INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

A GIBBS SAMPLER APPROACH TO ESTIMATION OF A MULTIPLE CORRELATION COEFFICIENT IN THE PRESENCE OF MISSING DATA

by

Theresa E. Perlis

A dissertation submitted to the Graduate Faculty in Educational Psychology in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York.

1996

This manuscript has been read and accepted for the Graduate Faculty in Educational Psychology in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

6/3/96
Date

Chair of Examining Committee
Alan L. Gross

6/3/96
Date

Executive Officer
Alan L. Gross

Roger Millsap

David Rindskopf

Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK

Abstract

A GIBBS SAMPLER APPROACH
TO ESTIMATION OF A MULTIPLE CORRELATION COEFFICIENT
IN THE PRESENCE OF MISSING DATA

by

Theresa E. Perlis

Adviser: Professor Alan Gross

A Bayesian approach to estimation of the multiple correlation coefficient when the data are partially missing is theoretically preferable to the commonly used ad hoc procedures or maximum likelihood estimation. However, in the presence of multiple missing data patterns, derivation of the marginal posterior density in closed form becomes analytically intractable, thus Monte Carlo methods must be used instead. The Gibbs sampler is an iterative simulation technique which avoids both complicated analytic derivation and numerical high-dimensional integration methods by generating random variables from a marginal distribution indirectly, without explicit calculation of the density. This paper demonstrates the use of the Gibbs sampler to approximate the marginal distribution of the multiple correlation coefficient conditional on the observed data, when the data are missing at random.

An executable computer program (written in FORTRAN) was developed for the PC to implement the technique for data sets containing up to 200 cases with 2 to 10 variables possessing any missing data structure (subject to certain minimal restrictions). These limits can be increased by

varying the parameters at the compilation stage of the FORTRAN program.

A simple example illustrating the steps in the estimation process performed by the Gibbs sampler is presented. The procedure is carried out on simulated and real data to yield point estimates and interval estimates in terms of highest density regions. Results are compared for alternative trials employing varying Gibbs sequence lengths and different sample sizes for simulation of the marginal distribution of the multiple correlation coefficient. Estimates on all data sets are compared with those obtained from use of a standard statistical computer package (SPSS). Interval estimate for the real data sets are compared with exact Bayesian intervals obtained analytically.

ACKNOWLEDGMENTS

I am extremely grateful to my advisor, Alan Gross, without whose immense help and exceptional patience I would not have been able to complete this dissertation. My thanks also go to the other members of my committee, Roger Millsap and David Rindskopf for their many constructive suggestions, and to Carole Kehr Tittle and Philip Ramsey who graciously agreed to read my dissertation.

Above all a huge thank you to my wonderful daughter Chloe, for your support and understanding throughout.

TABLE OF CONTENTS

LIST OF TABLES

Chapter I
INTRODUCTION

In multiple regression analysis, estimation of the multiple correlation coefficient becomes problematical when some cases in the data set have missing data points on the dependent and/or independent variables. Consider, for example, a study predicting college performance as a function of a number of background variables including SAT scores. For colleges which do not require SAT scores, these data may be missing for some students. Customary approaches to handling such "incomplete data sets" have relied mainly on ad hoc procedures to impute (i.e., fill in) missing values, or on model-based derivation of maximum likelihood estimates of the parameters of interest. A better, and more general, procedure stems from the Bayesian perspective which does not require the assumption of asymptotic normality underlying maximum likelihood estimation or the introduction of ad hoc procedures which are typically not statistically justifiable. The Bayesian approach attempts to derive the marginal density of a parameter given the observed data, and hence to construct an interval estimate. For example, when the distribution is bivariate normal with full response the marginal posterior distribution of the correlation coefficient can be calculated by integrating over the joint posterior distribution of the parameters expressed in terms of the joint prior density and the likelihood (Lee, 1989, pp. 169-171). Notwithstanding, for any data set with incomplete response, derivation of a posterior distribution becomes more complicated since all variables subject to missingness must be treated as additional random variables. In the bivariate normal situation with values missing on one of the variables,

Gross (1996) has derived an expression for the posterior distribution of the correlation coefficient. However, in a multivariate situation with multiple missing data patterns, the consequent increased dimension of the problem renders derivation of the posterior marginal densities in closed form analytically intractable. This paper examines an alternative procedure involving iterative numerical integration based on the Gibbs sampler (Gelfand & Smith, 1990; Geman & Geman, 1984) to approximate the marginal distribution of the multiple correlation coefficient conditional on the observed data, without explicit calculation of the density.

Standard statistical software packages typically circumvent the missing data problem by using "complete case" or "available case" analysis. For example, using the former approach the SPSS procedure for computing the sample correlation matrix excludes all cases with missing values on any of the variables included in the correlation matrix (termed "listwise deletion"). This choice could result in the inclusion of very few cases if many cases have missing data on some variables. Thus an alternative option in the same SPSS procedure ("pairwise deletion") permits the latter approach in which calculation of the correlation coefficient between any particular pair of variables is based on all cases with complete information for the two variables regardless of whether the cases have missing data on any other variables(s). More generally, with additional resources, various imputation techniques are available to fill in missing values to produce a "completed data set" before employing standard analysis methods (Little & Rubin, 1987, chap. 3). However, the drawback to using complete/available case analysis or ad hoc procedures such as mean imputation is that those procedures work only under limited conditions, or simply do not work at all. Unless there is a priori evidence

that the data distributions for respondents and nonrespondents are identical in the population and that the probability of response is independent of both fully and partially observed variables, the fully observed cases cannot be assumed to be a truly random subset of the original full data set. Hence the above-mentioned procedures may work in some situations for certain parameters but not others. Where systematic differences exist between respondents and non-respondents, inference may be biased with the problem exacerbated in smaller subgroups.

The maximum likelihood (ML) approach to parameter estimation initially postulates a general form for the population distribution, then uses the mode of the likelihood as the estimator. In other words, the parameter estimate is the value most likely to have generated the observed data under the proposed model. Under broad conditions a ML estimator is consistent and efficient, and distributed asymptotically normally (Kendall, 1991, p. 653), hence the asymptotic standard error of the ML estimate may be calculated from the information matrix to yield a confidence interval. The same estimation process can be applied to incomplete data, being formally no different from estimation for complete data; however it may be impossible to find a closed form solution to the likelihood equation, necessitating the use of iterative techniques instead. One of the most popular iterative methods is the Expectation-Maximization (EM) algorithm which can be employed in solutions to a broad range of problems. (More detail concerning the EM algorithm technique is provided in the literature survey). Nonetheless, despite the general appeal of ML estimation, it must be noted that the standard error obtained from the information matrix may not be valid in the presence of missing data, because the observed data are not generally distributed iid and the assumption of asymptotic normality

for maximum likelihood estimates may no longer be tenable (Little & Rubin, 1987, p. 88). Moreover the standard errors may be substantially underestimated in small samples.

The Bayesian approach to estimation provides exact results regardless of sample size and is preferable to ML estimation in any event, since inference does not depend on any underlying asymptotic theory. Moreover, the entire posterior distribution may be obtained instead of merely an interval estimate for the parameter. Another advantage is that prior information may be incorporated into the estimation process to increase precision. For small samples, where the information supplied solely by the data is even further limited by missing values, prior knowledge may prove particularly helpful. Notwithstanding the apparent simplicity of the Bayesian approach, for incomplete data the construction of a posterior distribution can involve complicated multiple integration which is too difficult to carry out analytically, thus Monte Carlo techniques must be utilized instead. However, even numerical integration becomes overly complex if the task involves generation of a sample from a distribution which has an unrecognizable form. The Gibbs sampler is one of a variety of numerical integration methods which can be implemented with the additional feature of "data augmentation" in order to simplify the analytic forms of distributions involved. Data augmentation in this context refers to the process of imputing (filling in) the missing data with predicted values based on current knowledge concerning the data structure. The algorithm specified by the Gibbs sampler, when carried out iteratively under certain regularity conditions permits simulation of a posterior marginal distribution (Casella & George, 1992; Gelfand & Smith, 1990). The technique is particularly useful for the multivariate situation with

missing data on many of the variables because, as noted above, these variables must be treated as unknown quantities along with the parameters of the distribution; as the dimension of the problem increases, calculation of the multiple integrals analytically or numerically becomes prohibitively difficult.

In the multiple regression problem under discussion, since the multiple correlation coefficient (ρ) is functionally related to the variance-covariance matrix (Σ), construction of the marginal posterior distribution of Σ suffices for estimation of ρ. Let $\underline{\mu}$ be the vector of means, and let the set of observed data be denoted by **O** and the missing data by **M**. Theoretically the marginal posterior distribution for Σ could be derived from the joint posterior distribution for both $\underline{\mu}$ and Σ:

$$p''(\Sigma \mid \mathbf{O}) = \int p''(\underline{\mu},\Sigma \mid \mathbf{O})\, d\underline{\mu} \qquad (1.1)$$

Similarly, the joint posterior could be derived as follows:

$$\begin{aligned} p''(\underline{\mu},\Sigma \mid \mathbf{O}) &= \int p(\underline{\mu},\Sigma,\mathbf{M} \mid \mathbf{O})\, d\mathbf{M} \\ &= \int p''(\underline{\mu},\Sigma \mid \mathbf{M},\mathbf{O})\, p(\mathbf{M} \mid \mathbf{O})\, d\mathbf{M} \end{aligned} \qquad (1.2)$$

where $p(\mathbf{M} \mid \mathbf{O})$ denotes the predictive density of the missing data given the observed data, and $p''(\underline{\mu},\Sigma \mid \mathbf{O},\mathbf{M})$ denotes the augmented-data posterior density of the population parameters using imputed values for the missing data. Conceptually, in simple terms the joint posterior density can be thought of as an average of augmented-data posterior densities over sets of imputed values generated from the predictive distribution of the missing data. However, the form of augmented posterior density in (1.2) is exceedingly complex in the presence of multiple missing data patterns, frequently rendering analytic derivation of the marginal posterior distribution of Σ impossible. On the other hand, the Gibbs approach is

very simple to carry out. The equation in (1.1) can be expressed alternatively as:

$$p''(\Sigma \mid \mathbf{O}) = \int p(\underline{\mu},\Sigma,\mathbf{M} \mid \mathbf{O})\, d\mathbf{M}\, d\underline{\mu} \qquad (1.4)$$

The Gibbs sampler produces a sample from the posterior distribution through a sequence of simple one-dimensional random variate generations by exploiting the fact that knowledge of the distribution of each variable in turn, conditional on all others in the set of random variables, is sufficient to determine the full joint density uniquely (Besag, 1974), and hence the marginal distributions.

The procedure is demonstrated on simulated and real data to yield interval estimates of the multiple correlation coefficient in terms of highest density regions. Results are compared across separate trials with varying specifications for the Gibbs sampling process. Gibbs intervals for the simulated data are evaluated in terms of coverage of the value of the population parameter. For the real data sets, the utility of the technique is examined by comparing the Gibbs interval estimates with estimates obtained analytically. Additionally, since point estimates are so commonly encountered in real life, posterior means, medians, and modes are calculated for all data sets to further demonstrate the loss of information suffered when intervals are not obtained. Gibbs estimates are also compared with point estimates computed by the SPSS correlation analysis procedure using listwise and pairwise deletion, and mean-substitution. For all data sets the performance of the Gibbs sampler is investigated for various Gibbs sample sizes to determine the number of iterations required to provide interval estimates of the multiple correlation coefficient to increasing degrees of accuracy.

Chapter II
REVIEW OF THE LITERATURE

INTRODUCTION

Until the early 1970's, attempts at valid inference in the presence of missing data usually involved ad hoc methods only. Then, in the first of a series of articles dealing with missing data problems, Rubin (1976) identified the conditions under which complete-data inference could be used in the presence of missing data. Attention was thus turned to developing more rigorous missing-data techniques. Using a likelihood-based approach, the adaptation of ML methods was explored, with a key problem being the development of algorithms to arrive at solutions. More recent approaches have involved the application of Bayesian methods and estimation of the entire posterior distribution instead of merely a point estimate of the unknown parameter. However, again due to the complexities involved, algorithm development was needed, leading to widespread utilization of methods employing iterative and/or Monte Carlo techniques. In particular, various iterative procedures have been developed which involve sampling from a sequence of conditional distributions, with much recent work devoted to the Gibbs sampler and its effectiveness in providing Bayesian approximations to models for which solutions had previously proved intractable (Gelfand & Smith, 1990).

This literature review is presented in six different sections:

1) Types of missing data processes
2) Ad hoc methods and their limitations
3) The likelihood approach and the EM algorithm
4) Bayesian techniques

5) The Gibbs Sampler

6) Practical applications of the Gibbs Sampler

1) TYPES OF MISSING DATA PROCESSES

The choice of techniques for handling missing data depends upon the underlying mechanism giving rise to missingness, since various assumptions must be made implicitly or explicitly. A typology of missing data mechanisms was created by Rubin (1976). When the probability of response is known to be independent of both fully and partially observed variables, the data are said to be "missing completely at random" (MCAR). Hence analysis can be carried out on the fully observed cases alone, since these form a random sample of the whole data set. A more common occurrence is "missing at random" (MAR), in which the probability of response depends on one or more fully observed variables, but is independent of the variables subject to nonresponse. Missing values are not randomly distributed across the whole data set, but occur randomly within subclasses defined by values of the fully observed variable(s). The most restrictive situation occurs where the probability of response is dependent on the unobserved value(s) (and possibly also the observed values). In this last situation, referred to as "not missing at random" (NMAR), the missing data mechanism is considered nonignorable.

For example, suppose in each of the following three situations we are attempting to use GRE scores to predict academic achievement as measured by overall GPA where GRE scores are available for all students but GPA is missing for some. a) Suppose a college admits all applicants regardless of GRE performance. A random sample of students is then selected and their GPAs recorded. The data are missing MCAR because the probability that

GPA is recorded is independent of GRE score and GPA. b) Suppose instead that the college selects applicants on the basis of high GRE scores. GPA scores are then recorded for all students. GPA will be missing for those students whose GRE was below the cutoff level and who consequently were not admitted to the program. The probability that GPA is recorded thus depends on the GRE score, hence data are missing MAR. c) Finally suppose that the college selects on the basis of high GRE scores, but some applicants who perform highly during the first year transfer to another school at the end of the first year. GPA is recorded for all remaining students. In this case the probability that GPA is recorded depends both on GRE score and on unmeasured GPA. The data are missing NMAR and the missing data mechanism is nonignorable.

Since missing data can be nonignorable in a myriad of ways, the missingness mechanism must be included in the modeling process. Inferences therefore depend on relationships about which the observed data provide no direct information whatsoever (i.e., the distribution of the response variables that are unobserved).

2) AD HOC METHODS AND THEIR LIMITATIONS

Ideally, appropriate handling of missing data should adjust for differences between complete and incomplete cases on fully observed variables in the estimation process, adjust standard errors of estimates to reflect both these differences and the reduced sample sizes resulting from nonresponse, and provide a measure of the sensitivity of inference to possible alternative models for nonresponse (Rubin, 1987, p.11). In order to be able to use standard complete-data methods of analysis, there are three general ad hoc approaches commonly taken: complete-case analysis

(listwise), available-case analysis (pairwise), and imputation procedures. However, none of these proves entirely adequate.

In complete-case analyses, cases that have missing data on one or more variables are excluded from all univariate and multivariate analyses. Thus only a subset of the original data set is analyzed. Where missingness is not MCAR, complete-case analysis can lead to serious biases depending on the underlying mechanism of missingness and the inference involved. In particular, even under MAR, estimates of the multiple correlation will be biased (Little & Rubin, 1987, p.41). Moreover, although implementation of the method is simple, the decrease in sample size may be substantial, especially where multiple variables have missing values. Consequently standard errors will be inflated, even when estimates are unbiased.

In available case analysis, cases are included in a specific analysis as long as they have complete response on the variables used in the current analysis. Thus different sets of analyses may be carried out on different subsets of data. Analytic results will be comparable and unbiased only if the data are missing MCAR. Typically, using available case analysis, covariances and correlations are derived using pairwise deletion - i.e., all cases with complete data on both the variables being correlated are used. However this approach can lead to covariance matrices that are not positive definite even under MCAR.

Since dropping cases with incomplete response forfeits information, an attractive alternative is to "complete" the data by imputing a value for each missing response. Each missing value is replaced with a real value, for example by filling in the mean of the observed values for a variable subject to missingness, or by imputing a value predicted by modeling the

partially observed variable given the fully observed variables using the observed data. The augmented data set can then be analyzed by standard complete-data methods. Valid imputation procedures should maintain the external relationships between the sample cases for a given variable, and the internal relationships between the variables for a given case (Ford, 1983). Moreover, imputed samples should be representative of the population covariance structure. Unfortunately, in practice, use of imputation procedures frequently relies more on readily available empirical evidence of validity than on a sound theoretical basis and there are many drawbacks to most of the techniques.

Mean-value imputation, which substitutes the mean of observed values for each variable in turn, leads to systematic underestimation of variances and covariances even under MCAR (Little & Rubin, 1987, p. 44), although the consequent attenuation of the correlation coefficient is relatively less, due to cancellation between the attenuations of the covariance and standard deviations (Kalton & Kasprzyk, 1982). An alternative mean-substitution method proposed by Buck (1960), imputes values on a case by case basis by regressing "missing" on "observed" variables for each pattern of missing data using estimates of the mean and covariance matrices from the fully observed data. Underestimation of variances and covariances are less than under the previous unconditional mean-substitution, provided the normality assumptions needed for regression hold, and MCAR is assumed. However, generally any form of mean imputation distorts the marginal distributions of the completed data. In particular, standard complete-data inference does not yield consistent estimates for parameters which are non-linear in the data, such as the correlation coefficient (Little & Rubin, 1987).

Other imputation methods, particularly the various hot-deck techniques, are typically used with survey data. Hot-deck methods involve the replacement of missing values by values from a subset of the observed values, chosen typically by various forms of matching between respondents and nonrespondents on observed characteristics. Some hot-deck methods are deterministic in that the choice of imputation value is determined solely by the matching rule. Others use a randomization process to select one of a group of candidate values proposed by the matching rule. In general, deterministic approaches tend to distort the distributions and attenuate variances of variables, whereas randomization approaches preserve the variability in the observed data but introduce imputation variance (Kalton & Kasprzyk, 1982). Covariances between variables subject to imputation and other variables are likely to be underestimated by all methods, hence leading to attenuation of bivariate or multiple correlations On the other hand it is possible that deterministic imputation methods employing auxiliary information may overestimate correlations, even under MCAR.

A further disadvantage is that most imputation methods do not allow for accurate assessment of the standard error of estimate using standard complete-data formulas (Ford, 1983). In general, the use of complete-data methods with imputed data sets treats missing values as if they are observed, thereby overstating the precision of the data (Rubin, 1987, p.12). Single imputations may yield reasonable parameter estimates with standard calculations, but the accompanying measures of precision do not convey the added uncertainty of having imputed rather than observed the missing values; nor do they indicate the sensitivity of inferences to the imputation scheme. More accurate variance approximations for imputed data sets can

be obtained by complicated methods such as balanced repeated replications (Rizvi, 1983), but these are rarely feasible to carry out.

Multiple imputation (Rubin, 1987) is an attempt to remedy the deficiencies of single imputation, by creating two or more plausible completed data sets, each with an independent draw from p(M | O), the predictive distribution of the missing data given the observed data. Inference that reflects sampling variability in the distribution of imputation values can be derived by combining results of separate complete-data inferences for each imputation set. This increases efficiency of estimation and allows for more accurate estimation of standard error. Although formation of a multiple imputation procedure can involve extremely complicated modeling, there are many instances in which a more feasible approach is merely to modify one of the commonly used single imputation methods. In any event, the additional benefit derived from using multiple instead of single imputation is dependent on the adequacy of the specific imputation technique employed. The increased amount or requisite data storage space and the extra work involved in analysing multiple data sets is also a major consideration.

3) THE LIKELIHOOD APPROACH AND THE EM ALGORITHM

Maximum likelihood methods of estimation adapted to missing data situations avoid problems associated with ad hoc methods and provide a more systematic approach. In "large" samples a ML estimator is approximately normal and is unbiased, hence standard methods of inference can be used to obtain interval estimates of the unknown parameter. For complete data the ML estimate of a distributional

parameter is the value for the parameter that maximizes the likelihood function $L(\theta \mid O)$ of the parameter (or equivalently the loglikelihood) given the observed data. Provided the mechanism leading to missing data can be ignored, the likelihood function takes the same form in the presence of missing data; in other words the likelihood is derived from only the observed portion of the full data set. The underlying mechanism leading to nonresponse can be ignored for likelihood-based inference when the missing data process is MCAR or MAR (Little & Rubin, 1987, p. 15) and the parameters underlying the data and response models are distinct. However for the NMAR situation the likelihood expression must incorporate a term representing the probability of response. The Heckman model (Heckman, 1976, 1979) is an example of NMAR in regression analyses, where the dependent variable has missing values which are missing as a probabilistic function of the dependent variable. A similar example is provided by the Tobit model (Amemiya, 1985) in which the dependent variable is subject to missingness and also censored (i.e., the variable y is observed only if $y \leq c$, otherwise y is recorded as c). The difficulty in working with NMAR data arises from the fact that although a model must be postulated, it is not possible to obtain the unobserved values in the data set to determine whether the model is realistic.

Moreover, even when the missing data mechanism is ignorable, the form of the observed data likelihood is nonstandard thus maximization can be difficult to accomplish. For example, in a bivariate normal problem the likelihood splits into three parts, and for the multivariate situation the function becomes correspondingly more complex. In response to the need for an algorithm to simplify the derivation process in a variety of missing data situations, the Expectation-Maximization (EM) algorithm was

presented by Dempster, Laird, and Rubin (1977). The EM algorithm is a very general iterative two-step technique to determine the ML estimate by quasi-augmentation of the observed data so that estimation can be based on the complete-data likelihood instead of the observed-data likelihood. Conceptually, the E step consists of finding the conditional expectation of the theoretical complete data likelihood given the observed data and an initial guess for the unknown parameter. The expectation of the likelihood is thus calculated with respect to the missing data. In the specific situation where the likelihood is linear in the missing data, direct substitution of expected values for missing values yields the expected complete-data loglikelihood. Otherwise, merely replacing the missing values by estimated values may lead to biased results (Little & Rubin, 1987). More generally, where the likelihood is linear in the sufficient statistics, the expected complete-data loglikelihood is obtained by computation of the expected values of the sufficient statistics in the E-step. Maximization of the expected complete-data loglikelihood constitutes the M step and the resulting updated parameter estimate is then used in the next E step. The algorithm is iterated until convergence. Dempster, Laird, and Rubin showed that under general conditions successive iterations always increase the likelihood, and that the sequence converges reliably at a linear rate to a local maximum or saddle point, although the convergence point may not be the global maximum if the loglikelihood possesses multiple maxima and/or saddle points. However, Wu (1983) subsequently proved that although any EM sequence converges to a stationary point, convergence to a local or global maximum likelihood estimate depends upon the choice of the initial value for the parameter, and is not guaranteed to take place. The limit point is the unique maximizer of the likelihood function only when the

likelihood function is unimodal and specific differentiability conditions exist.

Despite simplicity of implementation for the EM algorithm, there is currently very little software available to perform the technique (the BMDPAM program in Dixon, 1985, and LIMDEP are the two most commonly used), and considerable disadvantages exist. The rate of convergence may be exceedingly slow when a large proportion of the data is missing. Moreover, standard errors based on the observed or expected information matrix have to be derived separately after the last iteration, since the EM algorithm does not calculate either matrix. The major shortcoming of the approach however stems from the assumption, based on large samples with no missing data, that the ML estimator is approximately normally distributed. Since a ML estimate is a function of the observed values only, the presence of missing data diminishes the sample size. Unfortunately, it cannot be assumed that estimates are either unbiased or normally distributed in small samples, and estimated standard errors could be quite inaccurate. In particular, under NMAR, as sample size decreases and the proportion of missing data increases, the size of standard errors may rise beyond acceptable limits (Gross, 1990). Thus, although the EM algorithm may work well for the multivariate normal, for other distributions the large sample approximation can prove inadequate. In general, the use of the EM algorithm to generate an estimate of the multiple correlation coefficient is justifiable only in large samples with small amounts of missing data. Moreover, in practice the EM algorithm is generally most appropriate for situations where both the derivation of the conditional expectation of the complete data loglikelihood and the maximization process are straightforward.

4) BAYESIAN TECHNIQUES

An advantage to the Bayesian approach to estimation is that additional information is provided by specification of a prior probability distribution, whereas ML estimation is restricted to using information contained in the sample. Moreover, since Bayesian inference is not based on the asymptotic assumptions of the likelihood approach, it provides a viable alternative in situations where the likelihood is not closely approximated by the normal likelihood (particularly true in small samples). Using Bayesian analysis, one may attempt to derive the entire posterior distribution, and hence interval estimates for the parameters, or alternatively to obtain a point estimate by maximizing the posterior distribution. Frequently, however, neither approach proves analytically simple.

In general, with fully observed data, if we denote the population parameters by Θ, the posterior distribution of Θ is given by

$$p''(\Theta \mid O) = p'(\Theta)\, p(O \mid \Theta) / p(O)$$

where O denotes the observed data, $p''(\Theta \mid O)$ denotes the posterior density of the population parameter(s) given the data and $p'(\Theta)$ denotes the prior distribution of the parameter(s) (Lee, 1989). For many standard situations with no missing data the form of posterior is known, thus analytic derivation is not difficult. However, in the presence of missing data, the posterior distribution must be obtained by the integration

$$p''(\Theta \mid O) = \int p(\Theta'' \mid O,M)\, p(M \mid O)\, dM$$

which is tantamount to averaging the complete-data posterior over all possible values for the missing data. In certain simple missing-data situations an analytic solution is feasible. This has been done, for example, in estimation of the bivariate correlation coefficient with data missing on

one of the variables (Torres-Quevedo, 1993), also with data missing on both variables (Gross, 1996). Nonetheless, beyond the bivariate case, analytic derivation becomes intractable. When difficulty in partitioning the complete-data posterior or in evaluating the integral is a major impediment to derivation, solutions must be sought using analytical approximations or numerical integration techniques.

For example, minor modifications to the EM algorithm enable it to be used in a Bayesian framework to yield the mode of the posterior distribution of the parameter(s) in place of the ML estimate (Dempster et al., 1977). However, in many situations maximization of the posterior is too complicated to perform analytically, thus Monte Carlo methods must be used instead.

Monte Carlo EM Algorithm (MCEM)

Wei and Tanner (1990b) propose a simulation approach to expand the range of application of the EM algorithm in the Bayesian context when the complexity of the posterior density prohibits maximization. In this Monte Carlo EM algorithm (MCEM), given the current guess to the maximizer $\Theta^{(i)}$ of the observed posterior, $p''(\Theta \mid O)$, the E step is implemented by randomly generating a sample of k sets of "missing values" from the conditional predictive distribution $p(M \mid \Theta^{(i)},O)$ (the distribution of the missing data given the current parameter estimate and the observed data) to create an augmented data set. The log-posterior is then updated as a mixture of augmented log-posteriors, mixed over the k imputations:

$$\log p''(\Theta \mid O) = \left[\sum_{j=1}^{k} \log p''(\Theta \mid O,M_j)\right] / k$$

The M step consists of maximizing the mixture to yield an updated guess $\Theta^{(i+1)}$ to the mode of the observed posterior before a new iteration of the

algorithm begins. In the particular instance where k is equal to one, and a representative summary measure such as the mode or expected value of the conditional predictive distribution is chosen to be the imputed value instead of a random deviate, the MCEM reduces to a classic EM-type algorithm. Convergence of the algorithm can be monitored by plotting values of $\Theta^{(i)}$ versus iteration number.

Data Augmentation Algorithm

Various iteration procedures have been developed which involve sampling from a sequence of conditional distributions. A method termed the Data Augmentation Algorithm was proposed by Tanner and Wong (1987) to simulate the entire posterior distribution instead of just the mode. This procedure combines aspects of multiple imputation and the EM algorithm without the requirement for a large sample size. As noted above, the desired posterior distribution is given by

$$p''(\Theta \mid O) = \int p''(\Theta \mid O,M)\, p(M \mid O)\, dM \qquad (1)$$

where $p(\Theta'' \mid O,M)$ represents the complete-data posterior distribution, and $p(M \mid O)$ represents the predictive density of the missing data given the observed data. In turn, the predictive density of the missing data can be related to the posterior by

$$p(M \mid O) = \int p(M \mid \Theta,O)\, p''(\Theta \mid O)\, d\Theta \qquad (2)$$

where $p(M \mid \Theta,O)$ represents the conditional predictive distribution of the missing data given the parameter value(s) and the observed data. Substituting the right hand side of equation (2) into equation (1) yields

$$p''(\Theta \mid O) = \int \left[\int p''(\Theta \mid O,M)\, p(M \mid \Theta,O)\, dM\right] p''(\Theta \mid O)\, d\Theta \qquad (3)$$

It can be seen that (3) has the form of a fixed point equation with the posterior density appearing on both sides of the equation The justification for practical implementation of the Data Augmentation algorithm using

iteration is based on its similarity to the method of successive substitution for solving fixed point operator equations (Rall, 1969). Each iteration comprises two steps. At the inception of the iterative process, an initial estimation for $p''(\Theta \mid O)$ must be provided. Applying the method of composition to equation (2) a sample of k sets of "missing values" can be generated from the predictive density $p(M \mid O)$ by first generating a value $\Theta^{(i)}$ from $p''(\Theta \mid O)$, then generating a sample $M_1, M_2, M_3, ..., M_k$ from $p(M \mid \Theta^{(i)},O)$. This is called the "imputation" step. These values of M are used to create k augmented data sets. Next, in the "posterior" step, the current approximation to $p''(\Theta \mid O)$ is updated to be the mixture of augmented posteriors mixed over the k imputed data sets. i.e.,

$$p''(\Theta \mid O) = [\sum_{j=1}^{k} p''(\Theta \mid O,M_j)] / k$$

which is the Monte Carlo equivalent of integrating over the predictive density in equation (1). The process continues through many iterations of these two steps until convergence. The Data Augmentation Algorithm bears a resemblance to the MCEM discussed above, but the former allows for estimation of entire posterior, not merely the mode of the distribution. There is however a drawback to the procedure in that it requires knowledge concerning the form of all conditional densities, or at least the ability to sample numerically from the distributions. As we shall see later, the Gibbs Sampler has less stringent requirements for implementation.

Sampling Importance / Resampling Algorithm (SIR)

The Sampling Importance / Resampling algorithm (Rubin, 1987) is a noniterative algorithm which makes use of multiple imputations, for estimation of the entire posterior. In order to use this one must start with

good approximation to the posterior distribution $p(\Theta,M \mid O)$ of the parameters and missing data given the observed data, say

$$h(\Theta,M \mid O) = h(\Theta \mid O)\, h(M \mid \Theta,O)$$

where $h(\Theta \mid O)$ approximates $p(\Theta \mid O)$,

and $h(M \mid \Theta,O)$ approximates $p(M \mid \Theta,O)$.

Drawing on the fact that $p(\Theta,M \mid O) \propto p(\Theta)\, p(O,M \mid \Theta)$,

importance weights of the following form are proposed:

$$w(\Theta,M \mid O) = p(\Theta)\, p(O,M \mid \Theta) / h(\Theta,M \mid O)$$

Implementation of the algorithm takes place in three steps:

Step 1 Draw k values of (Θ,M) at random from $h(\Theta,M \mid O)$. Call these k values $(\Theta_{(j)},M_{(j)})$ $\quad j = 1,\ldots,k$

Step 2 Calculate importance ratios for each $(\Theta_{(j)},M_{(j)})$ by

$$w_j = w(\Theta_{(j)},M_{(j)} \mid O) = p(\Theta_{(j)})\, p(O,M_{(j)} \mid \Theta_{(j)}) / h(\Theta_{(j)},M_{(j)} \mid O)$$

Step 3 Calculate the posterior density $p(\Theta \mid O)$ from

$$p(\Theta'' \mid O) = \Big[\sum_{j=1}^{k} w_j\, p(\Theta'' \mid O,M)\Big] \Big/ \sum_{j=1}^{k} w_j$$

When the distribution of the missing data given the observed and the parameter values is known, or if one can sample from $p(M \mid \Theta,O)$, then

$$h(M \mid \Theta,O) = p(M \mid \Theta,O)$$

then the importance weights do not depend on M and reduce to

$$w_j = w(\Theta_{(j)} \mid O) = p(\Theta_{(j)})\, p(O \mid \Theta_{(j)}) / h(\Theta_{(j)} \mid O)$$

thus only $p(\Theta \mid O)$ needs to be approximated. Gelfand & Smith (1990) found that the Data Augmentation or Gibbs algorithms make more efficient use of the randomly generated variates than a non-iterative procedure such as SIR. The performance of the latter depends on the specification of $h(\Theta,M \mid O)$. Moreover, in practice it is best used when the fraction of missing information is modest, and a small number of multiple imputations

suffices, since the ratio of number of imputations to number of missing values affects performance (Tanner, 1992).

Poor Man's Data Augmentation Algorithms (PMDA)

The Poor Man's Data Augmentation Algorithms (PMDA) (Wei & Tanner, 1990b) are two modifications to the MCEM which are non-iterative, but which nevertheless permit simulation of the entire posterior distribution. In PMDA #1, having obtained the updated guess to the mode of the observed posterior from MCEM, a sample of k sets of "missing values" is generated from the conditional predictive distribution $p(M \mid \Theta^{(i)},O)$, then the observed posterior is approximated by the mixture of augmented posteriors, mixed over the k imputed data sets

$$p''(\Theta \mid O) = [\sum_{j=1}^{k} p''(\Theta \mid O,M_j)] / k$$

The approximation arises from the use of the conditional predictive rather than the predictive distribution $p(M \mid O)$ to generate values for imputation.

In situations where $p(M \mid O)$ can be easily evaluated, PMDA #2 calculates the observed posterior using the technique of importance sampling (Smith & Gelfand, 1992; Tanner, 1992). As in the first PMDA, using the updated guess $\Theta^{(i)}$ from MCEM, a sample of k sets of "missing values" is first generated from the conditional predictive distribution $p(M \mid \Theta^{(i)},O)$. Then weights

$$w_j = p(M_j \mid O) \; / \; p(M_j \mid \Theta^{(i)},O)$$

are assigned to the imputations to give an approximation to the observed posterior

$$p(\Theta'' \mid O) = [\sum_{j=1}^{k} w_j \, p(\Theta'' \mid O,M)] \; / \; \sum_{j=1}^{k} w_j$$

Since both of these algorithms are approximations, Wei and Tanner suggest each may serve best as starting point to one of the Data Augmentation, SIR, or Gibbs Sampler algorithms .

5) THE GIBBS SAMPLER

The Gibbs Sampler had its origins in an algorithm developed by Metropolis et al. (1953) to simplify computational aspects of sampling in an application in statistical physics. A generalization of the Metropolis algorithm, in which a sequence of samples from a statistical distribution is generated by Monte Carlo simulation of Markov chains, was presented by Hastings (1970), with specific application to the Poisson and normal distributions. (A sequence $\mathbf{x}_n$ of random variables is called Markov if for any n we have

$$p(x_n \mid x_{n-1}, x_{n-2}, \ldots, x_1) = p(x_n \mid x_{n-1}) \qquad \text{(Papoulis, 1965, p.529)}$$

In simple terms, at any stage in the sequence, future values of the random variable do not depend on values obtained prior to the current value.) The Gibbs Sampler algorithm was formally introduced in the Bayesian framework by Geman and Geman (1984) as an iterative algorithm, based only on elementary properties of Markov chains, to extract a marginal distribution from a set of conditional distributions without having to directly calculate the marginal density. Earlier work by Besag (1974) had proved that given a set of q variables for which every univariate conditional distribution (i.e., where <u>one</u> of the set of variables is conditioned on <u>all</u> of the remaining variables) is known, the joint density (and consequently the marginals) of the full set of variables is uniquely determined, provided that the joint density function is positive over the entire sample space. Geman and Geman thus proposed generation of a

"Gibbs sequence" of random variates by sampling from each of the conditional distributions in turn.

For example, assume

a) the joint distribution p(X,Y,Z) exists

b) the univariate conditional distributions

$$p(X \mid Y,Z) \qquad p(Y \mid Z,X) \qquad p(Z \mid X,Y)$$

are known.

Initial values $X^{(0)} = x^{(0)}$, and $Y^{(0)} = y^{(0)}$ are chosen.

First a value $Z^{(0)}$ is sampled from the distribution $p(Z \mid X=x^{(0)}, Y=y^{(0)})$.

Next a value $X^{(1)}$ is sampled from the distribution $p(X \mid Y^{(0)}=y^{(0)}, Z^{(0)}=z^{(0)})$.

Then a value $Y^{(1)}$ is sampled from the distribution $p(Y \mid Z^{(0)}=z^{(0)}, X^{(1)}=x^{(1)})$.

This process is iterated to yield a Gibbs sequence of random variates

$$X^{(0)}, Y^{(0)}, Z^{(0)}, X^{(1)}, Y^{(1)}, Z^{(1)}, ..., X^{(j)}, Y^{(j)}, Z^{(j)}, ..., X^{(k)}, Y^{(k)}, Z^{(k)}$$

The successive values of X (or Y or Z) represent a Markov chain, and for any starting value the distribution of $X^{(k)}$ tends to the true marginal distribution p(X) as $k \to \infty$ (Casella & George, 1992; Tanner, 1992). The process can be extended to any number of random variables. Moreover, Smith and Roberts (1993) demonstrate that the expected value of any function of interest can be estimated over realizations from a single run of the chain, since

$$\left[\sum_{j=1}^{k} f(X^{(j)})\right] / k \quad \xrightarrow[k\to\infty]{} \quad E\{f(X^{(j)})\} \quad \text{almost surely.}$$

It should be noted that the Data Augmentation algorithm is closely related to the Gibbs sampler, and both are practical when conditional distributions are available. The two algorithms are actually identical for the situation with two random variables. For q (>2) variables, the Gibbs requires all q univariate conditional distributions (i.e., one variable

conditioned on all the remaining variables) to be known, whereas Data Augmentation needs q(q-1) conditionals including all univariate conditionals. Both make use of imputation or augmentation of the data in order to be able to use the complete-data posterior distribution to facilitate the estimation process in a manner similar to the EM algorithm.

There has been much discussion concerning the desirability of basing inference on one long run of the chain generated by the Gibbs sampler or on multiple shorter runs. Geyer (1992) makes the case that if the runs are long enough, one long run is sufficient, whereas runs that are too short cannot anyway be used for valid inference. An advantage to using multiple runs is that discordant results provide direct evidence that the runs are too short, however concordance does not necessarily imply validity of results. Since successive $X^{(j)}$ are clearly correlated, Smith and Roberts (1993) suggest that, to approximate a random sample from the marginal distribution, realizations of $X^{(j)}$ can be collected from one long run at gaps spaced sufficiently to render serial correlation neglible. Alternatively, parallel independent runs of the chain can be performed and an approximate sample of m variates from p(X) obtained by using the final $X^{(k)}$ from each of the m independent Gibbs sequences of length k. The advantage to the former approach is that it involves only one initial transient phase in which the realizations of $X^{(j)}$ may remain heavily influenced by the starting value for many iterations. On the other hand, Gelman & Rubin (1992) emphasize the utility of multiple independent iterative sequences, and demonstrate how careful selection of a starting distribution for the Gibbs sampler can increase the efficiency of the iterative process. If the starting distribution is far from target the Gibbs sampler can take a long time to move from the region around one mode to

another, and hence may converge slowly. This may be particularly true when examining a lower dimensional summary of a simulated multidimensional random variable, even when the summary's target distribution is univariate and unimodal . The starting distribution is thus chosen to be centered around the modes of the desired posterior distribution, and overdispersed to ensure that the starting distribution covers the target distribution. (This is particularly important when the posterior is complicated enough so that it may be multimodal, although the approach may not work for every problem). By having several sequences use can be made of the variability present in starting distribution. MacEachern & Berliner (1994) however, demonstrate that subsampling from a single run of a long chain yields poorer estimates than using the full chain since stationarity implies that the correlation between two successive outcomes must lie between -.5 and .5. They advocate subsampling only when constrained by computational resources. Whatever the chosen approach, it is necessary to discard the initial outcomes of $X^{(j)}$ occurring in the "burn-in" period until the chain reaches equilibrium. Geyer suggests throwing out 1 or 2% of the runs to start, and inspection of the sample autocorrelation function to determine if more need to be discarded later.

Hastings (1970) points out that since samples are correlated, estimates of standard errors may need manipulation. A straightforward approach to estimation of the Monte Carlo error by invoking the Central Limit Theorem requires estimation of the asymptotic variance which may not necessarily be finite (Geyer, 1992). In his discussion of window estimators, the method of batch means, and specialized Markov chain estimators to provide variance estimation, Geyer points to limitations in all three methods. MacEachern & Berliner (1994) suggest that batching be

used to simplify estimation of the variance, although a single chain should be used to derive the parameter estimate.

Two important considerations regarding implementation are the monitoring convergence of the algorithm and the specification of m when employing parallel independent runs of the chain. Gelman & Rubin (1992) suggest a new approach to monitoring convergence of iterations, by monitoring change in relative size of the between sequence mean square and pooled within sequence mean square. Thus the uncertainty due to finite-length sequences is incorporated into distributional estimates. The independent replications permits estimation of sampling variability of estimators without the necessity to make inference concerning the structure of the time-series of simulations, which is all that it is generally possible to do from a single sequence

6) PRACTICAL APPLICATIONS OF THE GIBBS SAMPLER

The Gibbs sampler can be used in classical inference to calculate the likelihood function (Tanner, 1992), as well as in the Bayesian framework to generate the posterior distribution. Gelfand and Smith (1990) demonstrate the utility of the Gibbs algorithm for a wide range of statistical applications particularly calculation of posterior densities. Gelman & Rubin (1992) fit a multivariate random effects mixture model to psychological data from an experiment measuring reaction times of normal and schizophrenic patients. A model with 22 parameters was created to reflect the psychological theory. All of the univariate conditional posterior distributions have closed form so are easy to sample from. For each scalar estimand, an approximate distribution was created as the starting distribution for application of the Gibbs by locating the modes of the

posterior distributions and using importance resampling. Results show that the mean is close to target mean and the variance greater than the target variance.

Wakefield, Smith, Racine-Poon, and Gelfand (1994) investigate the problem of obtaining approximate posterior distributions for complex linear and non-linear population models with a hierarchical structure. The first example discussed is a linear population biological growth model representing the relationship between dental measurement and age in children, with an emphasis on making inferences concerning the difference in dental growth between boys and girls. The second example concerns computation of the predictive distribution for plasma concentration of the drug Cadralazine in cardiac failure patients by observing plasma concentration at periodic intervals after administration of a single dose. For these problems the Gibbs sampler is employed to facilitate a fully Bayesian analysis incorporating both inference and diagnostic checks regarding population outliers and inappropriate mean-variance relationships.

Chapter III
METHOD

The Gibbs sampler approach to estimation for multivariate data with incomplete observations is an iterative procedure, based on stepwise imputation of missing values, which simulates a marginal posterior distribution from a set of conditional distributions. This paper focuses on estimation of the multiple correlation coefficient (ρ) in a multivariate normal distribution with missing data, where the parameters of the distribution have unknown values. Since there is a one-to-one functional relationship between ρ and the variance-covariance matrix Σ (Winer, 1971, p. 107), ρ can be calculated directly when Σ is known. However in the current situation Σ is unknown. Thus the Gibbs sampler is employed to simulate the marginal posterior distribution for Σ, allowing a point or interval estimate of ρ to be subsequently obtained by calculation.

In a multivariate normal data set with missing values, with unknown population parameters Σ and $\underline{\mu}$ (the vector of means), let the observed data be denoted by **O** and the missing data be denoted by **M**. As discussed in the Chapter I, formula (1.4) the marginal posterior distribution for Σ can theoretically be derived from the joint distribution of $\underline{\mu}$, Σ, and **M** conditional on the observed data:

$$p''(\Sigma \mid \mathbf{O}) = \int p(\underline{\mu},\Sigma,\mathbf{M} \mid \mathbf{O})\, d\mathbf{M}\, d\underline{\mu}$$

(where $\underline{\mu}$, Σ, and **M** are all unknown).

Associated with this joint distribution there are three "univariate" conditional distributions (where the term "univariate" here indicates that one of the set of unknowns is conditioned on the other two unknowns and the observed data):

(a) The predictive distribution of the missing data conditional on current parameter values and the observed data, represented by

$$p(\mathbf{M} \mid \underline{\mu}, \Sigma, O)$$

(b) The posterior distribution of Σ conditional on the current values for both the population mean vector and the missing data, and the observed data, represented by

$$p''(\Sigma \mid \mathbf{M}, \underline{\mu}, \mathbf{O})$$

(c) The posterior distribution of $\underline{\mu}$ conditional on the current values for the population variance-covariance matrix and the missing data, and the observed data, represented by

$$p''(\underline{\mu} \mid \Sigma, \mathbf{M}, \mathbf{O})$$

The Gibbs sampler simulates a sample from $p''(\Sigma \mid \mathbf{O})$, by sampling repeatedly from the three conditional distributions

$$p(\mathbf{M} \mid \underline{\mu}, \Sigma, O) \qquad p''(\Sigma \mid \mathbf{M}, \underline{\mu}, \mathbf{O}) \qquad p''(\underline{\mu} \mid \Sigma, \mathbf{M}, \mathbf{O})$$

in turn, to yield a sequence of random variates

$$\mathbf{M}^{(1)}, \Sigma^{(1)}, \underline{\mu}^{(1)}, \mathbf{M}^{(2)}, \Sigma^{(2)}, \underline{\mu}^{(2)}, \mathbf{M}^{(3)}, \Sigma^{(3)}, \underline{\mu}^{(3)}, \ldots, \mathbf{M}^{(k)}, \Sigma^{(k)}$$

For large k, the final observation in the sequence represents a sample point from the marginal posterior distribution $p''(\Sigma \mid \mathbf{O})$ (Casella & George, 1992), and the sampled value for Σ determines a unique value for ρ. When m replicates of the Gibbs sampler are carried out, m values of ρ corresponding to the final values of Σ from the m individual Gibbs sequences are thus obtained. These m values for ρ represent a sample from the approximate marginal posterior distribution of $p''(\rho \mid \mathbf{O})$.

We start with the assumption that the full data set is multivariate normal, with q variables and n cases. The remainder of the this chapter is structured as follows:

1) The form of the three conditional distributions required by the Gibbs sampler is first discussed.

2) Next the logic and procedural details of the implementation of the iterative procedure and the estimation of ρ are explained.

3) The data sets used to demonstrate the Gibbs process are described.

4) Specifications for the production of alternative Gibbs samples are presented.

5) The process of monitoring convergence of the Gibbs algorithm and evaluating performance is described.

6) Finally a worked example is presented to illustrate the process of Gibbs sampling estimation.

All processing is performed by a FORTRAN program on a 486 PC.

1) FORM OF THE CONDITIONAL DISTRIBUTIONS

(a) $p(\mathbf{M} \mid \underline{\mu}, \Sigma, O)$

In general, for a multivariate normal distribution with parameters $\underline{\mu}$ and Σ, if the random variables $\underline{V}$ are partitioned into <u>any</u> two subsets $\underline{V_1}$ and $\underline{V_2}$, the joint distribution $p(\underline{V_1}, \underline{V_2} \mid \underline{\mu}, \Sigma)$ is normal with mean vector and variance-covariance matrix which can be partitioned as follows:

$$\underline{\mu} = \begin{bmatrix} \underline{\mu_1} \\ \underline{\mu_2} \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

where $\underline{\mu_1}$ and $\underline{\mu_2}$ represents the vectors of means for subsets $\underline{V_1}$ and $\underline{V_2}$ respectively, and Σ_{11} and Σ_{22} represent the variance-covariance matrices for $\underline{V_1}$ and $\underline{V_2}$ respectively.

The conditional distribution for $\underline{V_2}$ given $\underline{V_1}$ (i.e. $\underline{V_1}$ held constant) is also multivariate normal and can be written:

$$p(\underline{V}_2 \mid \underline{V}_1, \underline{\mu}, \Sigma) = N\,[E\,(\underline{V}_2 \mid \underline{V}_1, \underline{\mu}, \Sigma)\,,\ \Sigma_{v_2|v_1}] \tag{3.1}$$

with $$E\,(\underline{V}_2 \mid \underline{V}_1, \underline{\mu}, \Sigma) = \underline{\mu}_2 + B\,(\underline{V}_1 - \underline{\mu}_1) \tag{3.2}$$

and $$\Sigma_{v_2|v_1} = \Sigma_{22} - \Sigma_{21}\,\Sigma_{11}^{-1}\,\Sigma_{12} \tag{3.3}$$

where B is the matrix of regression coefficients for $\underline{V}_2$ regressed on $\underline{V}_1$ given by

$$B = \Sigma_{21}\,\Sigma_{11}^{-1} \tag{3.4}$$

Consider now an incompletely observed sample of data drawn from a multivariate normal population. Let us suppose that some cases are fully observed on all variables, and the remaining cases are missing all values on a particular subset of variables with the remainder of the variables fully observed. In other words the pattern of missingness is identical for all incompletely observed cases. The data random variables can be partitioned into the "observed" subset $\underline{V}_1$ that is fully observed on all cases, and the "missing" subset $\underline{V}_2$ that has missing values for a portion of the cases. The conditional distribution for $\underline{V}_2$ given $\underline{V}_1$ is thus multivariate normal and can be expressed by equation (3.1).

In reality, it is unlikely that the incompletely observed cases would <u>all</u> exhibit identical patterns of missingness, although there may be "Groups" of cases that do so. (Henceforth in this chapter the term "Group" is used to denote cases with an identical missing data pattern unique to that Group). Different missing data patterns require separate specifications for $p(\underline{V}_2 \mid \underline{V}_1, \underline{\mu}, \Sigma)$, since the subsets of "observed" and "missing" variables vary, as do the corresponding partitions of $\underline{\mu}$ and Σ. Thus a data set with multiple missing data patterns must be divided into Groups, each with its own simple missing data pattern, so that separate conditional distributions can be specified for each. Assuming that the data are missing at random, each conditional distribution is multivariate normal. For each individual

case i (i = 1,...,n) which is subject to missingness the case-specific conditional distribution is denoted by

$$p(\underline{M}_i \mid \underline{O}_i, \underline{\mu}, \Sigma) = N\,[E\,(\underline{M}_i \mid \underline{O}_i, \underline{\mu}, \Sigma)\,,\, \Sigma_{MIO}] \qquad (3.5)$$

where $\underline{M}_i$ and $\underline{O}_i$ represent the "missing" and "observed" variables respectively, and where it is understood that the partitions of $\underline{\mu}$ and Σ vary by Group. Additionally, it should be noted that since $\underline{\mu}$ and Σ are also unknown and (from the Bayesian point of view) considered to be random variables, the distribution of $\underline{M}_i$ is actually conditional on current values for these parameters at each iteration, not on fixed values. The notation $p(\mathbf{M} \mid \underline{\mu}, \Sigma, O)$ should henceforth be understood to represent the collection of conditional distributions $p(\underline{M}_i \mid \underline{O}_i, \underline{\mu}, \Sigma)$.

(b) $p''(\Sigma \mid \mathbf{M}, \underline{\mu}, \mathbf{O})$

The general form of any posterior distribution is specified by Bayes' theorem, which states that

posterior distribution $\propto$ prior distribution $\times$ likelihood

Thus, in the present situation, the posterior distribution of Σ is derived using

$$p''(\Sigma \mid \mathbf{M}, \underline{\mu}, \mathbf{O}) \propto p'(\Sigma)\, L(\Sigma \mid \mathbf{M}, \underline{\mu}, \mathbf{O}) \qquad (3.6)$$

Since we have no strong a priori beliefs concerning the value of Σ, and are willing to depend solely on the observed data to provide information, we adopt a neutral reference prior which is dominated by the likelihood. This choice relies on the reasonable assumption that the location and scale parameters are approximately independent a priori. Using Jeffrey's rule for selection of prior distributions for a set of

parameters, the prior distribution of Σ is given as (Box & Tiao, 1973, p. 426)

$$p'(\Sigma) \propto |\Sigma|^{-(q+1)/2} \tag{3.7}$$

Let $\mathbf{Y} = (\mathbf{M},\mathbf{O})$ represent the completed data set, incorporating current estimates for all missing values. The likelihood is given by

$$L(\Sigma \mid \underline{\mu},\mathbf{Y}) \propto |\Sigma|^{-n/2} \exp[-1/2 \sum_{i=1}^{n} (\underline{Y}_i - \underline{\mu})'\Sigma^{-1}(\underline{Y}_i - \underline{\mu})]$$

or equivalently, by

$$L(\Sigma \mid \underline{\mu},\mathbf{Y}) \propto |\Sigma|^{-n/2} \exp[-1/2 \text{ tr } \Sigma^{-1} \sum_{i=1}^{n} (\underline{Y}_i - \underline{\mu})(\underline{Y}_i - \underline{\mu})']$$

The likelihood can be written even more succinctly

$$L(\Sigma \mid \underline{\mu},\mathbf{Y}) \propto |\Sigma|^{-n/2} \exp[-1/2 \text{ tr } \Sigma^{-1} \mathbf{A}] \tag{3.8}$$

$$\text{where } \mathbf{A} = \sum_{i=1}^{n} (\underline{Y}_i - \underline{\mu})(\underline{Y}_i - \underline{\mu})' \tag{3.9}$$

(i.e., $\mathbf{A}$ represents the sum of squares and cross-products of the completed data based on deviations from the population mean, thus $\mathbf{A}$ depends on $\underline{\mu}$ and $\mathbf{M}$).

Combining expressions (3.6), (3.7), and (3.8) the posterior distribution can be written

$$p''(\Sigma \mid \mathbf{M},\underline{\mu},\mathbf{O}) \propto |\Sigma|^{-(q+1)/2} |\Sigma|^{-n/2} \exp[-1/2 \text{ tr } \Sigma^{-1} \mathbf{A}]$$

or equivalently

$$p''(\Sigma \mid \mathbf{M},\underline{\mu},\mathbf{O}) \propto |\Sigma|^{-(n+q+1)/2} \exp[-1/2 \text{ tr } \Sigma^{-1} \mathbf{A}] \tag{3.10}$$

The right hand side (RHS) of expression (3.10) has the form of an inverted Wishart distribution with n degrees of freedom and $q \times q$ precision matrix

A (Anderson, 1971, p.268), thus the posterior distribution of Σ can be denoted by

$$p''(\Sigma \mid \mathbf{M},\underline{\mu},\mathbf{O}) = W_q^{-1}(A,n) \qquad (3.11)$$

Let $\Psi = \Sigma^{-1}$, then the posterior distribution of Ψ has a Wishart distribution with n degrees of freedom, and $q \times q$ variance-covariance matrix $\mathbf{A}^{-1}$ represented as

$$p''(\Psi \mid \underline{\mathbf{M}},\underline{\mu},\mathbf{O}) = W_q(A^{-1},n) \qquad (3.12)$$

Thus, if a Ψ matrix is randomly generated from $W_q(A^{-1},n)$, the inverse of the generated value yields a value from the posterior distribution for Σ.

(c) $p''(\underline{\mu} \mid \Sigma,\mathbf{M},\mathbf{O})$

Again using Bayes theorem, the posterior distribution of $\underline{\mu}$ is derived using

$$p''(\underline{\mu} \mid \Sigma,\mathbf{M},\mathbf{O}) \propto p'(\underline{\mu})\, L(\underline{\mu} \mid \Sigma,\mathbf{M},\mathbf{O}) \qquad (3.13)$$

With the assumption of a priori independence between the parameters, a non-informative prior for $\underline{\mu}$ is obtained by taking $p'(\underline{\mu})$ locally uniform, and is given by

$$p'(\underline{\mu}) \propto \text{constant} \qquad (3.14)$$

Again, let $\mathbf{Y} = (\mathbf{M},\mathbf{O})$ represent the completed data set, incorporating current estimates for all missing values. The likelihood is

$$L(\underline{\mu} \mid \Sigma ,\mathbf{Y}) \propto |\Sigma|^{-n/2} \exp[-1/2 \sum_{i=1}^{n} (\underline{Y}_i - \underline{\mu})'\Sigma^{-1} (\underline{Y}_i - \underline{\mu})]$$

or equivalently,

$$L(\underline{\mu} \mid \Sigma ,\mathbf{Y}) \propto |\Sigma|^{-n/2} \exp[-1/2 \text{ tr } \Sigma^{-1} \sum_{i=1}^{n} (\underline{Y}_i - \underline{\mu})(\underline{Y}_i - \underline{\mu})'] \qquad (3.15)$$

The expression under the summation sign on the RHS of (3.15) can be rewritten as

$$\sum_{i=1}^{n} (\underline{Y}_i - \bar{\underline{Y}} - \underline{\mu} + \bar{\underline{Y}})(\underline{Y}_i - \bar{\underline{Y}} - \underline{\mu} + \bar{\underline{Y}})'$$

$$= \sum_{i=1}^{n} [(\underline{Y}_i - \bar{\underline{Y}}) - (\underline{\mu} - \bar{\underline{Y}})][(\underline{Y}_i - \bar{\underline{Y}})' - (\underline{\mu} - \bar{\underline{Y}})']$$

$$= \sum_{i=1}^{n} (\underline{Y}_i - \bar{\underline{Y}})(\underline{Y}_i - \bar{\underline{Y}})' + \sum_{i=1}^{n} (\underline{\mu} - \bar{\underline{Y}})(\underline{\mu} - \bar{\underline{Y}})'$$

$$- \sum_{i=1}^{n} (\underline{Y}_i - \bar{\underline{Y}})(\underline{\mu} - \bar{\underline{Y}})' - \sum_{i=1}^{n} (\underline{\mu} - \bar{\underline{Y}})(\underline{Y}_i - \bar{\underline{Y}})' \qquad (3.16)$$

The last two terms on the RHS in (3.16) sum to zero, thus

$$\sum_{i=1}^{n} (\underline{Y}_i - \underline{\mu})(\underline{Y}_i - \underline{\mu})'] = \sum_{i=1}^{n} (\underline{Y}_i - \underline{Y})(\underline{Y}_i - \underline{Y})' + n(\underline{\mu} - \underline{Y})(\underline{\mu} - \underline{Y})' \qquad (3.17)$$

The first term on the RHS in (3.17) is constant in the likelihood, therefore substituting from (3.17) into (3.15) the expression for the likelihood reduces to:

$$L(\underline{\mu} \mid \Sigma, \mathbf{Y}) \propto |\Sigma|^{-n/2} \exp[-1/2 \text{ tr } \Sigma^{-1} n(\underline{\mu} - \bar{\underline{Y}})(\underline{\mu} - \bar{\underline{Y}})'] \qquad (3.18)$$

Thus, combining expressions (3.13), (3.14) and (3.18) the posterior distribution of $\underline{\mu}$ is given by

$$p''(\underline{\mu} \mid \Sigma, \mathbf{Y}) \propto |\Sigma|^{-n/2} \exp[-1/2 \text{ tr } \Sigma^{-1} n(\underline{\mu} - \bar{\underline{Y}})(\underline{\mu} - \bar{\underline{Y}})'] \qquad (3.19)$$

which has the form of a multivariate normal with mean $\bar{\underline{Y}}$ and variance-covariance matrix Σ /n.

2) IMPLEMENTATION OF THE ITERATIVE PROCEDURE AND THE ESTIMATION OF ρ

There are three stages to the estimation process: a) Identification of missing data patterns and selection of starting values for the unknown parameters; b) Repeated applications of the Gibbs sampler to simulate the

marginal posterior distribution for Σ, and consequently the approximate marginal posterior distribution for ρ; and c) Calculation of an interval estimate for ρ from the approximate marginal posterior distribution.

(a) Two preparatory tasks are performed in the initial stage. First, all distinct missing data patterns present in the data are identified, and cases are grouped by patterns of missingness. Next, initial values of $\underline{\mu}$ and Σ are calculated from the subset of fully observed cases.

(b) In the next stage the Gibbs sampler produces a Gibbs sequence by sampling repeatedly from the three conditional distributions

$$p(\mathbf{M} \mid \underline{\mu},\Sigma,O) \qquad p''(\Sigma \mid \mathbf{M},\underline{\mu},\underline{\mathbf{O}}) \qquad p''(\underline{\mu} \mid \Sigma,\mathbf{M},\mathbf{O})$$

Thus each iteration comprises three steps. In the first step a set of imputations for the missing values are generated from $p(\underline{M}_i \mid \underline{\mu},\Sigma,\underline{O}_i)$ for each case separately. In other words, for each case within a particular Group i, the subset of imputation values is generated from the conditional distribution of $\underline{M}_i$ given current values for $\underline{\mu}$ and Σ, and observed values of $\underline{O}_i$ From equation (3.5) this is merely the task of generating a sample from a multivariate normal distribution with known mean vector and variance-covariance matrix, with the partitioning of $\underline{\mu}$ and Σ dependent on the specific Group. Calculation details for the appropriate mean vector and variance-covariance matrix for each Group are shown in equations (3.1) to (3.4). An IMSL subroutine generates random vectors from a multivariate normal distribution with zero mean and covariance matrix $\Sigma_{M_i|O_i}$, then the generated vectors are adjusted by $E(\underline{M}_i|\underline{O}_i)$ to yield deviates from the distribution $p(\underline{M}_i \mid \underline{\mu},\Sigma,\underline{O}_i)$. When the imputation procedure has been

carried out for each Group, we have a completed data set incorporating the current estimates for **M**.

In step two, using the completed data set denoted by **Y** = (**M**,**O**) and the current value for $\underline{\mu}$, according to formulae (3.9) through (3.12) a value for Σ can be generated by sampling from

$$p''(\Psi \mid \mathbf{M},\underline{\mu},\mathbf{O}) = W_q(A^{-1},n)$$

where $\Psi = \Sigma^{-1}$

and $$\mathbf{A} = \sum_{i=1}^{n} (\underline{Y}_i - \underline{\mu})(\underline{Y}_i - \underline{\mu})'$$

where $\underline{Y}_i$ represents the completed data for case i.

Using an algorithm developed in Johnson (1987), Ψ can be generated as follows:

$$\Psi = \mathbf{L}\ \mathbf{T}\ \mathbf{T}'\ \mathbf{L}'$$

where **L**' is the q × q upper triangular matrix obtained from the Cholesky factorization $\mathbf{A}^{-1} = \mathbf{L}\ \mathbf{L}'$

and where **T**' is a q × q upper triangular matrix such that

$$t_{ij} = z \sim N(0,1) \qquad \text{for } i < j,\ j = 2,...,q \qquad (3.20)$$

$$t_{ii} = [\ \chi^2_{(n-i+1)}\ ]^{-1/2} \qquad \text{for } i = 1,...q \qquad (3.21)$$

and t_{ij}'s are independent

In (3.20) z represents a standard univariate normal deviate, and in (3.21) $\chi^2_{(n-i+1)}$ represents a deviate from a chi-squared distribution with n-i+1 degrees of freedom. All components of the upper triangular matrix **T**' are independently generated using IMSL subroutines for random number generation of normal and chi-square variates.

The matrix **A** is first computed by adjusting the completed data set **Y** (incorporating current estimates for missing data) by subtracting the current value for $\underline{\mu}$, and then computing the sum of squares and cross-

products. The matrix **A** is inverted to give $\mathbf{A}^{-1}$, then the square-root factor of $\mathbf{A}^{-1}$ is computed to yield the upper triangular matrix **L'**, using IMSL subroutines. Finally Σ is computed as Ψ^{-1}

In the final step of an iteration, using the completed data set and the updated value for Σ , a vector $\underline{\mu}$ is randomly generated from the conditional distribution $p(\underline{\mu} \mid \Sigma, \mathbf{M}, \mathbf{O})$, where **M** represents the current set of imputations. As shown by formula (3.19) $\underline{\mu}$ has a multivariate normal distribution with mean $\underline{\bar{Y}}$ and variance-covariance matrix Σ/n. An IMSL subroutine is used to generate a random vector from a multivariate normal distribution with zero mean and covariance matrix Σ /n, then the generated vector is adjusted by adding $\underline{\bar{Y}}$ to yield an updated current value for $\underline{\mu}$ from the distribution $p(\underline{\mu} \mid \Sigma, \mathbf{M}, \mathbf{O})$. This step completes one iteration of the Gibbs sequence.

This cycle is iterated k times, to produce the sequence

$$\mathbf{M}^{(1)}, \Sigma^{(1)}, \underline{\mu}^{(1)}, \mathbf{M}^{(2)}, \Sigma^{(2)}, \underline{\mu}^{(2)}, \mathbf{M}^{(3)}, \Sigma^{(3)}, \underline{\mu}^{(3)}, \ldots, \mathbf{M}^{(k)}, \Sigma^{(k)}$$

and the kth Σ is used as the basis for calculating a value for ρ in the following way. Let $\underline{Y}$ represent the set of q variables. Assume that we are interested in the multiple correlation between Y_1 and the vector $(Y_2, Y_3, Y_4, \ldots, Y_q)'$.

Let σ_{11} represent the variance of Y_1

$\underline{\sigma}_{12}$ represent the vector of covariances of Y_1 with $(Y_2, Y_3, \ldots, Y_q)'$

Σ_{22} represent the variance-covariance matrix of variables $Y_2, Y_3, \ldots, Y_q$

The variance-covariance matrix Σ can be partitioned as follows:

$$\Sigma = \begin{bmatrix} \sigma_{11} & \underline{\sigma}'_{12} \\ \underline{\sigma}_{12} & \Sigma_{22} \end{bmatrix}$$

The multiple correlation ρ is given by

$$\rho = [(\underline{\sigma}'_{12} \Sigma_{22}^{-1} \underline{\sigma}_{12}) / \sigma_{11}]^{1/2}$$ (Anderson, 1984, p.134)

It should be noted that regardless of which of the q variables is selected as the dependent, the variables can always be renumbered so that it is possible to express the multiple correlation in this way.

After the first sequence of iterations, the final values obtained for $\underline{\mu}$ and Σ during the last iteration in the preceding sequence are used as start values for a new sequence. The entire sequence of iterations is repeated m times, to produce a sample of values for ρ which represent the approximate marginal posterior distribution of $p''(\rho \mid \mathbf{O})$. As discussed in the literature review, since the process is essentially one run of a single chain, ignoring the initial transient phase then collecting sample outcomes at spaced gaps minimizes serial correlation (Smith & Roberts, 1993). However, the viewpoint that subsampling yields poorer estimates is also strongly argued (MacEachern & Berliner, 1994). Thus, various options regarding choice of values for k and m are tested in independent trials.

(c) In the final stage, point (mean, median, and mode) and interval estimates for ρ are obtained from the approximate marginal posterior distribution of $p''(\rho \mid \mathbf{O})$. Since the posterior distribution is continuous, values are rounded to two decimal places before calculating the mode. Bayesian confidence intervals are usually represented by the highest density region (HDR). For a unimodal posterior distribution the HDR of the distribution is defined to be that interval which is the shortest possible interval containing a given probability and for which the density for every class of points inside the interval is greater than the density for every class outside the interval (Box & Tiao, 1973, p.85). Since the generated sample points for ρ are not grouped into classes, the individual sample values are

ordered by magnitude, and the lower and upper boundaries for all intervals containing 90% of the sample are identified. (e.g., for a sample of size 100, 90% of the sample is contained by the 1st and 90th values, by the 2nd and 91st values etc.) The interval corresponding to each pair of boundary values is calculated. The boundary points for the smallest interval represent the 90% HDR.

3) DESCRIPTION OF THE DATA

Eight sets of simulated data and two real data sets are used to exemplify the Gibbs estimation process. The executable FORTRAN program currently accepts a maximum of 10 variables and a maximum of 200 cases, although these limits can be increased (subject to PC memory limitations) by merely changing the FORTRAN parameter values at the compilation stage.

Two "fully observed" data sets, containing 30 and 60 cases respectively, were generated from each of four different multivariate normal populations. The generation process was performed by an IMSL subroutine for which it was necessary to specify the population correlation matrix. In order to provide a basis of comparability between data sets, some symmetry in the population intercorrelation structure was considered desirable, thus the following constraints were imposed on the correlation matrices: a) within each population, identical pairwise correlations between the dependent variable and each independent variable (ρ_{YX_i}); b) within each population, identical correlations between pairs of independent variables ($\rho_{X_iX_j}$); c) for one pair of populations, identical moderate (approximately .5) values for ρ with approximately equal values for ρ_{YX_i}; and d) for the other pair of populations, identical low (approximately .3)

values for ρ with approximately equal values for ρ_{YX_i}. Using an iterative procedure (in a FORTRAN program), correlation matrices that satisfied these criteria were identified. Four matrices were selected such that for the first pair of populations the value for $\rho_{X_iX_j}$ was low in one population (POP1) and moderate in the other (POP2). For the second pair of populations the value for $\rho_{X_iX_j}$ is moderate in one population (POP3) and high in the other population (POP4). Table 1 displays the intercorrelation structure for all four populations.

In order to satisfy computational requirements (positive definite matrices) the degree of missingness is subject to the constraint that each observation retain at least one observed variable, and that the number of fully observed cases exceed the number of variables by at least one. For each of the simulated data sets with 30 cases, values were deleted at random (MAR) for at most 1 variable on 8 cases, 2 variables on 7 cases, 3 variables on 6 cases, and 4 variables on 3 cases using the following random selection technique. First 8 cases were selected, using a set of 8 randomly generated uniform deviates between 1 and n, then for each of these 8 cases, one of the 5 variables was randomly selected using a randomly generated uniform deviate between 1 and 5. The value on the selected variable was deleted only if the value of at least one of the "observed" variables in the same case had a negative value. Next, from the remaining cases, 7 were randomly selected and for each of these cases two variables were selected randomly, using uniform random deviates as before. Again, the values on the selected variables were deleted only if the value of at least one of the "observed" variables was negative. The same process was repeated to create the remaining missing data patterns. For the data sets containing 60 cases, values were similarly deleted at random for at most 1 variable on 16 cases,

TABLE 1

POPULATION CHARACTERISTICS

FOR SIMULATED DATA SETS

POPULATION	CORRELATIONS		
	$\rho_{mult.}$	ρ_{yx_i}	$\rho_{x_ix_j}$
POP1	0.55	0.40	0.38
POP2	0.55	0.46	0.61
POP3	0.32	0.25	0.50
POP4	0.32	0.29	0.79

2 variables on 14 cases, 3 variables on 12 cases, and 4 variables on 6 cases. Since the missing values were created randomly depending on the values of observed variables for each case, the structure could not be known in advance; the resulting structure is presented in the next chapter. For each data set, a check was made to ensure that a sufficient number of values were set to missing.

The ninth (real) data set (AC) is well-known in the literature (Little & Rubin, 1987, p.101) and concerns size of apple crop. It is an example of bivariate normal data with a monotone pattern of missingness. Finally, the tenth (real) data set (ETS) represents educational test scores and is obtained from Gross (1996). The data set comprises 46 cases with missing data on both of the two variables.

4) SPECIFICATION OF SAMPLE SIZE AND SEQUENCE LENGTH FOR ALTERNATIVE TRIALS

Six independent interval estimates for ρ were produced for each data set. Full-chain sampling and subsampling were performed (discarding outcomes during the initial burn-in phase) for large, medium, and small samples drawn from the marginal posterior distribution of ρ. (The notation used to identify each Gibbs sample is of the form "GFm" or "GSm" where "F" or "S" represent full-chain or subsampling respectively, and the subsequent numeral (m) represents the sample size in thousands). It should be noted that although k (the number of iterations in a Gibbs sequence) needs to be large for the initial burn-in sequence, it need not remain constant, or even large, for successive sequences where subsampling takes place. The trials comprised the following:

I) Large sample size. m = 20,000. This should provide the most rigorous results and is be used as the "benchmark" against which results from the other two examples may be compared.

Full-chain (GF20). k=5,000 for the first Gibbs sequence, k=1 for the remaining 19,999 Gibbs sequences.

Subsampling (GS20). k=5,000 for the first Gibbs sequence, k=30 for the remaining 19,999 Gibbs sequences.

II) Medium sample size. m = 6000

Full-chain (GF6). k=1,000 for the first Gibbs sequence, k=1 for the remaining 5,999 Gibbs sequences

Subsampling (GS6). k=1,000 for the first Gibbs sequence, k=30 for the remaining 5,999 Gibbs sequences

III) Small sample size. m = 1000

Full-chain (GF1). k=1,000 for the first Gibbs sequence, k=1 for the remaining 999 Gibbs sequences

Subsampling (GS1). k=1,000 for the first Gibbs sequence, k=30 for the remaining 999 Gibbs sequences

5) MONITORING AND EVALUATION

A trial run on a few data sets was done first with a Gibbs sample of 2,000 values generated in one chain with no burn-in. Convergence of the algorithm was monitored by plotting the 2,000 sampled values of ρ in sampled order. Next, for each data set, the first Gibbs sample generated was G3F with m = 1,000, initial k = 1,000, and subsequent k = 1. The sample autocorrelation function was computed and plotted (STATGRAPHICS, 1989) to ensure that the autocorrelation coefficient fell

to within sampling error of zero before lag 30 (the chosen value for k in the subsampling examples).

Comparison of the 90% HDR estimates was made by examining overlap of the various intervals obtained from the independent Gibbs samples. The three posterior point estimates were compared with each other. In all examples, comparisons were also made between the "benchmark" intervals from the Gibbs large samples, and point estimates obtained from the SPSS regression procedure using listwise deletion, pairwise deletion, and mean-substitution.

For all simulated data sets, the HDRs obtained were examined for coverage of the value of the population parameter ρ. The posterior mean, median, and mode were also examined for their estimation "accuracy" (subjectively determined). The value of the population parameter for the real data sets was unavailable, however it was possible to calculate exact 90% intervals directly from the data. This was done using the formula derived by Gross (1966) for the posterior distribution of the correlation coefficient in bivariate data with missing values on both variables.

The Gibbs procedure was also carried out on each simulated complete data set (before deletion of "missing" values) to yield interval estimates without the constraint imposed by missing values. This allowed the performance of the Gibbs in the presence of missing data to be compared with the performance when all cases were fully observed.

6) EXAMPLE

In order to elucidate the estimation process using the Gibbs sampler, an unrealistic but simple example is now presented. It must be emphasized that this example in no way demonstrates the ability of the Gibbs sampler

to realistically produce the marginal posterior distribution, merely exemplifies the process. Full-chain Gibbs sampling is demonstrated, with m=10, k=2 for the first Gibbs sequence, k=1 for the remaining 9 Gibbs sequences, on a bivariate normal sample of size 10 with one case subject to missingness on one variable.

To create the sample, a set of 10 observations was generated from a bivariate normal population with population correlation coefficient (ρ) = 0.416, and mean vector and variance-covariance matrix:

$$\underline{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 4 & 3 \\ 3 & 13 \end{bmatrix}$$

and the value of the second variable on one case was set to missing. To make easier reading, all generated and calculated values are shown to two decimal places only, although actual precision is much greater.
The data set is as follows:

Case #	Y_1	Y_2
1	-0.16	-3.14
2	0.46	0.23
3	0.08	2.59
4	-2.21	-0.09
5	-2.08	-0.52
6	0.74	2.76
7	2.14	4.02
8	1.76	
9	0.97	-1.29
10	-0.18	-0.27

As discussed above there are three stages to the estimation process.

STAGE 1 This involves identification of the missing data patterns and Groups, and selection of starting values for the unknown parameters. In this example there is only one missing data pattern, thus only one Group

consisting solely of case #8. Initial values of $\underline{\mu}$ and Σ are calculated from the nine fully observed cases to be

$$\underline{\mu} = \begin{bmatrix} -0.03 \\ 0.47 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 1.94 & 1.45 \\ 1.45 & 5.03 \end{bmatrix}$$

STAGE 2 A Gibbs sequence is created by repeated sampling from the three conditional distributions

$$p(\mathbf{M} \mid \underline{\mu},\Sigma,O) \qquad p''(\Sigma \mid \mathbf{M},\underline{\mu},\mathbf{O}) \qquad p''(\underline{\mu} \mid \Sigma,\mathbf{M},\mathbf{O})$$

Step 1 In order to sample from $p(\mathbf{M} \mid \underline{\mu},\Sigma,O)$ we must first identify the subsets of "observed" and "missing" variables for each Group so as to appropriately partition $\underline{\mu}$ and Σ. In our one Group the "observed" variable subset contains Y_1 only, and the "missing" variable subset contains Y_2 only, so $\underline{\mu}$ and Σ are already appropriately partitioned. Using equations (3.1) to (3.4) we calculate the following:

$$B = 1.45 / 1.94 = 0.75$$

$$\Sigma_{Y_2|Y_1} = 5.03 - (1.45/1.94)*1.45 = 3.95$$

$$E(Y_2 \mid Y_1,\underline{\mu},\Sigma) = \mu_{Y_2} + B(Y_1 - \mu_{Y_1})$$

$$= 0.47 + 0.75(1.76+0.03) = 1.81$$

The conditional distribution for Y_2 given Y_1 for cases in the Group is thus given by

$$p(\underline{M}_i \mid \underline{O}_i,\underline{\mu},\Sigma) = N[1.81, 3.95]$$

Since there is only one case in the Group, just one normal random variate needs to be generated. In this example, since the conditional distribution is a univariate normal, we first use an IMSL routine to generate a variate from a normal distribution with a mean of zero and variance equal to 3.95. The obtained value (-2.45) is then adjusted by the mean (1.81) of the

desired distribution to yield a random variate from N [1.81, 3.95]. The current imputed value M for the missing value of Y_2 on case #8 is consequently calculated as

$$M^{(1)} = -2.45 + 1.81 = -0.64$$

(noting that M is a scalar is this example since only one value is imputed) to form a completed data set **Y** incorporating the current estimates for **M** as follows:

Case #	Y_1	Y_2	
1	-0.16	-3.14	
2	0.46	0.23	
3	0.08	2.59	
4	-2.21	-0.09	
5	-2.08	-0.52	
6	0.74	2.76	
7	2.14	4.02	
8	1.76	*-0.64*	*(imputed value)*
9	0.97	-1.29	
10	-0.18	-0.27	

Step 2 Using the completed data set and the current value for $\underline{\mu}$, a new value for Σ is generated from $p''(\Sigma \mid \mathbf{M},\underline{\mu},\mathbf{O})$ by first randomly drawing a value for Ψ from the Wishart distribution

$$p''(\Psi \mid \mathbf{M},\underline{\mu},\mathbf{O}) = W_q(A^{-1},n)$$

where $$\mathbf{A} = \sum_{i=1}^{n} (\underline{Y}_i - \underline{\mu})(\underline{Y}_i - \underline{\mu})'$$

and then computing $\Sigma = \Psi^{-1}$

Programming details for Wishart variate generation are fully explicated in the Implementation section above.

The new current value is

$$\Sigma^{(1)} = \begin{bmatrix} 7.27 & 7.11 \\ 7.11 & 10.15 \end{bmatrix}$$

Step 3 In this final step of one iteration of the Gibbs sequence, a value for $\underline{\mu}$ must be generated from $p''(\underline{\mu} \mid \Sigma, \mathbf{M}, \mathbf{O})$. Using current values for the completed data set $\mathbf{Y}$ and Σ, a random vector is first generated from the multivariate normal (MVN) distribution with zero mean and covariance matrix Σ/n, and is then adjusted by the vector of sample means from the current completed data set $\mathbf{Y}$ to yield a random vector from $p''(\underline{\mu} \mid \Sigma, \mathbf{M}, \mathbf{O})$. This becomes the new current value for $\underline{\mu}$.

We thus obtain

random vector from MVN($\underline{0}$,Σ/n)		$\bar{\underline{Y}}$		random vector from MVN($\bar{\underline{Y}}$,Σ/n)
$\underline{\mu}^{(1)} = \begin{bmatrix} -1.02 \\ -0.23 \end{bmatrix}$	+	$\begin{bmatrix} 0.15 \\ 0.37 \end{bmatrix}$	=	$\begin{bmatrix} -0.87 \\ 0.14 \end{bmatrix}$

Since we have chosen k to be 2 for the first Gibbs sequence, one more iteration of this three-step cycle is performed using the current values for $\underline{\mu}$ and Σ in Step 1. The resulting sequence is:

Iteration #	$M^{(i)}$	$\Sigma^{(i)}$	$\underline{\mu}^{(i)}$
1	-0.64	$\begin{bmatrix} 7.27 & 7.11 \\ 7.11 & 10.15 \end{bmatrix}$	$\begin{bmatrix} -0.87 \\ 0.14 \end{bmatrix}$
2	3.25	$\begin{bmatrix} 2.01 & 2.03 \\ 2.03 & 8.84 \end{bmatrix}$	$\begin{bmatrix} 0.45 \\ -0.61 \end{bmatrix}$

The final value of Σ in the above sequence is used as the basis for calculating a value for ρ as

$$\rho = [(2.03*2.03)/(2.01*8.84)]^{1/2} = 0.48$$

This value represents the first draw from the marginal posterior distribution of $p''(\rho \mid \mathbf{O})$.

The final values of Σ and $\underline{\mu}$ obtained in the above sequence become the current values used in sampling from $p(\mathbf{M} \mid \underline{\mu},\Sigma,O)$ at the beginning of the next Gibbs sequence. In this example we require a sample of size 10 draws from the marginal posterior distribution, so must produce 9 more values for ρ in the same way. However, since k has been chosen to be 1 for the remaining 9 Gibbs sequences, we need to perform only one iteration of Steps 1, 2, and 3 for each of the remaining 9 sequences. The 10 values of ρ thus obtained are

Sequence	ρ_i
1	0.48
2	-0.03
3	0.56
4	0.22
5	0.22
6	-0.37
7	-0.03
8	0.02
9	0.18
10	0.62

STAGE 3 In the final stage a point estimate for ρ is obtained by computing the average value of all the draws from the marginal posterior distribution to be

$$\rho = \sum_{i=1}^{10} \rho_i /10 = 0.17$$

The 90% HDR of the marginal distribution is determined in the following way. The 10 values for ρ are sorted into ascending order.

Sequence	ρ_i
1	-0.37
2	-0.03
3	-0.03
4	0.02
5	0.18
6	0.22
7	0.22
8	0.48
9	0.56
10	0.62

Since there are 10 values in the sample, 90% of the sample must include 9 values. Thus there are only two such intervals, one extending from the 1st to the 9th ordered values of ρ and one from the 2nd to the 10th values. The size of an interval is determined by the range between the boundary values; the first is 0.93 (the range between -0.37 and 0.56) and the other is 0.64 (the range between -0.03 and 0.62). Since the latter is the smaller of the two intervals the boundary points of the 90% HDR are [-0.03, 0.62]. This interval does in fact include the true value for ρ.

Chapter IV
RESULTS

This chapter is organized in sections as follows:

1) First for the eight simulated data sets, the patterns of missing data resulting from the missing at random (MAR) deletion process are summarized. The pattern of missingness existing in the two real data sets is also outlined.

2) Next, the results of the highest density region (HDR) estimation from all Gibbs samples for each of the ten data sets are presented. The performance of the Gibbs Sampler is assessed in various ways.

(a) Highest density regions (HDRs) from the eight simulated data sets are examined for coverage of the value of the parameter ρ.

(b) HDRs obtained for the two real data sets are compared with analytically derived interval estimates.

(c) For each of the ten data sets taken separately the HDRs from the six Gibbs Samples are compared.

(d) The posterior mean, median, and mode from each Gibbs sample are examined and compared with the corresponding HDR and the value of the parameter ρ for the simulated data or the analytically derived interval estimate for the real data

(e) HDR estimates for both simulated and real data sets are compared with SPSS point estimates.

(f) The effect of the value of the parameter ρ and data set size on interval coverage for the simulated data sets is examined.

(g) HDRs obtained for the simulated data sets are compared with those obtained from the same data sets when no values are missing.

This permits an assessment of differential estimation capability of the Gibbs on complete and incomplete data.

3) Finally, results of some ancillary investigations concerning the operating characteristics of the Gibbs Sampler are reported.

1) A DESCRIPTION OF THE MISSING DATA APPEARING IN THE DATA SETS

There are eight simulated data sets drawn from four populations. All data sets have 5 variables. The pattern of missing data in these data sets is summarized in Table 2. For example, the first two rows in the body of the table contain information for the simulated data sets 1a and 1b with 30 and 60 cases respectively. Both of these data sets are drawn from population POP1 with $\rho_{mult} = .55$, $\rho_{YX_i} = ..40$, and $\rho_{X_iX_j} = .38$. In data set 1a, 12 (40%) of the 30 cases have no missing data. Of the remaining 18 cases, 5 have missing values on one of the variables, 6 cases have missing values on 2 variables, 6 cases have missing values on 3 variables, and 1 case has missing values on 4 variables. Note that values may be missing on any of the five variables, and the pattern differs across cases. In general, about a third of the cases in the 60-case data sets (identified by the suffix "b") are fully observed, whereas the proportion is somewhat more in the 30-case data sets (identified by the suffix "a").

Each of the real data sets contains only two variables. The Apple Crop data set (AC) has 12 cases fully observed, and a monotonic pattern of missing data on one variable in 6 cases. The Educational Test Scores data set (ETS) has 18 fully observed cases and 28 cases with missing values on either of the variables.

TABLE 2

PATTERN OF MISSING DATA IN THE SAMPLE DATA SETS

				Data Set	Total n	Missing values per case				
						0	1	2	3	4
Simulated data										
	$\rho_{mult.}$	ρ_{Yx_i}	$\rho_{x_ix_j}$			(%of n)	Number of cases			
POP1	0.55	0.40	0.38							
				1a	30	12 (40%)	5	6	6	1
				1b	60	21 (35%)	15	12	9	3
POP2	0.55	0.46	0.61							
				2a	30	12 (40%)	5	6	6	1
				2b	60	20 (33%)	15	12	10	3
POP3	0.32	0.25	0.50							
				3a	30	12 (40%)	5	6	6	1
				3b	60	20 (33%)	15	12	9	4
POP4	0.32	0.29	0.79							
				4a	30	13 (43%)	5	5	6	1
				4b	60	19 (32%)	15	12	9	5
Real data										
				AC	18	12 (67%)	6			
				ETS	46	18 (39%)	28			

2) HDR ESTIMATES FOR EIGHT SIMULATED AND TWO REAL DATA SETS WITH MISSING DATA

The first three findings presented are of major importance; evidence for all three comes directly from Table 3. Other results of interest are presented subsequently. Table 3 shows the 90% highest density regions (HDRs) for the ten data sets. For each data set, six interval estimates from the six different Gibbs samples are shown. The Gibbs samples are identified by a letter (F) for full-chain sampling, S for sub-sampling) and a number (20 for 20,000 draws from the posterior distribution of ρ, 6 for 6,000 draws, and 1 for 1,000 draws). It should be noted that the burn-in sequence contains 5,000 iterations of the Gibbs cycle for Gibbs samples GF20 and GS20, whereas the burn-in sequence for each of the other Gibbs samples is 1,000 iterations. For example, in Table 3 the first column of HDRs are those produced by the Gibbs full-chain sample with a burn-in of 5,000 iterations of the Gibbs cycle (GF20) yielding 20,000 values from the marginal posterior distribution for ρ. Each row represents a different data set. Thus the interval (.42,.83) in the upper left hand cell of the body of the table represents the 90% HDR estimate of ρ produced by Gibbs sample GF20 for the data set 1a containing 30 cases drawn from population POP1.

(a) The first major result concerns coverage of ρ for the simulated data sets. As Table 3 shows, the intervals cover the true parameter values of ρ in all but two instances. (The intervals produced by Gibbs sample GF1 for data set 3b and Gibbs sample GS1 for data set 3a are the exceptions). The coverage appears marginal for the two populations with $\rho = .32$ since the value of ρ lies close to the interval boundary for all data sets drawn from these populations.

TABLE 3

90% HIGHEST DENSITY REGIONS

Population	ρ	Data Set	GIBBS SAMPLES GF20	GS20	GF6	GS6	GF1	GS1
Simulated data								
POP1	0.55							
		1a	(.42,.83)	(.43,.84)	.(43,.84)	(.43,.83)	(.43,.85)	(.43,.83)
		1b	(.43,.76)	(.43,.77)	(.42,.76)	(.43,.76)	(.43,.76)	(.44,.75)
POP2	0.55							
		2a	(.41,.83)	(.41,.83)	(.42,.84)	(.42,.83)	(.43,.84)	(.42,.82)
		2b	(.42,.76)	(.42,.76)	(.43,.77)	(.41,.75)	(.44,.75)	(.43,.75)
POP3	0.32							
		3a	(.32,.77)	(.32,.77)	(.31,.77)	(.32,.77)	(.32,.77)	(.33,.77)
		3b	(.30,.70)	(.30,.70)	(.30,.70)	(.30,.70)	(.34,.72)	(.30,.70)
POP4	0.32							
		4a	(.27,.73)	(.28,.73)	(.28,.74)	(.27,.73)	(.28,.74)	(.24,.71)
		4b	(.27,.68)	(.28,.69)	(.27,.69)	(.26,.67)	(.28,.68)	(.27,.67)
Real data								
Exact 90% HDR								
(-.97,-.78)		AC	(-.97,-.78)	(-.98,.-.78)	(-.97,-.79)	(-.97,.-.78)	(-.97,.-.79)	(-.97,-.78)
(-.31,.47)		ETS	(-.30,.47)	(-.32,.46)	(-.31,.47)	(-.30,.47)	(-.36,.45)	(-.31,.42)

(b) Secondly, for the two real data sets, the accuracy of the HDRs produced by the Gibbs Sampler is confirmed by their close match to HDRs calculated analytically from the posterior density of ρ. The formula for the posterior distribution of bivariate correlations with missing data on both variables was derived by Gross (1996). Hence, although the true parameter value of ρ in the real data sets is unknown, the exact 90% HDR for ρ can be constructed for each data set. The resulting intervals are (-.97,-.78) for the Apple Crop data and (-.31,.47) for the Educational Test Scores data. Referring to Table 3, it is apparent that most of the Gibbs samples produce intervals almost identical to the exact ones. The only exceptions are the intervals for the ETS data produced from Gibbs samples GF1 and GS1 which are not quite as close as those produced from the larger Gibbs samples.

(c) The third important finding is that on the whole, for each data set, there is very little variation between the HDRs produced by the separate Gibbs samples (Table 3). An alternative summary of the results of the Gibbs samples is given by Table 4, in which interval widths are displayed for each Gibbs sample and each data set. The row and column structure of the table matches that of Table 3, thus the upper left cell of the body of the table displays the width of the HDR interval for Gibbs sample GF20 on data set 1a. Within each data set, interval widths for all Gibbs samples are very similar. However, on the whole, intervals from the four larger Gibbs samples appear to exhibit slightly more consistency with each other in terms of boundary values and width. As noted above, the only two HDRs which do not cover ρ are produced by Gibbs sample GF1 for data set 3b and Gibbs sample GS1 for data set 3a. Each of these samples contain only 1,000 values drawn from the marginal posterior distribution.

TABLE 4

INTERVAL WIDTHS FOR
90% HIGHEST DENSITY REGIONS

Population	ρ	Data Set	GIBBS SAMPLES GF20	GS20	GF6	GS6	GF1	GS1
Simulated data								
POP1	0.55							
		1a	0.41	0.41	0.41	0.41	0.41	0.40
		1b	0.33	0.34	0.33	0.33	0.33	0.32
POP2	0.55							
		2a	0.42	0.42	0.42	0.41	0.41	0.40
		2b	0.34	0.34	0.34	0.34	0.31	0.33
POP3	0.32							
		3a	0.46	0.46	0.46	0.45	0.44	0.44
		3b	0.40	0.40	0.39	0.40	0.38	0.40
POP4	0.32							
		4a	0.45	0.46	0.46	0.46	0.45	0.46
		4b	0.41	0.41	0.41	0.41	0.40	0.40
Real data								
Exact 90% HDR								
(-.97,-.78)		AC	0.19	0.20	0.19	0.19	0.18	0.19
(-.31,.47)		ETS	0.77	0.78	0.79	0.77	0.81	0.73

(d) The posterior means, medians, and modes from each Gibbs sample are presented in Table 5. The body of the table comprises three blocks each with layout similar to that in Tables 3 and 4, except that instead of a 90% HDR a point estimate is displayed in each cell. The posterior means appear in the left-hand block, posterior medians in the central block, and posterior modes in the right-hand block. All mean and median values lie close to the center of the corresponding interval estimate. Although the posterior means and medians for data sets with $\rho = .55$ are accurate estimates of the population parameter to one decimal place, these same point estimates are extremely inaccurate for data sets with $\rho = .32$. (This is to be expected given the marginal coverage of the interval estimates for these data sets). The consistent small differences between means and medians reflects the slight skewness of all of the Gibbs samples. For each data set, the variability in the posterior mode between Gibbs samples indicates the unreliability of the measure; indeed in two instances, the Gibbs sample contains more than one mode.

(e) Comparisons with SPSS point estimates are made for both simulated and real data sets. The SPSS listwise, pairwise, and mean-substitution point estimates are shown in Table 6 along with HDRs from the benchmark Gibbs samples (GF20 and GS20 for comparison. Each row of the table contains estimates for a different data set. Thus, for example the first row in the body of the table contains estimates for the data set with 30 cases (1a) drawn from population POP1. The three SPSS point estimates are shown first, followed by the interval estimate from Gibbs sample GF20 (full-chain sampling of 20,000 values) and finally the interval estimate from Gibbs sample GS20 (subsampling of 20,000 values). Each SPSS method provides an estimate close to the population ρ for at least one of the

TABLE 5

POINT ESTIMATES OBTAINED FROM GIBBS SAMPLES

Population ρ	Data Set	GIBBS SAMPLES GF20	GS20	GF6	GS6	GF1	GS1	GIBBS SAMPLES GF20	GS20	GF6	GS6	GF1	GS1	GIBBS SAMPLES GF20	GS20	GF6	GS6	GF1	GS1
Simulated data		Posterior mean						Posterior median						Posterior mode					
POP1 0.55																			
	1a	0.63	0.63	0.63	0.63	0.62	0.63	0.64	0.64	0.64	0.64	0.63	0.65	0.66	0.66	0.63	0.71	0.58	0.66
	1b	0.59	0.59	0.59	0.59	0.60	0.59	0.60	0.60	0.60	0.60	0.61	0.60	0.61	0.62	0.61	0.60	0.56	0.60
POP2 0.55																			
	2a	0.62	0.62	0.62	0.62	0.61	0.63	0.63	0.63	0.63	0.63	0.62	0.64	0.64	0.66	0.66	0.66	0.71	0.70
	2b	0.59	0.59	0.59	0.59	0.59	0.59	0.60	0.60	0.60	0.59	0.60	0.60	0.62	0.61	0.61	0.62	0.60	0.63
POP3 0.32																			
	3a	0.54	0.54	0.54	0.54	0.53	0.55	0.55	0.55	0.55	0.55	0.54	0.56	0.59	0.56	0.59	0.59	0.52	0.54
	3b	0.50	0.49	0.50	0.49	0.51	0.49	0.50	0.50	0.51	0.50	0.52	0.50	0.52	0.49	0.54	0.52	0.52	0.52
POP4 0.32																			
	4a	0.50	0.50	0.50	0.50	0.50	0.50	0.51	0.51	0.51	0.51	0.51	0.51	0.50	0.51	0.51	0.57	*0.51	0.61
	4b	0.48	0.48	0.47	0.48	0.47	0.48	0.48	0.49	0.48	0.48	0.47	0.48	0.51	0.51	0.47	0.51	0.44	0.48
Real data																			
Exact 90% HDR																			
(-.97,-.78)	AC	-0.88	-0.88	-0.88	-0.88	-0.88	-0.88	-0.90	-0.90	-0.90	-0.90	-0.89	-0.90	-0.92	-0.92	-0.92	-0.93	-0.92	-0.93
(-.31,.47)	ETS	0.07	0.08	0.07	0.08	0.06	0.08	0.08	0.08	0.07	0.09	0.06	0.08	0.03	0.11	0.01	**0.10	0.01	-0.08

* Trimodal: 0.42, 0.51, 0.61
** Bimodal: 0.10, 0.12

TABLE 6

COMPARISON OF GIBBS INTERVAL ESTIMATES
WITH SPSS POINT ESTIMATES

Population	ρ	Data Set	SPSS POINT ESTIMATES Listwise	Pairwise	Mean-substitution	GIBBS SAMPLES GF20	GS20
Simulated data							
POP1	0.55						
		1a	0.82	0.64	0.53	(.42,.83)	(.43,.84)
		1b	0.57	0.53	0.45	(.43,.76)	(.43,.77)
POP2	0.55						
		2a	0.83	0.64	0.55	(.41,.83)	(.41,.83)
		2b	0.53	0.51	0.43	(.42,.86)	(.42,.76)
POP3	0.32						
		3a	0.67	0.52	0.42	(.32,.77)	(.32,.77)
		3b	0.44	0.36	0.29	(.30,.70)	(.30,.70)
POP4	0.32						
		4a	0.69	0.6	0.46	(.27,.73)	(.28,.73)
		4b	0.34	0.3	0.24	(.27,.68)	(.28,.69)
Real data							
Exact 90% HDR							
(-.97,-.78)		AC	-0.88	-0.88	-0.67	(-.97,-.78)	(-.98,.-.78)
(-.31,.47)		ETS	0.03	0.03	0.02	(-.30,.47)	(-.32,.46)

simulated data sets. However for some of these data sets none of the point estimates are close to the population parameter. Even where the SPSS estimate appears to be "good" it is difficult to assess the degree of accuracy. Interval estimates would be preferred, but these ad hoc methods can only produce point estimates.

(f) All Gibbs samples produce wider intervals for simulated data sets drawn from populations with $\rho = .32$ than for data sets drawn from populations with $\rho = .55$ (Table 4). However, as mentioned above, coverage of ρ when the true value is .55 appears better than when the true value is .32 (Table 3). In particular, for POP3, which has the weakest intercorrelation structure of all four populations (Table 2), the value of ρ lies close to, or at, the interval lower boundary. Also, as expected, within each population smaller intervals are produced for the 60-case data sets than for the 30-case data sets. This is primarily due to shifts in the interval upper boundary values which tend to be smaller (shifted to the left) in 60-case data sets than in 30-case data sets, whereas the lower boundaries do not differ much.

(g) The impact of the missing data on the estimation accuracy of the Gibbs Sampler is also of interest. Since the "missing" data structure in the simulated data sets was created by deletion of values (at random) from the complete data sets, additional Gibbs runs were conducted on the complete simulated data sets (before deletion of "missing" values) to yield interval estimates without the constraint imposed by missing values. As expected, all HDRs resulting from the complete data cover the population parameter value. There are however some systematic differences between the HDRs for the complete data and the HDRs produced when there is missing data.

Table 7 shows the increases in both interval size and boundary values that occur for Gibbs samples produced for incomplete data sets compared with intervals produced for the complete data sets. Thus, examining the upper left cell of the body of the table, the width of the 90% HDR obtained from Gibbs sample GF20 for data set 1a with missing data is .03 larger than the corresponding interval obtained when no values are missing. The central columns of the table show the increase in the value of the lower boundary that occurs when "missing" values are present, and the columns on the right hand side of the table show the increase in the value of the interval upper boundary. In general, although there is considerable overlap of the corresponding intervals produced from the complete and incomplete data sets, boundary values for the incomplete data sets shift to the right of comparable boundaries in the complete data sets. The upper boundary shifts more than the lower boundary, indicating that missing values also result in an increase in interval width. For all populations, there is a tendency for interval width to increase more for the 60-case data sets than for the 30-case data sets, when missing values are introduced into the data. It should be remembered (from Table 2) that there are more missing values in the 60-case data sets than in the 30-case data sets. Thus there are more values to be estimated in the 60-case data sets, which may explain the greater increase in interval width.

3) ANCILLARY INVESTIGATIONS OF THE GIBBS SAMPLER

It is important to consider the operating characteristics of the Gibbs Sampler, irrespective of estimate accuracy. At the inception of this study, test runs of the Gibbs process (a full-chain sample of 2000 values for ρ, with no burn-in) for data sets of size 30 for populations POP1 and POP3

TABLE 7

INCREASE IN HDR ATTRIBUTES
FOR 90% HDRs PRODUCED BY GIBBS WHEN MISSING DATA ARE PRESENT
COMPARED WITH HDRs FOR COMPLETE DATA SETS

Population ρ	Data Set	GIBBS SAMPLES GF20	GS20	GF6	GS6	GF1	GS1	GIBBS SAMPLES GF20	GS20	GF6	GS6	GF1	GS1	GIBBS SAMPLES GF20	GS20	GF6	GS6	GF1	GS1
		Increase in width of interval						Shift in HDR lower boundary						Shift in HDR upper boundary					
POP1 0.55																			
	1a	0.03	0.03	0.03	0.02	0.02	0.02	0.03	0.04	0.04	0.03	0.06	0.03	0.06	0.07	0.07	0.05	0.08	0.05
	1b	0.04	0.04	0.04	0.04	0.04	0.02	0.03	0.05	0.03	0.04	0.04	0.05	0.07	0.09	0.07	0.08	0.08	0.07
POP2 0.55																			
	2a	0.03	0.03	0.03	0.03	0.03	0.02	0.02	0.02	0.03	0.03	0.03	0.03	0.05	0.05	0.07	0.05	0.07	0.05
	2b	0.04	0.04	0.04	0.04	0.02	0.03	0.03	0.03	0.04	0.03	0.05	0.03	0.08	0.08	0.08	0.07	0.07	0.07
POP3 0.32																			
	3a	0.04	0.04	0.04	0.03	0.00	0.01	0.09	0.09	0.09	0.11	0.14	0.12	0.12	0.12	0.13	0.14	0.14	0.13
	3b	0.07	0.07	0.06	0.07	0.05	0.07	0.08	0.08	0.08	0.07	0.10	0.09	0.15	0.15	0.15	0.14	0.15	0.16
POP4 0.32																			
	4a	0.03	0.04	0.04	0.03	0.02	0.03	0.06	0.06	0.05	0.06	0.09	0.01	0.09	0.09	0.08	0.09	0.11	0.04
	4b	0.08	0.08	0.08	0.07	0.07	0.07	0.07	0.08	0.07	0.07	0.07	0.08	0.15	0.16	0.15	0.14	0.14	0.15

were checked for convergence of the algorithm. For each run, values of ρ in the sample were plotted in the order they were generated. There were no discernible trends, hence no evidence of slow convergence of the algorithm or poor starting values.

Next, before proceeding with the full complement of Gibbs samples, for each data set the Gibbs full-chain sample of 1000 values generated by sample GF1 was examined for serial correlation. In all instances, plots of the autocorrelation function (ACF) for lags 1 to 100 revealed large positive low-order correlations which damped out rapidly (after lag 5 or earlier) as lag length increased. Sporadic large ACF values were noted at a few longer lags, but it was determined that all correlations for lag multiples of 30 were within sampling error of zero; thus in subsampling every 30th outcome the effects of any autocorrelation should be eliminated. On the whole, results from the subsequent samples appear unaffected by the presence of autocorrelation since intervals obtained with and without subsampling do not differ substantially. In other words, even if serial correlation is present it does not have any noticeable impact on the boundary values or widths of the resulting HDRs.

Chapter V
SUMMARY AND DISCUSSION

This study presents a Bayesian approach to estimation of the multiple correlation coefficient (ρ) in the presence of missing data which is missing at random. In this situation construction of the posterior distribution function of ρ given the observed data is analytically intractable. Instead, the Gibbs sampler technique, a numerical integration method using an iterative algorithm, produces an approximation to the marginal posterior distribution.

Six Gibbs based interval estimates of the population multiple correlation coefficient (ρ) were obtained for each of eight simulated data sets drawn from four populations and for two real data sets. Performance is assessed in terms of coverage of the parameter value and width of the interval estimate. There are various issues to be considered. Do the Gibbs intervals cover the population multiple correlation coefficient? Does coverage vary for different values of the population multiple correlation? How much does data set size and amount of missing data affect performance? Are there differences in performance between the different length Gibbs samples and between full-chain and subsampling. If differences exist, what is the minimum length for a Gibbs chain to provide reliable (stable) estimates? Is the burn-in length adequate?

Results from the simulated data sets will be discussed first. Almost all (46 out of 48) of the 90% HDRs from the Gibbs samples cover the population parameter value. However, the parameter value always lies below the mid-point of the interval and in some instances is very close to the lower boundary. Thus, since all intervals provide wide coverage to the

right, the lower boundary of the interval is critical in determining whether or not coverage occurs. Overall it appears that narrower intervals with "better" coverage (i.e. ρ is more centered) are obtained when data sets come from a population with a higher value of ρ. Also the stronger the values of ρ_{YX_i} and $\rho_{X_iX_j}$ in the population, the smaller the value of the lower boundary, and hence the better the Gibbs interval coverage. It should be emphasized that any factor that shifts the lower boundary to the right risks non-coverage of ρ.

Within individual populations, narrower intervals resulting from a lowering of the upper boundary are obtained for the 60-case data sets than for the 30-case data sets. It should be noted that, for the data sets used in this study, data set size does not affect whether or not ρ is covered, merely the width of the interval estimate. The comparison of the Gibbs estimates produced for fully observed data sets with those produced for data sets with missing data gives some evidence that the interval width also depends upon the amount of missing data. Thus, although an increase in data set size tends to decrease variability in estimation of the parameter value, the decrease may be offset by an increase in variability resulting from the presence of a large number of missing values which also have to be estimated.

Concerning the results for the bivariate real data sets, it is important to note that the Gibbs results closely approximate the exact interval values derived analytically for both the AC and the ETS data. For example, the interval (-.97,-.78) produced by Gibbs sample GF20 for the AC data is identical to the analytic result. The Gibbs GF20 interval produced for the ETS data is (-.30,.47) whereas the exact interval is (-.31,.47).

In comparison with the simulated data sets, the HDR widths are much narrower for the Apple Crop data (AC). Among all of the data sets, the AC has the smallest number of cases; nevertheless it contains by far the smallest number of missing values. Thus, in accordance with the findings above, narrow intervals are not unexpected. In contrast to the AC data, interval widths for the ETS data are larger than those for any other data set. Based solely on data set size and number of missing values, we might have expected interval widths to fall between those obtained for 30-case data sets and those obtained for 60-case data sets in the simulated data. However, as discussed above, intervals tend to be wider for small values of ρ. Although the true value of ρ for the ETS data is unknown, the interval estimates indicate that it may be low. This might explain why the HDRs are wider for the ETS data than for any of the simulated data

Now we turn to the question of identifying differences between Gibbs samples with a view to finding the optimal Gibbs sample in terms of both performance and economy of resources in order to provide guidelines for future investigators. Overall, for each individual data set the interval width and boundary values do not differ substantially between samples. However, longer Gibbs runs tend to stabilize (usually by lowering) the lower boundary and hence the width of the interval since the upper boundary remains relatively constant. Further investigation subsequent to this study has provided some evidence that smaller Gibbs samples (GF1 and GS1) may not adequately represent the most extreme values in the marginal posterior distribution. This raised the question as to adequacy of the starting values for the variance-covariance matrix. However some additional trials with widely discrepant variance-covariance matrices to start resulted in precisely the same interval estimates. In other words, the

starting values have no effect on outcome. Moreover, one of the additional trials was conducted with no burn-in sequence, indicating that the burn-in sequence is probably unnecessary.

Since the shorter Gibbs runs do not always provide coverage of ρ, in this study only Gibbs samples containing at least 6000 values are considered reliable. However, it is entirely possible that a smaller sample (say 2,000 or 3,000) might prove adequate. It is important to note that subsampling provides no advantage over full-chain sampling in terms of performance. On an IBM-compatible 486 system (33 MHz), the computer run times for the full-chain Gibbs samples on the 30-case simulated data sets were approximately 7 minutes for GF6 and 40 minutes for GF20, whereas the corresponding subsamples took 2 hours and 7 hours respectively. The 60-case data sets required approximately 50% more time for all samples. Since subsampling does not improve performance and is far more costly in computer resources it should not be done.

Several questions have not been addressed by this study. For example, how does the pattern of missingness affect the performance of the Gibbs sampler? Are better estimates obtained when all of the missing data occurs on one or two variables, or when missingness is distributed evenly across all variables? What is the effect of changing the prior distribution for Σ? Additional investigations have indicated that coverage may be poor when the size of the data set is very small. Future research should address this issue. The problem may be due to the implied prior distribution for the multiple correlation coefficient. In this study we have relied on the one-to-one functional relationship between ρ and Σ to calculate ρ from Σ directly. The prior distribution for Σ is taken to be $p'(\Sigma) \propto |\Sigma|^{-(q+1)/2}$ where q represents the number of variables (Box & Tiao, 1973, p. 426). It

is not clear what this implies about the prior distribution for ρ. In the bivariate case, Gross and Torres-Quevedo (1965) showed that the implied prior distribution was

$$p'(\rho) \propto 1/(1 - \rho^2)^{3/2}$$

This prior heavily weights extreme values (positive or negative) for ρ. If the implied prior in the multivariate case were to similarly weight high values for the multiple correlation coefficient, this could produce intervals markedly shifted to the right. The solution may be to try to reparameterize the prior distribution and likelihood function so that ρ itself appears in the distribution function, instead of relying on the relationship between ρ and Σ. Then a reasonable prior could be directly applied to ρ.

The examples shown in this study demonstrate that it is feasible to obtain estimates in a multivariate normal data set where the data is thought to be missing at random. The FORTRAN program created for the study is easily adaptable for use with a data set of any size with any number of variables. Applications to real-world data are numerous. Many multivariate statistical analyses, such as multiple linear regression, principal component analysis, discriminant analysis, and canonical correlation analysis require the initial calculation of the sample variance-covariance matrix. When some of the data values are missing the Gibbs technique can be used to provide estimates for Σ or μ, or functions of these parameters such as the multiple correlation coefficient, by stopping at the appropriate stage in an iteration. Moreover, the missing data concept can be extended to include latent data.

BIBLIOGRAPHY

Amemiya, T. (1985). Advanced econometrics. Cambridge, MA: Harvard University Press.

Anderson, T.W. (1984). An introduction to multivariate statistical analysis. New York: Wiley.

Barry, D. & Hartigan, J.A. (1993). A Bayesian analysis for change point problems. Journal of the American Statistical Association, 88, 309-319.

Besag, J. (1974). Spatial interaction and the statistical analysis of life systems. Journal of the Royal Statistical Society, Ser. B, 36., 192-236.

Box, E.P. & Tiao, G.C. (1992). Bayesian inference in statistical analysis. New York: Wiley.

Buck, S.F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. Journal of the Royal Statistical Society, Ser. B, 22, 302-306.

Carlin, B.P., Gelfand, A.E., & Smith, A.F.M. (1992). Hierarchical Bayesian analysis of changepoint problems. Applied Statistics, 41(2), 389-405.

Casella, G. & George, E.I. (1992). Explaining the Gibbs sampler. The American Statistician, 46(3), 167-174.

Chan, K.S. (1993). Asymptotic behavior of the Gibbs sampler. Journal of the American Statistical Association, 88, 320-326.

Chen, M-H. & Schmeiser, B. (1993). Performance of the Gibbs, Hit-and-Run, and Metropolis samplers. Journal of Computational and Graphical Statistics, 2, 251-272.

De Groot, M.H. (1970). Optimal Statistical Decisions. New York: McGraw Hill.

Dellaportas, P. & Smith, A.F.M. (1993). Bayesian inference for generalized linear and proportional hazards models via Gibbs sampling. Applied Statistics, 42(3), 443-459.

Dempster, A.P., Laird, N.M., & 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.

Diebolt, J. & Robert, C.P. (1994). Estimation of finite mixture distributions through Bayesian sampling. Journal of the Royal Statistical Society, Series B, 56(2), 363-375.

Dixon, W.J. (Ed.) (1985). BMDP statistical software manual: 1985 reprinting. Berkeley: University of California Press.

Efron, B. (1994). Missing data, imputation, and the Bootstrap (with discussion). Journal of the American Statistical Association, 89, 463-479.

Evans, M., Guttman, I., & Olkin, I. (1992). Numerical aspects in estimating the parameters of a mixture of normal distributions. Journal of Computational and Graphical Statistics, 1, 351-365.

Ford, B.L. (1983). An overview of hot deck procedures, in Incomplete Data in Sample Surveys, Vol. II: Theory and Annotated Bibliography. (W. G. Madow, I Olkin, and D.B. Rubin, Eds.). New York: Academic Press.

Frigessi, A. & Staner, J. (1994). Informative priors for the Bayesian classification of satellite images. Journal of the American Statistical Association, 89, 703-709.

Gelfand, A.E., Hills, S.E., Racine-Poon, A., & Smith, A.F.M. (1990). Illustrations of Bayesian inference in normal data models using Gibbs sampling. Journal of the American Statistical Association, 85, 972-985.

Gelfand, A.E. & Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. Journal of the American Statistical Association, 85, 398-409.

Gelman, A. & Rubin, D.B. (1992). Inferences from iterative simulation using multiple sequences. Statistical Science, 7, 457-472.

Geman, S. & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 6, 721-741.

George, E.I. & McCulloch, R.E. (1991). Variable selection via Gibbs sampling. Journal of the American Statistical Association, 88, 881-889.

Geyer, C.J. (1992). Practical Markov chain Monte Carlo. Statistical Science, 7, 473-511.

Gilks, W.R., Clayton, D.G., Spiegelhalter, D.J., Best, N.G., McNeil, A.J., Sharples, L.D., & Kirby, A.J. (1993). Modelling complexity: applications of Gibbs sampling in medicine. Journal of the Royal Statistical Society, Series B, 55, 39-52.

Gilks, W.R. & Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. Applied Statistics, 41(2), 337-348.

Gross, A.L., (1990). A maximum likelihood approach to test validation with missing and censored dependent variables. Psychometrika, 55, 533-549.

Gross, A.L., (1996). Interval estimation of bivariate correlations with missing data on both variables, a Bayesian approach. Manuscript submitted for publication.

Gross, A.L. & Torres-Quevedo, R. (1995). Estimating correlations with missing data, a Bayesian approach. Psychometrika, 60, 341-354.

Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. Biometrika, 57, 97-109.

Heckman, J.J. (1976). The common structure of statistical models of truncation, sample selection, and limited dependent variables, and a simple estimator for such models. Annals of Economic and Social Measurement, 5, 475-492.

Heckman, J.J. (1979). Sample selection bias as a specification error. Econometrika, 47, 153-161.

Johnson, M.E. (1987). Multivariate statistical simulation. New York: Wiley.

Kalton, G. & Kasprzyk, D. (1982). Imputing for missing survey responses. American Statistical Association 1982, Proceedings of the Survey Research Methods Section, pp. 22-31.

Kendall, M.G. (1991). Advanced Theory of Statistics, Vol. 2. New York: Oxford University Press.

Kolassa, J.E. & Tanner, M.A. (1994). Approximate conditional inference in exponential families via the Gibbs sampler. Journal of the American Statistical Association, 89, 697-702.

Kong, A., Liu, J.S., & Wong, W.H. (1994). Sequential imputations and Bayesian missing data problems. Journal of the American Statistical Association, 89, 278-288.

Lee, P.M. (1989). Bayesian statistics: An introduction. New York: Halstead Press.

Little, J.A. & Rubin, D.B. (1987). Statistical analysis with missing data. New York: Wiley.

Liu J., Wong, W.H., & Kong, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. Biometrika, 81, 27-40.

MacEachern, S.N. & Berliner, L.M (1994). Subsampling the Gibbs sampler. The American Statistician, 48, 188-190.

McCulloch, R.E., & Tsay, R.S. (1993). Bayesian inference and prediction for mean and variance shifts in autoregressive time series. Journal of the American Statistical Association, 88, 968-978.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., & Teller, E. (1953). Equations of state calculations by fast computing machines. Journal of Chemical Physics, 21, 1087-1091.

Papoulis, A. (1965). Probability, random variables, and stochastic processes. New York: McGraw Hill.

Rall, L.B. (1969). Computational solution of nonlinear operator equations. New York: Wiley.

Ritter, C. & Tanner, M.A. (1992). Facilitating the Gibbs sampler: The Gibbs stopper and the Griddy-Gibbs sampler. Journal of the American Statistical Association, 87, 861-868.

Rizvi, M. H. (1983). An empirical investigation of some item nonresponse adjustment procedures, in Incomplete Data in Sample Surveys, Vol. I: Report and Case Studies. (W. G. Madow, H. Nisselson, and I. Olkin, Eds.). New York: Academic Press.

Roberts, G.O. & Polson, N.G. (1994). On the geometric convergence of the Gibbs sampler. Journal of the Royal Statistical Society, Series B, 56(2), 377-384.

Ross, S. (1984). A first course in probability. New York: Macmillan.

Rubin, D. B. (1976). Inference and missing data. Biometrika, 63, 581-592.

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. New York: Wiley.

Sinha, D. (1993). Semiparametric Bayesian anlysis of multiple event time data. Journal of the American Statistical Association, 88, 979-983.

Smith, A.F.M. & Gelfand, A.E. (1992). Bayesian statistics without tears: A sampling-resampling perspective. The American Statistician, 46, 84-88.

Smith, A.F.M. & Roberts, G.O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. Journal of the Royal Statistical Society, Series B, 55, 3-23.

Sobel, M.J. (1993). Bayes and empirical Bayes procedures for comparing parameters. Journal of the American Statistical Association, 88, 687-693.

STATGRAPHICS (1989). Users Manual. Rockville, MD: STSC. Inc.

Stevens, D.A. (1994). Bayesian retrospective multiple-changepoint identification. Applied Statistics, 43(1), 159-178.

Tanner, M.A. (1992). Tools for statistical inference. New York: Springer-Verlag.

Tanner, M.A. & Wong, W. (1987). The calculation of posterior distributions by data augmentation (with discussion). Journal of the American Statistical Association, 82, 528-550.

Torres-Quevedo, R. (1993). A Bayesian approach to estimating a correlation with missing data. Unpublished doctoral dissertation, The City University of New York.

Tu X.M. (1994). Analysis of data with censored initiating and terminating times, Journal of Computational and Graphical Statistics. 1, 97-112.

Wakefield, J.C., Smith, A.F.M., Racine-Poon, A., & Gelfand, A.E. (1994). Bayesian analysis of linear and non-linear population models by using the Gibbs sampler. Applied Statistics, 43(1), 201-221.

Wang, R., Sedransk, J. & Jinn, J.H. (1992). Secondary data analysis when there are missing observations. Journal of the American Statistical Association, 87, 952-961.

Wei, G.C.G. & Tanner, M.A. (1990a). Calculating the content and boundary of the highest posterior density region via data augmentation. Biometrika, 77(3), 649-652.

Wei, G.C.G. & Tanner, M.A. (1990b). 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.

Winer, B.J. (1971). Statistical principles in experimental design. New York: McGraw-Hill.

Wu, C.F. (1983). On the convergence properties of the EM algorithm. The Annals of Statistics, 11, 95-103.

Zeger, S. & Rizaul Karim, M. (1991). Generalized linear models with random effects: A Gibbs sampling approach. Journal of the American Statistical Association, 86, 79-86.