

# Analysis of Incomplete Multivariate Data

**J. L. Schafer**

Department of Statistics  
The Pennsylvania State University  
USA



**CHAPMAN & HALL/CRC**

---

A CRC Press Company  
Boca Raton London New York Washington, D.C.

## Library of Congress Cataloging-in-Publication Data

Catalog record is available from the Library of Congress.

This book contains information obtained from authentic and highly regarded sources. Reprinted material is quoted with permission, and sources are indicated. A wide variety of references are listed. Reasonable efforts have been made to publish reliable data and information, but the author and the publisher cannot assume responsibility for the validity of all materials or for the consequences of their use.

Apart from any fair dealing for the purpose of research or private study, or criticism or review, as permitted under the UK Copyright Designs and Patents Act, 1988, this publication may not be reproduced, stored or transmitted, in any form or by any means, electronic or mechanical, including photocopying, microfilming, and recording, or by any information storage or retrieval system, without the prior permission in writing of the publishers, or in the case of reprographic reproduction only in accordance with the terms of the licenses issued by the Copyright Licensing Agency in the UK, or in accordance with the terms of the license issued by the appropriate Reproduction Rights Organization outside the UK.

The consent of CRC Press LLC does not extend to copying for general distribution, for promotion, for creating new works, or for resale. Specific permission must be obtained in writing from CRC Press LLC for such copying.

Direct all inquiries to CRC Press LLC, 2000 N.W. Corporate Blvd., Boca Raton, Florida 33431.

**Trademark Notice:** Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation, without intent to infringe.

Visit the CRC Press Web site at [www.crcpress.com](http://www.crcpress.com)

1997 by Chapman & Hall/CRC

First edition 1997

First CRC Press reprint 1999

Originally published by Chapman & Hall

No claim to original U.S. Government works

International Standard Book Number 0-412-04061-1

Printed in the United States of America 3 4 5 6 7 8 9 0

Printed on acid-free paper

---

# Contents

---

## **Preface**

## **1 Introduction**

- 1.1 Purpose
- 1.2 Background
  - 1.2.1 The EM algorithm
  - 1.2.2 Markov chain Monte Carlo
- 1.3 Why analysis by simulation?
- 1.4 Looking ahead
  - 1.4.1 Scope of the rest of this book
  - 1.4.2 Knowledge assumed on the part of the reader
  - 1.4.3 Software and computational details
- 1.5 Bibliographic notes

## **2 Assumptions**

- 2.1 The complete-data model
- 2.2 Ignorability
  - 2.2.1 Missing at random
  - 2.2.2 Distinctness of parameters
- 2.3 The observed-data likelihood and posterior
  - 2.3.1 Observed-data likelihood
  - 2.3.2 Examples
  - 2.3.3 Observed-data posterior
- 2.4 Examining the ignorability assumption
  - 2.4.1 Examples where ignorability is known to hold
  - 2.4.2 Examples where ignorability is not known to hold
  - 2.4.3 Ignorability is relative
- 2.5 General ignorable procedures
  - 2.5.1 A simulated example
  - 2.5.2 Departures from ignorability
  - 2.5.3 Notes on nonignorable alternatives

- 2.6 The role of the complete-data model
  - 2.6.1 Departures from the data model
  - 2.6.2 Inference treating certain variables as fixed

### **3 EM and data augmentation**

- 3.1 Introduction
- 3.2 The EM algorithm
  - 3.2.1 Definition
  - 3.2.2 Examples
  - 3.2.3 EM for posterior modes
  - 3.2.4 Restrictions on the parameter space
  - 3.2.5 The ECM algorithm
- 3.3 Properties of EM
  - 3.3.1 Stationary values
  - 3.3.2 Rate of convergence
  - 3.3.3 Example
  - 3.3.4 Further comments on convergence
- 3.4 Markov chain Monte Carlo
  - 3.4.1 Gibbs sampling
  - 3.4.2 Data augmentation
  - 3.4.3 Examples of data augmentation
  - 3.4.4 The Metropolis-Hastings algorithm
  - 3.4.5 Generalizations and hybrid algorithms
- 3.5 Properties of Markov chain Monte Carlo
  - 3.5.1 The meaning of convergence
  - 3.5.2 Examples of nonconvergence
  - 3.5.3 Rates of convergence

### **4 Inference by data augmentation**

- 4.1 Introduction
- 4.2 Parameter simulation
  - 4.2.1 Dependent samples
  - 4.2.2 Summarizing a dependent sample
  - 4.2.3 Rao-Blackwellized estimates
- 4.3 Multiple imputation
  - 4.3.1 Bayesianly proper multiple imputations
  - 4.3.2 Inference for a scalar quantity
  - 4.3.3 Inference for multidimensional estimands
- 4.4 Assessing convergence
  - 4.4.1 Monitoring convergence in a single chain

- 4.4.2 Monitoring convergence with parallel chains
- 4.4.3 Choosing scalar functions of the parameter
- 4.4.4 Convergence of posterior summaries
- 4.5 Practical guidelines
  - 4.5.1 Choosing a method of inference
  - 4.5.2 Implementing a parameter-simulation experiment
  - 4.5.3 Generating multiple imputations
  - 4.5.4 Choosing an imputation model
  - 4.5.5 Further comments on imputation modeling

## **5 Methods for normal data**

- 5.1 Introduction
- 5.2 Relevant properties of the complete-data model
  - 5.2.1 Basic notation
  - 5.2.2 Bayesian inference under a conjugate prior
  - 5.2.3 Choosing the prior hyperparameters
  - 5.2.4 Alternative parameterizations and sweep
- 5.3 The EM algorithm
  - 5.3.1 Preliminary manipulations
  - 5.3.2 The E-step
  - 5.3.3 Implementation of the algorithm
  - 5.3.4 EM for posterior modes
  - 5.3.5 Calculating the observed-data loglikelihood
  - 5.3.6 Example: serum-cholesterol levels of heart attack patients
  - 5.3.7 Example: changes in heart rate due to marijuana use
- 5.4 Data augmentation
  - 5.4.1 The I-step
  - 5.4.2 The P-step
  - 5.4.3 Example: cholesterol levels of heart-Attack patients
  - 5.4.4 Example: changes in heart rate due to marijuana use

## **6 More on the normal model**

- 6.1 Introduction
- 6.2 Multiple imputation: example
  - 6.2.1 Cholesterol levels of heart-attack patients
  - 6.2.2 Generating the imputations
  - 6.2.3 Complete-data point and variance estimates
  - 6.2.4 Combining the estimates
  - 6.2.5 Alternative choices for the number of imputations

- 6.3 Multiple imputation: example 2
  - 6.3.1 Predicting achievement in foreign language study
  - 6.3.2 Applying the normal model
  - 6.3.3 Exploring the observed-data likelihood and posterior
  - 6.3.4 Overcoming the problem of inestimability
  - 6.3.5 Analysis by multiple imputation
- 6.4 A simulation study
  - 6.4.1 Simulation procedures
  - 6.4.2 Complete-data inferences
  - 6.4.3 Results
- 6.5 Fast algorithms based on factored likelihoods
  - 6.5.1 Monotone missingness patterns
  - 6.5.2 Computing alternative parameterizations
  - 6.5.3 Noniterative inference for monotone data
  - 6.5.4 Monotone data augmentation
  - 6.5.5 Implementation of the algorithm
  - 6.5.6 Uses and extensions
  - 6.5.7 Example

## **7 Methods for categorical data**

- 7.1 Introduction
- 7.2 The multinomial model and Dirichlet prior
  - 7.2.1 The multinomial distribution
  - 7.2.2 Collapsing and partitioning the multinomial
  - 7.2.3 The Dirichlet distribution
  - 7.2.4 Bayesian inference
  - 7.2.5 Choosing the prior hyperparameters
  - 7.2.6 Collapsing and partitioning the Dirichlet
- 7.3 Basic algorithms for the saturated model
  - 7.3.1 Characterizing an incomplete categorical dataset
  - 7.3.2 The EM algorithm
  - 7.3.3 Data augmentation
  - 7.3.4 Example: victimization status from the National Crime Survey
  - 7.3.5 Example: Protective Services Project for Older Persons
- 7.4 Fast algorithms for near-monotone patterns
  - 7.4.1 Factoring the likelihood and prior density
  - 7.4.2 Monotone data augmentation
  - 7.4.3 Example: driver injury and seatbelt use

## **8 Loglinear models**

- 8.1 Introduction
- 8.2 Overview of loglinear models
  - 8.2.1 Definition
  - 8.2.2 Eliminating associations
  - 8.2.3 Sufficient statistics
  - 8.2.4 Model interpretation
- 8.3 Likelihood-based inference with complete data
  - 8.3.1 Maximum-likelihood estimation
  - 8.3.2 Iterative proportional fitting
  - 8.3.3 Hypothesis testing and goodness of fit
  - 8.3.4 Example: misclassification of seatbelt use and injury
- 8.4 Bayesian inference with complete data
  - 8.4.1 Prior distributions for loglinear models
  - 8.4.2 Inference using posterior modes
  - 8.4.3 Inference by Bayesian IPF
  - 8.4.4 Why Bayesian IPF works
  - 8.4.5 Example: misclassification of seatbelt use and injury
- 8.5 Loglinear modeling with incomplete data
  - 8.5.1 ML estimates and posterior modes
  - 8.5.2 Goodness-of-fit statistics
  - 8.5.3 Data augmentation and Bayesian IPF
- 8.6 Examples
  - 8.6.1 Protective Services Project for Older Persons
  - 8.6.2 Driver injury and seatbelt use

## **9 Methods for mixed data**

- 9.1 Introduction
- 9.2 The general location model
  - 9.2.1 Definition
  - 9.2.2 Complete-data likelihood
  - 9.2.3 Example
  - 9.2.4 Complete-data Bayesian inference
- 9.3 Restricted models
  - 9.3.1 Reducing the number of parameters
  - 9.3.2 Likelihood inference for restricted models
  - 9.3.3 Bayesian inference
- 9.4 Algorithms for incomplete mixed data
  - 9.4.1 Predictive distributions

- 9.4.2 EM for the unrestricted model
- 9.4.3 Data augmentation
- 9.4.4 Algorithms for restricted models
- 9.5 Data examples
  - 9.5.1 St. Louis Risk Research Project
  - 9.5.2 Foreign Language Attitude Scale
  - 9.5.3 National Health and Nutrition Examination Survey

## **10 Further topics**

- 10.1 Introduction
- 10.2 Extensions of the normal model
  - 10.2.1 Restricted covariance structures
  - 10.2.2 Heavy-tailed distributions
  - 10.2.3 Interactions
  - 10.2.4 Semicontinuous variables
- 10.3 Random-effects models
- 10.4 Models for complex survey data
- 10.5 Nonignorable methods
- 10.6 Mixture models and latent variables
- 10.7 Coarsened data and outlier models
- 10.8 Diagnostics

## **Appendices**

- A Data examples
- B Storage of categorical data
- C Software

## **References**

---

## Preface

---

The last quarter of a century has seen enormous developments in general statistical methods for incomplete data. The EM algorithm and its extensions, multiple imputation and Markov chain Monte Carlo provide a set of flexible and reliable tools for inference in large classes of missing-data problems. Yet, in practical terms, these developments have had surprisingly little impact on the way most data analysts handle missing values on a routine basis. My hope is that this book will help to bridge the gap between theory and practice, making a multipurpose kit of missing-data tools accessible to anyone who may need them.

This book is intended for applied statisticians, graduate students and methodologically-oriented researchers in search of practical tools to handle missing data. The focus is applied rather than theoretical, but technical details have been included where necessary to help readers thoroughly understand the statistical properties of these methods and the behavior of the accompanying algorithms.

The methods presented here rely on three fully parametric models for multivariate data: the unrestricted multivariate normal distribution, loglinear models for cross-classified categorical data and the general location model for mixed continuous and categorical variables. In addition, the missing data are assumed to be missing at random, in the sense defined by Rubin (1976). My reviewers have correctly pointed out that many other vitally important topics could (and perhaps should) have been addressed: non-normal models such as the contaminated normal and multivariate-t; repeated measures and restricted covariance structures; censored and coarsened data; models for nonignorable nonresponse; latent variables; and hierarchical or random-effects models. Imputation for

complex surveys and censuses, a topic in which I am deeply interested, deserves much more attention than it received. For better or worse, I decided to limit the material to a few important subjects, but to treat these subjects thoroughly and illustrate them with non-trivial data examples.

This book would not have been possible without the generous support and encouragement of many friends, colleagues and agencies. Don Rubin, whose countless contributions to the area of missing data provided a springboard for this work, was the first to suggest publishing it as a book. The initial round of software development was sponsored by Frank Sulloway, whose wonderfully incomplete dataset provided the first and most colorful application of these methods. Additional support was provided by the Bureau of the Census, the United States Department of Agriculture and the National Center for Health Statistics. Many helpful comments and suggestions were given by John Barnard, Rose Brunner, Andrew Gelman, Bonnie Ghosh, Xiao-Li Meng, Susan Murphy, Maren Olsen, Ritz Scheuren, Stef van Buuren, Recai Yucel and Alan Zaslavsky, and the editorial and production staff at Chapman & Hall. Data on the Foreign Language Attitude Scale were contributed by Mark Raymond. My parents, Chester and Dolores Schafer, created a loving and stable childhood environment, and my wife Sharon did not fail to encourage and inspire. Prayer support was provided by Dr. Samuel C. Lee and members of University Bible Fellowship.

Finally, I must acknowledge my debt to the late Clifford C. Clogg, to whom this book is dedicated. Cliff's steady encouragement and careful review greatly improved the quality of the book, especially the first five chapters. His warmth, love for learning, hard work and faith continue to inspire the many who were close to him. Personally and professionally, it is most gratifying to know that Cliff regarded this book as 'good stuff'.

Joseph L. Schafer  
*University Park, Pennsylvania*  
*October 1996*