On Testing the Missing at Random Assumption

Manfred Jaeger

Institut for Datalogi, Aalborg Universitet, Fredrik Bajers Vej 7E, DK-9220 Aalborg Ø
jaeger@cs.aau.dk

Abstract. Most approaches to learning from incomplete data are based on the assumption that unobserved values are missing at random (mar). While the mar assumption, as such, is not testable, it can become testable in the context of other distributional assumptions, e.g. the naive Bayes assumption. In this paper we investigate a method for testing the mar assumption in the presence of other distributional constraints. We present methods to (approximately) compute a test statistic consisting of the ratio of two profile likelihood functions. This requires the optimization of the likelihood under no assumptions on the missingness mechanism, for which we use our recently proposed AI & M algorithm. We present experimental results on synthetic data that show that our approximate test statistic is a good indicator for whether data is mar relative to the given distributional assumptions.

1 Introduction

Most commonly used statistical learning methods are based on the assumption that missing values are *missing at random (mar)* [7]. For many datasets this assumption is not completely realistic. However, even when there are doubts as to the exact validity of the *mar* assumption, pragmatic considerations often lead one to adopt *mar*-based techniques like the ubiquitous EM algorithm. To help decide whether a method like EM should be applied, it would be very valuable to know whether the data at hand is *mar* or not. In this paper we investigate a method for performing statistical tests for *mar*.

We have to start with a caveat: *mar* is not testable [2, 6]. The exact technical content behind this statement has to be interpreted carefully: it only says that the *mar*-assumption per se – without any further assumptions about the data – cannot be refuted from the data. However, in many machine learning scenarios certain distributional assumptions about the data are being made. For example, when learning a Naive Bayes model, then the specific independence assumptions underlying Naive Bayes are made. As pointed out in [4], the *mar*-assumption can become refutable in the context of such existing assumptions on the underlying complete data distribution.

In this paper we show that a suitably defined likelihood ratio provides a test statistic that allows us to discriminate between *mar* and non-*mar* models relative to restricted parametric models for the complete data distribution. A crucial component in the computation of the likelihood ratio is the optimization of the likelihood under no assumptions on the coarsening mechanism. For this purpose we employ our recently introduced AI&M procedure [5].

2 The Likelihood Ratio Statistic

We shall work with the coarse data model [3], which allows to consider other forms of incompleteness than missing values. In the coarse data model the *mar*-assumption has its counterpart in the *coarsened at random (car)* assumption.

Incomplete data is a partial observation of some underlying complete data represented by a random variable X with values in a finite state space $W = \{x_1, \ldots, x_n\}$. X has a distribution P_θ for some θ in a parameter space Θ.

The value of X is observed only incompletely. In the general coarse data model such incomplete observations of X can be given by any subset of the state space W. Formally, these observations are the values of a random variable Y with state space 2^W. It is assumed that the observations Y always contain the true value of X. The joint distribution of X and Y, then can be parameterized by $\theta \in \Theta$ and a parameter vector λ from the parameter space

$$\Lambda_{sat} := \{(\lambda_{x,U})_{x \in W, U \in 2^W : x \in U} \mid \lambda_{x,U} \geq 0,\ \forall x \in W : \sum_{U : x \in U} \lambda_{x,U} = 1\},$$

so that $P_{\theta,\lambda}(X = x, Y = U) = P_\theta(X = x)\lambda_{x,U}$.

The parameter space Λ_{sat} represents the *saturated* coarsening model, i.e. the one that does not encode any assumptions on how the data is coarsened. Specific assumptions on the coarsening mechanism can be made by limiting admissible λ-parameters to some subset of Λ_{sat}. The *car* assumption corresponds to the subset

$$\Lambda_{car} := \{\lambda \in \Lambda_{sat} \mid \forall U \forall x, x' \in U : \lambda_{x,U} = \lambda_{x',U}\}.$$

When $\lambda \in \Lambda_{car}$ one can simply write λ_U for $\lambda_{x,U}$.

Let $\boldsymbol{U} = (U_1, \ldots, U_N)$ be a sample of realizations of Y. The log-likelihood ratio based on $\boldsymbol{U}$ for testing the *car*-assumption against the unrestricted alternative is

$$LR(\boldsymbol{U}) := \frac{1}{N}\Big(\max_{\theta \in \Theta} \max_{\lambda \in \Lambda_{car}} \sum_{i=1}^{N} log P_{\theta,\lambda}(Y = U_i) - \max_{\theta \in \Theta} \max_{\lambda \in \Lambda_{sat}} \sum_{i=1}^{N} log P_{\theta,\lambda}(Y = U_i)\Big) \tag{1}$$

(for convenience we normalize the ratio by the sample size).

The *profile log-likelihood* (over Θ) given a coarsening model $\Lambda \subseteq \Lambda_{sat}$ is defined as

$$LL_\Lambda(\theta \mid \boldsymbol{U}) := \max_{\lambda \in \Lambda} \sum_{i=1}^{N} log P_{\theta,\lambda}(Y = U_i).$$

In this paper we will only be concerned with $\Lambda = \Lambda_{sat}$ (no assumptions on the coarsening mechanism) and $\Lambda = \Lambda_{car}$ (*car* assumption). We call the resulting profile likelihoods simply *profile(sat)*-, respectively *profile(car)-likelihood*, denoted LL_{sat}, LL_{car}.

Using profile likelihoods, (1) can be rewritten as

$$LR(\boldsymbol{U}) := \frac{1}{N}(LL_{car}(\hat{\theta} \mid \boldsymbol{U}) - LL_{sat}(\hat{\hat{\theta}} \mid \boldsymbol{U})), \tag{2}$$

where $\hat{\theta}$ and $\hat{\hat{\theta}}$ are the maxima of $LL_{car}(\cdot \mid \boldsymbol{U})$, respectively $LL_{sat}(\cdot \mid \boldsymbol{U})$.

The profile(car) likelihood factors as

$$LL_{car}(\theta \mid \boldsymbol{U}) = \mathit{Lf}(\boldsymbol{U}) + LL_{FV}(\theta \mid \boldsymbol{U}),$$

where $LL_{FV}(\theta \mid \boldsymbol{U}) = \sum_{i=1}^{N} \mathit{log} P_\theta(X \in U_i)$ is the *face-value* log-likelihood [1], i.e. the likelihood obtained by ignoring the missingness mechanism, and

$$\mathit{Lf}(\boldsymbol{U}) := \max_{\lambda \in \Lambda_{car}} \sum_{i=1}^{N} \mathit{log} \lambda_{U_i}. \tag{3}$$

We wish to compute $\mathit{LR}(\boldsymbol{U})$ by computing the three components $\mathit{Lf}(\boldsymbol{U}),\ LL_{FV}(\hat{\theta} \mid \boldsymbol{U})$, and $LL_{sat}(\hat{\hat{\theta}} \mid \boldsymbol{U}))$. In most cases it will be impossible to obtain exact, closed-form solutions for any of these three terms. We therefore have to use approximate methods.

Approximating $\boldsymbol{LL_{FV}}$ To compute $LL_{FV}(\hat{\theta} \mid \boldsymbol{U})$ one has to find $\hat{\theta}$, i.e. optimize the face-value likelihood. This can typically be accomplished by some version of the EM algorithm. In our experiments we use the EM implementation for Bayesian networks provided by the Hugin system (www.hugin.com). Since we are not guaranteed to find a global maximum of LL_{FV}, we obtain only a lower bound on $LL_{FV}(\hat{\theta} \mid \boldsymbol{U})$.

Approximating $\boldsymbol{\mathit{Lf}(U)}$ To approximate the term $1/N\mathit{Lf}(\boldsymbol{U})$ in (2) we have to find

$$\max_{\lambda \in \Lambda_{car}} \frac{1}{N} \sum_{i=1}^{N} \mathit{log}(\lambda_{U_i}) = \max_{\lambda \in \Lambda_{car}} \sum_{j=1}^{K} m(\bar{U}_j) \mathit{log}(\lambda_{\bar{U}_j}), \tag{4}$$

where $\bar{\boldsymbol{U}} := \bar{U}_1, \ldots, \bar{U}_K$ is an enumeration of the distinct $U_i \in \boldsymbol{U}$, and $m(\bar{U}_j)$ is the empirical probability of $\bar{U}_j$ in $\boldsymbol{U}$. Thus, computing $\mathit{Lf}(\boldsymbol{U})$ is a convex optimization problem under linear constraints of the form $\lambda_U \geq 0$ and $\sum_{U : x \in U} \lambda_U = 1$. However, since there is one constraint of the latter form for each $x \in W$, the number of constraints is manageable only for very small state spaces W.

As a first simplification of the problem, we observe that since the objective function only depends on $\lambda_{\bar{U}}$ for $\bar{U} \in \bar{\boldsymbol{U}}$, we can restrict the optimization problem to these $\lambda_{\bar{U}}$ under the linear constraints

$$C(x) : \quad \sum_{\bar{U} \in \bar{\boldsymbol{U}} : x \in \bar{U}} \lambda_{\bar{U}} \leq 1 \qquad (x \in W). \tag{5}$$

An optimal solution $\hat{\lambda}$ for $(\lambda_{\bar{U}_1}, \ldots, \lambda_{\bar{U}_K})$ can be extended to an optimal solution $\hat{\lambda} \in \Lambda_{car}$ by setting $\hat{\lambda}_{\{x\}} := 1 - \sum_{\bar{U} \in \bar{\boldsymbol{U}} : x \in \bar{U}} \hat{\lambda}_{\bar{U}}$ for all $x \in W$ with $\{x\} \notin \bar{\boldsymbol{U}}$, and $\hat{\lambda}_U := 0$ for all other $U \notin \bar{\boldsymbol{U}}$.

The optimization problem now is reduced to a manageable number of parameters. In order to also obtain a manageable number of constraints, we perform the optimization of (4) only under a subset $C(x_1), \ldots, C(x_m)$ of the constraints (5), which is obtained by randomly sampling $x_i \in W$. Since we are thus relaxing the feasible region, we obtain an over-estimate of (4). In our experiments we found that it is sufficient to sample approximately $m = 2K$ constraints, i.e. adding more constraints tended not to change the computed maximum (4) significantly.

Once a set of constraints has been generated we employ the standard Lagrange multiplier approach to perform the optimization. For the unconstrained optimization of the dual function the PAL Java package is used (http://ftp.cse.sc.edu/bioinformatics/PAL/pal-1.4/).

Approximating $\boldsymbol{LL_{sat}}$ To compute $LL_{sat}(\hat{\hat{\theta}} \mid \boldsymbol{U}))$ we have to maximize the profile(sat)-likelihood. For this we use the AI&M procedure as introduced in [5]. AI&M resembles the EM procedure for maximizing LL_{FV}. Like EM, AI&M is a generic method that has to be implemented by concrete computational procedures for specific types of parametric models. In our experiments we use the AI&M implementation for Bayesian networks as described in [5]. The AI&M procedure will usually not find a global maximum of LL_{sat}, so that as for LL_{FV} we obtain a lower bound on the correct value.

Combining our approximations for the components of (2), we obtain an approximation $\bar{L}R(\boldsymbol{U})$ of $LR(\boldsymbol{U})$. Since we over-estimate $Lf(\boldsymbol{U})$, and under-estimate LL_{FV} and LL_{sat}, one cannot say whether $\bar{L}R(\boldsymbol{U})$ will over- or under-estimate $LR(\boldsymbol{U})$. However, when our approximation for LL_{FV} is better than our approximation for LL_{sat} (as can be expected), then we will obtain an over-estimate of $LR(\boldsymbol{U})$.

3 Generating Non-*car* Data

In our experiments we want to investigate how effective our computed $\bar{L}R(\boldsymbol{U})$ is for testing *car*. To this end we generate incomplete data from Bayesian network models following the general procedure described in [5]: to a Bayesian network with nodes $V_1, \ldots, V_k$ Boolean *observation nodes* $obsV_1, \ldots, obsV_k$ are added. The observation nodes are randomly connected with the original nodes and among themselves. The conditional probability tables for the observation nodes are randomly filled in by independently sampling the rows from a Beta distribution with mean μ and variance σ. Then complete instantiations of the extended network are sampled, giving an incomplete observations of $V_1, \ldots, V_k$ by omitting the values for which $obsV_i = false$.

This general procedure allows us to control in various ways how non-*car* the generated data will be. The first way is by setting the variance σ of the Beta distribution: $\sigma = 0$ means that all rows in all conditional probability tables will be identical, and so the $obsV_i$ nodes become actually independent of their parents, meaning that the data becomes *car* (indeed, missing completely at random). Large values of σ lead to nearly deterministic, complex dependency patterns of the $obsV_i$ variables on their parents, which allows for highly non-*car* mechanisms.

A second way of controlling *car* is by taking mixtures of several coarsening mechanisms: to generate a sample of size N, $l \geq 1$ different coarsening models are generated by our standard method, and from each a sample of size N/l is generated. By the following theorem, the data thus generated becomes *car* for $l \to \infty$.

Theorem 1. *Let μ be a probability distribution on Λ_{sat} such that*

$$\textit{for all } U \subseteq W,\ x, x' \in U:\ E_\mu[\lambda_{x,U}] = E_\mu[\lambda_{x',U}]. \tag{6}$$

There exists $\lambda_\infty \in \Lambda_{\mathrm{car}}$ such that for $\lambda_1, \lambda_2, \ldots \in \Lambda_{\mathrm{sat}}$ iid sampled according to μ: $P_\mu(lim_{n\to\infty} 1/n \sum_{i=1}^{n} \lambda_i = \lambda_\infty) = 1$. Furthermore, for fixed $\theta \in \Theta$: $P_\mu(lim_{n\to\infty} 1/n \sum_{i=1}^{n} P_{\theta,\lambda_i} = P_{\theta,\lambda_\infty}) = 1$.

The proof of the theorem is almost immediate by an appeal to the strong laws of large numbers. The theorem is of some independent interest in that it says that random mixtures of coarsening mechanisms tend to become *car*. This is relevant for real-life datasets, which often can be assumed to be produced by mixtures of coarsening mechanisms (for example, different employees entering customer's records into a database may exhibit different data coarsening mechanisms). However, one must also take into account that the symmetry condition (6) is satisfied for mathematically natural sampling distributions like Lebesgue measure, but not for most real-life sampling distributions over coarsening mechanisms.

Since our random construction of coarsening mechanisms is completely symmetric with respect to different values of the random variables, the symmetry condition (6) is satisfied, and our data becomes *car* for $l \to \infty$.

4 Experiments

In all our experiments we first select a Bayesian network from which incomplete data then is generated as described in the preceding section. In all experiments the structure of the network used for generating the data also defines the parametric model Θ used in computing $\bar{LR}(\boldsymbol{U})$. Thus, in our experiments the assumptions made for the underlying complete data distribution are actually correct. Most of our experiments are based on the standard benchmark 'Asia' and 'Alarm' Bayesian networks. Asia has 8 nodes, defining a state space W of size 256. Alarm has 37 nodes with $|W| = 1.7 \cdot 10^{16}$.

As a reference point for further experiments we use the following base experiment: using Asia as the underlying complete data model, 100 incomplete datasets are generated from 100 different coarsening models. Each dataset is of size 5000, and the parameters of the Beta distribution used in constructing the coarsening model are $\mu = 0.1, \sigma = 0.05$. This setting gives a distribution over parameters that is quite highly concentrated near extreme values 0 and 1, leading to an incomplete data distribution that is strongly non-*car* according to our heuristic described in Section 3.

Figure 1 a) shows the distribution over computed $\bar{LR}(\boldsymbol{U})$ values for the different datasets. The variance in the results is due to variations at three different levels: first, the different randomly generated coarsening models lead to a different expected value $E[LR(\boldsymbol{U})]$; second, the value $LR(\boldsymbol{U})$ for the actually sampled dataset varies from $E[LR(\boldsymbol{U})]$; third, our approximation $\bar{LR}(\boldsymbol{U})$ varies from $LR(\boldsymbol{U})$. Figure 1 b) shows the result of sampling 100 different datasets each from only three different coarsening models. This clearly indicates that the primary source of variance in a) is the difference in the coarsening models (of course, this can change for smaller sample-sizes). Finally, for the model inducing the leftmost cluster in b), and one dataset sampled from that

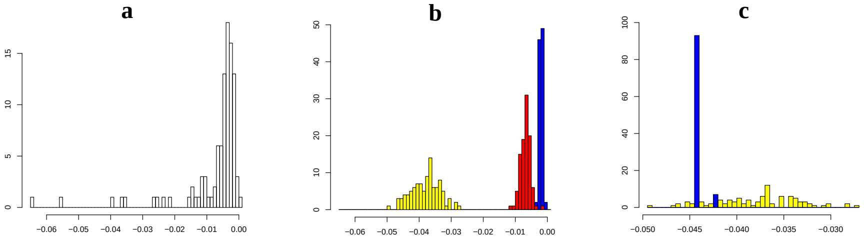

Fig. 1. Likelihood ratio distribution

model, the computation of $LR(\boldsymbol{U})$ was repeated 100 times. Figure 1 c) shows the result (black histogram), and, for comparison the result from 100 different datasets (light gray histogram – this is the same as in b)). From this we infer that the variance observed in b) for fixed models is mostly due to the sample variance of the datasets, and less to the variance in the randomized computation of $\bar{LR}(\boldsymbol{U})$. The bi-modality observed in the black histogram in c) can be traced back to the computation of $LL_{FV}(\hat{\theta}\boldsymbol{U})$, i.e. there appear to have been (at least) two different convergence points of EM.

In summary, the results shown in Figure 1 show that our computed $\bar{LR}(\boldsymbol{U})$ measure actual properties of the given model and data, and is not dominated by noise in the computation. We can now proceed to investigate how good an indicator for *car*-ness this value is. To this end we now vary the coarse data generation of the base experiment in two ways: in one experiment we use smaller variance parameters $\sigma = 0.02$ and $\sigma = 0$ in the Beta distribution. In a second experiment we leave the Beta distribution unchanged, but create mixtures with $l = 2$ and $l = 10$ components. Figure 2 shows the results. The left histograms in both rows are just the results from the base experiment again. As we move from left to right (in both rows), the data becomes more *car* according to our heuristic *car*-measures σ and l (it is truly *car* in the $\sigma = 0$ experiment). The corresponding increasing concentration of $\bar{LR}(\boldsymbol{U})$ near 0 shows that it can indeed serve as statistic for discriminating between *car* and non-*car* models.

Due to its quite small state space, models based on the Asia network do not pose the full computational challenge of computing $\bar{LR}(\boldsymbol{U})$. An experiment with two different variance settings has also been conducted for the Alarm network. Figure 3 shows the result. We observe that here the computed $\bar{LR}(\boldsymbol{U})$ values are all positive. Since the actual $LR(\boldsymbol{U})$ values must be ≤ 0, this means that the over-estimate of $Lf(\boldsymbol{U})$, combined with the under-estimate of LL_{sat} here lead to a significant over-estimate of $LR(\boldsymbol{U})$. Nevertheless, the computed $\bar{LR}(\boldsymbol{U})$ discriminates quite successfully between the *car* and non-*car* models.

As a final test for the $\bar{LR}(\boldsymbol{U})$ computation we use data from three different artificially constructed Bayesian networks: all networks contain seven binary nodes. The first network ('simple') contains no edges; the second network ('medium') contains 14 randomly inserted edges, and the third ('dense') is a fully connected network (21 edges). The conditional probability tables are randomly generated. The three models represent decreasingly restrictive distributional assumptions, with the dense network not encoding any restrictions. We again sample 100 datasets from 100 different random

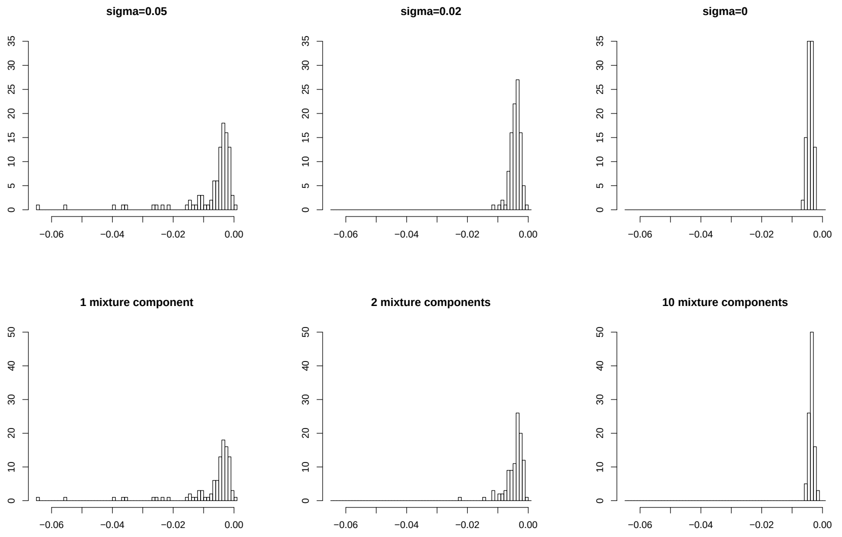

Fig. 2. Computed likelihood ratios and heuristic *car*-measures (Asia)

coarsening models (sample-size 5000, $\mu = 0.1$, $\sigma = 0.05$). Figure 4 shows the result. As required by the fact that *car* is not testable without any restrictive assumptions on the full data model, we observe that the $\bar{LR}(\boldsymbol{U})$-values for the dense network show no indication that the data is not *car*. The more restricted the model, the easier it becomes to refute the *car*-assumption based on $\bar{LR}(\boldsymbol{U})$.

5 Conclusion

Utilizing our recently introduced AI&M procedure for optimizing the profile(sat)-likelihood, we have shown how to compute an approximate likelihood-ratio statistic for test-

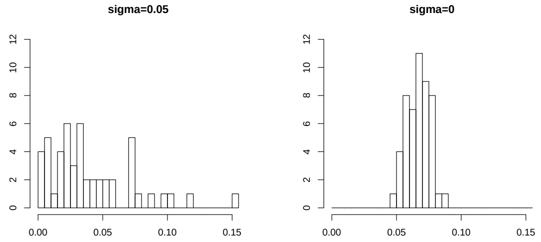

Fig. 3. Computed likelihood ratios and heuristic *car*-measures (Alarm)

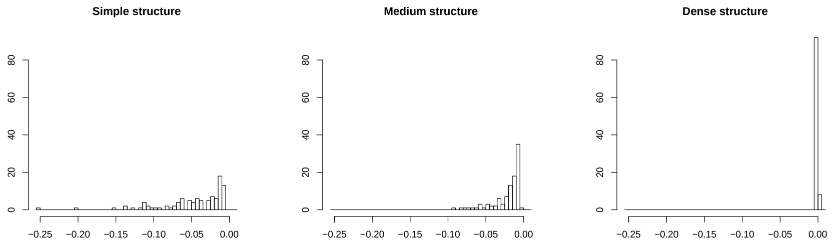

Fig. 4. Likelihood ratio and model structure

ing the *car* assumption in the context of distributional constraints on the underlying complete data distribution. Initial experiments show that we obtain a quite effective measure for discriminating between *car* and non-*car* incomplete data distributions.

To obtain a practical test for a particular dataset under consideration, one will also need a way to specify a critical value κ, so that the *car* hypothesis will be accepted iff $\bar{LR}(\boldsymbol{U}) \geq \kappa$. At this point there exist no general, theoretically well-founded rules for setting κ. The best way to proceed, therefore, is to empirically determine for a given state space W and a parametric model Θ, the sampling distribution of $\bar{LR}(\boldsymbol{U})$ under the *car* assumption, and to set κ according to the observed empirical distribution and the desired confidence level.

Tests for *car* can also play a role in model selection: when *car* is rejected relative to a current parametric complete data model Θ, one may either retain Θ and employ techniques not relying on *car*, or one can relax the parametric model, thus hoping to make it consistent with *car* (which ultimately it will, as illustrated in Figure 4).

References

1. A. P. Dawid and J. M. Dickey. Likelihood and bayesian inference from selectively reported data. *Journal of the American Statistical Association*, 72(360):845–850, 1977.
2. R. D. Gill, M. J van der Laan, and J. M. Robins. Coarsening at random: Characterizations, conjectures, counter-examples. In D. Y. Lin and T. R. Fleming, editors, *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis*, Lecture Notes in Statistics, pages 255–294. Springer-Verlag, 1997.
3. D. F. Heitjan and D. B. Rubin. Ignorability and coarse data. *The Annals of Statistics*, 19(4):2244–2253, 1991.
4. M. Jaeger. Ignorability for categorical data. *The Annals of Statistics*, 33(4):1964–1981, 2005.
5. M. Jaeger. The AI&M procedure for learning from incomplete data. In *Proceedings of UAI-06*, 2006. To appear.
6. C. Manski. *Partial Identification of Probability Distributions*. Springer, 2003.
7. D.B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.