HHS Public Access

Author manuscript

Stat Methodol. Author manuscript; available in PMC 2017 September 01.

Published in final edited form as:

Stat Methodol. 2016 September ; 32: 107–121. doi:10.1016/j.stamet.2016.04.004.

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

Chantal Larose[a], **Ofer Harel**[b,*], **Katarzyna Kordas**[c], and **Dipak K. Dey**[b]

[a]School of Business, State University of New York at New Paltz, New Paltz, NY, USA

[b]Department of Statistics, University of Connecticut, Storrs, CT, USA

[c]University of Bristol, Senate House, Tyndall Avenue, Bristol BS8 1TH, UK

Abstract

Latent class analysis is used to group categorical data into classes via a probability model. Model selection criteria then judge how well the model fits the data. When addressing incomplete data, the current methodology restricts the imputation to a single, pre-specified number of classes. We seek to develop an entropy-based model selection criterion that does not restrict the imputation to one number of clusters. Simulations show the new criterion performing well against the current standards of AIC and BIC, while a family studies application demonstrates how the criterion provides more detailed and useful results than AIC and BIC.

Keywords

entropy; latent class analysis; missing data; model selection; multiple imputation

1. Introduction

Latent class analysis (LCA) [1] is a model-based clustering methodology for categorical data. Variables in a data set are sometimes called “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.

Our goal is 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

*Corresponding author. ofer.harel@uconn.edu (Ofer Harel).

number of clusters. To do so, we first 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 construct our criterion.

There are methods for clustering categorical data using entropy [2, 3], 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 [4]. There are also ways to cluster incomplete categorical data using multiple imputation and latent class analysis (LCA) [5] using the fraction of missing information (FMI) as an LCA model selection criterion [6]. However, the methodology sets a fixed number of classes prior to imputation. 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 allows the use of LCA with multiple imputation without having to set a number of clusters beforehand.

Entropy [4, 7, 8] is a way to measure the variability, or chaos, in a stochastic system. Entropy depends on the probability density or mass function of the variables or model. One can calculate entropy of a mixture model, and thus entropy has been used as a model selection criterion in clustering scenarios. Typically, it is combined with the log-likelihood [9, 10], although it has been used on its own [3, 2]. Between two competing model based cluster solutions, the one with lower entropy means there is less variability within each cluster, and thus the clusters are more homogeneous. We are interested in looking at entropy itself as a model selection criterion.

Model selection in clustering also chooses the number of clusters. Existing model selection criteria have penalties to avoid choosing too many clusters, and overfitting the data. To determine what sort of penalty to introduce to our new model selection criterion, we need to understand how entropy of an LCA model behaves as the number of classes increases. Thus, 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 [11] 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 equaling zero when the number of classes approaches the number of unique 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 [11]. In addition, the performance of BIC and AIC breaks down after multiply imputing data sets in a regression context [12]. 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 paper is organized as follows. Sections 2, 3, and 4 discuss latent class analysis and model selection, entropy, and missing data respectively. Section 5 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 6 describes the methodology of Harel et al [5], and how we propose to extend the methodology. Section 7 presents our simulation study, in which we compare our entropy-based criterion to AIC and BIC. Section 7.2 demonstrates an application of our entropy-based criterion, and compared the results to those obtained by AIC and BIC. Section 8 wraps up the paper with our conclusions and directions for future work.

2. Latent Class Analysis

Clustering is the categorization of records into bunches (e.g., clusters) in order to describe grouping patterns in the data set. Latent Class Analysis (LCA) is a model-based clustering method which treats cluster (e.g., class) membership as a missing (e.g., latent) variable, used when the observed (e.g., manifest) variables are categorical [13, 1, 11, 14]. LCA assumes that, conditioned on the class membership, the manifest variables are independent. This assumption is typically referred to as the conditional independence assumption [13].

We utilize the notation from Harel et al [5]. The probability that a particular record belongs to a particular, fixed class is

$$f(\mathbf{y_i}|\pi_g)=\prod_{K}\left(\prod_{O_k}\pi_{k,o|g}^{y_{i,k,o}}\right), \quad (1)$$

where the index $i = 1, \ldots, N$ refers to records, $k = 1, \ldots, K$ refers to manifest variables, $o = 1, \ldots O_k$ refers to the possible values for each variable k, and $g = 1, \ldots, G$ refers to classes. The value $\boldsymbol{\pi}_{k,o|g}$ is the probability that a record i has value o while belonging to class g, and $y_{i,k,o}$ is an indicator that the i^{th} record has the o^{th} value in variable k. When we consider Equation 1 over all possible classes, we get the probability of a single record,

$$f(\mathbf{y_i}|\pi,\gamma)=\sum_{g=1}^{G}\gamma_g\left[\prod_{K}\left(\prod_{O_k}\pi_{k,o|g}^{y_{i,k,o}}\right)\right], \quad (2)$$

where the value $\boldsymbol{\gamma}_g$ is the probability of any randomly selected record belonging to class g.

When all observations are fully observed, and the number of desired classes is fixed, estimates for parameters and $\boldsymbol{\gamma}$ are $\boldsymbol{\pi}$ obtained via likelihood maximization. Once the parameter estimates are obtained, records are assigned to class g, where g is the class to which the record belongs with the largest probability [13].

2.1. Model Selection Criteria

As LCA is a model-based approach to clustering, model selection criteria are used to determine whether a particular LCA model describes the data in an accurate and useful way.

One popular criterion is the Bayesian Information Criterion (BIC) [15],

$$2 \times \ln\,[p(\boldsymbol{Y}|M_g)] \approx 2 \times \ln\,\left[p(\boldsymbol{Y}|\hat{\theta}_g, M_g)\right] - \nu_g \times \ln(n) = BIC_g, \quad (3)$$

where the data **Y** is fit to a model M_g, which contains ν_g parameters estimated by the $\hat{\theta}_g$ maximum likelihood estimates. The lowest BIC value indicates the best model. A closely-related criterion, Akaike Information Criterion (AIC) [16] uses the same formula as above, only changing $\nu_g \times ln(n)$ to $2 \times \nu_g$.

The entropy of a model is also used as a model selection criterion, either by itself or together with other statistics. The details are in the following section.

3. Entropy

Entropy as a general concept [7, 8] quantifies the uncertainty in a random variable or a model via the probability density or probability mass function of that random variable. The measure of uncertainty used in this paper is Shannon entropy [7], which defines the uncertainty of a discrete random variable **X** with distribution $p(\mathbf{X})$ as

$$H(\mathbf{X}) = -\sum p(\mathbf{X}) \times ln\,[p(\mathbf{X})]. \quad (4)$$

A common extension of Shannon entropy to continuous variables is differential entropy [8], where the summation is replaced by the integral over **X**. Other entropy measurements which extend Shannon entropy to the continuous case include [17].

Shannon entropy (referred to as “entropy” from this point onward) has many facets that make intuitive sense for using it to quantifying uncertainty. Entropy is non-negative, so the measure of variability is minimized at zero. Entropy is additive, so the entropy in two or more events will be at least the entropy in a single item. Entropy is maximized when all probabilities $p(\mathbf{X})$ are equal, since an outcome is most difficult to predict when all outcomes are equally likely. Entropy of a constant is zero, since no randomness is introduced. Finally, entropy is symmetric, meaning that a reordering of independent outcomes will not impact the entropy of the outcome itself [7].

Describing the entropy of a data set requires describing the entropy of the model which describes the records in the data set. This is due to the random nature of statistical models, which have variation and thus have entropy, compared to the static nature of observed values, which have no variation and thus no entropy. Therefore, we do not focus on the observed values of any particular record, but on the distribution of the variables which make up the record. For example, consider a data set made up of height and weight measurements.

If we allow height and weight of the records to be described by independent draws from the same bivariate normal model, then the entropy of every record in the data set is the entropy of the bivariate normal model, regardless of the realized value of the record.

Since we can quantify the entropy of a distribution, we can quantify the entropy of a mixture of distributions. This allows for entropy calculation of mixture models, such as LCA, quantified by Dias and Vermunt [18]

$$H(\alpha) = -\sum_{i=1}^{N} \alpha_{ig} \times \ln\,[\alpha_{ig}], \tag{5}$$

where $\alpha_{ig} = \sum_{g=1}^{G} \gamma_g \left[\prod_{K} \left(\prod_{O_k} \pi_{k,o|g}^{y_{i,k,o}} \right) \right]$, the distribution of the records into latent classes. Equation 5 is obtained by using α_{ig} in place of $p(\mathbf{X})$ in Equation 4. In other words, including the distribution of records as the probabilities when calculating the entropy of the variable.

Entropy has been used as a model selection criteria, in addition to the model selection criteria outlined previously. Examples includes when entropy is used on its own [3, 2] or in conjunction with the log-likelihood [10].

4. Missing Data

The world is full of incomplete data sets. For example, a study of 57 studies of HIV found that the average amount of missing data was 26% (median 23%), and that 74% of those studies used the outdated complete case analysis (CCA) method to address the missing values [19]. CCA occurs when incomplete records are excluded from an analysis. If there is any pattern to the missing values (common in real-world data [20]), that pattern is lost during CCA. If there happens to be no pattern, the data set is still reduced, and thus wastes the resources spent in collecting the data, and potentially inflating the standard errors of any subsequent estimates.

4.1. Three Missingness Mechanisms

Rubin [21, 22] described three mechanisms which quantify the three possible patterns of missingness in a data set. To present these mechanisms, we introduce notation common to the missing data field.

The data is written $\mathbf{Y} = (\mathbf{Y_{obs}}, \mathbf{Y_{mis}})$ to signify that the data can be bifurcated into observed values and missing values. The model for the data is written $P(\mathbf{Y}|\theta)$, where θ is a vector of parameters. Another matrix $\mathbf{R}$ is made as a reflection of the missingness in $\mathbf{Y}$, where every element of $\mathbf{R}$ is either zero or one, dependent on the missing or observed status of the corresponding element in $\mathbf{Y}$. The parameter which governs $\mathbf{R}$ is ϕ. All together, the model for $\mathbf{R}$ is $P(\mathbf{R}|\mathbf{Y_{obs}}, \mathbf{Y_{mis}}, \phi)$.

Using the above notation, we can describe the missingness mechanisms. The first, Missing Completely at Random (MCAR), describes $\mathbf{R}$ using $P(\mathbf{R}|\mathbf{Y_{obs}}, \mathbf{Y_{mis}}, \phi) = P(\mathbf{R}|\phi)$, where

nothing but the parameter describes the missingness. The second, Missing At Random (MAR), changes the model to $P(\mathbf{R}|\mathbf{Y}_{\mathbf{obs}},\mathbf{Y}_{\mathbf{mis}}, \phi) = P(\mathbf{R}|\mathbf{Y}_{\mathbf{obs}}, \phi)$, where the observed data and the parameter describe the missingness. The third and final mechanism, Missing Not At Random, extends the model to $P(\mathbf{R}|\mathbf{Y}_{\mathbf{obs}},\mathbf{Y}_{\mathbf{mis}}, \phi)$, where unobserved values also help describe the missingness.

4.2. Multiple Imputation

While a myriad of methods exist to address incomplete data, many have significant drawbacks. For example, complete case analysis requires deleting all incomplete records, thus sacrificing the partially observed information along with the missing values. More nuanced methods may use the estimates of variable means, or regress one variable on another, to obtain estimates of missing values. These approaches risk deflating the variability in the data, and may overemphasize relationships between particular variables. Methods which generate a single number to fill in a missing value treat the simulated value as a real, fixed data point, thus ignoring the variability in the imputation model. Multiple imputation, by contrast, retains all incomplete records while maintaining the integrity of relationships among variables and considering all sources of variability encountered in the data imputation [23, 24, 25].

Multiple imputation (MI) [26, 22, 27] is a three-stage, simulation-based procedure which fills in every missing value multiple times. The multiple substitutions create multiple data sets, which are analyzed individually and have their results combined to form a single coherent point estimate, with a standard error which accounts for both inter- and intra-imputation variability. What follows is a general overview of MI; details regarding how MI combines with LCA are given in the following section.

The Imputation stage begins MI by drawing $M > 1$ simulated values for each missing value from $\mathbf{Y}_{\mathbf{mis}} \sim P(\mathbf{Y}_{\mathbf{mis}}|\mathbf{Y}_{\mathbf{obs}},\mathbf{R})$. Different data sets are created by substituting each of the different values in turn. The result is M data sets, all of which contain different data values where the missing values were. Most software approaches missingness under the MAR assumption [23]. We continue this practice, noting that sensitivity analysis can be applied in future to examine the effect of values which are missing missing based on their own value.

The Analysis stage takes each of the M data sets and analyzes it as if no missing values had been found. Point estimates for parameters of interest, along with their standard errors, are obtained for all M data sets. Let these point estimates be written $\hat{Q}_m$, all of which estimate a parameter Q, and where $m = 1, \ldots, M$. Let the corresponding standard errors be written U_m. Following the recommendation by [28], we include all variables in both the imputation and analysis stages, which allows our methodologies for imputation and analysis to be congenial by the definition in [29].

The Combination stage takes the point estimates and standard errors from the Analysis stage and combines them into a single point estimate and standard error. The combination of these quantities follow Rubin's Rules [22]. Namely, the final point estimate is $\bar{Q}=\frac{1}{M}\sum \hat{Q}_m$,

with variance $T=\bar{U}+(1+\frac{1}{M})B$. The quantity $\bar{U}$ is the intra-imputation variation, $\bar{U}=\frac{1}{m}\sum U_m$, and B is the inter-imputation variation, $B=\frac{1}{M-1}\sum(\hat{Q}_m - \text{Çň})^2$. To perform hypothesis tests or confidence intervals for $\bar{Q}$, the statistic $\frac{Q-\text{Çň}}{T^{1/2}}\sim t_\nu$ is obtained, where $\nu=(m-1)\left[1+\frac{\bar{U}}{(1+m^{-1})B}\right]^{1/2}$.

4.3. Imputation via LCA

It is possible to obtain imputed values via an LCA model via an iterative Bayesian method, which imputes the latent class membership vector and the missing data values iteratively [5]. Briefly, the EM algorithm [30] obtains starting values, after which data augmentation allows draws of the necessary parameter values. Specifically, the LCA parameters (which are at first estimated using the EM algorithm) quantify how likely it is that each record belongs to each class. By assigning records to the class they have the highest probability of being a member, a class membership vector is created. Using the newly-imputed class membership information, one may create values of manifest variables within each class by choosing which categorical value is most likely. In this way, multiply imputed data sets with class membership can be created. For an application of this method to two-stage MI, see [5].

5. Limiting Behavior of LCA Entropy

To calculate the entropy of an LCA model [11, 18], we must consider the number of unique records possible instead of the number of records total. We limit our focus to unique records because the LCA formula describes the likelihood of each unique records. The probability of each unique record can, in turn, be considered a weighted sum of the individual records' probabilities.

To begin, let $\mathbf{Y}_{N\times K}$ represent a matrix of data, and $\mathbf{X}_{N\times K}$ be the subset of $\mathbf{Y}$ which contains only unique records. The entropy of an LCA model will use the probability mass function of unique records, $p(\mathbf{X} = \mathbf{x})$, where $\mathbf{x}$ is an individual unique record, thus:

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

The number of potentially unique records changes based on the number of variables K and the number of values each variable can take. Let the number of potential values for variable K be O_k. Then, the number of possible unique records is $N^u=\prod_{k=1}^{K} O_k$.

The entropy of an LCA model as the number of classes goes to the number of unique records has not, to our knowledge, been examined. To that end, we present a proof to describe that behavior. Using the work of Fruhwirth-Schnatter [11] and Dias and Vermunt [18], altered to consider only unique records, we begin with $p(\mathbf{X} = \mathbf{x_i})$:

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

The above $p(\mathbf{X} = \mathbf{x_i})$ is the distribution of the unique records in the data into latent classes, where $i \in (1, \ldots, N^u)$ denotes the possible unique records; $g \in (1, \ldots, G)$ denotes the classes; G denotes the maximum number of classes; $k \in (1, \ldots, K)$ denotes the variables; K denotes the maximum number of variables; $o \in (1, \ldots, O_k)$ denotes the categories specific to variable k; $\boldsymbol{\gamma}_g$ denotes the probability of being in class g; $\boldsymbol{\pi}_{k,o|g}$ denotes the probability that a record has value o in variable k, given it is in class g; and $x_{i,k,o}^u$ equals 1 if the variable k for unique record i has the o^{th} value.

From the above set-up, we obtain the following theorem:

Theorem 5.1

Assume that the i^{th} unique record is in the ith class, where $i = 1, \ldots, N^u$. The limit of the entropy of an LCA model as the number of classes approaches the number of unique records is

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

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.

Our take-awake from this theorem is that letting the number of classes in an LCA model approach the number of unique records will minimize the entropy of that model. This has direct implications for the remainder of our project. Minimum entropy is typically desired, but assigning each unique record its own class is tantamount to overfitting the data. Therefore, we must construct a penalty function into our entropy-based model selection criterion.

6. Incomplete LCA Methodology and Extensions

In this section, we detail the existing combination of MI and LCA, followed by our proposed extensions of the method.

6.1. Incomplete LCA Methodology

In the existing methodology (e.g. [5, 6]), the number of classes is chosen at the outset, before any values are imputed. Then, the EM algorithm is used to find the "best" starting values for parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\pi}$. From there, multiple imputations of both missing manifest values and class membership are obtained, following the steps outlined in Section 4.3. Once

multiple imputations are obtained, the parameter estimates are combined following Rubin's Rules (Section 4.2). Their averages describe a single LCA model, from which it is possible to obtain class membership for every record. The model, and its class membership assignments, may be analyzed as any LCA model for complete data would be.

Since the number of classes for the LCA model is fixed before the imputations begin, the resulting LCA model has that number of classes. In other words, there is no model selection taking place.

6.2. Extensions

In the above methodology, the selection of the number of classes occurs before imputation of data values, and thus it is implied that it occurs on a complete-case data set. We wish to free the methodology from the restrictions of (a) having to specify a single number of classes, and (b) doing so with only a complete-case data set to guide the decision. To this end, one may run the methodology multiple times on the incomplete data set, varying the number of classes in each run. The result is a series of LCA models, and the question becomes: Which model to use?

The classic model section criterion for LCA is BIC, which would choosen the model with the minimum of BIC values. The application of BIC to imputed results has been shown in a regression context, but demonstrated that the usefulness of BIC deteriorated when used with multiply imputed data [12]. Therefore, there is opportunity to construct a new model selection criterion, which seeks to ourperform BIC in the context of multiply imputed data.

Our new approach to selecting a multiply imputed LCA model begins by considering a range of possible classes. For each number of potential classes, impute missing manifest values and class membership as explained previously. However, instead of combining parameter estimates, calculate the entropy for each multiply imputed model. Thus, if there are M imputations for a set number of classes g, there will be M entropy estimates; one for each model. The average of these entropy values is the entropy statistic for that number of classes, g.

We have shown that the entropy of an LCA model goes to zero as the number of classes approaches the number of unique records. Therefore, we expect entropy values for models with higher values of g to be smaller than models with smaller values of g. To avoid overfitting the data, a penalty term must be introduced. We begin by calculating the relative change in entropy over all G classes, $\Delta H_g = \frac{H_g - H_{g+1}}{H_{g+1}}$. The trimmed standard deviation of the ΔH_g values, σ^*, is calculated. The trimmed standard deviation does not include the entropy in the two smallest numbers of classes. Thus, it quantifies the "tailing off" entropy values, instead of all entropy values.

If we had calculated the entropy values of all classes, the first few smallest classes, are likely to have more different entropy values than those classes which followed. In this case, the trimmed standard deviation avoids being infulenced by these unusual values. If the first few

classes do not have unusual values of entropy, the trimmed standard deviation still captures the overall variation in the values, and thus still serves its purpose.

We choose the "best" number of classes by observing when ΔH_g first falls below the threshold $\sigma_i^* = \sigma^*/t$, where t is a constant. An examination of which constant t to use occurs in the next section.

7. Simulation and Application

The simulations presented in this section quantify the ability of our new model selection criterion to identify the correct number of classes, and compares its performance to that of AIC and BIC. The comparison between our method, AIC, and BIC continues in the data application, where the classes themselves are examined to determine how useful the results are from each of the three methods.

7.1. Simulation

The simulation study seeks to determine whether our proposed entropy-based criterion outperforms AIC and BIC. The performance of each of the criteria are measured by their ability to choose the correct number of classes. The averages of the criteria are calculated in order to describe the overall performance of each criterion. The arithmetic and geometric means of AIC and BIC are also calculated, following the process in [12].

The foundation of our simulated data is the Zoo data set from the UCI Machine Learning Repository [31]. 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 take values zero or one, corresponding to a no or yes answer to whether record i has trait k. There is a class membership variable, Type, which correctly specifies which records belong to which class. Type breaks the data set into seven classes; thus, we know the true number of classes to be seven.

As there are only 101 records in the data, small sample size issues may occur when we consider splitting the data into seven classes. To avoid these issues, we generate $N = 1000$ simulated observations from an LCA model fit to the original data. Doing so will preserve the patterns inherent in the original data, while allowing us to work with a larger sample size. Referencing the Type variable, and its seven values, we fit an LCA model with seven classes. In order to run an LCA model, we run the [5] methodology on the data set with only one imputation. The LCA model parameters thus obtained

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

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

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

For every repetition of our simulation study, a new data set of a thousand records, seven variables, and seven classes was generated. The number of unique records in each data set is $2^7 = 128$.

Once the data is obtained, missingness was added using an MAR mechanism. If a record denoted that the animal did not produce milk, then the record's value for Predator was MCAR with probability ϕ_A. If instead the record denoted that the animal did produce milk, the corresponding value for Predator was MCAR with probability ϕ_B. The values of the two ϕ parameters were such that the total percent of missing values approximated 50%, 25%, and 10%.

Finally, different values of the threshold σ_t^* were studies. Values of t considered were $t = 2$, 4, 6, 8, 12, 14, and 20.

7.1.1. Simulation Results—Figure 1 displays the number of classes determined as the best by each criteria. AIC chose four classes, and BIC chose three; both criteria underestimate the number of classes in the data, illustrating how AIC and BIC performance deteriorates when applied to a multiply imputed data set. Our entropy-based criterion chose a number of classes that was much closer to the true number of classes. The accuracy of the average number of classes chosen by the entropy criterion was much higher when only 10% of the values were missing, as may be expected intuitively. However, when 25% and 50% of the values are missing, the average number of classes chosen by the entropy-criterion was only one away (six instead of seven), and the correct number of classes was always in the middle fifty percent of classes chosen.

We examine only entropy-based criteria in Table 1, as Figures 1a through 1c showed that AIC and BIC do not select the best number of classes and have nearly no variability. The table shows the the percent of choosing the correct number of classes (*Pct.*7), the percent of

choosing classes within one of the correct number (*Pct.*6 − 8), and the percent of simulation runs which never met the threshold (*Pct.NA*), for each percent of missing values and each entropy-based model selection criterion examined. Percent correct and percent near-correct in Table 1 are calculated relative to the number of non-NA repetitions per setting.

From Table 1, we can see the rate of correctly chosing the exact number of classes varied considerably, but is maximized for σ_4^* for all three levels of Percent Missingness (%*M is.*). The rate of the choice being within one of the correct number was maximized at σ_2^* for 10% and 25% missing, with σ_4^* coming in a close second; and σ_6^* for 50% missing, with σ_2^* and σ_4^* close behind. The percent of repetitions who had no entropy value below the threshold was maximized at σ_{20}^*, as expected. In other words, as the threshold becomes more and more strict, fewer and fewer cases meet the threshold.

We take these results as evidence that our method is outperforming AIC and BIC. We also consider the thresholds σ_2^* or σ_4^* to be the best for our entropy-based criterion.

7.2. Application

Rink et al [32] analyzed a subset of data collected by Kordas et al [33] which examined children's home environment, their parents' background, and the children's health status. The original data, described in Kordas et al [33], has 193 variables and 109 records. Questions include which parent the child lives with, whether the child has various health problems (hepatitis, anemia, etc.), the division of labor between the parents, and other topics related to the health and lifestyle of the child and parents.

Out of the 193 variables available, we look at the following: father's education (in years), parents' marital status (married, divorced, living together unmarried), mother's intelligence quotient (IQ), amount of hemoglobin in the child's blood (g/dL), socio-economic status (SES) of the child's home, HOME score, amount of manganese (Mn) in the child's hair (ug/g), and age of the child (in months). Socio-economic status is a summary of responses from twelve questions which inquired as to the furnishings of the child's home, including whether there was a refridgerator, television, etc. The HOME score was obtained via a survey of the home environment [32, pg. 48].

Figure 2 displays histograms of all observed values across the variables are are analyzing. All are continuous, except marital status, which has three values. Father's Education is skewed right; Managese Level is severely skewed right; Hemoglobin level, HOME score, and Mother's IQ are all mildly skewed left. SES and Child's Age are all approximately symmetric.

Table 2 gives us an idea of where the missing values are most commonly found in this data set. The frequency (*Freq.*) and percent (*Pct.*) of missing values for each variable are shown. The only complete variable is Child's Age, whereas the variable with the most missing values is Mother's IQ (23 values missing, 21.1%). Most other variables have about 8% – 10% missing values, except Marital Status (1 missing, 0.9%) and SES (2 missing, 1.8%). If we were to remove all incomplete records from the data, our sample size would reduce from 109 to 74, resulting in almost a third of records lost.

As observed, none of the variables are binary. In order to apply LCA analysis to this data set, we must bifurcate the variables' values into values of zero and one. This separation of values occured as follows. Father's Education became no college (years ≤ 12) or at least some college (years > 12). Marital Statis became two different indictor variables; one for divorced and another for living together unmarries, with married as the reference category. Mother's IQ was split into above and below 90. Hemoglobin was coded as above and below 11, since 11 is the average of two standard deviations below normal levels for children 0.5 – 2 years and 2 – 6 years [34]; below 11 thus indicated anemic children. SES and HOME values were coded as above and below their means of 6 and 9.1, respectively. Manganese levels were coded as above or below 1.2 ug/g, the third quartile for the variable; above 1.2 ug/g indicates concerningly high levels of manganese. Age was coded as above or below toddling at of 24 months.

7.2.1. Application Results—The first task is to construct a range of classes to explore. One class would tell us nothing, so our minimum number of classes was set at two. To include a variety of classes, the maximum was set at eight. For each of the seven different numbers of classes, the methodology of Harel et al [5] was performed. Entropy, AIC, and BIC values were obtained for each of the seven LCA class amounts. For the entropy criterion, three thresholds were examined, $t = 1, 2, 4$. AIC and BIC both chose two classes as the best.

Let us first choose a number of classes via the new entropy criterion. The entropy values, together with the relative change of entropy, is shown in Figure 3. The green lines overlayed on the graph are the three thresholds considered for the entropy criterion. Naturally, the thresholds are more restrictive as the value for t increases. The most generous threshold, σ^*, either includes or nearly includes all classes starting at $g = 4$. The two other thresholds only capture the final class. When we examine the models that have more than four classes, we discover that at least one class in each model is empty. Thus, we settle on four classes as that selected by the entropy criterion.

The fact that σ^* is our most successful criterion, as it chose for us the number of classes, does not agree with the results from our simulation study. The simulation suggested that σ_2^* or σ_4^* would be the best thresholds to use. This disagreement illustrates the inherent difference between simulation and data application. Recall that the simulation data was generated by an LCA model, and thus was easily analyzed by LCA methods. The data in the application section, by contrast, is messy. Thus, it makes sense that a more lenient threshold is the one that gives the best results.

Now we must compare the classes found by AIC and BIC to those found by our entropy criterion. Descriptions of each class, for both the AIC / BIC result and the entropy result, are in Table 3. The values inside the table are the percent of records that have the trait listed in the column header. For example, the values under "College" are the percent of records in that class which have Father's Education flagged as "At least some college." The percents are calculated with respect to the number of observed values within that variable.

We are interested to see whether using four classes instead of two adds anything to the interpretation of the results.

Let us begin in the AIC / BIC section of the table. Class 1 is entirely made up of divorced parents and father's with no college education, while Class 2 has 6.2% of fathers with some college education, and a close to 60–40 split across Living together unmarries and Married parents. High mothers' IQ levels in both classes are fairly low (10% and 24.2%), anemia rates are similar (20% and 17.3%), SES scores are close (34.6% and 45.7%), and manganese levels differ only by about 20% (40% to 22.7%). The only illustrative difference between the classes is in HOME score, where only 4% of Class 1 had a higher than average score, compared to 54.8% of Class 2. Overall, there is not a coherent story to be found in the AIC / BIC results. While some differences are apparent, there is nothing to help researchers identify children at risk of anemia or elevated levels of manganese.

Now we shift our focus to the second half of the table, with the four classes uncovered by the entropy criterion. Instead of comparing variables across classes, let us describe each class in turn.

Class 1 consists entirely of children who are healthy, in terms of iron levels and manganese levels. All of the children's mothers in this class have IQ over 90, while no father has any college education. The majority of children have above average SES (83.3%) and HOME (66.7%) values.

Class 2 also has no child with elevated levels of manganese, and the second smallest percent of anemic children (18.6%). It has the highest percent of children with above average HOME scores (70%), and the second highest percent of children with above average SES (51.5%).

Class 3 has the highest percentage of anemic children (22.7%) with elevated manganese levels (85%). These children all have parents that are living together but umarried, they have the smallest percent of above average SES (17.4%), and the second smallest percent of above average HOME scores (22.7%).

Class 4 has the second highest percentage of anemic (20%) and high manganese (41.7%) children. All children in this class have below average HOME scores, and nearly a third have below average SES (32%). All parents of children in this class are divorced.

There are several different stories that we can glean from the four-class solution. First, healthy children tend to have mothers with higher IQs and good home and socio-economic environments. When socioeconomic score and mother's IQ dips, children may be at risk for anemia, but manganese levels may remain within healthy limits. Low SES and HOME scores seem to be correlated with elevated anemia and high manganese rates. These stories could help identify at-risk children in future, and may be used to make sure parents are aware of the risks of anemia and elevated manganese. These stories, spelled out in the four-class model, were not uncovered using AIC or BIC.

8. Conclusion

We have presented a new entropy-based model selection criterion that chooses the correct number of classes more frequently than AIC and BIC in simulated data sets, and identifies useful class structures in a family studies data set.

We began by proving that the entropy of an LCA model approaches zero as the number of classes approaches the number of unique records; in effect, as the model begins to over-fit the data. Using this result, we developed an entropy-based criterion with a penalty function that recognized when the differences in entropy were beginning to tail off, thus signalling the beginning of over-fitting the data. Simulation studies showed the entropy-based criterion outperforming AIC and BIC by frequently selecting the correct, or close to correct, number of classes. Application to a family studies data set demonstrated that our criterion uncovered more nuanced, useful, and actionable classes than the AIC and BIC.

These results allow us to expand the current methodology of performing LCA on incomplete data from the restraint of selecting one fixed number of classes, chosen via complete-case analysis, to considering a range of classes which take into account imputed data.

Future directions for this work are many. We wish to leverage the multinomial LCA proof into an entropy-based criterion for multinomial data. In addition, there are many other model selection criteria apart from AIC and BIC, some of which involve information calculations just as entropy does [11]. We focus on AIC and BIC in this document as the first step toward comparing our criterion to these more nuanced model selection criteria. Finally, it is important to mention that we have not yet considered the question of $G = 1$ versus $G > 1$; that is, the presence or absence of any clustering behavior. Future interests include tackling this additional question.

Acknowledgments

The research was partially supported by Award Number K01MH087219 from the National Institute of Mental Health, the University of Connecticut CLAS Graduate Fellowship, and the University of Connecticut Department of Statistics Elizabeth Macfarlane Fellowship. The authors are entirely responsible for the contents of this paper. The paper does not reflect the official views of the National Institute of Mental Health nor those of the National Institutes of Health.

The computation was done partially on the Beowulf cluster of the University of Connecticut Department of Statistics. The cluster is partially financed by the NSF SCREMS (Scientific Computing Research Environments for the Mathematical Sciences) grant number 0723557. All tables were formatted using the R package xtable (Dahl, 2014).

The data collection for the application of our method was funded by a seed grant from the Children, Youth and Families Consortium at the Pennsylvania State University (Kordas, PI). Sincerest thanks to Dr. Elena Queirolo, Psychologist Graciela Ardoino, and Dr. Nelly Mañay for coordinating the study. In addition, many thanks to the Catholic University of Uruguay study research team for their data collection.

References

1. Hagenaars, JA.; McCutcheon, AL. Applied Latent Class Analysis. New York: Cambridge University Press; 2002.

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

3. Li, T.; Ma, S.; Ogihara, M. Entropy-based criterion in categorical clustering; Proceedings of the 2004 IEEE International Conference on Machine Learning; 2004. p. 536-543.

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

5. Harel O, Chung H, Miglioretti D. Latent class regression: Inference and estimation with two-stage multiple imputation. Biometrical Journal. 2013; 55:541–553. [PubMed: 23712802]

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

7. Shannon CE. A mathematical theory of communication. Bell System Technical Journal. 1948; 27:379–423.

8. Cover, TM.; Thomas, JA. Elements of Information Theory. Second. New York: Wiley-Interscience; 2006.

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

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

11. Fruhwirth-Schnatter, S. Finite Mixture and Markov Switching Models. Springer; 2006.

12. Chaurasia A, Harel O. Using AIC in multiple linear regression framework with multiply imputed data. Health Services and Outcomes Research Methodology. 2012; 12:219–233. [PubMed: 22879799]

13. Everitt, BS.; Landau, S.; Leese, M.; Stahl, S. Cluster Analysis. John Wiley & Sons, Inc.; 2011.

14. Linzer DA, Lewis JB. poLCA: An R package for polytomous variable latent class analysis. Journal of Statistical Software. 42

15. Banfield JD, Raftery AE. Model-based gaussian and non-gaussian clustering. Biometrics. 1993; 49(3):803–821.

16. Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control. 1974; 19:716–723.

17. Kittaneh OA, Khan MAU, Akbar M, Bayoud HA. Average entropy: a new uncertainty measure with application to image segmentation. The American Statistician. 2016; 70:18–24.

18. Dias JG, Vermunt JK. A bootsrap-based aggregate classifier for model-based clustering. Computational Statistics. 2008; 23:643–659.

19. Harel O, Pellowski J, Kalichman S. Are we missing the important of missing values in HIV prevention randomized clinical trials? review and recommendations. AIDS and Behavior. 2012; 16:1382–1393. [PubMed: 22223301]

20. Cranford JA, McCabe SE, Boyd CJ, Slayden J, Reed MB, Ketchie JM, Lange JE, Scott MS. Reasons for nonresponse in a web-based survey of alcohol involvement among first-year college students. Addictive Behaviors. 2008; 33:206–210. [PubMed: 17728069]

21. Rubin DB. Inference and missing data. Biometrika. 1976; 63:581–592.

22. Rubin, DB. Multiple Imputation for Nonresponse in Surveys. New York: J. Wiley & Sons; 1987.

23. Hua Zhou, X.; Zhou, C.; Lui, D.; Ding, X. Applied Missing Data Analysis in the Health Sciences. John Wiley & Sons; 2014.

24. Little, RJA.; Rubin, DB. Statistical Analysis with Missing Data. New York: Wiley-Interscience; 2002.

25. Schafer JL, Graham JW. Missing data: Out view of the state of the art. Psychological Methods. 2002; 7:147–177. [PubMed: 12090408]

26. Harel O, Zhou X-H. Multiple imputation: Review of theory, implementation, and software. Statistics in Medicine. 2007; 26:3057–3077. [PubMed: 17256804]

27. Schafer JL. Multiple imputation: A primer. Statistical Methods in Medical Research. 1999; 8:3–15. [PubMed: 10347857]

28. Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychological Methods. 2001; 6:330–351. [PubMed: 11778676]
29. Li Meng X. Multiple-Imputation inferences with uncongenial sources of input. Statistical Science. 1994; 9:538–573.
30. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society. Series B (Methodological). 1997; 39:138.
31. Bache K, Lichman M. UCI machine learning repository. donated by Richard Forsyth. 2013 URL http://archive.ics.uci.edu/ml.
32. Rink SM, Ardoino G, Queirolo EI, Cicariello D, Mañay N, Kordas K. Associations between hair manganese lvels and cognitive, language, and motor development in preschool children from montevideo, uruguay. Archives of Environmental and Occupational Health. 2014; 69:46–54. [PubMed: 23930796]
33. Kordas K, Ardoino G, Ciccariello D, Mañay N, Ettinger AS, Cook CA, Queirolo EI. Association of maternal and child blood lead and hemoglobin levels with maternal perceptions of parenting their young children. NeuroToxicology. 2011; 32:693–701. [PubMed: 21925208]
34. Janus J, Moerchel SK. Evaluation of anemia in children. American Family Physician. 2010; 81:1462–1471. [PubMed: 20540485]

Appendix

Proof of Theorem 5.1

Proof

Consider γ_g, the probability of a unique record x_i being in class g, where $i \in 1, \ldots, N_u$, 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:

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

where

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

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 $x_{i,k,o}^u$. If $x_{i,k,o}^u = 1$, then the i^{th} record has the o^{th} value in the k^{th} variable. For every record i and variable k, $(x_{i,k,1}^u, \ldots, x_{i,k,O^k}^u)$ 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 $x_{i,k}^u$, 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 $\pi_{k,o}^{x_{i,k,o}^u}$ which have an exponent of zero will equal one. There will be one single value of $\boldsymbol{\pi}_{k,o}$ where its exponent $x_{i,k,o}^u$ equals one instead of zero. In this case, the corresponding value of $\boldsymbol{\pi}_{k,o}$ 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, $\prod_{O^k} \left(\pi_{k,o}^{x_{i,k,o}^u} \right) = 1$ for fixed k.

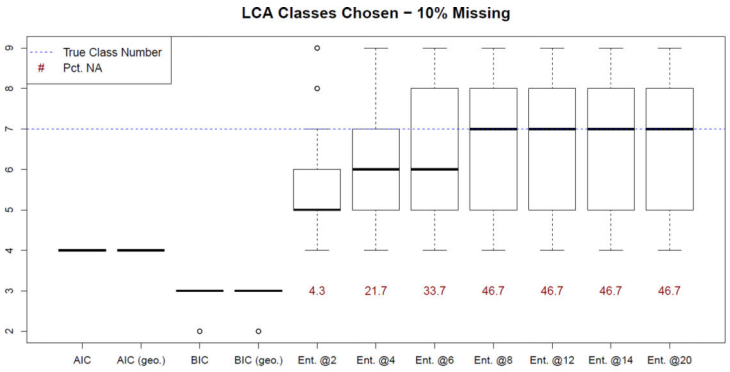
The results given above occur for each value of K. Thus,

$$\alpha_i = \prod_K \left[\prod_{O^k} \left(\pi_{k,o}^{x_{i,k,o}^u} \right) \right] = \prod_K [1] = 1.$$

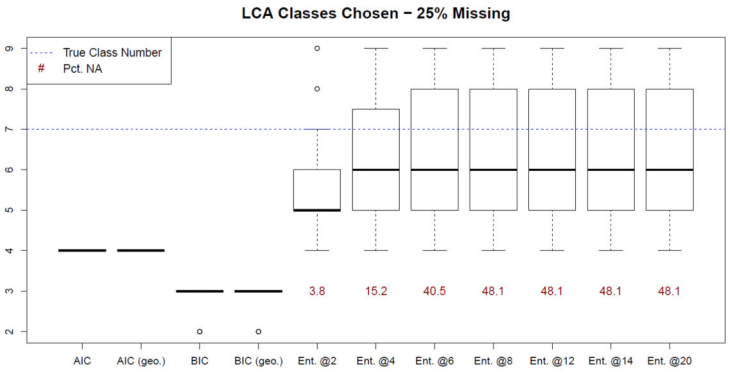
Finally, we have

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

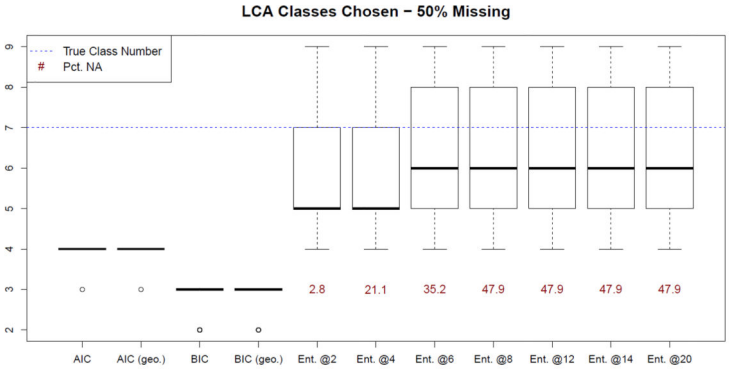
the smallest value entropy can take.



(a) Boxplots showing the number of classes chosen by each method for 10% missingness.



(b) Boxplots showing the number of classes chosen by each method for 25% missingness.



(c) Boxplots showing the number of classes chosen by each method for 50% missingness.

Figure 1.

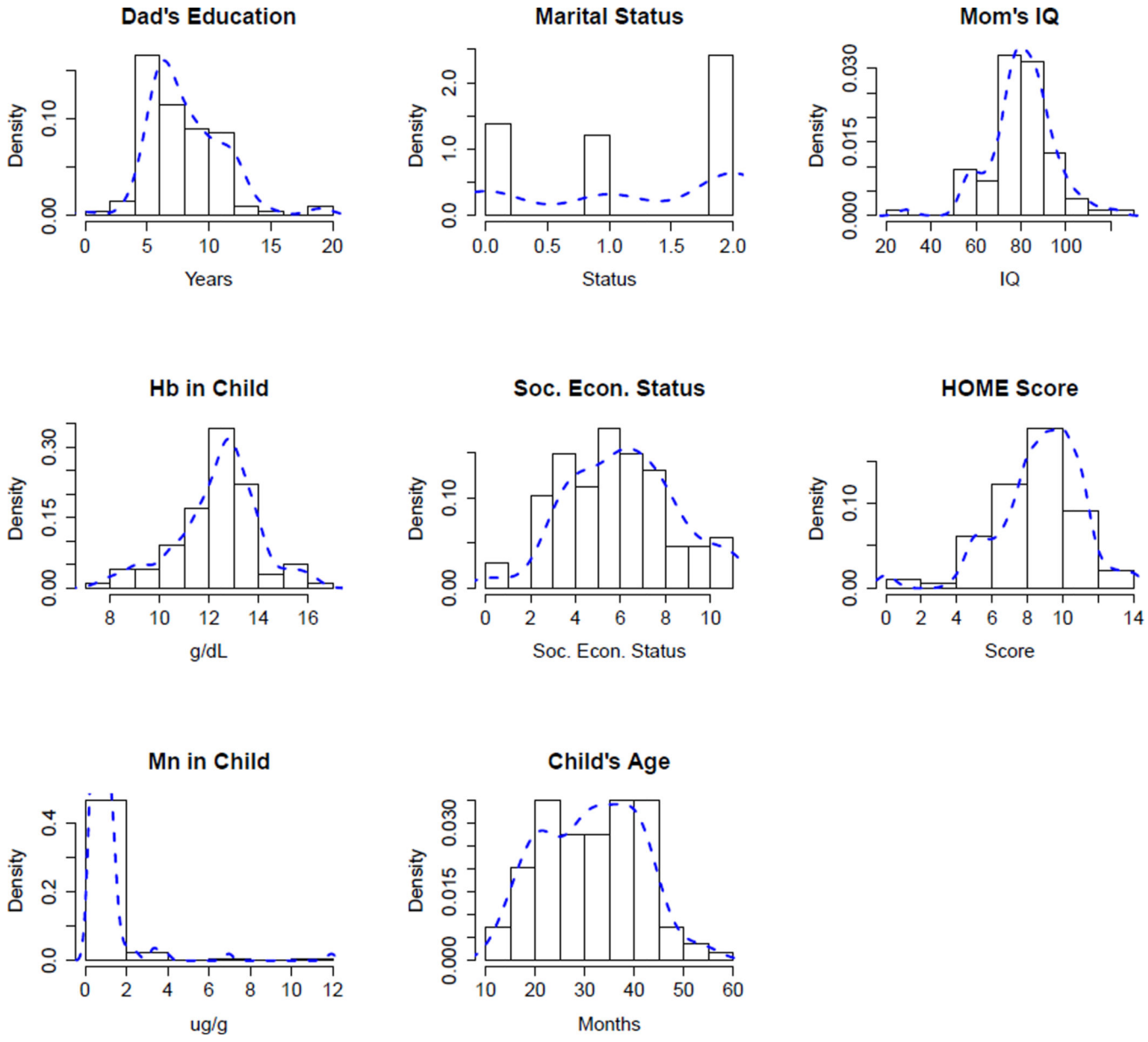


Figure 2.
Histograms of the eight variables analyzed in our application study.

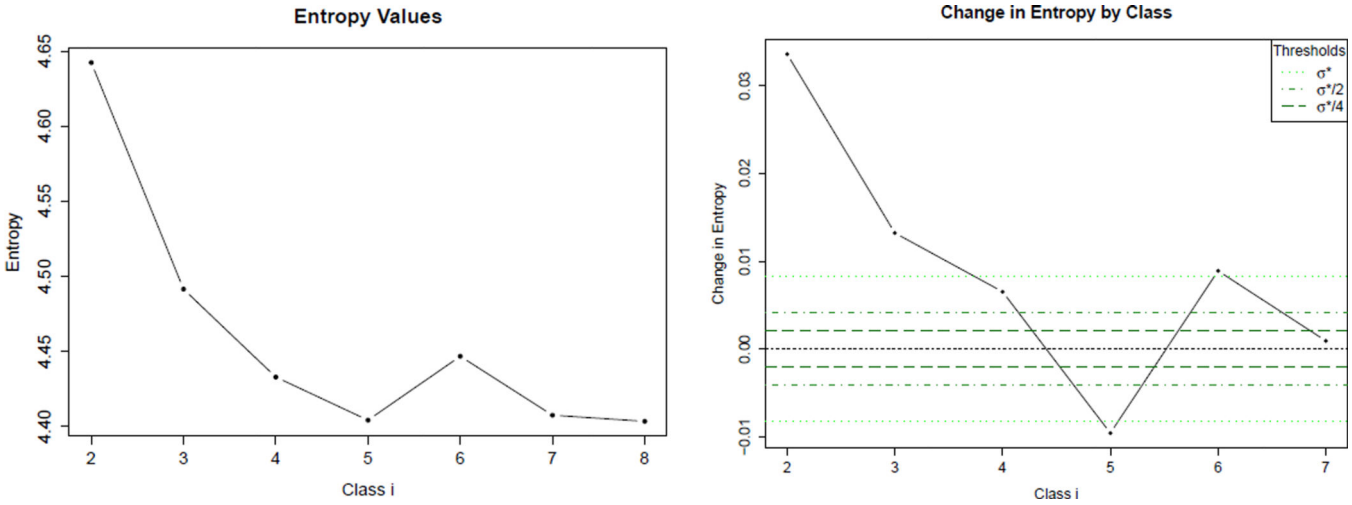


Figure 3.
Entropy values (left) and change in entropy values (right), with thresholds.

Table 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.

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 2

Frequency and percent of missing values in each variable.

	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 3

Percent “Yes” to each of nine categories. Two Classes found via AIC & BIC (top half of table). Four Classes found via the new entropy-based criterion (bottom half of table). Percents are calculated with respect to each variable’s count of observed values within each class.

% 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
% Records	College	Div.	Liv. Tog.	Marr.	IQ > 90	Anemic	SES > 6	HOME > 9.1	Mn High	Infant
21.4	0.0	8.3	41.7	50.0	100.0	0.0	83.3	66.7	0.0	25.0
30.1	8.7	0.0	50.0	50.0	10.8	18.6	51.1	70.0	0.0	37.5
24.0	4.3	0.0	100.0	0.0	15.0	22.7	17.4	22.7	85.0	17.4
24.5	0.0	100.0	0.0	0.0	10.0	20.0	32.0	0.0	41.7	23.1