Missing Data, Imputation, and the Bootstrap

Bradley EFRON*

Missing data refers to a class of problems made difficult by the absence of some portions of a familiar data structure. For example, a regression problem might have some missing values in the predictor vectors. This article concerns nonparametric approaches to assessing the accuracy of an estimator in a missing data situation. Three main topics are discussed: bootstrap methods for missing data, these methods' relationship to the theory of multiple imputation, and computationally efficient ways of executing them. The simplest form of nonparametric bootstrap confidence interval turns out to give convenient and accurate answers. There are interesting practical and theoretical differences between bootstrap methods and the multiple imputation approach, as well as some useful similarities.

KEY WORDS: Bayesian bootstrap; Bootstrap confidence intervals; Data augmentation; Ignorable nonresponse; Nonparametric MLE.

1. INTRODUCTION

Missing data refers to a class of problems made difficult by the absence of some part of a familiar data structure. In a correlation problem, for example, some of the data pairs might be missing one of the measurements. A substantial theory of imputation has been developed during the past 15 years to efficiently estimate a parameter of interest θ in a missing data situation and to assess the variability of the estimate $\hat{\theta}$. This theory, as developed by Rubin (1987), Tanner and Wong (1987), and several other authors mentioned in Section 3, is based on an appealing Bayesian analysis of the missing data structure. In this article the bootstrap, a frequentist device, is brought to bear on missing data problems, with a particular emphasis on nonparametric situations. There are interesting practical and theoretical differences between the bootstrap and imputation approaches, as well as some similarities.

The left panel of Table 1 presents a simple example of a missing data situation. Twenty-two students have each taken five exams, labeled A, B, C, D, E, but some of the A and the E scores marked "?," have been lost. We suppose that if there were no missing data, then we would consider the rows of the matrix to be a random sample of size $n = 22$ from an unknown five-dimensional probability distribution F, and that our goal is to estimate θ, the maximum eigenvalue of the covariance matrix of F,

$$\theta = \text{maximum eigenvalue of } \mathfrak{V}, \tag{1.1}$$

where $\mathfrak{V}$ is the covariance matrix of a single vector drawn from F.

The right panel of Table 1 shows an imputed data set in which the missing student scores have been replaced by estimates obtained from a two-way linear model. Parameter estimates $\hat{\nu}$, $\hat{\alpha}_i$, and $\hat{\beta}_j$ were obtained by minimizing the sum of squares

$$\sum_i \sum_j [o_{ij} - (\nu + \alpha_i + \beta_j)]^2, \qquad \left(\sum_1^{22} \alpha_i = 0, \sum_1^5 \beta_j = 0\right), \tag{1.2}$$

the sum being over the observed numerical scores o_{ij} in the left panel, those having a numerical value and not a question mark. Then the imputed values

$$\hat{x}_{ij} = \hat{\nu} + \hat{\alpha}_i + \hat{\beta}_j \tag{1.3}$$

were used to replace the missing scores $o_{ij} = ?$, of course setting $\hat{x}_{ij} = o_{ij}$ for the observed numerical scores. This is not necessarily a good imputation scheme, as will be discussed, but it is typical of "best-fit" imputation methods often used in practical situations (see Little and Rubin 1987, chap. 2). For now we will take it as a given part of the example under discussion.

The imputed data set consists of $n = 22$ rows (say, $\hat{x}_i$ for $i = 1, 2, \ldots, n$), from which we can calculate an empirical covariance matrix

$$\hat{\mathfrak{V}} = \frac{1}{n}\sum_{i=1}^{n} (\hat{x}_i - \hat{\mu})(\hat{x}_i - \hat{\mu})' \qquad \left(\hat{\mu} = \frac{1}{n}\sum_1^n \hat{x}_i\right). \tag{1.4}$$

The maximum eigenvalue of $\hat{\mathfrak{V}}$ is an obvious estimate $\hat{\theta}$ for the parameter of interest θ, in this case giving

$$\hat{\theta} = \text{maximum eigenvalue of } \hat{\mathfrak{V}} = 633.2. \tag{1.5}$$

We could just as well divide by $n - 1$ instead of n in defining $\hat{\mathfrak{V}}$, but (1.4) makes $\hat{\theta}$ equal the normal theory maximum likelihood estimate of θ, which is handy for later comparisons.

The calculation of $\hat{\theta}$ illustrates the traditional purpose of imputation: to replace missing values with numbers so that familiar statistical methods can be used. Here the familiar method is the covariance calculation (1.4). The question remains, "How good an estimate of θ is $\hat{\theta}$?" This article concerns bootstrap approaches to answering this question, particularly in nonparametric situations.

The simplest nonparametric bootstrap approach, sometimes called just the *nonparametric bootstrap* in what follows, begins by writing the observed data set as

$$\mathbf{o} = (o_1, o_2, \ldots, o_n). \tag{1.6}$$

Here o_i is the ith row of the matrix in the left panel of Table 1, including the question marks; for example, $o_1 = (?, 63, 65, 70, 63)$. A *nonparametric bootstrap sample,*

$$\mathbf{o}^* = (o_1^*, o_2^*, \ldots, o_n^*), \tag{1.7}$$

is obtained by drawing the o_i^* randomly and with replacement from the set $\{o_1, o_2, \ldots, o_n\}$ (see Efron and Tibshirani

* Bradley Efron is Professor of Statistics, Stanford University, Stanford, CA 94305. I am grateful to Paul Holland for several helpful discussions concerning missing data.

Journal of the American Statistical Association
June 1994, Vol. 89, No. 426, Theory and Methods

Table 1. The Student Score Data

	Observed Data **o**					Imputed Data $\hat{\mathbf{x}}$				
Student	A	B	C	D	E	A	B	C	D	E
1	?	63	65	70	63	56.21	63	65	70	63
2	53	61	72	64	73	53	61	72	64	73
3	51	67	65	65	?	51	67	65	65	58.94
4	?	69	53	53	53	47.96	69	53	53	53
5	?	69	61	55	45	48.46	69	61	55	45
6	?	49	62	63	62	49.96	49	62	63	62
7	44	61	52	62	?	44	61	52	62	51.69
8	49	41	61	49	?	49	41	61	49	46.94
9	30	69	50	52	45	30	69	50	52	45
10	?	59	51	45	51	42.46	59	51	45	51
11	?	40	56	54	?	39.54	40	56	54	44.33
12	42	60	54	49	?	42	60	54	49	48.19
13	?	63	53	54	?	46.21	63	53	54	50.99
14	?	55	59	53	?	45.21	55	59	53	49.99
15	?	49	45	48	?	36.87	49	45	48	41.66
16	17	53	57	43	51	17	53	57	43	51
17	39	46	46	32	?	39	46	46	32	37.69
18	48	38	41	44	33	48	38	41	44	33
19	46	40	47	29	?	46	40	47	29	37.44
20	30	34	43	46	18	30	34	43	46	18
21	?	30	32	35	21	20.46	30	32	35	21
22	?	26	15	20	?	9.87	26	15	20	14.66

NOTE: Left panel: 22 students have each taken 5 exams, labeled A, B, C, D, and E. Some of the scores for A and E, indicated by "?," are missing. Right panel: The missing data have been imputed from a two-way additive model. The full data set, taken from Mardia, Kent, and Bibby (1979), appears in table 1 of Efron (1992a).

1986). Then we can use the two-way fitting algorithm (1.2), (1.3) to impute the missing values in $\mathbf{o}^*$ (giving, say, $\hat{\mathbf{x}}^*$), and compute the bootstrap covariance matrix

$$\hat{\Sigma}^* = \frac{1}{n}\sum_{i=1}^{n}(\hat{x}_i^* - \hat{\mu}^*)(\hat{x}_i^* - \hat{\mu}^*)' \qquad \left(\hat{\mu}^* = \frac{1}{n}\sum \hat{x}_i^*\right). \tag{1.8}$$

Finally, we calculate $\hat{\theta}^*$, the maximum eigenvalue of $\hat{\Sigma}^*$, a *bootstrap replication* of $\hat{\theta}$.

A computer program drew 2,200 independent bootstrap samples $\mathbf{o}^*$ and evaluated $\hat{\theta}^*$ for each one. Figure 1 shows the histogram of the 2,200 $\hat{\theta}^*$ values, the histogram being notably long-tailed to the right. These have empirical standard deviation 212.0, which by definition is the bootstrap estimate of standard error for $\hat{\theta}$. The average $\hat{\theta}^*$ value is 610.3, giving a bootstrap bias estimate $-22.9 = 610.3 - \hat{\theta}$.

The number of bootstrap samples, 2,200, is ten times that needed for estimating the standard error of $\hat{\theta}$ (see Efron 1987, sec. 9). But we need that many for the more delicate task of forming an approximate confidence interval for θ. Row 1 of Table 2 gives α-level approximate confidence limits with $\alpha = .025, .05, \ldots, .975$, using the bootstrap confidence interval method called BC_a by Efron (1987), explained in Section 2. The 90% central interval, for example, runs from the .05 to .95 limit, $\theta \in [379, 1{,}164]$. Notice that this interval extends more than twice as far to the right of $\hat{\theta} = 633.2$ as to the left, reflecting the asymmetry of the bootstrap histogram as well as a bias correction.

Section 2 discusses the logic of the nonparametric bootstrap method (1.6), (1.7). More elaborate bootstraps are available, as will be discussed, but the simple method has much to recommend it. It is nonparametric, is applicable to any kind of imputation procedure, and requires no knowledge of the missing-data mechanism. Its main practical disadvantage is the computational expense of the 2,000 or so bootstrap replications required for reasonable numerical accuracy.

Row 2 of Table 2 is based on an analytic approximation to the BC_a confidence limits that requires no Monte Carlo replications. This method, called ABC by DiCiccio and Efron (1992), is discussed in Section 6. The computational burden is only a few percentage points of that for BC_a, a savings that can be crucial when using imputation methods more elaborate than (1.3).

Imputation method (1.2), (1.3) is suspect here because it imputes only best-fit values and ignores residual variability. It seems likely that $\hat{\mathbf{x}}$ will have a smaller empirical covariance matrix than the original unobservable data set $\mathbf{x}$. We can avoid this kind of bias by assuming a parametric model and estimating θ by maximum likelihood. A multivariate normal model applied to the observed data $\mathbf{o}$ in Table 1 gives maximum likelihood estimate (MLE) $\hat{\theta} = 631.3$—not much different than the previous estimate 633.2. The MLE was obtained using a variant of Dempster, Laird, and Rubin's (1977) EM algorithm.

The simple nonparametric bootstrap analysis described above can be applied just as well to the normal-theory MLE as to any other estimate. The computational economy of the ABC method is essential here because of difficulties in calculating $\hat{\theta}$. Row 4 of Table 2 shows the ABC limits for θ based on the MLE, these being somewhat wider than the limits based on (1.2)–(1.5). The delta method estimate of standard error, discussed in Section 5, was 253.9, about 15% bigger than the corresponding standard error for $\hat{\theta}$ given by (1.2)–(1.5). This is a price we pay for reducing bias. Shorter

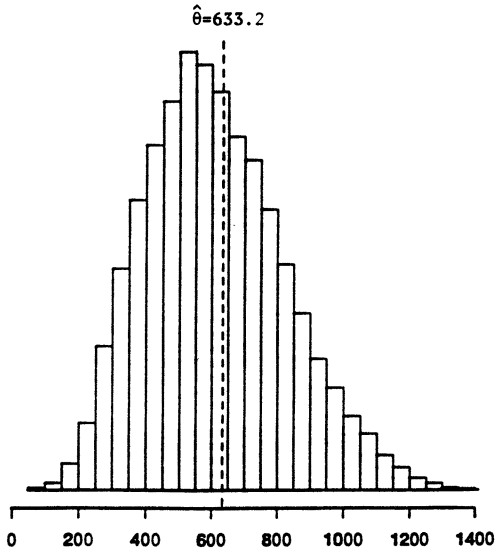

Figure 1. Histogram of 2,200 Nonparametric Bootstrap Replications of the Maximum Eigenvalue $\hat{\theta}$, Using Estimates Based on the Two-Way Fitting Algorithm (1.2), (1.3). This histogram has mean 610.3 and standard deviation $\hat{\sigma} = 212.0$, the bootstrap estimate of standard error. The first row of Table 2 gives approximate confidence limits for θ based on this histogram.

Table 2. Approximate Nonparametric Confidence Limits for the Maximum Eigenvalue θ, (1.1), Given the Observed Data **o** *in Table 1*

Confidence limit α:	.025	.050	.100	.160	.840	.900	.950	.975
1. BC_a	341	379	429	478	966	1,059	1,164	1,253
2. ABC	340	379	430	476	946	1,046	1,172	1,295
3. Full-mechanism BC_a:	349	387	439	490	970	1,074	1,213	1,300
4. ABC for MLE:	289	353	409	458	1,014	1,135	1,307	1,474
5. Multiple Imputation:	345	382	428	468	864	946	1,063	1,177

NOTE: Row 1: Nonparametric BC_a method based on the 2200 bootstrap replications of Figure 1. Row 2: An analytic approximation to row 1 called ABC, requiring no Monte Carlo replications. Row 3: A more elaborate bootstrap confidence method discussed in Section 2. Row 4: ABC limits for θ based on the normal theory MLE $\hat{\theta}$ instead of the best-fit imputation (1.3). Row 5: Multiple imputation, or data augmentation, limits. The five methods are explained in Sections 2, 3, and 5.

confidence intervals are not necessarily better confidence intervals, but in this case the normal theory MLE seems a little unrobust; see the end of Section 5.

Based on ideas suggested by the EM algorithm, Rubin proposed a theory of *multiple imputation* for assessing the variability of an estimator $\hat{\theta}$ obtained in a missing data situation. Some good references are Rubin (1987), Rubin and Schenker (1986), Tanner (1991), and others listed in Section 4. Tanner and Wong (1987) gave a neat computational description of the ideas involved, using the term *data augmentation.* The multiple imputation or data augmentation approach, described in Section 4, is quite different from the bootstrap approach of Table 2. It is based on Bayesian rather than frequentist theory. Nevertheless, Section 5 shows that bootstrap methods can be useful for implementing data augmentation. Row 5 of Table 2 refers to approximate confidence limits based on a data augmentation scheme. Section 6 briefly summarizes the advantages and disadvantages of the different approaches.

This article concerns nonparametric error estimates for missing data problems. There is an important practical difference between the parametric and nonparametric situations. It is natural in a parametric framework to estimate θ by its MLE, $\hat{\theta}$. We know that $\hat{\theta}$ has asymptotically optimal properties for estimating θ, in terms of bias, variance, or more general confidence statements. The only problem is to make such statements on the basis of a partially missing data set.

The choice of an estimator $\hat{\theta}$ is usually not so clear in a nonparametric setting. In our maximum eigenvalue problem, we might have considered three different estimators: (1) the best-fit estimator (1.2)–(1.5); (2) the Buck estimator (Buck 1960; Little 1983), in which $\hat{x}_{ij}$ in (1.3) is replaced by a linear regression predictor based on the complete cases and the elements of $\hat{\Sigma}$ in (1.4) are augmented by the addition of residual covariances (to add variability back into the imputed values $\hat{x}_{ij}$); and (3) the normal-theory MLE.

The best-fit estimator could easily be biased downward. The normal-theory estimate, which is nearly optimal in a normal sampling framework, seems to be overly variable for this data set. The Buck estimator is appealing here, being of intermediate complexity between the best-fit and normal-theory estimators. We could test the appeal by a nonparametric bootstrap analysis like that in Figure 1. The analysis might show that the Buck estimator is no more variable than the best-fit estimator, in which case it would be preferable in terms of smaller bias.

The point is this: In nonparametric situations it is useful to assess the statistical properties of a variety of estimators. The nonparametric bootstrap has the advantage of applying in a simple way to any estimator $\hat{\theta}$. Of course, this does not obviate the need to sensibly select $\hat{\theta}$. It does mean that the statistician can choose $\hat{\theta}$ using the usual variety of exploratory tools, including experience and intuition, with the assurance of being able to obtain reasonable estimates of $\hat{\theta}$'s variability.

The nonparametric and parametric situations converge when we have categorical data. In this case there is a nonparametric MLE $\hat{\theta}$, which is asymptotically optimal in terms of bias, variance, and so on. Now it is possible to directly compare the nonparametric bootstrap with multiple imputation. This comparison is made in Section 3.

2. NONPARAMETRIC BOOTSTRAP METHODS

This section discusses the logic of the simple nonparametric bootstrap method that produced Figure 1. For comparison, we also describe a different nonparametric bootstrap for missing data problems. The different structure of the two bootstraps manifests itself in what we need to assume about the missing data mechanism. Near the end of the section we briefly describe the BC_a system of approximate confidence intervals used in Table 2.

Figure 2 diagrams a missing data problem and its nonparametric bootstrap analysis. F is a population of units X_j,

$$F = \{X_j, j = 1, \ldots, N\}, \tag{2.1}$$

with N possibly infinite. A missing data mechanism, say $O_j = c(X_j)$, results in a population G of partially concealed objects

$$G = \{O_j = c(X_j), j = 1, 2, \ldots, N\}. \tag{2.2}$$

In the example of Section 1, X_j is a vector of five scores for one student and O_j is the same vector with some or perhaps none of the numerical values concealed by question marks.

We wish to infer the value of a parameter of the population F,

$$\theta_F = s(F). \tag{2.3}$$

In our example $s(F)$ is the maximum eigenvalue of the covariance matrix corresponding to F. A random sample of size n is obtained from G, $\mathbf{o} = (o_1, o_2, \ldots, o_n)$, as on the left side of Table 1. The nonparametric inference step estimates G by the empirical distribution $\hat{G}$ of $\mathbf{o}$,

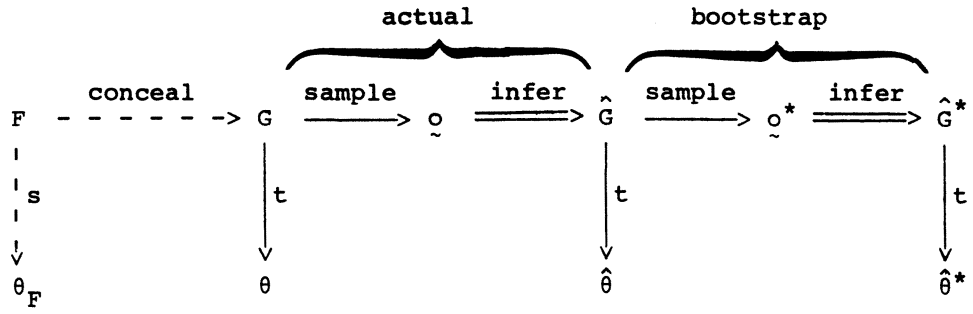

*Figure 2. Diagram of Nonparametric Bootstrap Applied to a Missing Data Problem. The individual members of a population of interest F are partially concealed according to a missing-data mechanism, giving a population G of partially concealed objects; **o** is a random sample of size n from G; $\hat{G}$ is the empirical distribution corresponding to **o**; and statistic $\hat{\theta} = t(\hat{G})$ estimates parameter $\theta = t(G)$, which is intended to be a good approximation to the actual parameter of interest $\theta_F = s(F)$. The bootstrap sampling and inference procedures duplicate those actually used, giving bootstrap replications $\hat{\theta}^* = t(\hat{G}^*)$. The BC_a and ABC methods, described later, give good approximate confidence intervals for θ based on the bootstrap replications.*

$$\hat{G}\text{: probability } 1/n \text{ on } o_i \quad \text{for} \quad i = 1, 2, \ldots, n. \tag{2.4}$$

We use some function of $\hat{G}$, say

$$\hat{\theta} = t(\hat{G}), \tag{2.5}$$

as an estimate of θ_F; for example, the bootstrap imputation estimate (1.2)–(1.5).

The nonparametric bootstrap procedure repeats the actual sampling, inference, and estimation steps, but beginning with $\hat{G}$ instead of G. The advantage of course is that $\hat{G}$, unlike G, is known, so that we can carry out an unlimited number of bootstrap replications. Each replication involves drawing a bootstrap sample $\mathbf{o}^*$ from $\hat{G}$, as at (1.7), forming the empirical distribution $\hat{G}^*$ corresponding to the $\mathbf{o}^*$ and calculating $\hat{\theta}^* = t(\hat{G}^*)$.

A principal advantage of the nonparametric bootstrap method is that it does not depend on the missing-data mechanism. The process of concealment does not affect the method, which conceptually begins with G rather than F. This is also a potential serious disadvantage. Nothing in Figure 2 connects the parameter of interest, $\theta_F = s(F)$, to the parameter being estimated, $\theta = t(G)$.

Some choices of $t(\hat{G})$ are better than others for reducing the possible bias of $\hat{\theta}$ as an estimate of θ_F. Row 4 of Table 4 refers to the normal-theory MLE estimator, for which $t(\hat{G})$ can be described as follows: assume that each student's score vector x_i is a random draw from a five-dimensional normal distribution

$$x_i \sim N_5(\mu, \Sigma), \tag{2.6}$$

some components of which have been concealed by the missing-data mechanism to give the observed vector o_i, $i = 1, 2, \ldots, n$. We estimate μ and Σ by normal-theory maximum likelihood and then take $\hat{\theta}$ to be the maximum eigenvalue of $\hat{\Sigma}$.

Under some circumstances this choice of $t(\cdot)$ will be *Fisher consistent* for estimating the maximum eigenvalue, in the sense that

$$\theta = t(G) = s(F) = \theta_F. \tag{2.7}$$

The circumstances are that F is multivariate normal and that the missing data mechanism does not affect the likelihood function. In other words, we assume that each o_i has the appropriate marginal normal distribution obtained from (2.6). Rubin (1987) called this latter assumption *ignorable nonresponse.*

Fisher consistency says that in large samples our estimate will tend toward the correct answer. This is an obviously desirable property, but it may be costly to insist upon it. In the example of Table 2 it seems to cost an extra 15%. This situation resembles the robust estimation of a population mean, where there is a trade-off between the bias and variance of possible estimators. Even if we completely trust the normality assumption (2.6), ignorable nonresponse is unverifiable, and is questionable in many realistic circumstances. It fails if the missing data depends on the missing values; for example, if the question marks in Table 1 occur more frequently at extreme values of A and E.

The estimation problem becomes simpler if we are dealing with categorical data points x_i, rather than with the five-dimensional vector-valued points of the eigenvalue example. In the categorical case it is easy to write down the nonparametric maximum likelihood estimator $\hat{\theta} = t(\hat{G})$. Fisher consistency (2.7) will now hold true for all distributions F, as long as we have ignorable nonresponse. The objection that the nonparametric bootstrap is estimating the wrong parameter is now totally removed, though possibly still at the expense of excess variability for $\hat{\theta}$ in reasonable-sized samples. The categorical case is discussed in Section 3, where it is used as a basis of comparison between the nonparametric bootstrap and multiple imputation.

Other bootstrap methods can be applied to missing-data problems. The *full-mechanism bootstrap* diagramed in Figure 3 begins more directly than the nonparametric bootstrap of Figure 2. The original, unobservable, data set $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is assumed to be obtained by random sampling from F. The missing-data mechanism is applied to the components of $\mathbf{x}$, $o_i = c(x_i)$, giving the observed data $\mathbf{o}$. Some method of inference, necessarily more complicated than (2.4), gives an estimate $\hat{F}$ for F based on $\mathbf{o}$. The parameter of interest, $\theta = s(F)$, is then estimated by $\hat{\theta} = s(\hat{F})$. The

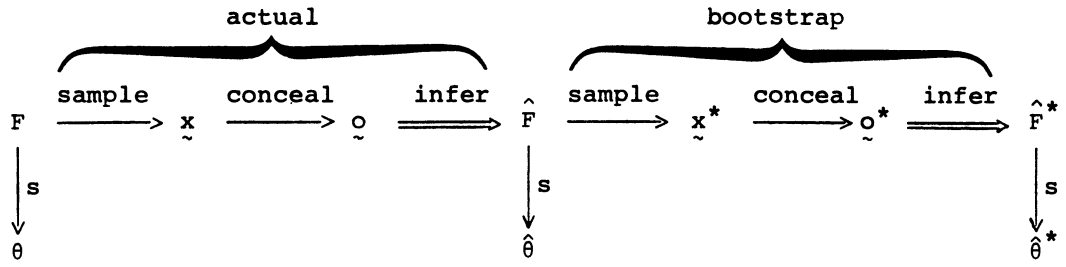

Figure 3. Full-Mechanism Bootstrap. $\mathbf{x}$ *is a random sample from population of interest* F*; members of* $\mathbf{x}$ *are partially concealed by the missing-data mechanism to give* $\mathbf{o}$*;* $\hat{F}$ *is an estimate of* F *based on* $\mathbf{o}$*; and the parameter of interest,* $\theta = s(F)$*, is estimated by* $\hat{\theta} = s(\hat{F})$*. The bootstrap sampling, missing-data, and inference procedures are supposed to duplicate those that actually occurred. This requires specification of the missing-data mechanism.*

full-mechanism bootstrap repeats the sampling, missing-data, and inference processes to yield bootstrap replications $\hat{\theta}^* = s(\hat{F}^*)$.

A total of 2,500 full-mechanism bootstrap replications $\hat{\theta}^*$ were obtained for the maximum eigenvalue problem of the Introduction. The inference method was taken to be as simple as possible: $\hat{F}$ equaled the empirical distribution of the best-fit imputation $\hat{\mathbf{x}}$ based on (1.2), (1.3), putting probability $\frac{1}{22}$ on each vector $\hat{x}_i$. (A more ambitious scheme might have obtained $\hat{F}$ by multiple imputations from $\mathbf{o}$, as discussed in Sec. 4.) The bootstrap samples $\mathbf{x}^* = (x_1^*, \ldots, x_n^*)$ were drawn by simple random sampling from $\hat{F}$ (i.e., from $\{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n\}$), and then $\mathbf{o}^*$ was obtained by concealing the same elements of $\hat{x}^*$ as those concealed by question marks on the left side of Table 1. For example, if x_1 were selected to be $\hat{x}_{22}$, then $o_1^* = (?, 26, 15, 20, 14.66)$. Best-fit imputation (1.2), (1.3) applied to $\mathbf{o}^*$ gave $\hat{\mathbf{x}}^*$, then $\hat{F}^*$, and finally $\hat{\theta}^*$.

The histogram of the 2,500 $\hat{\theta}^*$ values looked much the same as Figure 1, giving bootstrap standard error estimate $\hat{\sigma} = 219.8$ and bias estimate -23.2, compared with 212.0 and -22.9 previously. The BC_a confidence limits were also much the same as before, as seen in Row 3 of Table 2.

The full-mechanism bootstrap uses the same function $s(\cdot)$ to define and estimate the parameter of interest, avoiding the possibility of definitional bias that we worried about earlier. However a similar difficulty arises with the more complicated inference required in going from $\mathbf{o}$ to $\hat{F}$ rather than from $\mathbf{o}$ to $\hat{G}$.

There is a serious disadvantage: In Figure 3 we need to specify the missing-data mechanism $o_i^* = c(x_i^*)$ in order to obtain the bootstrap replications $\hat{\theta}^*$. In other words, we need to say what we would have observed in situations other than the one that actually occurred. Simply concealing the same elements as in $\mathbf{o}$ is allowable if the missing-data mechanism is what Little and Rubin (1987) called *missing completely at random,* but this is usually an unrealistic assumption. It is a much stronger assumption than ignorability.

On the other hand, if the missing data mechanism is *nonignorable,* then we may need to model it anyway, in which case the full-mechanism bootstrap becomes practical. Suppose that we believed that the observed first test scores in Table 1, the o_{1i}, were missing in a way that depended on the actual score x_{1i}, say

$$\mathrm{prob}\{o_{1i} = ? \mid x_i\} = \frac{1}{1 + e^{-(\lambda_0 + \lambda_1 x_{1i})}}. \tag{2.8}$$

Past experience or a preliminary data analysis might provide rough estimates of λ_0 and λ_1. Then we could use (2.8) to compute $\mathbf{o}^*$ from $\mathbf{x}^*$ in Figure 3, and so carry through the full-mechanism bootstrap.

Bootstrap theory aims to provide statistical inferences from a bootstrap histogram like Figure 1. The BC_a method used in Table 2 is a way of forming highly accurate approximate confidence intervals for the parameter

$$\theta = t(G)$$

from $\hat{\theta} = t(\hat{G})$ and bootstrap replications $\hat{\theta}^*$. It applies to both parametric and nonparametric situations (see Efron 1987). Typically the method is *second-order accurate;* if $\hat{\theta}(\alpha)$ is the endpoint of a one-sided BC_a interval of intended level α, then

$$\mathrm{prob}\{\theta < \hat{\theta}(\alpha)\} = \alpha + O_p(1/n) \tag{2.9}$$

as the sample size $n \to \infty$ (see Remark D). This compares with $\alpha + O_p(1/\sqrt{n})$ if we use the standard interval $\hat{\theta} + z^{(\alpha)}\hat{\sigma}$, where $\hat{\sigma}$ is an estimated standard error for $\hat{\theta}$ and $z^{(\alpha)}$ is the 100αth percentile of a $N(0, 1)$ distribution,

$$z^{(\alpha)} = \Phi^{-1}(\alpha), \tag{2.10}$$

Φ being the standard normal cdf. Remark B extends the BC_a method to the k-sample problem.

The BC_a interval endpoints are computed from the percentiles of the bootstrap histogram. Let $\hat{H}$ be the empirical cdf of the bootstrap replications, say B of them ($B = 2{,}200$ in Fig. 1):

$$\hat{H}(t) = \#\{\hat{\theta}^* < t\}/B. \tag{2.11}$$

Then $\hat{H}^{-1}(\alpha)$ is the 100αth bootstrap percentile. The α-level BC_a endpoint $\hat{\theta}(\alpha)$ is defined by

$$\hat{\theta}(\alpha) = \hat{H}^{-1}\Phi\left(z_0 + \frac{z_0 + z^{(\alpha)}}{1 - a(z_0 + z^{(\alpha)})}\right), \tag{2.12}$$

where z_0 is the *bias-correction constant*

$$z_0 = \Phi^{-1}\hat{H}(\hat{\theta}). \tag{2.13}$$

The computation of the *acceleration constant* a in (2.12) is described in Section 5 at (5.5).

If z_0 and a equal 0 and $\hat{H}$ is the normal distribution $N(\hat{\theta}, \hat{\sigma}^2)$, then $\hat{\theta}(\alpha)$ equals the standard confidence limit $\hat{\theta} + z^{(\alpha)}\hat{\sigma}$. Otherwise, formula (2.12) corrects all of the second-

order errors made by the standard intervals (see Efron 1987, sec. 2).

Formula (2.12) looks complicated, but it is easy to apply. The hard part is computing the 2,000 bootstrap replications necessary to give the formula sufficient accuracy. The ABC algorithm of Section 5 uses analytic approximations in place of Monte Carlo replications. Typically it requires only a few percent as much computational effort as BC_a.

Remark A. The full-mechanism and nonparametric bootstrap methods are identical in the important special case where we observe censored data from a survival analysis (see Efron 1981a).

Remark B. The basic idea of the nonparametric bootstrap of Figure 2 is to consider the o_i, including the missing values, as randomly sampled points from a population G. This idea has been used before in the sample survey literature, though typically with methods like balanced repeated replications or the jackknife, rather than the bootstrap. Fay (1986, sec. 4) used the jackknife on a complicated categorical data problem with missing data. Greenwood's formula for the variance of a survival curve is a delta method version of the same idea. Meng and Rubin (1989) suggested the possibility of a bootstrap approach to missing-data problems. Laird and Lewis' (1987) "Type I" and "Type II–III" bootstraps are examples, respectively, of the nonparametric and full-mechanism methods.

Remark C. The nonparametric bootstrap of Figure 2 can also be applied to K-sample problems

$$(G_1, G_2, \ldots, G_K) \to (\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_K), \qquad (2.14)$$

where the $\mathbf{o}_k$ are independent random samples of size n_k obtained from populations G_k, the parameter of interest being $\theta = t(G_1, G_2, \ldots, G_K)$. The empirical distributions $\hat{G}_k$ corresponding to the $\mathbf{o}_k$ give independent bootstrap samples $\mathbf{o}_k^*$ of size n_k, from which we calculate $\hat{G}^*$ and finally $\hat{\theta}^* = t(\hat{G}_1^*, \hat{G}_2^*, \ldots, \hat{G}_K^*)$. The BC_a intervals are calculated from (2.12), (2.13) as before, with the acceleration a obtained as in Remark J of Section 5.

Remark D. DiCiccio and Efron (1992) showed that the BC_a and ABC intervals are second-order accurate if the data set is obtained by random sampling from a multiparameter exponential family and θ is a smooth function of the expectation vector of the family. A discretization argument applies this result to the situation in Figure 2. We suppose that the sample space of the o_i can be discretized to a finite number of outcomes, say L of them. For the examples in Table 1, each of the five coordinates of an o_i vector takes its value in the 102-element set $\{?, 0, 1, 2, \ldots, 100\}$, so we can take $L = 102^5$. The multiparameter exponential family referred to above is the L-category family of multinomial distributions. See the comments on finite sample spaces of Efron (1987, sec. 8).

3. MULTIPLE IMPUTATION

Best-fit imputation, as illustrated on the right side of Table 1, conveys a false sense of accuracy if the imputed values are interpreted as ordinary observations. Rubin (1987, 1978) proposed drawing multiple random imputations of the missing data rather than a single best-fit imputation. Variability of results between the randomly imputed data sets can then be used to assess the true accuracy of an estimate $\hat{\theta}$. The variability calculation is carried out by means of a Bayesian updating scheme, quite different in concept from the bootstrap method of Section 2. This section briefly reviews multiple imputation, following the development of Tanner and Wong (1987). Section 4 presents a third bootstrap method for missing data, based on multiple imputation. At first we will use parametric notation, which is more natural in the Bayesian framework. Then we will discuss categorical data, for which it is easier to make a nonparametric comparison of multiple imputation with the bootstrap.

3.1 Data Augmentation

Let $\mathbf{o}$ indicate the observed data set and let $\mathbf{x}$ indicate any complete data set consonant with $\mathbf{o}$. In the example of Table 1, $\mathbf{x}$ could be any 22×5 matrix of numbers agreeing with $\mathbf{o}$ at all of its numerical entries. The actual complete data set $\mathbf{x}$ giving rise to $\mathbf{o}$, which would have been observed if there were no missing data, is assumed to be sampled from a parametric family with density function $f_\eta(\mathbf{x})$, with η being an unknown p-dimensional parameter vector. Starting with a prior density $\xi_0(\eta)$ on η, Bayes's theorem would give hypothetical posterior density $\xi(\eta|\mathbf{x})$ if $\mathbf{x}$ were observed and, more concretely, the actual posterior density $\xi(\eta|\mathbf{o})$ having observed $\mathbf{o}$. A standard probability calculation relates $\xi(\eta|\mathbf{o})$ to $\xi(\eta|\mathbf{x})$:

$$\xi(\eta|\mathbf{o}) = \int_{\mathbf{x}} \xi(\eta|\mathbf{x}) f(\mathbf{x}|\mathbf{o})\, d\mathbf{x}, \qquad (3.1)$$

where $f(\mathbf{x}|\mathbf{o})$ is the *predictive density* of $\mathbf{x}$ given $\mathbf{o}$, the conditional density integrating out η,

$$f(\mathbf{x}|\mathbf{o}) = \int_{\eta} f_\eta(\mathbf{x}|\mathbf{o}) \xi(\eta|\mathbf{o})\, d\eta. \qquad (3.2)$$

The integral in (3.1) is taken over all $\mathbf{x}$ consonant with $\mathbf{o}$.

Result (3.1), the *data augmentation identity,* can be stated as follows. The posterior density of η, given the observed data $\mathbf{o}$, is the average posterior density of η based on a complete data set $\mathbf{x}$. The average is taken over the predictive density of $\mathbf{x}$ given $\mathbf{o}$. In a typical missing-data problem, computing $\xi(\eta|\mathbf{x})$ is easy but computing $\xi(\eta|\mathbf{o})$ is difficult. If we can sample from the predictive density $f(\mathbf{x}|\mathbf{o})$, then (4.1) gives a practical way of approximating $\xi(\eta|\mathbf{o})$:

$$\hat{\xi}(\eta|\mathbf{o}) = \frac{1}{M} \sum_{m=1}^{M} \xi(\eta|\mathbf{x}^{(m)}), \qquad (3.3)$$

where $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(M)}$ are the multiple imputations, that is, independent draws from $f(\mathbf{x}|\mathbf{o})$. This argument has a circular look, because we need to know $\xi(\eta|\mathbf{o})$ to calculate $f(\mathbf{x}|\mathbf{o})$ in (3.2). Tanner and Wong (1987) investigated an iterative algorithm related to Gibbs' sampling for actually carrying out (3.3). Noniterative approximations are available, as discussed later.

Most often, inferences are desired for some real-valued function (or functions) of η,

$$\theta = s(\eta), \tag{3.4}$$

like the maximum eigenvalue in Section 1 rather than for the entire vector η. The marginal posterior densities of θ, say $\pi(\theta|\mathbf{o})$ and $\pi(\theta|\mathbf{x})$, are related by a marginalized version of (3.1),

$$\pi(\theta|\mathbf{o}) = \int_{\mathbf{x}} \pi(\theta|\mathbf{x}) f(\mathbf{x}|\mathbf{o})\, d\mathbf{x}, \tag{3.5}$$

with $f(\mathbf{x}|\mathbf{o})$ still being defined by (3.2).

The most obvious difficulty in applying (3.3)–(3.5) is the generation of imputations $\mathbf{x}^{(m)}$ from the predictive density $f(\mathbf{x}|\mathbf{o})$. A simple approach, called "poor man's data augmentation" by Wei and Tanner (1990), is to sample the $\mathbf{x}^{(m)}$ from $f_\eta(\mathbf{x}|\mathbf{o})$, with η set equal to the MLE $\hat{\eta}$:

$$\mathbf{x}^{(m)} \sim f_{\hat{\eta}}(\mathbf{x}|\mathbf{o}),$$

$$\text{independently for} \quad m = 1, 2, \ldots, M. \tag{3.6}$$

This could also be called a conditional parametric bootstrap sample. In many situations (3.6) is quite satisfactory, though it can underestimate variability if there is too much missing data.

A better approximation to the predictive density is often used. Each imputation $\mathbf{x}^{(m)}$ is drawn from its own bootstrapped choice of the parameter vector η:

$$\mathbf{x}^{(m)} \sim f_{\hat{\eta}^{*(m)}}(\mathbf{x}|\mathbf{o}),$$

$$\text{independently for} \quad m = 1, 2, \ldots, M. \tag{3.7}$$

A bootstrap parameter vector $\hat{\eta}^*$ is the MLE for η based on a bootstrap sample $\mathbf{o}^*$, (1.7). Rubin (1981) and Efron (1982, sec. 10.6) pointed out that the bootstrap distribution of $\hat{\eta}^*$ is quite close to the nonparametric Bayes posterior distribution of η given $\mathbf{o}$, starting from a vague Dirichlet prior for G, (2.2). The difference between (3.6) and (3.7) is the difference between a first-level and a second-level bootstrap sample. A nice application of (3.7) was presented by Heitjan and Little (1991). Little and Rubin (1987, sec. 12.4) call (3.7) a *proper* imputation method, as opposed to the *improper* method (3.6), which underestimates variability. They used the name *approximate Bayesian bootstrap* for procedures like (3.7).

The data augmentation identity requires the correct specification of $f_\eta(\mathbf{x}|\mathbf{o})$ in (3.2). In practice this means making some assumption about the missing-data mechanism, such as ignorable nonresponse, as discussed next.

3.2 Categorical Data

To compare multiple imputation with the nonparametric bootstrap, we need to have a nonparametric interpretation of (3.1)–(3.7). This is easy to do in the case of categorical data, where the original population units X_j in (2.1) have only a finite number of possible values, say $X_j \in \mathcal{X} = \{X(1), X(2), \ldots, X(L)\}$. Similarly, the partially concealed objects $O_j = c(X_j)$ produced by the missing-data mechanism are limited to a finite set $\mathcal{O} = \{O(1), O(2), \ldots, O(K)\}$. Each $O(k)$ is a subset of $\mathcal{X}$, so that observing $O_j = o_j$ means that the unobserved $X_j = x_j$ lies in the subset o_j. In what follows we will let $\delta(x, o)$ indicate whether or not x is among the values contained in o:

$$\begin{aligned} \delta(x, o) &= 1 \quad \text{if } x \in o \quad (x \in \mathcal{X}, o \in \mathcal{O}). \\ &= 0 \quad \text{if } x \notin o \end{aligned} \tag{3.8}$$

As a simple example, suppose that a human population is categorized by sex and handedness. There are $L = 4$ original population values:

$$\begin{aligned} &X(1) = (\text{male, left}), \qquad X(2) = (\text{male, right}), \\ &X(3) = (\text{female, left}), \quad \text{and} \\ &X(4) = (\text{female, right}). \end{aligned} \tag{3.9}$$

If there are difficulties in ascertaining handedness, we will need $K = 6$ states in $\mathcal{O}$: $O(k) = X(k)$ for $k = 1, 2, 3, 4$ and the two additional states

$$O(5) = (\text{male, ?}) \quad \text{and} \quad O(6) = (\text{female, ?}). \tag{3.10}$$

In this case $O(6)$ corresponds to $\{X(3), X(4)\}$, so that observing $o_j = O(6)$ means that x_j is either (female, left) or (female, right).

The missing-data mechanism can be described by the conditional probability density of O_j given X_j, say

$$c(o|x) = \text{prob}\{O_j = o | X_j = x\}$$

$$(\text{for } x \in \mathcal{X}, o \in \mathcal{O}). \tag{3.11}$$

Because X_j is a member of O_j, only those o containing x have positive probability:

$$c(o|x) = 0 \quad \text{if } \delta(x, o) = 0. \tag{3.12}$$

The populations F and G in (2.1), (2.2) correspond to densities $\mathbf{f} = (f_1, f_2, \ldots, f_L)$ and $\mathbf{g} = (g_1, g_2, \ldots, g_K)$ on $\mathcal{X}$ and $\mathcal{O}$ respectively. For example, f_3 is the proportion of units in F with X_j equalling $X(3)$. The two densities are related by $g_o = \sum_x c(o|x) f_x$, so $\mathbf{f}$ determines $\mathbf{g}$. Letting $\mathbf{C}$ be the $L \times K$ matrix with entries $c(o|x)$,

$$\mathbf{g} = \mathbf{f}\mathbf{C}. \tag{3.13}$$

In our nonparametric setting $\mathbf{f}$ plays the role of the parameter vector η in (3.1)–(3.7).

Bayes's rule gives $d_{\mathbf{f}}(x|o)$, the conditional probability density of x given o:

$$d_{\mathbf{f}}(x|o) = f_x c(o|x)/g_o. \tag{3.14}$$

The $L \times K$ matrix $\mathbf{D}_{\mathbf{f}} = (d_{\mathbf{f}}(x|o))$ inverts relationship (3.13),

$$\mathbf{f} = \mathbf{g}\mathbf{D}'_{\mathbf{f}}. \tag{3.15}$$

Because $\mathbf{D}_{\mathbf{f}}$ depends on $\mathbf{f}$, solving for $\mathbf{f}$ in (3.15) is not the same as using the linear inversion $\mathbf{f} = \mathbf{g}\mathbf{C}'(\mathbf{C}\mathbf{C}')^{-1}$. (We are assuming that $\mathbf{C}$ does not depend on $\mathbf{f}$. If it did, we would be in the more difficult situation where the missing-data mechanism itself provides information about $\mathbf{f}$.)

A random sample $\mathbf{o} = (o_1, o_2, \ldots, o_n)$ from G gives empirical probabilities $\hat{\mathbf{g}} = (\hat{g}_1, \hat{g}_2, \ldots, \hat{g}_K)$,

$$\hat{g}_k = \#\{o_i = O(k)\}/n. \tag{3.16}$$

The MLE $\hat{\mathbf{f}}$ of $\mathbf{f}$ is obtained from an empirical version of (3.15),

$$\hat{\mathbf{f}} = \hat{\mathbf{g}}\mathbf{D}'_{\hat{\mathbf{f}}}. \tag{3.17}$$

This is the self-consistency property of the MLE as explained by Dempster et al. (1977).

The trouble with using (3.17) to find the MLE $\hat{\mathbf{f}}$ is that $\mathbf{D}_{\mathbf{f}}$ depends on the missing-data density $c(o|x)$, which is usually unknown. This is where ignorability comes in. Let $p_{\mathbf{f}}(x|o)$ indicate the "obvious" conditional density of x given o:

$$p_{\mathbf{f}}(x|o) = \delta(x, o)f_x \Big/ \sum_{x'} \delta(x', o)f_{x'}; \tag{3.18}$$

$p_{\mathbf{f}}(x|o)$ puts probabilities proportional to f_x on each x in o. Suppose that the selection mechanism of o_i given x_i was *predetermined* in the following sense: First a disjoint partition $\mathcal{P}_i$ of $\mathcal{X}$ was defined, then x_i was selected according to density $\mathbf{f}$, and finally o_i was chosen to be that member of $\mathcal{P}_i$ into which x_i fell. In this case $p_{\mathbf{f}}(x_i|o_i)$ would equal the actual conditional density $d_{\mathbf{f}}(x_i|o_i)$.

Ignorability implies that we can ignore the missing-data mechanism and set $d_{\mathbf{f}}(x|o)$ equal to $p_{\mathbf{f}}(x|o)$ (see Rubin 1986, sec. 7). Letting $\mathbf{P}_{\mathbf{f}}$ be the $L \times K$ matrix $(p_{\mathbf{f}}(x|o))$, (3.18) becomes $\mathbf{f} = \mathbf{g}\mathbf{P}'_{\mathbf{f}}$, and (3.15) takes on the more practical form

$$\hat{\mathbf{f}} = \hat{\mathbf{g}}\mathbf{P}'_{\hat{\mathbf{f}}}. \tag{3.19}$$

We will write $\hat{\mathbf{f}} = \mathrm{MLE}(\hat{\mathbf{g}})$ to indicate the mapping from $\hat{\mathbf{g}}$ to $\hat{\mathbf{f}}$ implied by (3.19), ignoring questions of uniqueness and existence.

The nonparametric MLE for a parameter $\theta = s(\mathbf{f})$ is

$$\hat{\theta} = t(\hat{\mathbf{g}}) \equiv s(\mathrm{MLE}(\hat{\mathbf{g}})). \tag{3.20}$$

If (3.20) is the function t used in Figure 2, then we know we are estimating the correct parameter, because $\theta_F = s(\mathbf{f}) = t(\mathbf{g}) = \theta$ as in (2.7). Also, $\hat{\theta}$ is asymptotically efficient. This choice of t puts the nonparametric bootstrap on the same footing as multiple imputation.

Figure 4 is a comparison of the nonparametric bootstrap with multiple imputation in the nonparametric categorical data problem. The top line shows the computational steps involved in the nonparametric bootstrap. The resampling step $\mathbf{o} \to \mathbf{o}^*$ is the simple bootstrap defined at (1.7). Here $\mathbf{o}$ and $\mathbf{o}^*$ could just as well be labeled $\hat{G}$ and $\hat{G}^*$ as in Figure 2, or equivalently, $\hat{\mathbf{g}}$ and $\hat{\mathbf{g}}^*$. The bootstrap replication $\hat{\theta}^* = t(\hat{G}^*)$ in Figure 2 is given here by $\hat{\theta}^* = s(\mathrm{MLE}(\hat{\mathbf{g}}^*))$ as in (3.20).

The first three steps of the multiple imputation algorithm implement the approximate Bayesian bootstrap (3.7). Conditional resampling is done according to the conditional densities (3.8),

$$x_i^{**} | o_i \sim p_{\hat{\mathbf{f}}^*}(\cdot | o_i) \qquad (i = 1, 2, \ldots, n). \tag{3.21}$$

Having selected $\hat{\mathbf{f}}^*$, the sampling in (3.21) is conditionally independent for $i = 1, 2, \ldots, n$. The completed data set $\mathbf{x}^{**} = (x_1^{**}, x_2^{**}, \ldots, x_n^{**})$ is what we called $\mathbf{x}^{(m)}$ in (3.7). The double star notation emphasizes the two levels of bootstrap sampling involved.

A basic assumption of the multiple imputation approach is that inferences for θ would be easy to make if there were no missing data. We assume in Figure 4 that given $\mathbf{x}^{**}$, a completed data set, it is easy to calculate a posterior density $\xi(\mathbf{f}|\mathbf{x}^{**})$ and marginalize ξ to the appropriate posterior density $\pi(\theta|\mathbf{x}^{**})$ for θ. In contrast, the nonparametric bootstrap uses the replications $\hat{\theta}^*$ to directly make inferences about θ. The crucial marginalization step, from $\mathbf{f}$ to θ, is handled automatically by the bootstrap confidence algorithm BC_a or ABC.

Marginalization is a major difficulty in applying Bayesian methods to high-dimensional problems, even without missing data. In genuine Bayesian situations there can be no argument with inferences based on the posterior density $\pi(\theta|\mathbf{o})$. However multiple imputation is often applied in an objectivist framework, beginning, perhaps implicitly, with some form of uninformative prior. This can be tricky ground, where an apparently innocuous prior on the full parameter vector leads to unexpected biases for θ (see Bernardo and Berger 1991 and Tibshirani 1989). Section 4 concerns an easy and accurate marginalization technique, designed to simplify the use of multiple imputation.

Figure 4 shows that the nonparametric bootstrap is simpler and more direct than multiple imputation. On the other hand, multiple imputation can be applied to parametric problems and can incorporate Bayesian information (Lazzeroni, Schenker, and Taylor 1990). It more graphically conveys the effect of the missing data, as will be seen in Figure 5. The two methods are compared further in Section 6.

Remark F. Theoretically we could use the nonparametric MLE (3.20) to estimate $\hat{\theta}$ in the maximum eigenvalue problem, by discretizing the data in Table 1 as we did in Remark D. This is not at all practical, given only $n = 22$ points in a five-dimensional space. Ad hoc estimators like those in Section 1 arise in nonparametric problems because of the impracticality of full nonparametric maximum likelihood estimation.

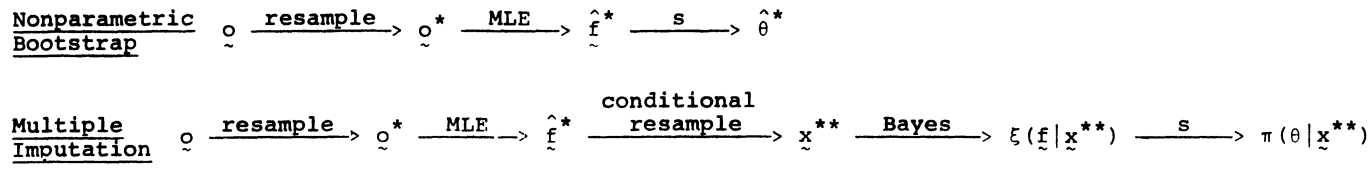

Figure 4. Comparison of Nonparametric Bootstrap With Multiple Imputation for Categorical Data. Resample indicates nonparametric bootstrap sample (1.7); MLE is the nonparametric maximum likelihood estimator for $\mathbf{f}$, (3.19); the parameter of interest is $\theta = s(\mathbf{f})$; and conditional resamples are obtained by the approximate Bayesian bootstrap method (3.21). Multiple imputation assumes that given the completed data set $\mathbf{x}^{**}$, it is easy to compute the Bayes posterior density $\xi(\mathbf{f}|\mathbf{x}^{**})$ for $\mathbf{f}$ and then marginalize ξ to the posterior density $\pi(\theta|\mathbf{x}^{**})$ for θ.

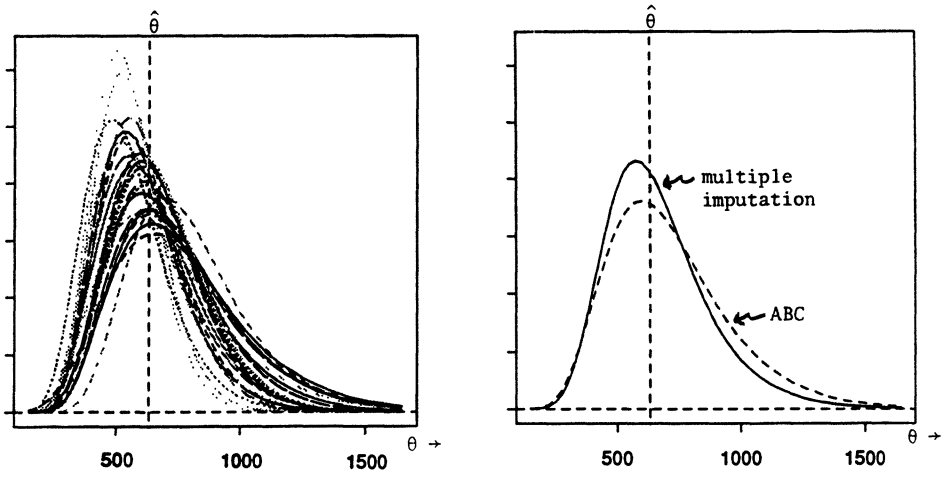

Figure 5. Multiple-Imputation Bootstrap for the Maximum Eigenvalue Problem of Section 1. The left panel shows ABC confidence densities $\pi^{\dagger}(\theta \mid \mathbf{x}^{(m)})$ for 25 imputed data sets $\mathbf{x}^{(m)}$, $m = 1, 2, \ldots, 25$. In the right panel, the solid line is $\hat{\pi}(\theta \mid \mathbf{o})$, (4.3), the average of $\pi^{\dagger}(\theta \mid \mathbf{x}^{(m)})$ for $m = 1, 2, \ldots, 50$, and the dashed line is $\pi^{\dagger}_{nonpar}(\theta \mid \mathbf{o})$, the ABC density for the nonparametric bootstrap of Section 2; $\hat{\pi}(\theta \mid \mathbf{o})$ has a shorter upper tail than $\pi^{\dagger}_{nonpar}(\theta \mid \mathbf{o})$.

4. MULTIPLE-IMPUTATION BOOTSTRAP

Suppose now that we have satisfactorily solved the problem of sampling from the predictive density $f(\mathbf{x}|\mathbf{o})$ to obtain imputations $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(M)}$ and wish to construct good approximate confidence intervals for a parameter of interest θ. The data augmentation identity (3.5) suggests using the percentiles of the estimated posterior density

$$\hat{\pi}(\theta|\mathbf{o}) = \frac{1}{M}\sum_{m=1}^{M} \pi(\theta|\mathbf{x}^{(m)}) \tag{4.1}$$

as the endpoints for such intervals. But to do so requires that the complete-data posterior density $\pi(\theta|\mathbf{x})$ enjoy good confidence properties. Choosing an appropriate *uninformative prior* for a high-dimensional vector parameter can be a difficult problem (see Berger and Bernardo 1991 and Tibshirani 1989). This section uses a bootstrap-based method developed by Efron (1992a) to avoid such choices and to simplify the calculation of (4.1).

Let $\theta_{\mathbf{x}}(\alpha)$ indicate the α-level endpoint of an exact or approximate system of confidence intervals for θ based on data $\mathbf{x}$. The *confidence density* for θ given $\mathbf{x}$ is defined to be

$$\pi^{\dagger}(\theta|\mathbf{x}) = 1 \Big/ \frac{d\theta_{\mathbf{x}}(\alpha)}{d\alpha}. \tag{4.2}$$

This density assigns probability .01 to θ lying between the .90 and .91 confidence limits, and so on. By definition, the 100αth percentile of $\pi^{\dagger}(\theta|\mathbf{x})$ is $\theta_{\mathbf{x}}(\alpha)$, so $\pi^{\dagger}(\theta|\mathbf{x})$ is just another way of describing the function $\theta_{\mathbf{x}}(\alpha)$. But the confidence density is convenient for use in (4.1), giving

$$\hat{\pi}^{\dagger}(\theta|\mathbf{o}) = \frac{1}{M}\sum_{m=1}^{M} \pi^{\dagger}(\theta|\mathbf{x}^{(m)}). \tag{4.3}$$

Confidence densities can be thought of as a way to automatically marginalize a high-dimensional posterior distribution to a single parameter of interest. If $\theta_{\mathbf{x}}(\alpha)$ represents a second-order accurate confidence interval endpoint, then $\pi^{\dagger}(\theta|\mathbf{x}^{(m)})$ will be a good approximation to the posterior density for θ, having begun with an appropriate uninformative prior for the full parameter vector (see Efron 1993). The ABC method described in Section 5 yields a very simple construction for $\pi^{\dagger}(\theta|\mathbf{x}^{(m)})$. We will call the construction of multiple-imputation confidence intervals via (4.3) and the ABC method the *multiple implementation bootstrap*. This method has an objectivist Bayesian rationale, like the Bayesian bootstrap with which it begins.

Figure 5 shows the application of the multiple-imputation bootstrap to the maximum eigenvalue problem of Section 1. $M = 50$ multiple imputations $\mathbf{x}^{(m)}$ were drawn using the approximate Bayesian bootstrap (3.7). The construction of $\pi^{\dagger}(\theta|\mathbf{x}^{(m)})$, described in Section 5, was only partially nonparametric; the family $f_{\eta}(\mathbf{x})$ used in (3.7) was the five-dimensional normal (2.6). Less parametric ways of drawing the $\mathbf{x}^{(m)}$ seemed impractical, given the small sample size and the five-dimensional sample space.

The left panel of Figure 5 shows the ABC densities $\pi^{\dagger}(\theta|\mathbf{x}^{(m)})$ for $m = 1, 2, \ldots, 25$. The solid curve in the left panel is the average of all 50 densities: $\hat{\pi}^{\dagger}(\theta|\mathbf{o})$, (4.3). Row 5 of Table 2 comprises the appropriate percentiles of $\hat{\pi}^{\dagger}(\theta|\mathbf{o})$.

The ABC method can also be used as a computationally efficient way to implement the nonparametric bootstrap of Section 2. The dashed curve in the right panel of Figure 5 is $\pi^{\dagger}_{\text{nonpar}}(\theta|\mathbf{o})$, the ABC confidence density appropriate to the nonparametric bootstrap. Row 2 of Table 2 comprises the percentiles of $\pi^{\dagger}_{\text{nonpar}}(\theta|\mathbf{o})$.

In this case the multiple imputation confidence limits are too short in the upper tail. Section 5 traces the difficulty to an overly influential student in Table 1, combined with the normal-theory imputations.

5. THE ABC ALGORITHM

The main disadvantage of the BC_a method is the large number of bootstrap replications required. This computational burden can often be avoided by using analytical ex-

pansions in place of the bootstrap Monte Carlo replications. DiCiccio and Efron (1992) developed an algorithm called ABC, standing for "approximate bootstrap confidence" intervals, that uses numerical second derivatives to accurately approximate the endpoints of the BC_a intervals. The development in that paper is mainly for parametric exponential family problems. Here the algorithm is adapted to nonparametric problems, actually simplifying the calculations. The algorithm is given as an *S* function in the Appendix.

Given a bootstrap sample $\mathbf{o}^* = (o_1, o_2, \ldots, x_n)$, (1.7), let P_i^* indicate the proportion of times o_i is represented in $\mathbf{o}^*$:

$$P_i^* = \#\{o_j^* = o_i\}/n \qquad (i = 1, 2, \ldots, n). \quad (5.1)$$

With the data $\mathbf{o}$ fixed, we can think of a bootstrap replication $\hat{\theta}^* = t(\hat{G}^*)$ in Figure 2 as a function of the resampling vector $\mathbf{P}^* = (P_1^*, P_2^*, \ldots, P_n^*)$, say $\hat{\theta}^* = T(\mathbf{P}^*)$. The resampling vector takes its value in the simplex

$$\mathcal{S}_n = \{\mathbf{P} : P_i \geq 0, \Sigma P_i = 1\}. \quad (5.2)$$

The resampled statistic $\hat{\theta}^* = T(\mathbf{P}^*)$ can be thought of as a function on the simplex, forming a *resampling surface* over $\mathcal{S}_n$, as in figure 6.1 of Efron (1982). The geometry of the resampling surface determines the bootstrap confidence intervals for θ. In the BC_a method the surface is explored by evaluating $T(\mathbf{P}^*)$ for some 2,000 random choices of $\mathbf{P}^*$.

The ABC algorithm approximates the BC_a interval endpoints by exploring the local geometry of the resampling surface, its slopes and curvatures, near the central point of the simplex $\mathbf{P}^o = 1/n$. This is done using numerical derivatives instead of Monte Carlo, enormously reducing the computational burden. This tactic fails for unsmooth statistics like the sample median, but it has worked well for a large number of examples in DiCiccio and Efron (1992), and also here for the maximum eigenvalue problem. The ABC intervals were proven to be second-order accurate for smooth statistics by DiCiccio and Efron.

Statistical error estimates based on derivatives are familiar from *delta method* or *influence function* calculations. For example, the nonparametric delta method estimate of standard error is

$$\hat{\sigma} = \left[\sum_{i=1}^{n} \dot{t}_i^2/n^2\right]^{1/2}. \quad (5.3)$$

The vector $\dot{\mathbf{t}} = (\dot{t}_1, \dot{t}_2, \ldots, \dot{t}_n)$ is the *empirical influence function,*

$$\dot{t}_i = \lim_{\varepsilon \to 0} \frac{T((1-\varepsilon)\mathbf{P}^o + \varepsilon \mathbf{e}_i) - T(\mathbf{P}^o)}{\varepsilon}, \quad (5.4)$$

where $T(\mathbf{P}^o) = t(\hat{F}) = \hat{\theta}$ and $\mathbf{e}_i$ is the ith coordinate vector $(0, 0, \ldots, 0, 1, 0, \ldots, 0)$. Having calculated $\dot{\mathbf{t}}$, the acceleration a in the BC_a formula (2.12) is given by

$$a = \frac{1}{6} \frac{\sum_{i=1}^{n} \dot{t}_i^3}{(\sum_{i=1}^{n} \dot{t}_i^2)^{3/2}} \quad (5.5)$$

(Efron 1987, sec. 7).

Definition (2.13) for z_0 looks simple compared to (5.5), but usually a is easier to estimate than z_0. The ABC algorithm quickly computes a en route to the confidence interval limits. The algorithm uses an analytic approximation for z_0 that is often more accurate than (2.13) even for very large numbers of bootstrap replications. In the maximum eigenvalue analysis of Figure 1, the algorithm gave $(z_0, a) = (.190, .099)$.

In addition to the MLE $\hat{\theta}$, the standard intervals $\hat{\theta} \pm z^{(\alpha)}\hat{\sigma}$ require only the calculation of $\hat{\sigma}$, often taken to be (5.3). The ABC intervals require three more constants, (a, z_0, c_q). Besides the acceleration a and the bias correction z_0, we need the *quadratic coefficient*

$$c_q = \lim_{\varepsilon \to 0} \frac{[T((1-\varepsilon)\mathbf{P}^o + \varepsilon \dot{\mathbf{t}}/(n^2\hat{\sigma})) - 2T(\mathbf{P}^o) + T((1-\varepsilon)\mathbf{P}^o - \varepsilon \dot{\mathbf{t}}/(n^2\hat{\sigma}))]}{\varepsilon^2}. \quad (5.6)$$

Computationally the ABC algorithm is only slightly more ambitious than a delta method analysis of standard error and bias for $\hat{\theta}$. It gives considerably more information, though, in the form of second-order accurate approximate confidence limits for θ. The definitions of a, z_0, and c_q were motivated and explained by DiCiccio and Efron (1992). The Appendix presents a nonparametric version of the ABC algorithm, written in the language S of Becker, Chambers, and Wilks (1988).

The ABC endpoints require a total of $2n + 2 + k$ recomputations of $T(\mathbf{P})$, with k being the number of endpoints desired. This amounts to 54 recomputations in rows 2 or 4 of Table 2, compared to some 2,000 recomputations for BC_a. The number can be further reduced by grouping the data points o_i, say into pairs.

The ABC algorithm requires that the statistic of interest be expressed in the resampling form $\hat{\theta}^* = T(\mathbf{P}^*)$. In the maximum eigenvalue example, calculations (1.2)–(1.5) are carried through with weight P_i^* on o_i, rather than weight $1/n$. We minimize $\sum_i \sum_j P_i^* [o_{ij} - (\mu + \alpha_i + \beta_j)]^2$ rather than (1.2), impute $\hat{x}_{ij}^* = \hat{\nu}^* + \hat{\alpha}_i^* + \hat{\beta}_j^*$ for the missing elements of $\mathbf{o}^*$, and calculate the weighted covariance matrix

$$\hat{\mathbf{\Sigma}}^* = \sum_{i=1}^{n} P_i^* (\hat{x}_i^* - \hat{\mu}^*)(\hat{x}_i^* - \hat{\mu}^*)' \qquad \left(\hat{\mu}^* = \sum_{i=1}^{n} P_i^* \hat{x}_i^*\right) \quad (5.7)$$

rather than (1.4); then $\hat{\theta}^* = T(\mathbf{P}^*)$ is the maximum eigenvalue of $\hat{\mathbf{\Sigma}}^*$. Usually the form of $T(\mathbf{P}^*)$ is obvious. Doubtful cases can be resolved by remembering that when $n\mathbf{P}^*$ is a vector of integers, say $(N_1^*, N_2^*, \ldots, N_n^*)$, then $T(\mathbf{P}^*)$ is the value of $\hat{\theta}$ applying to a sample of N_1^* copies of o_1^*, N_2^* copies of o_2^*, and so on. Remark I concerns $T(\mathbf{P}^*)$ for $\hat{\theta}$ the normal-theory MLE of θ.

It turns out to be easy to compute $\pi^\dagger(\theta|\mathbf{x})$ if $\theta_{\mathbf{x}}(\alpha)$ is the endpoint of the ABC interval (Efron 1993). For a given complete data set $\mathbf{x}$, let $(\hat{\theta}, \hat{\sigma}, a, z_0, c_q)$ be the five numbers required for the ABC endpoints. These numbers are calculated by the program abcnon in the Appendix. Let λ and w be defined as functions of θ in the following way:

$$\xi = \frac{\theta - \hat\theta}{\hat\sigma}, \qquad \lambda = \frac{2\xi}{1 + (1 + 4c_q\xi)^{1/2}},$$

$$w = \frac{2\lambda}{(1 + 2a\lambda) + (1 + 4a\lambda)^{1/2}}. \tag{5.8}$$

Then

$$\pi^\dagger(\theta|\mathbf{x}) = \frac{(1 - aw)^3}{(1 + aw)(1 + 2c_q\lambda)} \frac{e^{-(w - z_0)^2/2}}{(2\pi\hat\sigma^2)^{1/2}} \tag{5.9}$$

(see remark G and sec. 3).

Figure 5 was constructed by applying (4.3), (5.9) to the maximum eigenvalue problem of the Introduction. $M = 50$ multiple imputations $\mathbf{x}^{(m)}$ were constructed using the approximate Bayesian bootstrap (3.7). For each one, a nonparametric bootstrap sample $\mathbf{o}^{*(m)}$ gave $\hat{\mathbf{x}}^{*(m)}$ and then $\hat\eta^{*(m)} = (\hat\mu^{*(m)}, \hat{\mathfrak{V}}^{*(m)})$ as in (1.7), (1.8); then the missing components in the original data set $\mathbf{o}$ of Table 1 were filled in by sampling from the conditional normal distribution with expectation vector $\hat\mu^{*(m)}$ and covariance matrix $\hat{\mathfrak{V}}^{*(m)}$, say

$$x_i^{(m)} \,|\, o_i \sim f_{\hat\eta^{*(m)}}(x|o_i)$$

$$\text{independently for} \quad i = 1, 2, \ldots, 22. \tag{5.10}$$

Poor man's data augmentation (3.6) gave results similar to (5.10) in this example.

Each imputed data set $\mathbf{x}^{(m)}$ gave the five ABC numbers $(\hat\theta^{(m)}, \hat\sigma^{(m)}, a^{(m)}, z_0^{(m)}$, and $c_q^{(m)})$ obtained from the program abcnon in the Appendix, and these gave the confidence density $\pi^\dagger(\theta|\mathbf{x}^{(m)})$ based on $\mathbf{x}^{(m)}$, (5.6). The appropriate resampling function $T(\mathbf{P}^*)$ for the mth case, called "$tt(\mathbf{P})$" in the program, is defined as follows. Let $\hat{\mathfrak{V}}^*$ be the weighted covariance matrix

$$\hat{\mathfrak{V}}^* = \sum_{i=1}^n P_i^* (x_i^{(m)} - \hat\mu^*)(x_i^{(m)} - \hat\mu^*)'$$

$$(\hat\mu^* = \Sigma P_i^* x_i^{(m)}); \tag{5.11}$$

then $\hat\theta^* = T(\mathbf{P}^*)$ is the maximum eigenvalue of $\hat{\mathfrak{V}}^*$.

The curve $\pi^\dagger_{\text{nonpar}}(\theta|\mathbf{o})$ in Figure 5 was obtained by applying abcnon directly to the nonparametric bootstrap. The function $\hat\theta^* = T(\mathbf{P}^*)$ was now defined to be the maximum eigenvalue of $\hat{\mathfrak{V}}^*$ in (5.7). The five numbers $(\hat\theta, \hat\sigma, a, z_0$, and $c_q)$ were obtained from abcnon, giving $\pi^\dagger_{\text{nonpar}}(\theta|\mathbf{o})$ from (5.8), (5.9). Notice that in (5.7) the vectors $\hat x_i^*$ vary as functions of $\mathbf{P}^*$, whereas in (5.11) the $x_i^{(m)}$ do not.

Table 2 shows that the multiple imputation intervals are somewhat too short in the upper direction. The multiple imputation standard error estimate (Rubin and Schenker 1986), which does not involve (5.9), is similarly small:

$$\hat\sigma_{\text{mult}} = [35{,}150 + 2{,}897]^{1/2} = 195.1, \tag{5.12}$$

compared to the direct delta method estimate $\hat\sigma = 220.0$, obtained by applying (5.3) to $T(\mathbf{P}^*)$ defined from (5.7). There is no gold standard by which to judge Table 2, but the multiple imputation intervals are even 10% shorter than the intervals based on complete data for the 22 students (Efron 1992a, table 1). (The complete data intervals are about 5% shorter than those in row 1 or 2 of Table 2 here, which is consistent with having 10% more data.)

A possible source of the difficulty is the normal-theory imputation (5.10). The imputed data sets $\mathbf{x}^{(m)}$ were centered away from $\hat{\mathbf{x}}$ in a region where the complete-data delta method standard errors were noticeably smaller; $\hat\sigma(\mathbf{x}^{(m)})$ averaged 187.5, compared to either $\hat\sigma(\hat{\mathbf{x}}) = 219.8$ or $\hat\sigma(\mathbf{x}) = 214.4$ for the actual complete data set $\mathbf{x}$.

The twenty-second student in Table 1 has by far the greatest influence on the maximum eigenvalue estimate $\hat\theta = 633.2$. His empirical influence (5.10) was $t_{22} = 4{,}041.8$, compared to the next-largest values $t_{21} = 1{,}435.5$, $t_2 = 819.6$, $t_8 = -623.8, \ldots$. The two missing values in o_{22} were imputed in a noticeably different way by (5.10) as compared to (1.3), averaging (29.92, 18.44) over the 50 normal theory imputations compared to the best-fit imputation (9.87, 14.66), seen from $\hat{\mathbf{x}}$ on the right side of Table 1.

Figure 5 was recomputed after changing $x_{22}^{(m)}$ to $\hat x_{22}$ in all 50 imputations. Now $\hat\pi^\dagger(\theta|\mathbf{o})$ was in better agreement with $\pi^\dagger(\theta|\mathbf{o})$—in fact, it was slightly longer-tailed to the right; $\hat\sigma(\mathbf{x}^{(m)})$ now averaged 226.7.

Remark G. Formula (5.9) is actually the confidence density applying to the simpler approximate confidence interval method called ABC_q by DiCiccio and Efron (1992a). It is not much more difficult to compute the genuine ABC confidence density (Efron 1992, eq. 7.3), but the difference was not important here.

Remark H. It is possible to write down a K-sample version of the ABC algorithm, but the one-sample program given in the Appendix also handles the K-sample case, in the following way. The K-sample bootstrap replication $\hat\theta^* = t(\hat G_1^*, \hat G_2^*, \ldots, \hat G_K^*)$ has a resampling representation

$$\hat\theta^* = T(\mathbf{P}_1^*, \mathbf{P}_2^*, \ldots, \mathbf{P}_K^*), \tag{5.15}$$

where $\mathbf{P}_k^*$ is the kth resampling vector, $P_{ki}^* = \#\{o_{kj}^* = o_{ki}\}/n_k$, as in (5.1). Let $\mathbf{P}^*$ be a single long resampling vector of length $\sum_{k=1}^K n_k$,

$$\mathbf{P}^* = (P_{11}^*, \ldots, P_{1n_1}^*, P_{21}^*, \ldots, P_{2n_2}^*, \ldots, P_{K1}^*, \ldots, P_{K,n_K}^*), \tag{5.16}$$

and define the one-sample statistic

$$S(\mathbf{P}^*) = T(\mathbf{P}_1^*, \ldots, \mathbf{P}_K^*),$$

$$\text{where} \quad \mathbf{P}_k^* = (P_{k1}^*, P_{k2}^*, \ldots, P_{kn_k}^*) \Big/ \sum_{j=1}^{n_k} P_{kj}^*. \tag{5.17}$$

Then it can be shown that applying the one-sample abc algorithm of the Appendix to $S(\mathbf{P}^*)$ gives exactly the same confidence intervals as applying the appropriate K-sample abc program to (5.15). The acceleration a required for the BC_a intervals is obtained by applying (5.4), (5.5) to $S(\mathbf{P}^*)$.

6. SUMMARY

Three bootstrap methods for missing-data problems have been presented: nonparametric, full mechanism, and multiple imputation. Here is a brief summary of their advantages and drawbacks.

Nonparametric Bootstrap. This is easiest of the three methods to apply, both conceptually and, if the ABC algorithm is used, computationally. It requires no knowledge of the concealment mechanism leading to the observed pattern of missing data. The method applies just as well to ad hoc estimators like (1.2)–(1.5) and to MLE's. This can be convenient and efficient, as in the maximum eigenvalue problem, but opens the possibility of definitional bias. In missing-data problems this approach is limited to nonparametric settings, there being no obvious parametric equivalent of Figure 2. The method is also limited to simple random sampling situations, or multisample situations as in Remark B. The BC_a or ABC confidence intervals obtained from the nonparametric bootstrap are second-order accurate. If only a standard error is required, a bootstrap estimate can be obtained from just a couple hundred bootstrap replications.

Full-Mechanism Bootstrap. This is the approach that most closely resembles bootstrap methods for problems without missing data. It can be applied to parametric or nonparametric problems and to data situations more complicated than simple random sampling. It avoids the problem of definitional bias and can even be used to assess the definitional bias in estimators like (1.2)–(1.5). There is no equivalent of the ABC algorithm for reducing the computational burden. Nor is there a simple formula like (5.5) for the constant a used in the BC_a method (though using a based on (5.5) seems to give reasonable results).

The full-mechanism bootstrap requires knowledge of the concealment mechanism $\mathbf{x} \to \mathbf{o}$ in Figure 3. But it is sometimes of considerable interest, and even necessary to model the concealment mechanism (see Rubin 1987, chap. 6), in which case this is a less serious disadvantage.

Multiple Imputation Bootstrap. The basic data augmentation identity (3.1) is ideal for handling missing data problems for which there is a genuine Bayes prior. Its application to confidence intervals by means of confidence densities (4.3), (5.9) is computationally straightforward once the problem of sampling from the predictive density $f(\mathbf{x}|\mathbf{o})$ is solved. Here we require knowing the conditional density of $\mathbf{x}$ given $\mathbf{o}$, but not of $\mathbf{o}$ given $\mathbf{x}$ as with the full-mechanism bootstrap. Sampling methods like (3.6) or (3.7) are reasonable surrogates for the predictive density. However the maximum eigenvalue example suggests that (4.3), (5.9) may be uncomfortably vulnerable to failures in the parametric assumptions.

The multiple imputation bootstrap can be applied to parametric problems and to arbitrarily complicated data structures. Each multiple imputation $\mathbf{x}^{(m)}$ uses exactly the same set of observed data, with only the imputed numbers varying, so that the results are better conditioned on $\mathbf{o}$. The method fits in well with the EM algorithm, which is often used to find MLE's in missing data situations. Results like Figure 5 give an assessment of how much the missing data is affecting our answer. Asymptotic properties of the multiple imputation bootstrap, like second-order accuracy, have not yet been investigated.

APPENDIX: A NONPARAMETRIC ABC PROGRAM

The program abcnon, written in the language S of Becker et al. (1988), evaluates the ABC intervals described in Section 3; $tt(P)$ is the resampling function $T(\mathbf{P}^*)$.

```
''abcnon'' <-
function(tt, n, epsi = 0.001 alpha = c( .025,.05,.1,.16,.84,.9,.95,.975))
{
#abc for nonparametric problems, sample size n
#tt(P) is statistic in resampling form, where P[i] is weight on x[i]
  ep <- epsi/n; I<- diag(n); P0<- rep(1/n,n)
  t0 <- tt(P0)
#calculate t. and t.. ..........................................
  t. <- t.. <- numeric(n)
  for(i in 1:n) { di <- I[i, ] - P0
                  tp <- tt(P0 + ep * di)
                  tm <- tt(P0 - ep * di)
                  t.[i] <- (tp - tm)/(2 * ep)
                  t..[i] <- (tp - 2 * t0 + tm)/ep^2}
#calculate sighat,a,z0,and cq ..................................
  sighat <- sqrt(sum(t.^2))/n
  a <- (sum(t.^3))/6 * n^3 * sighat^3)
  delta <- t./(n^2 * sighat)
  cq <- (tt(P0+ep*delta) -2*t0 + tt(P0-ep*delta))/(2*sighat*ep^2)
  bhat <- sum(t..)/(2 * n^2)
  curv <- bhat/sighat - cq
  z0 <- qnorm(2 * pnorm(a) * pnorm( - curv))
#calculate interval endpoints.....................................
  w <- z0 + qnorm(alpha)
  lambda <- w/(1 - a * w)^2
  stan <- t0 + sighat * qnorm(alpha)
  abc <- seq(alpha)
  for(i in seq(alpha)) abc[i] <- tt(P0 + lambda[i] * delta)
  lims <- cbind(alpha, abc, stan)
#output in list form.............................................
  list(lims=lims, stats=c(t0,sighat,bhat), cons=(c(a,z0,cq)), t.=t.)
```

[Received July 1992. Received June 1993.]

REFERENCES

Becker, R. A., Chambers, J. M., and Wilks, A. R. (1988), *The New S Language,* Pacific Grove, CA: Wadsworth and Brooks/Cole.

Berger, J. O., and Bernardo, J. M. (1991), "On the Development of the Reference Prior Method," *Proceedings of the 4th Valencia International Meeting on Bayesian Statistics.*

Buck, S. F. (1960), "A Method of Estimation in Missing Values in Multivariate Data Suitable for Use With an Electronic Computer," *Journal of the Royal Statistical Society,* Ser. B, 22, 302–306.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood Estimation From Incomplete Data Via the EM Algorithm" (with discussion), *Journal of the Royal Statistical Society,* Ser. B, 39, 1–38.

DiCiccio, T. J., and Efron, B. (1992), "More Accurate Confidence Intervals in Exponential Families," *Biometrika,* 79.

Efron, B. (1981a), "Censored Data and the Boostrap," *Journal of the American Statistical Association,* 76, 312–319.

——— (1981b), "Nonparametric Estimates of Standard Error: The Jackknife, the Bootstrap, and Other Resampling Methods," *Biometrika,* 68, 589–599.

——— (1982), "The Jackknife, the Bootstrap, and Other Resampling Plans," *SIAM CBMS-NSF Monograph,* 38.

——— (1987), "Better Bootstrap Confidence Intervals and Bootstrap Approximations," *Journal of the American Statistical Association,* 82, 171–185.

——— (1992), "Jackknife-After-Bootstrap Standard Errors and Influence Functions," *Journal of the Royal Statistical Society,* Ser. B, 54, 83–127.

——— (1993), "Bayes and Likelihood Calculations From Confidence Intervals," *Biometrika,* 80, 3–26.

——— (1992b), "Six Questions Raised by the Bootstrap," in *Bootstrap Proceedings Volume,* ed. R. LePage, New York: John Wiley.

Efron, B., and Stein, C. (1981), "The Jackknife Estimate of Variance," *The Annals of Statistics,* 9, 586–596.

Efron, B., and Tibshirani, R. J. (1986), "Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy," *Statistical Science,* 1, 54–77.

Fay, R. E. (1986), "Causal Models for Patterns of Nonresponse," *Journal of the American Statistical Association,* 81, 354–365.

Hartigan, O. F., and Little, R. J. A. (1991), "Multiple Imputation for the Fatal Accident Reporting System," *Applied Statistics,* 40, 13–29.

Laird, N., and Lewis, T. A. (1987), "Empirical Bayes Confidence Intervals Based on Bootstrap Samples" (with discussion), *Journal of the American Statistical Association,* 82, 739–757.

Lazzeroni, L. C., Schenker, N., and Taylor, J. M. G. (1990), "Robustness of Multiple-Imputation Techniques to Model Misspecification," in *Proceedings of the Survey Research Methods Section, American Statistical Association.*

Little, R. J. A. (1983), "The Ignorable Case," in *Incomplete Data In Sample Surveys,* Vol. 2, Part VI, New York: Academic Press, pp. 341–382.
Little, R. J. A., and Rubin, D. B. (1987), *Statistical Analysis With Missing Data,* New York: John Wiley.
Louis, T. A. (1982), "Finding the Observed Information Matrix When Using the EM Algorithm," *Journal of the Royal Statistical Society,* Ser. B, 44, 226–233.
Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979), *Multivariate Analysis,* New York: Academic Press.
Rubin, D. B. (1976), "Inference and Missing Data," *Biometrika,* 63, 581–592.
——— (1978), "Multiple Imputations in Sample Surveys—A Phenomenological Bayesian Approach to Nonresponse," *Proceedings of the Survey Research Methods Section, American Statistical Association,* pp. 20–34.
——— (1981), "The Bayesian Bootstrap," *The Annals of Statistics,* 9, 130–134.
——— (1987), *Multiple Imputation for Nonresponse in Surveys,* New York: John Wiley.
Rubin, D. B., and Schenker, N. (1986), "Multiple Imputation for Interval Estimation From Simple Random Samples With Ignorable Nonresponse," *Journal of the American Statistical Association,* 81, 366–374.
Tanner, M. A. (1991), *Tools for Statistical Inference—Observed Data and Data Augmentation Schemes,* New York: Springer-Verlag.
Tanner, M. A., and Wong, W. H. (1987), "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association,* 82, 805–811.
Tibshirani, R. (1989), "Noninformative Priors for One Parameter of Many," *Biometrika,* 76, 604–608.
Wei, G. C. G., and Tanner, M. A. (1990), "A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms," *Journal of the American Statistical Association,* 85, 699–704.

COMMENT

Donald B. Rubin*

Efron's article is an interesting addition to the existing literature attempting to handle missing-data problems through frequentist resampling techniques. Other recent contributions include those of Rao and Shao (1992) on jackknife methods following single hot-deck imputation, which has been a survey practitioner's tool for bootstrap imputation for the better part of this century, and Fay (1993) on variants of such techniques, based on single and multiple imputation. Despite the fact that it is a pleasure to have Efron's exceptional technical adroitness and creativity brought to bear on the problem of missing data, several points of contention are raised by his article. Specifically, we have at least the following questions to consider:

1. Can we have confidence validity for an interval estimate without a well-defined estimand?
2. Can we claim distribution-free validity for a procedure whose operating characteristics vary critically with the underlying distributional assumptions?
3. Within the missing-data context, should the responsibilities and capabilities of the imputer and the ultimate user be assumed to be the same?
4. Is it acceptable to have strong hidden assumptions when creating imputations?
5. Was multiple imputation an outgrowth of EM or data augmentation?

I believe that the answer to each of these questions is "no," whereas it seems that Efron's answers would all be "yes," and so he and I should be able to provide a lively exchange of views for *JASA*'s readers.

1. CONFIDENCE VALIDITY WITHOUT A WELL-DEFINED POPULATION ESTIMAND?

Efron seems to depart from long-standing statistical tradition, dating back to Neyman's (1934) introduction of confidence intervals, by being willing to ascribe confidence validity to an interval estimate without requiring that the interval cover a well-defined population quantity (i.e., the estimand) over repeated samples with probability at least as great as the stated confidence coefficient. That is, he seems willing to find a consistent estimate of standard error for a statistic that is not consistent for a well-defined estimand and claim confidence validity for the resulting random interval.

In particular, consider Efron's motivating example with 5 variables, 22 students, and 22 missing values, where θ is the maximum eigenvalue of the population variance–covariance matrix and $\hat{\theta}$ is the maximum eigenvalue of the sample variance–covariance matrix of the data set with its missing values imputed by some method. Efron can be read as implying that confidence intervals based on the nonparametric bootstrap are valid whether or not the imputation method tracks the actual mechanism that created the missing data:

> More elaborate bootstraps are available, as will be discussed, but the simple method has much to recommend it. It is nonparametric, applicable to any kind of imputation procedure, and requires no knowledge of the missing-data mechanism. Its main practical disadvantage is the computational expense of the 2,000 or so bootstrap replications required for reasonable numerical accuracy. . . .

* Donald B. Rubin is Professor and Chairman, Department of Statistics, Harvard University, Cambridge, MA 02138.

Journal of the American Statistical Association
June 1994, Vol. 89, No. 426, Theory and Methods