Statistical Science
2011, Vol. 26, No. 2, 162–174
DOI: 10.1214/10-STS318

Calibrated Bayes, for Statistics in General, and Missing Data in Particular[1]

Roderick Little

Abstract. It is argued that the Calibrated Bayesian (CB) approach to statistical inference capitalizes on the strength of Bayesian and frequentist approaches to statistical inference. In the CB approach, inferences under a particular model are Bayesian, but frequentist methods are useful for model development and model checking. In this article the CB approach is outlined. Bayesian methods for missing data are then reviewed from a CB perspective. The basic theory of the Bayesian approach, and the closely related technique of multiple imputation, is described. Then applications of the Bayesian approach to normal models are described, both for monotone and nonmonotone missing data patterns. Sequential Regression Multivariate Imputation and Penalized Spline of Propensity Models are presented as two useful approaches for relaxing distributional assumptions.

Key words and phrases: Maximum likelihood, multiple imputation, penalized splines, propensity models, sequential regression multivariate imputation.

1. INTRODUCTION

There was clearly a time, perhaps not too far in the recent past, when Bayesian methods were considered "beyond the pail" by frequentist statisticians. But Bayesian methods have been resurgent in recent years, to the extent that few statisticians have no interest in them, even if they do not buy the complete philosophical package.

In this article I summarize my perspective on the role of Bayesian methods in statistics, borrowing from a more extensive discussion in Little (2006). I then provide a brief overview of Bayesian inference for missing data problems, both modeling and ignoring the missing data mechanism, and multiple imputation (MI), an important practical tool for dealing with missing data that has a Bayesian etiology. Finally, I give some examples of Bayesian missing-data methods which I believe frequentists could profitably add to their analytical toolkit.

Roderick Little is Richard D. Remington Collegiate Professor, Department of Biostatistics, University of Michigan, 1415 Washington Heights, Ann Arbor, Michigan 48109, USA (e-mail: rlittle@umich.edu).

[1]Discussed in 10.1214/10-STS318B and 10.1214/10-STS318A; rejoinder at 10.1214/10-STS318REJ.

Bayesian methods are particularly useful for handling missing data problems in statistics. Incomplete data problems are readily amenable to likelihood-based methods, since they do not require a rectangular data matrix. Maximum likelihood (ML) is an important approach, but the loglikelihoods corresponding to missing data problems are typically more complex than likelihoods for complete data, deviate more from the quadratic approximations that underlie asymptotic inferences, and are subject to under-identification or weak identification of parameters. Consequently, ML requires iterative calculations, information matrix-based standard errors are often difficult to compute in high-dimensional problems, and asymptotic ML inferences can have serious deficiencies, particularly with small fragmentary samples. In contrast, draws from the Bayesian posterior distribution can often be computed using direct simulation or Markov Chain Monte Carlo techniques, and these provide estimates of standard errors without the need to compute and invert the information matrix. The inferences based on the posterior distribution often have better frequentist properties than asymptotic inferences based on ML. Furthermore, multiple imputations can be generated as a byproduct of Bayesian calculations, and provide a practically useful and flexible tool for solving missing data problems.

These points are further developed in the material that follows.

2. WHY BAYES? THE CALIBRATED BAYES PHILOSOPHY

The statistics world is still largely divided into frequentists, who base inference for an unknown parameter θ on hypothesis tests or confidence intervals derived from the distribution of statistics in repeated sampling, and Bayesians, who formulate a model for the data and prior distribution for unknown parameters, and base inferences for unknowns on posterior distributions. Bayesians are also "subjective," as when proper priors are elicited, and "objective," as when conventional "reference priors" are adopted. Both these facets of the Bayesian paradigm have useful roles, depending on context. Asymptotic maximum likelihood inference can be seen as a form of large sample Bayes, with the interval for θ being interpreted as a posterior credibility interval rather than a confidence interval.

I believe both systems of statistical inference have strengths and weaknesses and, hence, the best course is to seek a compromise that combines them in a way that capitalizes on their strengths. The Bayesian paradigm is best suited for making statistical inference under an assumed model. Indeed, under full probability modeling, with prior distributions assigned to parameters, the Bayes theorem is indeed a theorem—the path determined by probability theory to inference about unknowns given the data, whether the targets are parameters or predictions of unobserved quantities.

The frequentist approach, on the other hand, has various well-known limitations regarding inference under an assumed model. First, it is not prescriptive: frequentist theory seems more like a set of concepts for assessing properties of inference procedures, rather than an inferential system *per se*. Under an agreed model, and assuming large samples, there is a relatively prescriptive path to inference based on maximum likelihood (ML) estimates and their large-sample distribution. However, other frequentist approaches are entertained in practice, like generalized estimating equations, based on robustness or other considerations. Also, there is no prescriptive frequentist approach to small sample problems. Indeed, for many problems, such as the Behrens–Fisher problem of comparing two means of normal distributions with different unknown variances, no procedure exists that has exact repeated-sampling properties, such as exact nominal confidence coverage for all values of the unknown parameter. Bayesian methods provide exact frequentist coverage for some complete-data problems—this occurs, in particular, in problems where the Bayesian and frequentist inferences are the same, as in t inference for normal multiple regression with a uniform prior on the regression coefficients and log variance. For more complex problems, including problems with missing data, Bayesian methods do not generally provide exact frequentist coverage, but they often improve on ML by providing better small-sample inferences, perhaps because Bayesian model shrinkage moderates inferences based on extreme parameters estimates. As just one example, consider the adjustment of estimates for categorical data motivated by Bayesian ideas (Agresti, 2002).

The Bayesian approach is prescriptive in the sense that, once a model and prior distribution are specified, there is a clear path to inferences based on the posterior distribution, or optimal estimates for a given choice of loss function. There is no prescription for choosing the model and prior distribution—that is what makes applied statistics interesting—but certain "reference" prior distributions for complete-data problems can be expected to yield good frequentist properties when applied to missing data problems; see, for example, Little (1988).

Frequentist inference violates the likelihood principle, and is ambiguous about whether to condition on ancillary or approximately ancillary statistics when performing repeated sampling calculations. Little (2006) provides more discussion, with examples.

An attractive feature of Bayesian methods in complete-data or missing-data problems is the treatment of nuisance parameters. Bayesian inference integrates over these parameters, rather than fixing them at their ML estimates. This tends to yield inferences with improved frequentist properties, since the uncertainty about these parameters is taken into account. For example, for complete or incomplete data problems, restricted ML, which integrates over location parameters, is generally viewed as superior to ML, which maximizes them.

If we were handed the model on a plate and told to do inference for unknowns, then Bayesian statistics is the clear winner. The problem, of course, is that we never know the true model. Bayesian inference requires and relies on a high degree of model specification (Efron, 1986)—full specification of a likelihood and prior. Developing a good model is challenging, particularly in complex problems. Furthermore, all models are wrong, and bad models lead to bad answers: under the frequentist paradigm, the search for procedures with good frequentist properties provides some degree of protection against model misspecifica-

tion, but there seems no such built-in protection under a strict Bayesian paradigm where frequentist calculations are not entertained.

Good principles for picking models are essential, and here I feel frequentist methods have an important role. We want models that yield inferences with good frequentist properties, such as 95% credibility intervals that cover the unknown parameter 95% of the time if the procedure was applied to repeated samples. The Bayesian has some tools for model development and checking, like Bayes factors and model averaging, but Bayesian hypothesis testing has well known problems, and, in my view, frequentist approaches are essential when it comes to model development and assessment.

To summarize, Bayesian statistics is strong for inference under an assumed model, but relatively weak for the development and assessment of models. Frequentist statistics provides useful tools for model development and assessment, but has weaknesses for inference under an assumed model. If this summary is accepted, then the natural compromise is to use frequentist methods for model development and assessment, and Bayesian methods for inference under a model. This capitalizes on the strengths of both paradigms, and is the essence of the approach known as Calibrated Bayes (CB).

Many statisticians have advanced CB ideas (e.g., Peers, 1965; Welch, 1965; Dawid, 1982), but I was particularly influenced by seminal papers by Box (1980) and Rubin (1984). Box (1980) wrote,

> "I believe that... sampling theory is needed for exploration and ultimate criticism of the entertained model in the light of the current data, while Bayes' theory is needed for estimation of parameters conditional on adequacy of the model."

He based his implementation of this idea on the factorization:

$$p(Y, \theta|\text{Model}) = p(Y|\text{Model})\, p(\theta|Y, \text{Model}),$$

where the second term on the right side is the posterior distribution of the parameter θ given data Y and Model, and is the basis for inference, and the first term on the right side is the marginal distribution of the data Y under the Model, and is used to assess the validity of the Model, with the aid of frequentist considerations. Specifically, discrepancy functions of the observed data $d(Y)$ are assessed from the perspective of realizations from their marginal distribution $p(d(Y)|\text{Model})$. A questionable feature of this "prior predictive checking" is that checks are sensitive to the choice of prior distribution even when this choice has limited impact on the posterior inference; in particular, it leads to problems with assessment of models involving noninformative priors.

Rubin (1984) wrote,

> "The applied statistician should be Bayesian in principle and calibrated to the real world in practice—appropriate frequency calculations help to define such a tie... frequency calculations are useful for making Bayesian statements scientific, scientific in the sense of capable of being shown wrong by empirical test; here the technique is the calibration of Bayesian probabilities to the frequencies of actual events."

Rubin (1984) advocated model checking based on a different distribution, namely, $p(Y^*, \theta^*|Y, \text{Model})$, the predictive distribution of future data Y^* and parameter values θ^* given the Model and observed data Y. This leads to posterior predictive checks (Rubin, 1984; Gelman, Meng and Stern, 1996), which extend frequentist checking methods by not limiting attention to checking statistics that have a known distribution under the model. These checks involve an amalgam of Bayesian and frequentist ideas, but are clearly frequentist in spirit in that they concern embedding the observed data within a sequence of unobserved data sets that could have been generated under the Model, and seeing whether the observed data are "reasonable."

Philosophy aside, perhaps the main reason why Bayesian methods have flourished in recent years is the development of powerful computational tools, like the Gibbs' sampler and other Markov Chain Monte Carlo (MCMC) methods. These, together with gains in computing power, have made it computationally feasible to carry out the high-dimensional integrations required. An important early breakthrough in MCMC methods actually occurred for a missing data problem, as I discuss in Example 2 below. Even if frequentists are completely against Bayes, they can apply these Bayesian computational tools with weak prior distributions, and interpret results as approximations to ML, with similar asymptotic properties.

3. A SHORT HISTORY OF STATISTICAL ANALYSIS WITH MISSING DATA

I divide the development of missing data methods in statistics into four eras:

3.1 Pre-EM Algorithm (Pre-1970s)

Early missing data methods involved complete-case analysis, that is, simply discarding data with any values missing, or simple imputation methods, which filled in missing values with best estimates and analyzed the filled-in data. The latter approach was developed quite extensively in the case of analysis of variance with missing outcomes, which were imputed to maintain a balanced design and hence an easily inverted design matrix (see Little and Rubin, 2002, Chapter 2). These ingenious methods are now mainly of historical interest, since inverting the design matrix corresponding to the unbalanced data is not a big problem given advances in modern computing. ML methods were developed for some simple missing data problems, notably bivariate normal data with missing values on one variable, which Anderson (1957) solved noniteratively by factoring the likelihood (see Example 1 below). ML for complex problems was iterative and generally too hard given limits of computation, although progress was made for contingency tables (Hartley, 1958) and normal models (Hartley and Hocking, 1971).

3.2 The Maximum Likelihood Era (1970s–Mid 1980s)

ML methods became popular and feasible in the mid-1970s with the development of the Expectation–Maximization (EM) algorithm. EM builds a link with complete-data ML and is simple to program in several important multivariate models, including the multivariate normal model with a general pattern of missing values. The term EM was coined in the famous paper by Dempster, Laird and Rubin (1977), which established some key properties of the method, including the fact that the likelihood does not decrease at each iteration. The EM algorithm had been previously discovered several times for particular models (e.g., McKendrick, 1926; Hartley, 1958; Baum et al., 1970), and had been formulated in some generality by Orchard and Woodbury (1972) and Sundberg (1974). The simplicity and versatility of EM motivated extensions of EM to handle more difficult problems, and applications to a variety of complete-data models for categorical and continuous data, as reviewed in Little and Rubin (1987), McLachlan and Krishnan (1997) and Meng and van Dyk (1997). For generalizations of the EM idea, see Lange (2004).

Another important development in this era was the formulation of models for the missing data mechanism, and associated sufficient conditions for when the missing data mechanism can be ignored for frequentist and Bayesian inference (Rubin, 1976).

3.3 Bayes and Multiple Imputation (Mid 1980s–Present)

The transition from ML to Bayesian methods in the missing data setting was initiated by Tanner and Wong (1987), who described data augmentation to generate the posterior distribution of the parameters of the multivariate normal model with missing data. Data augmentation is closely related to the Gibbs' sampler, as discussed below. Another important development was Rubin's (1978, 1987, 1996) proposal to handle missing data in public use data sets by multiple imputation (MI), motivated by Bayesian ideas. In its infancy, this proposal seemed very exotic and computationally impractical—not any more! MCMC facilitates Bayesian multiple imputation, and is now implemented in publicly-available software for the convenience of users of both Bayesian and frequentist persuasions.

3.4 Robustness Concerns (1990s–Present)

Model-based missing data methods are potentially vulnerable to model misspecification, although they tend to outperform naïve methods even when the model is misspecified (e.g., Little, 1979). Modern interest in limiting effects of model misspecification by adopting robust procedures has extended to missing data problems, notably with "doubly robust" procedures based on augmented inverse-probability weighted estimating equations (Robins, Rotnitsky and Zhao, 1994). From a more directly model-based Bayesian perspective, robustness takes the form of developing models that make relatively weak structural assumptions. A method based on one such model, Penalized Spline of Propensity Prediction (PSPP, Little and An, 2004), is discussed in Example 4 below. Another aspect of robustness concerns has been more interest in model checks for standard missing data models (Gelman et al., 2005).

I now sketch the likelihood and Bayesian theory for the analysis of data with missing values that underlies the methods in Sections 3.2 and 3.3. I then describe the transition from Sections 3.2 to 3.4 for the case of multivariate normal models, and elaborations for nonnormal data.

4. LIKELIHOOD-BASED METHODS WITH MISSING DATA

Likelihoods can be defined for nonrectangular data, so likelihood methods apply directly to missing-data problems:

$$\text{statistical model} + \text{incomplete data} \Rightarrow \text{likelihood}.$$

Given the likelihood function, standard approaches are to maximize it, leading to ML, with associated large sample standard errors based on the information, the sandwich estimator or the bootstrap; or to add a prior distribution and compute the posterior distribution of the parameters. Draws from the predictive distribution of the missing values can also be created as a basis for MI.

As described in Little and Rubin (2002, Chapter 6), let $Y = (y_{ij})_{n\times K}$ represent a data matrix with n rows (cases) and K columns (variables), and define the missing-data indicator matrix $M = (m_{ij})_{n\times K}$, with entries $m_{ij} = 0$ if y_{ij} is observed, $m_{ij} = 1$ if y_{ij} is missing. Also, write $Y = (Y_{\text{obs}}, Y_{\text{mis}})$, where Y_{obs} represents the observed components of Y and Y_{mis} the missing components. A full parametric model factors the distribution of (Y, M) into a distribution $f(Y|\theta)$ for Y indexed by unknown parameters θ, and a distribution $f(M|Y, \psi)$ for M given Y indexed by unknown parameter ψ. (This is called a selection model factorization; the alternative factorization into the marginal distribution of M and the conditional distribution of Y given M is called a pattern-mixture model.) If Y was fully observed, the posterior distribution of the parameters would be

$$p_{\text{complete}}(\theta, \psi|Y, M) = \text{const.} \times \pi(\theta, \psi) \times L(\theta, \psi|Y),$$

where $\pi(\theta, \psi)$ is the prior distribution of the parameters,

$$L(\theta, \psi|Y) = f(Y|\theta) \times f(M|Y, \psi)$$

is the complete-data likelihood and const. is a normalizing constant. With incomplete data, the full posterior distribution becomes

$$\begin{aligned} &p_{\text{full}}(\theta, \psi|Y_{\text{obs}}, M) \\ &\quad\propto \pi(\theta, \psi) \times L(\theta, \psi|Y_{\text{obs}}, M), \end{aligned} \tag{1}$$

where $L(\theta, \psi|Y_{\text{obs}}, M)$ is the observed likelihood, obtained by integrating the missing values out of the complete-data likelihood:

$$\begin{aligned} &f(Y_{\text{obs}}, M|\theta, \psi) \\ &\quad= \int f(Y_{\text{obs}}, Y_{\text{mis}}|\theta) f(M|Y_{\text{obs}}, Y_{\text{mis}}, \psi)\, dY_{\text{mis}}. \end{aligned}$$

A simpler posterior distribution of θ ignores the missing data mechanism, and is based on the likelihood given the observed data Y_{obs}:

$$p_{\text{ign}}(\theta|Y_{\text{obs}}) \propto \pi(\theta) \times L(\theta|Y_{\text{obs}}), \tag{2}$$

$$L(\theta|Y_{\text{obs}}) = \int f(Y_{\text{obs}}, Y_{\text{mis}}|\theta)\, dY_{\text{mis}},$$

which does not involve the model for the distribution of M.

Statistical analysis based on (2) is considerably easier than analysis based on (1), since (a) the model for the missing-data mechanism is hard to specify, and has a strong influence on inferences; (b) the integration over the missing data is often easier for equation (2) than for equation (1); and (c) the full model is often under-identified or poorly identified; identification is in some ways less of an issue in Bayesian inference, but results rest strongly on the choice of prior distribution. Thus, ignoring the missing-data mechanism is useful if it is justified. Sufficient conditions for ignoring the missing-data mechanism and basing inference on (2) (Rubin, 1976; Little and Rubin, 2002, Chapter 6) are as follows:

Missing at Random (MAR): $p(M|Y_{\text{obs}}, Y_{\text{mis}}, \psi) = p(M|Y_{\text{obs}}, \psi)$ for all Y_{mis},
A-priori independence: $\pi(\theta, \psi) = \pi(\theta) \times \pi(\psi)$.

Of these, MAR is the key condition in practice, and it implies that missingness can depend on values in the data set that are observed, but not on values that are missing.

The main challenges in developing posterior distributions based on (1) or (2) are the choice of model and computational issues, since the likelihood based on the data with missing values is typically much more complex than the complete-data likelihood. In the ML world, the expectation–maximization (EM) algorithm creates a tie between the complicated observed data likelihood and the simpler complete-data likelihood, facilitating this computational task. Specifically, suppose the missing-data mechanism is ignorable, and let $\theta^{(t)}$ be the current estimate of the parameter θ. The E-step of EM finds the expected complete-data loglikelihood if θ equaled $\theta^{(t)}$:

$$\begin{aligned} &Q\big(\theta|Y_{\text{obs}}, \theta^{(t)}\big) \\ &\quad= \int \log f(Y_{\text{obs}}, Y_{\text{mis}}; \theta) \\ &\qquad\cdot f\big(Y_{\text{mis}}|Y_{\text{obs}}, \theta = \theta^{(t)}\big)\, dY_{\text{mis}}. \end{aligned} \tag{3}$$

The M-step of EM determines $\theta^{(t+1)}$ by maximizing this expected complete-data loglikelihood:

$$Q\big(\theta^{(t+1)}|Y_{\text{obs}}, \theta^{(t)}\big) \geq Q\big(\theta|Y_{\text{obs}}, \theta^{(t)}\big) \quad \text{for all } \theta. \tag{4}$$

In the Bayesian world, the analog of EM is data augmentation, a variant of the Gibbs' sampler. (An even closer analog to the Gibbs' sampler is the Expectation Conditional Maximization algorithm, a variant of EM.)

The key idea is to iterate between draws of the missing values and draws of the parameters; draws of missing values replace expected values of functions of the missing values in the E-step of EM, and draws of the parameters replace maximization over the parameters in the M-step of EM. Specifically, suppose the missing data mechanism is ignorable, and let $(Y_{\text{mis}}^{(dt)}, \theta^{(dt)})$ be current draws of the missing data and parameters at iteration t; here and elsewhere a superscript d is used to denote a draw from a distribution. Then at iteration $t+1$, the analog of the E-step (3) of EM is to draw new values of the missing data:

$$Y_{\text{mis}}^{(d,t+1)} \sim p\big(Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(dt)}\big), \tag{5}$$

and the analog of the M-step (4) is to draw a new set of parameters from the completed-data posterior distribution:

$$\theta^{(d,t+1)} \sim p\big(\theta|Y_{\text{obs}}, Y_{\text{mis}}^{(d,t+1)}\big). \tag{6}$$

As t tends to infinity, this sequence converges to a draw $(Y_{\text{mis}}^{(d)}, \theta^{(d)})$ from the joint posterior distribution of Y_{mis} and θ. Those familiar with MCMC methods will recognize this as an application of the Gibbs' sampler to the pair of variables Y_{mis} and θ. The utility lies in the fact that (5) is often facilitated since the distribution conditions on the parameters, and (6) is a complete-data problem since it conditions on the imputations derived from (5). For more discussion of Bayesian computations for missing data, see Tan and Tian (2010).

5. MULTIPLE IMPUTATION

Draws $Y_{\text{mis}}^{(d)}$ of the missing data from equation (5) at convergence can be used to create $D > 1$ multiply-imputed data sets. Bayesian MI combining rules can then be used for inferences that propagate imputation uncertainty.

I outline the theory when the missing data mechanism is ignorable, although it readily extends to the case of nonignorable nonresponse. The idea is to relate the observed-data posterior distribution (1) to the "complete-data" posterior distribution that would have been obtained if we had observed the missing data Y_{mis}, namely,

$$p(\theta|Y_{\text{obs}}, Y_{\text{mis}}) \propto \pi(\theta) \times L(\theta|Y_{\text{obs}}, Y_{\text{mis}}). \tag{7}$$

Equations (2) and (7) can be related by standard probability theory as

$$\begin{aligned} &p_{\text{ign}}(\theta|Y_{\text{obs}}) \\ &\quad = \int p(\theta|Y_{\text{obs}}, Y_{\text{mis}})\, p(Y_{\text{mis}}|Y_{\text{obs}})\, dY_{\text{mis}}. \end{aligned} \tag{8}$$

Equation (8) implies that the posterior distribution $p_{\text{ign}}(\theta|Y_{\text{obs}})$ can be simulated by first drawing the missing values, $Y_{\text{mis}}^{(d)}$, from their posterior distribution, $p(Y_{\text{mis}}|Y_{\text{obs}})$, imputing the drawn values to complete the data set, and then drawing θ from its "completed-data" posterior distribution, $p(\theta|Y_{\text{obs}}, Y_{\text{mis}}^{(d)})$. That is,

$$p_{\text{ign}}(\theta|Y_{\text{obs}}) \approx \frac{1}{D}\sum_{d=1}^{D} p\big(\theta|Y_{\text{obs}}, Y_{\text{mis}}^{(d)}\big). \tag{9}$$

When the posterior mean and variance are adequate summaries of the posterior distribution, (9) can be effectively replaced by

$$E(\theta|Y_{\text{obs}}) = E[E(\theta|Y_{\text{obs}}, Y_{\text{mis}})|Y_{\text{obs}}], \tag{10}$$

and

$$\begin{aligned} \operatorname{Var}(\theta|Y_{\text{obs}}) &= E[\operatorname{Var}(\theta|Y_{\text{obs}}, Y_{\text{mis}})|Y_{\text{obs}}] \\ &\quad + \operatorname{Var}[E(\theta|Y_{\text{obs}}, Y_{\text{mis}})|Y_{\text{obs}}]. \end{aligned} \tag{11}$$

Approximating (10) and (11) using draws of Y_{mis} yields

$$E(\theta|Y_{\text{obs}}) \approx \bar{\theta} = \frac{1}{D}\sum_{d=1}^{D} \hat{\theta}^{(d)}, \tag{12}$$

where $\hat{\theta}^{(d)} = E(\theta|Y_{\text{obs}}, Y_{\text{mis}}^{(d)})$ is the posterior mean of θ from the dth completed data set, and

$$\operatorname{Var}(\theta|Y_{\text{obs}}) \approx \bar{V} + (1 + 1/D)B, \tag{13}$$

say, where $\bar{V} = D^{-1}\sum_{d=1}^{D} \operatorname{Var}(\theta|Y_{\text{obs}}, Y_{\text{mis}}^{(d)})$ is the average of the complete-data posterior covariance matrix of θ calculated for the dth data set $(Y_{\text{obs}}, Y_{\text{mis}}^{(d)})$, $B = \sum_{d=1}^{D}(\hat{\theta}_d - \bar{\theta})(\hat{\theta}_d - \bar{\theta})^T/(D-1)$ is the between-imputation covariance matrix, and $(1 + 1/D)$ is a correction to improve the approximation for small D. The quantity $(1 + 1/D)B$ in (13) estimates the contribution to the variance from imputation uncertainty, missed (i.e., set to zero) by single imputation methods.

Equations (12) and (13) are basic MI combining rules. Refinements that replace the normal reference distribution for scalar θ by a Student t distribution are given in Rubin and Schenker (1986), with further small-sample t refinements in Barnard and Rubin (1999); extensions to hypothesis testing are described in Rubin (1987) or Little and Rubin (2002, Chapter 10). Besides incorporating imputation uncertainty, another benefit of multiple imputation is that the averaging over data sets in (12) results in more efficient point estimates than does single random imputation. Often MI is not much more difficult than doing a

single imputation—the additional computing from repeating an analysis D times is not a major burden and methods for combining inferences are straightforward. Most of the work is in generating good predictive distributions for the missing values.

From a frequentist perspective, Bayesian MI for a parametric model has similar large-sample properties to ML, and it can be simpler computationally. Another attractive feature of MI is that the imputation model can differ from the analysis model by including variables not included in final analysis. Some examples follow:

(a) MI was originally proposed for public use files, where the imputer often has variables available for imputation, like geography, that are not available to the analyst because of confidentiality constraints. Such variables can be included in the imputation model, but will not be available for analysis. In other settings, auxiliary variables that are not suitable for inclusion in the final model, such as side-effect data for drugs in a clinical trial, may be useful predictors in an imputation model.

(b) For public use files, users perform analyses with different subsets of variables. Different ML analyses involving a variable with missing values imply different imputation models, to the extent that they involve different sets of variables. A more coherent approach is to multiply impute missing variables using a common model, and then apply MI methods to each of the analyses involving subsets of variables. This allows variables not in the subset to help predict the missing values.

(c) MI combining rules can also be applied when the complete-data inference is not Bayesian (for example, nonparametric tests or design-based survey inference). The assumptions contained in the imputation model are then confined to the imputations, and with small amounts of missing data, simple imputation models may suffice.

6. APPLICATIONS OF BAYESIAN METHODS TO MISSING DATA PROBLEMS

Sections 4 and 5 sketched the basic theory of Bayesian inference for missing data and the related method of MI. We now provide some examples of important models where these methods can be put to practical use. We focus mainly on continuous variables, although methods for categorical variables, and mixtures of continuous and categorical variables, are also available (Little and Rubin, 2002).

EXAMPLE 1 (Data with a monotone pattern of missing values). I mentioned in Section 3.1 the factored likelihood method of Anderson (1957). Consider bivariate normal data on two variables (Y_1, Y_2) where Y_1 is observed for all n observations, and Y_2 is observed for $r < n$ observations, that is, has $n - r$ missing values. Assume the missing-data mechanism is ignorable. The factored likelihood is obtained by transforming the joint normal distribution of (Y_1, Y_2) into the marginal distribution of Y_1, normal with mean μ_1 and variance σ_{11}, and the conditional distribution of Y_2 given Y_1, normal with mean $\beta_{20\cdot1} + \beta_{21\cdot1}Y_1$ and variance $\sigma_{22\cdot1}$. The likelihood then factorizes into the normal likelihood for $\phi_1 = (\mu_1, \sigma_{11})$ based on the n cases with Y_1 observed, and the normal likelihood for $\phi_2 = (\beta_{20\cdot1}, \beta_{21\cdot1}, \sigma_{22\cdot1})$ based on the r cases with both Y_1 and Y_2 observed. The ML estimates are immediate: the sample mean and sample variance of Y_1 (with denominator n) based on all n observations for ϕ_1, and the least squares estimates of the regression of Y_2 on Y_1 (with no degrees of freedom correction for $\hat{\sigma}_{22\cdot1}$) based on the r complete cases for ϕ_2. ML estimates of other parameters, such as the mean μ_2 of Y_2, are obtained by expressing them as functions of (ϕ_1, ϕ_2) and then substituting ML estimates of those parameters. In particular, this leads to the well-known regression estimate of μ_2:

$$\hat{\mu}_2 = \hat{\beta}_{20\cdot1} + \hat{\beta}_{21\cdot1}\hat{\mu}_1, \tag{14}$$

which is easily seen to be obtained when missing values of Y_2 are imputed as predictions from the regression of Y_2 on Y_1, with regression coefficients estimated on the complete cases.

A corresponding Bayesian analysis is obtained by adding conjugate prior distributions for ϕ_1 and ϕ_2, and computing draws from the posterior distributions of these parameters. The posterior distribution of ϕ_1 is based on standard Bayesian methods applied to the sample of n complete observations on Y_1—inverse-chi-squared for σ_{11}, normal for μ_1 given σ_{11}, and Student's t for μ_1. The posterior distribution of ϕ_2 is based on standard Bayesian methods for the regression of Y_2 on Y_1 applied to the r complete observations on Y_1 and Y_2—inverse chi-squared for $\sigma_{22\cdot1}$, normal for $(\beta_{20\cdot1}, \beta_{20\cdot1})$ given $\sigma_{22\cdot1}$, and multivariate t for $(\beta_{20\cdot1}, \beta_{20\cdot1})$. Draws $(\mu_1^{(d)}, \sigma_{11}^{(d)}, \beta_{20\cdot1}^{(d)}, \beta_{21.1}^{(d)}, \sigma_{22\cdot1}^{(d)})$ from these posterior distributions are simple to compute (see Little and Rubin, 2002, Chapter 7, for details). Draws from the posterior distribution of other parameters are then created in the same way as ML estimates, by expressing the parameters as functions

of (ϕ_1, ϕ_2) and then substituting draws. For example, a draw from the posterior distribution of μ_2 is

$$\mu_2^{(d)} = \beta_{20\cdot1}^{(d)} + \beta_{21\cdot1}^{(d)}\mu_1^{(d)}, \tag{15}$$

a formula that mirrors the ML formula (14).

This Bayesian approach is asymptotically equivalent to ML, but it has several useful features. First, prior knowledge about the parameters can be incorporated in the prior distribution if this is available; if not, noninformative reference prior distributions can be applied. Second, the posterior distributions do a better job of capturing uncertainty in small samples; for example, the draws (15) incorporate t-like corrections, which are not provided by standard asymptotic ML calculations. Third, the draws yield immediate estimates of uncertainty, such as the posterior variance, and 95% credibility intervals. The factored likelihood approach does not yield conveniently simple formulas for large sample variances based on the information matrix. These are easily approximated by draws (15), and are actually superior (in a frequentist sense) to asymptotic variances since they reflect the uncertainty better.

Computational advantages in simulating draws from the posterior distribution are modest in the current bivariate normal example, since there are not many parameters. These benefits are more substantial in larger problems where the factored likelihood trick can be applied. Suppose that there are $K > 2$ variables $(Y_1, Y_2, \ldots, Y_K)$ such that (a) the data have a monotone pattern, such that Y_k is always observed when Y_{k+1} is observed, for $k = 1, \ldots, K-1$; and (b) the conditional distribution of $(Y_k|Y_1, \ldots, Y_{k-1})$ has a distribution (not necessarily normal) with unknown parameters ϕ_k, for $k = 1, \ldots, K$; and (c) the parameters $(\phi_1, \ldots, \phi_K)$ are distinct and have independent prior distributions. Draws from the posterior distribution of $\phi = (\phi_1, \ldots, \phi_K)$ can then be obtaining from a sequence of complete-data posterior distributions, with the posterior distribution of ϕ_k based on the subset of data with $(Y_1, \ldots, Y_k)$ observed (Little and Rubin, 2002, Chapter 7). This elegant scheme forms the basis for MI in the case of a monotone pattern. In particular, SAS PROC MI yields multiple imputations for normal models, where the regressions of Y_k on $Y_1, \ldots, Y_{k-1}$ are not required to be linear and additive, as would be the case if the joint distribution was multivariate normal.

When the data are monotone but the parameters of the sequence of conditional distributions are not a-priori independent, or when the pattern is not monotone, these simple factored likelihood methods no longer apply, and draws from the posterior distribution need an iterative scheme. Markov Chain Monte Carlo methods then come to the rescue, as in the next example.

EXAMPLE 2 (The multivariate normal model with a general pattern of missing values). Suppose observations y_i are assumed to be randomly sampled from a multivariate normal distribution, that is,

$$y_i = (y_{i1}, \ldots, y_{ip}) \sim_{\text{ind}} N_p(\mu, \Sigma),$$

the normal distribution with mean μ and covariance matrix Σ. There are missing values, and let $y_{\text{obs},i}$, $y_{\text{mis},i}$ denote respectively the set of observed and missing values for observation i. Given current draw $\theta^{(dt)} = (\mu^{(dt)}, \Sigma^{(dt)})$ of the parameters, missing values (5) are drawn as

$$y_{\text{mis},i}^{(d,t+1)} \sim p\big(y_{\text{mis},i}|y_{\text{obs},i}, \theta^{(dt)}\big), \quad i = 1, \ldots, n, \tag{16}$$

which is the multivariate normal distribution of the missing variables given the observed variables in observation i, with parameters that are functions of $\theta^{(dt)}$, readily computed using the sweep operator (Little and Rubin, 2002, Section 7.4). New parameters (6) are drawn from the posterior distribution given the filled-in data, which is a standard Bayesian problem, namely,

$$\Sigma^{(d,t+1)} \sim p\big(\Sigma|Y_{\text{obs}}, Y_{\text{mis}}^{(d,t+1)}\big), \tag{17}$$

$$\big(\mu^{(d,t+1)}|\Sigma^{(d,t+1)}\big) \sim p\big(\mu|\Sigma^{(d,t+1)}, Y_{\text{obs}}, Y_{\text{mis}}^{(d,t+1)}\big), \tag{18}$$

where (17) is a draw from an inverse Wishart distribution, and (18) is a draw from a multivariate normal distribution. Details of these steps were originally described in Tanner and Wong (1987) as part of their data augmentation algorithm, and they form the basis for the multiple imputation algorithm in SAS PROC MI, originally developed by Schafer (Schafer, 1997). Steps (16)–(18) are closely related to the EM algorithm for ML estimation, except that they lead to draws from the posterior distribution. When feasible as here, it is recommended to first program EM, and correct programming errors by checking that the likelihood increases with each iteration, and then convert the EM algorithm into the Gibbs algorithm, essentially by replacing the conditional means of the missing values in the E-step by draws (16), and the complete-data ML parameters in the M-step by draws (17) and (18). MI based on this model is available in a variety of software, including SAS PROC MI.

A frequentist statistician might compute ML estimates and associated standard errors based on the information matrix. However, the Gibbs algorithm outlined above is simpler than computing information-matrix based standard errors, which are not an immediate output from EM. So a frequentist can use the draws from Gibbs' algorithm to compute tests and confidence intervals for the parameters, exploiting the asymptotic equivalence of Bayes and frequentist inferences (Little and Rubin, 2002, Chapter 6). As in the previous example, the Bayesian approach improves some aspects of small sample inference by including t-like corrections reflecting uncertainty in the variance parameters.

Example 2 allows missing data to be multiply imputed for a general pattern of missing values, rather than the monotone pattern in Example 1. A limitation is that it assumes a multivariate normal distribution for the set of variables with missing values (normality can be relaxed for variables that are completely observed). This is a relatively strong parametric assumption—in particular, it assumes that the regressions of missing variables on observed variables are normal, linear and additive, which is not very appealing when a missing variable is binary or regressions involve interactions, for example.

One approach to this problem is to modify the model to allow for mixtures of continuous and categorical variables. The general location model of Olkin and Tate (1961) provides a useful starting point (Little and Rubin, 2002, Chapter 14). This is useful, but the need to formulate a tractable joint distribution for the variables is restrictive. A more flexible approach is to apply MI for a sequence of conditional regression models for each missing variable, given all the other variables. This sequential regression multivariate imputation (SRMI) method is only approximately Bayes, but what it loses in theoretical coherence it gains in practical flexibility. It is the topic of the next example.

EXAMPLE 3 (Sequential regression multivariate imputation). Suppose we have a general pattern of missing values, and $(Y_1, Y_2, \ldots, Y_K)$ are the set of variables with missing values, and X represents a set of fully observed variables. SRMI (Raghunathan et al., 2001; Van Buuren et al., 2006) specifies models for the distribution of each variable Y_j given all the other variables $Y_1, \ldots, Y_{j-1}, Y_{j+1}, \ldots, Y_K$ and X, indexed by parameters ψ_j, with density $g_j(Y_j|X, Y_1, \ldots, Y_{j-1}, Y_{j+1}, \ldots, Y_K, \psi_j)$, and a noninformative prior distribution for ψ_j. Missing values of Y_j at iteration $t+1$ are imputed according to the following scheme: let $y_{ji}^{(t)}$ be the observed or imputed value of Y_j at iteration t, and let $Y^{(jt)}$ denote the filled-in data set with imputations of $Y_1, \ldots, Y_{j-1}$ from iteration $t+1$ and imputations of $Y_j, \ldots, Y_K$ from iteration t. For $j = 1, \ldots, K$, create new imputations of the missing values of Y_j as follows:

$\psi_j^{(t+1)} \sim p(\psi|X, Y^{(jt)})$, the posterior distribution of given data $(X, Y^{(jt)})$;

$y_{ji}^{(t+1)} \sim g_j(Y_j|X, y_{1i}^{(t+s1)}, \ldots, y_{j-1,i}^{(t+1)}, y_{j+1,i}^{(t)}, \ldots, y_{Ki}^{(t)}, \psi_j^{(t+1)})$, if y_{ji} is missing, $i = 1, \ldots, n$.

This algorithm is repeated for $t = 1, 2, 3, \ldots$ until the imputations are stable; typically, more than one chain is run to facilitate assessment of convergence (Gelman and Rubin, 1992). The algorithm is then repeated D times to create D multiply-imputed data sets, and inferences are based on standard MI combining rules.

The positive feature of SRMI is that it reduces the multivariate missing data problems into a set of univariate problems for each variable given all the other variables, allowing flexibility in the choice of model for each incomplete variable; that is, nonlinear and interaction terms are allowed in the regressions, and the error distribution can be chosen to match the nature of the outcome—logistic for a binary variable, and so on. The drawback is that the regression models for each variable given the others does not generally correspond to a coherent model for the joint distribution of $(Y_1, \ldots, Y_K)$ given X. Thus, the MI's are not draws from a well-defined posterior distribution. This does not seem to be a major problem in practice, and SRMI is a flexible and practical tool for handling a variety of missing data problems. Software is available (Raghunathan, Solenberger and Van Hoewyk, 2009; MICE, 2009).

The regression models in SRMI are parametric, and potentially vulnerable to model misspecification. As noted in Section 3.4, one recent interest in missing data research has been the development of robust methods that do not involve strong parametric assumptions. My last example concerns so-called "doubly-robust methods" for missing data.

EXAMPLE 4 (Robust modeling: Penalized spline of propensity prediction). For simplicity, we consider the case where missingness is confined to a single variable Y. Let $(x_{i1}, \ldots, x_{ip}, y_i)$ be a vector of variables for observation i, with y_i observed for $i = 1, \ldots, r$ and missing for $i = r+1, \ldots, n$, and $(x_{i1}, \ldots, x_{ip})$ observed for $i = 1, \ldots, n$. We assume that the probability that y_i is missing depends on $(x_{i1}, \ldots, x_{ip})$

but not y_i, so the missing data mechanism is MAR. We consider estimation and inference for the mean of Y, $\mu_y = E(Y)$. Let m_i denote the missing-data indicator for y_i, with $m_i = 1$ when y_i is missing and $m_i = 0$ when y_i is observed.

A number of robust methods involve the propensity to be observed, estimated by a logistic or probit regression of M on $X_1, \ldots, X_p$ (Rosenbaum and Rubin, 1983; Little, 1986). In particular, propensity weighting computes the mean of the complete cases, weighted by the inverse of the estimated probability that Y is observed. Propensity weighting can yield estimates with large variances, and more efficient estimates are obtained by predicting the missing values of Y based on a model, with robustness supplied by a calibration term that weights the residuals from the complete cases (Robins, Rotnitzky and Zhao, 1994; Rotnitzky, Robins and Scharfstein, 1998; Bang and Robins, 2005; Scharfstein, Rotnitsky and Robins, 1999; Kang and Schafer, 2007). In this context, an estimator is doubly robust (DR) if it is consistent if either (a) the prediction model relating Y to $X_1, \ldots, X_p$ is correctly specified or (b) the model for the propensity to respond is correctly specified. In my last example, we describe a Bayesian missing data method, called Penalized Spline of Propensity Prediction (PSPP), that has a DR property.

Define the logit of the propensity score for y_i to be observed as

$$p_i^*(\psi) = \text{logit}\big(\Pr(m_i = 0 | x_{i1}, \ldots, x_{ip}; \psi)\big), \tag{19}$$

where ψ denotes unknown parameters. The PSPP method is based on the balancing property of the propensity score, which means the missingness of y_i depends only on $(x_{i1}, \ldots, x_{ip})$ only through the propensity score, under the MAR assumption (Rosenbaum and Rubin, 1983). Given this property, the mean of Y can be written as

$$\mu_y = E[(1 - m_i)y_i] + E\big[m_i \times E\big(y_i | p_i^*(\psi)\big)\big]. \tag{20}$$

Thus, the missing data can be imputed conditioning on the propensity score. This leads to the Penalized Spline of Propensity Prediction Method (PSPP) (Little and An, 2004; Zhang and Little, 2009). Imputations in this method are predictions from the following model:

$$\begin{aligned}&\big(y_i | p_i^*(\psi), x_{i1}, \ldots, x_{ip}; \psi, \beta, \phi, \sigma^2\big)\\ &\quad \sim N\big(\text{spl}(p_i^*(\psi), \beta)\\ &\qquad + g(p_i^*, x_{i2}, \ldots x_{ip}; \phi), \sigma^2\big),\end{aligned} \tag{21}$$

where $N(\nu, \sigma^2)$ denotes the normal distribution with mean ν and constant variance σ^2. The first component of the mean function, $\text{spl}(p_i^*(\psi), \beta)$, is a spline function of the propensity score $p_i^*(\psi)$. The second component $g(p_i^*, x_{i2}, \ldots, x_{ip}; \phi)$ is a parametric function, which includes any covariates other than p_i^* that predict y_i. One of the predictors, here x_{i1}, is omitted from the g-function to avoid multicollinearity.

A variety of spline functions can be chosen; we choose a penalized spline (Eilers and Marx, 1996; Ruppert, Wand and Carroll, 2003) of the form

$$\begin{aligned}\text{spl}(p_i^*(\psi), \beta) &= \beta_0 + \beta_1 p_i^*(\psi)\\ &\quad + \sum_{k=1}^{K} \beta_{k+1}\big(p_i^*(\psi) - \kappa_k\big)_+,\end{aligned} \tag{22}$$

where $1, p_i^*(\psi), (p_i^*(\psi) - \kappa_1)_+, \ldots, (p_i^*(\psi) - \kappa_k)_+$ is the truncated linear basis; $\kappa_1 < \cdots < \kappa_K$ are selected fixed knots and K is the total number of knots, and $(\beta_2, \ldots, \beta_{K+1})$ are random effects, assumed normal with mean 0 and variance τ^2. This model can be fitted by ML using a number of existing software packages, such as PROC MIXED in SAS (SAS, 1992; Ngo and Wand, 2004; Littell et al., 1996) and lme$(\cdot)$ in S-plus (Pinheiro and Bates, 2000). The first step of fitting a PSPP model estimates the propensity score, for example, by a logistic regression model or probit model of M on $X_1, \ldots, X_p$; in the second step, the regression of Y on P^* is fit as a spline model with the other covariates included in the model parametrically in the g-function. When Y is a continuous variable we choose a normal distribution with constant variance. For other types of data, extensions of the PSPP can be formulated by using the generalized linear models with different link functions.

The average of the observed and imputed values of Y has a DR property, meaning that the predicted mean of Y is consistent if either (a) the mean of y_i given $(p_i^*(\psi), x_{i1}, \ldots, x_{ip})$ in model (3) is correctly specified, or (b1) the propensity $p_i^*(\psi)$ is correctly specified, and (b2) $E(y_i | p_i^*(\psi), \beta) = \text{spl}(p_i^*(\psi), \beta)$. The robustness feature derives from the fact that the regression function g does not have to be correctly specified, and the spline part of the regression function involves a weak parametric assumption, practically similar to the DR property mentioned above (Little and An, 2004; Zhang and Little, 2009).

How does Bayes play into these methods? The missing values of y_i can be multiply-imputed under this model, but note that these imputations involve a substantial number of unknown parameters, namely, the

regression coefficients and variances (β, τ^2) of the spline model, the regression coefficients ϕ in the parametric component g, the residual variance σ^2, and the nuisance parameters ψ in the propensity function. Uncertainty in these parameters is readily propagated under the Bayesian paradigm by drawing them from their posterior distributions, which is reasonably straightforward using a Gibbs' sampler.

7. CONCLUSION

I began this article by summarizing some arguments in favor of the CB approach to statistical inference, which to my mind incorporates the best features of both Bayesian and frequentist statistics. In short, inferences should be Bayesian, but model development and checking requires careful attention to frequentist properties of the resulting Bayesian inference. This CB "roadmap" is not a complete solution, since the interplay between model development and model inference, involving questions such as what range of model uncertainty should be included as part of formal statistical inference (Draper, 1995), remains illusive and hard to pin down. However, I find the CB approach a very satisfying basis for approaching practical statistical inference.

In the remainder of the article I argued that the CB approach is particularly attractive for dealing with problems of missing data. In a sense, all of inferential statistics is a missing data problem, since it involves making inferences about something that is unknown and hence "missing"; in that broad sense, I am merely restating the previous paragraph. However, if missing data is considered more restrictively as referring to situations where the data matrix is incomplete, or partial information is available on some variables, then the Bayesian paradigm is conceptually straightforward, since likelihoods do not require a fully-recorded rectangular data set.

Simply put, Bayesian statistics involves generating predictive distributions of unknowns given the data. Applied to missing data, it requires a predictive distribution for the missing data given the observed data. Multiple imputations are simply draws from this predictive distribution, and can be used for other analyses if a good model is chosen for the predictive distribution.

In Sections 4 and 5 I outlined the basic theory underlying Bayes inference and MI with missing data, emphasizing the role of the MAR assumption. An important extension of this theory is to problems of coarsened data, where some values in the data set are rounded, grouped or censored (Heitjan and Rubin, 1991; Heitjan, 1994). This theory has connections with the concept of noninformative censoring that underlies many methods of survival analysis with censored data. MI can be applied in these settings (Hsu et al., 2006), and is particularly useful for conditioning imputations on auxiliary variables not included in the primary analysis.

In Section 6 I illustrated Bayesian approaches to missing data, mainly for normal models, in view of their practical importance and historical interest. I emphasize that Bayesian methods are also useful for nonnormal missing data problems. The SRMI methods are not restricted to normal models, and Bayes and/or MI can be applied to categorical data models and mixtures of categorical and continuous variables (Little and Rubin, 2002, Chapters 13 and 14), and generalized linear models with missing covariates (Ibrahim et al., 2005). Bayesian hierarchical models are also natural for longitudinal data (Gilks et al., 1993; Ibrahim and Molenberghs, 2009) and small area estimation (Ghosh and Rao, 1994), with or without missing data.

In the examples I focused on MAR models, but Bayesian approaches to NMAR models are also very appealing. A key problem when data are not missing at random is lack of identifiability of the model, and Bayesian methods provide a formal solution to the problem by allowing the formulation of prior distributions for unidentified parameters that reflect the uncertainty (Rubin, 1977; Daniels and Hogan, 2008). Resulting inferences are arguably superior to frequentist methods based on sensitivity analysis, where unidentified parameters are varied over a range. For a recent illustration of these ideas, see Long, Little and Lin (2010), who apply Bayesian missing data methods to handle noncompliance in clinical trials.

I have focused here more on the Bayesian part of CB, given the emphasis of the workshop that motivated this article on Bayesian tools. Concerning the calibrated part, good frequentist properties require realistic models for the predictive distribution of the missing values, and this requires attention to checking the fit of the model to the observed data, and sensitivity analyses to assess the impact of departures from MAR. Gelman et al. (2005) and Abayomi, Gelman and Levy (2008) provide useful methods for model checking for multiple imputation, and more work in this area would be useful.

ACKNOWLEDGMENTS

Useful comments from two referees on an earlier draft are greatly appreciated, and helped to improve the article.

REFERENCES

ABAYOMI, K., GELMAN, A. and LEVY, M. (2008). Diagnostics for multivariate imputations. *Appl. Statist.* **57** 273–291. MR2440009

AGRESTI, A. (2002). *Categorical Data Analysis*, 2nd ed. Wiley, New York. MR1914507

ANDERSON, T. W. (1957). Maximum likelihood estimates for the multivariate normal distribution when some observations are missing. *J. Amer. Statist. Assoc.* **52** 200–203. MR0087286

BANG, H. and ROBINS, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61** 962–972. MR2216189

BARNARD, J. and RUBIN, D. B. (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika* **86** 949–955. MR1741991

BAUM, L. E., PETRIE, T., SOULES, G. and WEISS, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist.* **41** 164–171. MR0287613

BOX, G. E. P. (1980). Sampling and Bayes inference in scientific modelling and robustness (with discussion). *J. Roy. Statist. Soc. Ser. A* **143** 383–430. MR0603745

DANIELS, M. J. and HOGAN, J. W. (2008). *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis*. CRC Press, New York. MR2459796

DAWID, A. P. (1982). The well-calibrated Bayesian. *J. Amer. Statist. Assoc.* **77** 605–610. MR0675887

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B* **39** 1–38. MR0501537

DRAPER, D. (1995). Assessment and propagation of model uncertainty (with discussion). *J. Roy. Statist. Soc. Ser. B* **57** 45–97. MR1325378

EFRON, B. (1986). Why isn't everyone a Bayesian? (with discussion and rejoinder). *Amer. Statist.* **40** 1–11. MR0828575

EILERS, P. H. C. and MARX, B. D. (1996). Flexible smoothing with B-Splines and penalties. *Statist. Sci.* **11** 89–121. MR1435485

GELMAN, A. E., MECHELEN, I. V., VERBEKE, G., HEITJAN, D. F. and MEULDERS, M. (2005). Multiple imputation for model checking: Completed-data plots with missing and latent data. *Biometrics* **61** 74–85. MR2135847

GELMAN, A., MENG, X.-L. and STERN, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statist. Sinica* **6** 733–807. MR1422404

GELMAN, A. E. and RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statist. Sci.* **7** 457–472. MR1092987

GHOSH, M. and RAO, J. N. K. (1994). Small area estimation: An appraisal. *Statist. Sci.* **9** 55–76. MR1278679

GILKS, W. R., WANG, C. C., YVONNET, B. and COURSAGET, P. (1993). Random-effects models for longitudinal data using Gibbs' sampling. *Biometrics* **49** 441–453.

HARTLEY, H. O. (1958). Maximum likelihood estimation from incomplete data. *Biometrics* **14** 174–194.

HARTLEY, H. O. and HOCKING, R. R. (1971). The analysis of incomplete data. *Biometrics* **27** 783–808.

HEITJAN, D. F. (1994). Ignorability in general complete-data models. *Biometrika* **81** 701–708. MR1326420

HEITJAN, D. and RUBIN, D. B. (1991). Ignorability and coarse data. *Ann. Statist.* **19** 2244–2253. MR1135174

HSU, C. H., TAYLOR, J. M., MURRAY, S. and COMMENGES, D. (2006). Survival analysis using auxiliary variables via nonparametric multiple imputation. *Stat. Med.* **25** 3503–3517. MR2252407

IBRAHIM, J. G., CHEN, M.-H., LIPSITZ, S. R. and HERRING, A. H. (2005). Missing data methods in generalized linear models: A comparative review. *J. Amer. Statist. Assoc.* **100** 332–346. MR2166072

IBRAHIM, J. G. and MOLENBERGHS, G. (2009). Missing data methods in longitudinal studies: A review (with discussion). *Test* **18** 1–43. MR2495958

KANG, D. Y. and SCHAFER, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci.* **22** 523–539. MR2420458

LANGE, K. (2004). *Optimation*. Springer, New York. MR2072899

LITTELL, R. C., MILLIKEN, G. A., STROUP, W. W. and WOLNGER, R. D. (1996). *SAS System for Mixed Models*. SAS Institute Inc., Cary, NC.

LITTLE, R. J. (1979). Maximum likelihood inference for multiple regression with missing values: A simulation study. *J. Roy. Statist. Soc. Ser. B* **41** 76–87. MR0535548

LITTLE, R. J. A. (1986). Survey nonresponse adjustments for estimates of means. *Internat. Statist. Rev.* **54** 139–157.

LITTLE, R. J. A. (1988). Small sample inference about means from bivariate normal data with missing values. *Comput. Statist. Data Anal.* **7** 161–178. MR0991002

LITTLE, R. J. A. (2006). Calibrated Bayes: A Bayes/frequentist roadmap. *Amer. Statist.* **60** 213–223. MR2246754

LITTLE, R. J. A. and AN, H. (2004). Robust likelihood-based analysis of multivariate data with missing values. *Statist. Sinica* **14** 949–968. MR2089342

LITTLE, R. J. A. and RUBIN, D. B. (1987). *Statistical Analysis with Missing Data*, 1st ed. Wiley, New York. MR0890519

LITTLE, R. J. A. and RUBIN, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. Wiley, New York. MR1925014

LONG, Q., LITTLE, R. J. and LIN, X. (2010). Estimating the CACE in trials involving multi-treatment arms subject to noncompliance: A Bayesian framework. *J. Roy. Statist. Soc. Ser. C.* To appear.

MCKENDRICK, A. G. (1926). Applications of mathematics to medical problems. *Proc. Edinburgh Math. Soc.* **44** 98–130.

MCLACHLAN, G. J. and KRISHNAN, T. (1997). *The EM Algorithm and Extensions*. Wiley, New York. MR1417721

MENG, X. L. and VAN DYK, D. (1997). The EM algorithm—An old folk song sung to a fast new tune (with discussion). *J. Roy. Statist. Soc. Ser. B* **59** 511–567. MR1452025

MICE (2009). Multiple imputation via chained equations. Available at http://www.multiple-imputation.com.

NGO, L. and WAND, M. P. (2004). Smoothing with mixed model software. *J. Statist. Software* **V9** Issue 1.

OLKIN, I. and TATE, R. F. (1961). Multivariate correlation models with mixed discrete and continuous variables. *Ann. Math. Statist.* **32** 448–465. MR0152062

ORCHARD, T. and WOODBURY, M. A. (1972). A missing information principle: Theory and applications. In *Proc. 6th Berkeley Symposium on Mathematical Statistics and Probabality* **1** 697–715. Univ. California Press, Berkeley, CA. MR0400516

PEERS, H. W. (1965). On confidence points and Bayesian probability points in the case of several parameters. *J. Roy. Statist. Soc. Ser. B* **27** 9–16. MR0191029

PINHEIRO, J. C. and BATES, D. M. (2000). *Mixed-Effects Models in S and S-PLUS*. Springer, New York.

RAGHUNATHAN, T. E., LEPKOWSKI, J. M., VAN HOEWYK, J. and SOLENBERGER, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology* **27** 85–95.

RAGHUNATHAN, T. E., SOLENBERGER, P. W. and VAN HOEWYK, J. (2009). IVEware: Imputation and variance estimation software. Available at http://www.isr.umich.edu/src/smp/ive/.

ROBINS, J. M., ROTNITZKY, A. and ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89** 846–866. MR1294730

ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974

ROTNITZKY, A., ROBINS, J. M. and SCHARFSTEIN, D. O. (1998). Semiparametric regression for repeated measures outcomes with non-ignorable non-response. *J. Amer. Statist. Assoc.* **93** 1321–1339. MR1666631

RUBIN, D. B. (1976). Inference and missing data (with discussion). *Biometrika* **63** 581–592. MR0455196

RUBIN, D. B. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys. *J. Amer. Statist. Assoc.* **72** 538–543. MR0445668

RUBIN, D. B. (1978). Multiple imputations in sample surveys. In *Proceedings of the Survey Research Methods Section* 20–34. Amer. Statist. Assoc., Alexandria, VA.

RUBIN, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.* **12** 1151–1172. MR0760681

RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York. MR0899519

RUBIN, D. B. (1996). Multiple imputation after 18+ years. *J. Amer. Statist. Assoc.* **91** 473–489.

RUBIN, D. B. and SCHENKER, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *J. Amer. Statist. Assoc.* **81** 366–374. MR0845877

RUPPERT, D., WAND, M. P. and CARROLL, R. J. (2003). *Semiparametric Regression*. Cambridge Univ. Press, Cambridge. MR1998720

SAS (1992). The Mixed Procedure, Chapter 16 in SAS/STAT Software: Changes and Enhancements, Release 6.07. Technical Report P-229, SAS Institute, Inc., Cary, NC.

SCHAFER, J. L. (1997). *Analysis of Incomplete Multivariate Data.* CRC Press, New York. MR1692799

SCHARFSTEIN, D., ROTNITSKY, A. and ROBINS, J. (1999). Adjusting for nonignorable dropout using semiparametric models (with discussion). *J. Amer. Statist. Assoc.* **94** 1096–1146. MR1731478

SUNDBERG, R. (1974). Maximum likelihood theory for incomplete data from an exponential family. *Scand. J. Statist.* **1** 49–58. MR0381110

TAN, M. T. and TIAN, G.-L. (2010). *Bayesian Missing Data Problems: EM, Data Augmentation and Noniterative Computation*. CRC Press, New York. MR2562244

TANNER, M. A. and WONG, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *J. Amer. Statist. Assoc.* **82** 528–550. MR0898357

VAN BUUREN, S., BRAND, J. P. L., GROOTHUIS-OUDSHOORN, K. and RUBIN, D. B. (2006). Fully conditional specification in multivariate imputation. *J. Statist. Comput. Simul.* **76** 1049–1064. MR2307507

WELCH, B. L. (1965). On comparisons between confidence point procedures in the case of a single parameter. *J. Roy. Statist. Soc. Ser. B* **27** 1–8. MR0187327

ZHANG, G. and LITTLE, R. J. (2009). Extensions of the penalized spline propensity prediction method of imputation. *Biometrics* **65** 911–918. DOI: 10.1111/j.1541-0420.2008.01155. MR2649864