7-10-2015

Model-Based Clustering of Incomplete Data

Chantal Larose
University of Connecticut, chantal.larose@uconn.edu

Model-Based Clustering of Incomplete Data

Chantal Danielle Larose, Ph.D.

University of Connecticut, 2015

Several important questions have yet to be answered concerning clustering incomplete data. For example, how can disparate solutions from multiply imputed cluster results be resolved? Additionally, can a model-selection criterion be developed which can detect the correct number of LCA classes after multiple imputation has been performed? Finally, as cluster analysis depends on measures of uncertainty, what is the effect of missing values on such measures? This thesis presents new theorems, methodologies, and applications which demonstrate solutions to these pressing questions.

Model-Based Clustering of Incomplete Data

Chantal Danielle Larose

B.A., University of Connecticut, 2010

M.S., University of Connecticut, 2014

A Dissertation

Submitted in Partial Fullfilment of the

Requirements for the Degree of

Doctor of Philosophy

at the

University of Connecticut

2015

APPROVAL PAGE

Doctor of Philosophy Dissertation

Model-Based Clustering of Incomplete Data

Presented by

Chantal Danielle Larose, B.A., M.S.

Major Advisor

Dipak Dey

Major Advisor

Ofer Harel

Associate Advisor

Haim Bar

University of Connecticut

2015

DEDICATION

To my family.

To my father, Dr. Daniel T. Larose. Thank you for kindling my love of science, statistics, and The Adventures of Papageno and Friends. Your constant support, encouragement, and perspective has allowed me to achieve this wonderful dream.

To my sister, Ravel Larose, the coolest little geek I've ever met. Thanks for giggles, music night duets (*Ue ue!*), and for being your happy self.

To my mother, Debra Larose, and my brother, Tristan Larose, for putting up with a lot of statistical dinnertime conversations.

I learned this, at least, by my experiment; that if one advances confidently in the direction of his dreams, and endeavors to live the life which he has imagined, he will meet with a success unexpected in common hours.

- Henry David Thoreau, *Walden*

ACKNOWLEDGEMENTS

My eternal gratitude goes to my Ph.D. Co-Advisor, Dr. Dipak K. Dey. His support, patience, and intuitive understanding has meant more than words can say. He has touched many lives that are close to me, for which I cannot thank him enough. My heartfelt appreciation also goes to my other Ph.D. Co-Advisor, Dr. Ofer Harel. His care and supervision since my undergraduate days has opened doors for me that were not there before. Sincere thanks also to my Associate Advisor, Dr. Haim Bar. Our many conversations were invaluable to the completion of this thesis.

This research was supported in part by the National Institute of Mental Health, Award Number K01MH087219; the CLAS Graduate Fellowship from the University of Connecticut; and the Elizabeth Macfarlane Fellowship from Department of Statistics, University of Connecticut. The content of this dissertation is solely the responsibility of the authors, and it does not represent the official views of the National Institute of Mental Health or the National Institutes of Health.

Many thanks to Dr. Hwan Chung, Associate Professor in the Korea University Department of Statistics, for kindly supplying the R code for imputing LCA class membership and manifest variable values. Without him, the approach and comparisons in Chapter 4 would not have been possible.

Part of the computation was done on the Beowulf cluster of the Department of Statistics, University of Connecticut, partially financed by the NSF SCREMS (Scientific Computing Research Environments for the Mathematical Sciences) grant number 0723557.

Many thanks to the UConn Physics Department for supplying a LaTeX template for the UConn Ph.D. thesis. All tables were formatted using the R package *xtable* (Dahl, 2014).

The original data collection for the Application of Chapter 4 was funded by a seed grant from the Children, Youth and Families Consortium at the Pennsylvania State University (Kordas, PI). We would like to thank Dr. Elena Queirolo, Psychologist Graciela Ardoino, and Dr. Nelly Mañay for coordinating the study, and the study research team from the Catholic University of Uruguay for help in data collection. Also, sincerest thanks to Dr. Katarzyna Kordas, Head of Epidemiological Research, ALSPAC and Senior Lecturer in Epidemiology in the School of Social and Community Medicine, University of Bristol, for providing the data and translating the questionnaire.

Finally, a heartfelt thank-you to all the UConn faculty, staff, and graduate students. Their companionship has meant so much.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

1.1 Overview & Motivation

Cluster analysis, also called clustering, aims to uncover and describe patterns in a data set by separating the records into groups. Clustering methods have been used in social (Harding et al., 2012), behavioral (Bitsika et al., 2008), medical (Whitwell et al., 2099), environmental (Beck et al., 2013), public health (Fraga et al., 2010), and genetic (Yeung et al., 2001) research. Incomplete data, and patterns in the missing values, are common in these areas. However, default clustering methodologies in many statistical software packages require complete data. Other packages do address incomplete data, but require a single, pre-specified number of clusters. We seek to extend cluster analysis to incomplete data while considering a range of clusters.

Algorithms exist to cluster incomplete data, but many rely on single-imputation, and risk ignoring information in incomplete records (Schafer and Graham, 2002). For example, consider the K-Nearest Neighbors (KNN) methodology (Everitt

et al., 2011). KNN is a broad term for clustering methodologies which consider k records which are closest to an incomplete record, and use the information in these neighboring records to fill in the missing information in the record under consideration. While KNN is typically used in classification, where the missing value is only cluster membership and not a variable's value, variations of KNN have been used in an incomplete data clustering framework (Keerin and Kurutach, 2012; Bras and Menezes, 2007). However, by the nature of KNN, not all records are considered when filling in the missing values, and thus there is a risk of ignoring important information. Another example is Hathaway and Bezdek (2001), who perform modifications to the fuzzy c-means (FCM) clustering algorithm (Everitt et al., 2011) in order to account for incomplete records, all of which overlooking patterns in the missingness. Lagona and Picone (2012) discuss a maximum likelihood based approach to model-based clustering of skewed incomplete data. While parameter estimates can be obtained from maximum likelihood methods, the methodology can be inflexible, since each model under consideration requires its own process.

Basagaña et al. (2013) take the idea of clustering incomplete data into the realm of multiple imputation (MI) (Harel and Zhou, 2007; Rubin, 1987; Schafer, 1999). However, their methodology discards imputed data sets which do not supply the "best" cluster result, and thus does not address the question of how to combine clustering solutions with differing numbers of cluster centers.

Latent class analysis (Hagenaars and McCutcheon, 2002; Fruhwirth-Schnatter, 2006), another method of clustering data, has been incorporated into the multiple imputation framework. It is included both to impute missing values and to cluster records. Vermunt et al. (2008) used an LCA model with a large number of classes to impute categorical data using less computation time than other imputation models (e.g. Schafer (1999)). However, their focus is not finding clusters via LCA (Vermunt et al., 2008, pg. 378). Gebregziabher and DeSantis (2010) extended the methodology of Vermunt et al. (2008), but retained the emphasis on imputing, and not clustering, via latent class analysis. MI has been used to identify classes from LCA by treating the unobserved vector of class membership as a completely missing variable (Harel and Miglioretti, 2007). However, this methodology addressed an otherwise completely-observed data set. Harel et al. (2013) and Muthén and Muthén (2011) impute missing values via LCA and cluster the data, but both methodologies are limited to a single number of classes. Harel et al. (2013) chose the number of clusters before implementing MI. Muthén and Muthén (2011, Ch. 14, pg. 488) combines the parameter estimates of all multiply imputed models, which is only possible if the parameters describe the same number of clusters.

While the above are all valid methodologies, there are still important questions left unanswered when considering the intersection of cluster analysis and incomplete data. In this thesis, we focus on three such questions. First, how can we address incomplete normal mixture model data in a way that takes into

account all imputed clustering solutions without limiting ourselves to a single, pre-specified number of clusters? Second, how can we calculate the entropy of an incomplete data model, taking into account both the variability in the data and the additional uncertainty introduced by the missingness mechanism? Third, can we develop a model selection criterion for clustering incomplete categorical data that frees current methodologies from the restriction of pre-specifying the number of clusters, while also choosing the correct model more often than AIC and BIC?

In this thesis, we use the following notation. A capital letter represents a random variable (e.g., Y_1). A bold capital letter represents a matrix (e.g., $\mathbf{Y}$). A bold lower case letter represents a vector (e.g., $\mathbf{y_1}$). A lower case and unbold letter represents a scalar (e.g., y_{11}).

1.2 Literature Review

Now that we have our intended research directions clear, let us go back and detail the methodologies we have mentioned. In this section, we discuss cluster analysis, entropy, missing data, and multiple imputation.

1.2.1 Cluster Analysis

There are two general categories of cluster analysis: model-based clustering and deterministic clustering. For a thorough review of the methods listed in this section, please see Everitt et al. (2011) and Fruhwirth-Schnatter (2006).

Model-Based Clustering

Model-based clustering methods describe clusters using a mixture of probability distributions, where each distribution corresponds to a cluster (McLachlan and Peel, 2000; Banfield and Raftery, 1993; Fraley et al., 2012; Fraley and Raftery, 1998; Scott and Symons, 1971; Fraley and Raftery, 2002; Raftery and Dean, 2006; Wolfe, 1970). While each record is ultimately assigned to one cluster, it has a probability of belonging to every cluster. For this reason, model-based clustering is also referred to as "soft" clustering.

Scott and Symons (1971) and Wolfe (1970) first applied the method of likelihood maximization to the multivariate normal case. Banfield and Raftery (1993) improved the flexibility of the multivariate normal mixture model as applied to cluster analysis. One of the best known software packages for clustering normal mixture model data is Mclust (Fraley et al., 2012). Mclust, implemented in the apynomous R (R Core Team, 2014) package, has also been extended to include other distributions of continuous data (Lagona and Picone, 2012). Methods for clustering categorical data using mixture models include entropy-based methods (Liu et al., 2014) and latent class analysis (Hagenaars and McCutcheon, 2002; Fruhwirth-Schnatter, 2006).

Model-based clustering methods are able to consider a range of possible numbers of clusters, and determine the "best" one using familiar model-selection procedures (detailed in a following section). As model-based clustering has been

well-documented in the literature, this thesis uses notation from these publications whenever possible.

To illustrate, let $\mathbf{Y} = (\mathbf{y}_1, ..., \mathbf{y}_N)$ be a sample of N, K-variate records, $i = 1, ..., N$. Let f_g be the distribution which represents the g^{th} component, or cluster, in a mixture of G clusters. Each component has its own vector of parameters, written θ_g. The likelihood of the mixture is

$$L(\theta_1, ..., \theta_G|\mathbf{y}) = \Pi_{i=1}^{N} \Sigma_{g=1}^{G} \tau_g f_g(\mathbf{y}_i|\theta_g), \tag{1.1}$$

where τ_g is the probability of an observation $\mathbf{y}_i$ being from density f_g. These probabilities act as weights, therefore $\tau_g \geq 0$ for all g and $\sum_{g=1}^{G} \tau_g = 1$ (Fraley et al., 2012).

Mclust. The software Mclust (Fraley et al., 2012) is one approach to model-based clustering. It begins by replacing f_g in Equation 1.1 with the multivariate normal model

$$f(\mathbf{y}_i|\theta) = \frac{1}{(2\pi)^{-k/2}} |\Sigma|^{-1/2} e^{-\frac{1}{2}(\mathbf{y_i}-\mu)'\Sigma^{-1}(\mathbf{y_i}-\mu)}$$

which turns Equation 1.1 into a normal mixture model. For each of a finite range of G values, Mclust computes the corresponding likelihood using the EM algorithm (Dempster et al., 1997) to estimate parameters (Banfield and Raftery, 1993; Fraley et al., 2012).

Interestingly, while the EM algorithm can address missing values, the Mclust

algorithm does not handle missing values using the default settings. Mclust does contain a function to impute missing values using methods contained in Schafer's *mix* algorithm (Schafer, 1997), but multiple imputations are not performed, nor are the results combined. This practice amounts to single-imputation, the drawbacks of which are covered in the next section.

Latent Class Analysis. Latent class analysis (LCA) (Hagenaars and McCutcheon, 2002; Fruhwirth-Schnatter, 2006; Linzer and Lewis, 2011) is an approach to model-based clustering primarily used for categorical variables. LCA tries to explain the relationship between categorical variables by grouping the records into classes. In the language of the literature, the categorical variables are observed or "manifest" variables, while the class membership variable is an unobserved or "latent" variable whose values LCA seeks to find. The key assumption of LCA is that the categorical variables are independent, given that they are in the same class.

As shown in Harel et al. (2013), the LCA model specifies the probability that a record i is in class g as

$$f(\mathbf{y_i}|\pi_g) = \prod_{K} \left(\prod_{O_k} \left(\pi_{k,o|g}^{y_{i,k,o}} \right) \right),$$

where records are indexed $i = 1, ..., N$; variables are indexed $k = 1, ..., K$; classes (e.g. clusters) are indexed $g = 1, ..., G$; $\pi_{k,o|g}$ is the probability that a record is in class g, and has the o^{th} possible value ($o = 1, ..., O_k$) in the k^{th} variable; and $y_{i,k,o}$

equals 1 if the variable k for record i has the o^{th} value. The probability over all latent classes, G, is

$$P(\mathbf{y_i}|\pi,\gamma) = \sum_{g=1}^{G} \gamma_g \left[\prod_{K} \left(\prod_{O_k} \left(\pi_{k,o|g}^{y_{i,k,o}} \right) \right) \right], \tag{1.2}$$

where γ_g is the probability of being in the g^{th} latent class. Let $\gamma_g = \tau_g$ and $\prod_{K} \left(\prod_{O_k} \left(\pi_{k,o|g}^{y_{i,k,o}} \right) \right) = f_g(\mathbf{y}_i|\theta_g)$, and the connection between Equation 1.2 and Equation 1.1 is self-evident.

The two parameters of interest are the latent class probabilities, γ_g, and the conditional probabilities, π_{kgo}. The latent class probabilities give the weights of each class, while the conditional probabilities describe how likely it is that a variable has a particular value, given that it is in a specific class. The EM algorithm is used to estimate the values of γ_g and π_{kgo} (Harel et al., 2013; Hagenaars and McCutcheon, 2002).

Deterministic Clustering

Deterministic clustering breaks records into groups using measures of similarity. Deterministic clustering methods are often called "hard" clustering methods, because they consider a record belonging to one and only one cluster at a time, unlike model-based or "soft" clustering.

The goal of deterministic clustering is to find a clustering solution where all records which share a cluster are very similar, and all records which are in different

clusters are very dissimilar. Similarity is typically quantified using a distance measure, such as Euclidean distance. Some of the most popular deterministic clustering methods are K-Means (MacQueen, 1967) and K-Medoids (Kaufman and Rousseeuw, 1987). Other methods include hierarchical clustering (Everitt et al., 2011, Ch 4) and BIRCH (Zhang et al., 1997). There are also clustering methodologies which use entropy to assign records to clusters, in lieu of distance measures (Li et al., 2004; Barbara et al., 2002). However, due to the widespread use of K-Means and K-Medoids, those two methods are the only deterministic clustering methods used in this thesis.

The K-Means procedure begins by placing a pre-specified number of random, initial cluster centers into a data set. Clusters are formed by assigning records to the closest cluster center. Once these clusters are formed, the cluster means become the new cluster centers. Records are then re-assigned to the closest of the new cluster centers. Cluster centers and cluster membership are updated until no change occurs (MacQueen, 1967). The K-Medoids method uses a similarly structured algorithm, except it uses one of the records in each cluster as the cluster center (Kaufman and Rousseeuw, 1987).

Cluster Performance

There are different ways to determine which cluster solution should be considered the final, or "best," solution. Model-based clustering algorithms use model se-

lection criteria in order to identify the superior clustering model. For example, Mclust and LCA may use the *Bayesian Information Criterion* (BIC) (Banfield and Raftery, 1993):

$$2logp(\mathbf{Y}|M_g) \approx 2logp(\mathbf{Y}|\hat{\theta}_g, M_g) - \nu_g log(n) = BIC_g,$$

where $\mathbf{Y}$ is the data, M_g is the model, $\hat{\theta}_g$ is the maximum likelihood estimate of parameters θ_g, and ν_g is the count of parameters which were estimated. The value of BIC is calculated for each model under consideration, and the model with the lowest value is considered the best one. Since the clustering model, parameter estimates, and number of parameters to be estimated are all influenced by the number of clusters, choosing a model based on BIC will also choose a number of clusters to use.

Model-based clustering solutions may also be evaluated by an entropy criterion. Details are in the following section.

Deterministic clustering algorithms rely on other methods of determining the correct number of clusters, including silhouette values (Rousseeuw, 1987), adjusted rand index (ARI) (Rand, 1971; Hubert and Arabie, 1985), and the gap statistic (Tibshirani et al., 2001).

A record's silhouette value is $sil_i = \dfrac{b_i - a_i}{max(a_i, b_i)}$, where a_i is how far an observation i is from the center of the cluster it belongs to, and b_i is how far that same observation is from the center of the closest cluster which it does not belong

to. Silhouette values close to one indicate tight clusters with large spaces between them, which is desirable in a clustering solution (Rousseeuw, 1987).

The Rand Index (Rand, 1971) comapres how closely two clustering solutions agree on the clustering of pairs of observations. Agreement occurs when both clustering solutions place two records in the same cluster, or when they both place two records in different clusters. Disagreement occurs when one clustering solution says the records should be in the same cluster, and the other solution says they should be in separate clusters. Adjusted Rand Index (Hubert and Arabie, 1985) tweaks this measure to account for random agreement. If there is high agreement, prior or expert knowledge may influence the decision between two clustering solutions. If no such knowledge is available, the analyst may favor the simpler solution; e.g., the solution with fewer clusters.

The gap statistic measures the difference, or "gap," between the log of the observed variation within clusters and the log of the expectation of that variation. The number of clusters that results in the largest difference is considered the preferred number of clusters (Tibshirani et al., 2001).

1.2.2 Entropy

Entropy (Shannon, 1948; Cover and Thomas, 2006) measures the randomness (e.g. the uncertainty) of a stochastic system. It also measures the average information contained in a probability distribution, and as such relies on a probability

mass function (pmf) or probability density function (pdf). For a discrete random variable $\mathbf{A}$ with pmf $p(\mathbf{A})$, entropy is calculated by

$$H(\mathbf{A}) = -\sum_{\mathbf{A}} p(\mathbf{A}) ln\left[p(\mathbf{A})\right], \tag{1.3}$$

and for a continuous random variable with pdf $f(\mathbf{A})$, entropy is calculated by

$$H(\mathbf{A}) = -\int_{\mathbf{A}} f(\mathbf{A}) ln\left[f(\mathbf{A})\right]. \tag{1.4}$$

When we discuss the entropy of a data set, we are in fact discussing the entropy of the model or models which describe the unique records in the data. This thesis refers to such a quantity as the entropy of the data model. To illustrate, if we know that a record $\mathbf{x}_1$ has two values, which are jointly distributed bivariate normal (e.g. the height and weight of a single person), then the entropy of $\mathbf{x}_1$ is the entropy of a bivariate normal distribution. If all records in the data set are known to follow independent bivariate normal distributions, then the entropy of the data model is the sum of the entropy of the unique records.

As we can find the entropy of a probability distribution, so can we find the entropy of a mixture of distributions. Thus, we may calculate entropy of a model-based clustering solution, such as LCA. Entropy for LCA models is found using Dias and Vermunt (2008)

$$H(\alpha) = -\sum_{i=1}^{N} \alpha_{ig} ln(\alpha_{ig}),$$

where N is the number of records, and α_{ig} is

$$\alpha_{ig} = \sum_{g=1}^{G} \gamma_g \prod_{k=1}^{K} (\hat{\pi}_{kg1})^{y_{ik1}} (\hat{\pi}_{kg2})^{y_{ik2}},$$

the probability that a record i is in class g.

When looking for a number of classes to use in the final LCA model, one may consider the entropy values. Entropy has been used as a cluster goodness measure when combined with the log-likelihood (Biernacki and Govaert, 1997; Biernacki et al., 2000), as well as on its own (Li et al., 2004; Barbara et al., 2002).

1.2.3 Missing Data

If a person responding to a survey skips a question, their answer to that question becomes a missing value, and the data entry resulting from their response will be incomplete. The data set made up of all survey respondents will therefore also be incomplete. This simple example is one of a variety of ways missing values may occur.

How important is the missing data problem? How prevalent is the missing data problem in real-world data? Harel et al. (2012) examined prevalence of missing values in 57 HIV studies. The average amount of missing values across the studies was 26%, with a median of 23% and a range of 3 - 97%. These numbers show that missing data does, in fact, occur quite often in real data. Furthermore, 74% of the studies used CCA in their analyses. Since missing values tend to differ

by demographic subgroup (age, education level, etc.) (Cranford et al., 2008), we expect to have patterns in the missingness. These patterns are not accounted for in CCA. Therefore, results from such studies may be biased.

Missingness Mechanisms

There are three processes, commonly called missingness mechanisms, which generate missing values. These mechanism are Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR) (Rubin, 1976, 1987). MCAR assumes that the analyst can accurately describe how missing values occur by using a probability model which does not depend on the data set. MAR specifies that the pattern of missing values can be accurately described using a probability model based on observed values. MNAR requires that the probability model of missing values can only be accurately described using information from unobserved values.

To illustrate, let the data be written $\mathbf{Y} = (\mathbf{Y}_{obs}, \mathbf{Y}_{mis})$, where subscripts *obs* and *mis* denote the observed and missing values of the data, respectively. Let the data be described using $P(\mathbf{Y}|\theta)$, where θ is a vector of parameters. To describe the missingness in $\mathbf{Y}$, let there be a corresponding matrix $\mathbf{R}$ of the same dimensions as $\mathbf{Y}$, where each entry is equal to one if the corresponding entry in $\mathbf{Y}$ is observed and is equal to zero if the corresponding entry is missing. We may describe $\mathbf{R}$ using a probability model conditioned on a vector of parameters ϕ and

potentially conditioned on data values. Thus, we initially write the model for $\mathbf{R}$ as $P(\mathbf{R}|\phi, \mathbf{Y}_{obs}, \mathbf{Y}_{mis})$.

The missingness mechanism MCAR assumes that we may simplify $P(\mathbf{R}|\phi, \mathbf{Y}_{obs}, \mathbf{Y}_{mis})$ to $P(\mathbf{R}|\phi)$, since the missing values do not depend on observed or missing data. An example would be if $P(\mathbf{R}|\phi) = Bernoulli(0.5)$; here, a fair coin flip would determine whether each data value were observed. MAR assumes that we may simplify the model for $\mathbf{R}$ to $P(\mathbf{R}|\phi, \mathbf{Y}_{obs})$, since the missing values depend on observed (but not unobserved) data values. An example would be if a variable *Income* was missing for high values of a completely-observed variable *Age*. MNAR requires that the model for $\mathbf{R}$ remains $P(\mathbf{R}|\phi, \mathbf{Y}_{obs}, \mathbf{Y}_{mis})$, as the missingness depends on both observed and unobserved values. An example would be if a patient's reported pain level was missing because the pain was too great for the patient to report it.

Ignorability

When trying to obtain estimates from an incomplete data set, we can no longer focus solely on $\mathbf{Y}$, but must consider the joint distribution of $\mathbf{Y}$ and $\mathbf{R}$, written $P(\mathbf{Y}, \mathbf{R}|\theta, \phi)$. Modeling the joint behavior of $\mathbf{Y}$ and $\mathbf{R}$ is often difficult. However, under certain circumstances, the model for $\mathbf{R}$ can be ignored (Schafer, 1999; Rubin, 1976).

In general, there are two conditions under which you do not need to model

$\mathbf{R}$; for a detailed examination of when you can ignore $\mathbf{R}$, see Wood et al. (2005). First, MCAR or MAR mechanisms must have generated the missing values. Second, the parameters of $\mathbf{Y}$ and $\mathbf{R}$, θ and ϕ respectively, must be such that $f(\theta, \phi) = g(\theta)h(\phi)$. Under these conditions, we may rewrite the joint model of the observed data and the missingness, $P(\mathbf{Y}_{obs}, \mathbf{R}|\theta, \phi)$ as the product of $P(\mathbf{Y}_{obs}|\theta)$ and $P(\mathbf{R}|\mathbf{Y}_{obs}, \phi)$. Since we are only interested in θ, the parameter of the data, we may then focus only on $P(\mathbf{Y}_{obs}|\theta)$. We operate under the ignorability assumption throughout this thesis.

How to Handle Missing Data

The simplest method of addressing an incomplete data set is to use complete case analysis (CCA) and discard all incomplete records. If there is a pattern to the missing values, you risk overlooking that pattern by discarding incomplete records (Schafer and Graham, 2002). In our example of MAR missingness, *Income* was missing for large values of *Age*. If we wish to model the effect of *Age* on *Income*, CCA would delete records with high observed values of *Age* due to their missing *Income* value. Thus, the results would contain incorrect information about how the two variables interact. Even if there is no bias introduced by deleting incomplete records, CCA decreases sample size, thus wasting the resources spent in collecting the data.

Single imputation is another method for addressing missing values. Single

imputation replaces each missing value with a single estimate (Schafer and Graham, 2002). For example, the sample mean of observed values for *Income*, or the estimate obtained from regressing complete cases of *Income* on *Age*, could be used to fill in the missing values for *Income*. Creating these estimates retains the original sample size and allows the corresponding values of *Age* to be utilized, where they would otherwise have been discarded. However, it may overlook patterns in the missingness. In addition, using a single value such as the mean would decrease the variability in the incomplete variable; and using the regression estimate could overemphasize the relationship between the regressed variables. Finally, whichever single imputation estimate is used, the final result is a single data set, and every data value is treated as if it were the observed value. There is, in other words, no consideration for the variability within the imputation model, which means all observed variation is assumed to originate from the data. For imputed data, this is clearly not the case.

Multiple imputation (Harel and Zhou, 2007; Rubin, 1987; Schafer, 1999) generates multiple values to estimate each missing value. By substituting each value in turn, multiple data sets are generated. Not only is original sample size retained, but the variability among imputed values is easily calculated. Multiple imputation is thus our preferred method to address incomplete data sets.

1.2.4 Multiple Imputation

Multiple imputation (MI) (Harel and Zhou, 2007; Rubin, 1987; Schafer, 1999) produces $M > 1$ different, complete, simulated datasets, from which M sets of analysis results emerge. Rubin's rules (Rubin, 1987) combine point estimates from these analyses into a single estimate with a single measure of standard error that includes both the variation in the imputation model and variation in the data.

The Imputation stage of MI draws M sets of values from

$$\mathbf{Y}_{mis} \sim P(\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, \mathbf{R}),$$

the distribution of missing values based on observed values and the missingness mechanism. Under ignorability, the distribution becomes $\mathbf{Y}_{mis} \sim P(\mathbf{Y}_{mis}|\mathbf{Y}_{obs})$ (Rubin, 1976). Each set of simulated values is then incorporated into the original data set, resulting in M complete data sets that differ according to the set of imputed values that was used.

The Analysis stage of MI is when each of the M complete, simulated data sets is analyzed according to the analysis which was meant to be performed before encountering missing values. To illustrate, let Q be the parameter we want to estimate; for example μ or β_1. The Analysis stage results in point estimates $\hat{Q}_m$ (e.g. $\bar{Y}_m$ of $\hat{\beta}_{1,m}$) from $m = 1, ..., M$, along with the corresponding variances U_m.

The Combination stage uses Rubin's combining rules (Rubin, 1987) to meld the $\hat{Q}_m$ and U_m values into a single point estimate and variance. The point

estimate is calculated by $\bar{Q} = \frac{1}{M}\sum_{k=1}^{M}\hat{Q}_m$, and its variance is calculated $T = \bar{U} + (1 + \frac{1}{M})B$, where $\bar{U} = \frac{1}{M}\sum_{k=1}^{M}U_m$ and $B = \frac{1}{M-1}\sum_{m=1}^{M}(\hat{Q}_j - \bar{Q})^2$ (Rubin, 1987). To perform inference on $\bar{Q}$, we use the statistic $\dfrac{Q-\bar{Q}}{T^{1/2}}$, which has a t_ν distribution with degrees of freedom

$$\nu = (m-1)\left[1 + \frac{\bar{U}}{(1+m^{-1})B}\right]^{1/2}.$$

Fraction of Missing Information. Rubin (1987) bifurcated the formula for information of completely observed data into information from observed and unobserved parts of the data. The fraction of missing information (FMI) is the ratio of information in the unobserved data to the information in the complete data set. FMI measures the uncertainty in estimating a parameter. Thus, its value changes based on the parameter being estimated, even while estimating parameters for the same imputed data sets. To illustrate, recall the quantities given in the Combination Stage of MI. The estimate of FMI, $\hat{\lambda}$, is then

$$\hat{\lambda} = \frac{r + 2/(\nu+3)}{r+1}, \tag{1.5}$$

where $r = (1+m^{-1})B/\bar{U}$, and $\nu = (m-1)(1+r^{-1})^2$.

Ways to Multiply Impute Data. There are different methods of imputing data, the choice of which may depend on the data structure.

To generate simulated data values when the data are normally distributed, the EM algorithm and data augmentation can be used (Schafer, 1999) via the algorithm *Norm* (Schafer, 2008). Norm assumes each variable is marginally normally

distributed. Norm is reasonably robust, and may be used to imputed categorical data by rounding the imputed normal values. However, when data come from a normal mixture model, each cluster is normally distributed, but the variables themselves are not. Complications (e.g. introduction of false signals) arise if Norm is applied to such data without regard to this point of fact.

Predictive mean matching (Heitjan and Little, 1991; Schenker and Taylor, 1996) uses regression models to impute data. For each incomplete record, a number of complete records are identified whose observed values are close to the observed values of the incomplete record. Then a complete-case regression model is built, where the incomplete variable is regressed on the complete variables. The predicted values of the complete records are found, and one is chosen to be the imputed value of the incomplete record.

To impute categorical methods in a latent class framework, one can impute the class memberships and data values iteratively, as in Harel et al. (2013). By using the EM algorithm to find starting values of parameters for the LCA model, and MCMC to optimize the parameter values, Harel et al. (2013) imputes the class membership vector by observing which class each record has the largest probability of belonging to. Once classes are imputed, one can impute missing values of manifest variables by observing which categorical value is likely, given the class membership.

1.3 Outline

The thesis is organized in the following way.

Chapter 2 begins our venture into clustering incomplete data by presenting MICA: Multiply Imputed Cluster Analysis. MICA is a framework of steps which enables clustering of incomplete data sets while accounting for incomplete records and multiple clustering solutions. MICA utilizes a unique, two-stage clustering approach to rectify the issues encountered when using clustering with multiple imputation.

The chapter also includes the extension of MICA to normal mixture models, MICA-N. Two simulation studies determine which clustering algorithms perform best during each of the two MICA-N clustering stages, and shows how MICA-N outperforms complete case analysis. An application to genetic data illustrates MICA-N outperforming complete case analysis on a real data set.

Chapter 3 develops theorems that quantify the effect of missing values on Shannon entropy (Shannon, 1948; Cover and Thomas, 2006). Theorems show how MCAR and MAR missingness impact the estimators for entropy, and demonstrate how the new estimators approach their complete-case counterparts when the percent of missing values goes to zero. Simulations illustrate the behavior of entropy of the incomplete data model under MCAR and MAR missingness. Simulations also compare such behavior to that of the entropy of the fully observed and complete case data models, as well as to FMI.

Chapter 4 presents a new entropy-based LCA model selection criterion for multiply imputed data. We prove that entropy of an LCA model obtains its minimum of zero when the number of classes equals the number of unique records. We then use that information to build a model selection criterion with a penalty function which will recognize when entropy has begun to tail off toward zero. Different penalty thresholds are examined. The new entropy-based criterion outperforms AIC and BIC in simulation studies. The entropy-based criterion also discovers more nuanced and informative classes in a human development and family studies data set (Rink et al., 2014) than AIC and BIC.

Chapter 5 reviews the work presented in this thesis, highlighting new contributions to the literature. It underlines the three important questions left unanswered by previous methodology, and how each project in this thesis tackled one of those questions. Short-term and long-term research goals for each project are also discussed.

Chapter 2

Clustering Incomplete Data using Normal Mixture Models

2.1 Introduction

In this chapter we develop a new clustering methodology that can cluster multiply imputed data which follow a normal mixture model without being limited to a pre-specified number of clusters.

As detailed in the previous chapter, multiple imputation (MI) (Harel and Zhou, 2007; Rubin, 1987; Schafer, 1999) is the method most suitable for addressing incomplete data. MI has a simply-summarized and straightforward application, if the analysis which was intended before the question of missing values arose was an analysis which centered on point estimates. However, we encounter several problems when applying MI to cluster analysis. One such problem is that clustering multiply imputed data sets may result in cluster solutions with differing numbers of clusters. How can we resolve the contradictory information presented by differing cluster solutions, and obtain a single solution without discarding relevant information present in the imputed data? Another problem arises when imputing

the data. Normal mixture model data typically do not have marginally normal variables. If one uses an imputation model for normal data, the risk of introducing false signal to the data is high. To correctly impute data from a normal mixture model, additional steps must be taken.

We address these concerns while allowing consideration of a range of cluster amounts, instead of limiting the focus to a particular number of clusters. Imputation from a latent class model is not applicable in this scenario, as it would require prior specification of the number of clusters (Harel et al., 2013). For a new model-selection approach to latent class membership imputation for categorical data, which frees the current methodology from that restriction, please see Chapter 4.

The methodology we propose in this chapter is called MICA: Multiply Imputed Cluster Analysis. MICA is a general framework of steps whose steps can be specified to address particular data types. The specialized version of MICA for data that follows a normal mixture model is MICA-N.

This chapter is organized as follows. Section 2.2 begins by presenting MICA, and detailing the core ideas of the procedure. We continue in the same section to introduce the special algorithms and exact procedures which tailor our method to incomplete normal mixture model data. Section 2.3 investigates which clustering algorithms should be utilized in MICA, and whether they outperform CCA. Section 2.4 digs deeper into the MICA-N algorithm choices and its performance

against CCA with a different, more complex data set. Section 2.5 takes the finalized version of MICA-N and applies it to a genetic data set. Section 2.6 summarizes all results and discusses the next steps.

2.2 MICA: Multiply Imputed Cluster Analysis

The methodology we have crafted is called MICA: Multiply Imputed Cluster Analysis. MICA is performed in five steps: Preliminary Step, Imputation Step, Stage 1 Clustering, Stage 2 Clustering, and Cluster Membership Assignment. The general algorithm is presented first, followed by the specialization MICA-N.

2.2.1 MICA

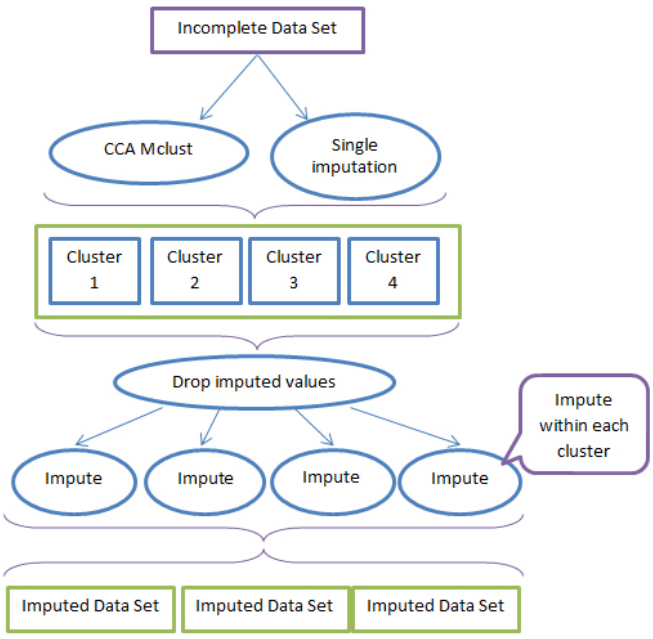


Fig. 2.1: Flowchart illustrating the Preliminary and Imputation steps of MICA

Here follows a description of the MICA framework. The Preliminary and Imputation steps are illustrated in Figure 2.1, while the remaining steps are illustrated in Figure 2.2.

Preliminary Step. The goal of the preliminary stage is to break the incomplete data set into incomplete clusters, with the sole purpose of allowing the Imputation Step to occur within each cluster. Two processes are run side-by-side: first, the complete case data is clustered; second, a single imputation step temporarily fills in the data. The temporarily-complete data is then assigned to the closest CCA cluster. Once this initial cluster membership information is obtained, the imputed values are dropped. The Preliminary stage results in a single set of incomplete clusters.

Concerns may exist about CCA clustering in the Preliminary step resulting in a biased set of clusters. However, simulations using MAR missingness - where bias in CCA analyses is likely to happen - show MICA obtaining lower misclassification proportions than CCA alone. See Section 2.4 for further discussion, and Section 2.6 for planned comparisons to alternative approaches.

Imputation Step. The goal of the Imputation stage is to impute missing values within each cluster, utilizing the cluster membership obtained in the Preliminary stage. Within each cluster, M sets of values are imputed and substituted for the missing values. Once imputed values are obtained, the cluster membership information from the Preliminary Step is dropped. The Imputation stage results

in M complete data sets, with no cluster membership information.

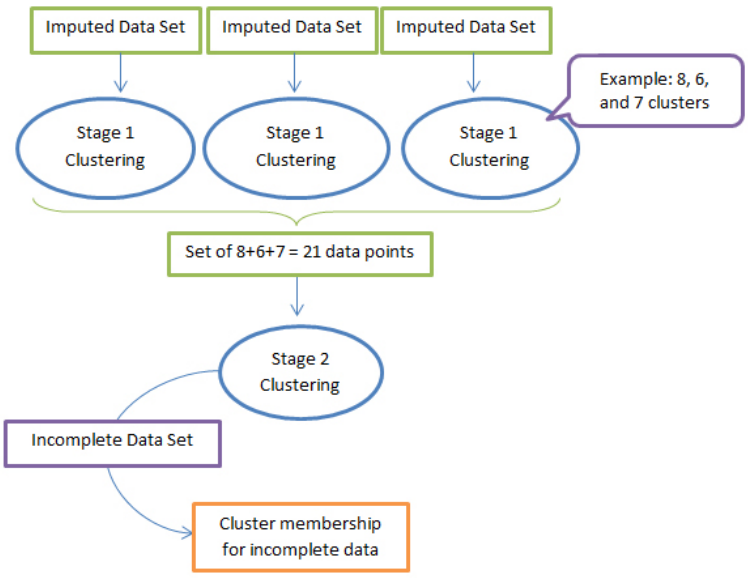


Fig. 2.2: Flowchart illustrating the Stage 1 Clustering, Stage 2 Clustering, and Cluster Membership steps of MICA

Stage 1 Clustering. The goal of Stage 1 Clustering is to obtain M clustering results from the M different, complete data sets. Each imputed data set from the previous step is clustered. The Stage 1 Clustering step results in M different clustering results, called Stage 1 clusters.

Stage 2 Clustering. The goal of Stage 2 Clustering is to retain information from all Stage 1 cluster analyses and yet consolidate the results into one set of clusters. To this end, the cluster centers from each Stage 1 clustering result are brought together into a single dataset. If there are three imputations in Stage 1, which result in eight, six, and seven clusters, then there will be $8 + 6 + 7 = 21$

records (i.e. cluster centers from Stage 1) clustered during this step. The result from Stage 2 Clustering is a single set of clusters, called Stage 2 clusters.

Cluster Membership Assignment. Arguably the most important result from cluster analysis is identifying cluster membership of all records. To assign membership to incomplete records, we look at the imputed versions of each record. These versions will be identical in the case of complete records, and different in the case of imputed records. We calculate the Euclidean distance from each imputed version of a record to all the Stage 2 cluster centers. The closest cluster center is the cluster which each imputed version of the record belongs to. The record is ultimately assigned to the cluster its imputed versions belong to most often.

As we are comparing our method to the complete case analysis equivalent, we must specify exactly what the CCA equivalent contains. The CCA version of MICA is made up solely of Stage 1 clustering, with an abbreviated Cluster Membership Assignment step afterwards. There is no Preliminary, Imputation, nor Stage 2 Clustering steps contained in the CCA version.

Finally, a word about label-switching. If you have two clusters in your data, Left Cluster and Right Cluster, running multiple cluster analyses on the data does not ensure that Left Cluster is labeled the same in every cluster result. The first cluster solution may call Left Cluster "Cluster 1," and the second cluster solution may call it "Cluster 2." How does one avoid the label switching problem?

MICA avoids label switching issues due to its use of Stage 2 cluster analysis. If analyses were halted at Stage 1, the particular labeling of clusters across imputations would introduce confusion. However, in Stage 2, all cluster centers from Stage 1 are gathered into a new dataset, which is then clustered only once. Since the final result relies on only one cluster solution, label switching is a non-issue.

2.2.2 MICA-N

When tailoring the MICA framework to the case of normal mixture models, the clustering and imputation algorithms must be specified. Much of the specification is aided by the knowledge that the clusters in our data are normally distributed.

Preliminary Step. The imputation method used is predictive mean matching (Heitjan and Little, 1991; Schenker and Taylor, 1996). Mclust (Fraley et al., 2012) is used in the CCA cluster analysis, to utilize our knowledge that clusters follow a normal mixture model.

Imputation Step. The imputation algorithm Norm (Schafer, 2008, 1997) is used, since we know that every cluster - and thus every group of data being imputed - is normally distributed.

Stage 1 Clustering. The Stage 1 cluster algorithm is Mclust, since the imputed data follows a normal mixture model. This choice of clustering algorithm is validated by the simulation study in Section 2.3.

Stage 2 Clustering. In Stage 2, the data consist of cluster centers. The data

are almost surely not normally distributed. Simulation studies (Sections 2.3 and 2.4) will determine which clustering algorithm under consideration performs best during Stage 2.

2.3 Iris Data Simulation Study

We begin investigating which clustering algorithms work best in the MICA-N framework by studying the Fisher's Iris data set (Anderson, 1935; Fisher, 1936). Simulations used the 100 records of Virginica and Versicolor flowers; Setosa flowers were omitted because they did not overlap with the other two species. The data variables are Sepal Width, Sepal Length, Petal Width, Petal Length, and Species. Species information was used only to verify cluster membership results, and was not included in the cluster analyses. The data was broken into two parts, Training and Testing data sets. Missingness was imposed only on the Training data. Testing data was used during Stage 2 to see how accurately the cluster results handled new observations.

Cluster performance was measured using *Proportion Misclassified* (PM) and *Proportion of data in the Versicolor cluster* (PV). Results were computed for Stage 1 and Stage 2. PM measured how many Versicolor records were assigned to the Virginica cluster and vice versa. PV was computed by dividing the number of records classified as Versicolor by the total number of records in the dataset. During MICA, *Fisher's Z Transformation* (Fisher, 1921) was used to transform

the point estimates to approximate a Normal distribution. The transformation allows us to apply Rubin's Rules (Rubin, 1987) to obtain a single point estimate and standard error of PM and PV.

These initial investigations used the same clustering methodology (K-Means, K-Medoids, or Mclust) during both stages of MICA-N. K-Means, K-Medoids, and Mclust were implemented using their respective R packages (R Core Team, 2014; Maechler et al., 2014; Fraley et al., 2012). Missingness was imposed on Sepal Width using MCAR, and MAR. Under MAR, Sepal Width was missing if the corresponding value of Petal Width was below a certain quantile. Training data percentages of 50% and 75% were used; 50% results were inconclusive and are thus omitted. Percent of missingness was set at 10%, 30%, and 50% to simulate a small, moderate, and large amounts of missing values. Imputations were set at 10, 50, and 100 to illustrate the effect of a small, moderate, and large amount of imputations. Different amounts of imputations resulted in very similar results, thus only results for 50 imputations are shown.

2.3.1 Results

Figure 2.3 illustrates the impact of MCAR and MAR on PM and PV during Stage 1. Small PM values indicate better clustering performance. Mclust has the lowest PM values. PV values closest to the true value, indicated by the horizontal dashed line, are the most accurate. Mclust is consistently the most accurate. Since Mclust

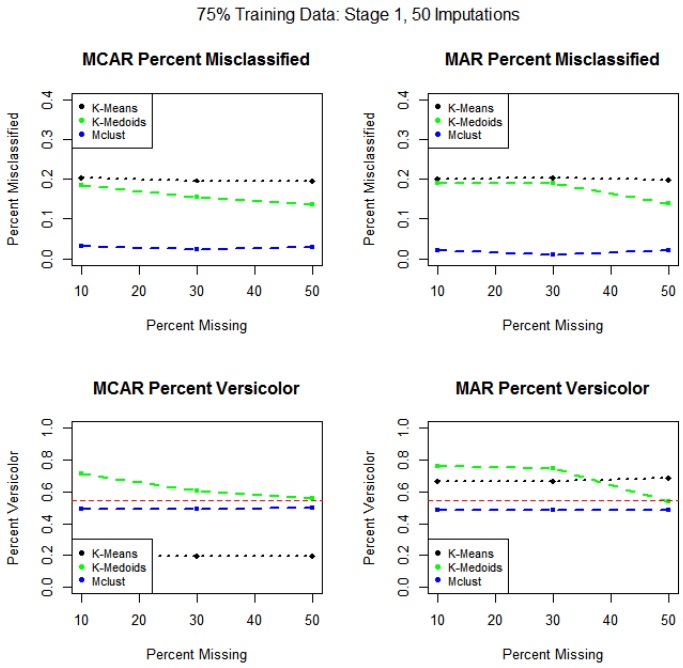


Fig. 2.3: PM and PV for Stage 1 clustering results under MCAR and MAR. Mclust has the lowest PM and most accurate PV, therefore we conclude that Mclust is the best clustering method for Stage 1.

has the lowest PM and most accurate PV, we prefer Mclust for Stage 1.

Figure 2.4 demonstrates the effect of MCAR and MAR on PM and PV during Stage 2. There is no one clustering algorithm which has the lowest PM. Similarly, there is no one clustering algorithm which has the most accurate PV.

We conclude that Mclust is the best clustering method for Stage 1, and that further study is necessary to determine the best algorithm for Stage 2.

Since we have hit upon the best clustering algorithm for Stage 1, we proceed to compare the performance of MICA-N's Stage 1 to the CCA equivalent. In

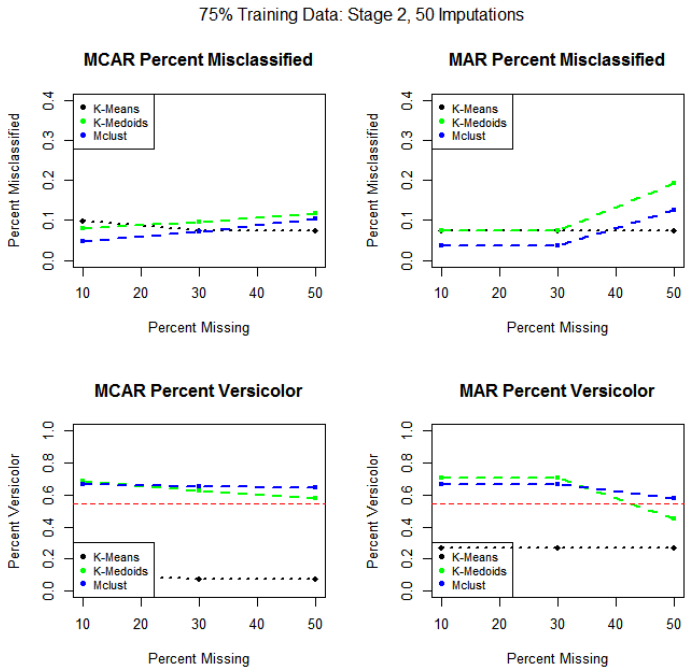
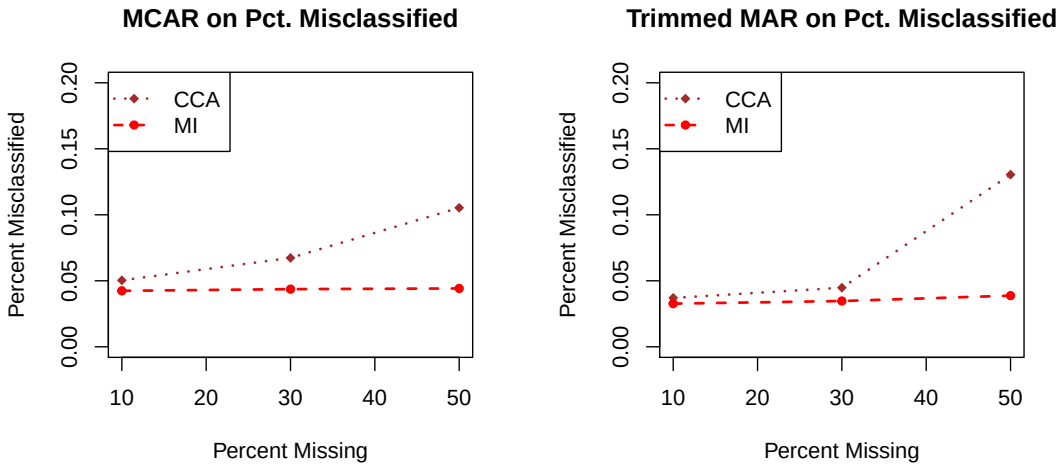


Fig. 2.4: PM and PV for Stage 2 clustering results under MCAR and MAR, when the same cluster method was used in both Stage 1 and Stage 2. Results are inconclusive.

this study, we use a new MAR mechanism which we call *Trimmed MAR*. To illustrate: If Versicolor records had values of Petal Width below a certain quantile, the corresponding values of Sepal Width were coded as missing. However, if Virginica records had values of Petal Width that were above a certain quantile, the corresponding values of Sepal Width were coded as missing. The result was that the two species' clusters were made harder to distinguish, by having their most unique and identifiable records eliminated from the data set. In these simulations, 90% Training data was used.



(a) **(b)**

Fig. 2.5: MCAR and Trimmed MAR on PM for CCA vs. Stage 1 MICA. PM steadily increases for CCA as percent missing increases, while MICA-N PM remains consistently low. We conclude MICA-N Stage 1 outperforms its CCA counterpart.

Figures 2.5 show Stage 1 PM and CCA PM for MCAR and Trimmed MAR missingness, as the percent of missing data increases. CCA PM values under both missingness mechanisms steadily increase, while MICA-N PM values remain below 5% regardless of the percent of missing data. Even though there are fewer records in CCA with 50% missing, the number of misclassified records is still higher than Stage 1 MICA-N. We conclude that Stage 1 of MICA-N outperforms its CCA counterpart.

2.4 Genetic Data Simulation Study

There are two goals to this second simulation study. First, determine which clustering method works best in Stage 2. Second, determine whether the combined Stage 1 and Stage 2 algorithms of MICA-N outperform CCA. Proportion Misclassified (PM) is used to measure cluster performance in both cases.

The data was based on six variables from Iyer et al. (1999). Originally, the data recorded human gene expression levels, with genes as the records and time points as the variables. The six variables we subset were the fifteen minute, one hour, four hour, eight hour, sixteen hour, and twenty-four hour measurements; chosen because they ran the breadth of the data set while reducing the dimension and avoiding temporally adjacent measurements.

The data is not organized into normally distributed clusters, nor is true cluster membership known. In order to obtain the data structure and information necessary to run our simulation study, Mclust was applied to obtain a normal mixture model approximation of the data. The parameters from this normal mixture model approximation (given below) were used to generate normal mixture model data for the simulation, and thus we consider these parameters the true structure of the data. Cluster membership information obtained during this step is used only to evaluate cluster membership accuracy after MICA-N is performed; it is not included in either the imputation or the cluster analyses.

The mixture model gave five clusters, with the following parameters:

$$\mu_{\mathbf{1}} = \begin{pmatrix} 0.94 \\ 0.88 \\ 0.67 \\ 0.49 \\ 0.73 \\ 0.79 \end{pmatrix}, \Sigma_1 = \begin{pmatrix} 0.04 & 0.02 & 0.01 & 0.01 & 0.00 & 0.00 \\ 0.02 & 0.04 & 0.02 & 0.01 & -0.00 & -0.01 \\ 0.01 & 0.02 & 0.04 & 0.01 & -0.00 & -0.01 \\ 0.01 & 0.01 & 0.01 & 0.03 & 0.01 & 0.01 \\ 0.00 & -0.00 & -0.00 & 0.01 & 0.05 & 0.05 \\ 0.00 & -0.01 & -0.01 & 0.01 & 0.05 & 0.10 \end{pmatrix},$$

$$\mu_{\mathbf{2}} = \begin{pmatrix} 1.07 \\ 2.04 \\ 1.32 \\ 1.41 \\ 1.25 \\ 1.36 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 0.11 & 0.03 & -0.03 & -0.01 & -0.01 & -0.04 \\ 0.03 & 1.12 & 0.12 & 0.19 & -0.16 & -0.26 \\ -0.03 & 0.12 & 0.18 & 0.19 & 0.11 & 0.17 \\ -0.01 & 0.19 & 0.19 & 0.80 & 0.56 & 0.63 \\ -0.01 & -0.16 & 0.11 & 0.56 & 0.82 & 0.98 \\ -0.04 & -0.26 & 0.17 & 0.63 & 0.98 & 1.21 \end{pmatrix},$$

$$\mu_{\mathbf{3}} = \begin{pmatrix} 1.40 \\ 3.72 \\ 5.27 \\ 6.38 \\ 3.46 \\ 2.89 \end{pmatrix}, \Sigma_3 = \begin{pmatrix} 0.54 & 0.95 & 0.84 & -0.50 & -0.54 & -0.62 \\ 0.95 & 9.68 & 5.01 & -3.81 & -3.98 & -1.60 \\ 0.84 & 5.01 & 50.22 & 29.17 & 4.87 & 3.84 \\ -0.50 & -3.81 & 29.17 & 32.23 & 12.76 & 6.74 \\ -0.54 & -3.98 & 4.87 & 12.76 & 11.73 & 6.64 \\ -0.62 & -1.60 & 3.84 & 6.74 & 6.64 & 5.96 \end{pmatrix},$$

$$\mu_{\mathbf{4}} = \begin{pmatrix} 1.05 \\ 0.97 \\ 1.02 \\ 1.20 \\ 2.07 \\ 3.51 \end{pmatrix}, \Sigma_4 = \begin{pmatrix} 0.05 & 0.03 & -0.02 & -0.07 & 0.01 & 0.23 \\ 0.03 & 0.08 & 0.01 & -0.04 & 0.02 & 0.31 \\ -0.02 & 0.01 & 0.11 & 0.13 & 0.07 & -0.04 \\ -0.07 & -0.04 & 0.13 & 0.31 & 0.09 & -0.26 \\ 0.01 & 0.02 & 0.07 & 0.09 & 0.46 & 0.73 \\ 0.23 & 0.31 & -0.04 & -0.26 & 0.73 & 4.63 \end{pmatrix},$$

$$\mu_{\mathbf{5}} = \begin{pmatrix} 0.96 \\ 1.26 \\ 2.74 \\ 2.76 \\ 1.69 \\ 1.42 \end{pmatrix}, \Sigma_5 = \begin{pmatrix} 0.05 & 0.03 & 0.03 & -0.01 & 0.02 & 0.01 \\ 0.03 & 0.23 & 0.21 & 0.02 & -0.04 & -0.05 \\ 0.03 & 0.21 & 1.86 & 0.19 & -0.19 & -0.11 \\ -0.01 & 0.02 & 0.19 & 1.00 & 0.26 & 0.13 \\ 0.02 & -0.04 & -0.19 & 0.26 & 0.50 & 0.37 \\ 0.01 & -0.05 & -0.11 & 0.13 & 0.37 & 0.35 \end{pmatrix},$$

and whose mixture percentage is:

$$\tau = (53.8, 8.7, 7.3, 13.6, 16.6)$$

For each of the 250 repetitions in the simulation study, a thousand observations were generated from the mixture model. This generated data was then broken into two parts. Approximately 75% of the data was put into a Training data set, while the remaining 25% was held out as a Testing data set. Missing

values were imposed only on the Training data. Percent of missingness was set at 10%, 30%, and 50% to simulate a small, moderate, and large amounts of missing values. Testing data was used during Stage 2 to see how accurately the cluster results handled new observations.

MAR missingness was imposed on the variable which represented the four hour measurement, which was made missing if the fifteen minute measurement was below the 10th, 25th, or 50th percentile, as dictated by the specified value of percent missing. Imputations were set at 50 to illustrate the effect of a moderate amount of imputations.

2.4.1 Results

Figure 2.6 illustrates notched boxplots of PM for CCA and three variations of MICA-N. The variations use K-Means, K-Medoids, and Mclust during Stage 2. Three plots show PM values at 10%, 25%, and 50% missingness. Non-overlapping notches show significant difference between medians (Chambers et al., 1983). Dashed lines indicate the bottom of the CCA notch. Our preferred Stage 2 clustering algorithm will have a notch which is entirely below the dashed line.

Across all figures, K-Means notches lie entirely above CCA notches, showing that K-Means medians are significantly higher than CCA at all levels of missing values. For 10% and 25% missing values, K-Medoids and Mclust medians have notches which lie entirely below the dashed line, therefore both have significantly

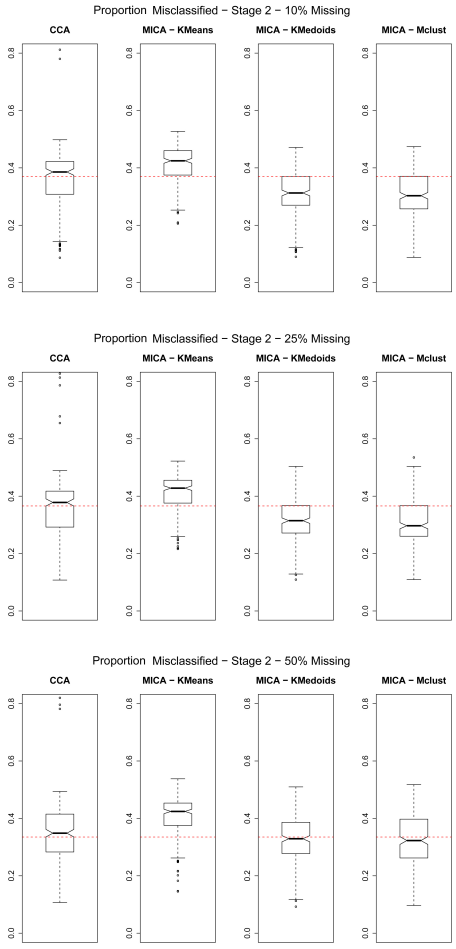


Fig. 2.6: Notched boxplots of Proportion Misclassified for CCA and variations of MICA-N using K-Means, K-Medoids, and Mclust during Stage 2.

lower medians than CCA in those cases. At 50% missing values, K-Medoids and Mclust have notches which overlap with the CCA notch. This suggests that 50% missing values in the four-hour measurement makes the clusters as difficult to distinguish via MI as via CCA. However, the MI methods do lack the outliers present in the CCA results.

Since the notched boxplots for K-Medoids and Mclust are nearly identical,

the two methods produce comparable PM results in Stage 2. We prefer K-Medoids over Mclust for the following reason. The records being clustered in Stage 2 are not necessarily normally distributed, therefore using Mclust in Stage 2 imposes an unrealistic assumption on the data. Since K-Medoids has values of PM comparable to Mclust and avoids imposing unrealistic assumptions, we prefer K-Medoids for Stage 2. We therefore decide that MICA-N will use Mclust in Stage 1 and K-Medoids in Stage 2. The combination of these two clustering algorithms outperformed CCA, as shown in Figure 2.6.

2.5 Application

The data contains gene expression levels of 6,178 yeast genes during a cell cycle (Cho et al., 1998; Yeung et al., 2001). Genes are the records in the data, and time points at which the gene expression levels were measured are the variables. A previous analysis of the data applied Mclust to complete records only, and found a five-component mixture which nicely paralleled the five phases of the cell cycle (Yeung et al., 2001). However, only complete records were used. In addition, only genes which were known to already have significant expression in particular cell cycle phases were used. We use all genes in the data set, including those with missing values. Since we include all genes, not only those known beforehand to have high expression values in particular cell cycle phases, our complete case data differs from the data analyzed previously.

From the original dataset of 6,178 records, there were only 1383 complete records. Using only complete data reduces the number of records by nearly 78%. A moderate 50 imputations was chosen to save on computational time while still introducing variability between imputations.

2.5.1 Results

MICA-N was performed on the entire incomplete data set. Stage 1 examined cluster solutions with between one and eleven clusters, inclusive, since previous analyses found that clustering results on this data do not supply more than eleven clusters (Larose, Dey, Harel, 2014). Stage 2 could not be automated, as in the Simulation studies, since the true cluster membership of the records is unknown. It is up to the analyst to decide what is the most appropriate and useful number of clusters to use. We examined K-Medoids results from two clusters (the smallest useful number) to twenty-two clusters (twice the highest number of clusters from Stage 1). The results are analyzed using silhouette values (Figure 2.7) (Rousseeuw, 1987), Adjusted Rand Indices (Table 2.1) (Rand, 1971; Hubert and Arabie, 1985), the Gap statistic (Figure 2.8) (Tibshirani et al., 2001), and expert knowledge (Yeung et al., 2001).

First, we utilize silhouette values. In Figure 2.7, Stage 2 records (i.e. the cluster centers from Stage 1) are ordered by silhouette value within each clustering result. The clustering solution with five clusters has continually high silhouette

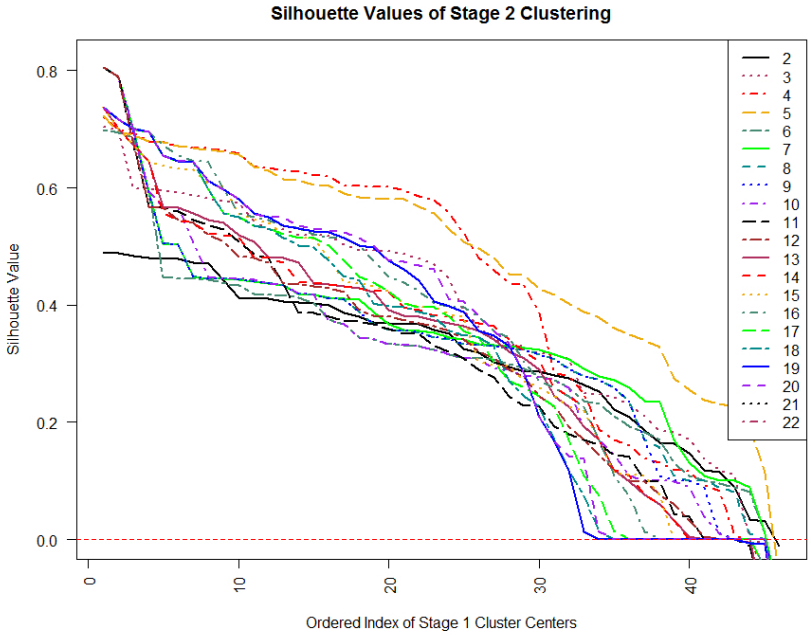


Fig. 2.7: Silhouette values for cluster solutions using two through twenty-two clusters. The five-cluster solution has consistently high silhouette values. The four-cluster solution also performs well. We decide to examine the four- and five-cluster solutions in more detail.

values. The next-best solution is the four-cluster solution. Using the information in the silhouette plot, we begin to suspect that four or five clusters may be the best solution.

Second, we consider Adjusted Rand Indices. In Table 2.1, the $(i, j)^{th}$ entries are Adjusted Rand Index values comparing cluster solutions with i and j clusters. If there is high agreement, the analyst may prefer the simpler solution or the solution which reflects prior knowledge. There is moderate agreement between

four and five clusters (0.81), but not enough to tempt us away from the five-cluster solution without further examination.

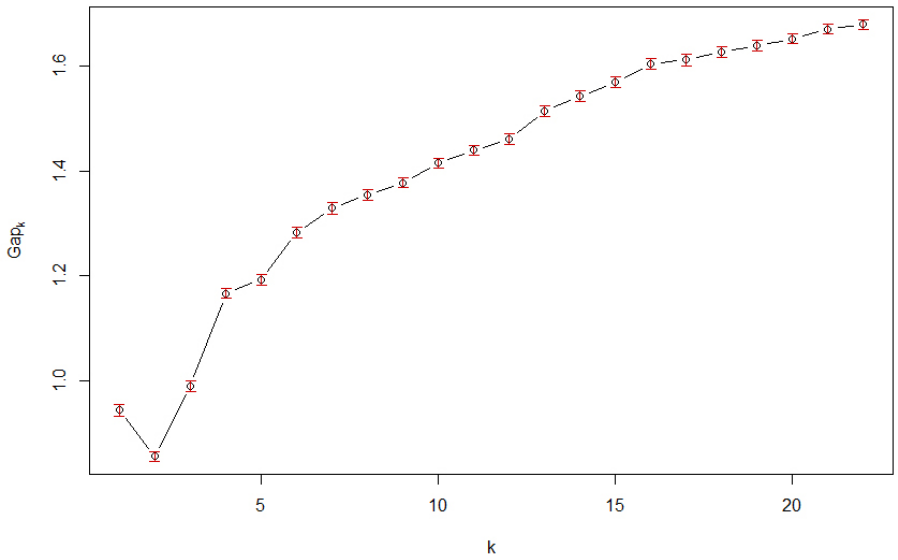


Fig. 2.8: Plot of the gap statistic Gap_g by the number of clusters g. Ideally we look for a maximized value of Gap_g. However, in this case we are left looking for an inflection point, which we find at $g = 5$.

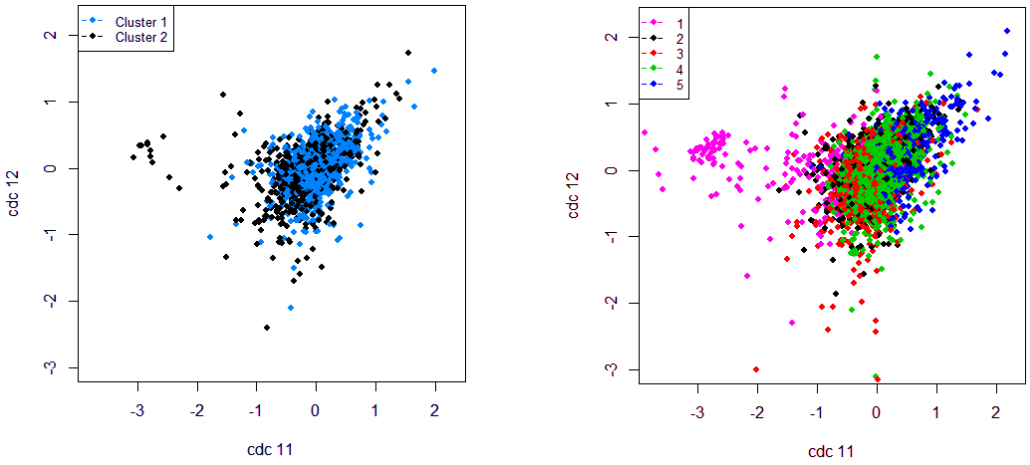
Third, the Gap statistic is calculated for one to twenty-two clusters. Graphical presentation of the Gap statistic is common, as the maximum values and general behavior of the statistic is easily understood. Figure 2.8 shows values of the gap statistic Gap_g by the number of clusters g. Ideally, analysts look for the maximized value of Gap_g. However, in this case we are left looking for an inflection point, which we find at $g = 5$. Therefore, we have increasing evidence that five clusters is the appropriate solution.

We make our final decision with help from expert knowledge. The gene

expression levels are known to be measured across five phases of the cell cycle. Therefore, one would expect to uncover five clusters in the data; one for each cell cycle. Let us combine this knowledge with the rest of our information. Silhouette values have demonstrated that five clusters gives the highest amount of cluster distinctness and tightness out of all cluster solutions considered. Adjusted Rand Indices have shown that no other cluster solution has high agreement with the five-cluster solution. The Gap Statistic reinforced our idea that $G = 5$ was best. Therefore, we decide to use five clusters during Stage 2.

CCA resulted in two clusters. Figure 2.9 shows two pairs of variables from the CCA data, colored by CCA cluster membership (2.9a); and the same pair of variables from one imputed dataset, colored by MICA-N Stage 2 cluster membership (2.9b). The CCA clusters do not describe any pattern in the dataset; cluster membership is arbitrary. For example, in Figure 2.9, there is a tail of data extending to the far left. This string of observations is not describes by the binary cluster membership. By contrast, MICA-N cluster membership tends to group chunks of the data which reside in distinct areas of the plot. In other words, MICA-N has successfully captured a pattern in the data that was lost in the CCA.

In summary, it is clear that MICA-N gives a more detailed and informative description of a real-world data set, as compared to the simplistic and uninformative CCA analysis.



(a) Two CCA clusters. Results show no discernable pattern.

(b) Five MICA-N clusters. Results clearly segment data into interpretable groups.

Fig. 2.9: Two pairs of variables from a single imputed dataset, color coded by records' membership to the two CCA Stage 2 clusters (2.9a) and five MICA Stage 2 clusters (2.9b).

2.6 Conclusion

We have developed MICA, a methodological framework which combines multiple imputation and clustering algorithms while retaining all imputed data sets. MICA-N applied the MICA methodology to normally distributed clusters. As shown in Sections 2.3, 2.4 and 2.5, MICA-N outperformed CCA in simulation studies and uncovered more nuanced patterns in a data application.

While other methodologies exist to cluster incomplete data, our methodology is the first to our knowledge that clusters multiply imputed data without pre-specifying a particular number of clusters to use.

Future comparisons are planned involving a maximum likelihood based approach to obtaining clustering model parameters and cluster membership. In addition, the current method is limited by being able to address normally distributed clusters: the imputation algorithm Norm is designed for normally distributed variables; and while the clustering algorithm Mclust is reasonably robust, it is nevertheless tailored to normal mixture models. While there exist methods to identify mixtures of skewed elliptical distributions in a data set, there is currently no commercial software which generates imputed values from such a distribution and clusters the data sets without the limitation to a single number of clusters. Developing such a model is a natural extension of our current method.

	Number of Clusters								
Number of Clusters	3	4	5	6	7	8	9	10	11
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.91	0.65	0.51	0.37	0.22	0.17	0.17	0.17	0.15
3	1.00	0.73	0.57	0.42	0.26	0.20	0.20	0.20	0.18
4		1.00	**0.81**	0.61	0.39	0.31	0.30	0.30	0.27
5			1.00	0.76	0.50	0.41	0.36	0.36	0.32
6				1.00	0.70	0.60	0.53	0.53	0.46
7					1.00	0.77	0.69	0.69	0.56
8						1.00	0.90	0.90	0.71
9							1.00	1.00	0.80
10								1.00	0.80

Table 2.1: Adjusted Rand Indices of Stage 2 clustering solutions from one to eleven clusters. There is moderate agreement between solutions which use four and five clusters (0.81), but not enough to sway us away from the five-cluster solution.

Chapter 3

The Impact of Missing Values on Measures of Uncertainty

3.1 Introduction

In this chapter, we develop various results concerning the effect of incomplete data on measures of uncertainty. Specifically, we look at how the Shannon entropy (Shannon, 1948; Cover and Thomas, 2006) of jointly normal data is affected by different patterns of missing values. The relationship between the fraction of missing information (FMI) (Rubin, 1987) and entropy is also studied to see whether connections exist between these two measures of variability.

Entropy is a measure of uncertainty in a stochastic system. Consider a data set, with a fixed set of realized values. Since the data values are held fixed, there is no randomness or uncertainty in the data itself; fixed data values do not change. However, let us assume that we may describe the data set as one realization of a distribution, e.g. a random sample from jointly bivariate normal variables. There is uncertainty in the random variables, which may be described using entropy.

It makes intuitive sense that missing values should increase the uncertainty

in the models describing the data by adding another random process; namely, the missingness mechanism. However, to our knowledge the impact of missing values on entropy has not been addressed. In addition, the missing mechanism may often be ignored; for example, when maximum likelihood estimates are of interest (Little and Rubin, 2002; Schafer, 1999). However, it is our belief that the random process which governs missing values must be taken into account when considering entropy, even under conditions where it can otherwise be ignored.

We choose to handle missing values using multiple imputation (MI) (Harel and Zhou, 2007; Rubin, 1987; Schafer, 1999). One bi-product of MI is the fraction of missing information (FMI) (Rubin, 1987), which is a measurement of how much information the parameter estimate is missing by having imputed values instead of observed values. As both FMI and entropy address uncertainty, albeit from different sources, it is of interest to see how their behaviors compare to each other over varying levels of missingness. As FMI is a product of MI, and thus subject to cases where parameters of interest are normally distributed, being able to capture the same information with the more flexible, more widely applicable entropy calculation would be a boon for the research field.

This chapter is constructed in the following manner. Section 3.2 contains our derivation of entropy of incomplete bivariate normal data under MCAR missingness, while Section 3.3 contains the same for the MAR case. Sections 3.4 and 3.5 discuss the extension of the bivariate case to the p-variate case for MCAR and

MAR respectively. Section 3.6 demonstrates our simulation results. Section 3.7 discusses our conclusions.

3.2 Bivariate MCAR Entropy

We begin by developing formulae for the entropy of bivariate normal data with MCAR missingness. Our data is a matrix $\mathbf{Y}_{n\times 2}$, which has two variables, $\mathbf{y_1}$ and $\mathbf{y_2}$, where $\mathbf{y_i} = (y_{i1}, ..., y_{in})$. Suppose the variables $\mathbf{y_1}$ and $\mathbf{y_2}$ follow a joint normal $N_2(\theta)$, where $\theta = (\mu, \Sigma)$. We impose missing values on $\mathbf{y_2}$, such that the first n_1 values of $\mathbf{y_2}$ are observed, and the remaining $n_1 + 1$ to n values are missing.

The missingness mechanism is represented as a matrix $\mathbf{R}_{n\times 2} = (\mathbf{r_1}, \mathbf{r_2})$, where $\mathbf{r_i} = (r_{i1}, ..., r_{in})$. Recall that MCAR imposes missing values based on "pure" randomness; i.e., the probability of an observation being missing depends only on a parameter, and not upon any observed or unobserved data. Since missing values are found only in $\mathbf{y_2}$, only the corresponding column of $\mathbf{R}$, namely $\mathbf{r_2}$, will be random. Let $\mathbf{r_2} \sim Bernoulli(\phi)$, where ϕ is a parameter which does not depend on observed or missing data. Since the values of $\mathbf{y_1}$ are completely observed, $\mathbf{r_1}$ will only contain the value 1. Since the entropy of a constant is zero, we do not include $\mathbf{r_1}$ in our calculations. Let the matrix of data and missingness be written $\mathbf{X} = (\mathbf{Y}, \mathbf{r_2}) = (\mathbf{y_1}, \mathbf{y_2}, \mathbf{r_2})$.

When calculating entropy, we consider component-wise entropy, which focuses on the entropy of one record at a time. In other words, we consider

$H(\mathbf{x_i}) = H(y_{1i}, y_{2i}, r_{2i})$, where $H(A)$ is the entropy of A. For *iid* records, such as under MCAR, the entropy of the model which describes $\mathbf{X}$ is the multiplication of the entropy of the model which describes the first record $\mathbf{x_1} = (y_{11}, y_{21}, r_{21})$ and the sample size, n. This brings us to Theorem 3.2.1.

Theorem 3.2.1. *For bivariate normal data* $\mathbf{Y}$ *with MCAR missingness Bernoulli*(ϕ) *in* $\mathbf{y_2}$*, the entropy is*

$$H(\mathbf{X}) = -\frac{n}{2}ln(2\pi\sigma_1^2) + \frac{n}{2}ln(2\pi\sigma_2^2(1-\rho^2)) - n(1-\phi)ln(1-\phi) - n\phi ln(\phi), \quad (3.1)$$

where ρ *is the correlation between* $\mathbf{y}_1$ *and* $\mathbf{y}_2$*. The proof is given in Appendix.*

Note that $\mathbf{X}$ represents incomplete data, and can be written $\mathbf{Y_{inc}}$. Following this change in notation, Equation 3.1 can be rewritten as

$$H(\mathbf{Y_{inc}}) = H(\mathbf{Y_{com}}) + H(\mathbf{r_2}),$$

where $H(\mathbf{Y_{inc}})$ is the entropy of the model of incomplete data, $H(\mathbf{Y_{com}})$ is the entropy of the model of completely observed data, and $H(\mathbf{r_2})$ is the entropy of the missingness mechanism for the incomplete variable. This form gives us the following relationship,

$$H(\mathbf{Y_{com}}) = H(\mathbf{Y_{inc}}) - H(\mathbf{r_2}),$$

which suggests the entropy of the model of completely observed data can be estimated by observing the entropy of the model of incomplete data and subtracting the entropy from the missingness mechanism, provided that the distribution of the MCAR mechanism is known.

3.2.1 Limiting Behavior of Bivariate MCAR Incomplete Entropy

We are interested to see the limiting behavior of Equation 3.1 as the percent of missing values approaches zero. This is equivalent to showing that the entropy of $\mathbf{r_2}$ converges to zero as the percent of missing values goes to zero.

Theorem 3.2.2. *For bivariate normal data* $\mathbf{Y}$ *with MCAR missingness* $\mathbf{r_2} \sim$ *Bernoulli*(ϕ) *in* $\mathbf{y_2}$,

$$\lim_{\phi \to 0} \Big(-\frac{n}{2} ln(2\pi\sigma_1^2) + \frac{n}{2} ln(2\pi\sigma_2^2(1-\rho^2)) - n(1-\phi)ln(1-\phi) - n\phi ln(\phi) \Big) =$$
$$-\frac{n}{2} ln(2\pi\sigma_1^2) + \frac{n}{2} ln(2\pi\sigma_2^2(1-\rho^2)).$$

The proof is given in the Appendix.

Theorem 3.2.2 tells us that entropy of the incomplete data model converges to the entropy of the complete data model as the percent of missing values approaches zero. In addition, since $H(\mathbf{r_2}) = -n(1-\phi)ln(1-\phi) - n\phi ln(\phi)$ is symmetric, and Theorem 3.2.2 states that $\lim_{\phi \to 0} H(\mathbf{r_2}) = 0$, then we also know that $\lim_{\phi \to 1} H(\mathbf{r_2}) = 0$.

To relate this to our data, keep in mind that this result discusses only the entropy of the missingness mechanism $\mathbf{r_2}$, separate from the entropy of the model for the data. As the percent of missing values approaches zero or one, the vector $\mathbf{r_2}$ will have nearly all zeroes (signifying that nearly all values are missing) or nearly all ones (signifying that nearly all values are observed), and the entropy of a constant (or vector of identical constants) is zero.

To put this result is useful terms, consider each case individually. If the percent of missing values approaches zero, all of $\mathbf{Y}$ is observed, and the entropy of the model for $\mathbf{Y}$ will ignore the missingness mechanism. If the percent of missing values approaches one, the number of values of $\mathbf{y_2}$ shrinks, until finally $\mathbf{y_2}$ is not observed ($\phi = 1$). In that case, the complete data becomes only $\mathbf{y_1}$, and the relationships described previously still hold.

3.3 Bivariate MAR Entropy

We extend our MCAR work to the MAR case. MAR missingness imposes missing values based on observed data. In our case, MAR missingness is imposed on $\mathbf{y_2}$ based on the values of $\mathbf{y_1}$. For each record, we set $r_{2i} \sim f(r_{2i}|y_{1i}) = Bernoulli(\phi^*)$, where $\phi^* = \dfrac{e^{\beta_0+y_{1i}}}{1+e^{\beta_0+y_{1i}}}$ and β_0 is a parameter. In this way, the probability of being observed increases with increasing values of y_{1i}. One can also introduce a coefficient for y_{1i}, β_1, to control the direction and strength of the effect $\mathbf{y_1}$ has on $\mathbf{r_2}$. Recall that our data is the combination of random variables and missingness

mechanism, $\mathbf{X} = (\mathbf{Y}, \mathbf{R})$. The result is the following theorem.

Theorem 3.3.1. *For bivariate normal data* $\mathbf{Y}$ *with MAR Bernoulli*(ϕ^*) *missingness, where* $\phi^* = \dfrac{e^{\beta_0 + y_{1i}}}{1 + e^{\beta_0 + y_{1i}}}$, *the entropy is*

$$H(\mathbf{X}) = \frac{n}{2} ln(2\pi e\sigma_1^2) + \frac{n}{2} ln(2\pi e\sigma_2^2(1-\rho^2)) - \sum_{i=1}^{n} \{(1-\phi_i^*) ln(1-\phi_i^*)\} - \sum_{i=1}^{n} \{\phi_i^* ln(\phi_i^*)\} \tag{3.2}$$

The proof is given in the Appendix.

Note that, again, $\mathbf{X}$ represents the incomplete data, and can be written $\mathbf{Y_{inc}}$. Therefore, Equation 3.2 can be rewritten as

$$H(\mathbf{Y_{inc}}) = H(\mathbf{Y_{com}}) + H(\mathbf{r_2})$$

$$= H(\mathbf{Y_{com}}) = H(\mathbf{Y_{inc}}) - H(\mathbf{r_2}),$$

which suggests entropy of the complete data model can be estimated from the entropy of the incomplete data model and the entropy of the missingness mechanism, provided that the distribution of the MAR mechanism is known.

3.3.1 Limiting Behavior of Bivariate MAR Incomplete Entropy Estimate

As before, we examine the limiting behavior of Equation 3.2 as the percent of missingness goes to zero.

Theorem 3.3.2. *For bivariate normal data* $\mathbf{Y}$ *with MAR missingness* $\mathbf{r_2} \sim$ *Bernoulli*(ϕ^*), *where* $\phi^* = \dfrac{e^{\beta_0+y_{1i}}}{1+e^{\beta_0+y_{1i}}}$,

$$\lim_{\phi_i^* \to 0} \left(\frac{n}{2} ln(2\pi e\sigma_1^2) + \frac{n}{2} ln(2\pi e\sigma_2^2(1-\rho^2)) - \sum_{i=1}^{n} \{(1-\phi_i^*) ln(1-\phi_i^*)\} - \sum_{i=1}^{n} \{\phi_i^* ln(\phi_i^*)\} \right) =$$

$$\frac{n}{2} ln(2\pi e\sigma_1^2) + \frac{n}{2} ln(2\pi e\sigma_2^2(1-\rho^2)).$$

The proof is in the Appendix.

Theorem 3.3.2 tells us that our incomplete entropy estimator converges to the complete data estimator as the percent of missing values approaches zero. The interpretations of this result carry over from the MCAR case.

3.4 *p*-variate MCAR Entropy

In this section, we consider a matrix of incomplete data, $\mathbf{Y_{nxp}} = (\mathbf{y_1}, ..., \mathbf{y_p})$. We examine so-called "block" missingness. Namely, the first k variables are completely observed; records 1 to n_1 of variables $\mathbf{y_{k+1}}, ..., \mathbf{y_p}$ are observed; and only values in records n_1+1 to n of variables $\mathbf{y_{k+1}}, ..., \mathbf{y_p}$ are missing, where $1 < k < p$. As a result, the corresponding matrix $\mathbf{R_{nxp}} = (\mathbf{r_1}, ..., \mathbf{r_p})$ will have columns of varying values only in $\mathbf{r_k}, ..., \mathbf{r_p}$. Since missing values occur in a block structure, we simplify notation by writing $\mathbf{Y} = (\mathbf{Y_A}, \mathbf{Y_B})$, where $\mathbf{Y_A} = (\mathbf{y_1}, ..., \mathbf{y_k})$ and $\mathbf{Y_B} = (\mathbf{y_{k+1}}, ..., \mathbf{y_p})$. Consequently, we also write $\mathbf{R} = (\mathbf{R_A}, \mathbf{R_B})$, where $\mathbf{R_A} = (\mathbf{r_1}, ..., \mathbf{r_k})$ and $\mathbf{R_B} = (\mathbf{r_{k+1}}, ..., \mathbf{r_p})$. Note that the vectors $\mathbf{r_{k+1}}$ to $\mathbf{r_p}$ are

identical, in order to reflect the "block" of missing values in the data. Our data is therefore $\mathbf{X} = (\mathbf{Y_A}, \mathbf{Y_B}, \mathbf{R_A}, \mathbf{R_B})$. We may exclude $\mathbf{R_A}$ from our entropy calculations, as all the values are equal to 1, and therefore have no variation. We again calculate entropy in a component-wise manner, focusing only on the i^{th} record $(\mathbf{y_{Ai}}, \mathbf{y_{Bi}}, \mathbf{r_{Bi}})$ before considering all records together.

Theorem 3.4.1. *The entropy of p-variate normal data* $\mathbf{Y_{nxp}} = (\mathbf{y_1}, ..., \mathbf{y_p})$ *with block MCAR missingness is*

$$H(\mathbf{X}) = \frac{nk}{2} ln(2\pi e) + \frac{1}{2} ln(|\Sigma_1|) + \frac{n(p-k)}{2} ln(2\pi e) + \frac{1}{2} ln(|\Sigma_{2|1}|) + nH(\mathbf{r_{B,i}}). \quad (3.3)$$

The proof is given in the Appendix.

3.5 *p*-variate MAR Entropy

The setup of the data in the p-variate MAR case, including the "block" missingness, is identical to the MCAR case. The only difference is in the missingness mechanism. In the MAR case, missing values are generated for one variable using Bernoulli(ϕ^*) missingness, where $\phi^* = \dfrac{e^{\beta_0 + y_{1i}}}{1 + e^{\beta_0 + y_{1i}}}$, and the other missingness patterns are identical to the first in order to create "block" missingness. The result is given below.

Theorem 3.5.1. *The entropy of p-variate normal data* $\mathbf{Y_{nxp}} = (\mathbf{y_1}, ..., \mathbf{y_p})$ *with block MAR missingness is*

$$H(\mathbf{X}) = nH(\mathbf{y}_{A,i}) + nH(\mathbf{y}_{B,i}|\mathbf{y}_{A,i}) + \sum_{i=1}^{n} H(\mathbf{r}_{B,i}|\mathbf{y}_{A,i,j}), \tag{3.4}$$

where $H(\mathbf{X})$ *is the entropy of* $\mathbf{X} = (\mathbf{Y_A}, \mathbf{Y_B}, \mathbf{R_A}, \mathbf{R_B})$*;* $H(\mathbf{y}_{A,i})$ *is the entropy of the* i^{th} *record of* $\mathbf{y_A}$*;* $H(\mathbf{y}_{B,i}|\mathbf{y}_{A,i})$ *is the entropy of the* i^{th} *record of* $\mathbf{y_A}|\mathbf{y_B}$*;* $H(\mathbf{r}_{B,i}|\mathbf{y}_{A,i,j})$ *is the entropy of the* i^{th} *record of* $H(\mathbf{r}_B|\mathbf{y}_{A_j})$*, where* j *is the column of* $\mathbf{y}_A$ *which determined the missingness; and* n *is the sample size.*

The proof is given in the Appendix.

3.6 Simulations

There are two sets of simulation studies: one for bivariate data with MCAR missingness, and the other for bivariate data with MAR missingness. Within each study, there are two research goals. The first goal is to compare the entropies of models for fully-observed and complete case with the newly-derived entropy of the model for incomplete data. Comparisons are made in terms of estimates, biases, and standard errors. The second goal is to compare the entropy of the model for incomplete data to the fraction of missing information obtained by (i) estimating the mean of the incomplete variable $\mathbf{y_2}$, and (ii) estimating the regression coefficients from regressing $\mathbf{y_2}$ on $\mathbf{y_1}$. Comparisons are made graphically and via correlation statistics.

Before we combine the incomplete entropy point estimates into their single point estimate using Rubin's Rules, we must address whether the assumptions of

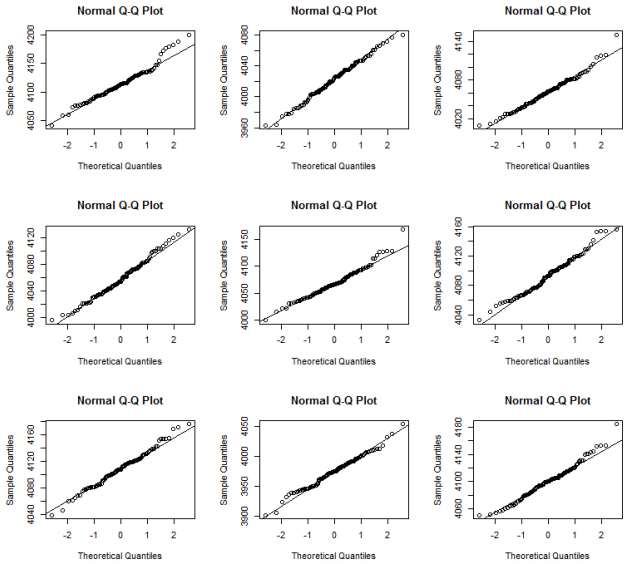


Fig. 3.1: 100 estimates of incomplete MCAR entropy, plotted on a normal QQ plot.

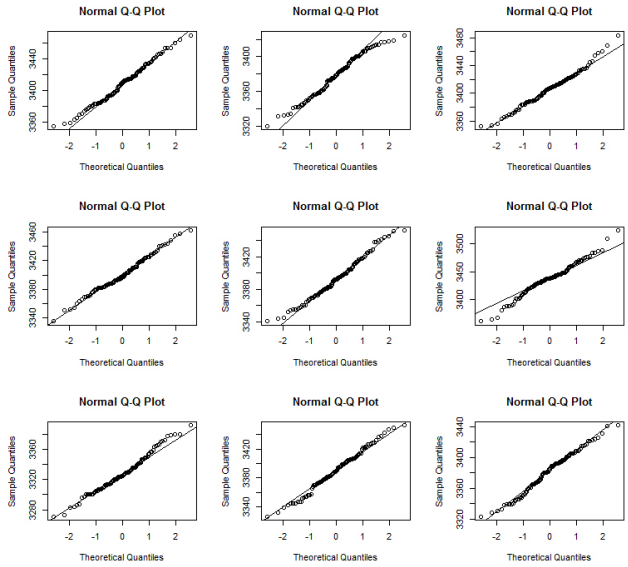


Fig. 3.2: 100 estimates of incomplete MAR entropy, plotted on a normal QQ plot.

Rubin's Rules allow us to combine the point estimates. Namely, we must determine whether the parameter of interest is normally distributed. To determine this, we looked at nine different runs of the MCAR (Figure 3.1) and MAR (Figure 3.2) simulations. From these figures, we can see that the entropy values are reasonably close to the 45-degree line in the normal QQ plot, which we take as evidence that we may apply Rubin's combining rules without using a transformation of the point estimates.

3.6.1 MCAR Entropy

For each of the 250 repetitions, a thousand observations of $(\mathbf{y_1}, \mathbf{y_2})$ were generated from $N_2\left(\mu = \begin{pmatrix} 0 \\ 5 \end{pmatrix}, \Sigma = \begin{pmatrix} 1 & .5 \\ .5 & 3 \end{pmatrix}\right)$ using the R package MASS (Venables and Ripley, 2002). The corresponding $\mathbf{r_2}$ values were simulated from $Bernoulli(\phi)$, in such a way to calculate 50%, 25%, 10%, 5%, 2.5%, and 1% missing values. The purpose of the small rates of missing observations (e.g. 2.5% and 1%) is to allow us to describe the behavior of the estimator of the incomplete data model as the percent of missing values approaches zero. This will serve as a complement to our mathematical derivations in Section 3.2.

3.6.2 How to Estimate MCAR Entropy

Some thought must be given to how one can estimate the entropy of models which describe fully observed, complete-case, and imputed data. For example,

while Equation 3.1 includes the entropy of the missingness mechanism, should it be included for the fully observed and complete-data cases? In this section, we give the estimators for entropy of fully observed, complete case, and imputed data models, and the reasoning behind the inclusion or exclusion of the entropy of the missingness mechanism.

Fully observed data, by definition, has no incomplete records. In this case, there is no need to estimate the entropy of the missingness mechanism. The entropy of the fully observed normal data model is therefore estimated by

$$\hat{H}(\mathbf{Y_f}) = \frac{n_f}{2} ln(2\pi e s_{1f}^2) + \frac{n_f}{2} ln(2\pi e s_{2f}^2 (1 - r_f^2)),$$

where σ and ρ have been estimated with s = sample standard deviation and r = sample correlation, respectively. The subscript f denotes estimates come from the fully observed data. Note that $n_f = n$, the original sample size.

The complete case analysis also involves records with no missing values. The intent of complete case analysis is to pretend there was no missingness, therefore we again do not estimate the entropy of the missingness mechanism. The complete case model's entropy is estimated by

$$\hat{H}(\mathbf{Y_{cca}}) = \frac{n_{cca}}{2} ln(2\pi e s_{1,cca}^2) + \frac{n_{cca}}{2} ln(2\pi e s_{2,cca}^2 (1 - r_{cca}^2)),$$

where the subscript *cca* denotes estimates come from the complete case data. Note that n_{cca} is the sample size of the complete case data, which will be smaller than n.

Estimating the entropy of incomplete data involves multiple steps. The process begins by imputing the missing data. As we are using bivariate normal data, we impute using R package *norm* (Schafer, 2008). Imputation supplies different complete data sets. The entropy of the model for each data set is obtained by

$$\hat{H}(\mathbf{Y_{m,j}}) = \frac{n}{2} ln(2\pi e s^2_{1,m,j}) + \frac{n}{2} ln(2\pi e s^2_{2,m,j}(1 - r^2_{m,j}))- \\ n(1 - p_{m,j}) ln(1 - p_{m,j}) - n p_{m,j} ln(p_{m,j}), \tag{3.5}$$

where the subscript m denotes estimates come from imputed data, and subscript j denotes estimates come from the j^{th} imputed data. Note that $n_{m,j} = n_f = n$, the original sample size. By applying Rubin's rules, we obtain the final estimate of entropy, $\hat{H}(\mathbf{Y_m})$

Results

Table 3.1 shows the entropy of models for fully observed, complete case, and imputed data, the FMI for estimating μ_2 (λ_{μ_2}), the FMI for estimating the beta coefficients for regressing $\mathbf{y_2}$ on $\mathbf{y_1}$ (λ_{β_0} and λ_{β_1}, respectively), and the differences and ratios between $\hat{H}(\mathbf{Y_m})$ and $\hat{H}(\mathbf{Y_f})$ and between $\hat{H}(\mathbf{Y_{cca}})$ and $\hat{H}(\mathbf{Y_f})$. For each value of Percent Missing ($\%Mis$), the first row is the average of all 250 values, and the second row is the standard deviation of those values.

The entropy estimate for imputed data slightly overestimates the entropy of the full data model (see $\Delta(m, f)$). However, it is at most 20% higher than the

entropy of the full data model (see $\hat{H}(\mathbf{Y_m})/\hat{H}(\mathbf{Y_f})$). The CCA estimate drastically underestimates the entropy of the full data model (see $\Delta(cca, f)$), and is at least half of the entropy of the full data model (see $\hat{H}(\mathbf{Y_{cca}})/\hat{H}(\mathbf{Y_f})$). This may be due to the large difference in sample sizes between the full data and complete-case data. Also, the ratio $\hat{H}(\mathbf{Y_{cca}})/\hat{H}(\mathbf{Y_f})$ is always approximately the percent of missing values.

A graphical comparison of entropy estimates for imputed, complete case, and fully observed data models is given in Figure 3.3. Let us examine the graphic in detail. The horizontal red line represents the theoretical value of fully observed bivariate normal entropy under the parameters we have specified. The black line represents the simulated estimate of that quantity. As we would expect, the estimate hovers around the theoretical value. The blue line represents the values of our new estimator for entropy of the incomplete data model. The brown line represents the CCA estimate. Bars around the point estimates indicate $\pm$ one standard deviation. The green dashed line is the length of the quantity $\hat{H}(\mathbf{r_2})$, the entropy of the missingness mechanism.

$\%Mis$	$\hat{H}(\mathbf{Y_f})$	$\hat{H}(\mathbf{Y_m})$	$\hat{H}(\mathbf{Y_{cca}})$	λ_{μ_2}	λ_{β_0}	λ_{β_1}	$\Delta(m,f)$	$\Delta(cca,f)$	$\frac{\hat{H}(\mathbf{Y_{cca}})}{\hat{H}(\mathbf{Y_f})}$	$\frac{\hat{H}(\mathbf{Y_m})}{\hat{H}(\mathbf{Y_f})}$
50	3343.62	4067.89	1666.61	0.54	0.54	0.32	724.27	-1677.01	0.50	1.22
	30.73	38.07	57.29	0.04	0.04	0.03	23.66	57.44	0.02	0.01
25	3342.17	3922.31	2504.03	0.33	0.34	0.20	580.15	-838.14	0.75	1.17
	29.84	35.92	53.33	0.03	0.03	0.03	21.38	48.15	0.01	0.01
10	3340.88	3672.64	3008.40	0.16	0.17	0.09	331.76	-332.48	0.90	1.10
	28.95	39.26	38.22	0.03	0.03	0.02	22.86	32.95	0.01	0.01
5	3340.94	3542.86	3175.64	0.08	0.09	0.05	201.92	-165.30	0.95	1.06
	32.04	38.66	36.06	0.02	0.02	0.01	20.19	21.72	0.01	0.01
1	3344.74	3401.74	3310.66	0.02	0.02	0.01	56.99	-34.08	0.99	1.02
	33.19	35.76	34.73	0.01	0.01	0.01	14.56	10.83	0.00	0.00

Table 3.1: Bivariate MCAR results. Imputations: 100. $\Delta(m,f) = H(\mathbf{Y_m}) - H(\mathbf{Y_f})$. $\Delta(cca,f) = H(\mathbf{Y_{cca}}) - H(\mathbf{Y_f})$. Standard errors are in parentheses.

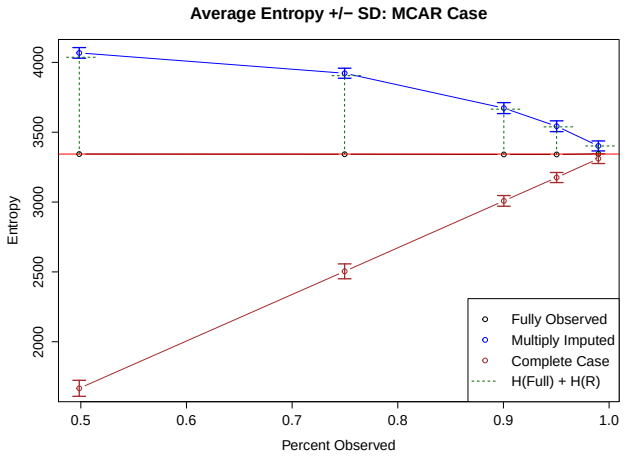


Fig. 3.3: MCAR case. Red line: theoretical value of entropy of fully observed data model. Black line: estimated entropy of fully observed data model. Blue line: estimated entropy of incomplete data model. Brown line: estimated entropy of CCA data model. Green dashed line: length represents $\hat{H}(\mathbf{r_2})$. Bars around point estimates: $\pm$ one standard deviation.

The entropy of the CCA data model is increasingly smaller than the entropy of the full data model as more and more data is missing. This result is expected, since more missingness means a smaller CCA data set, and thus less additive entropy. The bars around entropy of CCA data are also wider than the bars around the entropy of incomplete data. Two findings arise from this. First, the CCA estimates are less reliable. Second, the CCA estimate decreases when common sense (and our new formulae) dictate that it should increase, making it a misleading estimate for entropy of fully observed data.

Figure 3.4 compares the FMI for estimating μ_2 and the FMI for estimating

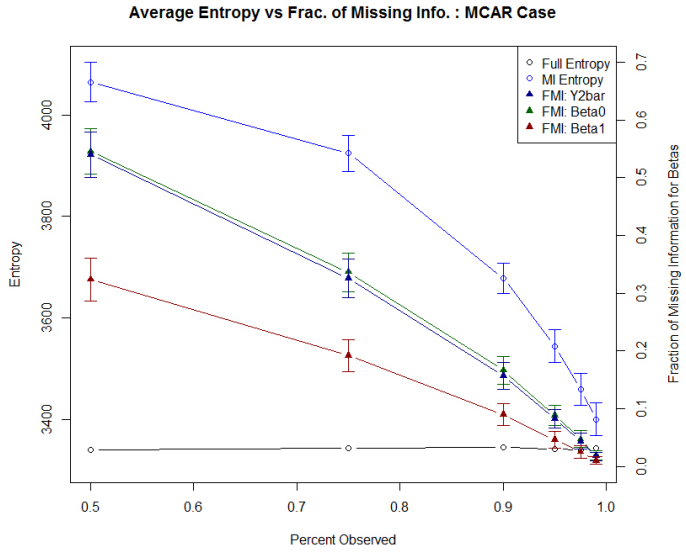


Fig. 3.4: MCAR case. Black line: estimated entropy of fully observed data model. Blue line: estimated entropy of incomplete data model. Dark blue line: FMI for estimating μ_2. Green and maroon lines: FMI for estimating β_0 and β_1 from regressing $\mathbf{y_2}$ on $\mathbf{y_1}$.

β_0 and β_1 from regressing $\mathbf{y_2}$ on $\mathbf{y_1}$ to the average entropy of imputed data models, while comparing both to the entropy of the fully observed data model. There is the suggestion of a relationship between the imputed estimate of entropy and the fraction of missing information. There is also the suggestion of a relationship between the imputed estimate of entropy and the fraction of missing information for estimating β_0, with possible correlation to the β_1 case as well.

To quantify the relationship uncovered by Figure 3.4, we calculate point estimates and p-values for Pearson correlation, Spearman correlation, and Kendall's tau via the *stats* R package. The correlations are calculated for entropy of the

data model with FMI, as well as the components of the formula for FMI, B and $\bar{U}$; and similarly for the entropy of the missingness mechanism.

Correlation records involving the entropy of the data model only are presented in Table 3.2, while those for the entropy of the missingness mechanism are in Table 3.3. The first, second, and third sections of the table contain point statistics and p-values for Pearson, Spearman, and Kendall statistics respectively. Within each section are results for 50, 25, 10, and 5% missingness. FMI, B, and $\bar{U}$ statistics are shown for estimating both μ_2 and β_1, and differentiated by their subscripts.

FMI for estimating both parameters tends not to be significantly associated with entropy of the data model ($\alpha = 0.05$). However, B and $\bar{U}$ are always significantly associated with entropy of the data model when estimating μ_2, and often associated with that quantity when estimating β_1. This pattern holds true for entropy of the missingness mechanism as well.

3.6.3 MAR Case

Recall that we are operating under MAR Bernoulli(ϕ^*) missingness, where $\phi^* = \frac{e^{\beta_0+y_{1i}}}{1+e^{\beta_0+y_{1i}}}$. Including the MAR mechanism makes our *iid* sample of $(\mathbf{y_1}, \mathbf{y_2})$ change to an independent but not identically distributed sample of $(\mathbf{y_1}, \mathbf{y_2}, \mathbf{r_2})$. Values of β_0 were such that the average percent missing in each case was about 50%, 30%, 7%, 5%, 2%, and 0.5%.

How to Estimate MAR Entropy

Estimating the entropy of incomplete data follows the same basic steps as the MCAR case: imputing the missing data, calculate entropy, and obtain the average. However, since the records are not *iid*, entropy must be estimated using a different formula:

$$\hat{H}(\mathbf{Y_{m,j}}) = \frac{n}{2} ln(2\pi e s^2_{1,m,j}) + \frac{n}{2} ln(2\pi e s^2_{2,m,j}(1 - r^2_{m,j})) -$$

$$\sum_{i=1}^{n} \{(1 - p^*_{m,j,i}) ln(1 - p^*_{m,j,i})\} - \sum_{i=1}^{n} \{p^*_{m,j,i} ln(p^*_{m,j,i}))\}, \tag{3.6}$$

where the subscript m denotes estimates come from imputed data, the subscript j denotes an estimate from the j^{th} imputed data set, and subscript i denotes an estimate from the i^{th} record. The individual point estimates are then averaged to obtain the final estimate of entropy, $\hat{H}(\mathbf{Y_m})$.

Results

Figure 3.5 shows the same information as Figure 3.3, now for the MAR missingness case. As before, entropy of the incomplete data model is larger than that of the fully observed model. Also, the entropy of the CCA data model is increasingly less than the entropy of the fully observed data model, and the bars around the CCA estimates are larger than the bars around the incomplete data estimates.

Table 3.2: MCAR - Correlations involving Entropy of the Data Model Only.

Corr.	$\%Mis.$	FMI_{μ_2}	B_{μ_2}	$\bar{U}\mu_2$	FMI_{β_1}	B_{β_1}	$\bar{U}_{\beta_1}$
Pearson	50	-0.05	0.27	0.82	0.12	0.25	0.35
p-value		0.42	0.00	0.00	0.06	0.00	0.00
	25	0.09	0.32	0.77	0.07	0.11	0.12
		0.15	0.00	0.00	0.24	0.07	0.06
	10	0.02	0.20	0.73	0.14	0.16	0.13
		0.73	0.00	0.00	0.03	0.01	0.04
	5	0.12	0.26	0.73	0.04	0.05	0.04
		0.06	0.00	0.00	0.50	0.42	0.56
Spearman	50	-0.06	0.24	0.81	0.10	0.24	0.35
p-value		0.31	0.00	0.00	0.12	0.00	0.00
	25	0.10	0.33	0.76	0.06	0.09	0.14
		0.12	0.00	0.00	0.32	0.15	0.03
	10	0.02	0.20	0.74	0.15	0.19	0.13
		0.77	0.00	0.00	0.02	0.00	0.04
	5	0.12	0.26	0.72	0.06	0.07	0.04
		0.05	0.00	0.00	0.34	0.28	0.55
Kendall	50	-0.05	0.17	0.62	0.06	0.16	0.25
p-value		0.28	0.00	0.00	0.13	0.00	0.00
	25	0.07	0.22	0.57	0.04	0.07	0.09
		0.12	0.00	0.00	0.30	0.12	0.03
	10	0.01	0.13	0.54	0.10	0.13	0.09
		0.74	0.00	0.00	0.02	0.00	0.04
	5	0.08	0.18	0.53	0.04	0.04	0.02
		0.05	0.00	0.00	0.34	0.30	0.56

Table 3.3: MCAR Correlations with Entropy of the Missingness Mechanism.

Corr.	$\%Mis.$	FMI_{μ_2}	B_{μ_2}	$\bar{U}\mu_2$	FMI_{β_1}	B_{β_1}	$\bar{U}_{\beta_1}$
Pearson	50	-0.20	-0.20	-0.00	0.07	0.06	-0.02
p-value		0.00	0.00	0.98	0.27	0.33	0.79
	25	0.37	0.36	0.01	0.29	0.28	0.07
		0.15	0.00	0.00	0.24	0.07	0.06
	10	0.60	0.60	0.12	0.41	0.40	0.01
		0.73	0.00	0.00	0.03	0.01	0.04
	5	0.65	0.65	0.08	0.38	0.38	0.05
		0.06	0.00	0.00	0.50	0.42	0.56
Spearman	50	-0.13	-0.09	0.05	0.09	0.07	-0.01
p-value		0.04	0.18	0.47	0.14	0.24	0.88
	25	0.38	0.35	0.00	0.31	0.32	0.08
		0.12	0.00	0.00	0.32	0.15	0.03
	10	0.56	0.58	0.16	0.38	0.39	0.00
		0.77	0.00	0.00	0.02	0.00	0.04
	5	0.62	0.62	0.10	0.40	0.39	0.05
		0.05	0.00	0.00	0.34	0.28	0.55
Kendall	50	-0.09	-0.06	0.03	0.07	0.05	-0.01
p-value		0.04	0.20	0.50	0.13	0.25	0.90
	25	0.26	0.24	-0.00	0.21	0.22	0.05
		0.12	0.00	0.00	0.30	0.12	0.03
	10	0.40	0.41	0.11	0.27	0.27	0.00
		0.74	0.00	0.00	0.02	0.00	0.04
	5	0.45	0.46	0.07	0.28	0.27	0.04
		0.05	0.00	0.00	0.34	0.30	0.56

$PctMis$	$\hat{H}(\mathbf{Y_f})$	$\hat{H}(\mathbf{Y_m})$	$\hat{H}(\mathbf{Y_{cca}})$	λ_{μ_2}	λ_{β_0}	λ_{β_1}	$\Delta(m,f)$	$\Delta(cca,f)$	$\frac{\hat{H}(\mathbf{Y_{cca}})}{\hat{H}(\mathbf{Y_f})}$	$\frac{\hat{H}(\mathbf{Y_m})}{\hat{H}(\mathbf{Y_f})}$
50	3343.50	3369.03	1620.15	0.61	0.62	0.40	25.53	-1723.34	0.48	1.01
	(31.60)	(39.12)	(53.08)	(0.04)	(0.04)	(0.04)	(22.50)	(53.43)	(0.02)	(0.01)
30	3341.72	3443.38	2279.42	0.41	0.42	0.33	101.66	-1062.30	0.68	1.03
	(33.48)	(35.35)	(58.59)	(0.04)	(0.04)	(0.04)	(14.05)	(58.15)	(0.02)	(0.00)
7	3342.51	3442.54	3085.10	0.12	0.13	0.14	100.03	-257.42	0.92	1.03
	(29.98)	(30.87)	(39.10)	(0.02)	(0.02)	(0.03)	(6.72)	(29.93)	(0.01)	(0.00)
4	3342.61	3424.16	3173.37	0.08	0.08	0.10	81.55	-169.24	0.95	1.02
	(31.02)	(31.35)	(36.73)	(0.02)	(0.02)	(0.03)	(5.30)	(23.62)	(0.01)	(0.00)
2	3341.17	3388.55	3275.69	0.03	0.03	0.05	47.38	-65.48	0.98	1.01
	(28.94)	(29.24)	(32.18)	(0.01)	(0.01)	(0.02)	(3.11)	(17.42)	(0.01)	(0.00)
1	3340.46	3364.36	3316.36	0.01	0.01	0.02	23.89	-24.10	0.99	1.01
	(33.09)	(33.34)	(34.51)	(0.01)	(0.01)	(0.01)	(2.23)	(10.82)	(0.00)	(0.00)

Table 3.4: Bivariate MAR results. Imputations: 100. λ: FMI estimating μ_2. $\Delta(m,f) = H(\mathbf{Y_m}) - H(\mathbf{Y_f})$. $\Delta(cca,f) = H(\mathbf{Y_{cca}}) - H(\mathbf{Y_f})$. Standard errors are in parentheses.

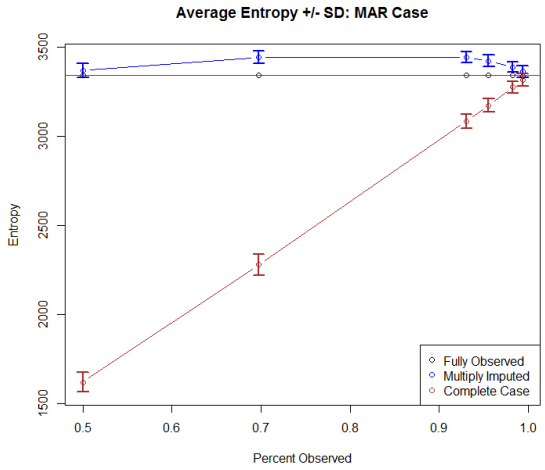


Fig. 3.5: MAR case. Red line: theoretical entropy of fully observed data model. Black line: estimated entropy of fully observed data model. Blue line: estimated entropy of incomplete data model. Brown line: estimated entropy of CCA data model. Bars around point estimates: $\pm$ one standard deviation.

For the MCAR case, the amount by which entropy of the incomplete data model increased over the entropy of the fully observed data model was almost exactly the value of $\hat{H}(\mathbf{r_2})$. Figure 3.6 shows a similar comparison for the MAR case. It is clear that the value of $\hat{H}(\mathbf{r_2}|\mathbf{y_1})$ exceeds the difference between $\hat{H}(\mathbf{Y_f})$ and $\hat{H}(\mathbf{Y_m})$. The reason may be because $\hat{H}(\mathbf{r_2}|\mathbf{y_1})$ takes into account the entropy of $\mathbf{y_1}$ in addition to the entropy of $\mathbf{r_2}$, since the values of $\mathbf{r_2}$ are determined by values of $\mathbf{y_1}$.

Figure 3.7 has the same information as Figure 3.4, this time for the MAR case. While the MCAR case showed a similar curved pattern between the MI

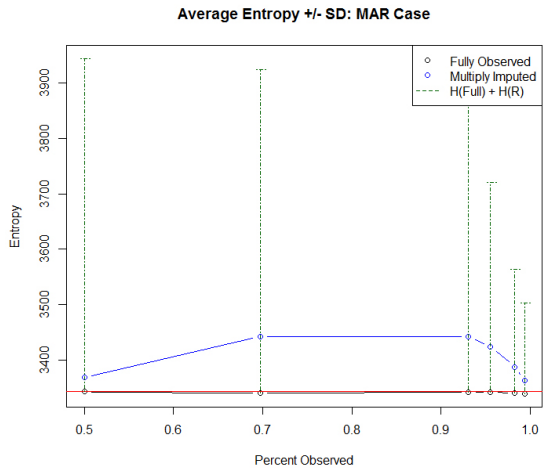


Fig. 3.6: MAR case. Red line: theoretical entropy of fully observed data model. Black line: estimated entropy of fully observed data model. Blue line: estimated entropy of incomplete data model. Green lines: length represents the entropy of $\mathbf{r_2}|\mathbf{y_1}$.

entropy estimates and the fraction of missing information, we do not detect a similar pattern here. This may be because, as previously mentioned, the value for the MI entropy estimate is taking the entropy of $\mathbf{y_1}$ into account.

We again quantify the relationship between FMI and entropy by calculating point estimates and p-values for different measures of association.

Correlation results for entropy of the data model are presented in Table 3.5, while those for the entropy of the missingness mechanism are in Table 3.6. The structures of the tables are similar to Tables 3.2 and 3.3. As in the MCAR case, FMI for estimating both parameters tends not to be significantly associated with entropy of the data model nor the entropy of the missingness mechanism. B and

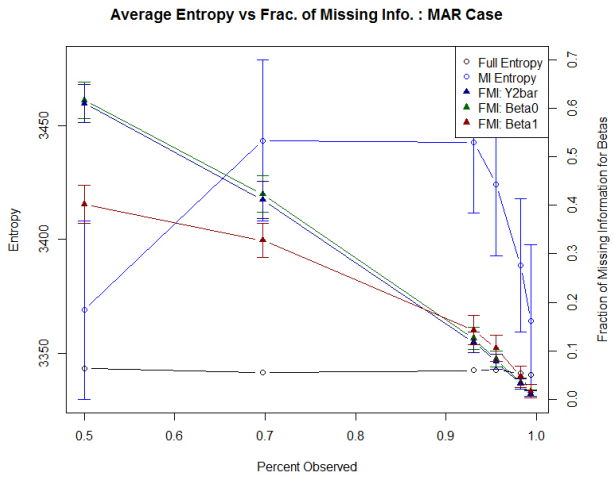


Fig. 3.7: MAR case. Black line: estimated entropy of fully observed data model. Dark blue line: FMI estimating $\bar{\mathbf{y}}_\mathbf{2}$. Blue line: estimated entropy of incomplete data model. Green and maroon lines: FMI for estimating β_0 and β_1. Bars around point estimates: $\pm$ one standard deviation.

$\bar{U}$ are significantly associated with entropy of the data model when estimating μ_2, and often associated with that quantity when estimating β_1. However, B and $\bar{U}$ are only infrequently significantly associated with the entropy of the missingness mechanism. The cause of the drop in significant association may be due to the nature of the missingness mechanism. The Bernoulli model which generated the MAR missing values used values of $\mathbf{y}_1$, so that the missingness mechanism still takes the data into account.

3.7 Conclusions

We have made the first of many strides into the realm of entropy of incomplete data. Our work has focused on the normal case, which allowed us to tackle unknown territory in a well-behaved setting.

Our theoretical work includes (a) the derivation of new theorems to describe entropy of the incomplete data model when missingness is generated by MCAR and MAR mechanisms; and (b) proofs that the new theorems for entropy of incomplete data models converge to their previously-known complete data counterparts when the percent of missing values go to zero.

From simulation studies we have also found that (c) that the entropy of the incomplete data model remains closer to the entropy of the fully observed data model than the CCA model; and (d) in the MCAR case, the entropy of the incomplete data model is larger than the entropy of fully observed data model by nearly the exact value of entropy in the missingness mechanism, and (e) while FMI is not significantly associated with incomplete entropy, the between- and within-imputation variances which make up the formula for FMI are often significantly associated with incomplete entropy.

Extensions of the work presented here include extending the formulae to other distributions, such as binomial and multinomial. In addition, teasing apart the entropy of the MAR missingness mechanism into Bernoulli-only and data-only components will shed further light on the reasons behind our results.

Table 3.5: MAR - Correlations with Entropy of the Data Model

Corr.	$\%Mis.$	FMI_{μ_2}	B_{μ_2}	$\bar{U}\mu_2$	FMI_{β_1}	B_{β_1}	$\bar{U}_{\beta_1}$
Pearson	50	0.12	0.41	0.85	0.09	0.21	0.30
p-value		0.07	0.00	0.00	0.16	0.00	0.00
	30	0.07	0.32	0.76	0.09	0.13	0.12
		0.27	0.00	0.00	0.17	0.04	0.06
	15	0.06	0.25	0.69	0.18	0.16	-0.10
		0.33	0.00	0.00	0.00	0.01	0.13
	5	0.08	0.23	0.73	0.07	0.07	-0.03
		0.18	0.00	0.00	0.27	0.29	0.58
Spearman	50	0.08	0.38	0.83	0.07	0.19	0.29
p-value		0.19	0.00	0.00	0.24	0.00	0.00
	30	0.05	0.22	0.66	0.14	0.11	-0.10
		0.41	0.00	0.00	0.02	0.07	0.13
	15	0.05	0.28	0.75	0.06	0.11	0.14
		0.44	0.00	0.00	0.31	0.09	0.03
	5	0.11	0.24	0.71	0.12	0.11	-0.04
		0.07	0.00	0.00	0.06	0.08	0.50
Kendall	50	0.06	0.26	0.64	0.05	0.12	0.20
p-value		0.17	0.00	0.00	0.28	0.00	0.00
	30	0.03	0.19	0.56	0.04	0.07	0.09
		0.46	0.00	0.00	0.31	0.09	0.03
	15	0.04	0.15	0.48	0.09	0.07	-0.06
		0.41	0.00	0.00	0.03	0.08	0.13
	5	0.08	0.16	0.51	0.08	0.08	-0.03
		0.07	0.00	0.00	0.06	0.07	0.49

Table 3.6: MAR - Correlations with Entropy of the Missingness Mechanism

Corr.	$\%Mis.$	FMI_{μ_2}	B_{μ_2}	$\bar{U}\mu_2$	FMI_{β_1}	B_{β_1}	$\bar{U}_{\beta_1}$
Pearson	50	-0.03	-0.11	-0.22	-0.06	0.15	0.48
p-value		0.61	0.09	0.00	0.36	0.02	0.00
	30	0.05	0.04	-0.04	-0.00	0.09	0.25
		0.47	0.54	0.51	0.94	0.15	0.00
	15	0.12	0.11	0.01	-0.05	-0.07	-0.05
		0.07	0.07	0.90	0.46	0.27	0.40
	5	0.10	0.11	0.03	0.10	0.06	-0.16
		0.11	0.09	0.61	0.12	0.35	0.01
Spearman	50	-0.01	-0.08	-0.19	-0.05	0.17	0.47
p-value		0.84	0.19	0.00	0.42	0.01	0.00
	30	0.00	0.01	-0.04	-0.02	0.08	0.21
		0.95	0.93	0.48	0.77	0.21	0.00
	15	0.10	0.09	-0.02	-0.05	-0.06	-0.06
		0.11	0.15	0.79	0.46	0.38	0.36
	5	0.10	0.11	0.05	0.13	0.11	-0.13
		0.10	0.09	0.41	0.05	0.09	0.04
Kendall	50	-0.01	-0.06	-0.13	-0.03	0.11	0.33
p-value		0.80	0.19	0.00	0.42	0.01	0.00
	30	0.00	0.00	-0.03	-0.01	0.05	0.14
		0.99	0.96	0.45	0.79	0.21	0.00
	15	0.07	0.06	-0.01	-0.03	-0.04	-0.04
		0.12	0.16	0.82	0.46	0.36	0.31
	5	0.07	0.07	0.03	0.08	0.07	-0.09
		0.09	0.09	0.45	0.05	0.08	0.04

Chapter 4

Latent Class Analysis of Incomplete Data via an Entropy-Based Criterion

4.1 Introduction

Latent class analysis (LCA) (Hagenaars and McCutcheon, 2002) is a model-based clustering methodology for categorical data. Variables in a data set are "manifest" variables, while the unknown vector of class membership is the "latent" variable. LCA breaks the data into classes (e.g., clusters) via two parameters: latent class probabilities and conditional probabilities. The former dictates how likely it is that a record belongs to each class, while the latter describes the probability of a particular variable having a particular value given that it is in a certain class. LCA assumes that the relationships between manifest variables are accounted for by their class membership. Thus, conditioning on class membership makes manifest variables independent.

In this chapter, we prove that the entropy of an LCA model decreases to zero as the number of classes increases to the number of unique records. We then

use this knowledge to develop a new model selection criterion in order to utilize methods for clustering incomplete categorical data using MI without having to limit ourselves to a single number of clusters.

There are methods for clustering categorical data using entropy (Barbara et al., 2002; Li et al., 2004), but they do not address incomplete data. There is also an entropy-based criterion for mixture model data, but it was applied to complete, normal mixture model data (Celeux and Soromenho, 1996), but this methodology requires specification of a single number of clusters prior to imputing the data. There are also ways to cluster incomplete categorical data using multiple imputation and latent class analysis (LCA) (Harel et al., 2013) of the fraction of missing information (FMI) as an LCA model selection criterion (Harel and Miglioretti, 2007), which addresses the case where class membership is the only missing component; in other words, it addresses the complete data case. We propose a new, entropy-based model selection criterion method for the case where the manifest variables are incomplete and class membership is unknown, which will allow the use of LCA with multiple imputation without having to set a number of clusters beforehand.

Entropy (Celeux and Soromenho, 1996; Shannon, 1948; Cover and Thomas, 2006) has been used as a model selection criterion. Typically, it is combined with the log-likelihood (Biernacki and Govaert, 1997; Biernacki et al., 2000), although it has been used on its own (Li et al., 2004; Barbara et al., 2002). We are interested

in looking at entropy itself as a model selection criterion.

To begin, we prove that the entropy of an LCA model with G classes goes to zero as G goes to the number of unique records in the data. Fruhwirth-Schnatter (2006) describes entropy of a mixture model as equaling zero if each record belongs to its cluster with probability one. Realistically, this is not likely to happen unless every record has its own cluster. We are unaware of a proof that shows entropy equalling zero when the number of classes approaches the number of records. Therefore, we begin by providing such a proof.

Using a number of classes equal to the number of unique records is akin to over-fitting the model; it tells you almost nothing about the grouping patterns in your data. Since we seek to build an entropy-based model selection criterion, we introduce a penalty function, aimed at choosing the best number of classes before encountering the tailing-off effect in entropy, which occurs as more and more unnecessary classes are used.

Moreover, we are interested in the performance of an entropy-based criterion as a model selection tool after multiple imputation has been implemented. BIC or AIC are often used to choose a model, though they do not take into account the need for a well-separated cluster solution (Fruhwirth-Schnatter, 2006). In addition, the performance of BIC and AIC breaks down after multiply imputing data sets in a regression context (Chaurasia and Harel, 2012). This leaves the field open for a new model selection criterion. Therefore, we set out to build

an entropy-based model selection criterion which outperforms BIC and AIC after multiple imputation, while considering more than one number of classes at a time.

The chapter is organized as follows. Section 4.2 details LCA entropy, and showcases our proof that the entropy of an LCA model goes to zero as the number of classes approaches the number of possible unique records. Section 4.3 describes the methodology of Harel et al. (2013), and how we propose to extend the methodology. Section 4.4 presents our simulation study, in which we compare our entropy-based criterion to AIC and BIC. Section 4.5 demonstrates an application of our entropy-based criterion, and compared the results to those obtained by AIC and BIC. Section 4.6 wraps up the chapter with our conclusions and directions for future work.

4.2 Limiting Behavior of LCA Entropy

When considering the entropy of an LCA model (Fruhwirth-Schnatter, 2006; Dias and Vermunt, 2008), we focus on the number of possible unique records. This is because the LCA model describes the probability distribution of the unique records, which can be considered a weighted sum of the probabilities of the individual records. Let $\mathbf{Y}$ be an $N \times K$ matrix of categorical data. Then let $\mathbf{X}$ be an $N^u \times K$ matrix of the unique records. Note how $\mathbf{X}$ is a subset of $\mathbf{Y}$, and retains all K variables. The probability of each row of $\mathbf{X}$ is one value of the probability mass function $p(\mathbf{X} = \mathbf{x})$, where $\mathbf{X}$ is the matrix of all potential outcomes. To find

the entropy of an LCA model, we calculate

$$H(\mathbf{X}) = -\sum_{i=1}^{N^u} p(\mathbf{X} = \mathbf{x_i}) \times ln[p(\mathbf{X} = \mathbf{x_i})],$$

The number of possible unique records is a function of the number of variables in a data set and the number of levels each variable may take. Namely, $N^u = \prod_{k=1}^{K} O_k$, where K is the number of variables and O_k is the number of levels in variable k.

To our knowledge, there does not exist a proof for entropy of an LCA model as the number of classes approaches the number of unique records. We therefore present one. To begin, we write $p(\mathbf{X} = \mathbf{x_i})$ as presented in Fruhwirth-Schnatter (2006) and Dias and Vermunt (2008), tweaked to specify unique records:

$$p(\mathbf{X} = \mathbf{x_i}) = \alpha_{ig} = \sum_{G} \gamma_g \left[\prod_{K} \left(\prod_{O_k} \left(\pi_{k,o|g}^{x_{i,k,o}^u} \right) \right) \right],$$

where $i \in (1, ..., N^u)$ is the index of possible unique records; $g \in (1, ..., G)$ is the index of classes, where G is the total number of classes; $k \in (1, ..., K)$ is the index of variables, where K is the total number of variables; $o \in (1, ..., O_k)$ is the index of categories specific to variable k, where O_k is the total number of categories for variable k; γ_g is the probability of being in class g, where $\sum_{g=1}^{G} \gamma_g = 1$; $\pi_{k,o|g}$ is the probability that a record has the o^{th} value in the k^{th} variable, given it is in class g, where $\sum_{O_k} \pi_{k,o|g} = 1$; and $x_{i,k,o}^u$ equals 1 if the variable k for unique record i has the o^{th} value.

Assume that the i^{th} unique record is in the i^{th} class, where $i = 1, ..., N^u$. If

this assumption is not true at first, a permutation of class labels will make it true. The result is the following theorem.

Theorem 4.2.1.

$$\lim_{G \to N} \left(- \sum_{N} (\alpha_{ig} \times ln(\alpha_{ig})) \right) = 0,$$

where $\alpha_{ig} = \sum_{G} \gamma_g \left[\prod_{K} \left(\prod_{O_k} \left(\pi_{k,o|g}^{x_{i,k,o}^u} \right) \right) \right]$.

The proof is given in the Appendix.

Therefore, entropy of an LCA model is minimized when the number of classes equals the number of unique records.

We want our criterion to choose a number of classes which corresponds to a low entropy value, without being fooled into over-fitting the data. This means that when we make our entropy-based criterion, we cannot rely on the absolute minimum value of entropy to point us toward the best number of classes. Instead, we must build in a penalty in order for our criterion to recognize when entropy has stopped significantly decreasing, and instead has begun to "tail off" toward zero.

4.3 Incomplete LCA Methodology and Extensions

The clustering methods used in this chapter are a blend of MI, LCA, and entropy, all of which have been discussed previously. In this section, we clarify how these topics combine to cluster incomplete categorical data, and how our new entropy-

based criterion extends the methodology. Since our goal is to develop a new model selection criterion, we assume that no prior or expert knowledge exists about the best number of classes.

4.3.1 Incomplete LCA Methodology

In Harel et al. (2013), LCA class membership and missing values of manifest variables are imputed in order to identify in which class the records belonged, and to identify the parameters associated with the LCA model. The number of classes was decided during the application, before imputation began.

Class membership and missing data values are imputed as follows. The EM algorithm finds the best starting values for the γ and π parameters of the LCA model (see Equation 1.2). Once these estimates are obtained, MI is used to multiply impute missing data values and class membership. After the imputations are complete, the imputed parameters are combined using Rubin's Rules to obtain a single LCA model, with class membership, which may then be analyzed as any LCA model would be. Since the number of classes was fixed, there is no model selection performed after imputation.

4.3.2 Extensions

In the case where the number of classes is not pre-specified, one may run the previously described methodology on the same incomplete data set, and vary

the numbers of LCA classes each model requires. This extra step will supply the analyst with a variety of LCA models to choose from; one model for every number of classes. From here, the question becomes: How does one pick a single imputed model?

The typical model selection approach in LCA is BIC. Therefore, the best model determined via this method is chosen by the number of classes which minimizes the mean of the BIC values over imputed data sets. The procedure has been demonstrated in a multiple regression framework by Chaurasia and Harel (2012). However, the usefulness of the BIC deteriorates when applied to multiply imputed data (Chaurasia and Harel, 2012). Therefore, an opportunity is presented to develop a new model selection criterion.

We have developed a new model selection criterion to replace BIC in the context detailed above. To begin, let there be a range for the potential number of classes. For each number, we impute class membership and missing manifest variables as detailed previously. However, we do not immediately combine parameter estimates. Instead, the entropy of each imputed LCA model is obtained, and the average over all imputed models is calculated. The result is that, for each specific number of classes, there is a corresponding average entropy statistic.

As we have proven in the previous section, the entropy of an LCA model will go to zero as the number of classes approaches the number of unique records. Therefore, choosing the best number of classes via minimum entropy will tend

to choose the largest number of classes. To remedy this, we first calculate the relative change in entropy over G classes, $\Delta H_g = \frac{H_g - H_{g+1}}{H_{g+1}}$. We then calculate a trimmed standard deviation of the ΔH_g values, σ^*, which does not consider the changes in entropy of the two smallest numbers of classes. This trimmed standard deviation will be more likely to capture the 'tail' effect of the change in entropy as the number of classes increases than if we had included the entropy of all classes. To choose the best number of classes, we select the number of classes g which corresponds to when ΔH_g first dips below a threshold: σ_t^*, σ^* divided by a constant t. We examine the best constant to use in the following section.

4.4 Simulation Study

The goal of the simulation study is to show that our entropy-based criterion outperforms the AIC and BIC criteria. Performance is measured by the number of times the entropy-based criterion, AIC, and BIC criteria select the correct number of classes. The average number of classes chosen is also included, to describe the overall behavior of the criteria. Following Chaurasia and Harel (2012), the arithmetic and geometric means of AIC and BIC are calculated over imputed data sets to obtain a single value.

We generated data based on the Zoo data from the UCI Machine Learning Repository (Bache and Lichman, 2013). The Zoo data has 101 records and 18 variables. Of the 18 variables, we subset seven: Hair, Feathers, Eggs, Milk,

Airborne, Predator, and Backbone. All variables are binary, with values 0/1 representing No/Yes answers to the question "Does record i have trait k?" for records $i = 1, ..., 101$ and variables $k = 1, ..., 7$. The original data set also includes a Type variable which specifies the correct class labels. Using this variable, we know there are seven classes in the original data. To avoid small sample size issues, we fit an LCA model to the complete Zoo data. Since the Zoo data has seven classes, the LCA model was fit using seven classes. To obtain our LCA model, we ran the methodology of Harel et al. (2013) on the complete data set, specifying one imputation, thus obtaining a single copy of the model parameters. The parameters obtained are:

$$\pi_{k|g=1} = \begin{pmatrix} 1.00 & 0.00 \\ 0.97 & 0.03 \\ 0.06 & 0.94 \\ 1.00 & 0.00 \\ 1.00 & 0.00 \\ 0.27 & 0.73 \\ 0.36 & 0.64 \end{pmatrix}, \pi_{k|g=2} = \begin{pmatrix} 0.05 & 0.95 \\ 1.00 & 0.00 \\ 0.97 & 0.02 \\ 0.00 & 1.00 \\ 0.95 & 0.05 \\ 0.47 & 0.52 \\ 0.00 & 1.00 \end{pmatrix}, \pi_{k|g=3} = \begin{pmatrix} 0.00 & 1.00 \\ 1.00 & 0.00 \\ 0.00 & 1.00 \\ 1.00 & 0.00 \\ 0.00 & 1.00 \\ 1.00 & 0.00 \\ 1.00 & 0.00 \end{pmatrix}$$

$$\pi_{k|g=4} = \begin{pmatrix} 1.00 & 0.00 \\ 1.00 & 0.00 \\ 0.00 & 1.00 \\ 1.00 & 0.00 \\ 1.00 & 0.00 \\ 0.50 & 0.50 \\ 0.00 & 1.00 \end{pmatrix}, \pi_{k|g=5} = \begin{pmatrix} 1.00 & 0.00 \\ 1.00 & 0.00 \\ 0.00 & 1.00 \\ 1.00 & 0.00 \\ 0.00 & 1.00 \\ 1.00 & 0.00 \\ 1.00 & 0.00 \end{pmatrix}, \pi_{k|g=6} = \begin{pmatrix} 0.00 & 1.00 \\ 1.00 & 0.00 \\ 1.00 & 0.00 \\ 0.00 & 1.00 \\ 1.00 & 0.00 \\ 0.00 & 1.00 \\ 0.00 & 1.00 \end{pmatrix}$$

$$\pi_{k|g=7} = \begin{pmatrix} 1.00 & 0.00 \\ 0.05 & 0.95 \\ 0.00 & 1.00 \\ 1.00 & 0.00 \\ 0.15 & 0.85 \\ 0.55 & 0.45 \\ 0.05 & 0.95 \end{pmatrix}, \gamma = \begin{pmatrix} 0.3267 \\ 0.3960 \\ 0.0396 \\ 0.0198 \\ 0.0099 \\ 0.0099 \\ 0.1980 \end{pmatrix}$$

We used the above parameters to generate a thousand observations, together with their class membership. Therefore, each repetition of our simulation study has a new data set with $N = 1000$ records, seven variables, and seven classes. There are $2^7 = 128$ possible unique records in the data.

Missingness was added using MAR mechanisms. Specifically, if a record had the value zero for Milk (e.g., the animal does not produce milk), then the

record's value of Predator was made missing following an MCAR mechanism with parameter ϕ_A, while if the record had value one for Milk (e.g., the animal produces milk) then the value for Predator was made missing following an MCAR mechanism with parameter ϕ_B. The ϕ values were chosen so that the total amount of missing values in Predator approximated 50%, 25%, and 10% missingness. Different thresholds were examined during the simulation study. Specifically, σ^* was divided by 2, 4, 6, 8, 12, 14, and 20.

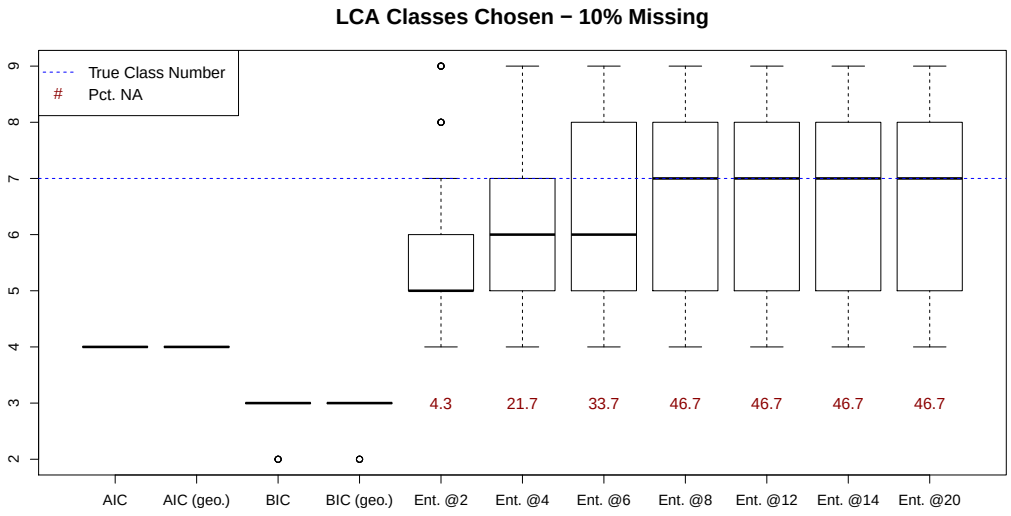


Fig. 4.1: Boxplots showing the number of classes chosen by each method for 10% missingness. Reps: 92

Results are shown in Figure 4.1, Figure 4.2, Figure 4.3, and Table 4.1. Figures 4.1 through Figure 4.3 show the number of classes chosen by each of the considered model selection criteria. The arithmetic and geometric means of AIC and BIC showed no difference in their selection of the number of classes. Both

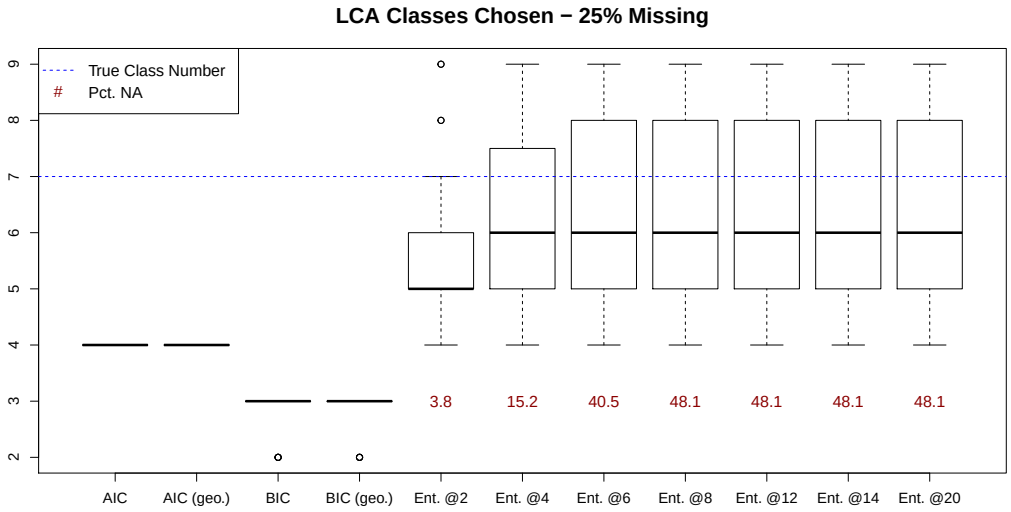


Fig. 4.2: Boxplots showing the number of classes chosen by each method for 25% missingness. Reps: 79

AIC and BIC underestimated the number of classes, selecting four classes and three classes respectively, with a outlier of two classes chosen by BIC. Entropy, meanwhile, had a much higher variability compared to AIC and BIC, but also chose numbers of classes which were much closer to the true number of classes.

Table 4.1 shows the average number of classes chosen, standard deviation of the number of classes chosen, percent of times the correct number of classes was chosen, and the percent of times the number of classes chosen was within one of the correct number, for each entropy-based model selection criterion examined. We examine only entropy-based criteria in this table, as Figures 4.1 through 4.3 showed that AIC and BIC do not select the best number of classes and have nearly no variability. Percents in Table 4.1 are calculated relative to the number

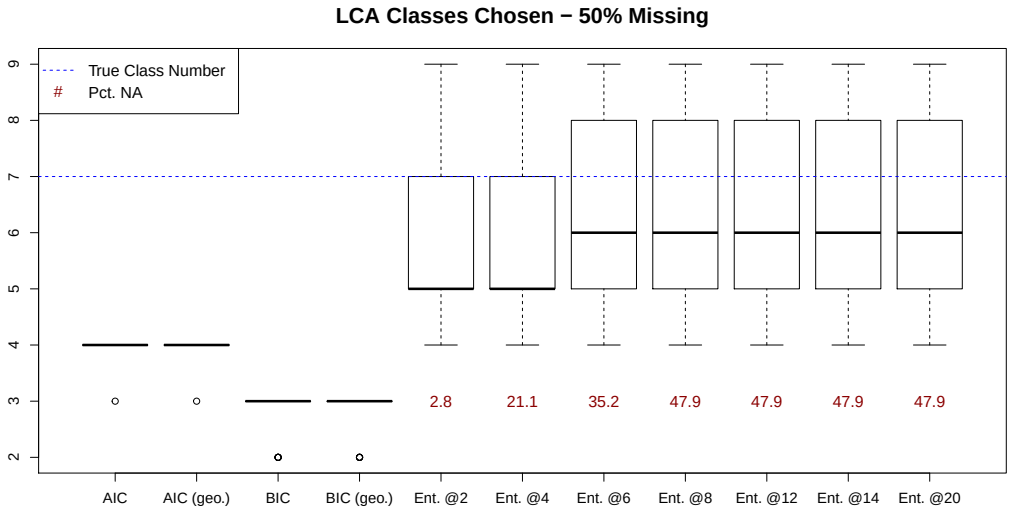


Fig. 4.3: Boxplots showing the number of classes chosen by each method for 50% missingness. Reps: 71

of non-NA repetitions per setting.

From Table 4.1, we can see the percent of times the correct number of classes was chosen (Pct. 7) varied considerably, but was maximized for σ_4^* for all three levels of Percent Missingness ($\%Mis.$). The percent of being within one of the correct number of classes (Pct. 6-8) was maximized at σ_2^*, with σ_4^* coming in a close second. Additionally, the percent of repetitions who had no entropy value below the threshold was maximized at σ_{20}^*, as was expected. In other words, as the threshold becomes more and more strict, fewer and fewer cases meet the threshold.

We take the above results to indicate that our method is performing better than AIC and BIC. Additionally, we begin to suspect that σ_2^* or σ_4^* is the best

Results	%Mis.	σ_2^*	σ_4^*	σ_6^*	σ_8^*	σ_{12}^*	σ_{14}^*	σ_{20}^*
Pct. 7	10	9.3	13.95	13.95	5.56	2.78	0	3.03
	25	11.63	13.95	8.33	2.78	2.78	3.03	3.03
	50	13.95	13.95	11.11	2.78	9.09	3.03	3.03
Pct. 6-8	10	32.56	27.91	27.91	25	22.22	21.21	30.3
	25	30.23	27.91	16.67	22.22	22.22	30.3	30.3
	50	30.23	27.91	30.56	22.22	27.27	30.3	30.3
Pct. NA	10	2.33	37.21	37.21	36.11	41.67	21.21	45.45
	25	16.28	37.21	5.56	41.67	41.67	33.33	45.45
	50	25.58	37.21	19.44	41.67	6.06	45.45	45.45

Table 4.1: Percent correct (Pct. 7), percent near-correct (Pct. 6-8), and percent which never met the threshold (Pct. NA) for 10%, 25%, and 50% Missing. Results for entropy values only. Column headers σ_t^* indicate thresholds are divided by $t = 2, 4$, and so on.

penalty for the entropy-based criterion. However, we will revisit this idea in the data application.

4.5 Application

Rink et al. (2014) analyzed a subset of data collected and described in Kordas et al. (2011); we look at a subset of the data in Rink et al. (2014). Specifically, we look

at father's education (in years), parents marital status (married, divorced, living together unmarried), mother's intelligent quotient (IQ), amount of hemoglobin in the child's blood (ug/dl), socio-economic status (SES) of the child's home, HOME score, amount of manganese (Mn) in the child's blood(ug/g), and age of the child (in months). Socio-economic status was derived as the sum of twelve indicators, each asking whether the household contained items such as a TV, refridgerator, and DVD player. HOME score was derived using a survey constructed to assess the home environment; for a more detailed variable description, see Rink et al. (2014, pg. 48).

Table 4.2: Missing Values

	Dad's Edu.	Marital	Mom's IQ	Child Hb	SES	HOME	Child Mn	Child Age
Freq.	9	1	23	9	2	11	9	0
Pct.	8.3	0.9	21.1	8.3	1.8	10.1	8.3	0.0

Table 4.3: Frequency and percent of missing values in each variable.

Table 4.3 shows a table of the frequency (top row) and respective percentage (bottom row) of missing values in the data. For each variable, there is between 0 to 23 missing values, representing 0% to 21.1% missingness. The pattern of missingness for this data is shown in Figure 4.4, where grey sections of the graph indicate observed values, and black sections indicate missing values (generated by the R package Amelia (Honaker et al., 2011)). If we were to remove incomplete

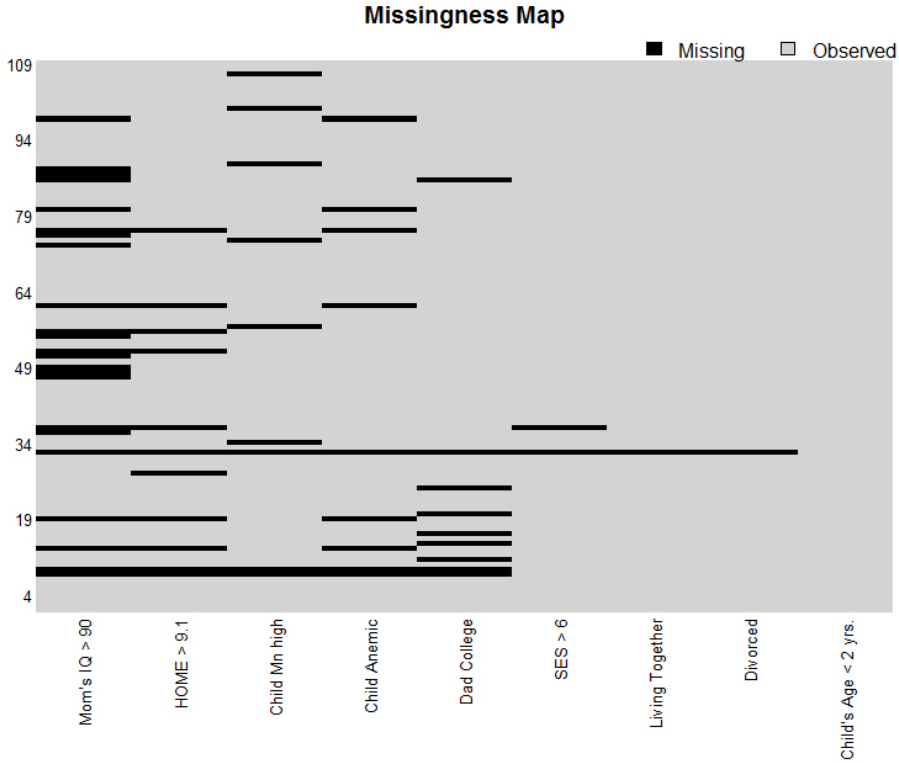


Fig. 4.4: Plot of observed (grey) and missing (black) values in our subset of variables.

records from this data set, our sample of 109 records would decrease to 74, a reduction in sample size of almost a third.

Histograms of all observed values of the variables are shown in Figure 4.5. Marital status is a three-level categorical variable, while the others are all continuous. Many follow a bell-shaped distribution, while only Father's Education and Manganese Level show severe skewness.

As most of these variables are continuous, and we address binary data, we created binary versions of each variable in the following manner. Father's

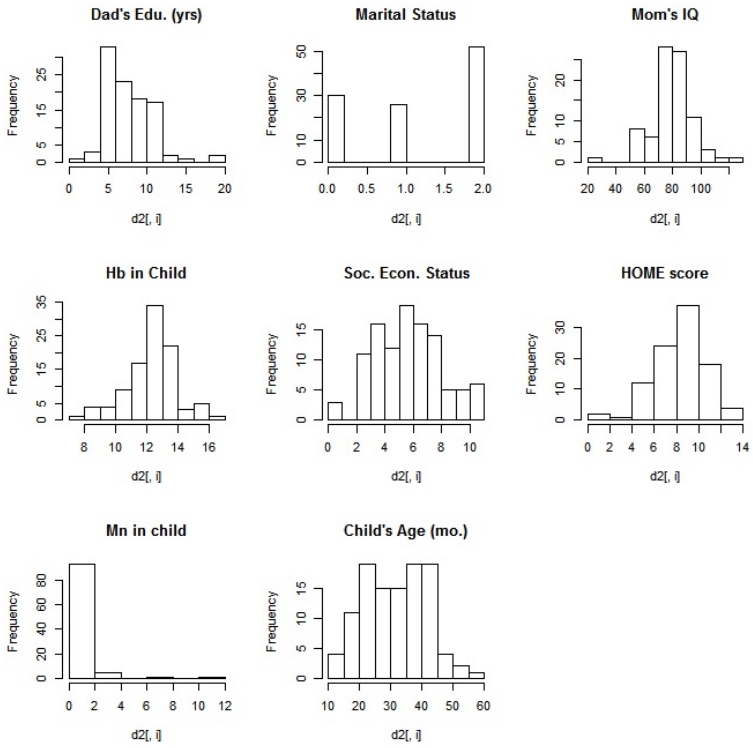


Fig. 4.5: Histograms of the eight variables analyzed in our application study.

education was flagged as either no college (years ≤ 12), or at least some college (years > 12). Parents' marital status was made into two indicator variables, one each for divorced and living together unmarried, with married as the reference category. As the sample had the majority of Mother's IQ values below 100, the cutoff point for below-average IQ was set at 90 instead of 100. Hemoglobin levels below 11 (the mean of two standard deviations below normal levels for children six months to two years (10.5) and two to six years (11.5) (Janus and Moerchel, 2010)) were coded to indicate anemic children. The thresholds for socio-economic status and HOME score were approximately the respective means, 6 and 9.1, indicating

above- and below-average socio-economic status and HOME score. Manganese levels were coded as above or below the recommended safety level of 1.2 ug/g for children one to three years old (Chen and Copes, 2011). Age was made into an indicator for below toddling age (months $\leq$ 24) or at-or-above toddling age (months $>$ 24).

4.5.1 Results

We applied the methodology of Harel et al. (2013) to impute and classify records into classes. The number of classes ranged from two to eight. We obtained AIC, BIC, and entropy values for each of the seven considered class numbers. The minimum AIC and BIC values indicated the best number of classes, according to those criteria. Three thresholds were explored: σ^*, $\sigma^*/2$, and $\sigma^*/4$. The entropy values, and relative change in entropy values with three thresholds, are shown in Figure 4.6.

From Figure 4.6, we can see that the most lenient penalty, σ^*, captures any of the entropy values. Having σ^* as our best penalty is a different result than our simulation findings, which suggested σ_2^* or σ_4^* as the best penalty function. The difference between the simulation and application results indicates that a more lenient penalty may be necessary for a data applications, which are typically more chaotic than simulation data. Specifically, the application results show that four classes is the first to fall within the bounds of our penalty, and seven classes is

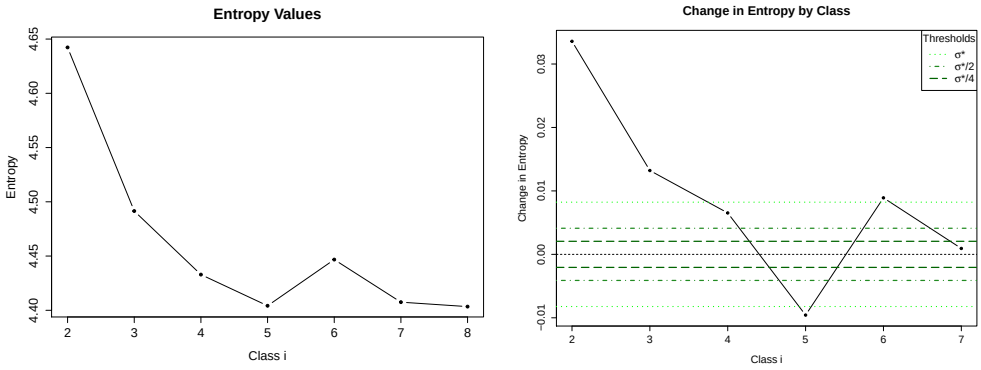


Fig. 4.6: Entropy values (left) and change in entropy values (right), with thresholds.

the last. However, when we examine more than four classes, we find many classes empty. We then settle on four classes.

Our entropy-based criterion found four classes, while AIC and BIC agreed that there were two classes. Variable descriptions for each entropy-based criterion and AIC class are shown in Table 4.4. The percents are calculated with respect to each variable's count of observed values. The entropy-based criterion found more classes than AIC, but does this extra information tell us anything?

Let's start with the AIC results. All parents are divorced in Class 1, compared with a 63%-37% spread across Living Together and Married in Class 2. Mother's IQ is predominantly under 90 in both categories, with little distinction between classes (10% and 24.2% above 90). Results for anemia (20% and 17.3%) and socio-economic standard (34.6% and 45.7%) are barely distinguishable between classes. Over half of families in Class 2 have high HOME scores,

% Records	College	Div.	Liv. Tog.	Marr.	IQ > 90	Anemic	SES > 6	HOME > 9.1	Mn High	Infant
24.5	0.0	100.0	0.0	0.0	10.0	20.0	32.0	0.0	41.7	23.1
24.0	4.3	0.0	100.0	0.0	15.0	22.7	17.4	22.7	85.0	17.4
30.1	8.7	0.0	50.0	50.0	10.8	18.6	51.1	70.0	0.0	37.5
21.4	0.0	8.3	41.7	50.0	100.0	0.0	83.3	66.7	0.0	25.0
% Records	College	Div.	Liv. Tog.	Marr.	IQ > 90	Anemic	SES > 6	HOME > 9.1	Mn High	Infant
24.5	0.0	100.0	0.0	0.0	10.0	20.0	34.6	4.0	40.0	26.9
75.5	6.2	0.0	63.4	36.6	24.2	17.3	45.7	54.8	22.7	28.9

Table 4.4: Percent "Yes" to each of nine categories. Four Classes found via the new entropy-based criterion (EBC). Two Classes found via AIC & BIC. Percents are calculated with respect to each variable's count of observed values within each class.

compared to only 4% in Class 1. Manganese levels are more elevated in Class 1; 40% compared to 23%. Overall, we can glean some information, such as differences between HOME scores, but other distinctions between the classes are not clear.

Now let us move on to the entropy-based criterion results. Class 1 consists entirely of divorced parents with below-average HOME scores, and with almost 42% of children having high manganese levels. Class 2 has only living-together parents, with homes primary below average for both SES and HOME scores, and with 85% of children having high manganese levels. Class 3 is evenly divided between living-together and married parents, with 51% and 70% of homes having above-average SES and HOME scores respectively, and with no children having high manganese levels. Class 4 has a mix of parental married statuses, with all mothers having IQs above 90, no anemic children, 83.3% and 66.7% of homes having high SES and HOME scores respectively, and no children with high manganese levels.

Obtaining four classes instead of two has brought several things to light. The entropy-based criterion identified one class where almost all mothers have IQs higher than 90, which the two-class dichotomy did not capture. In this class, no child was anemic or had high manganese levels. Our new criterion also found one class where 85% of children had high manganese levels, which was not captured by using two classes. Perhaps this class could be targeted for high manganese information in future. These differences show that the five-class solution is better

for describing patterns in the data set. In other words, our new entropy-based criterion has given more useful results than AIC.

4.6 Conclusion

There is a need for a model selection criterion which performs well after MI has been implemented. In this chapter, we have presented an entropy-based criterion which accurately chooses the number of classes and outperforms AIC and BIC.

We have derived a proof which shows that the entropy of an LCA model, where each variable has its own number of categories, goes to zero as the number of classes approaches the number of unique records. With this knowledge, we have developed an entropy-based model selection criterion which uses a penalty to recognize when entropy has turned from clearly separating data into clusters to over-fitting the data by needless divisions.

Simulation studies have shown that our proposed entropy-based criterion outperforms the arithmetic and geometric means of AIC and BIC for selecting a model with the correct or near-correct number of classes. This result means that we may utilize our entropy-based criterion to extend cluster analysis of incomplete data from the current state of the art, which requires a pre-specified number of clusters, to an selection from among a set of possible numbers of clusters. Applying our entropy-based criterion to a family studies data set brought to light unique and descriptive strata which AIC and BIC failed to find.

Extensions include building upon our multinomial LCA proof by applying our entropy-based criterion to multinomial data. Applying our new criterion to a GIS data set, in order to capitalize on the categorical nature of that data, is of great interest. In addition, we wish to extend the current work to LC cluster analysis for continuous data.

While AIC and BIC are common LCA model selection criteria, there are other information-based criteria which involve entropy (Fruhwirth-Schnatter, 2006). In future, we would like to compare our criterion to such competitors.

The application relied on binary conversion of continuous data. We would like to study the sensitivity of the binary classifications. For example, would the structure of the classes found by our criterion change if we placed the cut-off for IQ at 100 instead of 90?

Also, note that we were not interested in $G = 1$ versus $G > 1$. In other words, we assume that there is some clustering behavior in the data. In future, we would like to develop inference with our estimator to test the presence or absence of clustering behavior in the data.

Chapter 5

Conclusions & Future Work

5.1 Conclusion

This thesis tackles three important questions that arise when clustering incomplete data. First, how can researchers address incomplete normal mixture model data in a way that includes incomplete records and takes into account all imputed clustering solutions? Second, how can we calculate the entropy of an incomplete data model, taking into account both the variability in the data and the additional uncertainty introduced by the missingness mechanism? And lastly, can we develop a model selection criterion for clustering incomplete categorical data that frees current methodologies from the restriction of pre-specifying the number of clusters, while also choosing the correct model more often than AIC and BIC?

The answer to the first question is in Chapter 2, which presents MICA: Multiply Imputed Cluster Analysis. MICA is a multi-step framework which can be tailored to specific data structures. MICA has a unique two-stage clustering approach, which allows us to combine clustering results from multiply imputed

data sets which disagree as to the number of clusters. Such an approach allows us to consider more than one number of clusters during the analysis, thus freeing us from the restrictions of previous methodologies.

Chapter 2 also demonstrates MICA-N, a specialized version of MICA for data that follow a normal mixture model. Simulations show that MICA-N outperforms its CCA counterpart. A genetic data application showed that MICA-N uncovered more nuanced and useful patterns than the complete case analysis.

Comparison against the complete case analysis may seem trivial to some missing data analysts. The authors wish to emphasize that this thesis chapter documents the first of many steps in demonstrating the applicability of MICA, in which case it is appropriate to begin with a comparison to CCA before proceeding to other, more complex methods of handling missing values. If MICA-N had not compared favorably to CCA, there would be no point in continuing the research in this direction. Since MICA-N has proven itself against CCA, the authors look forward to more advanced comparison studies in future.

The second question is answered in Chapter 3, where we derive brand new bivariate and multivariate estimators for measuring the entropy of a normally distributed data model under MCAR and MAR missingness. These theorems show that, even under ignorability, the pattern of missingness is important to accurately estimating the entropy of an incomplete data set.

For both bivariate and multivariate cases, we demonstrate how a researcher

can estimate the entropy of the complete data model from the entropy of the incomplete data model if he or she knows the missingness mechanism. We showed that the bivariate estimators converge to their complete-data counterparts as the percent of missing values approaches zero. Simulations showed that the new bivariate estimators produced values that were closer to the entropy of the complete data model than the CCA counterpart, and had smaller standard deviations.

Correlations between incomplete entropy and the fraction of missing information (FMI), as well as the individual components of the FMI formula, were also analyzed. FMi is not significantly associates with incomplete entropy, but its component parts tend to be significantly associated, especially under MCAR missingness and when estimating the mean of the incomplete variable.

The third question is answered in Chapter 4, where we prove that entropy of an LCA model is minimized when the number of classes equals the number of unique records. We then apply this knowledge to build a new model selection criterion which selects the best number of classes for an LCA model after imputing missing data. The new model selection criterion allows LCA analysis on incomplete data, such as Harel et al. (2013), to be rid of the requirement of pre-specifying the number of classes the LCA model describes.

We analyze our criterion under a number of different penalty functions, and observe the trade-off between accuracy of the criterion and strictness of the penalty function. Our new model selection criterion outperforms the AIC and

BIC criteria, which are typically used to choose which number of classes best fits an LCA model.

An application of our new criterion to a human development and family studies data set uncovered four classes which identified groups with healthy or concerning levels of anemia and manganese. These groups were further described by mother's IQ and parental marital status. Using AIC and BIC uncovered two groups with little difference between them. Implementing our criterion thus uncovered more nuanced and more useful strata in the data set.

Taken all together, the chapters of this thesis have answered three key questions which arise when attempting to cluster incomplete data.

There are several directions in which we would like to extend our current research. One direction of interest is extending MICA to other data structures by building new algorithms to impute missing values from a skewed-t distribution, and to cluster them appropriately. These new algorithms would extend MICA to skewed-t mixture model data, while at the same time allowing different numbers of clusters in each imputed clustering solution.

Another direction is to extend our work on entropy of an incomplete data model to non-normal data, and to non-ignorable missingness. Tackling non-normal data cases would increase the applicability of our new understanding of incomplete data model entropy, while considering MNAR missingness is of interest due to our realization that entropy calculation requires knowledge of the

missingness mechanism, even under conditions where it is otherwise ignorable. We are also interested in teasing apart the data- and missingness-based parts of the MAR entropy results, to lend a theoretical foundation to the simulation results we observed.

One final direction is to extend our new entropy-based model selection criterion from the binary variable case to the multinomial case, and to the case where different variables have different numbers of levels. Such an extension would support the new theorem we have already developed, and extend the usefulness of our criterion to more complex data cases. A comparison between the performance of the binary cirterion in the bifurcated data set, and the continuous-data methodology of MICA-N, is also of great interest.

Chapter 6

References

6.1 *

References

Anderson, E. (1935). The irises of the gaspe peninsula. *Bulletin of the American Iris Society*, 59:2–5.

Bache, K. and Lichman, M. (2013). UCI machine learning repository. donated by Richard Forsyth.

Banfield, J. D. and Raftery, A. E. (1993). Model-based gaussian and non-gaussian clustering. *Biometrics*, 49(3):803 – 821.

Barbara, D., Li, Y., and Couto, J. (2002). COOLCAT: an entropy-based algorithm for categorical clustering. *Proceedings of the eleventh international conference on information and knowledge management*, pages 582–589.

Basagaña, X., Barrera-Gomez, J., Benet, M., Anto, J. M., and Garcia-Aymerich, J. (2013). A framework for multiple imputation in cluster analysis. *Practice of Epidemiology*, 177:718–725.

Beck, M. W., Vondracek, B., and Hatch, L. K. (2013). Environmental clustering of lakes to evaluate performance of a macrophyte index of biotic integrity. *Aquatic Botany*, 108:16–25.

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated classification likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:719–725.

Biernacki, C. and Govaert, G. (1997). Using the classification likelihood to choose the number of clusters. *Computing Science and Statistics*, 29:451–457.

Bitsika, V., Sharpley, C. F., and Orapeleng, S. (2008). An exploratory analysis of the use of cognitive, adaptive, and behavioural indices for cluster analysis of asd subgroups. *Journal of Intellectual Disability Research*, 52:973–985.

Bras, L. P. and Menezes, J. C. (2007). Improving cluster-based missing value estimation of DNA microarray data. *Biomolecular Engineering*, 24:273–282.

Celeux, G. and Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, 13:195–202.

Chambers, J. M., Cleveland, W. S., Tukey, P. A., and Kleiner, B. (1983). *Graphical Methods for Data Analysis*. Wadsworth & Brooks/Cole.

Chaurasia, A. and Harel, O. (2012). Using AIC in multiple linear regression framework with multiply imputed data. *Health Services and Outcomes Research Methodology*, 12:219–233.

Chen, H. and Copes, R. (2011). Manganese in drinking water and intellectual impairment in school-age children. *Environmental Health Perspectives*, 119.

Cho, R. J., Campbell, M. J., Winseler, E. A., Steinmetz, L., Conway, A., Wodicka, L., et al. (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell*, 2(1):65–73.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory, Second Edition*. Wiley-Interscience, New York.

Cranford, J. A., McCabe, S. E., Boyd, C. J., Slayden, J., Reed, M. B., Ketchie, J. M., Lange, J. E., and Scott, M. S. (2008). Reasons for nonresponse in a web-based survey of alcohol involvement among first-year college students. *Addictive Behaviors*, 33:206–210.

Dahl, D. B. (2014). *xtable: Export tables to LaTeX or HTML*. R package version 1.7-3.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1997). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39:138.

Dias, J. G. and Vermunt, J. K. (2008). A bootsrap-based aggregate classifier for model-based clustering. *Computational Statistics*, 23:643–659.

Everitt, B. S., Landau, S., Leese, M., and Stahl, S. (2011). *Cluster Analysis*. John Wiley & Sons, Inc.

Fisher, R. A. (1921). On the 'probable error' of a coefficient of correlation deduced from a small sample. *Metron 1*, pages 3–32.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188.

Fraga, S., Severo, M., Costa, D., Lopes, C., and Ramos, E. (2010). Clustering behaviors among 13-year-old Portuguese adolescents. *Journal of Public Health*, 19:21–27.

Fraley, C. and Raftery, A. E. (1998). How many clusters? which clustering method? answers via model-based cluster analysis. *The Computer Journal*, 41:578–588.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97:611–631.

Fraley, C., Raftery, A. E., and Scrucca, L. (2012). mclust version 4 for r: Normal mixture modeling for model-based clustering, classification, and density estimation. *Technical Report No. 597, Department of Statistics, University of Washington*.

Fruhwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer.

Gebregziabher, M. and DeSantis, S. M. (2010). Latent class based multiple imputation approach for missing categorical data. *Journal of Statistical Planning and Inference*, 140:3252–3262.

Hagenaars, J. A. and McCutcheon, A. L. (2002). *Applied Latent Class Analysis*. Cambridge University Press, New York.

Harding, H. G., Burns, E. E., and Jackson, J. L. (2012). Identification of child sexual abuse survivor subgroups based on early maladaptive schemas: Implications for understanding differences in posttraumatic stress disorder symptom severity. *Cognitive Therapy and Research*, 36:560–575.

Harel, O., Chung, H., and Miglioretti, D. (2013). Latent class regression: Inference and estimation with two-stage multiple imputation. *Biometrical Journal*, 55:541–553.

Harel, O. and Miglioretti, D. (2007). Missing information as a diagnostic tool for latent class analysis. *Journal of Data Science*, 5:269–288.

Harel, O., Pellowski, J., and Kalichman, S. (2012). Are we missing the important of missing values in HIV prevention randomized clinical trials? review and recommendations. *AIDS and Behavior*, 16:1382–1393.

Harel, O. and Zhou, X.-H. (2007). Multiple imputation: Review of theory, implementation, and software. *Statistics in Medicine*, 26:3057–3077.

Hathaway, R. J. and Bezdek, J. C. (2001). Fuzzy c-means clustering of incomplete data. *IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics*, 31:735–744.

Heitjan, D. F. and Little, R. J. A. (1991). Multiple imputation for the fatal accident reporting system. *Journal of the Royal Statistical Society C*, 01:13–29.

Honaker, J., King, G., and Blackwell, M. (2011). Amelia II: A program for missing data. *Journal of Statistical Software*, 45(7):1–47.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2:193–218.

Iyer, V. R., Eisen, M. B., Ross, D. T., Schuler, G., Moore, T., Lee, J. C. F., Trent, J. M., Staudt, L. M., Jr., J. H., Boguski, M. S., Lashkari, D., Shalon, D., Botstein, D., and Brown, P. O. (1999). The transcriptional program in the response of human fibroblasts to serum. *Science*, 283:83–87.

Janus, J. and Moerchel, S. K. (2010). Evaluation of anemia in children. *American Family Physician*, 81:1462–1471.

Kaufman, L. and Rousseeuw, P. J. (1987). Clustering by means of medoids. *Statistical Data Analysis based on the L_1 Norm*, pages 405–416.

Keerin, P. and Kurutach, W. (2012). Cluster-based knn missing value imputation for DNA microarray data. *IEEE International Conference on Systems, Man, and Cybernetics*.

Kordas, K., Ardoino, G., Ciccariello, D., nay, N. M., Ettinger, A. S., Cook, C. A., and Queirolo, E. I. (2011). Association of maternal and child blood lead and hemoglobin levels with maternal perceptions of parenting their young children. *NeuroToxicology*, 32:693–701.

Lagona, F. and Picone, M. (2012). Model-based clustering of multivariate skew data with circular components and missing values. *Journal of Applied Statistics*, 39:927–945.

Li, T., Ma, S., and Ogihara, M. (2004). Entropy-based criterion in categorical clustering. *Proceedings of the 2004 IEEE International Conference on Machine Learning*, pages 536–543.

Linzer, D. A. and Lewis, J. B. (2011). poLCA: An R package for polytomous variable latent class analysis. *Journal of Statistical Software*, 42.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Wiley-Interscience, New York.

Liu, R., Anand, A., Dey, D. K., and Javidi, B. (2014). Entropy-based clustering of embryonic stem cells using digital holographic microscopy. *Journal of the Optical Society of America A*, 31:677–684.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K. (2014). *cluster: Cluster Analysis Basics and Extensions*. R package version 1.15.2 — For new features, see the 'Changelog' file (in the package source).

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. Wiley-Interscience.

Muthén, L. K. and Muthén, B. O. (2011). *Mplus User's Guide, Sixth Edition*. Los Angeles, CA: Muthén & Muthén.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raftery, A. E. and Dean, N. (2006). Variable selection for model-based clustering. *Journal of the American Statistical Association*, 101(473):168–178.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66:846–850.

Rink, S. M., Ardoino, G., Queirolo, E. I., Cicariella, D., nay, N. M., and Kordas, K. (2014). Associations between hair manganese lvels and cognitive, language, and motor development in preschool children from montevideo, uruguay. *Archives of Environmental and Occupational Health*, 69:46–54.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63:581–592.

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

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall.

Schafer, J. L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research*, 8:3–15.

Schafer, J. L. (2008). *NORM: Analysis of incomplete multivariate data under a normal model, Version 3*. Software package for R. University Park, PA: The Methodology Center, The Pennsylvania State University.

Schafer, J. L. and Graham, J. W. (2002). Missing data: Out view of the state of the art. *Psychological Methods*, 7:147–177.

Schenker, N. and Taylor, J. M. G. (1996). Partially parametric techniques for multiple imputation. *Computational Statistics and Data Analysis*, 22:425–446.

Scott, A. J. and Symons, M. J. (1971). Clustering methods based on likelihood ratio criteria. *Biometrics*, 27:387–397.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423.

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of data clusters via the gap statistic. *Journal of the Royal Statistical Society B*, 63:411–423.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.

Vermunt, J. K., van Ginkel, J. R., van der Ark, L. A., and Sijtsma, K. (2008). Multiple imputation of incomplete categorical data using latent class analysis. *Sociological Methodology*, 38:369–397.

Whitwell, J. L., Przybelski, S. A., Weigand, S. D., Ivnik, R. J., Vemuri, P., Gunter, J. L., Senjem, M. L., Shiung, M. M., Boeve, B. F., Knopman, D. S., Parisi, J. E., Dickson, D. W., Petersen, R. C., Jr., C. R. J., and Josephs, K. A. (2099). Distinct anatomical subtypes of the behavioral variant of frontotemporal dementia: A cluster analysis study. *Brain*, 132:2932–2946.

Wolfe, J. H. (1970). Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research*, 5:329–350.

Wood, A. M., White, I. R., Hillsdon, M., and Carpenter, J. (2005). Comparison of imputation and modelling methods in the analysis of a physical activity trial with missing outcomes. *International Journal of Epidemiology*, 34:89–99.

Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E., and Ruzzo, W. L. (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics*, 17:977–987.

Zhang, T., Ramakrishnan, R., and Livny, M. (1997). BIRCH: A new data clustering algorithm and its applications. *Data Mining and Knowledge Discovery*, 1:141–182.

Appendix A

Appendix

A.1 Proof of Theorem 3.2.1

Proof. Component-wise, the distribution of missingness in y_{2i} is written $r_{2i} \sim f(r_{2i}|y_{1i}, y_{2i}, \phi)$, where ϕ is a parameter of r_2. Since we are assuming MCAR, r_2 cannot depend on y_1 nor on y_2. Therefore $f(r_{2i}|y_{1i}, y_{2i}, \phi) = f(r_{2i}|\phi)$. Specifically, $r_{2i} \sim Bernoulli(\phi)$, where ϕ is the complement of the percent missing in the data. Parameters of y_1, y_2 and r (θ and ϕ respectively) have been suppressed in the following derivations.

To begin, the entropy of one record in our data is

$$\begin{aligned} H(x_i) &= -\int_{y_{1i},y_{2i},r_{2i}} f(y_{1i}, y_{2i}, r_{2i}) ln f(y_{1i}, y_{2i}, r_{2i}) d(y_{1i}, y_{2i}, r_{2i}) \\ &= -\int_{y_{1i}}\int_{y_{2i}}\int_{r_{2i}} f(y_{1i}, y_{2i}, r_{2i}) ln f(y_{1i}, y_{2i}, r_{2i}) dr_{2i} dy_{2i} dy_{1i} \end{aligned}$$

To separate the joint distribution of y_{1i}, y_{2i}, and r_{2i} into something which might factor out of an integral, we use the identity $f(y_{1i}, y_{2i}, r_{2i}) = f(y_{1i}) f(y_{2i}|y_{1i}) f(r_{2i})$. Note that this identity is only true under the MCAR assumption. Using this identity, the above entropy is reduced to

$$\begin{aligned} H(x_i) &= -\int_{y_{1i}}\int_{y_{2i}}\int_{r_{2i}} f(y_{1i}) f(y_{2i}|y_{1i}) f(r_{2i}) ln\left[f(y_{1i}) f(y_{2i}|y_{1i}) f(r_{2i})\right] dr_{2i} dy_{2i} dy_{1i} \\ &= -\int_{y_{1i}}\int_{y_{2i}}\int_{r_{2i}} f(y_{1i}) f(y_{2i}|y_{1i}) f(r_{2i}) \left[ln f(y_{1i}) + ln f(y_{2i}|y_{1i}) + ln f(r_{2i})\right] dr_{2i} dy_{2i} dy_{1i} \\ &= -\int_{y_{1i}}\int_{y_{2i}}\int_{r_{2i}} f(y_{1i}) f(y_{2i}|y_{1i}) f(r_{2i}) ln f(y_{1i}) dr_{2i} dy_{2i} dy_{1i} \\ &\quad -\int_{y_{1i}}\int_{y_{2i}}\int_{r_{2i}} f(y_{1i}) f(y_{2i}|y_{1i}) f(r_{2i}) ln f(y_{2i}|y_{1i}) dr_{2i} dy_{2i} dy_{1i} \\ &\quad -\int_{y_{1i}}\int_{y_{2i}}\int_{r_{2i}} f(y_{1i}) f(y_{2i}|y_{1i}) f(r_{2i}) ln f(r_{2i}) dr_{2i} dy_{2i} dy_{1i} \end{aligned}$$

$$
\begin{aligned}
H(x_i) = &- \int_{y_{1i}} \int_{y_{2i}} f(y_{1i}) f(y_{2i}|y_{1i}) lnf(y_{1i}) \left(\int_{r_{2i}} f(r_{2i}) dr_{2i} \right) dy_{2i} dy_{1i} \\
&- \int_{y_{1i}} \int_{y_{2i}} f(y_{1i}) f(y_{2i}|y_{1i}) lnf(y_{2i}|y_{1i}) \left(\int_{r_{2i}} f(r_{2i}) dr_{2i} \right) dy_{2i} dy_{1i} \\
&- \int_{y_{1i}} \int_{y_{2i}} f(y_{1i}) f(y_{2i}|y_{1i}) \left(\int_{r_{2i}} f(r_{2i}) lnf(r_{2i}) dr_{2i} \right) dy_{2i} dy_{1i},
\end{aligned}
$$

The first and second integral over r_{2i} equal one; the third is the entropy of the distribution of r_{2i}, which will be denoted $H(r_{2i})$. The result is:

$$
\begin{aligned}
H(x_i) = &- \int_{y_{1i}} \int_{y_{2i}} f(y_{1i}) f(y_{2i}|y_{1i}) lnf(y_{1i}) dy_{2i} dy_{1i} \\
&- \int_{y_{1i}} \int_{y_{2i}} f(y_{1i}) f(y_{2i}|y_{1i}) lnf(y_{2i}|y_{1i}) dy_{2i} dy_{1i} \\
&+ H(r_{2i}) \int_{y_{1i}} \int_{y_{2i}} f(y_{1i}) f(y_{2i}|y_{1i}) dy_{2i} dy_{1i}.
\end{aligned}
$$

Pulling terms out of the integral over y_{2i} gives:

$$
\begin{aligned}
H(x_i) = &- \int_{y_{1i}} f(y_{1i}) lnf(y_{1i}) \left(\int_{y_{2i}} f(y_{2i}|y_{1i}) dy_{2i} \right) dy_{1i} \\
&- \int_{y_{1i}} f(y_{1i}) \left(\int_{y_{2i}} f(y_{2i}|y_{1i}) lnf(y_{2i}|y_{1i}) dy_{2i} \right) dy_{1i} \\
&+ H(r_{2i}) \int_{y_{1i}} f(y_{1i}) \left(\int_{y_{2i}} f(y_{2i}|y_{1i}) dy_{2i} \right) dy_{1i}.
\end{aligned}
$$

The first and third integral over y_{2i} equal one; the second is the entropy of the distribution of $y_{2i}|y_{1i}$, which we will denote $H(y_{2i}|y_{1i})$. The result is:

$$
\begin{aligned}
H(x_i) = &- \int_{y_{1i}} f(y_{1i}) lnf(y_{1i}) dy_{1i} \\
&+ H(y_{2i}|y_{1i}) \int_{y_{1i}} f(y_{1i}) dy_{1i} \\
&+ H(r_{2i}) \int_{y_{1i}} f(y_{1i}) dy_{1i}.
\end{aligned}
$$

The second and third integrals equal one, while the third is the entropy of the distribution of y_{1i}, which we will denote $H(y_{1i})$. Therefore, the entropy of one record of an incomplete data set as defined above is

$$H(x_i) = H(y_{1i}) + H(y_{2i}|y_{1i}) + H(r_{2i}). \tag{A.1}$$

The records in the MCAR case are independent and identically distributed. Therefore, we may sum Equation A.1 n times to account for the n iid records, and obtain

$$H(x_i) = \sum_{i=1}^{n} (H(y_{1i}) + H(y_{2i}|y_{1i}) + H(r_{2i})) = nH(y_{1i}) + nH(y_{2i}|y_{1i}) + nH(r_{2i}) \tag{A.2}$$

as a framework for entropy of an MCAR incomplete bivariate normal dataset.

The following are clear from our assumptions, introductory mathematical statistics, and textbook entropy derivations:

- $y_{1i} \sim N_1(\mu_1, \sigma_1^2)$, therefore $H(y_{1i}) = \frac{1}{2} ln(2\pi e \sigma_1^2)$
- $y_{2i}|y_{1i} \sim N_1(\mu_{2.1}, \sigma_{2.1}^2 = \sigma_2^2(1-\rho^2))$, therefore $H(y_{2i}|y_{1i}) = \frac{1}{2} ln(2\pi e \sigma_2^2(1-\rho^2))$
- $r_{2i} \sim Bern(\phi)$, therefore $H(r_{2i}) = -(1-\phi)ln(1-\phi) - \phi ln(\phi)$,

If we plug these values into Equation A.2 and simplify the expression, we obtain Theorem 3.2.1. □

A.2 Proof of Theorem 3.2.2

Proof. Our formula for bivariate normal data with Bernoulli missingness is

$$\frac{n}{2} ln(2\pi e \sigma_1^2) + \frac{n}{2} ln(2\pi e \sigma_2^2(1-\rho^2)) - n(1-\phi)ln(1-\phi) - n\phi ln(\phi).$$

We are interested to see whether $\lim_{\phi \to 0}[n(1-\phi)ln(1-\phi) + n\phi ln(\phi)] = 0$:

$$\begin{aligned}
\lim_{\phi \to 0}[n(1-\phi)ln(1-\phi) + n\phi ln(\phi)] &= n \lim_{\phi \to 0}[(1-\phi)ln(1-\phi)] + n \lim_{\phi \to 0}[\phi ln(\phi)] \\
&= n \lim_{\phi \to 0}[(1-\phi)] \lim_{\phi \to 0}[ln(1-\phi)] + n \lim_{\phi \to 0}[\phi ln(\phi)] \\
&= n \lim_{\phi \to 0}[ln(1-\phi)] + n \lim_{\phi \to 0}[\phi ln(\phi)],
\end{aligned}$$

since $\lim_{\phi \to 0}[(1-\phi)] = 1$.

$$n \lim_{\phi \to 0}[ln(1-\phi)] + n \lim_{\phi \to 0}[\phi ln(\phi)] = n \lim_{\phi \to 0}[\phi ln(\phi)]$$

since $ln(1-\phi) = 0$ as ϕ goes to zero.

$$n \lim_{\phi \to 0}[\phi ln(\phi)] = n \lim_{\phi \to 0} \left[\frac{ln(\phi)}{\phi^{-1}} \right],$$

the limit of the numerator is $-\infty$ and the limit of the denominator is ∞. Applying L'Hospital's rule gives:

$$n \lim_{\phi \to 0} \left[\frac{\phi^{-1}}{(-1)\phi^{-2}} \right] = n \lim_{\phi \to 0} [-\phi] \to 0.$$

□

A.3 Proof of Theorem 3.3.1

Proof. The derivation of entropy under this scenario follows the same logic as the previous derivation. It begins by considering the formula of a single record with MAR missingness, which involves a single density function and the log of that function. Using the conditional probability formula, the single density is broken into conditional and marginal densities. The integral is then resolved, resulting in the final formula.

Note that we are able to pull $H(r_{2i}|y_{1i})$ out of the integral over y_1 because entropy focuses on the distribution of $r_{2i}|y_{1i}$ and not the realized values. Since $H(r_{2i}|y_{1i}) = -(1-\phi^*)ln(1-\phi^*) - \phi^* ln(\phi^*)$, we may pull the entropy formula out of the integral over y_{1i}. Our formula so far is

$$H(y_{1i}) + H(y_{2i}|y_{1i}) + H(r_{2i}|y_{1i}). \tag{A.3}$$

However, while the records in the MAR case are independent, they are no longer identically distributed. This is due to the realized values of y_{1i} impacting the value of ϕ^*. Therefore, when we sum Equation A.3 n times to account for the n records, we obtain

$$\begin{aligned} &\sum_{i=1}^{n} \left(H(y_{1i}) + H(y_{2i}|y_{1i}) + H(r_{2i}|y_{1i}) \right) \\ =&\sum_{i=1}^{n} \left(H(y_{1i}) \right) + \sum_{i=1}^{n} \left(H(y_{2i}|y_{1i}) \right) + \sum_{i=1}^{n} \left(H(r_{2i}|y_{1i}) \right) \\ &= nH(y_{1i}) + nH(y_{2i}|y_{1i}) + \sum_{i=1}^{n} \left(H(r_{2i}|y_{1i}) \right), \end{aligned} \tag{A.4}$$

as a framework for entropy of an MAR incomplete bivariate normal dataset.

Once again, the following relationships are clear:

- $y_{1i} \sim N_1(\mu_1, \sigma_1^2)$, therefore $H(y_{1i}) = \frac{1}{2}ln(2\pi e\sigma_1^2)$
- $y_{2i}|y_{1i} \sim N_1(\mu_{2.1}, \sigma_{2.1}^2 = \sigma_2^2(1-\rho^2))$, therefore $H(y_{2i}|y_{1i}) = \frac{1}{2}ln(2\pi e\sigma_2^2(1-\rho^2))$
- $r_{2i} \sim Bern(\phi^*)$, therefore $H(r_2) = -(1-\phi^*)ln(1-\phi^*) - \phi^* ln(\phi^*)$.

Plugging the above formulae into Equation A.4, we obtain

$$\frac{n}{2}ln(2\pi e\sigma_1^2) + \frac{n}{2}ln(2\pi e\sigma_2^2(1-\rho^2)) - \sum_{i=1}^{n}((1-\phi_i^*)ln(1-\phi_i^*) + \phi_i^* ln(\phi_i^*)))$$
$$= \frac{n}{2}ln(2\pi e\sigma_1^2) + \frac{n}{2}ln(2\pi e\sigma_2^2(1-\rho^2)) - \sum_{i=1}^{n}\{(1-\phi_i^*)ln(1-\phi_i^*)\} - \sum_{i=1}^{n}\{\phi_i^* ln(\phi_i^*)\},$$

which is our final formula for incomplete bivariate normal entropy with MAR Bernoulli(ϕ^*) missingness, where $\phi^* = \dfrac{e^{\beta_0+y_{1i}}}{1+e^{\beta_0+y_{1i}}}$. □

A.4 Proof of Theorem 3.3.2

Proof. As before, we are interested in

$$\lim_{\phi^*\to 0}(\sum_{i=1}^{n}\{(1-\phi_i^*)ln(1-\phi_i^*)\} - \sum_{i=1}^{n}\{\phi_i^* ln(\phi_i^*)\}).$$

However, recall that $\phi^* = \dfrac{e^{\beta_0+y_{1i}}}{1+e^{\beta_0+y_{1i}}}$, where y_{1i} are the values in the data set. Therefore, we cannot arbitrarily let $\phi^* \to 0$, since the y_{1i} values are fixed. We must instead examine the behavior of the only arbitrary parameter in ϕ^*, namely β_0.

For ϕ^* to go to zero, all elements ϕ_i^* must also go to zero. This is satisfied using known properties of the logistic function. For a realized value of y_{1i}, ϕ_i^* goes to zero when β_0 goes to negative infinity, regardless of the fixed values of y_{1i}. Therefore, instead of considering $\lim_{\phi^*\to 0}$, we consider $\lim_{\beta_0\to-\infty}$, which results in $\phi_i^* \to 0$ for all i.

Our limit equation is now

$$\lim_{\phi_i^*\to 0}(\sum_{i=1}^{n}\{(1-\phi_i^*)ln(1-\phi_i^*)\} - \sum_{i=1}^{n}\{\phi_i^* ln(\phi_i^*)\}),$$

which goes to zero using the same logic presented in Theorem 3.2.2.

□

A.5 Proof of Theorem 3.4.1

Proof. To begin, the entropy of the i^{th} record in our data is

$$\begin{aligned}
&-\int_{\mathbf{y_{Ai}},\mathbf{y_{Bi}},\mathbf{r_{Bi}}} f(\mathbf{y_{Ai}},\mathbf{y_{Bi}},\mathbf{r_{Bi}})lnf(\mathbf{y_{Ai}},\mathbf{y_{Bi}},\mathbf{r_{Bi}})d(\mathbf{y_{Ai}},\mathbf{y_{Bi}},\mathbf{r_{Bi}}) \\
=&-\int_{\mathbf{y_{Ai}}}\int_{\mathbf{y_{Bi}}}\int_{\mathbf{r_{Bi}}} f(\mathbf{y_{Ai}},\mathbf{y_{Bi}},\mathbf{r_{Bi}})lnf(\mathbf{y_{Ai}},\mathbf{y_{Bi}},\mathbf{r_{Bi}})d\mathbf{r_{Bi}}d\mathbf{y_{Bi}}d\mathbf{y_{Ai}}.
\end{aligned}$$

To separate the joint distribution we use the identity $f(\mathbf{y_{Ai}},\mathbf{y_{Bi}},\mathbf{r_{Bi}}) = f(\mathbf{y_{Ai}})f(\mathbf{y_{Bi}}|\mathbf{y_{Ai}})f(\mathbf{r_{Bi}})$. Again, the distribution of $\mathbf{r_{Bi}}$ is unconditional because we are assuming MCAR. We plug in the identity into the expression above expression to obtain

$$\begin{aligned}
=&-\int_{\mathbf{y_{Ai}}}\int_{\mathbf{y_{Bi}}}\int_{\mathbf{r_{Bi}}} f(\mathbf{y_{Ai}})f(\mathbf{y_{Bi}}|\mathbf{y_{Ai}})f(\mathbf{r_{Bi}})ln\left[f(\mathbf{y_{Ai}})f(\mathbf{y_{Bi}}|\mathbf{y_{Ai}})f(\mathbf{r_{Bi}})\right]d\mathbf{r_{Bi}}d\mathbf{y_{Bi}}d\mathbf{y_{Ai}} \\
=&-\int_{\mathbf{y_{Ai}}}\int_{\mathbf{y_{Bi}}}\int_{\mathbf{r_{Bi}}} f(\mathbf{y_{Ai}})f(\mathbf{y_{Bi}}|\mathbf{y_{Ai}})f(\mathbf{r_{Bi}})\left[lnf(\mathbf{y_{Ai}})+lnf(\mathbf{y_{Bi}}|\mathbf{y_{Ai}})+lnf(\mathbf{r_{Bi}})\right]d\mathbf{r_{Bi}}d\mathbf{y_{Bi}}d\mathbf{y_{Ai}}
\end{aligned}$$

$$\begin{aligned}
=&-\int_{\mathbf{y_{Ai}}}\int_{\mathbf{y_{Bi}}}\int_{\mathbf{r_{Bi}}} f(\mathbf{y_{Ai}})f(\mathbf{y_{Bi}}|\mathbf{y_{Ai}})f(\mathbf{r_{Bi}})lnf(\mathbf{y_{Ai}})d\mathbf{r_{Bi}}d\mathbf{y_{Bi}}d\mathbf{y_{Ai}} \\
&-\int_{\mathbf{y_{Ai}}}\int_{\mathbf{y_{Bi}}}\int_{\mathbf{r_{Bi}}} f(\mathbf{y_{Ai}})f(\mathbf{y_{Bi}}|\mathbf{y_{Ai}})f(\mathbf{r_{Bi}})lnf(\mathbf{y_{Bi}}|\mathbf{y_{Ai}})d\mathbf{r_{Bi}}d\mathbf{y_{Bi}}d\mathbf{y_{Ai}} \\
&-\int_{\mathbf{y_{Ai}}}\int_{\mathbf{y_{Bi}}}\int_{\mathbf{r_{Bi}}} f(\mathbf{y_{Ai}})f(\mathbf{y_{Bi}}|\mathbf{y_{Ai}})f(\mathbf{r_{Bi}})lnf(\mathbf{r_{Bi}})d\mathbf{r_{Bi}}d\mathbf{y_{Bi}}d\mathbf{y_{Ai}}
\end{aligned}$$

$$\begin{aligned}
=&-\int_{\mathbf{y_{Ai}}}\int_{\mathbf{y_{Bi}}} f(\mathbf{y_{Ai}})f(\mathbf{y_{Bi}}|\mathbf{y_{Ai}})lnf(\mathbf{y_{Ai}})\left(\int_{\mathbf{r_{Bi}}} f(\mathbf{r_{Bi}})d\mathbf{r_{Bi}}\right)d\mathbf{y_{Bi}}d\mathbf{y_{Ai}} \\
&-\int_{\mathbf{y_{Ai}}}\int_{\mathbf{y_{Bi}}} f(\mathbf{y_{Ai}})f(\mathbf{y_{Bi}}|\mathbf{y_{Ai}})lnf(\mathbf{y_{Bi}}|\mathbf{y_{Ai}})\left(\int_{\mathbf{r_{Bi}}} f(\mathbf{r_{Bi}})d\mathbf{r_{Bi}}\right)d\mathbf{y_{Bi}}d\mathbf{y_{Ai}} \\
&-\int_{\mathbf{y_{Ai}}}\int_{\mathbf{y_{Bi}}} f(\mathbf{y_{Ai}})f(\mathbf{y_{Bi}}|\mathbf{y_{Ai}})\left(\int_{\mathbf{r_{Bi}}} f(\mathbf{r_{Bi}})lnf(\mathbf{r_{Bi}})d\mathbf{r_{Bi}}\right)d\mathbf{y_{Bi}}d\mathbf{y_{Ai}}.
\end{aligned}$$

The first and second integral over $\mathbf{r_{Bi}}$ equal one; the third is the entropy of the distribution of $\mathbf{r_{Bi}}$, which will be denoted $H(\mathbf{r_{Bi}})$. The result is:

$$
\begin{aligned}
&= -\int_{\mathbf{y_{Ai}}}\int_{\mathbf{y_{Bi}}} f(\mathbf{y_{Ai}})f(\mathbf{y_{Bi}}|\mathbf{y_{Ai}})lnf(\mathbf{y_{Ai}})d\mathbf{y_{Bi}}d\mathbf{y_{Ai}} \\
&\quad -\int_{\mathbf{y_{Ai}}}\int_{\mathbf{y_{Bi}}} f(\mathbf{y_{Ai}})f(\mathbf{y_{Bi}}|\mathbf{y_{Ai}})lnf(\mathbf{y_{Bi}}|\mathbf{y_{Ai}})d\mathbf{y_{Bi}}d\mathbf{y_{Ai}} \\
&\quad + H(\mathbf{r_{Bi}})\int_{\mathbf{y_{Ai}}}\int_{\mathbf{y_{Bi}}} f(\mathbf{y_{Ai}})f(\mathbf{y_{Bi}}|\mathbf{y_{Ai}})d\mathbf{y_{Bi}}d\mathbf{y_{Ai}}
\end{aligned}
$$

$$
\begin{aligned}
&= -\int_{\mathbf{y_{Ai}}} f(\mathbf{y_{Ai}})lnf(\mathbf{y_{Ai}})\left(\int_{\mathbf{y_{Bi}}} f(\mathbf{y_{Bi}}|\mathbf{y_{Ai}})d\mathbf{y_{Bi}}\right)d\mathbf{y_{Ai}} \\
&\quad -\int_{\mathbf{y_{Ai}}} f(\mathbf{y_{Ai}})\left(\int_{\mathbf{y_{Bi}}} f(\mathbf{y_{Bi}}|\mathbf{y_{Ai}})lnf(\mathbf{y_{Bi}}|\mathbf{y_{Ai}})d\mathbf{y_{Bi}}\right)d\mathbf{y_{Ai}} \\
&\quad + H(\mathbf{r_{Bi}})\int_{\mathbf{y_{Ai}}} f(\mathbf{y_{Ai}})\left(\int_{\mathbf{Y_2}} f(\mathbf{y_{Bi}}|\mathbf{y_{Ai}})d\mathbf{y_{Bi}}\right)d\mathbf{y_{Ai}}.
\end{aligned}
$$

The first and third integral over $\mathbf{y_{Bi}}$ equal one; the second is the entropy of the distribution $\mathbf{y_{Bi}}|\mathbf{y_{Ai}}$, which we will denote $H(\mathbf{y_{Bi}}|\mathbf{y_{Ai}})$. The result is:

$$
\begin{aligned}
&= -\int_{\mathbf{y_{Ai}}} f(\mathbf{y_{Ai}})lnf(\mathbf{y_{Ai}})d\mathbf{y_{Ai}} \\
&\quad + H(\mathbf{y_{Bi}}|\mathbf{y_{Ai}})\int_{\mathbf{y_{Ai}}} f(\mathbf{y_{Ai}})d\mathbf{y_{Ai}} \\
&\quad + H(\mathbf{r_{Bi}})\int_{\mathbf{y_{Ai}}} f(\mathbf{y_{Ai}})d\mathbf{y_{Ai}}.
\end{aligned}
$$

The second and third integrals equal one, while the third is the entropy of the distribution of $\mathbf{y_{Ai}}$, which we will denote $H(\mathbf{y_{Ai}})$. Therefore, the entropy of one record of an incomplete data set as defined above is

$$H(\mathbf{y_{Ai}}) + H(\mathbf{y_{Bi}}|\mathbf{y_{Ai}}) + H(\mathbf{r_{Bi}}). \tag{A.5}$$

The records in the MCAR case are independent and identically distributed. Therefore, we may sum Equation A.5 n times to account for the n iid records, and obtain

$$nH(\mathbf{y_{Ai}}) + nH(\mathbf{y_{Bi}}|\mathbf{y_{Ai}}) + nH(\mathbf{r_{Bi}}) \tag{A.6}$$

as a framework for entropy of MCAR incomplete p-variate normal data where there is a block of missing values.

The following are clear:

- $\mathbf{y_{Ai}} \sim N_k(\mu_\mathbf{1}, \Sigma_1)$, therefore $H(\mathbf{y_{Ai}}) = \frac{k}{2}ln(2\pi e) + \frac{1}{2}ln(|\Sigma_1|)$
- $\mathbf{y_{Bi}}|\mathbf{y_{Ai}} \sim N_{p-k}(\mu_{2|1}, \Sigma_{2|1} = \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$, therefore $H(\mathbf{y_{Bi}}|\mathbf{y_{Ai}}) = \frac{p-k}{2}ln(2\pi e) + \frac{1}{2}ln(|\Sigma_{2|1}|)$
- $H(\mathbf{r_{Bi}})$ is the entropy of the $(p-k)$-variate Bernoulli distribution with probabilities of success ϕ.

  If we plug these values into the above, we obtain our final formula.

□

A.6 Proof of Theorem 3.5.1

Proof. We must first address the definition of the parameter which governs the missingness mechanism. In the p-variate MCAR case, $\mathbf{r}_{B,i} \sim Bern_{p-k}(\phi)$, where ϕ was a vector of probabilities independent of observed or unobserved data. However, in the bivariate MAR case, ϕ changed to ϕ^*, where $\phi_i^* = \dfrac{e^{\beta_0+y_{1i}}}{1+e^{\beta_0+y_{1i}}}$.

The first concern is how to obtain a block of missing values. In other words, how to ensure all $n_1 + 1$ to n records in $\mathbf{Y_B}$ will be missing. We accomplish this by basing the $p-k$ vectors of $\mathbf{R_B}$ on a single column of $\mathbf{Y_A}$. Namely, $\mathbf{R_B} = (\mathbf{r_{k+1}}, ..., \mathbf{r_p})$ has $p-k$ n-dimensional vectors, but every one of those $p-k$ vectors is the same n-dimensional vector, ϕ^*, which is obtained using:

$$\phi^* = \frac{e^{\beta_0+\mathbf{y}_{A,j}}}{1+e^{\beta_0+\mathbf{y}_{A,j}}},$$

where $\mathbf{y}_{A,j}$ denotes a single, pre-determined vector of $\mathbf{Y_A}$.

Now that the missingness mechanism is addressed, we may calculate p-variate MAR entropy. The calculations follow those for MCAR p-variate entropy, excepting that the notation $\mathbf{r}_{B,i}$ becomes $\mathbf{r}_{B,i}|\mathbf{y}_{A,ij}$, since the missingness depends on the i^{th} value of the vector $\mathbf{y}_{A,j}$.

Our final component-wise p-variate MAR entropy formula is

$$H(\mathbf{x_i}) = H(\mathbf{y_{Ai}}) + H(\mathbf{y}_{Bi}|\mathbf{y}_{Ai}) + H(\mathbf{r}_{B,i}|\mathbf{y}_{A,ij}),$$

and the entropy for the entire data set is

$$H(\mathbf{X}) = nH(\mathbf{y_{Ai}}) + nH(\mathbf{y}_{Bi}|\mathbf{y}_{Ai}) + \sum_{i=1}^{n} H(\mathbf{r}_{B,i}|\mathbf{y}_{A,ij}).$$

□

A.7 Proof of Theorem 4.2.1

Proof. Consider γ_g, the probability of a unique record x_i being in class g, where $i \in 1, ..., N_u$, and N_u is the number of unique records. For all $i \neq g, \gamma_g = 0$, because record i must be in class i (by our assumption). Since $\sum_{g=1}^{G} \gamma_g = 1$, it remains that $\gamma_g = 1$ when $i = g$.

Therefore, we eliminate the summation over G and the term γ_g. Our formula becomes:

$$Entropy = -\sum_{N} \alpha_i \times ln(\alpha_i),$$

where

$$\alpha_i = \prod_{K} \left[\prod_{O^k} \left(\pi_{k,o}^{x^u_{i,k,o}} \right) \right]$$

Now that class g corresponds to record i (from our assumption), there is only one record being considered in α_i. Let us consider each possible value for $x^u_{i,k,o}$. If $x^u_{i,k,o} = 1$, then the i^{th} record has the o^{th} value in the k^{th} variable. For every record i and variable k, $(x^u_{i,k,1}, ..., x^u_{i,k,O^k})$ can only equal 1 for one of the values, and must equal 0 for the other values. This is because there is only a single number in $x^u_{i,k}$, and therefore can only take one value of o.

We are left with $x^0 = 1 \forall x$, even $x = 0$. Therefore, all values of $\pi_{k,o}^{x^u_{i,k,o}}$ which have an exponent of zero will equal one. There will be one single value of $\pi_{k,o}$ where its exponent $x^u_{i,k,o}$ equals one instead of zero. In this case, the corresponding value of $\pi_{k,o}$ must also equal one. This is due to the fact that $\sum_{O^k} \pi_{k,o} = 1$. This is also due to the fact that the probability of a single number being equal to its value is one. Therefore, $\prod_{O^k} \left(\pi_{k,o}^{x^u_{i,k,o}} \right) = 1$ for fixed k.

The results given above occur for each value of K. Thus,

$$\alpha_i = \prod_{K} \left[\prod_{O^k} \left(\pi_{k,o}^{x^u_{i,k,o}} \right) \right] = \prod_{K} [1] = 1.$$

Finally, we have

$$Entropy = -\sum_{N} \alpha_i \times ln(\alpha_i) = -\sum 1 \times ln(1) = 0,$$

the smallest value entropy can take. □