

# Missing Data: Dial M for ???

Xiao-Li MENG

The question mark is common notation for the missing data that occur in most applied statistical analyses. Over the past century, statisticians and other scientists not only have invented numerous methods for handling missing/incomplete data, but also have invented many forms of missing data, including data augmentation, hidden states, latent variables, potential outcome, and auxiliary variables. Purposely constructing unobserved/unobservable variables offers an extraordinarily flexible and powerful framework for both scientific modeling and computation and is one of the central statistical contributions to natural, engineering, and social sciences. In parallel, much research has been devoted to better understanding and modeling of real-life missing-data mechanisms; that is, the unintended data selection process that prevents us from observing our intended data. This article is a very brief and personal tour of these developments, and thus necessarily has much missing history and citations. The tour consists of a number of M's, starting with a historic story of the mysterious method of McKendrick for analyzing an epidemic study and its link to the EM algorithm, the most popular and powerful method of the twentieth century for fitting models involving missing data and latent variables. The remaining M's touch on theoretical, methodological, and practical aspects of missing-data problems, highlighted with some common applications in social, computational, biological, medical, and physical sciences.

## 1. McKENDRICK, A MYSTERY, AND EM

Table 1, adopted from Meng (1997), tells a fascinating story of missing data from the early part of the twentieth century. The first two rows describe an epidemic of cholera in an Indian village, where $x$ represents the number of cholera cases within a household and $n_x$ is the corresponding observed number of such households. Prior to presenting this example, McKendrick (1926) derived a Poisson model for such data. However, the direct Poisson fit, reported in the third row, is so poor that any goodness-of-fit method that fails to reject the Poisson model must itself be rejected.

Had McKendrick (1926) settled for the simple Poisson model, it would not have been the earliest citation in the seminal paper on the EM algorithm by Dempster, Laird, and Rubin (1977), nor would it have been reprinted in *Breakthroughs in Statistics*, Vol. III (Kotz and Johnson 1997). McKendrick's approach seems rather mysterious, especially because he did not provide a derivation. He first calculated $s_1 = \sum_x x n_x = 86, s_2 = \sum_x x^2 n_x = 166$, and

$$\hat{n} = \frac{s_1^2}{s_2 - s_1} = 92.45. \tag{1}$$

Next, he treated $\hat{n} \approx 93$ as the Poisson sample size, and thus estimated the Poisson mean by $\tilde{\lambda} = s_1/\hat{n} = .93$. The fitted counts were then calculated via $\hat{n}\tilde{\lambda}^x \exp(-\tilde{\lambda})/x!$, as given in the fourth row of Table 1. The fit is evidently very good for $x \geq 1$, but exhibits an astonishingly large discrepancy for $x = 0$.

This large discrepancy gives a clue to McKendrick's approach—there were too many 0's for the simple Poisson model to fit. Earlier a lieutenant-colonel in the Indian Medical Service and then a curator of the College of Physicians at Edinburgh, McKendrick had astute insight into the excess 0's, as he wrote that "[T]his suggests that the disease was probably water borne, that there were a number of wells, and that the inhabitants of 93 out of 223 houses drank from one well which was infected." In other words, a household can have no cases of cholera either because it was never exposed to cholera or because it was exposed but no member of it was infected. The existence of these unexposed households complicates the analysis, for without external information, one cannot distinguish an unexposed from an exposed but uninfected household. To a modern statistician, this immediately suggests using the binomial/Poisson mixture model, also known as the zero-inflated Poisson (ZIP) model (see, e.g., Böhning, Dietz, and Schlattmann 1999), which models a binomial indicator for the exposure status and, conditional on being exposed, a Poisson variable as before. Although McKendrick (1926) was not explicit, he fit a zero-truncated Poisson (ZTP) model; that is, he ignored the observed $n_0 = 168$ zero-class count and used remaining data under the Poisson model to *impute* the unobserved zero-class count from the exposed population. Once he had the imputed total, the rest is history. The ingenious part of McKendrick's approach is his imputation of the total Poisson size $n$ via (1), which equates the sample variance with the sample mean. Neither $s_1$ nor $s_2$ is affected by the actual value of $n_0$, yet $\lim_{n\to\infty} \hat{n}/n = \lambda^2/(\lambda^2 + \lambda - \lambda) = 1$, and thus $\hat{n}$ is a consistent imputation/estimate of the true Poisson sample size $n$. The mystery is then unfolded.

The key ingredients of McKendrick's approach are to first identify a missing data structure, perhaps constructed, then impute the missing data, and finally analyze the completed data set as if there were no missing data. This procedure is a predecessor of many modern missing-data methods. A key advance of modern methods, thanks to enormously improved computing power, is iterative repetition

Xiao-Li Meng is Professor, Department of Statistics, University of Chicago, IL 60637 (E-mail: *meng@galton.uchicago.edu*). This research is supported in part by National Science Foundation grant DMS 96-26691 and National Security Agency grant MDA 904-99-1-0067. The author thanks George Casella for the invitation to write this Y2K vignette and for comments, and also thanks Radu Craiu, Dan Heitjan, Mary Sara McPeek, Dan Nicolae, William Rosenberger, Jay Servidea, and David van Dyk for comments and/or proofreading.


Journal of the American Statistical Association
December 2000, Vol. 95, No. 452, Vignettes

*Table 1. Data and Fitted Values For McKendrick's Problem*

| $x$ | 0 | 1 | 2 | 3 | 4 | ≥5 | Total |
|---|---|---|---|---|---|---|---|
| $n_x$ | 168 | 32 | 16 | 6 | 1 | 0 | 223 |
| Direct Poisson fit | 151.64 | 58.48 | 11.28 | 1.45 | 0.00 | .01 | 223 |
| McKendrick's fit | 36.89 | 34.11 | 15.77 | 4.86 | 1.12 | .24 | 93 |
| MLE fit | 33.46 | 32.53 | 15.81 | 5.12 | 1.25 | .29 | 88.46 |

of such types of processes, as in the EM algorithm, or multiple repetitions, as with multiple imputation (Rubin 1987). The need for this iteration/repetition was recognized by Irwin (1963), who noted that once an estimator of $\lambda$ was obtained via McKendrick's approach, $n$ can be reimputed by $n^{(t+1)} = n^{(t)} \exp(-\lambda^{(t)}) + n_{\text{obs}}$, where $n_{\text{obs}} = \sum_{x\geq 1} n_x$ and $t$ indexes iteration, and in turn $\lambda$ can be reestimated via $\lambda^{(t+1)} = s_1/n^{(t+1)}$. Irwin's method, though not a special case, resembles the two-step EM algorithm. In the Expectation step, the complete-data log-likelihood $l(\theta|Y_{\text{com}})$ is imputed by its conditional expectation $Q(\theta|\theta^{(t)}) = E[l(\theta|Y_{\text{com}})|\theta^{(t)}, Y_{\text{obs}}]$, where $Y_{\text{obs}}$ is the observed data. In the Maximization step, $Q(\theta|\theta^{(t)})$ is maximized as a function of $\theta$ to determine $\theta^{(t+1)}$. For the ZTP model, $\theta = \lambda$, and the M step is the same as McKendrick's or Irwin's; that is, $\lambda^{(t+1)} = s_1/n^{(t+1)}$. The E step is an improved version of Irwin's imputation; that is, $n^{(t+1)} = n_{\text{obs}}/(1 - \exp(-\lambda^{(t)}))$. Combining the E and M steps yields $\lambda^{(t+1)} = (s_1/n_{\text{obs}})(1 - \exp(-\lambda^{(t)}))$, which, like Irwin's method, converges to the maximum likelihood estimate, $\hat{\lambda} = .972$, as long as $\lambda^{(0)} > 0$. The fifth row of Table 1 gives the corresponding fit.

McKendrick's problem also highlights the celebrated idea of data augmentation when one adopts the binomial/Poisson mixture model. Because a complete sample is available from this model, there are no missing data in the traditional sense. Nonetheless, we can view the mixture/exposure indicator as missing data and construct the corresponding EM algorithm (Meng 1997). Purposely constructing missing data, such as mixture indicators, random effects, and latent factors, is a key contribution of Dempster et al. (1977) and has seen an enormous number of applications in statistical and scientific studies, as illustrated in Sections 4–6. This, along with a large number of recent improvements and extensions of EM (see Liu, Rubin, and Wu 1998; McLaughlan and Krishnan 1997; Meng and van Dyk 1997; and the references therein) have served to substantially increase the applicability and speed of EM-type algorithms.

## 2. MISSING-DATA MECHANISM

A profound difficulty in dealing with real-life missing-data problems is to reasonably understand and model the *missing-data mechanism* (MDM), namely the process that prevents us from observing our intended data. This process is a data selection process, like a sampling process, yet because it is typically not controlled by or even unknown to the data collector, it can be subject to all kinds of (hidden) biases, known collectively as *nonresponse bias.* Although the general theoretical foundation of sampling processes existed in the early part of the twentieth century (e.g., Neyman 1934) and the impact of selection bias (e.g., from a purposive selection) has long been understood, the corresponding foundation for MDM was not formally developed until much later, starting with Rubin (1976a). Two key mechanisms introduced by Rubin, namely missing at random (MAR) and missing completely at random (MCAR), now appear in most statistical articles that contain analyses of incomplete data, often even without citation. These concepts have also been extended (see, e.g., Heitjan 1997, 1999; Heitjan and Rubin 1991).

Assuming MCAR basically means that we believe the observed data are a random subsample of the intended sample, and thus we can analyze it just as we analyze the intended sample, only with reduced size. Because this assumption is generally very far from the truth, common convenient approaches such as ignoring any case with missing values can be strongly biased (see, e.g., Little and Rubin 1987). MAR is a much weaker assumption, which allows the MDM to depend on observed quantities, but not on unobserved quantities. Under MAR, we can ignore the MDM in a likelihood inference based on the observed data without inducing nonresponse bias (but possibly inducing inefficiency when there is a priori dependence between the estimand and parameters governing the MDM, i.e., when the *parameter distinctness* assumption of Rubin (1976a) is violated; see Shih 1994). However, for sampling-based inference, it generally requires MCAR to ignore the MDM (see Heitjan and Basu (1996) for illustrations.)

When the MDM is not MAR (and thus not MCAR), the probability of missingness depends on the unobserved values themselves. The MDM is then generally not ignorable, meaning that the validity of our inference depends crucially on the particular model of the MDM we adopt. Because ignorability is fundamentally untestable from the observed data alone, one must exercise great caution when drawing substantive conclusions from any inference under a nonignorable model. Sensitivity analysis to the specification of an MDM model is a necessity, and subjective knowledge can play a critical role, as illustrated by Molenberghs, Goetghebeur, Lipsitz, and Kenward (1999). Modeling nonignorable MDMs is currently a very active research area with many open problems (see, e.g., Heitjan 1999; Ibrahim, Lipsitz, and Chen 1999; Mohlenberghs, Kenward, and Lesaffre 1997; Scharfstein, Rotnitzky, and Robins 1999).

## 3. MULTIPLE IMPUTATION AND UNCONGENIALITY

The common usage of *nonresponse bias* for general biases induced by an MDM reflects the historical fact that nonresponse in sample surveys is the most visible missing-data problem in general practice, especially in social sciences. Thanks to the efforts made by many statisticians and social scientists throughout the twentieth century, we are seeing fewer and fewer articles using convenient missing-data "methods" such as mean imputation and complete-case analyses without acknowledging their potential serious flaws. On the other hand, the simplicity of these "methods" is so attractive that preventing practitioners from being seduced requires scientifically and statistically more defensi-

ble methods with comparable simplicity. Multiple imputation (Rubin 1987) was motivated by this need. In the context of public-use or shared databases, the first step of Rubin's multiple imputation is to have the data collector build a sensible imputation model given available data and knowledge about the MDM, which are typically far more comprehensive than what could possibly be available to an average user (e.g., Barnard and Meng 1999; Meng 1994a; Rubin 1996). The data collector then draws $M$ (e.g., 5–10) independent samples of all the missing values, as a set, from the imputation model, thereby creating $M$ completed-data sets and thus permitting general users to directly assess and account for the increased variability/uncertainties due to nonresponse. For a user, analyzing a multiply imputed dataset means conducting $M$ separate complete-data analyses, one for each of the $M$ completed-data sets, and then combining these $M$ completed-data analysis outputs using a few general rules. Readers are referred to Schafer (1999) for an updated tutorial; Gelman, King, and Liu (1998) for a recent application in public opinion polls; and Schafer (1997) for a comprehensive treatment of the practical implementation of Rubin's multiple imputation, including software.

In the context of public-use data files, there is a crucial separation between the data collector/imputer and general users. The two parties typically have different goals, data, information, and assessments and thus often adopt different models or even different modes of inference. Consequently, the imputation model is usually *uncongenial* to the user's analysis procedure; that is, the latter cannot be embedded into a (Bayesian) model that is compatible with the imputation model (Meng 1994a). One way to reduce this uncongeniality, of course, is to encourage more information exchange, such as having the imputer provide additional imputation quantities beyond the imputed datasets (e.g., Meng 1994a; Robins and Wang 2000; Schafer and Schenker 2000). Although this is clearly a direction for more research, the practical constraints (e.g., confidentiality, a user's choice of inferential mode) ensure the issue of uncongeniality will always remain. Consequently, much research is needed to establish a more flexible "multiparty" paradigm for comparing and evaluating statistical procedures, a paradigm that promotes most effective procedures given resource and practical constraints, rather than those that are misguided by impossible idealizations (see Rubin 1996), even if such idealizations are sensible in a congenial environment.

## 4. MCMC AND PERFECT SIMULATION

Constructing unobserved variables—namely, the method of data augmentation (e.g., Tanner and Wong 1987) or of auxiliary variables (e.g., Besag and Green 1993; Edwards and Sokal 1988)—has played a critical role in the development of efficient Markov chain Monte Carlo (MCMC) algorithms. Some recent findings (e.g., Liu and Wu 1999; Meng and van Dyk 1999; van Dyk and Meng 2000) demonstrate the seemingly limitless potential of this method. Here I briefly describe one of its uses for perfect simulation, a rapidly growing area of MCMC—the list of references in *http://dimacs.rutgers.edu/~dbwilson/exact* is updated constantly.

Perfect simulation, or exact sampling, refers to a class of MCMC algorithms that in finite time provide genuine and independent draws from the limiting (i.e., stationary) distribution of a Markov chain. This seemingly impossible task was made possible by the *backward coupling* method of Propp and Wilson (1996) which, in a very rough sense, is a clever stochastic counterpart of the deterministic method for finding the optimizer by comparing the value of the objective function at each point. Consequently, this class of methods is most effective with finite-state chains, though they are by no means restricted to such chains (e.g., Green and Murdoch 1999; Murdoch 2000; Murdoch and Green 1998).

Indeed, data augmentation can help us to transform a continuous state-space problem into a finite one. For example, suppose that we are interested in simulating from $p(X)$ and we can augment this model to $p(X, Y)$ such that both $p(X|Y)$ and $p(Y|X)$ are easy to sample and the augmented variable $Y$ is discrete. We can then implement a two-step Gibbs sampler, which induces a marginal Markov chain on $Y$. Because $Y$ has a finite state space, in some cases we can directly implement the backward coupling method with this discrete chain to obtain iid draws from $p(Y)$. The desired iid draws from $p(X)$ are then obtained easily by drawing from $p(X|Y)$ given the draws of $Y$'s. If the state space of $Y$ is too large for the direct backward coupling method, then one can try multistage backward coupling (Meng 2000). An immediate application of this approach is to Bayesian finite mixtures (joint work with D. Murdoch), where $Y$ is the subpopulation indicator. Readers are referred to Møller and Nicholls (1999) and the references therein for other methods of using discrete hidden variables to make perfect simulation effective for routine applications in statistics.

## 5. MENDELIAN LIKELIHOODS AND RELATIVE INFORMATION

It was exactly one century ago when Mendel's basic theory of heredity was rediscovered and gained general recognition (e.g., McPeek 1996; Thompson 1996), marking the real birth of modern genetics. The theory of Mendel (1866) provides general principles for probabilistic modeling of the inheritance of genes from parents to offspring. However, in common pedigree analyses, we typically miss some data on genotype information (e.g., allele types at some genetic markers), on the genealogical tree (e.g., whether an allele was from the paternal side or the maternal side), and even on phenotype (e.g., a disease status of an ancestor). Consequently, Mendelian modeling and the associated computation are intrinsically problems of missing data, typically of very high dimension with exceedingly complex structures. Monte Carlo simulation is an effective general approach for such problems, but its efficiency depends on the choice of underlying data augmentation scheme. Two common, and sometime competing, choices are genotypes and inheritance vectors/meiosis indicators, which indicate whether the origin of the gene is grandpaternal or grandmaternal (see Lander and Green 1987; Thompson 1994, 2000). Once we ob-

tain draws of the missing data $Y_{\text{mis}}$ from $p(Y_{\text{mis}}|Y_{\text{obs}}, \theta)$ (e.g., via MCMC), the computation of the observed-data Mendelian likelihood ratio $L(\theta_1|Y_{\text{obs}})/L(\theta_2|Y_{\text{obs}})$ can be dealt with effectively via bridge sampling (Bennett 1976; Jensen and Kong 1999; Meng and Wong 1996) and the reweighted mixture method of Geyer (1991). A currently challenging problem is to make such methods more accurate for computing the likelihood ratio when $\theta_1$ and $\theta_2$ belong to different marker regions (Thompson 2000), and the warp bridge sampling method (Meng 1999; Meng and Schilling 1999) is a possible direction to explore, because it increases efficiency by increasing the overlaps of the underlying densities by warping their shapes. The use of bridge sampling for assessing the convergence of Monte Carlo EM (Wei and Tanner 1990), which is useful for genetic linkage analysis (Guo and Thompson 1992), was detailed by Meng and Schilling (1996).

Another important missing-data problem in genetic linkage analysis is estimating the amount of information in the observed data *relative* to the total amount of information that would have been available had there been no missing data. The statistical literature on the fraction of missing information has been largely focused on its theoretical properties (e.g., Dempster et al. 1977; Liu 1994; Meng 1994b) and methodological uses (e.g., Meng and Rubin 1991) in computation and estimation. However, the focus here is more on design, with the aim of directly guiding follow-up strategies; for example, using more genetic markers with existing DNA samples versus collecting DNA samples from additional families, by assessing how much more information could be obtained if, say, we add more markers. An additional difference is that, because hypothesis testing is a useful screening tool for linkage studies (e.g., Thompson 1996), we need to measure the relative information in the context of hypothesis testing. This requires considering the roles of both the null hypothesis (i.e., no linkage) and the alternative hypothesis (as specified by a trait model). Although this issue does not arise in the estimation context, the basic identities given by Dempster et al. (1977) are fundamental for establishing a general theoretical framework for studying relative information in the context of hypothesis testing; details will be given elsewhere (as a joint paper with A. Kong and D. Nicolae).

## 6. MAPPING THE BRAIN AND THE UNIVERSE

Image reconstruction, a critical component in many medical and physical studies, is fundamentally another class of missing data problems. In the medical imaging context, perhaps the best-known example to statisticians is positron emission tomography (PET), for which the use of the EM algorithm signifies statisticians' direct involvement in the developing stage of the technique (e.g., Lange and Carson 1984; Shepp and Vardi 1982). The overview given by Vardi, Shepp, and Kaufman (1985), using brain mapping as an example, showed the intrinsic missing-data nature of PET, for we cannot directly observe the count of photons emitted from each pixel (i.e., a location in the brain). In addition, we face missing-data problems such as attenuation by the body's tissues and the escape of photons that travel along lines that do not intersect with any detector. As with linkage analysis, the choice of data augmentation schemes, or hidden-data spaces, has a direct impact on the speed of computation. As an example, Fessler and Hero (1995) discussed clever choices of hidden-data spaces that have made EM-type reconstructions more practical, overcoming the slowness of early EM reconstructions that use individual pixel counts as hidden states. An algorithmic analysis of the method of Fessler and Hero was given by Meng and van Dyk (1997) in the framework of the AECM algorithm.

Similar imaging techniques also play an important role in astrophysics, where the use of EM-type algorithms, such as the Richardson–Lucy algorithm (Lucy 1974; Richardson 1972), predates the publication of Dempster et al. (1977), though the development of fast EM and related Bayesian imaging algorithms has just begun (e.g., van Dyk 1999; van Dyk, Connors, Kashyap, and Siemiginowska 2001). The Poisson spectral imaging model, designed to analyze data from the Chandra observatory (lunched on the space shuttle *Columbia,* July 1999) and other upcoming detectors, is an example of needing efficient methodologies for handling data from the new generation of high-resolution satellite telescopes. The Poisson model here is designed to summarize the relative frequency of photon energies (x-ray or $\gamma$-ray), collected as counts in a number of bins, arriving at a detector. The detected photons originate from many sources (e.g., a "continuum" and a number of "line profiles") and have been subject to background contamination, instrument response, and stochastic absorption. Each of these distortions requires a layer of modeling (e.g., Poisson, multinomial), forming an overall *multilevel* hierarchical model for the observed binned energies, a typical situation with real-data latent-variable modeling. Each of these levels, as well as any combination or function of them, is a candidate for data augmentation in fitting the model. An efficient choice can substantially improve the computational speed; van Dyk (1999) and van Dyk and Meng (1999) gave details and empirical evidences.

## 7. MILLENNIUM WISHES

The topic of missing data is as old and as extensive as statistics itself—after all, statistics is about knowing the unknowns. It is thus impossible in a few pages to discuss all of the main areas of past and present research. Areas not discussed here include, among many others, noniterative methods (e.g., Baker, Rosenberger, and DerSimonian 1992; Rubin 1976b), direct maximization of observed-data likelihoods (e.g., Molenberghs and Goetghebeur 1997), pattern-mixture models (e.g., Little 1993), bootstrap methods (e.g., Efron 1994), estimating equation approaches (e.g., Heyde and Morton 1996; Robins, Rotnitzky, and Zhao 1994; Lipsitz, Ibrahim and Zhao 1999), and potential outcome in causal inferences (e.g., Barnard, Du, Hill, and Rubin 1998; Rubin 1978). Consequently, the 82 references listed in this article are really just the tip of the iceberg—even with many missing articles, Meng and Pedlow (1992) found more than 1,000 EM-related articles, about 85% of which were in non-statistical journals. The number must have doubled by now.

Much remains to be done, however. The most pressing task, in my opinion, is placing further emphasis on the general recognition and understanding, at a conceptual level, of the necessity of properly dealing with the missing-data mechanism, as part of our ongoing emphasis on the importance of the data collection process in any meaningful statistical analysis. The missing-data mechanism is in the blood of statistics, and it is the nastiest and the most deceptive cell, especially for nonstatisticians—why on earth should anyone be concerned with data that one does not even have? I conclude with an excerpt from a referee's report of Tu, Meng, and Pagano (1994), to make one of my wishes for the new millennium. Reports like this will soon be of great value, but only on auction.

> The statement, "The naive approach of ignoring the missing data and using only the observed portion could provide very misleading conclusions" is nonsense to me (and I think the authors should also recognize it as nonsense in the real world). Similarly, what does it mean, "When analyzing such missing data, . . ."; if the data are missing, you can't analyze them. Except for old, rigid, demanding, clunky data treatment methods like the Yates algorithm (and except for the ridge problems discussed), it is unlikely that ". . . the analysis could still be very complicated due to the unbalanced structure of the observed data . . ." (page 4). Does any chemometrician every (sic) worry about making "it possible to utilize computer routines already developed for complete-data maximization"? I don't think any chemometricians every (sic) use data-specific data-treatment methods.

(To purchase a copy of this referee report, please dial M for Meng!)

## REFERENCES


Baker, S. G., Rosenberger, W. F., and DerSimonian, R. (1992), "Closed-Form Estimates for Missing Counts in Two-Way Contingency Tables," *Statistics in Medicine*, 11, 643–657.

Barnard, J., Du, J., Hill, J. L., and Rubin, D. B. (1998), "A Broader Template for Analyzing Broken Randomized Experiments," *Sociological Methods and Research*, 27, 285–317.

Barnard, J., and Meng, X-L. (1999), "Applications of Multiple Imputation in Medical Studies: From AIDS to NHANES," *Statistical Methods in Medical Research*, 8, 17–36.

Bennett, C. H. (1976), "Efficient Estimation of Free Energy Differences From Monte Carlo Data," *Journal of Computational Physics*, 22, 245–268.

Besag, J., and Green, P. J. (1993), "Spatial Statistics and Bayesian Computation," *Journal of the Royal Statistical Society*, Ser. B, *Methodological*, 55, 25–37.

Böhning, D., Dietz, E., and Schlattmann, P. (1999), "The Zero-Inflated Poisson Model and the Decayed, Missing and Filled Teeth Index in Dental Epidemiology," *Journal of the Royal Statistical Society*, Ser. B, 162, 195–209.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 39, 1–37.

Efron, B. (1994), "Missing Data, Imputation, and the Bootstrap," *Journal of the American Statistical Association*, 89, 463–475.

Edwards, R., and Sokal, A. (1988), "Generalization of the Fortuin–Kasteleyn–Swendsen–Wang Representation and Monte Carlo Algorithm," *Physical Review Letters*, 28, 2009–2012.

Fessler, J. A., and Hero, A. O. (1995), "Penalized Maximum-Likelihood Image Reconstruction Using Space-Alternating Generalized EM Algorithm," *IEEE Transactions on Image Processing*, 4, 1417–1438.

Gelman, A., King, G., and Liu, C. (1998), "Not Asked and Not Answered: Multiple Imputation for Multiple Surveys" (with discussion), *Journal of the American Statistical Association*, 93, 846–874.

Geyer, C. (1991), "Reweighting Monte Carlo Mixtures," Technical Report 568, University of Minnesota, School of Statistics.

Green, P. J., and Murdoch, D. J. (1999), "Exact Sampling for Bayesian Inference: Towards General Purpose Algorithms," in eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Bayesian Statistics 6, London: Oxford University Press.

Guo, S. W., and Thompson, E. A. (1992), "A Monte Carlo Method for Combined Segregation and Linkage Analysis," *American Journal of Human Genetics*, 51, 1111–1126.

Heitjan, D. F. (1994), "Ignorability in General Incomplete-Data Models," *Biometrika*, 81, 701–708.

——— (1997), "Ignorability, Sufficiency and Ancillarity," *Journal of the Royal Statistical Society*, Ser. B, 59, 375–381.

——— (1999), "Causal Inference in Clinical Trials, A Comparative Example," *Controlled Clinical Trials*, 20, 309–318.

Heitjan, D. F., and Basu, S. (1996), "Distinguishing 'missing at random' and 'missing completely at random'," *The American Statistician*, 50, 207–213.

Heitjan, D. F., and Rubin, D. B. (1991), "Ignorability and Coarse Data," *The Annals of Statistics*, 19, 2244–2253.

Heyde, C. C., and Morton, R. (1996), "Quasi-likelihood and Generalizing the EM Algorithm," *Journal of the Royal Statistical Society*, Ser. B, 58, 317–327.

Ibrahim, J. G., Lipsitz, S. R., and Chen, M-H. (1999), "Missing Covariates in Generalized Linear Models When the Missing Data Mechanism is Nonignorable," *Journal of the Royal Statistical Society*, Ser. B, 61, 173–190.

Irwin, J. O. (1963), "The Place of Mathematics in Medical and Biological Statistics," *Journal of the Royal Statistical Society*, Ser. B, 126, 1–45.

Jensen, C. S., and Kong, A. (1999), "Blocking Gibbs Sampling for Linkage Analysis in Large Pedigrees With Many Loops," *American Journal of Human Genetics*, 65, 885–901.

Kotz, S., and Johnson, N. L. (1997), *Breakthroughs in Statistics, Vol. III*, New York: Springer.

Lander, E. S., and Green, P. (1987), "Construction of Multilocus Genetic Linkage Maps in Humans," *Proceedings of the National Academy of Sciences*, 84, 2363–2367.

Lange, K., and Carson, R. (1984), "EM Reconstruction Algorithms for Emission and Transmission Tomography," *Journal of Computer Assisted Tomography*, 8, 306–316.

Lipsitz, S. R., Ibrahim, J. G., and Zhao, L. P. (1999), "A Weighted Estimating Equation for Missing Covariate Data With Properties Similar to Maximum Likelihood," *Journal of the American Statistical Association*, 94, 1147–1160.

Little, R. J. (1993), "Pattern-Mixture Models for Multivariate Incomplete Data," *Journal of the American Statistical Association*, 88, 125–134.

Little, R. J., and Rubin, D. B. (1987), *Statistical Analysis With Missing Data*, New York: Wiley.

Liu, C., Rubin, D. B., and Wu, Y. N. (1998), "Parameter Expansion for EM Acceleration—the PXEM Algorithm," *Biometrika*, 75, 755–770.

Liu, J. S. (1994), "Fraction of Missing Information and Convergence Rate of Data Augmentation," in *Computationally Intensive Statistical Methods: Proceedings of the 26th Symposium Interface*, pp. 490–497.

Liu, J. S., and Wu, Y. N. (1999), "Parameter Expansion Scheme for Data Augmentation," *Journal of the American Statistical Association*, 94, 1264–1274.

Lucy, L. B. (1974), "An Iterative Technique for the Rectification of Observed Distributions," *Astronomical Journal*, 79, 745–754.

McKendrick, A. G. (1926), "Applications of Mathematics to Medical Problems," *Proceedings of the Edinburgh Mathematics Society*, 44, 98–130.

McLaughlan, G. J., and Krishnan, T. (1997), *The EM Algorithm and Extensions*, New York: Wiley.

McPeek, M. S. (1996), "An Introduction to Recombination and Linkage Analysis," in *Genetic Mapping and DNA Sequencing*, eds. T. Speed and M. Waterman, New York: Springer, pp. 1–14.

Mendel, G. J. (1866), *Experiments in Plant Hybridisation*, English trans., Edinburgh: Oliver and Boyd, 1965.

Meng, X-L. (1994a), "Multiple Imputation Inferences With Uncongenial Sources of Input" (with discussion), *Statistical Sciences*, 9, 538–573.

——— (1994b), "On the Rate of Convergence of the ECM Algorithm," *The Annals of Statistics*, 22, 326–339.

——— (1997), "The EM Algorithm and Medical Studies: A Historical Link," *Statistical Methods in Medical Research*, 6, 3–23.

——— (1999), "Invited discussions of Matther Stephens's and Simon Tavaré's papers on Statistical and Computational Approaches to Genetic Evolution," in *Bulletin of the International Statistical Institute; 52nd Session, Helsinki*. Available at http://www.stat.fi/isi99/proceedings.html

——— (2000), "Toward a More General Propp–Wilson Algorithm: Multistage Backward Coupling," *Fields Institute Communications Series Vol. 26: Monte Carlo Methods*, American Mathematical Society.

Meng, X-L., and Pedlow, S. (1992), "EM: A Bibliographic Review With Missing Articles," in *Proceedings of the Statistical Computing Section, American Statistical Association*, pp. 24–27.

Meng, X-L., and Rubin, D. B. (1991), "Using EM to Obtain Asymptotic Variance-Covariance Matrices: The SEM Algorithm," *Journal of the American Statistical Association*, 86, 899–909.

Meng, X-L., and Schilling, S. (1996), "Fitting Full-Information Item Factor Models and an Empirical Investigation of Bridge Sampling," *Journal of the American Statistical Association*, 91, 1254–1267.

——— (1999), "Warp Bridge Sampling," *Revised for The Journal of Computational and Graphical Statistics*.

Meng, X-L., and van Dyk, D. A. (1997), "The EM Algorithm—An Old Folk Song Sung to a Fast New Tune" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 59, 511–567.

——— (1999), "Seeking Efficient Data Augmentation Schemes via Conditional and Marginal Augmentation," *Biometrika*, 86, 301–320.

Meng, X-L., and Wong, W. H. (1996), "Simulating Ratios of Normalizing Constants via a Simple Identity: A Theoretical Explanation," *Statistica Sinica*, 6, 831–860.

Molenberghs, G., and Goetghebeur, E. (1997), "Simple Fitting Algorithms for Incomplete Categorical Data," *Journal of the Royal Statistical Society*, Ser. B, 59, 401–414.

Molenberghs, G., Goetghebeur, E. J., Lipsitz, S. R., and Kenward, M. G. (1999), "Nonrandom Missingness in Categorical Data: Strengths and Limitations," *The American Statistician*, 53, 110–118.

Molenberghs, G., Kenward, M. G., and Lesaffre, E. (1997), "The Analysis of Longitudinal Ordinal Data With Nonrandom Drop-Out," *Biometrika*, 84, 33–44.

Møller, J., and Nicholls, G. K. (1999), "Perfect Simulation for Sample-Based Inference," Preprint.

Murdoch, D. J. (2000), "Exact Sampling for Bayesian Inference: Unbounded State Space," *Fields Institute Communications Series Vol. 26: Monte Carlo Methods*, American Mathematical Association.

Murdoch, D. J., and Green, P. J. (1998), "Exact Sampling From a Continuous State Space," *Scandinavian Journal of Statistics*, 25, 483–502.

Neyman, J. (1934), "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection" (with discussion), *Journal of the Royal Statistical Society*, Ser. A, 97, 558–625.

Propp, J. G., and Wilson, D. B. (1996), "Exact Sampling With Coupled Markov Chains and Applications to Statistical Mechanics," *Random Structures and Algorithms*, 9, 1, 2, 223–252.

Richardson, W. H. (1972), "Bayesian-Based Iterative Methods of Image Restoration," *Journal of the Optical Society of America*, 62, 55–59.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994), "Estimation of Regression Coefficients When Some Regressors are not Always Observed," *Journal of the American Statistical Association*, 89, 846–866.

Robins, J. M., and Wang, N. (2000), "Inference for Imputation Estimators," *Biometrika*, 87, 113–124.

Rubin, D. B. (1976a), "Inference and Missing Data," *Biometrika*, 63, 581–592.

——— (1976b), "Noniterative Least Squares Estimates, Standard Errors and $F$ Tests for Analyses of Variance With Missing Data," *Journal of the Royal Statistical Society*, Ser. B, 38, 270–274.

——— (1978), "Bayesian Inference for Causal Effects: The Role of Randomization," *The Annals of Statistics*, 6, 34–58.

——— (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley.

——— (1996), "Multiple Imputation After 18+ Years" (with discussion), *Journal of the American Statistical Association*, 91, 473–489.

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

——— (1999), "Multiple Imputation: A Primer," *Statistical Methods in Medical Research*, 8, 3–15.

Schafer, J. L., and Schenker, N. (2000), "Inference With Imputed Conditional Means," *Journal of the American Statistical Association*, 95, 144–154.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999), "Adjusting for Nonignorable Drop-out Using Semiparametric Nonresponse Models" (with discussion), *Journal of the American Statistical Association*, 94, 1096–1146.

Shepp, L. A., and Vardi, Y. (1982), "Maximum Likelihood Reconstruction for Emission Tomography," *IEEE Transactions on Image Processing*, 2, 113–122.

Shih, W. J. (1994), Discussion of "Informative Drop-Out in Longitudinal Data Analysis," by Diggle and Kenward, *Applied Statistics*, 43, 87.

Tanner, M. A., and Wong, W. H. (1987), "The Calculation of Posterior Distributions by Data Augmentation" (with discussion), *Journal of the American Statistical Association*, 82, 528–550.

Thompson, E. A. (1994), "Monte Carlo Likelihood in Genetic Mapping," *Statistical Science*, 9, 355–366.

——— (1996), "Likelihood and Linkage: From Fisher to the Future," *The Annals of Statistics*, 24, 449–465.

——— (2000), *Statistical Inference From Genetic Data*, Heyward, CA: Institute of Mathematical Statistics.

Tu, X. M., Meng, X-L., and Pagano, M. (1994), "On the Use of Conditional Maximization in Chemometrics," *Journal of Chemometrics*, 8, 365–370.

van Dyk, D. A. (1999), "Fast New EM-Type Algorithms With Applications in Astrophysics," technical report, Department of Statistics, Harvard University.

van Dyk, D. A., Connors, A., Kashyap, V. L., and Siemiginowska, A. (2001), "Analysis of Energy Spectrum With Low Photon Counts via Bayesian Posterior Simulation," *Astrophysical Journal*, to appear.

van Dyk, D. A., and Meng, X-L. (1999a), "Algorithms Based on Data Augmentation: A Graphical Representation and Comparison," in *Models, Predictions, and Computing: Proceedings of the 31st Symposium on the Interface*, eds. M. Pourahmadi and K. Berk, pp. 230–239.

——— (2001), "The Art of Data Augmentation" (with discussion), *Journal of Computational and Graphical Statistics*, to appear.

Vardi, Y., Shepp, L. A., and Kaufman, L. (1985), "A Statistical Model for Positron Emission Tomography," *Journal of the American Statistical Association*, 80, 8–19.

Wei, G., and Tanner, M. A. (1990), "A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithm," *Journal of the American Statistical Association*, 85, 699–704.