MISSING DATA, IMPUTATION, AND THE BOOTSTRAP

BY

BRADLEY EFRON

TECHNICAL REPORT NO. 153

JULY 1992

PREPARED UNDER THE AUSPICES

OF

PUBLIC HEALTH SERVICE GRANTS
5 R01 GM21215-17 AND 5 R01 CA55325

DIVISION OF BIOSTATISTICS

STANFORD UNIVERSITY

STANFORD, CALIFORNIA
Missing Data, Imputation, and The Bootstrap

By

Bradley Efron

Technical Report No. 153

July 1992

Prepared Under the Auspices

Of

Public Health Service Grants

5 R01 GM21215-17 and 5 R01 CA55325

Also supported by National Science Foundation Grant DMS89-05874 and issued as Technical Report No. 397, Department of Statistics, Stanford University.

Division of Biostatistics

Stanford University

Stanford, California
Missing Data, Imputation, and The Bootstrap

By

Bradley Efron

Technical Report No. 153

July 1992

Prepared Under the Auspices

Of

Public Health Service Grants

5 R01 GM21215-17 and 5 R01 CA55325

Also supported by National Science Foundation Grant DMS89-05874 and issued as Technical Report No. 397, Department of Statistics, Stanford University.

Division of Biostatistics

Stanford University

Stanford, California
MISSING DATA, IMPUTATION, AND THE BOOTSTRAP

Bradley Efron

Abstract

Missing data refers to a class of problems made difficult by the absence of some portions of a familiar data structure. A regression problem might have some missing values in the predictor vectors, for example. A substantial theory of imputation has been developed to estimate a parameter of interest $\theta$ in a missing data situation. Here we bring bootstrap methods to bear on the question of assigning confidence intervals for $\theta$. Nonparametric bootstrap intervals based on a missing-data estimator $\hat{\theta}$ give convenient and accurate answers. There are interesting practical and theoretical differences between bootstrap methods and the multiple imputation approach, as well as some similarities.
Missing Data, Imputation, and the Bootstrap

Bradley Efron

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 \( \theta \) 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 4, is based on an appealing Bayesian analysis of the missing data structure. In this paper the bootstrap, a frequentist device, is brought to bear on missing data problems. 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 5 exams, labelled 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 we would consider the rows of the matrix to be a random sample of size \( n = 22 \) from an unknown 5-dimensional probability distribution \( F \); and that our goal is to estimate \( \theta \), the maximum eigenvalue of the covariance matrix of \( F \),

\[
\theta = \text{maximum eigenvalue of } \Sigma,
\]

where \( \Sigma \) 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

\[
\sum_i \sum_j [o_{ij} - (\nu + \alpha_i + \beta_j)]^2, \quad \left( \sum_i \alpha_i = 0, \sum_j \beta_j = 0 \right)
\]

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

\[
\tilde{x}_{ij} = \hat{\nu} + \hat{\alpha}_i + \hat{\beta}_j
\]

were used to replace the missing scores \( o_{ij} = ? \), of course setting \( \tilde{x}_{ij} = o_{ij} \) 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 Chapter 2 of Little and Rubin (1987). For now we will take it as a given part of the example under discussion.
<table>
<thead>
<tr>
<th>student</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>?</td>
<td>63</td>
<td>65</td>
<td>70</td>
<td>63</td>
<td>56.21</td>
<td>63</td>
<td>65</td>
<td>70</td>
<td>63</td>
</tr>
<tr>
<td>2</td>
<td>53</td>
<td>61</td>
<td>72</td>
<td>64</td>
<td>73</td>
<td>53</td>
<td>61</td>
<td>72</td>
<td>64</td>
<td>73</td>
</tr>
<tr>
<td>3</td>
<td>51</td>
<td>67</td>
<td>65</td>
<td>65</td>
<td>?</td>
<td>51</td>
<td>67</td>
<td>65</td>
<td>65</td>
<td>58.94</td>
</tr>
<tr>
<td>4</td>
<td>?</td>
<td>69</td>
<td>53</td>
<td>53</td>
<td>53</td>
<td>47.96</td>
<td>69</td>
<td>53</td>
<td>53</td>
<td>53</td>
</tr>
<tr>
<td>5</td>
<td>?</td>
<td>69</td>
<td>61</td>
<td>55</td>
<td>45</td>
<td>48.46</td>
<td>69</td>
<td>61</td>
<td>55</td>
<td>45</td>
</tr>
<tr>
<td>6</td>
<td>?</td>
<td>49</td>
<td>62</td>
<td>63</td>
<td>62</td>
<td>49.96</td>
<td>49</td>
<td>62</td>
<td>63</td>
<td>62</td>
</tr>
<tr>
<td>7</td>
<td>44</td>
<td>61</td>
<td>52</td>
<td>62</td>
<td>?</td>
<td>44</td>
<td>61</td>
<td>52</td>
<td>62</td>
<td>51.69</td>
</tr>
<tr>
<td>8</td>
<td>49</td>
<td>41</td>
<td>61</td>
<td>49</td>
<td>?</td>
<td>49</td>
<td>41</td>
<td>61</td>
<td>49</td>
<td>46.94</td>
</tr>
<tr>
<td>9</td>
<td>30</td>
<td>69</td>
<td>50</td>
<td>52</td>
<td>45</td>
<td>30</td>
<td>69</td>
<td>50</td>
<td>52</td>
<td>45</td>
</tr>
<tr>
<td>10</td>
<td>?</td>
<td>59</td>
<td>51</td>
<td>45</td>
<td>51</td>
<td>42.46</td>
<td>59</td>
<td>51</td>
<td>45</td>
<td>51</td>
</tr>
<tr>
<td>11</td>
<td>?</td>
<td>40</td>
<td>56</td>
<td>54</td>
<td>?</td>
<td>39.54</td>
<td>40</td>
<td>56</td>
<td>54</td>
<td>44.33</td>
</tr>
<tr>
<td>12</td>
<td>42</td>
<td>60</td>
<td>54</td>
<td>49</td>
<td>?</td>
<td>42</td>
<td>60</td>
<td>54</td>
<td>49</td>
<td>48.19</td>
</tr>
<tr>
<td>13</td>
<td>?</td>
<td>63</td>
<td>53</td>
<td>54</td>
<td>?</td>
<td>46.21</td>
<td>63</td>
<td>53</td>
<td>54</td>
<td>50.99</td>
</tr>
<tr>
<td>14</td>
<td>?</td>
<td>55</td>
<td>59</td>
<td>53</td>
<td>?</td>
<td>45.21</td>
<td>55</td>
<td>59</td>
<td>53</td>
<td>49.99</td>
</tr>
<tr>
<td>15</td>
<td>?</td>
<td>49</td>
<td>45</td>
<td>48</td>
<td>?</td>
<td>36.87</td>
<td>49</td>
<td>45</td>
<td>48</td>
<td>41.66</td>
</tr>
<tr>
<td>16</td>
<td>17</td>
<td>53</td>
<td>57</td>
<td>43</td>
<td>51</td>
<td>17</td>
<td>53</td>
<td>57</td>
<td>43</td>
<td>51</td>
</tr>
<tr>
<td>17</td>
<td>39</td>
<td>46</td>
<td>46</td>
<td>32</td>
<td>?</td>
<td>39</td>
<td>46</td>
<td>46</td>
<td>32</td>
<td>37.69</td>
</tr>
<tr>
<td>18</td>
<td>48</td>
<td>38</td>
<td>41</td>
<td>44</td>
<td>33</td>
<td>48</td>
<td>38</td>
<td>41</td>
<td>44</td>
<td>33</td>
</tr>
<tr>
<td>19</td>
<td>46</td>
<td>40</td>
<td>47</td>
<td>29</td>
<td>?</td>
<td>46</td>
<td>40</td>
<td>47</td>
<td>29</td>
<td>37.44</td>
</tr>
<tr>
<td>20</td>
<td>30</td>
<td>34</td>
<td>43</td>
<td>46</td>
<td>18</td>
<td>30</td>
<td>34</td>
<td>43</td>
<td>46</td>
<td>18</td>
</tr>
<tr>
<td>21</td>
<td>?</td>
<td>30</td>
<td>32</td>
<td>35</td>
<td>21</td>
<td>20.46</td>
<td>30</td>
<td>32</td>
<td>35</td>
<td>21</td>
</tr>
</tbody>
</table>

Table 1. Left panel: 22 students have each taken 5 exams, labelled A, B, C, D, 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).

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

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (\tilde{x}_i - \tilde{\mu})(\tilde{x}_i - \tilde{\mu})' \quad (\tilde{\mu} = \frac{1}{n} \sum_{i=1}^{n} \tilde{x}_i).$$

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

$$\hat{\theta} = \text{maximum eigenvalue of } \hat{\Sigma} = 633.2.$$  

We could just as well divide by $n-1$ instead of $n$ in defining $\hat{\Sigma}$, but (1.4) makes $\hat{\theta}$ equal the normal theory maximum likelihood estimate of $\theta$, 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 employed. Here the familiar method is the covariance calculation (1.4). The question remains how good an estimate of $\theta$ is $\hat{\theta}$? This paper concerns bootstrap approaches to answering this question.

The simplest bootstrap approach begins by writing the observed data set as

$$\mathbf{x} = (x_1, x_2, \ldots, x_n),$$

(1.6)
"o" standing for observed, where o₁ is the ith row of the matrix in the left panel of Table 1, including the question marks. For example o₁ = (?, 63, 65, 70, 63). A nonparametric bootstrap sample

\[ o^* = (o_1^*, o_2^*, \ldots, o_n^*) \]  

(1.7)

is obtained by drawing the o₁ᵢᵣ randomly and with replacement from the set \{o₁, o₂, \ldots, oᵢᵣ\}, see Efron and Tibshirani (1986). Then we can use the two-way fitting algorithm (1.2), (1.3) to impute the missing values in o*, giving say \( \hat{\mathbf{x}}^* \), and compute the bootstrap covariance matrix

\[ \hat{\Sigma}^* = \frac{1}{n} \sum_{i=1}^{n} (\hat{\mathbf{x}}_i^* - \hat{\mu}^*) (\hat{\mathbf{x}}_i^* - \hat{\mu}^*)' \quad (\hat{\mu}^* = \frac{1}{n} \sum \hat{\mathbf{x}}_i^*). \]  

(1.8)

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

A computer program drew 2200 independent bootstrap samples o*, and evaluated \( \hat{\theta}^* \) for each one. Figure 1 shows the histogram of the 2200 \( \hat{\theta}^* \) 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 bootstrap bias estimate -22.9 = 610.3 - \( \hat{\theta} \).

![Histogram](image)

Figure 1. Histogram of 2200 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 1 gives approximate confidence limits for \( \theta \) based on this histogram.
The number of bootstrap samples, 2200, is ten times that needed for estimating the standard error of \( \hat{\theta} \), see Section 9 of Efron (1987). However we need this many for the more delicate task of forming an approximate confidence interval for \( \theta \). Row 1 of Table 2 gives \( \alpha \)-level approximate confidence limits with \( \alpha = .025, .05, \ldots , .975 \), using the bootstrap confidence interval method called BC\(_a\) in Efron (1987). The 90\% central interval, for example, runs from the .05 to the .95 limit, \( \theta \in [379, 1164] \). 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.

<table>
<thead>
<tr>
<th>Confidence limit ( \alpha ):</th>
<th>.025</th>
<th>.050</th>
<th>.100</th>
<th>.160</th>
<th>.840</th>
<th>.900</th>
<th>.950</th>
<th>.975</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. BC(_a)</td>
<td>341</td>
<td>379</td>
<td>429</td>
<td>478</td>
<td>966</td>
<td>1059</td>
<td>1164</td>
<td>1253</td>
</tr>
<tr>
<td>2. ABC</td>
<td>340</td>
<td>379</td>
<td>430</td>
<td>476</td>
<td>946</td>
<td>1046</td>
<td>1172</td>
<td>1295</td>
</tr>
<tr>
<td>3. Full-mechanism BC(_a):</td>
<td>349</td>
<td>387</td>
<td>439</td>
<td>490</td>
<td>970</td>
<td>1074</td>
<td>1213</td>
<td>1300</td>
</tr>
<tr>
<td>4. ABC for MLE:</td>
<td>289</td>
<td>353</td>
<td>409</td>
<td>458</td>
<td>1014</td>
<td>1135</td>
<td>1307</td>
<td>1474</td>
</tr>
<tr>
<td>5. Multiple Imputation:</td>
<td>345</td>
<td>382</td>
<td>428</td>
<td>468</td>
<td>864</td>
<td>946</td>
<td>1063</td>
<td>1177</td>
</tr>
</tbody>
</table>

**Table 2.** Approximate nonparametric confidence limits for the maximum eigenvalue \( \theta \), (1.1), given the observed data \( o \) in Table 1. 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. Row 4: ABC limits for \( \hat{\theta} \) based on the normal-theory maximum likelihood estimate \( \hat{\theta} \) 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.

Section 2 discusses the logic of the nonparametric bootstrap method (1.6). More elaborate bootstraps are available, as will be discussed, but the simple method has a lot to recommend it. It is nonparametric, applicable to any kind of imputation procedure, and requires no knowledge of the missing-data mechanism. It’s main disadvantage is the computational expense of the 2000 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 in DiCiccio and Efron (1992), is discussed in Section 3. The computational burden is only a few percent 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 since it only imputes best-fit values, and ignores residual variability. It seems likely that \( \bar{x} \) will have a smaller empirical covariance matrix than the original, unobservable, data set \( x \). We can avoid this kind of bias by assuming a parametric model and estimating \( \theta \) by maximum likelihood. A multivariate normal model applied to the observed data \( o \) 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 EM algorithm, (1977).
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 $\theta$ 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 3, 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.

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) give 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 ideas. 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.

Most of our discussion will focus on nonparametric methods. The summary in Section 6 briefly mentions parametric confidence intervals.

2. Nonparametric Bootstrap Confidence Intervals. This section discusses the logic of the nonparametric bootstrap method as it applied to missing data problems, and describes the BC$_a$ system of approximate bootstrap confidence intervals. Section 3 shows how most of the bootstrap computational effort can be avoided using analytic approximations instead of Monte Carlo, employing the method called ABC.

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

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

with $N$ possibly infinite. A concealment process, say $O_j = c(X_j)$, results in a population $G$ of partially concealed subjects

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

In the example of Section 1, $X_j$ is a vector of 5 scores for one student, while $O_j$ 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 $F$,

$$\theta_F = s(F).$$  \hspace{1cm} (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 $G$, $o = (o_1, o_2, \ldots, o_n)$, as on the left side of Table 1. The nonparametric inference step estimates $G$ by the empirical distribution $\hat{G}$ of $o$. 

\( \hat{G} \): probability \( 1/n \) on \( o_i \) for \( i = 1, 2, \ldots, n \). (2.4)

We use some function of \( \hat{G} \), say
\[
\hat{\theta} = t(\hat{G}),
\] (2.5)
as an estimate of \( \theta_F \), for example the two-way 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 \( o^* \) from \( \hat{G} \), as at (1.7), forming the empirical distribution \( \hat{G}^* \) corresponding to \( o^* \), and calculating \( \hat{\theta}^* = t(\hat{G}^*) \).

The BC\( _\alpha \) method, described below, is a way of forming highly accurate approximate confidence intervals for the parameter
\[
\theta = t(G),
\] from the bootstrap replications \( \hat{\theta}^* \). Typically the method is second-order accurate: if \( \hat{\theta}(\alpha) \) is the endpoint of a one-sided BC\( _\alpha \) interval of intended level \( \alpha \), then
\[
\text{Prob}\{\theta < \hat{\theta}(\alpha)\} = \alpha + O_p(1/n)
\] (2.6)
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 \( z^{(\alpha)} \) is the 100\( \alpha \)th percentile of a \( N(0,1) \) distribution. Remark B extends the BC\( _\alpha \) method to the \( k \)-sample problem.

---

**Figure 2.** Diagram of nonparametric bootstrap applied to a missing-data problem; the individual members of a population of interest \( F \) are partially concealed, giving a population \( \hat{G} \) of partially concealed objects; \( o \) is a random sample of size \( n \) from \( G \); \( \hat{G} \) is the empirical distribution corresponding to \( o \); 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\( _\alpha \) and ABC methods give good approximate confidence intervals for \( \theta \) based on the bootstrap replications.
A principal advantage of the nonparametric bootstrap method is that it does not depend on
the missing-data mechanism. The process of concealment, or data loss, does not affect the method,
which conceptually begins with \( G \) rather than \( F \). This is also a potential disadvantage. There is
nothing in the method connecting the parameter of interest \( \theta_F = s(F) \) with the parameter being
estimated \( \theta = t(\hat{G}) \). The connection depends considerably on the good sense of the statistician in
choosing an estimator \( \hat{\theta} = t(\hat{G}) \).

Some choices of \( t(\hat{G}) \) are better than others for reducing the possible bias of \( \hat{\theta} \) as an estimate of
\( \theta_F \). Row 4 of Table 2 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 5-dimensional
normal distribution
\[
x_i \sim N_5(\mu, \Sigma),
\]
(2.7)
some components of which have been concealed to give the observed vector \( o_i, i = 1, 2, \cdots, n \); we
estimate \( \mu \) and \( \Sigma \) 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 max-
imum eigenvalue, in the sense that
\[
\theta = t(\hat{G}) = s(F) = \theta_F.
\]
(2.8)
The circumstances are that \( F \) is multivariate normal, and that the concealment mechanism does not
affect the likelihood function. In other words, we assume that each \( o_i \) has the appropriate marginal
normal distribution obtained from (2.7). Rubin (1987) calls this last assumption ignorable non-
response.

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%. Even if we completely trust the normality assumption
(2.7), ignorable nonresponse is unverifiable, and questionable in many realistic circumstances. It
fails if the data concealed depends on the concealed values, for example if the question marks in
Table 1 occur more frequently at extreme values of A and E. The situation resembles the robust
estimation of a population mean, where there is a trade-off between the bias and variance of possible
estimators.

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

$$
\begin{array}{c}
\text{actual} \\
\begin{array}{c}
\text{sample} \\
F \\
\downarrow s \\
\theta
\end{array} \rightarrow \\
\text{conceal} \\
\rightarrow \quad \rightarrow \quad \rightarrow \\
\text{infer} \\
\hat{F} \\
\downarrow \hat{\theta} \\
\hat{\theta}^* \\
\text{bootstrap} \\
\begin{array}{c}
\text{sample} \\
\rightarrow x^* \\
\downarrow s \\
\hat{\theta}^*
\end{array} \rightarrow \\
\text{conceal} \\
\rightarrow o^* \\
\text{infer} \\
\hat{F}^* \rightarrow \\
\downarrow \hat{\theta}^*
\end{array}
$$

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

2500 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 $\hat{x}$ based on (1.2), (1.3), putting probability 1/22 on each vector $\hat{x}_i$. (A more ambitious scheme might have obtained $\hat{F}$ by multiple imputations from $o$, as discussed in Section 4.) The bootstrap samples $x^* = (x^*_1, \cdots, x^*_n)$ were drawn by simple random sampling from $\hat{F}$, i.e. from $\{\hat{x}_1, \hat{x}_2, \cdots, \hat{x}_n\}$, and then $o^*$ was obtained by concealing the same elements of $\hat{x}^*$ as those concealed by question marks on the left side of Table 1. Best-fit imputation (1.2), (1.3) applied to $o^*$ gave $\hat{x}^*$, then $\hat{F}^*$, and finally $\hat{\theta}^*$. The histogram of the 2500 $\hat{\theta}^*$ 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 we worried about earlier. This advantage is partially negated by the more complicated inference required in going from $o$ to $\hat{F}$ rather than from $o$ to $\hat{\theta}$. There is a more serious disadvantage: here we need to specify the concealment mechanism $o_i^* = c(x_i^*)$ in order to obtain the bootstrap replications $\hat{\theta}^*$. Simply concealing the same elements as in $o$ is allowable if we are in the situation Little and Rubin (1987) called missing at random, but this is often an unrealistic assumption.

A third bootstrap method for missing data is discussed in Section 5, based directly on the multiple imputation approach. Section 6 summarizes the pros and cons of all the methods.

$BC_a$ is a method of constructing second-order accurate confidence intervals from either parametric or nonparametric bootstrap replications. See Efron (1987).
percentiles, the method requires two derived constants described below for the nonparametric case, the acceleration \( \hat{a} \) and the bias-correction \( \hat{z}_0 \). The \( \alpha \)-level BC\(_{\alpha} \) endpoint, say \( \hat{\theta}[\alpha] \), is defined by

\[
\hat{\theta}(\alpha) = \hat{H}^{-1}\Phi(\hat{z}_0 + \frac{z(\alpha)}{1 - \hat{a}(\hat{z}_0 + z(\alpha))}),
\]

where \( \Phi \) is the standard normal cdf, \( z(\alpha) = \Phi^{-1}(\alpha) \), and \( \hat{H} \) is the empirical cdf of the bootstrap replication, say \( B \) of them,

\[
\hat{H}(t) = \#\{\hat{\theta}^* < t\}/B,
\]

Central intervals for \( \theta \) are obtained from pairs of \( \alpha \) levels, for example the approximate 90\% interval \([\hat{\theta}(0.05), \hat{\theta}(0.95)]\). Definition (2.9) looks complicated, but it is computationally easy to apply. If \( \hat{a} = 0 \), \( \hat{z}_0 = 0 \), and \( \hat{H} \) is perfectly normal, then \( \hat{\theta}(\alpha) \) equals the standard confidence limit \( \hat{\theta} + z(\alpha)\hat{\sigma} \). In general, (2.9) corrects all of the second-order errors made by the standard intervals, see Section 2 of Efron (1987).

The bias-correction constant \( \hat{z}_0 \) is given by

\[
\hat{z}_0 = \Phi^{-1} \hat{H}(\hat{\theta}),
\]

\( \hat{\theta} \) being the original estimate \( \hat{\theta} = t(\hat{G}) \) in Figure 2. The acceleration \( \hat{a} \) is calculated from the empirical influence function of \( \hat{\theta} \), described as follows. Given a bootstrap sample \( o^* = (o_1^*, o_2^*, \cdots, o_n^*) \), let the resampling vector

\[
P^* = (P_1^*, P_2^*, \cdots, P_n^*)
\]

indicate the proportion of times each \( o_i \) is represented in \( o^* \),

\[
P_i^* = \#\{o_i^* = o_i\}/n.
\]

With the data \( o \) fixed, we can think of \( \hat{\theta}^* = t(\hat{F}^*) \) as a function of \( P^* \), say

\[
\hat{\theta}^* = T(P^*).
\]

Assume that \( T(P^*) \) can be smoothly defined for values of the vector \( P^* \) in a neighborhood of the central point

\[
P^o = (1/n, 1/n, \cdots, 1/n).
\]

The empirical influence function is defined to be the vector \( \hat{t} = (\hat{t}_1, \hat{t}_2, \cdots, \hat{t}_n) \), with

\[
\hat{t}_i = \lim_{\epsilon \to 0} \frac{T((1-\epsilon)P^o + \epsilon e_i) - T(P^o)}{\epsilon},
\]

where \( T(P^o) = t(\hat{F}) = \hat{\theta} \), and \( e_i \) is the \( i \)-th coordinate vector \( (0, 0, \cdots, 0, 1, 0, \cdots, 0) \). Having calculated \( \hat{t} \), the acceleration \( \hat{a} \) is given by

\[
\hat{a} = \frac{1}{6} \frac{\sum_{i=1}^n t_i^3}{(\sum_{i=1}^n t_i^2)^{3/2}},
\]
Efron (Section 7, 1987).

Definition (2.11) looks simple compared to (2.17), but usually \( \hat{a} \) is easier to estimate than \( \hat{u}_0 \). The ABC algorithm of Section 3 quickly computes \( \hat{a} \) en route to the confidence interval limits. The algorithm uses an analytic approximation for \( \hat{u}_0 \) which is often more accurate than (2.11) even for very large numbers \( B \) of bootstrap replications.

**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).

**3. The ABC Algorithm.** The main disadvantage of the BC\( \alpha \) method is the large number of bootstrap replications required. Very often, this computational burden can be avoided by using analytical expansions in place of the bootstrap Monte Carlo replications. Diciccio and Efron (1992) develop a simple algorithm called ABC, standing for “approximate bootstrap confidence” intervals, that uses numerical second derivatives to accurately approximate the endpoints of the BC\( \alpha \) 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 resampling vector \( \mathbf{P}^* \), (2.12), takes its value in the simplex

\[
S_n = \{ \mathbf{P} : P_i \geq 0, \Sigma P_i = 1 \}.
\]  

The resampled statistic \( \hat{\theta}^* = T(\mathbf{P}^*) \) can be thought of as a function on the simplex, forming a resampling surface over \( S_n \) as in Figure (6.1) of Efron (1982). The geometry of the resampling surface determines the bootstrap confidence intervals for \( \theta \). In the BC\( \alpha \) method, the surface is explored by evaluating \( T(\mathbf{P}^*) \) for some 2000 random choices of \( \mathbf{P}^* \).

The ABC algorithm approximates the BC\( \alpha \) interval endpoints by exploring the local geometry of the resampling surface, its slopes and curvatures, near the central point of the simplex \( \mathbf{P}^0 = 1/n \). 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, including the maximum eigenvalue problem. The ABC intervals are proved to be second-order accurate, for smooth statistics, in 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

\[
\bar{\sigma}_{d} = \left[ \sum_{i=1}^{n} \frac{t_i^2}{n^2} \right]^{1/2},
\]  

\( t \) being the empirical influence function (2.16). See Section 6.5 of Efron (1982) and Efron (1981b).

There is also a nonparametric delta method estimate for the bias of \( \hat{\theta} \),

\[
\bar{\text{bias}} = \sum_{i=1}^{n} \frac{\bar{t}_i}{(2n^2)},
\]  

10
where $\ddot{t}_i$ is an element of the second-order influence function,

$$
\ddot{t}_i = \lim_{\epsilon \to 0} \frac{T((1-\epsilon)P^o + \epsilon e_i) - 2T(P^o) + T((1-\epsilon)P^o - \epsilon e_i)}{\epsilon^2}.
$$

(3.4)

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 $\hat{\sigma}_{\text{delta}}$. The ABC intervals require three more constants, $(\hat{\alpha}, \hat{z}_0, \hat{c}_q)$. Besides the acceleration $\hat{\alpha}$ and the bias-correction $\hat{z}_0$, (2.17) and (2.11), we need the quadratic coefficient

$$
\hat{c}_q = \lim_{\epsilon \to 0} \frac{T((1-\epsilon)P^o + \epsilon \hat{t}/(n^2\hat{\sigma})) - 2T(P^o) + T((1-\epsilon)P^o - \epsilon \hat{t}/(n^2\hat{\sigma}))}{\epsilon^2}.
$$

(3.5)

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 $\theta$. The definitions of $\hat{\alpha}, \hat{z}_0$, and $\hat{\theta}_0$ are motivated and explained in 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 program begins by numerically evaluating $\hat{t}_i$ and $\ddot{t}_i$ for $i = 1, 2, \ldots, n$. This is most of the computational work, requiring 2n recomputations of $T(P)$. It next computes $\hat{\theta}, \hat{\alpha},$ and $\hat{\text{bias}}$ from (3.2), (2.17) and (3.3). Then $\hat{c}_q$ is evaluated from (3.5), requiring 2 more recomputations. The bias-correction $\hat{z}_0$ is calculated from an analytical approximation to (3.11)

$$
\hat{z}_0 = \Phi^{-1}\{2\Phi(\hat{\alpha}) \cdot \Phi(\text{curv})\}
$$

(3.6)

The ABC interval endpoints are computed directly from $(\hat{\theta}, \hat{\sigma})$ and $(\hat{\alpha}, \hat{z}_0, \hat{c}_q)$, using formulas developed in DiCiccio and Efron (1992), as listed in the algorithm.

The ABC endpoints require a total of $2n + 2 + k$ recomputations of $T(P)$, $k$ being the number of endpoints desired. This amounts to 54 recomputations in rows 2 or 4 of Table 2, compared to some 2000 recomputations for BCa. The number can be further reduced by grouping the data points $o_i$, say into pairs.

The ABC algorithm requires the statistic of interest to be expressed in the resampling form $\hat{\theta}^* = T(P^*)$. 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 $\bar{e}_{ij}^* = \bar{\theta}^* + \bar{\alpha}_i^* + \bar{\beta}_j^*$ for the missing elements of $o^*$, and calculate the weighted covariance matrix

$$
\hat{\Sigma}^* = \sum_{i=1}^{n} P_i^* (\hat{\hat{\theta}}_i^* - \hat{\mu}^*) (\hat{\hat{\theta}}_i^* - \hat{\mu}^*)'
$$

(3.7)

$$
\hat{\mu}^* = \sum_{i=1}^{n} P_i^* \hat{\bar{e}}_i^*
$$

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

11
As a second example we consider a missing data regression problem investigated in Section 3 of Lazzeroni et al (1990). The observed data \( o = (o_1, o_2, \cdots, o_{100}) \) consists of \( n = 100 \) observations from a two-predictor regression problem where the response variable has been lost in 20 cases, say

\[
o_i = \begin{cases} (u_{1i}, u_{2i}, y_i) & \text{for } i = 1, 2, \cdots, 80 \\ (u_{1i}, u_{2i}, ?) & \text{for } i = 81, \cdots, 100. \end{cases}
\]

(3.8)

The parameter of interest is taken here to be the probability that the response variable exceeds 15,000,

\[
\theta = \text{Prob}\{Y > 15000\}.
\]

(3.9)

Figure 4 shows the data used in this example. For plotting purposes, each pair \((u_{1i}, u_{2i})\) has been reduced to the single predictor

\[
\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 u_{1i} + \hat{\beta}_2 u_{2i},
\]

(3.10)

**Figure 4.** Missing data regression example; \( o_i = (u_{1i}, u_{2i}, y_i) \) for 80 cases, \( o_i = (u_{1i}, u_{2i}, ?) \) for 20 cases; horizontal axis is predictor \( \hat{y}_i = -21671 + 3386u_{1i} + 3246u_{2i} \), based on least-squares fit to the 80 complete cases; vertical axis is \( y \). The 80 complete cases indicated by + marks; the 20 incomplete cases indicated by dotted lines, representing the imputed distribution of \( y_i \) given \((u_{1i}, u_{2i})\), (3.12). Parameter of interest is \( \theta = \text{Prob}\{Y > 15000\} \). Model suggested by Monte Carlo study in Lazzeroni et al. (1990).
with \((\beta_0, \beta_1, \beta_2) = (-21671, 3386, 3246)\), the least-squares coefficients for predicting \(y\) based on the 80 complete cases. This also gives the 80 residuals

\[
\hat{e}_j = y_j - \hat{y}_j \quad j = 1, 2, \ldots, 80.
\]  

(3.11)

Plus marks in Figure 4 indicate the 80 complete case \((\hat{y}_i, y_i)\). The 20 incomplete cases are indicated by dotted lines, each dot being a point \((\hat{y}_i, y_{ij})\), with

\[
y_{ij} = \hat{y}_i + \hat{e}_j \quad i = 81, \ldots, 100, \text{ and } j = 1, \ldots, 80,
\]  

(3.12)

(3.12) represents a simple form of multiple imputation for the missing \(y\)-values.

Figure 4 suggests a reasonable estimate for \(\theta = \text{Prob}\{Y > 15000\}\),

\[
\hat{\theta} = \frac{\#\{y_i > 15000; i = 1, 2, \ldots, 80\} + \#\{y_{ij} > 15000; i = 81, \ldots, 100 \text{ and } j = 1, 2, \ldots, 80\}}{100}
\]  

(3.13)

equalling \(\hat{\theta} = .240\) for our data. This compares with the naive estimate based on just the 80 complete observations, \(16/80 = .200\).

How accurate is \(\hat{\theta}\) as an estimate of \(\theta\)? BC\(_a\) and ABC confidence interval endpoints for \(\theta\) are shown in Table 3. The approximate central 90\% ABC interval is

\[
\theta \in (.160, 334),
\]  

(3.14)

compared with the BC\(_a\) interval (.155, 336) obtained from 2000 nonparametric bootstrap replications of \(\hat{\theta}\). The bootstrap c.d.f. \(\hat{F}\) was close to normal in this case, and \(\hat{\alpha}\) and \(\hat{\beta}_0\) were small, so that the bootstrap intervals were not much different than the standard intervals.

<table>
<thead>
<tr>
<th>Confidence limit (\alpha):</th>
<th>.025</th>
<th>.050</th>
<th>.100</th>
<th>.160</th>
<th>.840</th>
<th>.900</th>
<th>.950</th>
<th>.975</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. BC(_a)</td>
<td>.140</td>
<td>.155</td>
<td>.174</td>
<td>.187</td>
<td>.297</td>
<td>.315</td>
<td>.336</td>
<td>.357</td>
</tr>
<tr>
<td>2. ABC</td>
<td>.147</td>
<td>.160</td>
<td>.177</td>
<td>.191</td>
<td>.300</td>
<td>.313</td>
<td>.334</td>
<td>.352</td>
</tr>
</tbody>
</table>

Table 3. BC\(_a\) and ABC limits for \(\theta = \text{Prob}\{Y > 15000\}; \hat{\theta} = .240\), (3.13); BC\(_a\) results based on 2000 nonparametric bootstrap replications, shown in Figure 2. Standard error estimates for \(\hat{\theta}: BC\(_a\) .055; ABC .053; delta method .044. Other constants: \(\hat{\alpha} = .021, \hat{\alpha}_y = .006, \hat{\beta}_0 = .040\) (2.11) or .029 (2.6), bias = .000, (3.3).

The bootstrap standard error estimate for \(\hat{\theta}\) was \(\hat{\sigma}_{\text{boot}} = .055\), this being the empirical standard deviation of the 2000 \(\hat{\theta}^*\) values. The distance between the ABC interval endpoints gives an ABC estimate of standard error, for instance

\[
\hat{\sigma}_{\text{abc}} = .053 = \frac{334 - .160}{2 \cdot 1.645}
\]  

(3.15)
based on (3.14). As a point of comparison, the delta-method estimate (3.2) equalled

$$\hat{\sigma}_{\text{delta}} = .044.$$  \hfill (3.16)

Delta method standard errors seem often to be too small, while jackknife standard error estimates are biased upwards. See Efron and Stein (1981) and Efron (1992b). If \( \hat{\theta}_{(i)} \) is the statistic \( \hat{\theta} \) based on the \((n - 1)\) element data set excluding point \( o_i \), then

$$\hat{\sigma}_{\text{jack}} = \left\{ \frac{n - 1}{n} \sum_{i=1}^{n} [\hat{\theta}_{(i)} - \hat{\theta}(\cdot)]^2 \right\}^{1/2} \quad (\hat{\theta}(\cdot) = \sum_{i=1}^{n} \hat{\theta}_{(i)}/n) \hfill (3.17)$$

is the jackknife estimate of standard error. For the regression example

$$\hat{\sigma}_{\text{jack}} = .057.$$  \hfill (3.18)

In the author's experience, the order of performance for standard error estimation is \( \hat{\sigma}_{\text{boot}} \) or \( \hat{\sigma}_{\text{abc}}, \hat{\sigma}_{\text{jack}}, \hat{\sigma}_{\text{delta}}, \) see Efron (1981) and Remark H.

Influence functions indicate the effect of individual data points on the statistic \( \hat{\theta} \). The empirical influence function (2.16) is calculated as part of the abc algorithm. The left panel of Figure 5 shows the empirical influences for the missing data regression example. The right panel shows the jackknife influence function,

![Empirical Influence Function](image1)

![Jackknife Influence Function](image2)

**Figure 5.** Influence functions for the missing data regression example; “m” indicates points \( o_i \) missing the response variable \( y_i \). **Left panel:** empirical influence (2.16). **Right panel:** jackknife influence (3.19). The jackknife is greater than the empirical influence for complete points \( o_i \) having large values of \( \hat{y}_i \).
\[ i_{i,jack} = (n - 1)(\hat{\theta}(i) - \hat{\theta}(i)), \quad (3.19) \]

Efron (1991). We see that the 20 data points with missing \( y \) values tend to have influence nearer zero than the complete points. The jackknife influences for the complete points are more extreme than the \( i_i \) for points on the right end of the \( \hat{y} \) scale.

**Remark B.** The nonparametric bootstrap of Figure 2 can also be applied to \( K \)-sample problems

\[ (G_1, G_2, \ldots, G_K) \rightarrow (o_1, o_2, \ldots, o_K), \quad (3.20) \]

where the \( o_k \) 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 \( o_k \) give independent bootstrap samples \( o_k^* \) of size \( n_k \), from which we calculate \( \hat{\theta}^* \) and finally \( \hat{\theta}^* = t(\hat{G}_1^*, \hat{G}_2^*, \ldots, \hat{G}_K^*) \). The BC_\alpha intervals are calculated from (2.9), (2.15) as before, with the acceleration \( \hat{\alpha} \) obtained as in Remark C.

**Remark C.** It is possible to write down a \( K \)-sample version of the ABC algorithm, but for nonparametric problems 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

\[ \hat{\theta}^* = T(P_1^*, P_2^*, \ldots, P_K^*), \quad (3.21) \]

where \( P_k^* \) is the \( k \)th resampling vector, \( P_{kj}^* = \#\{o_{kj}^* = o_{kj}\}/n_i \), as in (2.13). Let \( P^* \) be a single long resampling vector of length \( \sum_{k=1}^{K} n_k \),

\[ P^* = (P_{11}^*, \ldots, P_{1n_1}^*, P_{21}^*, \ldots, P_{2n_2}^*, \ldots, P_{K1}^*, \ldots, P_{Kn_K}^*), \quad (3.22) \]

and define the one-sample statistic

\[ S(P^*) = T(P_1^*, \ldots, P_K^*) \text{ where } P_k^* = (P_{k1}^*, P_{k2}^*, \ldots, P_{kn_k}^*)/\sum_{j=1}^{n_k} P_{kj}^*. \quad (3.23) \]

Then it can be shown that applying the one-sample abc algorithm of the Appendix to \( S(P^*) \) gives exactly the same confidence intervals as applying the appropriate \( K \)-sample abc program to (3.20). The acceleration \( \hat{\alpha} \) required for the BC_\alpha intervals is obtained by applying (2.16), (2.17) to \( S(P^*) \).

**Remark D.** DiCiccio and Efron (1992) show that the BC_\alpha and ABC intervals are second-order accurate if the data set is obtained by random sampling from a multi-parameter exponential family, and \( \theta \) 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 example in Table 1, each
of the 5 coordinates of an \( \alpha_i \) 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 in Section 8 of Efron (1987). More direct proofs of second-order accuracy for nonparametric bootstrap intervals are being pursued by DiCiccio and Efron, both to avoid the artificial aspect of discretization and to give a realistic assessment of how well the approximations work for small values of \( n \).

**Remark E.** Row 4 of Table 2 refers to the normal-theory MLE for \( \theta \): let \( f_\eta(x) \) represent the 5-dimensional normal density \( N_5(\mu, \Sigma) \), \( \eta \equiv (\mu, \Sigma) \), let \( \hat{\ell}_\eta(x) \) be the score vector \( \frac{\partial}{\partial \eta} \log f_\eta(x) \), and define \( E_i(\eta) = E_\eta(\hat{\ell}_\eta(x_i)|\alpha_i) \); then the MLE \( \hat{\eta} = (\hat{\mu}, \hat{\Sigma}) \) is the solution of the score equation

\[
\hat{\eta} : \frac{1}{n} \sum_{i=1}^{n} E_i(\eta) = 0, \tag{3.24}
\]

and \( \hat{\theta} \) is the maximum eigenvalue of \( \hat{\Sigma} \). In this case, the function \( \hat{\theta}^* = T(P^*) \) required to apply the ABC algorithm takes \( \hat{\theta}^* \) equal to the maximum eigenvalue of \( \hat{\Sigma}^* \), where \( \hat{\Sigma}^* = (\hat{\mu}^*, \hat{\Sigma}^*) \) solves

\[
\hat{\eta}^* : \sum_{i=1}^{n} P_i^* E_i(\eta) = 0. \tag{3.25}
\]

Because the ABC algorithm uses resampling vectors \( P^* \) very near \( P^0 = (1/n, \ldots, 1/n) \), it is easy to find \( \hat{\eta}^* \) by a few Newton-Raphson steps starting at \( \hat{\eta} \). If iterative Monte Carlo methods are used to solve (3.24), as in Wei and Tanner (1990), importance sampling methods can be used to solve (3.25) based on the same set of Monte Carlo replications.

4. 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 in Tanner and Wong (1987). For comparison, the method is applied to the regression problem of Section 3. Section 5 presents a different bootstrap method, based on multiple imputation.

Let \( o \) indicate the observed data set and \( x \) indicate any complete data set consonant with \( o \). In the example of Table 1, \( x \) could be any \( 22 \times 5 \) matrix of numbers agreeing with \( o \) at all of its numerical entries. The actual complete data set \( x \) giving rise to \( o \), which would have been observed if there were no missing data, is assumed to be sampled from a parametric family with density function \( f_\eta(x) \), \( \eta \) being an unknown \( p \)-dimensional parameter vector. Starting with a prior density \( \pi_0(\eta) \) on \( \eta \), Bayes' theorem would give hypothetical posterior density \( \pi(\eta|x) \) if \( x \) were observed,
and, more concretely, the actual posterior density $\pi(\eta|o)$ having observed $o$. A standard probability calculation relates $\pi(\eta|o)$ to $\pi(\eta|x)$,

$$
\pi(\eta|o) = \int_x \pi(\eta|x)f(x|o)dx,
$$

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

$$
f(x|o) = \int_\eta f_\eta(x|o)\pi(\eta|o)d\eta
$$

The integral in (4.1) is taken over all $x$ consonant with $o$.

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

$$
\hat{\pi}(\eta|o) = \frac{1}{M} \sum_{m=1}^M \pi(\eta|x^{(m)}),
$$

where $x^{(1)}, x^{(2)}, \ldots, x^{(M)}$ are the multiple imputations, i.e. independent draws from $f(x|o)$. This argument has a circular look since we need to know $\pi(\eta|o)$ in order to calculate $f(x|o)$ in (4.2). Tanner and Wong (1987) investigate an iterative algorithm, related to Gibbs’ sampling, for actually carrying out (4.3). Non-iterative approximations are available. One such approximation based on the bootstrap is used in the example below.

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

$$
\theta = t(\eta),
$$

like the maximum eigenvalue in Section 1, rather than for the entire vector $\eta$. The marginal posterior densities of $\theta$, say $\pi^\theta(\theta|o)$ and $\pi^\theta(\theta|x)$, are related by a marginalized version of (4.1),

$$
\pi^\theta(\theta|o) = \int_x \pi^\theta(\theta|x)f(x|o)dx,
$$

$f(x|o)$ still being defined by (4.2).

For practical purposes it may be enough to know the posterior mean and variance of $\theta$ given $o$. Result (4.5) allows us to compute these in terms of the posterior mean and variance given $x$,

$$
E(\theta|o) = \int_x E(\theta|x)f(x|o)dx,
$$

and

$$
\text{var}(\theta|o) = \int_x [\text{var}(\theta|x) + (E(\theta|x) - E(\theta|o))^2]f(x|o)dx.
$$
For a given imputed data set \( x^{(m)} \) let \( E^{(m)} = E\{\theta|x^{(m)}\} \) and \( v^{(m)} = \text{var}\{\theta|x^{(m)}\} \). Then we can estimate \( E\{\theta|o\} \) and \( \text{var}\{\theta|o\} \) as in (4.3),

\[
\hat{E}\{\theta|o\} = E^{(\cdot)} = \frac{1}{M} \sum_{m=1}^{M} E^{(m)}
\]

and

\[
\hat{\text{var}}\{\theta|o\} = \frac{1}{M} \sum_{m=1}^{M} v^{(m)} + \frac{1}{M-1} \sum_{m=1}^{M} (E^{(m)} - E^{(\cdot)})^2.
\]

(4.8)

(4.9)

Rubin and Schenker (1986), use a slightly different version of (4.9).

The discussion so far in this section has been purely Bayesian. Frequentist results are obtained in the multiple imputation literature by starting with some sort of uninformative prior for the parameter vector \( \eta \). Nonparametric results can be obtained in some situations by broadening the family \( f_{\eta}(x) \) to include all possible distributions. As an example we consider the missing-data regression problem of Figure 4 from the multiple imputation point of view.

The first difficulty in applying (4.3) is the generation of imputations \( x^{(m)} \) from the predictive density \( f(x|o) \), (4.2). A simple approach, called "poor man's data augmentation" in Wei and Tanner (1990), is to sample the \( x^{(m)} \) from \( f_{\eta}(x|o) \) with \( \eta \) set equal to the MLE \( \hat{\eta} \),

\[
x^{(m)} \sim f_{\hat{\eta}}(x|o), \quad \text{independently for } m = 1, 2, \ldots, M.
\]

(4.10)

This could also be called a conditional parametric bootstrap sample. In many situations (4.10) is quite satisfactory, though it can underestimate variability if there is too much missing data. See Remark F.

A better approximation to the predictive density is obtained by drawing each imputation \( x^{(m)} \) from its own bootstrapped choice of the parameter vector \( \eta \),

\[
x^{(m)} \sim f_{\hat{\eta}^{(m)}}(x|o), \quad \text{independently for } m = 1, 2, \ldots, M.
\]

(4.11)

A bootstrap parameter vector \( \hat{\eta}^{*} \) is the MLE for \( \eta \) based on a bootstrap sample \( o^{*} \), (1.7). Rubin (1981) and Efron (1982, Section 10.6) point out that the bootstrap distribution of \( \hat{\eta}^{*} \) is quite close to the nonparametric Bayes posterior distribution of \( \eta \) given \( o \), starting from a vague Dirichlet prior for \( G \), (2.2). The difference between (4.10) and (4.11) is the difference between a first and second level bootstrap sample.

It is easy to apply (4.11) to the missing data regression problem, (3.8)-(3.11), as long as we interpret \( f_{\hat{\eta}^{*}}(x|o) \) in a nonparametric sense: the bootstrap sample \( o^{*} = (o_{1}^{*}, o_{2}^{*}, \ldots, o_{100}^{*}) \) gives least-squares coefficients \( \hat{\beta}^{*} \) and a vector of residuals \( \hat{e}^{*} \) based on the complete cases in \( o^{*} \), and then a predicted value \( \hat{y}_{i}^{*} = \hat{\beta}_{0}^{*} + \hat{\beta}_{1}^{*} u_{1i} + \hat{\beta}_{2}^{*} u_{2i} \) for each of the 20 original incomplete cases in \( o \); the imputation \( x^{(m)} \) assigns the \( i \)th incomplete case in \( o \) the imputed value

\[
x_{i}^{(m)} = (u_{1i}, u_{2i}, \hat{y}_{i}^{*} + \hat{e}_{i}^{*}),
\]

(4.12)
where $\hat{\psi}$ is a random draw from $\tilde{\psi}$. Of course $x_i^{(m)}$ equals $o_i$ for the 80 complete cases.

In this way, 200 imputed data sets $x^{(1)}, x^{(2)}, \ldots, x^{(200)}$ were drawn, and for each one the estimate

$$\hat{\theta}^{(m)} = \# \{ x_i^{(m)} \text{ with } y_i^{(m)} > 15000 \} / 100$$  \hspace{1cm} (4.13)

was calculated. There is no missing data now, so $\hat{\theta}^{(m)}$ would be assigned binomial variance

$$\text{var}^{(m)} = \theta(m)(1 - \theta(m))/100$$  \hspace{1cm} (4.14)

in a standard frequentist analysis.

It seems reasonable to take $E^{(m)} = \hat{\theta}^{(m)}$ and $v^{(m)} = \text{var}^{(m)}$ in order to get approximate frequentist results from the Bayesian relationships (4.8), (4.9). The 200 imputations gave point estimate

$$\hat{\theta} = .241$$  \hspace{1cm} (4.15)

based on (4.8), with a standard error based on (4.9) of

$$\left[ \text{var}^{(c)} + \frac{1}{199} \sum (\hat{\theta}^{(m)} - \hat{\theta}^{(c)})^2 \right]^{1/2} = [.00182 + .00069]^{1/2} = .050,$$

agreeing only moderately well with $\hat{\sigma}_{\text{boot}} = .055$ or $\hat{\sigma}_{\text{abc}} = .053$. As a benchmark, the usual binomial standard error estimate for 100 complete cases would be $[.241 \cdot 759/100]^{1/2} = .043$.

Here we have used multiple imputation to estimate a standard error for $\theta$. Section 5 uses a different kind of bootstrap method to derive approximate confidence intervals for $\theta$ based on the data augmentation identity.

**Remark E.** Wei and Tanner (1990) suggest two "poor man’s data augmentation" schemes, the easiest of which is (4.10). This is a useful suggestion, but the following simple example shows that using (4.10) instead of $f(x|o)$ in (4.1) can substantively underestimate the variance of the posterior density, if the proportion of missing data is too large. Let $x = (x_1, x_2, \ldots, x_n)$ be an i.i.d. sequence of $n$ real-valued observations from a $N(0, 1)$ distribution, and suppose we observe $o = (x_1, \ldots, x_{n_1})$, the first $n_1$ of them, losing the last $n_2 = n - n_1$. Starting with a uniform prior density $\pi_0(\theta) = 1$ gives posterior distribution

$$\pi(\theta|o) \sim N(\bar{x}_1, \frac{1}{n_1}) \quad (\bar{x}_1 = \sum_{i=1}^{n_1} x_i / n_1).$$  \hspace{1cm} (4.17)

The MLE is $\hat{\theta} = \bar{x}_1$. Using $f_{\theta}(x|o)$ instead of $f(x|o)$ in (4.1) leads to the poor man’s posterior distribution

$$\pi(\theta|o) \sim N(\bar{x}_1, \frac{1}{n_1}[1 - r^2]) \quad \text{where} \quad r = \frac{n_2}{n}.$$  \hspace{1cm} (4.18)

The difference between $\pi(\theta|o)$ and $\pi(\theta|o)$ is tolerable for most purposes if $r$ is less than .1 or .2, but not so for larger proportions of missing data. The example of Figure 4 is near the boundary of tolerance.
Remark G. Louis (1982) derives an interesting likelihood-based variant of the data augmentation identity (4.1),

$$-\tilde{L}_n(\eta) = E_\eta\{-\tilde{L}_n(x)\mid o\} - \text{cov}_n\{\tilde{L}_n(x)\mid o\}. \quad (4.19)$$

Here $\tilde{L}_n(x)$ is the score vector based on $x$, with $j$th component $\partial \log f_n(x) / \partial \eta_j$, $\tilde{L}_n(o)$ is the score vector based on $o$, and $-\tilde{L}_n(x)$ is the observed Fisher information matrix, with $jk$th entry $-\partial^2 \log f_n(x) / \partial \eta_j \partial \eta_k$. This result extends the Fisherian identity $\tilde{L}_n(o) = E_\eta\{\tilde{L}_n(x)\mid o\}$ underlying the EM algorithm, see Efron’s commentary on Dempster, Laird and Rubin (1977).

In a parametric problem we can use (4.19) to approximate $-\tilde{L}_n(o)$, and then estimate the standard error of a function of interest $\hat{\theta} = t(\tilde{\eta})$ by

$$\hat{\sigma}_{\text{fisher}} = \left[\hat{t}(\tilde{\eta})'\{-\tilde{L}_n(o)\}^{-1}\hat{t}(\tilde{\eta})\right]^{1/2}, \quad (4.20)$$

$\hat{t}(\tilde{\eta})' \equiv (\cdots \partial t(\eta) / \partial \eta_j \cdots)|_{\tilde{\eta}}$. To do so, let $x^{(1)}, x^{(2)}, \ldots, x^{(M)}$ be a poor man’s data augmentation sample (4.10). Then $-\tilde{L}_n(o)$ can be estimated by substituting the empirical mean of the matrices $-\tilde{L}_n^{(m)}(x^{(m)})$ and empirical covariance of the vectors $\tilde{L}_n^{(m)}(x^{(m)})$ into (4.19). Notice that (4.19) allows us to use poor man’s data augmentation without apology.

Remark H. Other useful methods of estimating the standard error of a missing-data estimate $\hat{\theta}$ are developed in Meng and Rubin (1989), Meilijson (1989), and Carlin (1987). These, and (4.20), are delta methods, and as such are subject to the warning following (3.16).

Remark I. The data augmentation identity requires the correct specification of $f_n(x\mid o)$ in (4.2). In practice this usually means making some assumption about the concealment mechanism giving $o$ from $x$, such as ignorable nonresponse.

5. Multiple-Imputation Bootstrap. Suppose now that we have satisfactorily solved the problem of sampling from the predictive density $f(x\mid o)$ to obtain imputations $x^{(1)}, x^{(2)}, \ldots, x^{(M)}$, and wish to construct good approximate confidence intervals for a parameter of interest $\theta = t(\eta)$. The data augmentation identity (4.5) suggests using the percentiles of the estimated posterior density

$$\hat{\pi}_\theta(\theta\mid o) = \frac{1}{M} \sum_{m=1}^{M} \pi_\theta(\theta\mid x^{(m)}) \quad (5.1)$$

as the endpoints for such intervals. However to do so requires that the complete-data posterior density $\pi_\theta(\theta\mid x)$ enjoy good confidence properties. Choosing an appropriate uninformative prior $\pi_0(\eta)$ for the vector parameter $\eta$ can be a difficult problem, see Berger and Bernardo (1991). This section uses a bootstrap-based method developed in Efron (1992a) to avoid such choices and to simplify the calculation of (5.1).
Let \( \theta(x)(\alpha) \) indicate the \( \alpha \)-level endpoint of an exact or approximate system of confidence intervals for \( \theta \) based on data \( x \). The confidence density for \( \theta \) given \( x \) is defined to be

\[
\pi^\dagger(\theta|x) = 1 - \frac{d \theta_x(\alpha)}{d \alpha}.
\] (5.2)

This density assigns probability .01 to \( \theta \) lying between the .90 and .91 confidence limits, etc. By definition, the 100\( \alpha \)th percentile of \( \pi^\dagger(\theta|x) \) is \( \theta_x(\alpha) \), so \( \pi^\dagger(\theta|x) \) is just another way of describing the function \( \theta_x(\alpha) \). However the confidence density is particularly convenient for use in (5.1), giving

\[
\hat{\pi}^\dagger(\theta|o) = \frac{1}{M} \sum_{m=1}^{M} \pi^\dagger(\theta|x^{(m)}).
\] (5.3)

It turns out to be easy to compute \( \pi^\dagger(\theta|x) \) if \( \theta_x(\alpha) \) is the endpoint of the ABC interval, Efron (1992a). For a given complete data set \( x \) let \( (\hat{\theta}, \hat{\sigma}, \hat{\alpha}, \hat{\tau}_0, \hat{\sigma}_t) \) be the 5 numbers required for the ABC endpoints. These numbers are calculated by the program abcn on in the Appendix. Let \( \lambda \) and \( w \) be defined as follows,

\[
\theta \rightarrow \xi = \frac{\theta - \hat{\theta}}{\hat{\sigma}} \rightarrow \lambda = \frac{2 \xi}{1 + (1 + 4 \hat{\sigma}_t \xi)^{1/2}} \rightarrow w = \frac{2 \lambda}{(1 + 2 \lambda) + (1 + 4 \lambda)^{1/2}}.
\] (5.4)

Then

\[
\pi^\dagger(\theta|x) = \frac{(1 - \hat{\alpha}w)^3}{(1 + \hat{\omega})(1 + 2 \hat{\sigma}_t \lambda)} \cdot e^{-\frac{1}{2}(w - \hat{\omega})^2}
\frac{(2\pi \hat{\sigma}_t)^{1/2}}{e^{(w - \hat{\omega})^2}}
\] (5.5)

See Remark L.

Row 5 of Table 1 was constructed by applying (5.3), (5.5) to the maximum eigenvalue problem of the Introduction. \( M = 50 \) multiple imputations \( x^{(m)} \) were constructed using the bootstrap method (4.11). For each one, a nonparametric bootstrap sample \( \alpha^{(m)} \) gave \( \hat{\pi}^{\dagger^{(m)}} \) and then \( \hat{\pi}^{\dagger^{(m)}} = (\hat{\mu}^{\dagger^{(m)}}, \hat{\Sigma}^{\dagger^{(m)}}) \) as in (1.7), (1.8); the missing components in the original data set \( o \) of Table 1 were filled in by sampling from the conditional normal distribution with expectation vector \( \hat{\mu}^{\dagger^{(m)}} \) and covariance matrix \( \hat{\Sigma}^{\dagger^{(m)}} \), say

\[
x^{(m)}_i o_i \sim f_{n^{\dagger^{(m)}}(m)}(x|o_i) \quad \text{independently for } i = 1, 2, \cdots, 22.
\] (5.6)

Poor man’s data augmentation (4.10) gave results similar to (5.6) in this example.

Each imputed data set \( x^{(m)} \) gave the five ABC numbers \( (\hat{\mu}^{(m)}, \hat{\sigma}^{(m)}, \hat{\alpha}^{(m)}, \hat{\tau}_0^{(m)}, \hat{\sigma}_t^{(m)}) \), obtained from the program abcn on in the Appendix, and then the confidence density \( \pi^\dagger(\theta|x^{(m)}) \) based on \( x^{(m)}, (5.5) \). The appropriate resampling function \( T(P^*) \) for the \( m \)th case, (2.14), called “tt(P)” in the program, is defined as follows: let \( \hat{\Sigma}^* \) be the weighted covariance matrix

\[
\hat{\Sigma}^* = \sum_{i=1}^{n} P_i^*(x_i^{(m)} - \hat{\mu}^*)(x_i^{(m)} - \hat{\mu}^*)'
\quad (\hat{\mu}^* = \Sigma P_i^* x_i^{(m)});
\] (5.7)
then $\hat{\theta}^* = T(P^*)$ is the maximum eigenvalue of $\hat{S}^*$.

The left panel of Figure 6 shows $\pi^t(\theta|x^{(m)})$ for $m = 1, 2, \ldots, 25$. The solid curve in the right panel is $\hat{S}^t(\theta|x)$, (5.3), the average of all $M = 50$ densities $\pi^t(\theta|x^{(m)})$. For comparison, the dashed line shows $\hat{S}^t(\theta|x)$, the ABC density (5.5) based directly on $(\hat{\theta}, \hat{\sigma}, \hat{a}, \hat{Z}_0, \hat{Z}_1)$ for the incomplete data $o$, as in Section 3. These $5$ numbers were obtained by applying abcmn to $T(P^*)$ as defined by (3.7). Notice that in (3.7) the vectors $\hat{x}_i^{(m)}$ vary as functions of $P^*$, while in (5.7) the $x_i^{(m)}$ do not. Row 5 of Table 2 consists of the appropriate percentiles of $\hat{S}^t(\theta|x)$. Row 2 is the percentiles of $\pi^t(\theta|x)$.

Table 2 shows that the multiple-imputation intervals are somewhat too short in the upper direction. The multiple imputation standard error estimate based on (4.9), which does not involve (5.5), is similarly small:

$$\hat{S}_{\text{mult}} = \sqrt{35150 + 2897} = 195.1,$$

(5.8)

compared to the direct delta-method estimate $\hat{S}_{\text{delta}} = 220.0$, obtained by applying (3.2) to $T(P^*)$ defined from (3.7). 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, Table 1 of Efron (1992a). (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 (5.6). The imputed data sets $x^{(m)}$ were centered away from $\hat{x}$, in a region where the complete-data delta-method standard errors were noticeably smaller: $\hat{S}_{\text{delta}}(x^{(m)})$ averaged 187.5, compared to either $\hat{S}_{\text{delta}}(\hat{x}) = 219.8$, or to $\hat{S}_{\text{delta}}(\hat{x}) = 214.4$ for the actual complete data set $x$.

**Remark J.** The 22nd student in Table 1 has by far the greatest influence on the maximum eigenvalue estimate $\hat{\theta} = 633.2$. His empirical influence (2.16) was $t_{22} = 4041.8$ compared to the next largest values $t_{21} = 1435.5$, $t_2 = 819.6$, $t_4 = -623.8$, $\ldots$. The two missing values in $o_{22}$ were imputed in a noticeably different way by (5.6) 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), as seen in $\hat{x}$ on the right side of Table 1.

Figure 6 was recomputed after changing $x_{22}^{(m)}$ to $\hat{x}_{22}$ in all 50 imputations. Now $\hat{S}^t(\theta|x)$ was in better agreement with $\pi^t(\theta|x)$, in fact being slightly longer-tailed to the right; $\hat{S}_{\text{delta}}(x^{(m)})$ now averaged 226.7.

**Remark K.** The normal-theory imputation (5.6), which is suspect here, might have been replaced by a less parametric method, as in (4.12). The smaller sample size and more complex missing data patterns make this more difficult to do than in the regression problem of Section 4.
Figure 6. Multiple-Imputation bootstrap for the maximum eigenvalue problem of Section 1. Left Panel: ABC confidence densities $\pi^*(\theta|x^{(m)})$ for 25 imputed data sets $x^{(m)}$, $m = 1, 2, \ldots, 25$; Right Panel: solid line is the multiple-imputation bootstrap density $\hat{\pi}^*(\theta|o)$, (5.3), average of $\pi^*(\theta|x^{(m)})$ for $m = 1, 2, \ldots, 50$. Dashed curve is $\pi^*(\theta|o)$, the ABC density (5.5) based directly on $o$, (3.7); $\hat{\pi}^*(\theta|o)$ has a shorter upper tail than $\pi^*(\theta|o)$.

Remark L. Formula (5.5) is actually the confidence density applying to the simpler approximate confidence interval method called $ABC_q$ in DiCiccio and Efron (1992a). It is not much more difficult to compute the genuine ABC confidence density, equation (7.3) of Efron (1992), but the difference was not important here.

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) as to MLEs. 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 multi-sample
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, the bootstrap estimate $\hat{\sigma}_{\text{boot}}$ can be obtained from just a couple hundred bootstrap replications.

The nonparametric influence function $\hat{\iota}$ computed in the ABC algorithm usefully indicates the effect of individual data points on $\hat{\theta}$. The components of $\hat{\iota}$ also give the standard error estimate $\hat{\sigma}_{\text{delta}}$, (3.2). However the nonparametric jackknife estimate $\hat{\sigma}_{\text{jack}}$ (3.17) is preferred to $\hat{\sigma}_{\text{delta}}$ as is $\hat{\sigma}_{\text{boot}}$ and $\hat{\sigma}_{\text{ABC}}$.

**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 (2.17) for the constant $\tilde{a}$ used in the $BC_a$ method (though using $\tilde{a}$ based on (2.17) seems to give reasonable results).

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

**Multiple Imputation Bootstrap.** The basic data augmentation identity (4.1) 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 (5.3), (5.5) 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 (4.10) or (4.11) are reasonable surrogates for the predictive density. However the maximum eigenvalue example suggests that (5.3), (5.5) 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, 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 maximum likelihood estimates in missing-data situations. Results like (4.16) give a quantitative 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

The program abcnon, written in the language S of Becker, Chambers and Wilks (1988), evaluates the ABC intervals described in Section 3; tt(P) is the resampling function T(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
  curv <- bhat/sighat - cq
  z0 <- qnorm(2 * pnorm(a) * pnorm(-curv))
  #calculate interval endpoints..........................
  Z <- z0 + qnorm(alpha)
  za <- Z/(1 - a * Z)^2
  stan <- t0 + sighat * qnorm(alpha)
  abc <- seq(alpha)
  for(i in seq(alpha)) abc[i] <- tt(P0 + za[i] * delta)
  lims <- cbind(alpha, abc, stan)
  #output in list form.................................
  list(lims=lims, stats=c(t0, sighat, bhat), cons=c(c(a, z0, cq)), t.=t.)
}

References


