMM-LD, Move, Stay, and Truth, respectively. The difference between this and Figure 4.16 is small, which does not influence our main conclusion in Section 4.4.
Figure B.3: The contour plots from the kernel density of log(1+INCTOT) (standardized) and PROTPX99 for the missing observations in the study of HMI. For each imputation model, the contour plots of all its imputed kernel densities are identically and independently drawn and look similar, so that we randomly select one imputation for the purpose of display. The only difference between this and Figure 4.17 seems to be from the model HCMM-LD. However, this does not affect our main conclusion in Section 4.4.
Figure B.4: The contour plots from the kernel density of OCCSCORE (standardized) and PROTPX99 for the missing observations in the study HMI. For each imputation model, the contour plots of all its imputed kernel densities are identically and independently drawn and look similar, so that we randomly select one imputation for the purpose of display. There is no significant difference between this and Figure 4.18.
Figure B.5: The means of the standardized log(1+INCTOT) by EDUC in the study of HMI. T, H, M, and S stand for True, HCMM-LD, HCMM-FNF with Stay, and HCMM-FNF with Move. For each imputation model, the error bars are obtained by applying the combination rule on all its imputations. There is no significant difference between this and Figure 4.19.
Figure B.6: The means of the standardized OCCSCORE by EDUC in the study of HMI. T, H, M, and S stand for True, HCMM-LD, HCMM-FNF with Stay, and HCMM-FNF with Move. For each imputation model, the error bars are obtained by applying the combination rule on all its imputations. The difference between this and Figure 4.20 is small, which does not influence our main conclusion in Section 4.4.
Figure B.7: The means of PROPTX99 by EDUC in the study of HMI. T, H, M, and S stand for True, HCMM-LD, HCMM-FNF with Stay, and HCMM-FNF with Move. For each imputation model, the error bars are obtained by applying the combination rule on all its imputations. The difference between this and Figure 4.21 is small, which does not affect our main conclusion in Section 4.4.

B.2.2 Low Mutual Information (LMI) Study
Figure B.8: The marginal probability distributions of log(1+INCTOT) (standardized) in the study of LMI. In each plot, the black solid line with gray area is from the truth with its 95% bootstrap confidence interval. The other lines are from the completed datasets after imputation. There is no significant difference between this and Figure 4.23.
Figure B.9: The marginal probability distributions of OCCSCORE (standardized) in the study of LMI. In each plot, the black solid line with gray area is from the truth with its 95% bootstrap confidence interval. The other lines are from the completed datasets after imputation. There is no significant difference between this and Figure 4.24.
Figure B.10: The marginal boxplots of EDUC in the study of LMI. Each panel corresponds to a factor level in EDUC. And each boxplot is a summary of all 10 imputed counts in that factor level from a particular imputation model. Here, H, M, S, T stand for HCMM-LD, Move, Stay, and Truth, respectively. There is some difference between this and Figure 4.25. However, such a difference does not affect the main conclusion we get in Section 4.4.
Figure B.11: The marginal boxplots of DIFFSENS in the study of LMI. Each panel corresponds to a factor level in DIFFSENS. And each boxplot is a summary of all 10 imputed counts in that factor level from a particular imputation model. Here, H, M, S, T stand for HCMM-LD, Move, Stay, and Truth, respectively. There is no significant difference between this and Figure 4.26.
Figure B.12: The contour plots of the kernel density for log(1+INCTOT) and OCCSCORE (standardized) for the missing observations in the study of LMI. For each imputation model, the contour plots of all its imputed kernel densities are identically and independently drawn and look similar, so that we randomly select one imputation for the purpose of display. There is no big difference between this and Figure 4.27.
Figure B.13: The contour plots of the kernel density for PROPTX99 and the standardized log(1+INCTOT) for the missing observations in the study of LMI. For each imputation model, the contour plots of all its imputed kernel densities are identically and independently drawn and look similar, so that we randomly select one imputation for the purpose of display. There is no significant difference between this and Figure 4.28.
Figure B.14: The contour plots of the kernel density for PROPTX99 and OCC-Score (standardized) for the missing observations in the study of LMI. For each imputation model, the contour plots of all its imputed kernel densities are identically and independently drawn and look similar, so that we randomly select one imputation for the purpose of display. There is no significant difference between this and Figure 4.29.
Figure B.15: The means of the standardized log(1+INCTOT) by EDUC in the study of LMI. T, H, M, and S stand for True, HCMM-LD, Stay, and Move. For each imputation model, the error bars are obtained by applying the combination rule on all its imputations. There is some difference between this and Figure 4.30. However, such a difference does not influence the main conclusion in Section 4.4.

Figure B.16: The means of the standardized OCCSCORE by EDUC in the study of LMI. T, H, M, and S stand for True, HCMM-LD, Stay, and Move. For each imputation model, the error bars are obtained by applying the combination rule on all its imputations. The difference between this and Figure 4.31 is small, which does not affect the main conclusion in Section 4.4.
Figure B.17: The means of PROPTX99 by EDUC in the study of LMI. T, H, M, and S stand for True, HCMM-LD, Stay, and Move. For each imputation model, the error bars are obtained by applying the combination rule on all its imputations. There is some difference between this and Figure 4.32. However, such a difference does not influence the main conclusion we get in Section 4.4.

Figure B.18: The means of the standardized log(1+INCTOT) by DIFFSENS in the study of LMI. T, H, M, and S stand for True, HCMM-LD, Stay, and Move. For each imputation model, the error bars are obtained by applying the combination rule on all its imputations. There is some difference between this and Figure 4.33. However, such a difference does not affect the main conclusion we get in Section 4.4.
Figure B.19: The means of the standardized OCCSCORE by DIFFSENS in the study of LMI. T, H, M, and S stand for True, HCMM-LD, Stay, and Move. For each imputation model, the error bars are obtained by applying the combination rule on all its imputations. There is some difference between this and Figure 4.34. However, such a difference does not influence the main conclusion we get in Section 4.4.

Figure B.20: The means of PROPTX99 by DIFFSENS in the study of LMI. T, H, M, and S stand for True, HCMM-LD, Stay, and Move. For each imputation model, the error bars are obtained by applying the combination rule on all its imputations. There is some difference between this and Figure 4.35. However, such a difference does not affect the main conclusion we get in Section 4.4.
Bibliography


Wei, L. and Reiter, J. P. (2016), “Releasing synthetic magnitude microdata con-
strained to fixed marginal totals,” Statistical Journal of the International Associ-

Willenborg, L. and De Waal, T. (2012), Elements of Statistical Disclosure Control,

Yang, M., Buso, M., Butani, S., Hiles, D., Mushtaq, A., Pramanik, S., and Scheuren,
Biography

Lan Wei was born in Shanghai, China. Before coming to the U.S., she graduated from Shanghai Jiao Tong University. Then, she attended the University of Chicago and received her master degree in statistics in 2011. After that, she attended Duke University where she plans to graduate with her Ph.D. in statistics in 2016 under the supervision of Professor Jerome P. Reiter.