Statistica Sinica **29** (2019), 551-566
doi:https://doi.org/10.5705/ss.202016.0073

THE IMPACT OF MISSING VALUES ON DIFFERENT MEASURES OF UNCERTAINTY

Chantal Larose[1], Dipak K. Dey[2] and Ofer Harel[2]

[1]*Eastern Connecticut State University and* [2]*University of Connecticut*

Abstract: Entropy quantifies uncertainty in a data set. Intuition tells us that missing values should increase the uncertainty in a data set, but the affect of missing values on entropy has never been quantified. This paper develops formulae for the entropy of incomplete normal data under different missingness mechanisms. The results are compared to the fraction of missing information, which quantifies uncertainty in parameter estimates due to missing values, to compare the two measurements of uncertainty.

Key words and phrases: Entropy, fraction of missing information, missing data, multiple imputation.

1. Introduction

Entropy (Shannon (1948), Cover and Thomas (2006)) is a distribution-based measure of uncertainty in a model, that can quantify the uncertainty in a data set. Missing values are frequently found in data sets, and it makes sense that holes in the data would affect - and, most likely, increase - the uncertainty in the data set. However, the interplay between missing values and entropy has not been quantified.

Multiple imputation (MI) (Rubin (1987), Schafer (1999), Harel and Zhou (2007)) is one method for addressing missing values. MI creates multiple complete data sets, which are then individually analyzed. Estimates of parameters of interest are obtained and combined to form final estimates of the parameters. These final estimates take into account the variability in the data and the fact that the data is incomplete. MI also quantifies the fraction of missing information (FMI) (Rubin (1987)), which measures the information lost to a parameter estimate due to the incomplete data. It is clear that FMI and entropy both measure a type of uncertainty, albeit from different sources. It is of interest to compare the statistics to each other over varying levels of missingness, to identify any relationship between the two. Such a comparison is a first in the literature.

We need some notation. A boldface capital letter (e.g. $\mathbf{Y}$) denotes a matrix. A capital letter without boldface (e.g. Y_1) denotes a random variable. A boldface lower case letter (e.g. $\mathbf{y_1}$) denotes a vector. A lower case letter without boldface (e.g. y_{11}) denotes a scalar quantity.

Sections 2 and 3 review entropy and missing data. Sections 4 and 5 debut our work on incomplete normal data entropy theorems for the MCAR and MAR missingness mechanisms, respectively. Section 6 displays the results of applying our work to simulated data. Section 7 summarizes our findings.

2. Entropy

Entropy is a numeric description of the randomness inherent in a stochastic system. It utilizes the probability mass function (pmf) or probability density function (pdf) of the random variable in question. Let random variable A be discrete with support S_A. The entropy of A is

$$-\sum_{S_A} p(A)\ln(p(A)). \tag{2.1}$$

Similarly, let random variable B be continuous with support S_B. The entropy of B is

$$-\int_{S_B} f(B)\ln(f(B)). \tag{2.2}$$

When calculating the entropy of a data set, it is necessary to consider the entropy of each individual record. Thus, the random variables in the previous definitions denote the records in a data set. If we know that a single record has a bivariate normal distribution, the entropy of that record is the entropy of the bivariate normal distribution.

To calculate the entropy of all records in a data set, we must consider whether the records are independent and identically distributed. If the records are independent and identically distributed, the entropy of the data set is the entropy of one record multiplied by the number of records. If the records are independent but not identically distributed, the entropy of the data set is the sum of the entropy of the records.

3. Missing Data

There are three mechanisms that describe the possible patterns of missing values (Rubin (1987), Rubin (1976)): Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR). MCAR indicates missingness is independent of any data values, MAR indicates missingness may depend on observed values, and MNAR indicates missingness may depend on unobserved values. Since this paper documents the first look at entropy of incomplete data, and MNAR requires additional considerations above and beyond that of MCAR and MAR, we focus on the MCAR and MAR mechanisms.

3.1. Missing data notation

We describe the missingness in a data set $\mathbf{Y}$, with observed and missing elements $\mathbf{Y}_{obs}$ and $\mathbf{Y}_{mis}$, using a corresponding matrix $\mathbf{R}$. In $\mathbf{R}$, the elements are equal to one or zero if the element in $\mathbf{Y}$ is observed or missing, respectively. The data matrix is described using the model $P(\mathbf{Y}|\theta)$, and the missingness matrix is described using the model $P(\mathbf{R}|\phi, \mathbf{Y}_{obs}, \mathbf{Y}_{mis})$.

The model for $\mathbf{R}$ may be simplified under MCAR and MAR mechanisms. If the missing values are MCAR, the model for $\mathbf{R}$ is $P(\mathbf{R}|\phi)$. Similarly, if the missing values are MAR, the model for $\mathbf{R}$ is $P(\mathbf{R}|\phi, \mathbf{Y}_{obs})$.

3.2. Ignorability

Estimating parameters from incomplete data requires consideration of the joint distribution of $\mathbf{Y}$ and $\mathbf{R}$, which can be challenging. However, if MCAR or MAR mechanisms underly the missing values, and the parameters that govern the data are distinct from the parameters that govern the missingness mechanism, the model for $\mathbf{R}$ may be ignored (Schafer (1999), Rubin (1976)). Ignorability allows us to let $P(\mathbf{Y}_{obs}, \mathbf{R}|\theta, \phi) = P(\mathbf{Y}_{obs}|\theta) \times P(\mathbf{R}|\mathbf{Y}_{obs}, \phi)$. Since the only parameter of interest is θ, we may then consider only $P(\mathbf{Y}_{obs}|\theta)$. Ignorability is assumed throughout this work.

One interesting point concerning ignorability is whether it applies to entropy estimation. While ignorability holds when maximum likelihood estimates of interest (Little and Rubin (2002), Schafer (1999)), the particular details of $\mathbf{R}$ directly affect entropy calculation. This quandary and its solution need to be detailed to more thoroughly quantify how incomplete data affects different kinds of analyses.

3.3. Multiple imputation

Multiple imputation allows researchers to address incomplete data sets by simulating many possible values to complete the data, analyzing the resulting complete data sets, and combining the point estimates and standard errors of the analyses into a single result.

In this paper, our focus is on normally distributed data. The algorithm to impute missing values for normally distributed data is well established. The process: Norm (Schafer (2008)), employs the EM algorithm (Dempster, Laird and Rubin (1997)) and data augmentation to generate sets of replacement data values. Sets are denoted $m = 1, \ldots, M$. While imputing normal data typically requires only five to ten imputations (Rubin (1987), Schafer (1997)), more are necessary when imputing in order to estimate FMI (Harel (2007), Schafer (1997)). This will be addressed in the simulation sections.

Once multiple data sets are created, the originally-intended analyses can take place. Let the parameter of interest be Q. Each analysis begets a point estimate $\hat{Q}_m$, with corresponding variance $\hat{U}_m$. We consider several quantities of interest: estimates of entropy, estimate of a variable mean, and estimate of a vector of regression coefficients.

Rubin's Rules (Rubin (1987)) are used to combine the point estimates and variables into a single result. The combination rules are

$$\bar{Q} = \frac{1}{M}\sum_{i=1}^{m}\hat{Q}_m, \qquad \bar{U} = \frac{1}{M}\sum_{i=1}^{m}\hat{U}_m,$$
$$B = \frac{1}{M-1}\sum_{i=1}^{m}\left(\hat{Q}_m - \bar{Q}\right)^2, \quad T = \bar{U} + \left(1 + \frac{1}{M}\right)B.$$

The variable $\bar{Q}$ is the final estimate of parameter Q. The value of T is the final variance estimate of Q.

3.4. Fraction of missing information

The fraction of missing information (FMI) measures the information in a parameter estimate that is lost due to missing values, relative to the information in the estimate if we had complete data (Rubin (1987)). FMI can be used to measure how the missing values affect the uncertainty in a point estimate (Harel (2007)). The estimate of FMI, $\hat{\lambda}$, is

$$\hat{\lambda} = \frac{r + 2/(\nu+3)}{r+1}, \tag{3.1}$$

where $r = (1+m^{-1})B/\bar{U}$, $\nu = (m-1)(1+r^{-1})^2$, and B and $\bar{U}$ are from Rubin's Rules.

4. Entropy under MCAR

The goal of this paper is to illustrate the behavior of entropy and compare it to the behavior of FMI. In order to do so, we must first calculate the entropy of an incomplete data set. We begin by developing a formula for the entropy of bivariate normal data entropy with MCAR missingness.

Let $\mathbf{Y}_{n\times 2} = (\mathbf{y_1}, \mathbf{y_2})$ be jointly distributed $N_2(\theta)$, with $\theta = (\mu, \Sigma)$. MCAR missingness is imposed on $\mathbf{y_2}$. Regardless of the missingness pattern, the records can be rearranged so that the first n_1 values of $\mathbf{y_2}$ are observed, and the remaining $n_1 + 1$ to n values are missing. The MCAR mechanism itself is $\mathbf{R}_{n\times 2} = (\mathbf{r_1}, \mathbf{r_2})$, where the first n_1 values of $\mathbf{r_2}$ equal 1 and the remainder equal 0. Thus, only $\mathbf{r_2}$ is random, while $\mathbf{r_1}$ is a constant vector of zeros. Under the MCAR assumption, the values of $\mathbf{r_2}$ equal one with equal probability, which allows us to model the behavior of $\mathbf{r_2}$ using a *Bernoulli*(ϕ) distribution, where ϕ is the probability of a single observation in $\mathbf{r_2}$ being equal to one. When we consider the data and its missingness together, the notation is $\mathbf{Y_{inc}} = (\mathbf{Y}, \mathbf{r_2}) = (\mathbf{y_1}, \mathbf{y_2}, \mathbf{r_2})$.

The records in $\mathbf{Y_{inc}}$ are independently and identically distributed. Thus, the entropy of the entire data set is the entropy of a single record times the sample size n. What follows is Theorem 1, whose proof is given in the supplementary materials.

Theorem 1. *For bivariate normal data* $\mathbf{Y}$ *with MCAR missingness Bernoulli*(ϕ) *in* $\mathbf{y_2}$*, the entropy is*

$$H(\mathbf{Y_{inc}}) = -\frac{n}{2}\ln\left(2\pi e\sigma_1^2\right) + \frac{n}{2}\ln\left(2\pi e\sigma_2^2\left(1-\rho^2\right)\right) - n(1-\phi)\ln(1-\phi) - n\phi\ln(\phi), \tag{4.1}$$

where $\ln(e) = 1$.

By definition, we know the entropy of the complete, bivariate normal data $\mathbf{Y_{com}}$ to be $H(\mathbf{Y_{com}}) = -(n/2)\ln(2\pi e\sigma_1^2) + (n/2)\ln(2\pi e\sigma_2^2(1-\rho^2))$; and the entropy of the MCAR missingness mechanism $\mathbf{r_2}$ to be $H(\mathbf{r_2}) = -n(1-\phi)\ln(1-\phi) - n\phi\ln(\phi)$. These facts then give us

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

which suggests that researchers can work backwards from the entropy of a complete data set to obtain the value of entropy if there were no missing values, assuming the MCAR missingness is known.

4.1. Limiting behavior of the estimate of entropy of incomplete data

To observe the limiting behavior of Equation (4.1) as the percent of missing values goes to zero, we calculate the limit of $H(\mathbf{r_2})$ as ϕ approaches zero. The proof is given in the supplementary material.

Theorem 2. *For bivariate normal data* $\mathbf{Y}$ *with MCAR missingness* $\mathbf{r_2} \sim$ *Bernoulli*(ϕ) *in* $\mathbf{y_2}$, $\lim_{\phi \to 0} H(\mathbf{r_2}) = 0$.

Thus, the entropy of incomplete MCAR data converges to the complete data entropy as the percent of missing values decreases.

5. Entropy under MAR

We now address the MAR missingness mechanism. In our study, values of $\mathbf{y_2}$ are missing based on the values of $\mathbf{y_1}$. Let $r_{2i} \sim f(r_{2i}|y_{1i}) = Bernoulli(\phi_i^*)$, where $\phi_i^* = e^{\beta_0 + y_{1i}^o}/(1 + e^{\beta_0 + y_{1i}^o})$ and β_0 is a parameter. Here y_{1i}^o is a fixed realization and thus ϕ_1^* is not random, ϕ_1^* is a parameter. Thus, the probability of $\mathbf{y_2}$ being observed increases as the corresponding value of y_{1i} increases. What follows is Theorem 3, whose proof is given in the supplementary materials.

Theorem 3. *For bivariate normal data* $\mathbf{Y}$ *with MAR Bernoulli*(ϕ_i^*) *missingness, where* $\phi_i^* = e^{\beta_0 + y_{1i}^o}/(1 + e^{\beta_0 + y_{1i}^o})$, *the entropy is*

$$H(\mathbf{Y_{inc}}) = \frac{n}{2} \ln\left(2\pi e \sigma_1^2\right) + \frac{n}{2} \ln\left(2\pi e \sigma_2^2 \left(1 - \rho^2\right)\right) - \sum_{i=1}^{n} \left\{(1 - \phi^*) \ln\left(1 - \phi^*\right)\right\} - \sum_{i=1}^{n} \left\{\phi^* \ln\left(\phi^*\right)\right\}. \tag{5.1}$$

Equation (5.1) can be rewritten as

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

which suggests completely observed entropy can be obtained if incomplete data entropy and the missingness mechanism are known.

5.1. Limiting behavior of bivariate estimate

We determine the limit of Equation (5.1) as the percent of missing values goes to zero. The proof is in the supplementary material.

Theorem 4. *For bivariate normal data* $\mathbf{Y}$ *with MAR missingness* $\mathbf{r_2} \sim$ *Bernoulli*(ϕ^*), *where* $\phi^* = e^{\beta_0 + y_{1i}^o}/(1 + e^{\beta_0 + y_{1i}^o})$, *and* y_{i1}^o *is fixed, then*

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

In this, ϕ^* *does not depend on* i, *since* y_{i1}^o *is a fixed realization.*

Thus, the incomplete data entropy formula approaches the complete data case as the percent of missing values goes to zero.

6. Simulations

There are two sets to our simulation studies: MCAR missingness and MAR missingness. Within each set, there are two research topics. The first compares fully-observed data entropy, complete case data entropy, and the new incomplete data entropy via estimates, biases, and standard errors. The second compares the new incomplete data entropy to the fraction of missing information when estimating the mean of the incomplete variable $\mathbf{y_2}$, and when estimating the regression coefficients from regressing $\mathbf{y_2}$ on $\mathbf{y_1}$.

The simulations used MI to complete the incomplete data sets. Here, we must ensure that the assumptions of Rubin's Rules are met before we combine point estimates. To this end, we looked at nine different runs of the MCAR (Figure 1) and MAR (Figure 2) simulations. Clearly, the entropy values are close to the 45-degree line in the normal QQ plots, and thus we can apply Rubin's combining rules to the point estimates.

6.1. MCAR entropy

For each of the 250 repetitions, a thousand observations were generated from $N_2\left(\mu = \begin{pmatrix} 0 \\ 5 \end{pmatrix}, \Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 3 \end{pmatrix}\right)$. The $\mathbf{r_2}$ values were simulated from $Bernoulli(\phi)$ in such a way to calculate 50%, 25%, 10%, 5%, 2.5%, and 1% missing values. The small rates of missing observations are to allow us to identify the behavior of our estimator as the percent of missing values approaches zero, to complement the mathematical derivations in previous sections.

6.1.1. How to estimate MCAR entropy

We need to identify when is it appropriate to estimate the entropy of the missingness mechanism when estimating entropy for fully observed, complete case, and imputed data entropy.

First, we address fully observed data. Completely observed data sets have no incomplete records, and thus there is no need to consider the missingness mechanism. Fully observed normal data entropy is therefore estimated by

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

where σ and ρ have been estimated with s, the sample standard deviation, and

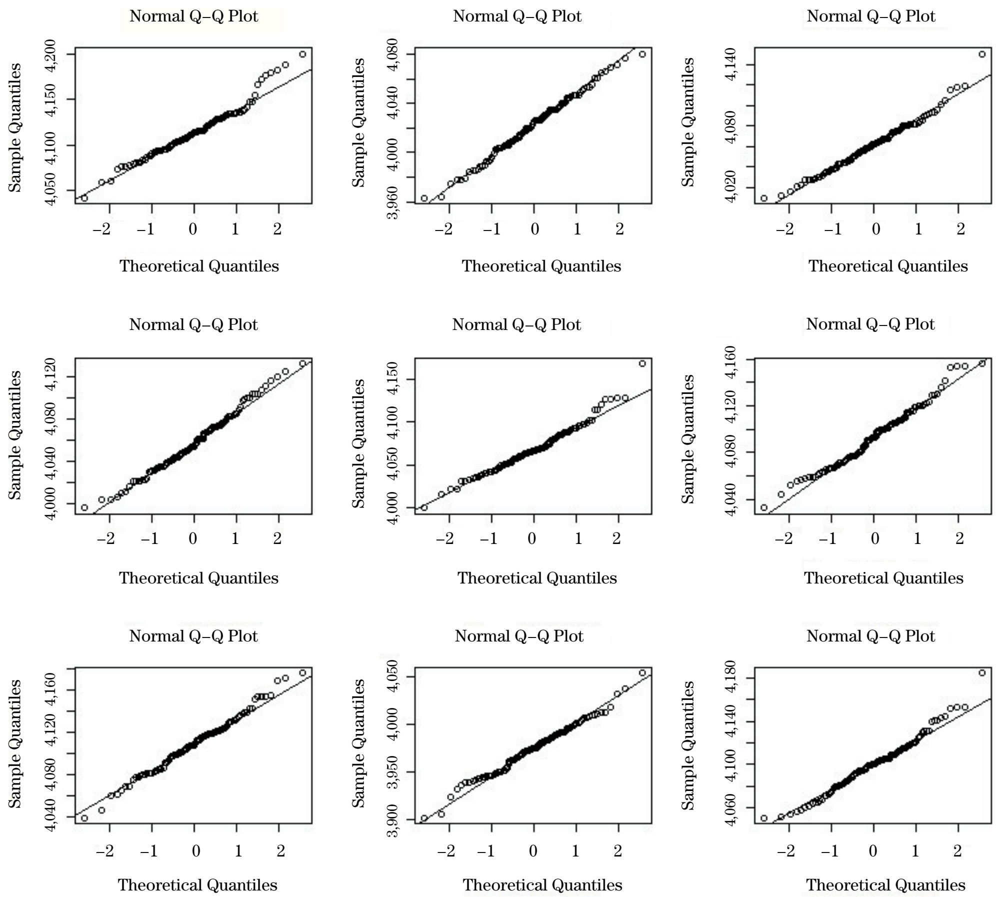

Figure 1. Normal QQ plot of a hundred incomplete MCAR entropy estimates, all reasonably close to the 45-degree line.

c, the sample correlation. The subscript f denotes estimates come from the fully observed data. Here $n_f = n$, the original sample size.

Second, we address complete case data. Complete case analyses involves records with no missing values, since all incomplete records have been removed. Therefore, we again do not consider the missingness mechanism. The complete case entropy is estimated by

$$\hat{H}(\mathbf{Y_{cca}}) = \frac{n_{cca}}{2} \ln\left(2\pi e s^2_{1,cca}\right) + \frac{n_{cca}}{2} \ln\left(2\pi e s^2_{2cca}\left(1 - c^2_{cca}\right)\right),$$

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.

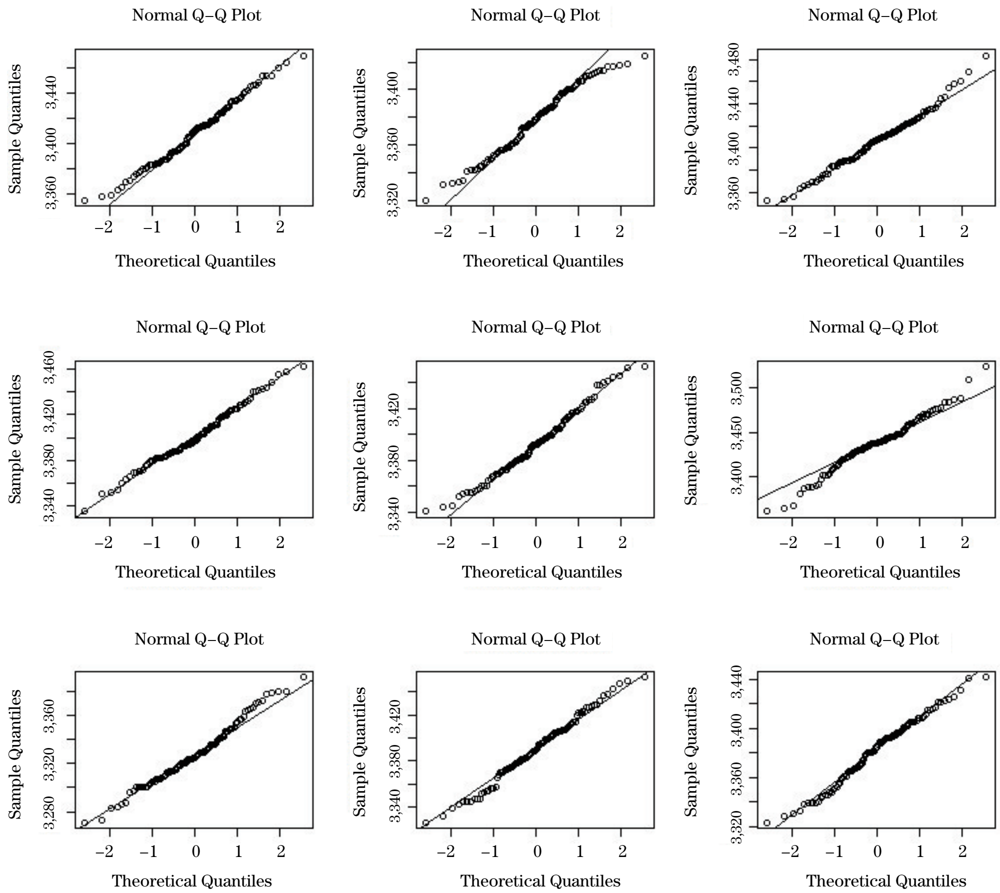

Figure 2. Normal QQ plot of a hundred incomplete MAR entropy estimates, all reasonably close to the 45-degree line.

Third, we address incomplete data. Estimating the entropy of incomplete data requires multiple steps. We begin by imputing the missing data using the R package *norm* (Schafer (2008)), which supplies multiple data sets. The entropy of each data set is obtained using

$$\hat{H}\left(\mathbf{Y}_{m_i}\right) = \frac{n_m}{2}\ln\left(2\pi e s_{1m}^2\right) + \frac{n_m}{2}\ln\left(2\pi e s_{2m}^2\left(1-c_m^2\right)\right) + n_m\left(1-p_m\right)\ln\left(1-p_m\right) - n_m p_m \ln(p_m), \tag{6.1}$$

where the subscript m_i denotes estimates come from the m^{th} imputed data set, and $n_m = n_f = n$, the original sample size. Rubin's Rules give the final estimate of entropy, $\hat{H}(\mathbf{Y_m})$.

Table 1. Bivariate MCAR results. Imputations: 100. λ: FMI for 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 of the point estimates are in parentheses.

$\%Mis$	$\hat{H}(\mathbf{Y_f})$	$\hat{H}(\mathbf{Y_m})$	$\hat{H}(\mathbf{Y_{cca}})$	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	3,343.62	4,067.89	1,666.61	0.54	0.54	0.32	724.27	−1,677.01	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	3,342.17	3,922.31	2,504.03	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	3,340.88	3,672.64	3,008.40	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	3,340.94	3,542.86	3,175.64	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	3,344.74	3,401.74	3,310.66	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)

6.1.2. Results

Table 1 shows the entropy of completely observed data, $\hat{H}(\mathbf{Y_f})$; the entropy of complete case data, $\hat{H}(\mathbf{Y_{cca}})$; the average entropy of imputed data, $\hat{H}(\mathbf{Y_m})$; the FMI for estimating μ_2, the mean of $\mathbf{y}_2$ (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 repetitions, and the second row is the standard deviation of the 250 repetitions.

MI estimate of entropy overestimates the entropy of the full data set (see $\Delta(m,f)$), but only by at most 20% (see $\hat{H}(\mathbf{Y_m})/\hat{H}(\mathbf{Y_f})$). The CCA estimate of entropy underestimates the entropy of the full data set (see $\Delta(cca,f)$); at its lowest point it is half the value of fully observed data entropy (see $\hat{H}(\mathbf{Y_{cca}})/\hat{H}(\mathbf{Y_f})$). This is most likely due to the large difference in sample sizes between the full data and complete-case data.

The tabulated values are graphically compared in Figure 3. The horizontal line shows the theoretical value of fully observed bivariate normal entropy under the parameters we have specified. The black line represents the simulated estimate of the fully observed data entropy. As expected, the estimate hovers around the theoretical value. The dark grey line represents the new MI-based estimate of entropy, which overestimates the entropy of the fully observed data. The light grey line represents the CCA estimate of entropy, which underestimates the fully observed entropy estimate. The bars around each point estimate indicate ± one standard deviation. The dashed line has length equal to $\hat{H}(\mathbf{r_2})$, the entropy of the missingness mechanism.

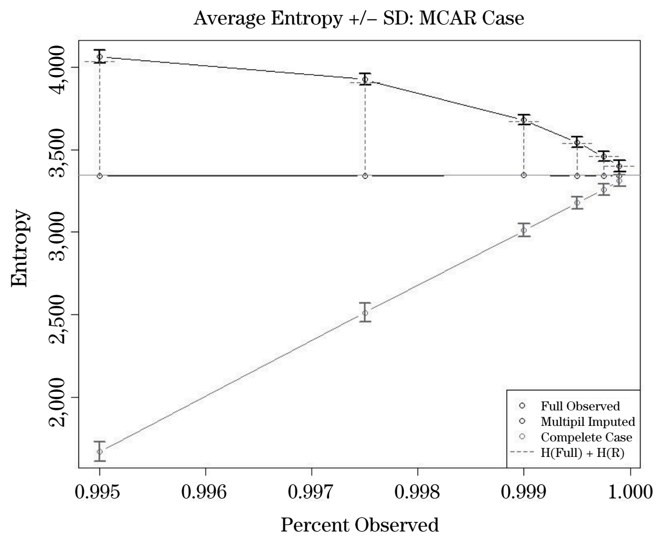

Figure 3. MCAR case. Horizontal line: theoretical fully observed entropy. Center circles: estimated fully observed entropy. Dark grey: MI-based estimated entropy. Light grey: CCA estimated entropy. Dotted lines: $\hat{H}(R)$. Bars: $\pm$ one standard deviation.

The fact that the MI-based estimate overestimates the fully observed data entropy is expected, since it has been shown that additional uncertainty is introduced when values are missing. In addition, the MI-based estimate overestimates the fully observed entropy by almost the exact value of $\hat{H}(\mathbf{r_2})$, as illustrated by the green bars. This supports the initial idea that $\mathbf{r_2}$ is important in the estimate of incomplete data entropy, whereas $\mathbf{r_2}$ can be ignored (under certain conditions) when estimating other quantities (e.g. sample mean).

The result that CCA estimate increasingly underestimates the entropy of the full data set as more data is missing is expected, as more missingness means a smaller complete case data set, and thus less additive entropy. However, comparison of CCA and MI estimates imply two things. First, the CCA error bars are wider and thus the estimate is less reliable. Second, the CCA estimate decreases when common sense (and our new formulae) says that it should increase, making the CCA estimate for entropy a misleading estimate.

Figure 4 compares FMI when estimating μ_2 and FMI when estimating β_0 and β_1 from regressing $\mathbf{y_2}$ on $\mathbf{y_1}$ to the average of the MI estimate of entropy, while comparing both to the simulated estimate of fully observed entropy. The graph suggests a relationship between the imputed estimate of entropy and the fraction of missing information, and a relationship between the imputed estimate

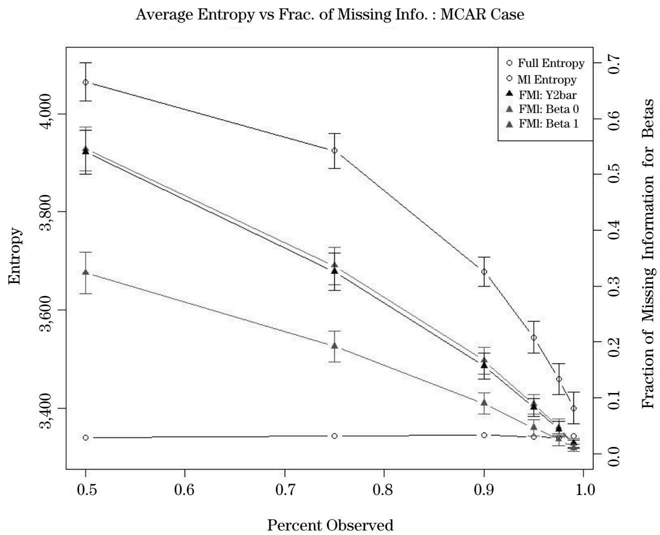

Figure 4. MCAR case. Black with circles: estimated fully observed entropy. Black with triangles: FMI for estimating μ_2. Grey with circles: MI-based estimated entropy. Dark and light grey with triangles: FMI for estimating β_0 and β_1 from regressing $\mathbf{y_2}$ on $\mathbf{y_1}$.

of entropy and the fraction of missing information for estimating β_0, especially for the case in which Percent Observed is above 40%, with possible correlation to the β_1 case as well.

6.2. MAR case

The MAR missingness mechanism is Bernoulli(ϕ^*) missingness, where $\phi^* = e^{\beta_0 + y_{1i}^o}/(1 + e^{\beta_0 + y_{1i}^o})$, for fixed y_{i1}^o. Using the MAR mechanism makes the records of the data set an independent but not identically distributed sample of ($\mathbf{y_1}, \mathbf{y_2}$, $\mathbf{r_2}$). Values of β_0 were set such that the average percent missing for each case was about 50%, 30%, 7%, 5%, 2%, and 0.5%.

6.2.1. How to estimate MAR entropy

As the records are no longer *iid*, entropy must be estimated using a new formula:

$$\hat{H}(\mathbf{Y}_{m_j}) = \frac{n_m}{2} \ln\left(2\pi e s_{1m}^2\right) + \frac{n_m}{2} \ln\left(2\pi e s_{2m}^2 \left(1 - c_m^2\right)\right) - \sum_{i=1}^{n} \left\{\left(1 - p_{i,m}^*\right) \ln\left(1 - p_{i,m}^*\right)\right\} - \sum_{i=1}^{n} \left\{p_{i,m}^* \ln\left(p_{i,m}^*\right)\right\}, \qquad (6.2)$$

where the subscript m_j denotes estimates come from the m^{th} imputed data set,

Table 2. 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 of the point estimates are in parentheses.

$PctMis$	$\hat{H}(\mathbf{Y_f})$	$\hat{H}(\mathbf{Y_m})$	$\hat{H}(\mathbf{Y_{cca}})$	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	3,343.50	3,369.03	1,620.15	0.61	0.62	0.40	25.53	−1,723.34	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	3,341.72	3,443.38	2,279.42	0.41	0.42	0.33	101.66	−1,062.30	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	3,342.51	3,442.54	3,085.10	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	3,342.61	3,424.16	3,173.37	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	3,341.17	3,388.55	3,275.69	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	3,340.46	3,364.36	3,316.36	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)

and the subscript i, m denotes an estimate from the i^{th} record in an imputed data set. Here $n_m = n_f = n$ the original sample size. Rubin's Rules supplies the the final estimate of entropy, $\hat{H}(\mathbf{Y_m})$.

6.2.2. Results

Table 2 contains the results for the MAR case with the same structure as Table 1. Figure 5 shows the same information as Figure 3, now for the MAR case. The MI-based estimate overestimates the fully observed data entropy, but to a lesser degree. The CCA estimate still underestimates the entropy of the full data set, and the bars around the CCA estimates are still larger than the bars around the MI-based estimate.

Figure 6 compares the gap between the MI-based estimate of entropy and the fully observed data entropy to the value of $\hat{H}(\mathbf{r_2})$. 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 $\hat{H}(\mathbf{Y_m})$, possibly due to the fact that $\hat{H}(\mathbf{r_2}|\mathbf{y_1})$ 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 $\mathbf{y_1}$.

Similar to the MCAR case, Figure 7 compares FMI when estimating μ_2, FMI when estimating β_0, and the estimate of fully observed entropy. While the MCAR case shows a similarity between the MI entropy estimates and FMI, no similar pattern exists for the MAR case. This may be, as before, because the value for the MI entropy estimate considers the entropy of y_1.

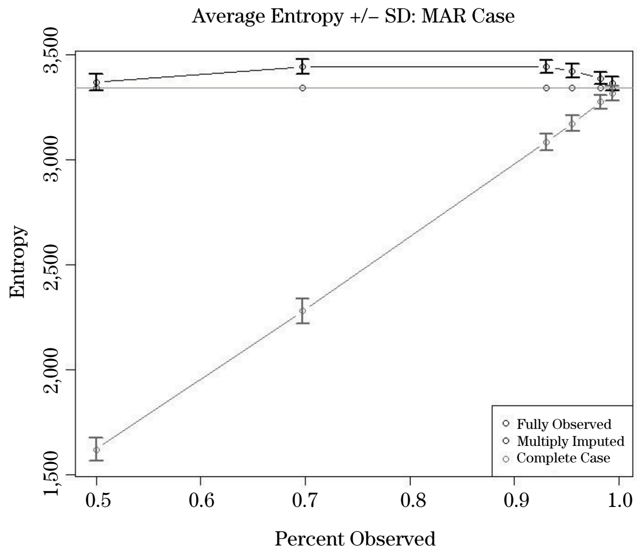

Figure 5. MAR case. Horizontal line: theoretical fully observed entropy. Center circles: estimated fully observed entropy. Dark grey: MI-based estimated entropy. Light grey: CCA estimated entropy. Bars: $\pm$ one standard deviation.

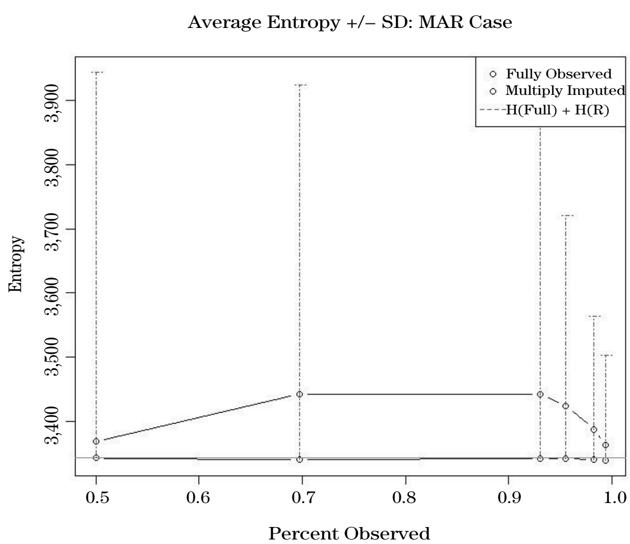

Figure 6. MAR case. Horizontal line: theoretical fully observed entropy. Center circles: estimated fully observed entropy. Dark grey: MI-based estimated entropy. Dashed line: $\hat{H}(R)$.

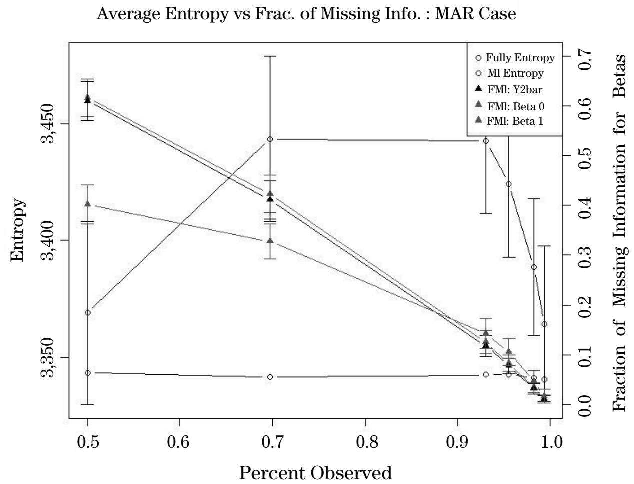

Figure 7. MAR case. Black with circles: fully observed entropy. Black with triangles: FMI estimating $\bar{\mu}_2$. Grey with circles: entropy of incomplete data. Dark and light grey with triangles: FMI for estimating β_0 and β_1. Bars: $\pm$ one standard deviation.

7. Conclusions

This paper documents the first of many strides into the realm of entropy of incomplete data. Our work focused on the bivariate normal case, which allows this new work to begin by considering a well-behaved data structure.

The natural extension of the work presented here is the p-variate normal data case. Work is already well progressed, with theorems established for the MCAR and MAR cases. The limiting behaviors of these estimators, and related simulation studies, are now underway.

Supplementary Materials

Supplementary materials contain the proofs of Theorems 1 through 4.

Acknowledgment

The research was partially supported by Award Number K01MH087219 from the National Institute of Mental Health, the University of Connecticut Department of Statistics Elizabeth Macfarlane Fellowship, and the University of Connecticut CLAS Graduate 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.

References

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*, 2nd Edition. Wiley-Interscience, New York.

Dempster, A. P. and 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 (Statistical Methodological).* **39** 138.

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

Harel, O. (2007). Inferences on missing information under multiple imputation and two-stage multiple imputation. *Statistical Methodology.* **4** 75-89.

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

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

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

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal* **265**, 379-423.

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.* The Methodology Center, The Pennsylvania State University.

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

Mathematical Sciences Department, Eastern Connecticut State University, 83 Windham Street, Willimantic, CT 06226, USA.

E-mail: larosec@easternct.edu

Department of Statistics, 215 Glenbrook Road, University of Connecticut, Storrs, CT 06269-4120, USA.

E-mail: dipak.dey@uconn.edu

Department of Statistics, 215 Glenbrook Road, University of Connecticut, Storrs, CT 06269-4120, USA.

E-mail: ofer.harel@uconn.edu

(Received February 2016; accepted November 2017)