②

Iterative Algorithms for Integral Equations of the First Kind with Applications to Statistics

Mark Vangel
Harvard University

Technical Report No. ONR-C-12

October, 1992

Reproduction in whole or in part is permitted for any purpose of the United States Government.

This document has been approved for public release and sale, its distribution is unlimited.

Iterative Algorithms for Integral Equations of the First Kind with Applications to Statistics

Mark Vangel
Harvard University

ABSTRACT

This dissertation explores the use of a preconditioned Richardson iterative algorithm for the solution of linear and nonlinear ill-posed integral equations of the first kind. The discussion consists of three parts, which can be roughly categorized as: numerical analysis, applications to statistical methodology, and an application to an inverse problem.

In the first part, singular matrix equations that result from discretizing ill-posed integral equations of the first kind are considered. Sufficient conditions for the convergence of Richardson's algorithm to a solution are established, and necessary and sufficient conditions are proved for special cases. The inconsistent case is also discussed. A preconditioning for equations with positive kernels leads to the *Conditional Expectation algorithm*, which is discussed in detail. A notion of 'iterative regularization' is introduced and related to the more usual penalized least squares approach to regularization.

In the second part two problems in statistical methodology are considered which involve the solution of nonlinear integral equations of the first kind. The first is the Behrens-Fisher problem. Trickett and Welch (Biometrika, 1954) determined a very nearly similar test for the Behrens-Fisher problem having reasonable power by numerically 'solving' a nonlinear integral equation. The Trickett-Welch method is examined, and a version of the Conditional Expectation algorithm for nonlinear equations is applied to the Behrens-Fisher problem. The second methodological problem that is considered is that of β-content tolerance limits involving data from a one-way balanced random effects model. The Conditional Expectation algorithm is used to approximately solve a nonlinear equation of the first kind numerically, and to thereby derive a new tolerance limit procedure which is shown to be a substantial improvement over the only other method in the statistics literature.

In the third part an inverse problem is discussed in which the right hand side of the integral equation is estimated. In this example, the objective is to infer the probability density of the radii of random spheres in a two-phase medium from radii of circles in cross-sectional slices of this medium. The Conditional Expectation algorithm leads to an effective technique for solving this problem.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION
Unclassified

1b. RESTRICTIVE MARKINGS

2a. SECURITY CLASSIFICATION AUTHORITY

2b. DECLASSIFICATION / DOWNGRADING SCHEDULE

3. DISTRIBUTION / AVAILABILITY OF REPORT

4. PERFORMING ORGANIZATION REPORT NUMBER(S)
TR No. ONR-C-12

5. MONITORING ORGANIZATION REPORT NUMBER(S)

6a. NAME OF PERFORMING ORGANIZATION
Dept. of Statistics
Harvard University

6b. OFFICE SYMBOL *(If applicable)*

7a. NAME OF MONITORING ORGANIZATION

6c. ADDRESS *(City, State, and ZIP Code)*
Department of Statistics, Room SC713
Harvard University
Cambridge, MA 02138

7b. ADDRESS *(City, State, and ZIP Code)*

8a. NAME OF FUNDING / SPONSORING ORGANIZATION

8b. OFFICE SYMBOL *(If applicable)*
Code 1111

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
N00014-91-J1005

8c. ADDRESS *(City, State, and ZIP Code)*
Office of Naval Research
Arlington, VA 22217-5000

10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO

11. TITLE *(Include Security Classification)*
Iterative Algorithms for Integral Equatins of the First Kind With Applications to Statistics

12. PERSONAL AUTHOR(S)
Mark Vangel

13a. TYPE OF REPORT
Technical

13b. TIME COVERED
FROM ______ TO ______

14. DATE OF REPORT *(Year, Month, Day)*
October 1992

15. PAGE COUNT
177

16. SUPPLEMENTARY NOTATION

17. COSATI CODES

FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*

19. ABSTRACT *(Continue on reverse if necessary and identify by block number)*

See reverse side.

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS

21. ABSTRACT SECURITY CLASSIFICATION

22a. NAME OF RESPONSIBLE INDIVIDUAL
Herman Chernoff

22b. TELEPHONE *(Include Area Code)*
617-495-5462

22c. OFFICE SYMBOL

83 APR edition may be used until exhausted
All other editions are obsolete

SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED

Abstract

This dissertation explores the use of a preconditioned Richardson iterative algorithm for the solution of linear and nonlinear ill-posed integral equations of the first kind. The discussion consists of three parts, which can be roughly categorized as: numerical analysis, applications to statistical methodology, and an application to an inverse problem.

In the first part, singular matrix equations that result from discretizing ill-posed integral equations of the first kind are considered. Sufficient conditions for the convergence of Richardson's algorithm to a solution are established, and necessary and sufficient conditions are proved for special cases. The inconsistent case is also discussed. A preconditioning for equations with positive kernels leads to the *Conditional Expectation algorithm*, which is discussed in detail. A notion of 'iterative regularization' is introduced and related to the more usual penalized least squares approach to regularization.

In the second part two problems in statistical methodology are considered which involve the solution of nonlinear integral equations of the first kind. The first is the Behrens-Fisher problem. Trickett and Welch (Biometrika, 1954) determined a very nearly similar test for the Behrens-Fisher problem having reasonable power by numerically 'solving' a nonlinear integral equation. The Trickett-Welch method is examined, and a version of the Conditional Expectation algorithm for nonlinear equations is applied to the Behrens-Fisher problem. The second methodological problem that is considered is that of β-content tolerance limits involving data from a one-way balanced random effects model. The Conditional Expectation algorithm is used to approximately solve a nonlinear equation of the first kind numerically, and to thereby derive a new tolerance limit procedure which is shown to be a substantial improvement over the only other method in the statistics literature.

In the third part an inverse problem is discussed in which the right hand side of the integral equation is estimated. In this example, the objective is to infer the probability density of the radii of random spheres in a two-phase medium from radii of circles in cross-sectional slices of this medium. The Conditional Expectation algorithm leads to an effective technique for solving this problem.

Contents

Acknowledgements

I am indebted to many individuals who helped, either directly or indirectly, to make this thesis possible. Most important of all was the guidance of my advisor, Professor Hernam Chernoff of the Department of Statistics, who has provided encouragement and detailed criticism for almost four years. Like many theses, this document is the tip of an iceberg. Many ideas and applications which do not appear in the final document were explored in detail, and this required considerable time and patience on the part of my advisor, who was working with me in an area not directly related to his own research interests. He taught me a great deal about posing questions, seeking answers, and presenting results. He worked with me extensively during the last year, even though he was on sabbatical leave, and without this effort on his part I would have not graduated this June. Working with Professor Chernoff was the most rewarding educational experience that I have had, and I was always be grateful to him for it.

Professor Donald Anderson of the Division of Applied Sciences, though formally a second reader, voluntarily took on the unofficial role of 'second advisor'. During the 1991-1992 academic year, while Professor Chernoff was on sabbatical leave, I met with Professor Anderson nearly every week. He took an active interest in my research, and provided detailed comments at every step in the final year of thesis writing. Through his criticism of my often clumsy attempts to prove the principal theoretical results of this thesis, Professor Anderson patiently taught me a considerable amount of functional analysis and linear algebra, providing a foundation which I hope to build on in the years to come.

During the years of research and writing, I worked full time as a statistician at the Army Materials Technology Laboratory in Watertown, Massachusetts. Obviously, I could not have both fulfilled my duties to this organization and written a thesis simultaneously without considerable encouragement and support from individuals at the Laboratory. I would like to single out for special thanks among these individuals Donald Neal, who always believed in what I was capable of doing, even at times when I had serious doubts, and Colin Freese, who has been my friend and mentor for over a decade.

I will never forget the students of the Statistics department, who together make a friendly, supportive community within which it was a pleasure to work and learn. Among these students, Pat Meehan, Connie Brown, Tom Blackwell, Chris Schmid, Andrew Gelman, and Xiao-Li Meng come to mind as friends who always provided enthusiastic support and encouargement, without which I might not have made it through.

Finally, I would like to thank my parents, who taught me throughout my life that the opportunity to learn is a privilege, and that the fulfillment of ones potential to learn is of primary importance in life. This thesis is dedicated to them.

Chapter 1

Ill-Posed Integral Equation Problems in Statistics

1.1 Introduction

Many problems of interest either in mathematical statistics or in applications can be formulated as integral equations. We are concerned in this dissertation with the common situation where the integral equation is ill-posed. It is the nature of an ill-posed problem that slight changes in given functions or data cause large changes in the solution. Typically, even changes due to discretization or roundoff error in the computer representation of a function can cause instability when attempts are made to solve the problem numerically.

The main objective of this thesis is to indicate how a simple iterative method, with an appealing probabilistic interpretation, can be used for the numerical solution of what are generally perceived to be difficult integral equation problems.

Let the (possibly nonlinear) integral equation to be solved be

$$\int_0^1 k\left\{x, y, f[\phi(x,y)]\right\} dy = g(x), \tag{1.1}$$

where k, ϕ, and g are known functions. The linearization of this integral equation, which follows from the *Fréchet derivative* of the nonlinear integral operator, is a linear integral equation with *kernel* equal to the derivative of k with respect to its third argument, which we will denote as $k'(x,y,f)$.

Let f^0 be a first approximation to a solution of (1.1); often we will choose $f^0 = 0$. One form of the iteration that we will propose relates f^{n+1} to f^n by

$$f^{n+1} = f^n + \frac{g - \int_0^1 k(x,y,f^n)dy}{\int_0^1 k'(x,y,f^n)dy}. \tag{1.2}$$

An example of a problem which leads to an integral equation of the form (1.1) which does not have a solution in the usual sense of the word, but which can be easily treated numerically by the iteration (1.2), is the Behrens-Fisher problem. Trickett and Welch (1954) apply an iteration, which can be regarded as an approximation to (1.2), to the nonlinear integral equation formulation of this classical problem with amazingly good results. For sample sizes $n_1 = n_2 = 20$, Trickett and Welch provide details of *hand calculations* of five iterations which lead to a smooth critical value statistic which provides a test differing from the nominal size by no more than ± *.000002*, regardless of the value of the variance ratio. We will discuss the Behrens-Fisher problem from the point of view of integral equations in Chapter 5.

Maric and Graybill (1979), independently of Trickett and Welch (1954), applied the same algorithm to a variant of the Behrens-Fisher problem. Wang (1989), using the Trickett-Welch approach, iteratively solved a β-expectation tolerance limit problem for a normal random-effects model. In Chapter 6, we discuss the solution of Vangel (1987, 1990, 1992) to a normal random effects model β-content tolerance limit problem, a problem which can also be formulated as a nonlinear integral equation.

In all four of these cases, the authors use iterative algorithms to 'solve' nonlinear ill-posed problems numerically, problems which have long been known to most likely possess either no solutions, or else only pathological solutions (Linnik, 1968). It is also significant that in none of the above articles is there a single mention of the *ill-posed nature* of the problems being treated numerically.

On the other hand, in the current literature on ill-posed integral equation problems iterative algorithms are scarcely mentioned. Regularization methods dominate this landscape. The usual regularization methods (see, e.g., Tikhonov and Arsenin, 1977) introduce a penalty term which causes a solution to be more or less smooth depending on the value of a parameter. Since any linear smoother solves a certain penalized least squares problem (Hastie and Tibshirani, 1990, p.72), there is regularization implicit in using an iterative method on a problem in which the kernel acts as a smoother, a point which we will take up in Chapter 3.

In addition to interesting problems in mathematical statistics, the methods of this thesis may prove useful in the solution of many ill-posed problems in applied statistics. Although our emphasis will be on problems for which g in (1.1) is known without error, we will also consider, in Chapter 7, a classical inverse problem of stereology where g is either a function observed with error, or else an estimate of a probability density.

1.2 The Ill-Posed Nature of Integral Equations of the First Kind

In this section, we introduce some terminology from the theory of integral equations, and we discuss the concept of an ill-posed problem. A review of the classical theory of integral equations of the first kind along with a discussion of the ill-posedness of these integral equations appears in Chapter 2.

1.2.1 Classification of Integral equations

A linear *Fredholm* integral equation of the *first kind* is an equation of the form

$$\int_0^1 k(x,y)f(y)dy = g(x), \tag{1.3}$$

where k, f and g are functions in L_2. The function $k(x,y)$ is called the *kernel* of the integral equation. For nonlinear integral equations (e.g., 1.1), the kernel is also a function of the unknown f. All of the nonlinear examples which we will consider are special cases of the equation

$$\int_0^1 k\{x, y, f[\phi(\cdot, y)]\}dy = g(x), \tag{1.4}$$

where k, g, and ϕ are known functions.

The general linear Fredholm equation of the *second kind* is

$$g(x) + \lambda \int_0^1 k(x,y)f(y)dy = f(x), \tag{1.5}$$

where λ is a constant. The methods to be discussed in this thesis are also applicable to the serond kind equation (1.5). However, we will not consider the second kind equation further since it is generally well posed (see section 1.2.2 below) and more efficiently solved by methods which exploit the special structure of second kind equations (the classical Fredholm theorems, see, e.g., Smithies, 1958).

If the upper limits in the above integrals are replaced by x, then these equations become equations of the *Volterra* type. A linear Volterra equation of the first kind

$$\int_0^x k(x,y)f(y)dy = g(x), \tag{1.6}$$

can be regarded as a Fredholm equation with the kernel

$$k_*(x,y) = \begin{cases} k(x,y) & \text{if } y \leq x \\ 0 & \text{if } y > x \end{cases} . \tag{1.7}$$

Alternatively, with the change of variable $y = xw$, (1.6) becomes

$$x \int_0^1 k(x,xw)f(xw)dw = g(x), \tag{1.8}$$

an equation with constant limits of integration.

1.2.2 Ill-Posed Problems

Hadamard originated the classification of inverse problems as well- and ill-posed; a general discussion appears in Tikhonov and Arsenin (1977, pp. 7-8). We consider here the nonlinear operator equation $A(f) = g$, where f is to be found in terms of given data g. If $A(f) = g$ for some function f, then we write $f = A^-(g)$. The problem of determining f is *well posed* if the following three conditions are satisfied:

1. For every g there exists a solution f,
2. this solution is unique, and
3. the inverse operator A^- is continuous.

Problems which do not satisfy all of these conditions (particularly condition (3)) are said to be *ill-posed.*

1.2.3 *Near Solutions* and *Near Convergence* of Ill-Posed Integral Equations of the First Kind

When treating an ill-posed integral equation of the first kind numerically, we are usually not interested in obtaining an exact 'solution', because a solution which corresponds to *exactly* the right hand side in a numerical representation of the integral equation can be *very different* from a solution to the original functional equation. The reason for this is that a representation of the right hand side on a computer will always differ (because of discretization error, roundoff error, and possibly noise) from the true function g. We therefore introduce the notion of a *near-solution* for a smooth, well behaved function which results in a right hand side close to the actual right hand side. An iterative algorithm which results in near-solutions after a moderate number of iterations will sometimes be referred to as *nearly convergent*. The iterative algorithms discussed in this thesis can produce near-solutions in practice, even when the matrix discretization, or perhaps the original integral equation, has *no solution*. In practice, one stops after at most a few dozen iterations. In theory, one considers infinitely many iterations, and the nearly convergent algorithm will either converge, possibly to an exact solution (which is likely not to be smooth) or else, in the inconsistent case, the iteration diverges.

1.3 An Example: Deflection of a Simply Supported Beam

We present the following example both to illustrate the ill-posed nature of the integral equation of the first kind and to introduce a very simple integral equation which will be referred to repeatedly in later chapters as a model problem. The problem introduced here is widely used as an example in the literature on numerical methods for integral equations of the first kind.

Consider a thin, elastic beam of unit length 'hinged' at the ends so that bending moments cannot be transmitted from the supports. Let a continuous force be applied perpendicular to the beam, and let this force as a function of position be denoted $f(x)$. The relationship between f and the displacement g that it causes is, for an appropriate choice of material constants,

$$\int_0^1 k(x,y)f(y)dy = g(x), \tag{1.9}$$

where the kernel k (a Green's function) is

$$k(x,y) = \begin{cases} y(1-x) & \text{if } y \leq x \\ x(1-y) & \text{if } y > x. \end{cases} \tag{1.10}$$

The integral equation (1.9) is equivalent to the boundary value problem

$$\frac{d^2g}{dx^2} + f(x) = 0, \tag{1.11}$$

$$g(0) = g(1) = 0,$$

and its solution is

$$f(x) = -\frac{d^2g}{dx^2} \tag{1.12}$$

(see, e.g., Tricomi, 1957, pp. 116-117).

We consider here right hand sides of the form

$$g_l(x) = g_0(x)(1 + \sin(\pi l x)/l). \tag{1.13}$$

The L_2 norm of g_l is

$$||g_l|| \leq ||g_0||(1 + O(1/l)), \tag{1.14}$$

so by making l large enough we have, for arbitrary positive ϵ, that

$$|||g_l|| - ||g_0||| < \epsilon. \tag{1.15}$$

The solution to (1.9), however, is

$$\begin{aligned} f_l(x) &= -g_0''(x)[1 + \sin(\pi l x)/l] + g_0(x) l \pi^2 \sin(\pi l x) \\ &- 2\pi g_0'(x)\cos(\pi l x), \end{aligned} \tag{1.16}$$

and as $l \to \infty$, the difference in norms $|||f_l|| - ||f_0|||$ is unbounded.

1.4 Integral Equations of the First Kind in Statistics

There are several sources of integral equations of the first kind in statistics. These include problems of

1. unbiased estimation,
2. estimating a prior distribution on a parameter given the marginal distribution of the data and the likelihood,
3. similar tests for normal theory problems, and
4. inverse problems of indirect measurement.

We will introduce 1) and 3) in Chapter 2, with detailed discussion of particular examples of 3) (the Behrens-Fisher problem and a tolerance limit problem) to follow in Chapters 5 and 6. An inverse problem of stereology provides an example of 4) which we will consider in Chapter 7. The empirical Bayes problem of estimating a prior distribution is formally very much like 1), and we will not discuss this problem in this thesis.

1.5 An Outline of the Remaining Chapters

Chapter 2 consists of review material from linear algebra, matrix analysis, functional analysis, probability, statistics, and the theory of linear operator equations of the first kind. Most readers will find some of this material helpful, although probably no one will find all of this material new. Because this thesis is partly numerical analysis and partly statistics, it is necessary to consider readers from each of these fields who might not have a strong background in the other discipline.

Chapter 3 contains most of the theoretical discussion of this thesis. We begin by introducing the Richardson and preconditioned Richardson iterative algorithms for linear operator equations of the first kind. We then briefly review the literature on convergence of some basic iterative algorithms in L_2.

This thesis is concerned almost exclusively with matrix equations which arise from the discretization of integral equations. Since the integral equations which we shall consider are ill-posed, the matrix equations which result from discretizations will usually be numerically singular, and often also inconsistent. Even though, because of roundoff error, the discretizations will almost never be *exactly* singular, the study of the singular case helps throw light on the situation where one has *almost* singular matrices, and on the original problem in function space, where one can have nonuniqueness or inconsistency.

In Chapter 3, we establish a sufficient condition for convergence of Richardson's algorithm for consistent matrix equations, and we also prove that the conditions are necessary for an important class of problems. We also discuss the inconsistent case qualitatively, and argue that the proposed algorithms are robust with respect to moderate violation of the consistency assumption.

The proposed iterative algorithms tend to produce smooth approximate solutions; hence there is regularization implicit in using these iterative methods. In Chapter 3, we introduce the notion of *iterative regularization* and relate it to penalized least squares.

Although Richardson's algorithm tends to produce smooth near-solutions in many situations, this algorithm can converge very slowly. The objective of *preconditioning* is to produce a modified algorithm which converges more rapidly. We examine a form of preconditioning of the Richardson iterates for matrix equations with positive matrices. This preconditioning consists of operating on both sides of the equation on the left so as to make the matrix stochastic, hence the name *stochastic preconditioning.* We use the Perron-Frobenius theory of positive matrices to suggest under what conditions our proposed preconditioning can be expected to work well. Several heuristic motivations are also provided; one probabilistic motivation leads to the suggested name *Conditional Expectation Algorithm* for the proposed preconditioned Richardson algorithm and its nonlinear generalizations.

Chapter 3 concludes with linear Fredholm and Volterra examples. A careful discussion of the discretization process is given in an appendix, so that all of the numerical examples in this thesis can be readily duplicated and extended by the interested reader.

In Chapter 4, we consider nonlinear equations. This is a short chapter which serves mostly to establish notation and to generalize the Conditional Expectation algorithm, introduced only for linear problems in the previous chapter, to nonlinear integral equations of the first kind.

In Chapter 5, we begin the discussion of applications to statistics by reviewing the Behrens-Fisher problem, with an emphasis on on the Trickett and Welch (1954) solution. Most of the results of this chapter are not new, but the perspective on the problem is. We are as concerned with the *method* of solution as with the results. Also, unlike Trickett and Welch, we are aware of Linnik's (1968) demonstration that only pathological exact solutions exist. The algorithm which Trickett and Welch use, with much success, can be regarded as a very good approximation to a Conditional Expectation algorithm. In fact, the differences between the iterates produced by the Trickett-Welch and Conditional Expectation algorithms are negligable. However, the Conditional Expectation algorithm can work in situations where the Trickett-Welch approach is not useful, as we show in Chapter 6.

In Chapter 6, we discuss one-sided β-content tolerance limits for a normal population with two components of variance estimated by data from a one-way balanced random-effects ANOVA model. By numerically approximating the solution to a nonlinear integral equation using a Conditional Expectation algorithm, we develop a tolerance limit procedure which provides the appropriate confidence level *almost* independently of the unknown ratio of within- to between-group variances. It is very likely the case that, as with the Behrens-Fisher problem, this tolerance limit problem has either none or else only pathological exact solutions. However, by numerically 'solving' an integral equation of the first kind, using the Conditional Expectation algorithm, we obtain near solutions and are able to develop a method which represents a substantial improvement over the Mee-Owen (1983) approach, which is the only competing procedure in the statistics literature. For ease of computation, we provide coefficients for polynomials fit to the integral equation solutions for two important cases. We also suggest another very simple alternative to the Mee-Owen method.

In Chapter 7, we discuss an interesting example of an ill-posed inverse problem. Consider a two-phase medium where the first phase consists of spherical inclusions of random radius randomly distributed in a second phase. The radii of these spheres are assumed to follow a probability distribution which has a density, and we would like to estimate this density. The available data are *circle* radii measured on *cross-sections* of the material. The density of the circle radii is related to the density of the sphere radii by an Abel integral equation of the first kind. This problem of indirect measurement is typical of the inverse problems of *stereology*, the science of inferring higher dimensional structure from lower dimensional data. In this chapter, we derive the Abel equation (first reported in Wicksell (1925)) and briefly review the extensive literature on this problem. We then proceed to apply the Conditional Expectation algorithm in order to develop a method for solving this equation. This apparently new approach is demonstrated on both simulated and real data. For the simulated data, we consider both the case where the density of the circle radii is a function observed with noise, and where the circle radius density is estimated by a sample from this probability density.

Chapter 2

A Review of Background Material from Linear Algebra, Functional Analysis, Probability, and Statistics

2.1 Matrix Algebra

We review here those concepts from matrix algebra which will be used in this thesis. We assume familiarity with topics generally covered in a first course in this subject, although we will briefly review some of these ideas (eigenvalue, similarity, etc.) for completeness. The definition of a vector space, and a discussion of the important notions of range and nullspace are deferred until Section 2.3, where we take these topics up in a more general Hilbert space setting.

In this thesis, we will denote m-dimensional complex Euclidean space by $\mathcal{C}^m$, and m-dimensional real Euclidean space by $\mathcal{R}^m$.

2.1.1 Elementary Notions

Let A be an arbitrary $m \times n$ matrix with elements $a_{ij} \in \mathcal{C}$. The entry in the ith row and jth column of A is a_{ij}, and we write this as $A_{ij} = a_{ij}$. The *transpose* of A, A^T, has typical element $A^T_{ij} = a_{ji}$, and the *adjoint* of A, A^*, has typical element $A^*_{ij} = \bar{a}_{ji}$, where the overbar denotes complex conjugation. If $A = A^T$, then A is *symmetric*; if $A = A^*$, then A is *Hermitian*. The *rank* of a matrix A is the number of linearly independent rows, which equals the number of linearly independent columns.

A scalar λ is an *eigenvalue*, and a nonzero vector x is a corresponding *eigenvector*, of a square matrix A if

$$Ax = \lambda x. \tag{2.1}$$

If A is Hermitian, then λ is real. If A is Hermitian, and for all $x \in \mathcal{C}^m$, $x \neq 0$,

$$x^* Ax \geq 0, \tag{2.2}$$

then A is *positive semi-definite*, and all eigenvalues of A are nonnegative. If the inequality in (2.2) is strict, then A is *positive definite* and all of the eigenvalues of A are positive.

The eigenvalues of a square matrix A are the roots of the *characteristic polynomial*

$$d(\lambda) \equiv |A - \lambda I|, \tag{2.3}$$

where $|\cdot|$ denotes the determinant. If zero is an eigenvalue, then $|A| = 0$ and the matrix A is said to be *singular*; otherwise A is *nonsingular*. The multiplicity of an eigenvalue as a root of $d(\lambda)$ is called the *algebraic multiplicity* of the eigenvalue. The dimension of the subspace of eigenvectors corresponding to an eigenvalue is called the *geometric multiplicity* of the eigenvalue. The geometric multiplicity of an eigenvalue is always less than or equal to its algebraic multiplicity.

Matrices for which the algebraic and geometric multiplicities of at least one eigenvalue are not equal are said to be *defective*, or *non-diagonalizable*. When the multiplicity of an eigenvalue is referred to without a modifier, algebraic multiplicity is implied. When we refer to a set of eigenvalues, or to the cardinality of such a set, without explicitly stating that we mean *distinct* eigenvalues, then it is to be understood that we have in mind eigenvalues repeated according to their algebraic multiplicities. Sometimes we will state this idea briefly by using the phrase 'counting multiplicities'.

Two square matrices which represent the same linear transformation, possibly with respect to different bases, are said to be *similar*. In particular, similar matrices have the same eigenvalues. We state this formally as

Definition 2.1.1 (Similarity) *Two $m \times m$ matrices A and B are said to be* **similar** *if there exists a nonsingular matrix S such that*

$$A = S^{-1}BS.$$

If A is Hermitian, then A is similar to a diagonal matrix (which must have the eigenvalues of A as its diagonal elements). There is a matrix S which provides the similarity transformation and is unitary, i.e. $S^{-1} = S^*$. (A matrix U for which $U^*U = I$ is said to be *unitary*; if $U^TU = I$, then U is *orthogonal*.) The following result is the *spectral theorem for Hermitian matrices:*

Theorem 2.1.1 (Spectral theorem for Hermitian matrices) *Let A be a Hermitian matrix. Then there is a matrix U such that $U^*U = I$ and*

$$A = U\Lambda U^*,$$

where Λ is a real diagonal matrix whose diagonal elements are eigenvalues of A, and where the columns of U are corresponding eigenvectors.

If A is Hermitian, then there is a set of orthonormal eigenvectors (i.e. eigenvectors $\{u_i\}$ for which $u_i^*u_j$ equals one if $i = j$, and zero otherwise); the columns of U form one such set. In general, the matrix U of the theorem is not unique. If A is real and symmetric, then U can be selected to be real also.

2.1.2 The Singular Value Decomposition

Assume that A is an $m \times n$ matrix with $m \leq n$, and that the rank of A is $q \leq m$. Then A^*A and AA^* are Hermitian, positive semidefinite matrices. The eigenvalues of AA^*, which we denote σ_i^2, where

$$\sigma_1^2 \geq \sigma_2^2 \geq \ldots \geq \sigma_m^2 \geq 0, \tag{2.4}$$

are also eigenvalues of A^*A. If $n > m$, then the $n \times n$ matrix A^*A has $n - m$ additional eigenvalues which equal zero. The nonnegative numbers $\{\sigma_i\}_{i=1}^m$ are called *singular values*, and we have the following result, which is, in a sense, one possible extension of Theorem 2.1.1 to general matrices (Horn and Johnson, 1985, p. 414):

Theorem 2.1.2 (Singular Value Decomposition) *Let A be an arbitrary $m \times n$ matrix, with $m \leq n$, and let the rank of A be $q \leq m$. Then, there exist unitary matrices U and V, where U is $m \times m$ and V is $n \times n$, and an $m \times n$ diagonal matrix Σ, such that*

$$A = U\Sigma V^*.$$

The m diagonal elements of Σ are the singular values of A, denoted $\{\sigma_i\}_{i=1}^m$, where

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_q > 0 = \sigma_{q+1} = \ldots = \sigma_m.$$

The columns of U and V are called left and right **singular vectors**, *respectively, of A. The columns of U are eigenvectors of AA^*, and the columns of V are eigenvectors of A^*A (arranged in the same order as the corresponding eigenvalues σ_i^2).*

Here again, if A is real, we can find real orthogonal matrices U and V. If A is a square matrix with eigenvalues λ_i and singular values σ_i, then

$$\max_i \sigma_i \geq \max_i |\lambda_i|. \tag{2.5}$$

2.1.3 The Jordan Canonical Form

Not all matrices are *diagonalizable*, that is, similar to a diagonal matrix. Any square matrix, however, is similar to a matrix which is *nearly* diagonal, and this near-diagonal representation is referred to as the *Jordan Canonical Form* (Horn and Johnson, 1985, Chapter 3).

Let A be an arbitrary $m \times m$ matrix of rank $q \leq m$. There exists a nonsingular matrix S such that

$$J = S^{-1}AS, \tag{2.6}$$

where

$$J = \operatorname{diag}(J_1, \ldots, J_r), \tag{2.7}$$

a *Jordan form* matrix, is block diagonal, with $r \leq m$ blocks. Each *Jordan block* J_i has an eigenvalue of A, λ_i, on its main diagonal, ones on the diagonal for which the column index is one greater than the row index, and zeros everywhere else. For example, if J_i happens to be 4×4, then it will be a matrix of the form

$$J_i(\lambda_i) = \begin{bmatrix} \lambda_i & 1 & & \\ & \lambda_i & 1 & \\ & & \lambda_i & 1 \\ & & & \lambda_i \end{bmatrix}. \tag{2.8}$$

If the dimension of J_i is $n_i \times n_i$, then $\sum_{i=1}^r n_i = m$, where the sum of the geometric multiplicities of distinct eigenvalues is r. The Jordan form exists for any square matrix, and it is, except for permutations of rows and columns, unique.

2.1.4 Matrices Having a Diagonalizable Nullspace

Let A be an $m \times m$ matrix of rank q. If A is nonsingular, then $q = m$ and there exists a nonsingular matrix B and a nonsingular Jordan form matrix J such that $A = B^{-1}JB$. If A is singular, then $q < m$ and there exists a nonsingular matrix B such that

$$A = B^{-1} \begin{bmatrix} J_{s\times s} & 0_{s\times(m-s)} \\ 0_{(m-s)\times s} & N_{(m-s)\times(m-s)} \end{bmatrix} B, \tag{2.9}$$

where: $s \leq q$, J is a nonsingular $s \times s$ Jordan form matrix, and N is a matrix of Jordan blocks corresponding to a zero eigenvalue. The rank of J, s, is equal to the rank of A, q, if and only if $N = 0$. If $N \neq 0$, then $s < q$, since the nonzero rows of BAB^{-1} corresponding to rows of N are each linearly independent of the rows of BAB^{-1} corresponding to rows of J.

Consider the submatrix N in (2.9), and let the typical element of this matrix be denoted n_{ij}. For $l = 0, \ldots, m-s-1$, define the lth *super-diagonal* to be the set of entries $s_l = \{n_{i,i+l}\}_{i=1}^{m-s-l}$. The only nonzero elements of N are on the first super-diagonal, and these values equal one. It is easy to show by direct calculation that any nonzero elements of N^2 must be ones on the *second* super-diagonal. To see this, compute the square of any Jordan block corresponding to a zero eigenvalue, for example (2.8) with $\lambda_i = 0$. Similarly, N^l, for $l < m-s$, must be zero everywhere except possibly on the lth super-diagonal. It follows that $N^l = 0$ for all $l \geq m-s$.

A matrix which when raised to some power is equal to a zero matrix is said to be *nilpotent*, which is the reason why the letter 'N' is used in (2.9). The smallest positive integer ι such that $N^\iota = 0$ is called the *index* of both the matrix N and the matrix A in the Jordan form representation (2.9). If a matrix A is nonsingular, we define it to have index $\iota = 0$.

When N in (2.9) is a zero matrix, then the geometric multiplicity of zero as an eigenvlaue of a singular matrix equals its algebraic multiplicity. If N has a nonzero block, then this is no longer the case. We will refer to the class of singular matrices for which

$$A = B^{-1} \begin{bmatrix} J_{q\times q} & 0_{q\times(m-q)} \\ 0_{(m-q)\times q} & 0_{(m-q)\times(m-q)} \end{bmatrix} B, \tag{2.10}$$

for nonsingular J and B, as matrices *having a diagonalizable nullspace*. We are introducing this *nonstandard* terminology in this thesis, since, for our purposes, it is more suggestive than the usual definition: i.e. that a matrix A is of the form (2.10) if and only if A has a *group inverse* (Campbell and Meyer, Chapter 7). However, it will be convenient to express certain results in terms of the group inverse of a matrix, so we define this concept next.

Definition 2.1.2 (Group Inverse) *Let A be an arbitrary square matrix. A generalized inverse matrix, $A^\#$, such that*

1. $A^\# A A^\# = A^\#$,

2. $AA^\# A = A$, *and*

3. $AA^\# = A^\# A$

is called the **group inverse** *of the matrix A.*

If $A^{\#}$ exists, then it is unique. If A is singular and $A^{\#}$ exists, then A must be of the form (2.10). We can see by direct calculation that

$$A^{\#} = B^{-1}\begin{bmatrix} J^{-1}_{q\times q} & 0_{q\times(m-q)} \\ 0_{(m-q)\times q} & 0_{(m-q)\times(m-q)} \end{bmatrix} B \tag{2.11}$$

is the group inverse of A. For a detailed discussion of the properties of the group inverse, see Chapter 7 of Campbell and Meyer (1979).

2.1.5 Congruence

Definition 2.1.3 (Congruence) *A square matrix B is said to be* **congruent** *to a matrix A if there exists a nonsingular matrix S such that*

$$B = SAS^*.$$

It is easy to show that the properties of being positive definite and positive semi-definite are preserved by a congruence transformation.

Lemma 2.1.1 *Let A be positive semi-definite, and let B be congruent to A. Then B is positive semi-definite. If A is positive definite, then B is also.*

Proof: For some nonsingular matrix S, and any nonzero vector x, we have that

$$x^*Bx = x^*SAS^*x = (S^*x)^*A(S^*x) = y^*Ay \geq 0, \tag{2.12}$$

since A is positive semi-definite by hypothesis. If A is positive definite and $x \neq 0$, then $y = S^*x$ is not zero, the inequality (2.12) is strict, and hence B is positive definite. ∎

2.1.6 Nonnegative Matrices

There is an extensive theory for matrices having nonnegative elements (e.g., Horn and Johnson, 1985, chapter 9). A matrix A is said to be *positive*, and we write $A > 0$, if all of the elements of A are strictly positive. Similarly, if A has only nonnegative elements, we say that A is *nonnegative*, and we write $A \geq 0$. The fundamental theorem in the theory of nonnegative matrices is the Perron-Frobenius theorem, a special case of which we state below.

The maximum of the moduli of the eigenvalues of a matrix is called the *spectral radius*, and denoted $\rho(A)$. Since this is an important notion, we give a formal definition:

Definition 2.1.4 (Spectral Radius) *Let A be an $m\times m$ matrix with eigenvalues $\{\lambda_i\}_{i=1}^{m}$, where the λ_i need not all be distinct. The spectral radius of A is defined by*

$$\rho(A) \equiv \max_{1\leq i\leq m} |\lambda_i|.$$

The spectral radius is the radius of the smallest circle, centered at the origin, which contains all of the eigenvalues. We can now state a version of the Perron-Frobenius theorem.

Theorem 2.1.3 (Perron-Frobenius) *Let $A > 0$ be a positive $m \times m$ matrix, and assume that the eigenvectors of A have norm one. Then the following are among the properties of A:*

1. *The spectral radius of A is equal to p, where p is a real eigenvalue of A with algebraic multiplicity one, and p is the unique eigenvalue of modulus p.*

2. *The matrix A has a positive eigenvector x corresponding to p.*

3. *Denote the sums of the values in the ith row of A by r_i, and the* ordered *row sums, from smallest to largest, by $r_{(i)}$. Then*

$$r_{(1)} \leq p \leq r_{(m)}.$$

We will refer to the positive eigenvalue p and the corresponding positive eigenvector x (of norm one) of the above theorem as the *Perron-Frobenius eigenvalue* and *Perron-Frobenius eigenvector* respectively.

A nonnegative matrix for which all of the row sums equal one is called *stochastic*. It follows immediately from the Perron-Frobenius theorem that for a stochastic matrix the Perron-Frobenius eigenvalue and eigenvector are $p = 1$ and

$$x = \frac{1}{\sqrt{m}} \begin{bmatrix} 1 \\ 1 \\ \cdot \\ \cdot \\ 1 \end{bmatrix}, \qquad (2.13)$$

respectively.

We have as another consequence of the Perron-Frobenius theorem the following result relating positive matrices, Perron-Frobenius eigenvectors, and stochastic matrices:

Lemma 2.1.2 *Every positive matrix is similar to a matrix proportional to a stochastic matrix*

Proof: Let A be positive, and let D_x be the diagonal matrix whose diagonal elements are those of the Perron-Frobenius eigenvector x of A. Then

$$B = D_x^{-1} A D_x$$

has constant row sums. To see this, let a_{ij} denote the typical element of A, let r_i denote the sum of the elements in the ith row of B, and let x_i denote the ith diagonal entry in D_x. Then, for any i,

$$r_i = \sum_{j=1}^{m} \frac{a_{ij} x_j}{x_i} = \rho(A) \frac{x_i}{x_i} = \rho(A),$$

therefore, $B/\rho(A)$ is stochastic. ∎

Knowing that the spectral radius is bounded by the extremal row sums often provides useful upper, but not lower, bounds on $\rho(A)$. One reason for this is that positive kernels often decrease to zero at the boundaries of their domain, and the corresponding row sums of a discretized matrix will be near zero. Actually, the *average* row sum is still a lower bound for the spectral radius of a positive symmetric matrix A, as we show below:

Lemma 2.1.3 *The Perron-Frobenius eigenvalue of a positive symmetric matrix is bounded below by the* **average** *row sum*

Proof: Let A be positive and $m \times m$ with elements a_{ij}, and Perron-Frobenius eigenvalue and eigenvector denoted by p and x, respectively. Since p is larger in modulus than any other eigenvalue of A, we have that (Strang, 1976, p. 253)

$$\sup_{y \neq 0} \frac{y^T A y}{y^T y} = p.$$

Define the unit vector

$$z = [1, ..., 1]^T / \sqrt{m}.$$

Then

$$z^T A z = \sum_{i=1}^{m} \sum_{j=1}^{m} a_{ij}/m = \sum_{i=1}^{m} r_i/m \leq p.$$

∎

2.2 Matrix Analysis

There are many ways to define a norm for square matrices, and corresponding to each norm there is a metric on the space of square matrices. There is, therefore, a theory of *matrix analysis*, for which the two volume work of Horn and Johnson (1985, 1990) is an excellent reference.

2.2.1 Matrix Norms

A matrix norm satisfies the following five axioms (Horn and Johnson, 1985, p. 290):

Definition 2.2.1 (Matrix Norm) *Let $\|\cdot\|$ be a mapping from the space of square matrices, with elements in $\mathcal{C}$, to $\mathcal{R}$. The function $\|\cdot\|$ is a* **matrix norm** *if, for all $m \times m$ matrices A and B,*

1. $\|A\| \geq 0$
2. $\|A\| = 0$ *if and only if* $A = 0$
3. $\|cA\| = |c|\|A\|$ *for all scalars* $c \in \mathcal{C}$
4. $\|A + B\| \leq \|A\| + \|B\|$.
5. $\|AB\| \leq \|A\|\|B\|$.

Properties (1-4) are the axioms of a *vector norm*; a norm with property (5) is called *submultiplicative*.

The largest singular value of a matrix provides a matrix norm, the *spectral norm* (Horn and Johnson, 1985, p. 295):

Lemma 2.2.1 (Spectral Norm) *The largest singular value of a matrix A is a matrix norm, called the* **spectral** *or l_2 norm and denoted*

$$\|A\|_2 \equiv [\rho(A^* A)]^{1/2} = \sigma_1.$$

On occasion, we will use another matrix norm, the l_∞ *norm*, which is easily expressed in terms of the elements of a matrix (Horn and Johnson, 1985, p. 295):

Lemma 2.2.2 (l_∞ norm) *Let A be an $m \times m$ matrix with typical element a_{ij}. The function $\| \cdot \|_\infty$, defined by*

$$\|A\|_\infty \equiv \max_{1 \leq i \leq m} \sum_{j=1}^{m} |a_{ij}|$$

is a matrix norm, called the l_∞ **norm,** *or simply the* **infinity norm.**

The *spectral norm* should not be confused with the *spectral radius.* In general, the spectral radius is not a norm, but for each fixed square matrix A it is the greatest lower bound for the values of all matrix norms of A (Horn and Johnson, 1985, p. 297).

Theorem 2.2.1 *Let a matrix A and $\epsilon > 0$ be given. Then*

1. *For any matrix norm $\| \cdot \|_\alpha$,*

$$\rho(A) \leq \|A\|_\alpha.$$

2. *There exists a matrix norm $\| \cdot \|_\beta$ such that*

$$\rho(A) \leq \|A\|_\beta \leq \rho(A) + \epsilon$$

2.2.2 Convergent Matrices

A square matrix A is said to be *convergent* (Horn and Johnson, 1985, p. 298) if

$$\lim_{k \to \infty} A^k = 0, \tag{2.14}$$

that is, if all of the elements of A^k decrease to zero in absolute value as $k \to \infty$. Another definition of a convergent matrix, easily shown to be equivalent to (2.14), is

Definition 2.2.2 (Convergent Matrix) *An $m \times m$ matrix A is* **convergent** *if, for all vectors $v \in C^m$,*

$$\lim_{k \to \infty} A^k v = 0. \tag{2.15}$$

A necessary and sufficient condition for a matrix to be convergent is given by the following theorem (Horn and Johnson, 1985, p. 138):

Theorem 2.2.2 *A square matrix A is convergent if and only if $\rho(A) < 1$.*

If $\rho(A) = 1$, then the powers of A can converge to a *nonzero* matrix. A matrix A for which this is the case is sometimes referred to as *semi-convergent.* We discuss this idea in more detail below.

If $\rho(A) = 1$ and A has a Jordan block which is an identity submatrix, then there is a corresponding subspace $\mathcal{U}$ such that for $u \in \mathcal{U}$, $A^n u$ does not diverge, although $A^n u \not\to 0$ unless $u = 0$. If the Jordan form has a block $I + N$, where $N \neq 0$ is nilpotent, then $A^n u$ blows up for u in a corresponding space.

The following theorem (Horn and Johnson, 1985, p. 299) says something about the rate at which a convergent matrix approaches zero:

Theorem 2.2.3 *Let A be an $m \times m$ matrix, and let $\epsilon > 0$ be given. Then, there exists a constant $C = C(A, \epsilon)$ such that*

$$\left|(A^k)_{ij}\right| \leq C(\rho(A) + \epsilon)^k, \tag{2.16}$$

for all $k = 1, 2, 3, \ldots$, and for all $i, j = 1, 2, \ldots, m$.

We will need to sum series of powers of matrices in Chapter 3. The following useful lemma follows immediately from Theorem 2.2.2:

Lemma 2.2.3 (Geometric Series) *If A is a square matrix and $\rho(I - A) < 1$, then A is nonsingular and*

$$\sum_{i=0}^{\infty}(I - A)^i = A^{-1}. \tag{2.17}$$

If $\rho(I - A) = 1$, then $(I - A)^i \not\to 0$, so (2.17) cannot be a convergent series. However, the partial sums of (2.17) may remain bounded, as can be seen from the example

$$A = \begin{bmatrix} 1-i & 0 \\ 0 & 1/2 \end{bmatrix}. \tag{2.18}$$

2.2.3 Condition Numbers

With respect to any matrix norm, the *condition number* of a nonsingular matrix A is defined as

$$\kappa(A) \equiv \|A\|\|A^{-1}\|. \tag{2.19}$$

If A is singular, then $\kappa(A) \equiv \infty$. Note that, for any matrix norm, and any nonsingular matrix A,

$$\kappa(A) \geq \|AA^{-1}\| = \|I\| \geq \rho(I) = 1. \tag{2.20}$$

A condition number provides a measure of how nearly singular a matrix is, with a large condition number suggesting that a matrix is 'nearly' singular. Let $Kf = g$ be a matrix equation. If κ is a condition number with respect to a norm $\|\cdot\|_*$, then for any two vectors f and $\tilde{f}$ (Stoer and Bulirsch, 1980, p. 179),

$$\frac{\|\tilde{f} - f\|_*}{\|f\|_*} \leq \kappa \frac{\|K\tilde{f} - Kf\|_*}{\|Kf\|_*}, \tag{2.21}$$

so a condition number relates the relative change in the right hand side of an equation to the relative change in a solution. The most often used condition number is defined in terms of the l_2 norm. Let A be a nonsingular $m \times m$ matrix with largest and smallest singular values given by σ_1 and σ_m, respectively. Then

$$\kappa_2(A) = \frac{\sigma_1}{\sigma_m} \tag{2.22}$$

is the condition number of A with respect to the l_2 norm.

2.3 Elementary Notions of Functional Analysis

Since we are ultimately interested in approximating integral equations by matrix equations, and attempting to solve these resulting matrix equations on a computer, most of the theoretical discussions in this thesis will be in m-dimensional space which we take, for flexibility, to be $\mathcal{C}^m$ rather than $\mathcal{R}^m$. However, we will make use of some function-space results concerning integral equations in a Hilbert space, and so we review here the functional analysis that we will require.

2.3.1 Normed Vector Spaces

A *vector space*, $\mathcal{H}$, is a set of elements, called *vectors*, together with the operations of *vector addition* and *scalar multiplication*, over a scalar field. We will take this scalar field to be either the complex numbers $\mathcal{C}$, or the real numbers $\mathcal{R}$. The defining properties of a vector space are as follows:

Definition 2.3.1 (Vector Space) *Let $\mathcal{H}$ be a nonempty set, let C be a scalar field, and let there be two binary operations '$+$' and '$\times$', corresponding to vector addition and scalar multiplication, respectively. Let x, y, z be arbitrary points in $\mathcal{H}$, and let $\alpha, \beta, \gamma \in C$ be arbitrary scalars. Then $\mathcal{H}$ is a* **vector space**, *and the points in $\mathcal{H}$ are called* **vectors**, *if all of the following properties are satisfied:*

1. *There is a binary operation, called* **vector addition**, *that assigns to each pair of elements $x, y \in \mathcal{H}$ a unique element of $\mathcal{H}$ called their sum, and denoted $x + y$. For all $x, y, z \in \mathcal{H}$:*

   (a) $x + y = y + x$,

   (b) $x + (y + z) = (x + y) + z$,

   (c) there is an element $0 \in \mathcal{H}$ such that $x + 0 = x$, and

   (d) there is an element $-x \in \mathcal{H}$ such that $x + (-x) = 0$.

2. *There is a rule which assigns to each pair $\alpha \in C$ and $x \in \mathcal{H}$ a unique vector, called the* **scalar product** *of α and x, and denoted αx. For arbitrary $\alpha, \beta \in C$ and $x, y \in \mathcal{H}$, the scalar product has the following properties:*

   (a) $\alpha(\beta x) = (\alpha\beta)x$,

   (b) $\alpha(x + y) = \alpha x + \alpha y$,

   (c) $(\alpha + \beta)x = \alpha x + \beta x$, *and*

   (d) $1x = x$.

For a function f, which assigns for each $x \in A$ an element $y = f(x) \in B$, we write $f : A \to B$. The set A is called the *domain* of f; the set of all $y = f(x)$ for $x \in A$ is called the *range* of f and denoted $\mathcal{R}(f)$. A function is also sometimes called an *operator* or a *mapping*, and we will use these terms interchangeably, although different terms are customary in different contexts.

Functional analysis is concerned with analysis on vector spaces. In order to do analysis, we need a generalization of the idea of the *distance* between two vectors of an arbitrary vector space. This leads to the concept of a *metric*:

Definition 2.3.2 (Metric) *Let X be a set, and let $x, y, z \in X$ be arbitrary points. A* **metric**, *$d(x, y) : X \times X \to \mathcal{R}$ is defined by the following properties:*

1. *d is finite and nonnegative,*

2. $d(x, y) = 0 \iff x = y$,

3. $d(x, y) = d(y, x)$, *and*

4. $d(x, z) \leq d(x, y) + d(y, z)$.

We define next a function mapping vectors into nonnegative scalars called a *norm*, thereby generalizing the notion of length to vectors in abstract spaces:

Definition 2.3.3 (Norm) *Let $\mathcal{H}$ be a vector space, and let $x, y \in \mathcal{H}$, and $\alpha \in C$ be arbitrary. A* **norm**, $\|\cdot\| : \mathcal{H} \rightarrow \mathcal{R}$, *is defined by the following four properties:*

1. $\|x\| \geq 0$,
2. $\|x\| = 0 \iff x = 0$,
3. $\|\alpha x\| = |\alpha|\|x\|$, *and*
4. $\|x + y\| \leq \|x\| + \|y\|$.

A metric can always be defined in terms of a norm, for example

$$d(x, y) \equiv \|x - y\|. \tag{2.23}$$

A *normed space* is a vector space together with a norm, $(\mathcal{H}, \|\cdot\|)$. Usually the norm is understood, and the normed space is denoted simply $\mathcal{H}$. We can do analysis in general normed spaces; in particular, we can define limits and Cauchy sequences.

Definition 2.3.4 (Limit, Convergence) *A sequence $\{x_n\}$ in a normed vector space $(\mathcal{H}, \|\cdot\|)$* **converges** *to a* **limit** *x if $x \in \mathcal{H}$, and for every $\epsilon > 0$ there exists an $N = N(\epsilon)$ such that for all $n > N$*

$$\|x - x_n\| < \epsilon.$$

Definition 2.3.5 (Cauchy sequence) *A sequence $\{x_n\}$ in a normed vector space $(\mathcal{H}, \|\cdot\|)$ is called a* **Cauchy sequence** *if, for every $\epsilon > 0$, there exists an $N = N(\epsilon)$ such that for all $m, n > N$*

$$\|x_m - x_n\| < \epsilon.$$

A normed vector space in which all Cauchy sequences converge (to vectors in $\mathcal{H}$) is called *complete*. A complete normed vector space is a *Banach space.*

We can discuss limits and continuity in Banach space, but we have no notion of orthogonality, and so most of the geometry of finite dimensional Euclidean space does not apply to a general Banach space. However, if the additional structure of an *inner product* is imposed on a Banach space, the *complete inner product space*, or *Hilbert space*, which results has a geometry which is in some ways very much like Euclidean space. An inner product is defined as follows:

Definition 2.3.6 (Inner product) *Let X be a vector space, and let $x, y, z \in X$ and $\alpha, \beta \in C$ be arbitrary. An* **inner product** *is a function $(\cdot, \cdot) : X \times X \rightarrow C$ with the following properties:*

1. $(x + y, z) = (x, z) + (y, z)$
2. $(\alpha x, y) = \alpha(x, y)$,
3. $(x, y) = \overline{(y, x)}$, *and*
4. $(x, x) \geq 0$; $(x, x) = 0 \iff x = 0$.

A fundamental inequality for inner products is the *Cauchy-Schwarz inequality:*

Lemma 2.3.1 (Cauchy-Schwarz Inequality) *Let x and y be any vectors in an inner product space. Then*

$$|(x,y)|^2 \leq (x,x)(y,y), \tag{2.24}$$

with equality if and only if either $x = 0$, or $y = 0$, or $y = \alpha x$ for some constant α.

An inner product determines a norm,

$$\|x\| \equiv (x,x)^{1/2}, \tag{2.25}$$

and the Cauchy-Schwarz inequality relates this norm to the corresponding inner product.

Another important result which holds in a general inner product space is the Pythagorean Theorem:

Theorem 2.3.1 (Pythagorean Theorem) *Let x, x_1, and x_2 be vectors in an inner product space, where $x = x_1 + x_2$ and $(x_1, x_2) = 0$. Then*

$$\|x\|^2 = \|x_1\|^2 + \|x_2\|^2.$$

Proof:

$$\begin{aligned} \|x\|^2 &= (x_1,x_1) + (x_2,x_2) + (x_1,x_2) + (x_2,x_1) \\ &= \|x_1\|^2 + \|x_2\|^2 \blacksquare \end{aligned}$$

2.3.2 Hilbert Space

A Banach space in which the norm is determined by an inner product is called a *Hilbert space*. An example of a Hilbert space with scalar field $\mathcal{R}$ is $\mathcal{R}^m$, with inner product $(x,y) \equiv x^T y$ and norm $\|x\| = (x^T x)^{1/2}$. Another important example is the space of square integrable complex valued functions, L_2.

Let us define an inner product (f,g) on the vector space L_2 of all Lebesgue measurable complex valued functions for which

$$\int_{-\infty}^{\infty} |f(x)|^2 dx < \infty \tag{2.26}$$

as follows

Definition 2.3.7 (Inner Product in L_2) *Let $f, g \in L_2$. The* **inner product** (f,g) *is*

$$(f,g) \equiv \int_{-\infty}^{\infty} f(x)\overline{g(x)}dx,$$

where the integral is a Lebesgue integral.

This inner product determines the norm

$$\|f\|_2 \equiv (f,f)^{1/2} = \sqrt{\int_{-\infty}^{\infty} |f(x)|^2 dx} \tag{2.27}$$

(Kreyszig, 1978, p. 62). Strictly speaking, by a function f we mean an *equivalence class* of functions which are equal almost everywhere. It turns out that L_2 is complete with respect to this norm, and hence is a Hilbert space.

We will refer to this space as L_2, and to the norm $\|\cdot\|_2$ as the L_2 *norm*. We will be concerned primarily with functions of a real variable, and we will sometimes refer to this *real* function space as L_2. There are corresponding spaces for functions with other domains which we will also refer to as L_2. Sometimes the domain of the functions in the space is included in the notation, for example $L_2[0,1]$ is the space of square integrable functions of a single variable on the unit interval. We will use the notation which doesn't indicate the domain where there is no risk of confusion.

By analogy with the inner product on $\mathcal{R}^m$, we say that two vectors x and y are *orthogonal* if $(x, y) = 0$. If e is a unit vector and x is any vector, then we call (x, e) the *orthogonal projection*, or simply the *projection*, of x onto e. A sequence of unit vectors $\{e_i\}$ is called an *orthonormal sequence* if

$$(e_i, e_j) = \delta_{ij}, \tag{2.28}$$

where δ_{ij} is the *Kronecker* δ,

$$\delta_{ij} \equiv \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} . \tag{2.29}$$

An orthonormal sequence is called *complete* if *any* vector in the space can be expressed as a limit of linear combinations of elements in this sequence. A complete orthonormal sequence is also called an *orthonormai ɥasis*, or simply a *basis*. In particular, L_2 has such bases. A Hilbert space which, like L_2, has a countable orthonormal basis is said to be *separable*. Although we will state some results more generally, we will confine our attention primarily to L_2.

'Complete' thus has two meanings. A normed vector space is *complete* if all Cauchy sequences converge; an orthonormal system in a Hilbert space is *complete* if all vectors in the space can be expressed as limits of linear combinations of vectors in the orthonormal system. It will always be clear from the context which notion of completeness is to be used.

For $\mathcal{R}^m$ there is a natural notion of dimension. We can now provide a general definition for Hilbert space. If the number of vectors in a basis is finite, then this number is the same for any basis, and is called the *dimension* of the space. If a space has an orthonormal basis consisting of infinitely many vectors, then all bases consist of infinitely many vectors, and we say that the space is *infinite dimensional.*

For any $f \in \mathcal{H}$, we represent f formally in terms of a basis $\{e_i\}_{i=1}^{\infty}$ as a *Fourier series*

$$f = \sum_{i=1}^{\infty} \alpha_i e_i, \tag{2.30}$$

where, for every i, $\alpha_i \equiv (f, e_i)$. The α_i are called *Fourier coefficients* of f with respect to the basis $\{e_i\}$. We will always interpret an infinite sum such as (2.30) to mean that

$$\lim_{N\to\infty} \left\| f - \sum_{i=1}^{N} \alpha_i e_i \right\|^2 = 0. \tag{2.31}$$

Let $\mathcal{H}$ be a Hilbert space. Under what conditions does every vector $f \in \mathcal{H}$ have a Fourier series representation (2.30)? Bessel's inequality provides a first step toward an answer to this question:

Lemma 2.3.2 (Bessel Inequality) *Let $\{e_i\}$ be any orthonormal sequence in c Hilbert space $\mathcal{H}$. Let $x \in \mathcal{H}$ be arbitrary. Then*

$$\sum_i |(x, e_i)|^2 \leq \|x\|^2 \tag{2.32}$$

If $\mathcal{H}$ is separable, then we can say more:

Lemma 2.3.3 (Parseval Identity) *Let $\{e_i\}$ be an orthonormal basis in a separable Hilbert space $\mathcal{H}$. Let $x \in \mathcal{H}$ be arbitrary. Then*

$$\sum_i |(x, e_i)|^2 = \|x\|^2. \tag{2.33}$$

If we are working in a separable Hilbert space, then we can use Parseval's identity to show that (2.31) holds for any vector f. Therefore, if we interpret convergence in the sense of convergence in norm, f has, for a given basis, a Fourier series (2.30). It can be shown that this series is unique. A good discussion, in the context of integral equations, of the material of this paragraph is in Tricomi (1957, pp. 83-88).

2.3.3 Linear Operators

We will be concerned with linear operators in L_2, so we give a formal definition of a linear operator:

Definition 2.3.8 (Linear Operator) *Let U_1 and U_2 be vector spaces. A* **linear operator** *$K : U_1 \rightarrow U_2$ is an operator (i.e., a mapping) such that for any x and y in U_1, and for any scalars $\alpha, \beta \in C$,*

$$K(\alpha x + \beta y) = \alpha K(x) + \beta K(y).$$

We will be interested exclusively in the case where U_1 and U_2 are Hilbert spaces. We will adopt the conventional notation Kx for $K(x)$. We collect here some definitions for classes of linear operators which will be needed in this and subsequent chapters:

Definition 2.3.9 (Bounded Operator) *A linear operator $K : U_1 \rightarrow U_2$ between Hilbert spaces is* **bounded** *if there exists a real number c such that, for all $x \in U_1$,*

$$\|Kx\| \leq c\|x\|. \tag{2.34}$$

Definition 2.3.10 (Continuous Operator) *Let $K : U_1 \rightarrow U_2$ be a linear operator between Hilbert spaces. K is said to be* **continuous** *if for any $\epsilon > 0$, there exists a $\delta > 0$ such that for any vectors x_1 and x_2 in U_1,*

$$\|x_1 - x_2\| < \delta \Rightarrow \|Kx_1 - Kx_2\| < \epsilon.$$

It can be shown (Kreyszig, 1978, p. 97) that a linear operator is continuous if and only if it is bounded. It is customary to talk in this context about bounded, not continuous, operators.

Let $\mathcal{H}$ be a Hilbert space, and let $y_0 \in \mathcal{H}$ be arbitrary. The special linear operator $K : \mathcal{H} \rightarrow C$ defined by $Kx \equiv (x, y_0)$ is bounded. Also, linear operators on $\mathcal{R}^m$ can be represented by matrices, and are necessarily bounded.

Definition 2.3.11 (Operator Norm) *Let $K : U_1 \rightarrow U_2$ be a bounded linear operator between Hilbert spaces. The norm of the operator K, $\|K\|$, is*

$$\|K\| = \sup_{x \in U_1, x \neq 0} \frac{\|Kx\|}{\|x\|}.$$

Definition 2.3.12 (Adjoint Operator) *Let $K : U_1 \rightarrow U_2$ be a bounded linear operator between Hilbert spaces. The* **adjoint** *of K is the operator $K^* : U_2 \rightarrow U_1$ such that, for all $x \in U_1$ and $y \in U_2$,*

$$(Kx, y) = (x, K^*y).$$

We take for granted here that this definition makes sense: that is, that the adjoint exists. A proof that the adjoint K^* of a bounded linear operator K exists, is bounded and unique, and that $\|K\| = \|K^*\|$ can be found in Kreyszig (1978, pp. 196-197).

Definition 2.3.13 (Self-Adjoint Operator) *A bounded linear operator K is said to be* **self-adjoint** *if $U_1 = U_2$ and $K = K^*$.*

Definition 2.3.14 (Positive Operator) *Let $K : U \rightarrow U$ be a self-adjoint linear operator on a Hilbert space. K is said to be* **positive** *if, for all $x \in U$,*

$$(Kx, x) \geq 0.$$

For $U = \mathcal{R}^m$, with scalar field $\mathcal{R}$, positive operators correspond to positive semi-definite matrices.

Eigenvalues and eigenvectors can also be defined for general linear operators:

Definition 2.3.15 (Eigenvalue and Eigenvector) *Let $K : U \rightarrow U$ be a linear operator. The vector $x \in U$ is an* **eigenvector**, *and the scalar $\lambda \in C$ is an* **eigenvalue**, *if*

$$Kx = \lambda x.$$

A positive, self-adjoint linear operator has only real, nonnegative eigenvalues (Kreyszig, 1978, p. 475, problem 5). In matrix analysis, the adjoint corresponds to the transposed complex conjugate (or transpose, for symmetric matrices), and self-adjoint operators correspond to Hermitian (or symmetric, in the real case) matrices.

2.3.4 Orthogonal Complements in Hilbert Space

We review in this subsection some basic ideas about the geometry of Hilbert space which we will make extensive use of in Chapter 3. A detailed exposition of this material appears in Kreyszig (1978, Chapter 3).

We begin with some elementary notions. Let $\mathcal{H}$ be a Hilbert space, and let $\mathcal{H}_1 \subseteq \mathcal{H}$ be an arbitrary subset of $\mathcal{H}$. We say that $\mathcal{H}_1$ is a *subspace* of $\mathcal{H}$ if it is a vector space. If $\mathcal{H}_1$ contains all of its limit points (with respect to the norm induced by the inner product on $\mathcal{H}$), then $\mathcal{H}_1$ is said to be a *closed* subspace of $\mathcal{H}$. A subspace of a complete metric space is itself complete if and only if it is closed (Kreyszig, 1978, p. 30), and hence $\mathcal{H}_1$ is a Hilbert space with respect to the inner product on $\mathcal{H}$ if and only if $\mathcal{H}_1$ is closed.

Let $\mathcal{H}$ be a Hilbert space, and let $\mathcal{H}_1 \subseteq \mathcal{H}$ and $\mathcal{H}_2 \subseteq \mathcal{H}$ be arbitrary subspaces. The subspaces $\mathcal{H}_1$ and $\mathcal{H}_2$ are said to be *orthogonal* if, for any $h_1 \in \mathcal{H}_1$ and $h_2 \in \mathcal{H}_2$, $(h_1, h_2) = (h_2, h_1) = 0$. We write this as $\mathcal{H}_1 \perp \mathcal{H}_2$.

The set of all vectors orthogonal to a subspace $\mathcal{H}_1$,

$$\mathcal{H}_1^{\perp} \equiv \{h \in \mathcal{H} | h \perp \mathcal{H}_1\}, \tag{2.35}$$

is a subspace called the *orthogonal complement* of $\mathcal{H}_1$. The orthogonal complement $\mathcal{H}_1^{\perp}$ is closed, and if $\mathcal{H}_1$ is closed, then $\mathcal{H}_1^{\perp\perp} = \mathcal{H}_1$ (Kreyszig, 1978, p. 149). In general, we have that $\mathcal{H}_1^{\perp\perp} = \bar{\mathcal{H}}_1$, where we denote the *closure* of a space by an overbar.

If every $h \in \mathcal{H}$ can be expressed uniquely as $h = h_1 + h_2$, where $h_1 \in \mathcal{H}_1$ and $h_2 \in \mathcal{H}_2$, then $\mathcal{H}$ is equal to the *direct sum* of the subspaces $\mathcal{H}_1$ and $\mathcal{H}_2$, and we write

$$\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2. \tag{2.36}$$

Assume now that $\mathcal{H}_1$ is an arbitrary *closed* subspace, and that $\mathcal{H}_2 = \mathcal{H}_1^{\perp}$. Then $\mathcal{H}$ can be written as the direct sum (2.36) (Kreyszig, 1978, p. 146). This result is referred to as the *projection theorem*. If $h = h_1 + h_2$, where $h_1 \in \mathcal{H}_1$ and $h_2 \in \mathcal{H}_2$, we say that h_1 is the *orthogonal projection* (or briefly, the *projection*) of h onto the closed subspace $\mathcal{H}_1$.

Given a linear operator between Hilbert spaces, $K : U_1 \to U_2$, define the *nullspace* of K by

$$\mathcal{N}(K) = \{x \in U_1 | Kx = 0\}. \tag{2.37}$$

It is easy to show that $\mathcal{N}(K)$ is a closed subspace. The *range* of K is

$$\mathcal{R}(K) = \{y \in U_2 | y = Kx \text{ for some } x \in U_1\}. \tag{2.38}$$

The operator K is said to be of *infinite rank* if the dimension of $\mathcal{R}(K)$ is infinite, otherwise K is said to be of *finite rank.*

The nullspace and range for the adjoint operator, $K^* : U_2 \to U_1$, are

$$\mathcal{N}(K^*) = \{x \in U_2 | K^* x = 0\}, \tag{2.39}$$

and

$$\mathcal{R}(K^*) = \{y \in U_1 | y = K^* x \text{ for some } x \in U_2\}. \tag{2.40}$$

If U_2 is infinite dimensional, then $\mathcal{R}(K)$ need not be closed (and similarly for U_1 and $\mathcal{R}(K^*)$).

It is not difficult to establish (e.g., Kress, 1989, p. 226) that

$$\overline{\mathcal{R}(K)}^{\perp} = \mathcal{N}(K^*), \tag{2.41}$$

$$\overline{\mathcal{R}(K^*)}^{\perp} = \mathcal{N}(K), \tag{2.42}$$

$$U_2 = \overline{\mathcal{R}(K)} \oplus \mathcal{N}(K^*), \tag{2.43}$$

and

$$U_1 = \overline{\mathcal{R}(K^*)} \oplus \mathcal{N}(K). \tag{2.44}$$

If K is self-adjoint, then $U_1 = U_2$, $K = K^*$, and $\overline{\mathcal{R}(K)}^{\perp} = \mathcal{N}(K)$.

2.3.5 Compact Linear Operators

Compact operators on an infinite dimensional Hilbert space have a structure that is in many ways similar to that of matrices in a finite dimensional space. The prototypical compact operators are integral operators. We begin with a definition:

Definition 2.3.16 (Compact Operator) *Let $K : U_1 \rightarrow U_2$ be a linear operator between separable Hilbert spaces. K is said to be* **compact** *if for every bounded sequence $\{x_i\}_{i=1}^{\infty}$ in U_1 the sequence $\{Kx_i\}_{i=1}^{\infty}$ has a convergent subsequence.*

A compact operator is necessarily bounded, since otherwise there would exist a bounded sequence $\{x_i\}_{i=1}^{\infty}$ such that $\|Kx_i\| \rightarrow \infty$, and for which $\{Kx_i\}_{i=1}^{\infty}$ has no convergent subsequence. Since an operator is bounded if and only if it is continuous, it follows that a compact operator must be continuous.

The most important properties of compact operators for our purposes are the spectral properties; and with respect to spectral properties, compact operators behave very much like matrices. We state, without proof, the spectral theorem for compact, self-adjoint operators

Theorem 2.3.2 (Spectral Theorem for Compact Self-Adjoint Operators) *Let $K : U \rightarrow U$ be a compact, self-adjoint operator on a Hilbert space. There exists a sequence of vectors $\{\phi_i\}$, such that*

$$(\phi_i, \phi_j) = \delta_{ij};$$

and a bounded sequence of nonzero real scalars $\{\lambda_i\}$, such that, for all i,

$$K\phi_i = \lambda_i\phi_i.$$

Each eigenvalue can correspond to at most a finite number of ϕ_i. Thus we can, without loss of generality, label the λ_i in nonincreasing order of absolute value, so that

$$|\lambda_1| \geq |\lambda_2| \geq \cdots,$$

are the nonzero eigenvalues of the operator K, and the corresponding orthonormal vectors $\{\phi_i\}$ are eigenvectors. If there are infinitely many distinct and nonzero eigenvalues, then these eigenvalues must have zero as an accumulation point.

For any $x \in U$, we have that

$$Kx = \sum_i \lambda_i(x, \phi_i)\phi_i,$$

where by '$\sum_i$' we mean the sum over all (finite or infinitely many) nonzero *eigenvalues.*

Zero may also be an eigenvalue of a compact operator K, and, if so, we denote this eigenvalue of special importance by λ_0. A good source for the spectral theory of compact operators is Kreyszig (1978, Chapter 8).

If a linear operator is compact but not self-adjoint, then the eigenvalues need not be real or even exist. For any compact operator, the subspace spanned by the eigenvectors corresponding to a single eigenvalue can have dimension greater than one, but must be finite dimensional. When discussing compact operators which are not self-adjoint, we will make use of the *singular vector expansion*, which is a natural extension of the singular value decomposition to compact operators in a Hilbert space (Smithies, 1958, Chapter

8). If K is compact, then the operators K^*K and KK^* are compact, self-adjoint, and positive, with the same eigenvalues. It turns out that these eigenvalues are the squares of *singulnr values* of K, defined in a way exactly analogous to the singular values of a matrix.

Theorem 2.3.3 (Singular Vector Expansion) *Let $K : U_1 \rightarrow U_2$ be a compact linear operator between Hilbert spaces. The operators K^*K and KK^* are compact, self-adjoint and positive, each with nonzero eigenvalues*

$$\sigma_1^2 \geq \sigma_2^2 \geq \ldots > 0,$$

where

$$K\psi_i = \sigma_i\phi_i,$$
$$K^*\phi_i = \sigma\psi_i,$$

and hence

$$KK^*\phi_i = \sigma_i^2\phi_i,$$
$$K^*K\psi_i = \sigma_i^2\psi_i.$$

For all i and j,

$$(\phi_i, \phi_j) = \delta_{ij},$$

and

$$(\psi_i, \psi_j) = \delta_{ij}.$$

For any $x \in U_1$, we have

$$Kx = \sum_i \sigma_i(x, \psi_i)\phi_i,$$

where, as in Theorem 2.3.2, we interpret this sum to be over the (finite or infinitely many) singular values.

The *positive* constants $\{\sigma_i\}$ are called the *singular values* of K, and the two orthonormal sequences $\{\phi_i\}$ and $\{\psi_i\}$ are called *singular vectors.* We will, on occasion, find it convenient to refer to $\{\phi_i, \psi_i; \sigma_i\}$ as a *singular system.* It is customary to define the singular values of infinite rank operators to be *positive*, in contrast to the singular values of matrices, which are *nonnegative*, and which can be zero.

2.4 Fredholm Integral Equations of the First Kind in L_2

Let the function $k(x,y) \in L_2\{[0,1] \times [0,1]\}$ be the kernel of a Fredholm integral equation of the first kind:

$$\int_0^1 k(x,y)f(y)dy = g(x). \tag{2.45}$$

If, in (2.45), $f \in L_2$, then it can be shown that $g \in L_2$. The linear operator equation corresponding to (2.45) can be written as $Kf = g$, where $K : L_2[0,1] \rightarrow L_2[0,1]$, given by

$$(Kf)(x) \equiv \int_0^1 k(x,y)f(y)dy, \tag{2.46}$$

is compact (Young, 1988, p.93). If $k(x,y) = \overline{k(y,x)}$, then we say that the kernel is *self-adjoint*; if $k(x,y) = k(y,x)$ then we say $k(x,y)$ is *symmetric.* It is easy to see that linear operators K corresponding to self-adjoint (in particular, real symmetric) kernels are self-adjoint. We will consider next the equation $Kf = g$ where $k(x,y)$ is not necessarily self-adjoint. The self-adjoint case will not be discussed; it follows easily, by means of Theorem 2.3.2 from the more general results of this subsection.

Since K is compact, from Theorem 2.3.3 there exists a set of singular values of K, $\{\sigma_i\}$, an orthonormal basis $\{\phi_i\}$ for $\mathcal{R}(K)$, and an orthonormal basis $\{\psi_i\}$ for $\mathcal{R}(K^*)$. It follows easily from this that

$$k(x,y) = \sum_i \sigma_i \phi_i(x)\psi_i(y). \tag{2.47}$$

We will be prima ily concerned with the case where $k(x,y)$ is real, in which case $\{\phi_i\}$ and $\{\psi_i\}$ can be taken to be real as well. If there are infinitely many terms in the sum (2.47), then by the equal sign we mean that

$$\lim_{N\to\infty} \left\| k(x,y) - \sum_{i=1}^{N} \sigma_i \phi_i(x)\psi_i(y) \right\|^2 = 0. \tag{2.48}$$

Let $\{\tilde{\phi}_j\}$ be an orthonormal basis for $\overline{\mathcal{R}(K)}^{\perp} = \mathcal{N}(K^*)$; and let $\{\tilde{\psi}_j\}$ be an orthonormal basis for $\overline{\mathcal{R}(K^*)}^{\perp} = \mathcal{N}(K)$. Then f and g can be written as

$$f = \sum_i a_i \psi_i + \sum_j \tilde{a}_j \tilde{\psi}_j \tag{2.49}$$

and

$$g = \sum_i b_i \phi_i + \sum_j \tilde{b}_j \tilde{\phi}_j. \tag{2.50}$$

The *Fourier coefficients* a_i, $\tilde{a}_j$, b_i, and $\tilde{b}_j$ are easily shown to be projections of f cr g onto basis functions; for example $a_i = (f, \psi_i)$.

Since

$$Kf = \sum_i \sigma_i \phi_i (f, \psi_i) = \sum_i \sigma_i a_i \phi_i, \tag{2.51}$$

in order for $Kf = g$ to have a solution, we must have that, for all i, $\tilde{b}_i = 0$ and $a_i \sigma_i = b_i$.

Any solution must have the form

$$f = \sum_i \frac{b_i}{\sigma_i} \psi_i + \sum_j \tilde{a}_j \tilde{\psi}_j \equiv \sum_i \frac{b_i}{\sigma_i} \psi_i + h, \tag{2.52}$$

for arbitrary $h \in \mathcal{N}(K)$. For f to be a solution, we must have that

$$\sum_i \frac{b_i}{\sigma_i} \psi_i \in L_2. \tag{2.53}$$

It can be shown that a necessary and sufficient condition for (2.53) is that

$$\sum_i \frac{|b_i|^2}{\sigma_i^2} < \infty. \tag{2.54}$$

Thus we have the following theorem, proved by Picard (1910) for linear Fredholm integral equations of the first kind, and later extended by others (e.g., Groetsch, 1980, pp. 156-157) to arbitrary compact linear operators.

Theorem 2.4.1 (Picard) *Let K be compact, with singular system $\{\phi_i, \psi_i; \sigma_i\}$, and let $g \in L_2$ be given. There exists a function f such that $Kf = g$ if and only if*

1. $\sum_{i=1}^{\infty} \frac{|(g,\phi_i)|^2}{\sigma_i^2} < \infty$, *and*

2. $(g, u) = 0$ *for all u such that $K^*u = 0$.*

2.4.1 Existence and Uniqueness of Solutions of Linear Operator Equations

We summarize next the conditions under which a solution to a linear operator equation of the first kind exists, and the conditions under which it is unique.

A solution to $Kf = g$ exists if and only if $g \in \mathcal{R}(K)$. If a solution f exists, then it is unique if and only if $\mathcal{N}(K) = \{0\}$. If more then one solution exists, then the difference between any two solutions is in $\mathcal{N}(K)$, and therefore as a consequence of the projection theorem, there exists exactly one solution $f_1 \in \overline{\mathcal{R}(K^*)} = \mathcal{N}(K)^{\perp}$. Let $g \in \mathcal{R}(K)$, and let f_1 be the unique $f_1 \perp \mathcal{N}(K)$ such that $Kf_1 = g$. The set of all solutions to $Kf = g$ is given by

$$\mathcal{F} = \{f = f_1 + f_2 | Kf_1 = g, f_1 \in \mathcal{N}(K)^{\perp}, f_2 \in \mathcal{N}(K)\}. \tag{2.55}$$

By the Pythagorean theorem,

$$\|f\|^2 = \|f_1\|^2 + \|f_2\|^2. \tag{2.56}$$

Since for any solution f, $\|f\| \geq f_1$, $f_1 \in \mathcal{F}$ is the *minimum norm solution.*

2.4.2 Infinite Rank Compact Operator Equations of the First Kind are Ill-Posed

In Chapter 1, we defined what it means for an equation to be ill-posed, and we provided some intuition for why integral equations of the first kind are often ill-posed. We now use the theory outlined in the present chapter to build on this intuition in a more general context.

The Nature of the Spectrum of Infinite Rank Compact Linear Operator Equations

Let $K : \mathcal{H} \to \mathcal{H}$ be a compact, positive, self-adjoint linear operator on a separable Hilbert space. Then K has a finite or countable spectrum of positive eigenvalues (Theorem 2.3.2). If K is of infinite rank, then K has infinitely many nonzero eigenvalues, and these eigenvalues must have zero as an accumulation point. In particular, $K = T^*T$ is compact, positive, and self-adjoint for any bounded linear operator T, and the nonzero eigenvalues of K are the squares of the singular values of T (Theorem 2.3.3). Therefore if K is compact, of infinite rank, but not necessarily self-adjoint, then the singular values of K (eigenvalues, if $K = K^*$) will have zero as an accumulation point. It can be shown that K cannot have a bounded inverse, and hence that the linear operator equation $Kf = g$ is ill-posed.

Because of this, a necessary condition for this equation to have a solution is that the Fourier coefficients in the expansion of g must decrease in absolute values sufficiently rapidly as the corresponding singular values approach zero, a result made precise by Picard's Theorem (2.4.1).

$\mathcal{R}(K) \neq \overline{\mathcal{R}(K)}$ if K is Compact and of Infinite Rank: Some Implications for Ill-Posedness

The second part of Theorem 2.4.1 states that $g \perp \mathcal{N}(K^*)$ (or, if K is self-adjoint, $g \perp \mathcal{N}(K)$). This ensures that $g \in \overline{\mathcal{R}(K)}$. But, as we shall see in this subsection, if K has infinitely many non-zero eigenvalues, then $\mathcal{R}(K)$ is not closed. Therefore, the first condition in Theorem 2.4.1 is required in order to demonstrate that $g \in \mathcal{R}(K)$. If g is observed with error and/or represented on a computer, the fact that $\mathcal{R}(K) \neq \overline{\mathcal{R}(K)}$ has important consequences, as can be seen by the following result of Strand (1974).

Theorem 2.4.2 (Strand, 1974, p. 801) *Let K be compact, with eigenvalues $\{\lambda_i\}_{i=1}^{\infty}$, where*

$$|\lambda_1| \geq |\lambda_2| \geq \ldots \geq 0. \tag{2.57}$$

Assume that infinitely many of these eigenvalues are nonzero. Let $g \in \mathcal{R}(K)$ and $\epsilon > 0$ be arbitrary. Then there exists a function $\tilde{g} \in \mathcal{H}$ such that:

1. *$\tilde{g} \notin \mathcal{R}(K)$,*

2. *$\tilde{g} \perp \mathcal{N}(K^*)$, and*

3. *$\|g - \tilde{g}\| < \epsilon$.*

By definition $\mathcal{R}(K)$ is dense in its closure, $\overline{\mathcal{R}(K)}$. Theorem 2.4.2 states that $\overline{\mathcal{R}(K)} - \mathcal{R}(K)$ is dense in $\mathcal{R}(K)$.

This result provides one way of understanding what it means for an integral equation to be ill-posed. One can always find a perturbation of the right hand side of arbitrarily small norm which changes a solvable integral equation into an equation with *no solution.*

Actually, the consequences of ill-posedness for the numerical solution of integral equations of the first kind is somewhat different. When an equation with a reasonably smooth kernel is discretized for solution on a computer, the resulting system of algebraic equations has many small eigenvalues, and hence is very nearly singular. The exact right hand side $g(x)$ and a representation of $g(x)$ on a computer will always be slightly different, because of inevitable roundoff and discretization error. The solution of the matrix equation corresponding to this slightly perturbed right hand side will very likely exist, however it will often be very different from the exact solution f.

2.5 Probability Theory

One empirical basis for mathematical probability lies in the observation of the *long range relative frequency* of 'favorable' events in the repetition of a random experiment. The theory originated with the investigation of games of chance in the seventeenth century, where a set of *elementary* outcomes were treated as equally likely. A. N. Kolmogorov provided an axiomatic foundation for probability in 1933, making use of the theory of measure and integration. The present section is a very brief outline of the principal ideas of probability theory, along with the definitions and some important properties of certain probability distributions. There are many introductory books at various levels which the reader can turn to for details; the present discussion follows Tucker (1967).

2.5.1 Probability Spaces

In order to have a rigorous discussion of probability, it is necessary to define a set of possible outcomes of a random phenomenon, called a *sample space.*

Definition 2.5.1 (Sample Space, Elementary Event) *A* **sample space** Ω *is a set of elements or points* $\omega \in \Omega$*, called* **elementary events,** *each of which is a possible outcome of a random phenomenon under consideration.*

Probability is a set function which associates subsets of Ω with numbers in the unit interval. If the sample space is uncountable, then it is necessary to restrict this set function to a class of subsets which satisfies the properties of a *σ-field* :

Definition 2.5.2 (σ-field) *A set of subsets $\mathcal{S}$ of Ω is called a* **σ-field** *if*

1. *For every $A \in \mathcal{S}$, $A^c \in \mathcal{S}$,*

2. *if $A_1, A_2, \ldots, A_n, \ldots$ is a countable sequence of elements of $\mathcal{S}$, then $\cup_n A_n \in \mathcal{S}$, and*

3. *$\emptyset \in \mathcal{S}$.*

Subsets $A \in \mathcal{S}$ are called *events.* The pair $(\Omega, \mathcal{S})$ is sometimes called a *measurable space.*

In order for a set function to be a *probability* or *probability measure,* this function must be as defined in the following:

Definition 2.5.3 (Probability) *A* **probability** *P is a normed measure over a measurable space $(\Omega, \mathcal{S})$; that is P is a real-valued function which assigns to every $A \in \mathcal{S}$ a number $P(A)$ such that*

1. *$P(A) \geq 0$ for every $A \in \mathcal{S}$,*

2. *$P(\Omega) = 1$, and*

3. *if $\{A_n\}_{n=1}^{\infty}$ is any countable sequence of disjoint events, then*

$$P(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n).$$

A *probability space* can now be defined.

Definition 2.5.4 (Probability Space) *A* **probability space** *is a triple $(\Omega, \mathcal{S}, P)$, where Ω is a sample space, $\mathcal{S}$ is a σ-field of subsets of Ω, and P is a probability measure on the measurable space $(\Omega, \mathcal{S})$.*

2.5.2 Random Variables and Probability Distributions

Often one cannot, or does not want to, observe directly $\omega \in \Omega$. Instead, what is measured or studied is the value of a function on the sample space. Such an $\mathcal{S}$-measurable function is called a *random variable.*

Definition 2.5.5 (Random Variable) *Let $(\Omega, \mathcal{S}, P)$ be a probability space. A* **random variable,** *$X : \Omega \rightarrow \mathcal{R}$ is a real-valued $\mathcal{S}$-measurable function. That is, for every real number x,*

$$\{\omega \in \Omega | X(\omega) \leq x\} \in \mathcal{S}.$$

We will adopt the convention of using the notation X both for the random variable X and for a value of this random variable $X(\omega)$. We will denote the event $\{\omega \in \Omega | X(\omega) \leq x\}$ by $\{X \leq x\}$, and its probability by $P(X \leq x)$.

Associated with every random variable X is a *distribution function* (also called a *cumulative distribution function*, a *cdf*, or simply a *distribution*), $F_X(x)$, which gives the probability that X is less than or equal to any real number x.

Definition 2.5.6 (Distribution Function) *If X is a random variable, its* **distribution function** F_X *is defined by*

$$F_X(x) \equiv P(X \leq x).$$

It can be shown that F_X is monotone nondecreasing, right-continuous, and that

$$\lim_{x \rightarrow -\infty} F_X(x) = 0,$$

and

$$\lim_{x \rightarrow \infty} F_X(x) = 1.$$

It is straightforward to extend the definition of distribution to the *joint distribution* of several random variables.

Definition 2.5.7 (Multivariate Distribution Function) *Let $X_1, \ldots, X_n$ be random variables, where $n \geq 1$. The* joint distribution function *of $\{X_1, \ldots, X_n\}$ is defined by*

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) \equiv P\left(\cap_{i=1}^n \{X_i \leq x_i\}\right),$$

where $-\infty < x_i < \infty$, for $1 \leq i \leq n$.

A related concept is the *probability density*, defined in the univariate case as follows:

Definition 2.5.8 (Probability Density) *Let $F_X(x)$ be an absolutely continuous distribution function. Then*

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt$$

for some function $f_X(t)$, called the **probability density** *of the random variable X.*

A random variable X which has a density

$$f_X(x) = \frac{dF_X(x)}{dx} \tag{2.58}$$

is said to be *continuous.* A random variable which takes on values in a finite or countable set is said to be *discrete.* The notions of distribution and density can be generalized to random variables which assume values in more general spaces.

Corresponding to joint distribution functions, there can be *joint probability densities.* We will only need to make use of bivariate densities. For example, let X and Y be two continuous random variables with joint density $f_{X,Y}(x,y)$. The univariate *marginal density* of either random variable is obtained by 'integrating out' the other variable, for example

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy \tag{2.59}$$

is the marginal density of X.

2.5.3 Expectation and Moments

Mathematical *expectation* is a linear functional of a random variable which models the empirical fact of *long run averages.*

Definition 2.5.9 (Expectation) *The* **expectation** *of a random variable X, denoted $E(X)$, is defined to be the Lebesgue integral of X with respect to the probability measure P,*

$$E(X) \equiv \int X P(d\omega),$$

provided that this integral exists.

Usually it is more convenient to write this integral either as a Lebesgue-Stieltjes integral with respect to the distribution of a random variable, or else as an integral involving a probability density. If X has a density $f_X(x)$, then the following are equal:

$$E(X) = \int X P(d\omega) = \int_{-\infty}^{\infty} x dF_X(x) = \int_{-\infty}^{\infty} x f_X(x) dx. \tag{2.60}$$

The expectations of X^n are of particular importance. When these expectations exist, they are called *moments* of the random variable X.

Definition 2.5.10 (Moments, Central Moments) *The nth* **moment** *of a random variable X is defined to be the expectation $E(X^n)$, provided that this expectation exists. If $E(X) = \mu$, then the nth* **central moment** *is defined to be $E\{(X-\mu)^n\}$.*

The *mean* of a random variable X is $E(X)$, and it is usually denoted μ. If X is a random variable with mean μ, then the *variance* of X, usually denoted σ^2, is $E[(X-\mu)^2]$.

Expectation is a linear functional; that is, if X and Y are any random variables for which $E(X)$ and $E(Y)$ exist, and α and β are constants, then

$$E(\alpha X + \beta Y) = \alpha E(X) + \beta E(Y). \tag{2.61}$$

2.5.4 Conditional Probability and Independence

Intuitively, if we toss a coin twice, the result of the first toss has 'no effect' on the result of the second toss. We would say that these tosses are 'independent'. This provides a motivation for the concept of *independence* in probability theory.

Independence is, of course, a special situation. For example, one might ask how one would estimate the probability of drawing the ace of spades as a second card given each of the three following situations:

1. that the first card drawn is the ace of spades,

2. that the first card drawn is the eight of hearts, or

3. no information on the first card.

This leads naturally to the notion of *conditional probability.*

Definition 2.5.11 (Conditional Probability) *If $(\Omega, \mathcal{S}, P)$ is a probability space, and $A, B \in \mathcal{S}$, with $P(A) > 0$, then*

$$P(B|A) \equiv \frac{P(A \cap B)}{P(A)}$$

is called the **conditional probability of B given A.**

It is easy to show that $P(\cdot|A)$ is a probability measure.

In the case of discrete random variables, it is easy to define $P(X = x|Y = y)$ if $P(Y = y)$ is not zero. In the case of continuous random variables, we always have that $P(Y = y) = 0$, and conditional probability (as well as *conditional expectation*, to be defined below) raises measure theoretic problems. These are treated rigorously using the Radon-Nikodym theorem (e.g., Chung, 1974, Chapter 9). It is not necessary to discuss these technical issues here, as long as we use certain basic properties.

If X and Y are random variables with a joint density $f_{X,Y}(x,y)$, we will define the *conditional density* of X given Y, $f_{X|Y}(x|y)$, in terms of which we can compute conditional probabilities and conditional expectations.

Definition 2.5.12 (Conditional Density) *Let X and Y be continuous random variables, with marginal probability densities $f_X(x)$ and $f_Y(y)$, and joint density $f_{X,Y}(x,y)$. Then, the* **conditional density** *of the random variable X given that the random variable Y equals y is*

$$f_{X|Y}(x|y) \equiv \frac{f_{X,Y}(x,y)}{f_Y(y)},$$

provided that $f_Y(y) \neq 0$.

An expectation with respect to a conditional distribution is called a *conditional expectation.* We assume that (X,Y) is continuous, so that $f_{X|Y}(x|y)$ exists.

Definition 2.5.13 (Conditional Expectation) *Let (X,Y) be continuous random variables, and assume that the conditional density $f_{X|Y}(x|y)$ exists. Then the* **conditional expectation** *of X given that $Y = y$ is defined to be*

$$E(X|y) = E(X|Y = y) = \int x f_{X|Y}(x|y)dx,$$

provided that this integral exists.

We write the random variable $E(X|Y)$ by substituting Y for y in the right hand side of the defining equation.

More generally, we have the following properties of $E(X|Y)$. Let X_1, X_2 and Y be random variables, let g_1 and g_2 be functions such that $E(|g_1(X_1)|) < \infty$ and $E(|g_2(X_2)|) < \infty$, and let α and β be constants. Then

$$E[\alpha g_1(X_1) + \beta g_2(X_2)|Y] = \alpha E[g_1(X_1)|Y] + \beta E[g_2(X_2)|Y], \tag{2.62}$$

$$E\{E[g_1(X_1)|Y]\} = E[g_1(X_1)], \tag{2.63}$$

$$E[g_2(Y)g_1(X_1)|Y] = g_2(Y)E[g_1(X_1)|Y], \tag{2.64}$$

and also

$$P(A|Y) = E(1_A|Y), \tag{2.65}$$

where 1_A is the *indicator random variable* corresponding to the event A, defined by

$$1_A(\omega) \equiv \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A. \end{cases} \tag{2.66}$$

If two events are such that $P(A|B) = P(A)$, or equivalently $P(A \cap B) = P(A)P(B)$, then the events A and B are said to be *independent.* More generally, we have the following definition:

Definition 2.5.14 (Independent Events) *Let $\mathcal{B} = \{B_\alpha, \alpha \in I\}$ be a set of events. These events are said to be* **independent** *if for every positive integer n and every n distinct elements $\alpha_1, \ldots, \alpha_n$ in the indexing set I, we have that*

$$P(B_{\alpha_1} \cap \ldots \cap B_{\alpha_n}) = \prod_{i=1}^{n} P(B_{\alpha_i}).$$

If all events involving X are independent of those involving Y, i.e. $\{X \in A\}$ and $\{Y \in B\}$ are independent for all sets A and B, then X and Y are said to be *independent random variables.* In this case, $F_{X,Y}(x,y) = F_X(x)F_Y(y)$, and, if X and Y are continuous, $f_{X,Y}(x,y) = f_X(x)f_Y(y)$, and $f_{X|Y}(x|y) = f_X(x)$. More generally, we have that:

Definition 2.5.15 (Independent Random Variables) *Let $\{X_\alpha, \alpha \in I\}$ be a family of random variables. These random variables are said to be* **independent** *if, for every positive integer n and every n distinct elements $\alpha_1 \ldots \alpha_n$ in the indexing set I, we have that*

$$F_{X_{\alpha_1},\ldots,X_{\alpha_n}}(x_1, \ldots, x_n) = \prod_{i=1}^{n} F_{X_{\alpha_i}}(x_i).$$

If X and Y are independent random variables, then, for any functions $h_1(x)$ and $h_2(y)$ for which $E[|h_1(X)|] < \infty$, $E[|h_2(Y)|] < \infty$

$$E[h_1(X)h_2(Y)] = E[h_1(X)]E[h_2(Y)], \tag{2.67}$$

and

$$E(X|Y) = E(X). \tag{2.68}$$

We also will be making use of the following results for independent random variables:

$$P[X \leq h(Y)] = E\{F_X[h(Y)]\} = \int F_X[h(y)]f_Y(y)dy, \tag{2.69}$$

and the variance of $X + Y$ is the sum of the variances of X and Y.

2.5.5 Some Distribution Theory for Statistics

We will make extensive use of several special continuous probability distributions of importance to statistics. In this section, we define those probability distributions which we will use in this thesis, and we state some important properties and relations.

All of these distributions are related, directly or indirectly, to the *standard normal distribution*, denoted $\Phi(x)$ and defined by

$$\Phi(x) \equiv \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt. \tag{2.70}$$

The corresponding *standard normal density* is

$$\phi(x) \equiv \frac{d\Phi(x)}{dx} = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}. \tag{2.71}$$

If X has a standard normal distribution, we indicate this by $X \sim \mathrm{N}(0,1)$, where '$\sim$' is read 'is distributed as' and the arguments of N indicate that X has a mean of zero and a variance of one.

The six densities which we will use are defined below, where we adopt the convention of separating,by a semicolon, parameters which define special cases of a class of distributions, from the possible value x of the random variable.

$$\begin{aligned}
f_1(x;\mu,\sigma^2) &\equiv \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right) && (2.72)\\
f_2(x;\nu) &\equiv \frac{x^{\nu/2-1}e^{-x/2}}{\Gamma(\nu/2)2^{\nu/2}} && (2.73)\\
f_3(x;\lambda_1,\lambda_2) &\equiv \frac{\Gamma(\lambda_1+\lambda_2)}{\Gamma(\lambda_1)\Gamma(\lambda_2)}x^{\lambda_1-1}(1-x)^{\lambda_2-1} && (2.74)\\
f_4(x;\nu_1,\nu_2) &\equiv \frac{\Gamma[(\nu_1+\nu_2)/2]}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)}\left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} && (2.75)\\
&\quad \cdot\frac{x^{\nu_1/2-1}}{[1+(\nu_1/\nu_2)\,x]^{(\nu_1+\nu_2)/2}}\\
f_5(x;\nu) &\equiv \frac{\Gamma[(\nu+1)/2]}{\sqrt{\nu\pi}\Gamma(\nu/2)}(1+x^2/\nu)^{-(\nu+1)/2} && (2.76)\\
f_6(x;\nu,\delta) &\equiv \sqrt{\frac{2\pi}{\nu}}\left[\Gamma(\nu/2)2^{\nu/2-1}\right]^{-1} && (2.77)\\
&\quad \cdot\int_0^\infty \phi\left(\frac{tx}{\sqrt{\nu}}-\delta\right)\phi(t)t^\nu dt
\end{aligned}$$

The following are listed below for each of these densities:

- Notation for the corresponding distribution,
- The interval over which the density is nonzero (the *support*), and
- The mean and variance, if necessary:

1. f_1 is the *normal density*, with distribution denoted $\mathrm{N}(\mu,\sigma^2)$, with support the real line, and with mean μ and variance σ^2;
2. f_2 is the *χ^2 density with ν degrees of freedom*, with distribution denoted χ^2_ν, with support the positive reals, and with mean ν and variance 2ν;
3. f_3 is the *Beta density*, with distribution denoted Beta (λ_1,λ_2), with support [0,1], and with mean $\lambda_1/(\lambda_1+\lambda_2)$;
4. f_4 is the F *density with ν_1 and ν_2 degrees of freedom*, with distribution denoted F_{ν_1,ν_2}, and with support the positive reals;
5. f_5 is the *t density with ν degrees of freedom*, with distribution denoted T_ν, with support the real line, and with mean zero;
6. f_6 is the *noncentral t density with ν degrees of freedom and noncentrality parameter* δ, with distribution function denoted $T_\nu(\delta)$, and with support the real line.

In addition, we note that if $Z \sim \mathrm{N}(\mu,\sigma^2)$, then if n is an integer,

$$E[(Z-\mu)^n] = 0 \qquad (2.78)$$

for n odd, and

$$E[(Z-\mu)^n] = \frac{n!}{(n/2)!}\frac{\sigma^n}{2^{n/2}} \tag{2.79}$$

for n even.

Let $\{X_i\}_{i=1}^n$ denote a sequence of random variables. We call such a sequence a *random sample*. If the X_i are independent and identically distributed, we use the notation *iid*. The sample mean and variance are defined as follows:

Definition 2.5.16 (Mean and Variance of a Sample) *Let $\{X_i\}_{i=1}^n$ be a random sample. The* **sample mean** *and* **sample variance** *are*

$$\bar{X} \equiv \sum_{i=1}^{n} X_i/n \tag{2.80}$$

and

$$S^2 \equiv \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1), \tag{2.81}$$

respectively.

If the $\{X_i\}$ are *iid* normally distributed, then the following important result holds:

Theorem 2.5.1 (Distribution of the Mean and Variance of a Normal Sample) *Assume $\{X_i\}_{i=1}^n$ are iid* $\mathrm{N}(\mu,\sigma^2)$. *The sample mean and variance, $\bar{X}$ and S^2 respectively, are independent,*

$$\bar{X} \sim \mathrm{N}(\mu, \sigma^2/n),$$

and

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

The following lemmas relate some of the random variables whose distributions were defined above. These results are important, and the proofs are omitted here. In a more leisurely presentation, many of these 'results' would be used as defining the corresponding random variables, and the distributions would be derived from those definitions.

Lemma 2.5.1 (Sums of Normal Random Variables) *If $X \sim \mathrm{N}(\mu_1,\sigma_1^2)$ and $Y \sim \mathrm{N}(\mu_2,\sigma_2^2)$, where X and Y are independent, and $a, b \in \mathcal{R}$ are arbitrary constants, then*

$$aX + bY \sim \mathrm{N}(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2).$$

and

$$aX + b \sim \mathrm{N}(a\mu_1 + b, a^2\sigma_1^2).$$

Lemma 2.5.2 (Sums of Squares of Normal Random Variables) *If $\{X_i\}_{i=1}^n$ are iid* $\mathrm{N}(0,1)$, *then*

$$\sum_{i=1}^{n} X_i^2 \sim \chi_n^2.$$

Lemma 2.5.3 (Sums and Quotients of χ^2 Random Variables) *Let $X \sim \chi^2_{\nu_1}$ and $Y \sim \chi^2_{\nu_2}$, where X and Y are independent. Define the three random variables $Z_1 \equiv X + Y$, $Z_2 \equiv X/(X+Y)$, and $Z_3 \equiv (X/\nu_1)/(Y/\nu_2)$. Then*

1. Z_1 *and* Z_2 *are independent,*

2. $Z_1 \sim \chi^2_{\nu_1+\nu_2}$*,*

3. $Z_2 \sim$ Beta $(\nu_1/2, \nu_2/2)$*, and*

4. $Z_3 \sim \mathrm{F}_{\nu_1,\nu_2}$.

Lemma 2.5.4 (Student's t Distribution) *If* $Z \sim \mathrm{N}(0,1)$ *and* $Y \sim \chi^2_\nu$*, where* Z *and* Y *are independent, then*

$$\frac{Z+\delta}{\sqrt{Y/\nu}} \sim T_\nu(\delta),$$

and

$$\frac{Z}{\sqrt{Y/\nu}} \sim T_\nu.$$

It is customary to use capitals for random variables and Greek letters for parameter values. There are exceptions often due to ancient conventions.

2.5.6 Some Limit Theorems

Two important limit theorems concerning the behavior of the average $\bar{X}$ of n *iid* random variables $X_1, X_2, \ldots, X_n$ are the *Law of Large Numbers* and the *Central Limit Theorem.* In order to state these, we need to define two forms of convergence for a sequence of random variables.

Definition 2.5.17 (Convergence in Probability) *Let* $\{X_n\}$ *be a sequence of random variables. We say that* $\{X_n\}$ **converges in probability** *to* X *if for every* $\epsilon > 0$

$$P(|X_n - X| \geq \epsilon) \to 0$$

as $n \to \infty$*, and we write* $X_n \xrightarrow{\mathrm{P}} X$*.* X *can be either a constant or a random variable.*

Definition 2.5.18 (Convergence in Distribution) *If* X *is a random variable with distribution* $F_X(x)$*, and if* $\{X_n\}$ *is a sequence of random variables with distributions* $\{F_{X_n}(x_n)\}$*, then we say that* X_n **converges in distribution** *to* X*, and we write* $F_{X_n} \xrightarrow{\mathcal{D}} F_X$*, if for all points of continuity* x *of* $F_X(x)$

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x).$$

The Law of Large Numbers states that the average of *iid* random variables with finite mean converges in probability to that mean; in other words, that expectation has been properly defined to model long run averages.

Theorem 2.5.2 (Law of Large Numbers) *Let* $\{X_i\}$ *be a sequence of iid random variables with mean* μ*, and let* $\bar{X}_n$ *be given by*

$$\bar{X}_n \equiv \sum_{i=1}^{n} X_i/n.$$

Then $\bar{X}_n \xrightarrow{\mathrm{P}} \mu$*.*

If the variance is finite, then the Central Limit Theorem tells us more, i.e. that $\bar{X}_n$ is approximately normally distributed with mean μ and variance σ^2/n.

Theorem 2.5.3 (Central Limit Theorem) *Let $\{X_i\}$ be a sequence of iid random variables with mean μ and variance $\sigma^2 < \infty$. Then*

$$\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \xrightarrow{\mathcal{D}} \mathrm{N}(0,1).$$

We can now immediately derive several important results involving some of the distributions introduced in the previous subsection:

Lemma 2.5.5 *Let $A_n \sim \chi^2_n$, and let T_n and $T_n(\delta)$ denote Student t random variables. Then,*

$$A_n = n + \sqrt{2n}U_n$$

where $U_n \xrightarrow{\mathcal{D}} \mathrm{N}(0,1)$ as $n \to \infty$. Also, as $n \to \infty$, we have the following:

1. $A_n/n \xrightarrow{\mathrm{P}} 1$,

2. $T_n \xrightarrow{\mathcal{D}} \mathrm{N}(0,1)$, *and*

3. $T_n(\delta) \xrightarrow{\mathcal{D}} \mathrm{N}(\delta,1)$.

2.6 A Decision-Theoretic Approach to Estimation and Hypothesis Testing

Statistical decision theory, a theory of decision making in the presence of uncertainty, extends and unifies much of classical statistical inference. Statistical decision theory was first studied extensively by Abraham Wald in the 1940's. Two useful texts which were consulted in the preparation of this section are Chernoff and Moses (1957) and Berger (1985).

In the present section, after introducing some of the ideas of statistical decision theory we show how the classical statistical problems of *estimation* and *hypothesis testing*, which will concern us in this thesis, can be regarded as special cases of this general theory. Finally, we will illustrate each of these two classes of problems with an example.

2.6.1 Decision-Making Under Uncertainty

A simple decision-making problem under uncertainty can be modeled as follows. Given a set $\mathcal{A}$ of possible *actions* $a \in \mathcal{A}$, a choice of action, or *decision*, has to be made. The consequence of this decision depends on the unknown *state of nature* $\theta \in \Theta$. Thus, for each action a and state θ, there is a consequence (which may depend in part on chance).

For any individual whose preferences satisfy some modest assumptions, consequences can be represented by a real valued *utility* measure which has the following properties (e.g., Chernoff and Moses, 1957, Chapter 4):

1. The higher utility goes to the preferred consequence

2. If the consequences have random components, then the utility for a random situation can be evaluated as the mathematical expectation of the corresponding utilities, even though we are not involved in a long run average situation.

Because statisticians prefer to measure how much they lose because of ignorance, it is conventional to use losses in place of utilities, where we can define loss as negative utility. Thus the consequences can be represented by a loss function $L(\theta, a)$. Now we are in the position of having a game of a statistician with nature. Nature picks $\theta \in \Theta$, and the statistician *in ignorance of* θ picks $a \in \mathcal{A}$. The game (in *normal form*) is represented by $L(\theta, a)$. By performing an *experiment*, the statistician has an opportunity to obtain information about the state of nature. Unfortunately, most experiments are less than fully informative; they do not tell us θ, but rather they provide data in the form of a random variable X which takes on values in $\mathcal{X}$, the distribution of which depends on the state. The help that we get from the data depends on the extent to which the distribution of the data depends on θ. Having observed the data, the statistician must incorporate that information in his decision making. He does so by selecting his action as a function of X. Thus we have the *decision function* $\delta : \mathcal{X} \longrightarrow \mathcal{A}$ or $\delta(X) = A$, where the resulting action A is ordinarily random, since it depends on the data X. Occasionally we will use the terminology of game theory and refer to a decision function as a *strategy*. The consequence of using δ when the state of nature is θ is measured by the expected loss as a function of θ, called the *risk*:

$$R(\theta, \delta) \equiv E_\theta[L(\theta, A)] = E_\theta[L(\theta, \delta(X))], \tag{2.82}$$

where the subscript represents expectation with respect to the distribution of X, when θ is the state of nature.

By introducing the experiment we have changed our relatively simple problem into a more complicated looking problem of the same form: where the statistician chooses the decision function while nature still chooses the state. However, we have lost nothing and possibly gained something, because among our decision functions are those which ignore the data. Typically, with informative experiments, we can do better than before.

2.6.2 Admissibility and Bayes Risk

Let X be a random variable, with distribution function $F_X(x; \theta)$. On the basis of X, we choose an action by means of the decision function $\delta : \mathcal{X} \longrightarrow \mathcal{A}$. We would like to choose a δ which makes $R(\theta, \delta)$ small for all $\theta \in \Theta$. Of course, we may have two decision functions δ_1 and δ_2, for which $R(\theta, \delta_1) < R(\theta, \delta_2)$ for some values of θ, but for which $R(\theta, \delta_1) > R(\theta, \delta_2)$ for some other values of θ. In this case, we cannot say which of δ_1 and δ_2 is preferable on the basis of $R(\theta, \delta)$ alone. However, if $R(\theta, \delta_1) < R(\theta, \delta_2)$ for *all* θ, then δ_1 is clearly preferable. A decision function δ_* *dominates* a decision function δ if $R(\theta, \delta_*) \leq R(\theta, \delta$ for all θ and $R(\theta, \delta_*) < R(\theta, \delta)$ for some θ. A decision function is *inadmissible* if it is dominated by some other strategy, and *admissible* otherwise.

It is natural for an *optimizer* to insist that we select only admissible strategies, but that rarely solves the dilemma of how to select a decision function. Occasionally, however, we do have a situation where a certain type of problem recurs frequently, and through past experience we learn that the θ values behave like random variables with a known probability distribution $\Pi(\theta)$; for simplicity of presentation we will take θ to be continuous, with density $\pi(\theta)$. In those cases we can evaluate our decision function by minimizing the

Bayes Risk

$$r(\delta) \equiv E\{E_\theta[L(\theta,\delta)]\} = \int E_\theta[L(\theta,\delta)]\pi(\theta)d\theta. \tag{2.83}$$

A decision function which minimizes the Bayes risk is called a *Bayes strategy.*

Two facts concerning Bayes strategies are of particular importance. The first of these is a theorem which states that under suitable regularity conditions every admissible strategy is a Bayes strategy or a limit of Bayes strategies. The second asserts that it is often relatively easy to find a Bayes strategy by using the data X to replace the *prior* distribution π by a *posterior* distribution

$$\pi^*(\theta) = \frac{\pi(\theta)f_X(X;\theta)}{\int f_X(X;\tilde{\theta})\pi(\tilde{\theta})d\tilde{\theta}}, \tag{2.84}$$

where $f_X(x;\theta)$ is the density of X. Then we select $A = \delta(X)$ as a value a which minimizes the posterior risk, conditional on the data X,

$$E^*[L(\theta,a)] = \int L(\theta,a)\pi^*(\theta)d\theta, \tag{2.85}$$

where θ is a random variable with posterior distribution π^*. Here X is present implicitly, since $\pi^*(\theta)$ depends on X.

2.6.3 Philosophies of Inference

The data X may have reduced uncertainty due to the unknown state of nature, but it is seldom the case that there exists a decision function $\delta(X)$ which dominates all others. There are two primary (and several secondary) schools of thought on how to select 'good' decision functions when such a selection cannot be done on the basis of $R(\theta,\delta)$ alone: the *frequentist* and *Bayesian* philosophies of inference.

The term 'frequentist' is a misnomer. It suggests the use of long run average which is not relevant. A distinction between the two schools is, rather, that the frequentist tries to be objective while the Bayesian is subjective. There is a theorem, very much like the theorem that gives rise to utility, that states that if a decision maker acts *coherently* on related problems, he must be acting as though he has a prior probability (Ferguson, 1967, pp. 17-22). Using conditional probability, we can show how this prior changes with additional information, but this theorem does not say where the prior comes from. A weakness of the Bayesian philosophy is that when we replace our vague feelings about the prior by some approximation, that approximation may carry more information than we really feel we have. The solution based on the approximation may be far from an approximation to the solution, and there is a resulting lack of *robustness.* The fact that Bayesians are subjective is also perceived by many to be a weakness. On the other hand, frequentists try to find a procedure which will not do poorly no matter what the true state of nature. A shortcoming of this approach is that whatever criterion that a frequentist might suggest, it will either be equivalent to a Bayesian criterion, or else it will lead to paradoxes because of the theorem on coherent decision making. If the criterion is equivalent to a Bayesian one, then the prior is likely to have been chosen for mathematical convenience, and it might not be a reasonable reflection of prior experience.

2.6.4 Estimation

In problems of statistical inference, functions of the observed data X are usually called *statistics*, and the state of nature θ is called a *parameter* in a *parameter space* Θ. Two broad classes of statistical problems are problems of *estimation* and *hypothesis testing*, and we briefly consider estimation next.

A decision problem for which knowledge of θ would suggest that the best action to take is $g(\theta)$ is called an *estimation* problem, and the corresponding decision function is called an *estimator*. Typically, for such a problem the loss will depend on how close a is to $g(\theta)$. Ordinarily a smooth loss function can then be approximated by *squared-error loss*, i.e.,

$$L(\theta, a) = (a - g(\theta))^2, \tag{2.86}$$

In those cases we want a decision function δ for which the *mean square error*

$$R(\theta, \delta) = E_\theta[(\delta(X) - g(\theta))^2] \tag{2.87}$$

is small. Let the expected value of an estimator $\delta(X)$ be denoted $\mu_\delta(\theta)$. Then the risk $R(\theta, \delta)$ can be written as a sum of two terms

$$R(\theta, \delta) = E_\theta[(\delta(X) - \mu_\delta(\theta))^2] + [g(\theta) - \mu_\delta(\theta)]^2. \tag{2.88}$$

The first term in (2.88) is the variance of the estimator $\delta(X)$, and the second term is the square of the *bias* of $\delta(X)$.

Admissibility does not do much to help reduce the class of available estimators in this case. To see this, let $\delta_0 \equiv \theta_0$, for any value θ_0 of θ, and note that, for the loss (2.86), $R(\theta_0, \delta_0) = 0$. Although δ_0 makes no use of the data X, it is at least as good as *any* decision function when the true parameter is θ_0.

For the Bayesian, the Bayes strategy for squared-error loss would be the mean of the posterior distribution of $g(\theta)$. A non-Bayesian can eliminate ridiculous strategies such as the guess $\theta = \theta_0$ above by restricting the class of decision functions to be considered. Often this is done by restricting consideration to unbiased estimators. An estimator δ is an *unbiased estimator* of $g(\theta)$ if

$$E_\theta[\delta(X)] = g(\theta) \tag{2.89}$$

for all θ.

Among unbiased estimators for a particular estimation problem, one can often determine an estimator $\delta_U(X)$ which minimizes the risk (2.88). Since squared-error loss for an unbiased estimator is the same as variance, we call such a $\delta_U(X)$ a *minimum variance unbiased estimator.*

2.6.5 Hypothesis Testing

A decision problem with only two actions is called a *hypothesis testing problem* for reasons that will become clear shortly. We can divide up the class Θ of states of nature into two sets: one set, Θ_0, for which one of the actions, say a_0, is the best action, and another set, $\Theta_1 = \Theta - \Theta_0$, for which the other action, say a_1, is the best action. Thus, we can identify a_0 with accepting the hypothesis

$$H_0 : \theta \in \Theta_0, \tag{2.90}$$

and a_1 with accepting the alternative hypothesis

$$H_1 : \theta \in \Theta_1. \tag{2.91}$$

Any decision function δ consists of dividing up the set $\mathcal{X}$ of possible observations into two subsets: U_0 and $U_1 = \mathcal{X} - U_0$. Observations in U_0 lead to accepting H_0, and observations in U_1 lead to accepting H_1. The risk $R(\theta, \delta)$ depends on both the cost of making the wrong decision and on the probability of making the wrong decision, when θ is the state of nature. For example, if we associate a loss of zero with a correct decision, then

$$\begin{aligned} R(\theta,\delta) &= L(\theta,a_1)P_\theta(X \in U_1) \text{ for } \theta \in \Theta_0, \\ R(\theta,\delta) &= L(\theta,a_0)P_\theta(X \in U_0) \text{ for } \theta \in \Theta_1. \end{aligned} \tag{2.92}$$

Historically, the theory of hypothesis testing developed slowly in several stages, before the introduction of decision theory. In the first stage of *significance testing*, the formulation was incomplete and no attention was paid to the alternative hypothesis nor to the cost of making the wrong decision. Typically one wished to establish that some treatment had an effect. A *null hypothesis* H_0 would be formulated to state that the treatment had no effect. (The action in the real world corresponding to rejecting the hypothesis that there is no effect would be to continue research in that direction or to decide tc apply the treatment. Accepting the hypothesis would presumably lead to giving up on the treatment.) A statistic T would be introduced which would measure how inconsistent the data are with the null hypothesis, and would lead to rejection if T were large enough.

For example, assume that our experiment consists of n *iid* observations $X_1, X_2, \ldots, X_n$ from a $N(\mu, \sigma^2)$ distribution where σ^2 is known, and that our null hypothesis is

$$H_0 : \mu = 0. \tag{2.93}$$

A reasonable statistic to use in assessing evidence against H_0 appears to be the absolute value of the sample mean, $|\bar{X}|$, since $|\bar{X}|$ estimates $|\mu|$, and so large values of $|\bar{X}|$ suggest that the data are inconsistent with H_0. We propose the test 'reject H_0 if

$$T = |\bar{X}| > 1.96\sigma/\sqrt{n}\text{'}. \tag{2.94}$$

The probability of rejecting the null hypothesis when the null hypothesis is true is called the *significance level* or the *size* of a hypothesis test, and usually denoted α. For our example, the constant 1.96 was chosen so that $\alpha \doteq .05$. It is important to choose a significance level before examining the data. A measure of the consistency of the data with a null hypothesis is the *P-value*, which is the smallest significance level for which the null hypothesis can be rejected. Thus, a test statistic which would yield a P-value of less than .05 would be regarded as *significant at the .05 level* and lead to rejection if a .05 level test were used. In this case a P-value of .0001 would be regarded as highly significant, and would be of interest to the statistician who isn't completely bound by formalism, but in principle it would lead to the same conclusion as a P-value of .0499.

When a test is of the form 'reject H_0 if $T \geq k$', k is sometimes called a *critical value*. Traditionally k is a constant, although we will consider situations in which k is a function of the data, and we will approximate the functional form of this *statistic* k, in order to acheive certain as yet unspecified aims, by attempting to solve an integral equation.

The above example problem becomes more complicated if, as is common in real applications, σ is unknown.

Then a particular test of the form 'reject H_0 if $T > 1.645\sigma_0/\sqrt{n}$', where σ_0 is some constant, has the undesirable result that the probability of rejecting the hypothesis depends on the *nuisance parameter* σ, which is not of major interest in itself. In fact the probability of rejecting the hypothesis $H_0 : X \sim N(0,\sigma^2)$, with unknown positive σ, varies from 0 to 1 as σ varies over the interval $(0,\infty)$. This problem was resolved by W. S. Gossett, using the pseudonym 'Student', who suggested the use of the test procedure: 'reject H_0 if

$$T = \frac{|\bar{X}|}{S/\sqrt{n}} > k\text{'}. \tag{2.95}$$

Here k is a constant critical value, and the denominator is an estimate of $\sigma/\sqrt{n}$. The test (2.95) resembles the previous test (2.94), with the known standard deviation replaced by its estimate. When H_0 is true, the probability of falsely rejecting H_0 is determined from Student's-t distribution, and depends only on the choice of k and $n-1$; it is independent of the nuisance parameter. Test procedures for which the probability of rejection when the hypothesis is true does not depend on the nuisance parameter are called *similar*.

A later stage in the development of the theory of hypothesis testing came out of the realization that the significance theory did not give any formal suggestions for selecting one test statistic over another. Neyman and Pearson introduced the notion of *alternative hypotheses*. They formulated the problem of minimizing the probability of accepting the hypothesis when it is false, given the size or significance level of the test. Then the above problem could be stated as one where we observe *iid* observations which are $N(\mu,\sigma^2)$, where $\theta = (\mu,\sigma)$ and it is desired to test

$$H_0 : \theta \in \Theta_0 = \{\theta : \mu = 0, 0 < \sigma < \infty\} \tag{2.96}$$

against the alternative

$$H_1 : \theta \in \Theta_1 = \{\theta : \mu \neq 0, 0 < \sigma < \infty\}. \tag{2.97}$$

Here, one is interested in the *power function* which measures the probability of rejecting the hypothesis for all possible values of θ. In our example above the power function of the t-test suggested depends only on k, n, and the noncentrality parameter $\delta = \sqrt{n}|\mu|/\sigma$. To see this, note that (2.95) can be written as

$$\begin{aligned} T &= \left|\frac{\bar{X}}{S/\sqrt{n}}\right| \qquad (2.98) \\ &= \left|\frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} + \frac{\sqrt{n}\mu}{\sigma}}{S/\sigma}\right| \\ &\equiv \left|\frac{Z+\delta}{Y}\right|, \end{aligned}$$

where $\delta = \sqrt{n}|\mu|/\sigma$, $Z \sim \mathrm{N}(0,1)$, $(n-1)Y^2 \sim \chi^2_{n-1}$, and Y^2 (hence Y) is independent of Z. Therefore, T is distributed as the absolute value of a noncentral-t random variable, with $(n-1)$ degrees of freedom and noncentrality parameter $|\delta|$.

In general, for composite hypotheses, the size of a test is

$$\alpha = \sup_{\theta \in \Theta_0} P_\theta(\text{ Reject } H_0). \tag{2.99}$$

The Student t test, described in (2.95), can be shown to be optimal among size α tests for which the power is symmetric in the parameter μ.

This theory fails to give formal consideration to the cost of incorrect decisions, but there was always some sort of informal attention paid to cost, in order to rationalize the selection of good significance levels of the test procedures. It is implicit in that the Neyman-Pearson theory tends to treat the two hypotheses asymmetrically.

2.6.6 Confidence Intervals

So far the theory of estimation, as expressed above, does not pay much attention to how reliable the estimates are. Ordinarily, the statistician or scientist wants to know, for his real decision making, which depends only in part on his estimate, how reliable this estimate is. Traditionally, one accompanies an estimate of $g(\theta)$ with an estimate of how variable that estimate is. Philosophically this puts us in a problem of estimating the variance of the estimate of the variance of the ...of the estimate. That problem can be resolved by the use of confidence intervals or regions.

A *confidence interval*, or more generally, a *confidence region*, is a random set which contains the true value of a (scalar or vector) parameter with at least a specified probability, or *confidence.* Let $\mathcal{U}(X)$ be a subset of the parameter space Θ which depends on the data X. If, for all $\theta \in \Theta$

$$P_\theta[g(\theta) \in \mathcal{U}(X)] \geq \gamma, \tag{2.100}$$

where the probability is determined from the distribution $F_X(x;\theta)$ of the data, then the region $\mathcal{U}(X)$ is called a confidence region for $g(\theta)$ of confidence at least γ.

For example, if $X_i \sim \mathrm{N}(\mu,\sigma^2)$ for $i = 1,\ldots,n$, $\theta = (\mu,\sigma^2)$, and $\bar{X}$ and S^2 are the sample mean and variance, then

$$T = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim T_{n-1}, \tag{2.101}$$

where t_{n-1} denotes the Student-t distribution with $n-1$ degrees of freedom, and hence, for all μ

$$P_\theta[\bar{X} - t_{n-1}(\alpha/2)S/\sqrt{n} \leq \mu \leq \bar{X} + t_{n-1}(\alpha/2)S/\sqrt{n}] = \gamma, \tag{2.102}$$

where $\alpha \equiv 1-\gamma$ and $P(T \geq t_{n-1}(\alpha/2)) = \alpha/2$. The random interval

$$\Gamma : (\bar{X} - t_{n-1}(\alpha/2)S/\sqrt{n},\ \bar{X} + t_{n-1}(\alpha/2)S/\sqrt{n}) \tag{2.103}$$

contains μ with probability γ, and we say that Γ is a $100\gamma\%$ confidence interval for μ. To be more specific, Γ is a *two-sided interval*; we can also construct *one-sided intervals* if we are interested only in a lower, or upper, *confidence limit* on μ.

A hypothesis under which the parameter equals a specific point in the parameter space is called a *simple hypothesis*; the complementary situation is called a *composite hypothesis.* There is a one-to-one relationship between simple hypotheses and confidence intervals: given a confidence interval of confidence $1-\alpha$ for θ_0 the test 'reject $H_0 : \theta = \theta_0$ if θ_0 is not in this confidence interval' is a hypothesis test of size α.

Actually, there can be a one-to-one relationship between confidence intervals and hypothesis tests even when the null hypothesis is composite, and the confidence interval (2.103) provides one such example. The interval (2.103) corresponds to the composite null hypothesis

$$H_0 : \{(\mu,\sigma^2) : \mu = \mu_0, 0 < \sigma^2 < \infty\} \tag{2.104}$$

together with the composite alternative

$$H_1 : \{(\mu,\sigma^2) : \mu \neq \mu_0, 0 < \sigma^2 < \infty\}. \tag{2.105}$$

A test of H_0 with alternative H_1 of size α is provided by the criterion 'reject H_0 if (2.103) does not contain μ_0'. The reason why the interval (2.103) corresponds to a hypothesis test is that the relevant test statistic does not depend on the nuisance parameter σ^2, and were it not for this parameter H_0 would be simple.

2.6.7 Examples of Integral Equations in Estimation and Hypothesis Testing

In this subsection, we provide examples of problems in unbiased estimation and hypothesis testing which give rise to integral equations of the first kind. The first example is a problem of unbiased estimation chosen because it is simple and because it illustrates an iterative algorithm which we will discuss in later chapters. The hypothesis testing example provides a preview of the Behrens-Fisher problem, to be presented in much more detail in Chapter 5.

Determining an Unbiased Estimator

Let X be a random variable with probability density

$$f(x;\theta) = \begin{cases} e^{-(x-\theta)} & \text{for } x \geq \theta \\ 0 & \text{for } x < \theta. \end{cases} \tag{2.106}$$

We will determine an unbiased estimator of θ^2, that is, a function $h(X)$ such that

$$E_\theta[h(X)] = \int_\theta^\infty h(x)f(x;\theta)dx = \theta^2. \tag{2.107}$$

If such an estimator exists, it can be shown to be the unique minimum variance unbiased estimator of θ^2.

We will solve this problem by employing an iterative algorithm which is a special case of the method to be considered in later chapters. Given an approximation $h^n(X)$ to $h(X)$, we define $h^{n+1}(X)$ to be

$$h^{n+1}(x) \equiv h^n(x) + \left[\theta^2 - \int_\theta^\infty h^n(y)f(y;\theta)dy\right]_{\theta=x}, \tag{2.108}$$

where once the function of θ in the square brackets is calculated, θ is to be replaced with x.

Let $h^0(x) = 0$. We can easily calculate the first two moments of X,

$$E_\theta(X) = \theta + 1, \tag{2.109}$$

and

$$E_\theta(X^2) = \theta^2 + 2\theta + 2, \tag{2.110}$$

and use these moments to show that

$$\begin{aligned} h^0(x) &= 0, \\ h^1(x) &= x^2, \\ h^2(x) &= x^2 - 2x - 2, \text{ and} \\ h^n(x) &= x^2 - 2x, \end{aligned} \tag{2.111}$$

for $n > 2$. The random variable $h(X) = X^2 - 2X$ is the unbiased estimator; the algorithm converged to the exact solution in three iterations.

The Behrens-Fisher Problem

Let $X_{1i}, i = 1, \ldots, n_1$ and $X_{2i}, i = 1, \ldots, n_2$ denote random samples from normal populations with means and variances (μ_1, σ_1^2) and (μ_2, σ_2^2), respectively, and let the sample means and variances be $\bar{X}_j$ and S_j^2, for $j = 1,2$.

Consider the problem of testing the composite null hypothesis

$$H_0 : \mu_1 = \mu_2 \tag{2.112}$$

against the alternative

$$H_1 : \mu_1 > \mu_2. \tag{2.113}$$

If the variance ratio, $r \equiv \sigma_1^2/\sigma_2^2$, is known then since

$$D \equiv \bar{X}_1 - \bar{X}_2 \sim \mathrm{N}(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2), \tag{2.114}$$

and D is independent of the estimate of $\sigma_1^2 = r\sigma_2^2$

$$K \equiv \frac{(n_1 - 1)S_1^2 + r(n_2 - 1)S_2^2}{n_1 + n_2 - 2}, \tag{2.115}$$

which is proportional to a $\chi^2_{n_1+n_2-2}$ random variable, we have that

$$\frac{D - (\mu_1 - \mu_2)}{\sqrt{[n_1^{-1} + (rn_2)^{-1}]K}} \sim T_{n_1+n_2-2}. \tag{2.116}$$

Thus, a simple extension of the student-t hypothesis test, discussed in Section 2.6.5 for a single sample situation, provides an effective means of performing hypothesis tests and obtaining confidence intervals for this two-sample case where r is known.

The situation where the variance ratio is unknown is usually referred to as the *Behrens-Fisher problem*. This problem, the main topic of Chapter 5, has been controversial and important to the theory of statistics.

A natural test statistic to consider for the Behrens-Fisher problem is

$$U = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}, \tag{2.117}$$

where the null hypothesis is rejected when U is observed to be greater than a critical value, which is a function of the data to be determined. We will allow the critical value of this test statistic to depend, in an unspecified way, on the sample variances.

We would like the size of this hypothesis test to not depend on the nuisance parameter r, or equivalently, on

$$\theta = \frac{\sigma_1^2/n_1}{\sigma_1^2/n_1 + \sigma_2^2/n_2}. \tag{2.118}$$

A sample estimate of this parameter is the statistic

$$R = \frac{S_1^2/n_1}{S_1^2/n_1 + S_2^2/n_2}. \tag{2.119}$$

We pose the following mathematical problem: determine a function, d, of the random variable R so that, given that the null hypothesis is true,

$$P(U \geq d(R)|\theta) = \alpha, \tag{2.120}$$

for all θ. A function $d(R)$ satisfying (2.120) would provide a critical value statistic for a similar test of the hypothesis (2.112) against the alternative (2.113) of size α.

This function d, if indeed it exists, will be shown in Chapter 5 to be a solution of the following nonlinear integral equation:

$$E_{\nu_1/2,\nu_2/2}\left\{T_{\nu_1+\nu_2}\left[d(W)\sqrt{\nu_1+\nu_2}\sqrt{\frac{X\theta}{\nu_1}+\frac{(1-X)(1-\theta)}{\nu_2}}\right]\right\}=1-\alpha, \qquad (2.121)$$

where $\nu_j \equiv n_j - 1$; the expectation is with respect to X, a Beta random variable with parameters $\nu_1/2$ and $\nu_2/2$; $T_\eta(\cdot)$ denotes the t distribution with η degrees of freedom; and W denotes the random variable

$$W \equiv \frac{X\theta/\nu_1}{X\theta/\nu_1+(1-X)(1-\theta)/\nu_2}. \qquad (2.122)$$

Chapter 3

Richardson's Algorithm, Preconditioning, and Iterative Regularization

3.1 Richardson's Algorithm

Let $K : L_2[0,1] \rightarrow L_2[0,1]$ be compact. Consider the linear equation of the first kind

$$Kf = g, \tag{3.1}$$

where f and g are in L_2, and g is a known function. Define the iteration

$$f^{n+1} = f^n + \theta D(g - Kf^n), \quad \text{for } n = 0,1,2,\ldots, \tag{3.2}$$

where θ is a positive constant, $f^0 : L_2 \rightarrow L_2$ is arbitrary, and D is a known invertible linear operator with a bounded inverse. This is the iterative algorithm which will concern us for most of this chapter. When $D = I$, the identity operator, (3.2) is *Richardson's algorithm*,

$$f^{n+1} = f^n + \theta(g - Kf^n), \quad \text{for } n = 0,1,2,\ldots, \tag{3.3}$$

proposed by Richardson (1910) for the iterative solution of sparse linear systems. We would like to choose D to accelerate convergence to a vector f such that $\|Kf - g\|$ is sufficiently small, a practice known as *preconditioning*, where D will be referred to as the *preconditioning operator.*

The plan of this chapter is as follows. We consider first the behavior of (3.3) in various situations. Except for a literature review. we do not consider the convergence of this algorithm to a solution in L_2. Since we are ultimately interested in solving discretizations of integral equations on a computer, we are more interested in singular, and possibly inconsistent, matrix equations than in functional equations, and we are more interested in how close this algorithm comes to a smooth *near-solution* in a few dozen iterations than in ultimate convergence to a solution. We introduced the concept of a near-solution in Chapter 1. Having established some elementary ideas of functional analysis in Chapter 2, we can now be more specific. We will say that a function f is a near-solution to an equation $Kf = g$ if

$$\|Kf - g\| < \tau\|g\|, \tag{3.4}$$

where the constant τ is application dependent. A choice of τ corresponds to a decision concerning what is considered to be a 'small' residual.

Next, we consider the choice of a preconditioning operator, and our interest shifts to (3.2). A particular choice of D leads t) the *Conditional Expcctation* algorithm, which is motivated in several ways and illustrated on various examples. We will demonstrate that the Conditional Expectation algorithm can quickly lead to near-solutions.

If K is an integral operator with a smooth kernel, than (3.3) will tend to produce smooth approximate solutions. There is regularization implicit in the iteration, and, following the discussion of convergence theory and preconditioning, we present the idea of ***iterative regularization.***

This chapter concludes with the discussion of examples; details of the numerical implementation of the algorithms are provided in Appendix A.

3.1.1 Convergence of the Richardson and Landweber Algorithms in L_2

Proofs of the convergence of (3.3), under various conditions on the operator K, appear in the literature, which we briefly review here. For more information, a good place to start is Patterson (1974) and the references there.

If the operator K is positive and compact, then it is necessarily self-adjoint and it has a denumerable set of nonnegative eigenvalues. Moreover, $\|K\| = \lambda_1$, where λ_1 is the largest eigenvalue of K. If K is only assumed to be compact, then K^*K is positive and compact. It is not difficult to show (Patterson, 1974, p. 7) that if (3.1) is solvable, then it has the same solutions as

$$K^*Kf = K^*g. \tag{3.5}$$

Landweber (1951) considered the iteration

$$f^{n+1} = f^n + \theta K^*(g - Kf^n) \tag{3.6}$$

where $0 < \theta < 2/\lambda_1^2$, and proved convergence for a Fredholm integral operator having a continuous, real kernel. If a solution exists, then (3.6) converges to a solution, otherwise this iteration converges to a function which minimizes $\|g - Kf\|$. If K is positive and compact, then Landweber has also (trivially) proved convergence of Richardson's algorithm (3.3) for $0 < \theta < 2/\lambda_1$, although he did not comment on this fact. Bialy (1959, see also Patterson, 1974, pp. 33-41) generalized Landweber's results to K bounded, but not necessarily compact.

3.1.2 Richardson's Algorithm for Matrix Equations

Usually, the iterations of Richardson's and Landweber's algorithms cannot be performed analytically. Instead one discretizes an integral equation in order to obtain an approximating matrix equation. The discretization schemes used in this thesis for integral equations of the first kind are described in Appendix A.

We therefore consider in this section Richardson's algorithm (3.3) applied to the matrix equation $Kf = g$, where K is square and possibly singular, and g is not necessarily in the range of K; i.e. the equation might be *inconsistent.*

We are interested in matrix equations which are approximations to ill-posed integral equations, so situations where K is singular and/or the matrix equation is inconsistent are particularly important. The L_2 convergence theory reviewed in the previous subsection is of little use here. Indeed, for reasons discussed in Section 2.4.2, we are less interested in 'solving' the equation than in finding a smooth 'near-solution': hence, an algorithm which ultimately diverges might still be of considerable use if it quickly leads to such a near-solution, at which point the iteration can be terminated.

Some Notation and a Preliminary Lemma

Let K be an $m \times m$ matrix of rank $q < m$. The Jordan form of the matrix K will be written as

$$K = B^{-1} \begin{bmatrix} J_{11} & 0_{12} \\ 0_{21} & N_{22} \end{bmatrix} B \equiv \tag{3.7}$$

$$\begin{bmatrix} B^{\cdot 1}_{m\times s} & B^{\cdot 2}_{m\times(m-s)} \end{bmatrix} \begin{bmatrix} J_{11 s\times s} & 0_{12 s\times(m-s)} \\ 0_{21(m-s)\times s} & N_{22(m-s)\times(m-s)} \end{bmatrix} \begin{bmatrix} B_{1\cdot s\times m} \\ B_{2\cdot (m-s)\times m} \end{bmatrix},$$

where J_{11} is a nonsingular matrix of Jordan blocks, and N_{22} is a nilpotent matrix of index $\iota \geq 1$ of Jordan blocks corresponding to a zero eigenvalue. The dimensions of the submatrices are as indicated, and $s \leq q$. Either the row or the column dimension of each block in the partitioned matrices B and B^{-1} is equal to m; the dot indicates which. Also, the use of superscripts and subscripts on these blocks is intended to aid in identifying, at a glance, that a product such as $B^{\cdot 1}B_{1\cdot}$ is conformable.

We will also make use of the partitioned identity matrix

$$I_{m\times m} = \begin{bmatrix} I_{11 s\times s} & 0_{12 s\times(m-s)} \\ 0_{21(m-s)\times s} & I_{22(m-s)\times(m-s)} \end{bmatrix}. \tag{3.8}$$

Since $BB^{-1} = B^{-1}B = I$, we have the identities

$$\begin{aligned} B^{\cdot 1}B_{1\cdot} + B^{\cdot 2}B_{2\cdot} &= I, \\ B_{1\cdot}B^{\cdot 1} &= I_{11}, \\ B_{1\cdot}B^{\cdot 2} &= 0_{12}, \\ B_{2\cdot}B^{\cdot 1} &= 0_{21}, \text{ and} \\ B_{2\cdot}B^{\cdot 2} &= I_{22}. \end{aligned} \tag{3.9}$$

We will use the above notation and identities in the following lemma. The various parts of this lemma are either well known, or else follow directly from well known results in texts such as Campbell and Meyer (1979).

Lemma 3.1.1 *Let K be an $m \times m$ square matrix of index $\iota > 0$ with Jordan form (3.7). Let $V \equiv B^{\cdot 1}B_{1\cdot}$. Then*

a) V and $I - V$ are projections onto $\mathcal{R}(V)$ and $\mathcal{R}(I - V)$, respectively,

b) $V(I - V) = (I - V)V = 0$,

c) $x = Vx + (I - V)x$ is the unique decomposition $x = x_1 + x_2$ for which $x_1 \in \mathcal{R}(V)$ and $x_2 \in \mathcal{R}(I - V)$, and

d)

$$\mathcal{R}(V) = \mathcal{N}(I - V) = \mathcal{R}(B^{\cdot 1}) = \mathcal{N}(B_{2\cdot}) = \mathcal{R}(K^{\iota}), \text{ and} \tag{3.10}$$

$$\mathcal{R}(I - V) = \mathcal{N}(V) = \mathcal{N}(B_{1\cdot}) = \mathcal{R}(B^{\cdot 2}) = \mathcal{N}(K^{\iota}). \tag{3.11}$$

Proof: (a) Since

$$V^2 = B^{\cdot 1}(B_{1\cdot}B^{\cdot 1})B_{1\cdot} = B^{\cdot 1}I_{11}B_{1\cdot} = V \tag{3.12}$$

and

$$(I - V)^2 = I + V^2 - 2V = I + V - 2V = I - V, \tag{3.13}$$

both V and $I - V$ are idempotent, and hence projection matrices. We say that V projects onto $\mathcal{R}(V)$ *along* $\mathcal{N}(V)$, and that $I - V$ projects onto $\mathcal{N}(V)$ along $\mathcal{R}(V)$. Note that these projections are in general not orthogonal.

(b) This follows immediately from (a):

$$V(I - V) = V - V^2 = V - V = 0; \tag{3.14}$$

similarly $(I - V)V = 0$.

(c) Of course, $x = Vx + (I - V)x$ is one such decomposition. Let $x = x_1 + x_2$, where $x_1 \in \mathcal{R}(V)$ and $x_2 \in \mathcal{R}(I - V)$. Then there exist vectors y_1 and y_2 such that $x_1 = Vy_1$ and $x_2 = (I - V)y_2$. So

$$Vx = V^2y_1 + V(I - V)y_2 = Vy_1 = x_1, \tag{3.15}$$

and

$$(I - V)x = (I - V)Vy_1 + (I - V)^2y_2 = (I - V)y_2 = x_2, \tag{3.16}$$

where we have made use of (a) and (b). Therefore, the decomposition is unique.

(d) We note first that

$$K^{\iota} = \begin{bmatrix} B^{\cdot 1} & B^{\cdot 2} \end{bmatrix} \begin{bmatrix} J_{11}^{\iota} & 0_{12} \\ 0_{21} & N_{22}^{\iota} \end{bmatrix} \begin{bmatrix} B_{1\cdot} \\ B_{2\cdot} \end{bmatrix} = B^{\cdot 1}J_{11}^{\iota}B_{1\cdot}, \tag{3.17}$$

since $N_{22}^{\iota} = 0$ by the definition of the index ι.

If $x \in \mathcal{R}(K^{\iota})$, then there exists a y such that

$$x = K^{\iota}y = B^{\cdot 1}(J_{11}^{\iota}B_{1\cdot}y) \equiv B^{\cdot 1}z, \tag{3.18}$$

for some vector z, so $\mathcal{R}(K^{\iota}) \subset \mathcal{R}(B^{\cdot 1})$. Now let $x \in \mathcal{R}(B^{\cdot 1})$, so that

$$x = B^{\cdot 1}y = B^{\cdot 1}(J_{11}^{\iota}B_{1\cdot}B^{\cdot 1}J_{11}^{-\iota})y \equiv B^{\cdot 1}J_{11}^{\iota}B_{1\cdot}z = K^{\iota}z, \tag{3.19}$$

so $\mathcal{R}(B^{\cdot 1}) \subset \mathcal{R}(K^{\iota})$, and hence $\mathcal{R}(B^{\cdot 1}) = \mathcal{R}(K^{\iota})$.

Obviously,

$$\mathcal{N}(B_{1\cdot}) \subset \mathcal{N}(B^{\cdot 1}J_{11}^{\iota}B_{1\cdot}) = \mathcal{N}(K^{\iota}). \tag{3.20}$$

Let $x \in \mathcal{N}(K^{\iota})$. Then,

$$K^{\iota}x = B^{\cdot 1}J_{11}^{\iota}B_{1\cdot}x = 0 \Rightarrow J_{11}^{-\iota}(B_{1\cdot}B^{\cdot 1})J_{11}^{\iota}B_{1\cdot}x = B_{1\cdot}x = 0, \tag{3.21}$$

so $\mathcal{N}(K^{\iota}) \subset \mathcal{N}(B_{1\cdot})$, and hence $\mathcal{N}(B_{1\cdot}) = \mathcal{N}(K^{\iota})$.

We show next that $\mathcal{R}(V) = \mathcal{N}(I - V)$ and $\mathcal{R}(I - V) = \mathcal{N}(V)$. Assume that $x \in \mathcal{R}(I - V)$. Then, using (b),

$$x = (I - V)y \Rightarrow Vx = V(I - V)y = 0 \Rightarrow x \in \mathcal{N}(V). \tag{3.22}$$

Conversely, if $x \in \mathcal{N}(V)$ then

$$Vx = 0 \Rightarrow x - Vx = (I - V)x = x \Rightarrow x \in \mathcal{R}(I - V). \tag{3.23}$$

Therefore, $\mathcal{R}(I - V) = \mathcal{N}(V)$.

Similarly, assume that $x \in \mathcal{R}(V)$. Then

$$x = Vy \Rightarrow (I - V)x = (I - V)Vy = 0 \Rightarrow x \in \mathcal{N}(I - V). \tag{3.24}$$

If $x \in \mathcal{N}(I - V)$, then

$$(I - V)x = 0 \Rightarrow x = Vx \Rightarrow x \in \mathcal{R}(V). \tag{3.25}$$

Therefore, $\mathcal{R}(V) = \mathcal{N}(I - V)$.

Now we show that $\mathcal{R}(V) = \mathcal{R}(B^{-1})$ and $\mathcal{N}(V) = \mathcal{N}(B_1.)$. The first of these follows from

$$y \in \mathcal{R}(V) \Rightarrow y = Vx = B^{-1}B_1.x \equiv B^{-1}z \Rightarrow y \in \mathcal{R}(B^{-1}) \tag{3.26}$$

and

$$y \in \mathcal{R}(B^{-1}) \Rightarrow y = B^{-1}x \Rightarrow y = B^{-1}(B_1.B^{-1})x = (B^{-1}B_1.)B^{-1}x \equiv Vz \Rightarrow y \in \mathcal{R}(V). \tag{3.27}$$

The identity $\mathcal{N}(V) = \mathcal{N}(B_1.)$ follows similarly from

$$x \in \mathcal{N}(V) \Rightarrow Vx = B^{-1}B_1.x = 0 \Rightarrow (B_1.B^{-1})B_1.x = B_1.x = 0 \Rightarrow x \in \mathcal{N}(B_1.) \tag{3.28}$$

and

$$x \in \mathcal{N}(B_1.) \Rightarrow B_1.x = 0 \Rightarrow B^{-1}B_1.x = Vx = 0 \Rightarrow x \in \mathcal{N}(V). \tag{3.29}$$

We complete the proof of this lemma by showing that $\mathcal{N}(B_2.) = \mathcal{N}(I - V)$ and $\mathcal{R}(B^{-2}) = \mathcal{R}(I - V)$. Since $I - V = B^{-2}B_2.$, we have immediately that $\mathcal{N}(B_2.) \subset \mathcal{N}(I - V)$. If $x \in \mathcal{N}(I - V)$, then

$$(I - V)x = B^{-2}B_2.x = 0 \Rightarrow (B_2.B^{-2})B_2.x = B_2.x = 0, \tag{3.30}$$

and so $\mathcal{N}(I - V) \subset \mathcal{N}(B_2.)$, and hence $\mathcal{N}(B_2.) = \mathcal{N}(I - V)$. Finally, if $x \in \mathcal{R}(I - V)$, then

$$x = (I - V)y = B^{-2}B_2.y \equiv B^{-2}z, \tag{3.31}$$

so $\mathcal{R}(I - V) \subset \mathcal{R}(B^{-2})$. Conversely, if $x \in \mathcal{R}(B^{-2})$, then

$$x = B^{-2}y = B^{-2}(B_2.B^{-2})y = (I - V)(B^{-2}y) \equiv (I - V)z, \tag{3.32}$$

therefore $\mathcal{R}(B^{-2}) \subset \mathcal{R}(I - V)$, so $\mathcal{R}(B^{-2}) = \mathcal{R}(I - V)$. ∎

Convergence of Richardson's Algorithm for Nonsingular Matrix Equations

Convergence of Richardson's algorithm (3.3) to the unique solution $f = K^{-1}g$ of the equation $Kf = g$ where K is nonsingular depends on the spectral radius of the *iteration matrix*

$$G \equiv I - \theta K. \tag{3.33}$$

To see this, let $Kf = g$ and note that (3.3) leads to

$$(f - f^n) = (f - f^{n-1}) - \theta K(f - f^{n-1}) = G(f - f^{n-1}). \tag{3.34}$$

and hence, if we let

$$u^n \equiv f - f^n, \tag{3.35}$$

then

$$u^n = G^n u^0. \tag{3.36}$$

If $\rho(G) < 1$, then by Theorem 2.2.2, $G^n \to 0$, so $u^n \to 0$ for all initial approximations f^0 and all right hand sides g. If $\rho(G) \geq 1$, then $G^k \not\to 0$. So, by the definition of a convergent matrix, there must exist vectors u^0 for which $u^n = G^n u^0 \not\to 0$, hence there exist initial vectors f^0 such that $f^n \not\to f$. We have established the following theorem:

Theorem 3.1.1 (Convergence for The Nonsingular Case) *Let $f = K^{-1}g$. A necessary and sufficient condition for the iteration (3.3) to converge to f for all f^0 is that $\rho(I - \theta K) < 1$.*

If K is positive definite then, because of the spectral theorem (Theorem 2.1.1), the behavior of Richardson's algorithm is particularly transparent. Let K be $m \times m$ and positive definite, with eigenvalues

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m > 0, \tag{3.37}$$

and corresponding orthonormal eigenvectors $\{v_i\}_{i=1}^m$. The condition $\rho(I - \theta K) < 1$ translates to

$$-1 < 1 - \theta\lambda_1 \leq \ldots \leq 1 - \theta\lambda_m < 1, \tag{3.38}$$

and we have

$$0 < \theta < 2/\lambda_1 \tag{3.39}$$

as the necessary and sufficient condition for convergence for arbitrary f^0. The solution f, the right hand side g, and the iterates f^n can be expressed in terms of these eigenvectors as, say,

$$f = \sum_{i=1}^m c_i v_i, \tag{3.40}$$

$$g = \sum_{i=1}^m h_i v_i, \text{ and} \tag{3.41}$$

$$f^n = \sum_{i=1}^m c_i^n v_i. \tag{3.42}$$

Because of the orthonormality of the $\{v_i\}$, the Richardson iteration (3.3), in the form (3.34), leads to the following expression for the coefficients $\{c_i^n\}$:

$$c_i - c_i^n = (1 - \theta\lambda_i)^n (c_i - c_i^0) \tag{3.43}$$

for $i = 1, \ldots, m$ and $n \geq 0$. If the condition (3.39) holds, then, for each i,

$$\lim_{n\to\infty} c_i^n = c_i, \tag{3.44}$$

so $f^n \to f$. Note that since $Kf = g$,

$$Kf = \sum_{i=1}^m \sum_{j=1}^m \lambda_i v_i v_i^* c_j v_j = \sum_{i=1}^m \lambda_i c_i v_i \tag{3.45}$$

implies that

$$c_i = h_i/\lambda_i. \tag{3.46}$$

for $i = 1, \ldots, m$.

Convergence of Richardson's Algorithm for Singular Matrix Equations

We come now to the central results of this section. What if the square matrix K in the equation $Kf = g$ is singular, so that this equation has either zero or else infinitely many solutions? We consider in this subsection conditions under which Richardson's algorithm applied to a singular matrix equation converges to a solution; a necessary condition for this is that $g \in \mathcal{R}(K)$. The geometry underlying Lemma 3.1.1 leads directly to the following sufficient conditions for convergence:

Theorem 3.1.2 (Convergence for The Singular Case) *Let K be a square, singular matrix with index ι. Richardson's algorithm (3.3) converges to a solution f of the equation $Kf = g$ if the following conditions are satisfied:*

1. *All of the nonzero eigenvalues of the matrix θK are contained in the interior of the unit circle with center (1,0) in the complex plane,*
2. *$f^0 \in \mathcal{R}(K^{\iota-1})$, and*
3. *$g \in \mathcal{R}(K^{\iota})$,*

where we interpret $\mathcal{R}(K^0)$ to mean $\mathcal{R}(I)$.

Proof: Using the notation of Section 3.1.2, note that the index $\iota \geq 1$, and write (3.3) in the form

$$\begin{aligned}
f^n &= (I - \theta K)^n f^0 + \sum_{i=0}^{n-1} (I - \theta K)^i \theta g \qquad (3.47)\\
&= \begin{bmatrix} B^{\cdot 1} & B^{\cdot 2} \end{bmatrix} \begin{bmatrix} (I_{11} - \theta J_{11})^n & 0_{12} \\ 0_{21} & (I_{22} - \theta N_{22})^n \end{bmatrix} \begin{bmatrix} B_{1\cdot} \\ B_{2\cdot} \end{bmatrix} f^0 \\
&+ \begin{bmatrix} B^{\cdot 1} & B^{\cdot 2} \end{bmatrix} \begin{bmatrix} \sum_{i=0}^{n-1} (I_{11} - \theta J_{11})^i & 0_{12} \\ 0_{21} & \sum_{i=0}^{n-1} (I_{22} - \theta N_{22})^i \end{bmatrix} \begin{bmatrix} B_{1\cdot} \\ B_{2\cdot} \end{bmatrix} \theta g \\
&= B^{\cdot 1} (I_{11} - \theta J_{11})^n B_{1\cdot} f^0 + B^{\cdot 1} \sum_{i=0}^{n-1} (I_{11} - \theta J_{11})^i B_{1\cdot} \theta g \\
&+ B^{\cdot 2} (I_{22} - \theta N_{22})^n B_{2\cdot} f^0 + B^{\cdot 2} \sum_{i=0}^{n-1} (I_{22} - \theta N_{22})^i B_{2\cdot} \theta g \\
&\equiv a_1^n + a_2^n + a_3^n + a_4^n.
\end{aligned}$$

The nonzero eigenvalues of θK correspond to the eigenvalues of the nonsingular matrix $I_{11} - \theta J_{11}$. If λ is any eigenvalue of θK in the interior of the circle specified in the statement of the theorem, then $1 - \theta\lambda$ is contained in the interior of the unit circle centered at the origin, so condition 1) implies that

$$\rho(I_{11} - \theta J_{11}) < 1. \qquad (3.48)$$

Hence, by Theorem 2.2.2,

$$\lim_{n\to\infty} a_1^n = \lim_{n\to\infty} B^{\cdot 1} (I_{11} - \theta J_{11})^n B_{1\cdot} f^0 = 0 \qquad (3.49)$$

and by Lemma 2.2.3

$$\mathop{\text{iim}}_{n\to\infty} a_2^n = \lim_{n\to\infty} B^{\cdot 1} \sum_{i=0}^{n-1} (I_{11} - \theta J_{11})^i B_{1\cdot} \theta g = B^{\cdot 1} J_{11}^{-1} B_{1\cdot} g. \qquad (3.50)$$

Next, assume that $f^0 \in \mathcal{R}(K^{\iota-1})$. Then there exists a vector y such that

$$f^0 = B^{\cdot 1} J_{11}^{\iota-1} B_{1\cdot} y + B^{\cdot 2} N_{22}^{\iota-1} B_{2\cdot} y, \tag{3.51}$$

and hence, for all $n \geq 0$,

$$\begin{aligned} a_3^n = {} & B^{\cdot 2}(I_{22} - \theta N_{22})^n B_{2\cdot} f^0 \\ = {} & B^{\cdot 2}(I_{22} - \theta N_{22})^n B_{2\cdot} \left(B^{\cdot 1} J_{11}^{\iota-1} B_{1\cdot} y + B^{\cdot 2} N_{22}^{\iota-1} B_{2\cdot} y \right) \\ = {} & \sum_{j=0}^{n} \left(\begin{array}{c} n \\ j \end{array} \right) B^{\cdot 2}(-\theta N_{22})^j B_{2\cdot} \left(B^{\cdot 1} J_{11}^{\iota-1} B_{1\cdot} y + B^{\cdot 2} N_{22}^{\iota-1} B_{2\cdot} y \right) \\ = {} & \sum_{j=0}^{n} \left(\begin{array}{c} n \\ j \end{array} \right) B^{\cdot 2}(-\theta)^j N_{22}^{j+\iota-1} B_{2\cdot} y \\ = {} & B^{\cdot 2} N_{22}^{\iota-1} B_{2\cdot} y, \end{aligned} \tag{3.52}$$

where we have used the identitie; (3.9) and the definition of the index ι. Note that a_3^n is independent of n. The remaining condition and Lemma 3.1.1 (d) together lead to

$$g \in \mathcal{R}(K^{\iota}) = \mathcal{N}(B_{2\cdot}), \tag{3.53}$$

therefore the term a_4^n of (3.47) is equal to zero for all n.

If the conditions of the theorem are satisfied, then

$$\lim_{n\to\infty} f^n = \lim_{n\to\infty} (a_1^n + a_2^n + a_3^n + a_4^n) = B^{\cdot 1} J_{11}^{-1} B_{1\cdot} g + B^{\cdot 2} N_{22}^{\iota-1} B_{2\cdot} y, \tag{3.54}$$

for some vector y. We demonstrate that the algorithm leads to convergence to a solution by evaluating $K(\lim_{n\to\infty} f^n)$:

$$\begin{aligned} & K \left(B^{\cdot 1} J_{11}^{-1} B_{1\cdot} g + B^{\cdot 2} N_{22}^{\iota-1} B_{2\cdot} y \right) \\ = {} & \left(B^{\cdot 1} J_{11} B_{1\cdot} + B^{\cdot 2} N_{22} B_{2\cdot} \right) \left(B^{\cdot 1} J_{11}^{-1} B_{1\cdot} g + B^{\cdot 2} N_{22}^{\iota-1} B_{2\cdot} y \right) \\ = {} & B^{\cdot 1} J_{11} B_{1\cdot} B^{\cdot 1} J_{11}^{-1} B_{1\cdot} g + B^{\cdot 2} N_{22} B_{2\cdot} B^{\cdot 2} N_{22}^{\iota-1} B_{2\cdot} y \\ = {} & B^{\cdot 1} J_{11} J_{11}^{-1} B_{1\cdot} g + B^{\cdot 2} N_{22}^{\iota} B_{2\cdot} y \\ = {} & B^{\cdot 1} B_{1\cdot} g = V g = g, \end{aligned} \tag{3.55}$$

where we have used the identities (3.9) and Lemma 3.1.1. Note that the term involving y of (3.54) depends on f^0, and will not be unique since $\iota > 0$. ∎

Corollary 3.1.1 (Global Convergence) *If K is a singular matrix with a diagonalizable nullspace, then conditions 1) and 3) of Theorem 3.1.2 are necessary and sufficient for Richardson's algorithm (3.3) to converge to a solution for all initial vectors f^0 (which is now the content of condition 2)).*

Proof: We will use the notation of Theorem 3.1.2. If K has a diagonalizable nullspace, then $\iota = 1$. Conditions 3) and 2) of Theorem 3.1.2 become

$$g \in \mathcal{R}(K^{\iota}) = \mathcal{R}(K) \tag{3.56}$$

and

$$f^0 \in \mathcal{R}(K^{\iota-1}) = \mathcal{R}(I). \tag{3.57}$$

Sufficiency thus follows from Theorem 3.1.2 with $\iota = 1$.

To prove necessity, first note that if condition 3) is violated, then a solution does not exist, so Richardson's algorithm cannot converge to a solution. Next, assume that condition 3) holds, but that condition 1) does not. Then $g \in \mathcal{R}(K)$, so a solution exists. We will show that for any g there exist vectors f^0 for which Richardson's iteration does not converge.

From Lemma (3.1.1) (d), observe that

$$\mathcal{R}(K) = \mathcal{N}(B_{2\cdot}), \tag{3.58}$$

hence $B_{2\cdot}g = 0$. Using (3.58), together with the hypothesis $\iota = 1$, the expression for the nth iterate (3.47) becomes

$$\begin{aligned} f^n &= B^{\cdot 1}(I_{11} - \theta J_{11})^n B_{1\cdot}f^0 + B^{\cdot 1}\sum_{i=0}^{n-1}(I_{11} - \theta J_{11})^i B_{1\cdot}\theta g \\ &+ n\theta B^{\cdot 2}B_{2\cdot}f^0 = a_1^n + a_2^n + n\theta(I - V)f^0. \end{aligned} \tag{3.59}$$

Since condition 1) does not hold,

$$\rho(I_{11} - \theta J_{11}) \geq 1, \tag{3.60}$$

and, by Theorem 2.2.2, $(I_{11} - \theta J_{11})^n \not\to 0$. Hence, there exist vectors f^0 such that $a_1^n \not\to 0$, as well as vectors f^0 (for example, $f^0 = 0$) for which a_1^n *does* converge to zero.

For any g, either a_2^n converges, or it does not. Assume that a_2^n converges. Then choose any $f^0 \in \mathcal{R}(K)$ for which a_1^n does *not* converge. To see that such an f^0 must exist, begin by choosing any vector $\tilde{f}^0$ for which a_1^n does not converge. Such a $\tilde{f}^0$ cannot be in $\mathcal{N}(B_{1\cdot})$, which equals $\mathcal{N}(K)$ by Lemma 3.1.1 (d). Lemma 3.1.1 (c) implies that

$$\tilde{f}^0 = V\tilde{f}^0 + (I - V)\tilde{f}^0 \equiv f_R^0 + f_N^0, \tag{3.61}$$

and this decomposition is unique. Since

$$\tilde{f}^0 \notin \mathcal{N}(K) = \mathcal{R}(I - V), \tag{3.62}$$

$f_R^0 \neq 0$. Let

$$f^0 = f_R^0 \in \mathcal{R}(K). \tag{3.63}$$

But $(I - V)$ projects onto $\mathcal{N}(K)$, hence $\theta n(I - V)f^0 = 0$ and

$$f^n = a_1^n + a_2^n, \tag{3.64}$$

where a_1^n does not converge, but a_2^n does. Therefore, f^n does not converge.

Assume that a_2^n does not converge. Then let $f^0 = 0$, so that

$$f^n = a_2^n, \tag{3.65}$$

which diverges. Conditions 1) and 3) are therefore necessary, and the proof of the corollary is complete. ∎

Inconsistent Equations

An inconsistent equation has no solution. However, if the right hand side of such an equation is replaced with *any* projection onto $\mathcal{R}(K)$, then the equation which results will have *many* solutions. Usually, one considers orthogonal projections, but we will find the generally non-orthogonal projection provided by the matrix V of Lemma 3.1.1 to be more convenient for our purposes.

Let K be an $m \times m$ singular matrix of index ι. Let $g \in \mathcal{C}^m$ be an arbitrary vector. From Lemma 3.1.1 (a-c)

$$\mathcal{C}^m = \mathcal{R}(K^\iota) \oplus \mathcal{N}(K^\iota), \tag{3.66}$$

so we can express g as

$$g = g_R + g_N, \tag{3.67}$$

where, $g_R \in \mathcal{R}(K^\iota)$, and $g_N \in \mathcal{N}(K^\iota)$. From Lemma 3.1.1, we note that g_R and g_N are uniquely determined by

$$g_R = Vg \tag{3.68}$$

and

$$g_N = (I - V)g, \tag{3.69}$$

respectively, where $V \equiv B^{\cdot 1}B_{1\cdot}$. We call any vector f such that $Kf = Vg$ a *generalized solution*, and we write

$$Kf \overset{\text{gen}}{=} g. \tag{3.70}$$

If K is a singular matrix of index ι, and if $Kf = g$ is inconsistent, then since

$$g \notin \mathcal{R}(K) \Rightarrow g \notin \mathcal{R}(K^\iota), \tag{3.71}$$

we have as a consequence of the proof of Theorem 3.1.2 that $\{f^n\}$ is not expected to converge. Using the notation of this theorem, we will examine the rate of divergence of the sequence $\{f^n\}$ for the case where $g \notin \mathcal{R}(K)$ in order to develop some understanding of how useful Richardson's algorithm can be if g_N is small, but nonzero.

From equation (3.47), we have that

$$f^n = a_1^n + a_2^n + a_3^n + a_4^n. \tag{3.72}$$

Assume that $\rho(I_{11} - \theta J_{11}) < 1$. Then

$$\lim_{n\to\infty} a_1^n = 0, \tag{3.73}$$

and

$$\lim_{n\to\infty} a_2^n = B^{\cdot 1}J_{11}^{-1}B_{1\cdot}g \equiv \tilde{f}, \tag{3.74}$$

where $K\tilde{f} \overset{\text{gen}}{=} g$, since from (3.55) we see that

$$\begin{aligned} K\tilde{f} &= (B^{\cdot 1}J_{11}B_{1\cdot} + B^{\cdot 2}N_{22}B_{2\cdot})B^{\cdot 1}J_{11}^{-1}B_{1\cdot}g \\ &= B^{\cdot 1}J_{11}B_{1\cdot}B^{\cdot 1}J_{11}^{-1}B_{1\cdot}g = Vg = g_R. \end{aligned} \tag{3.75}$$

The sequence a_1^n converges to zero at the rate $\rho\left[(I_{11} - J_{11})^n\right]$.

The remaining terms a_3^n and a_4^n do not, in general, have finite limits. If we choose $f^0 \in \mathcal{R}(K^{\iota-1})$, which we can always do by taking $f^0 = 0$, then a_3^n is equal to zero for all n. However, a_4^n can increase without bound if $g_N \neq 0$.

The maximum rates at which the terms a_3^n and a_4^n can go to infinity follow from the following results:

$$(I_{22} - \theta N_{22})^n = \sum_{i=0}^{\min(\iota-1,n)} \binom{n}{i} (-\theta N_{22})^i = O(n^{\iota-1}) \tag{3.76}$$

and

$$\sum_{i=0}^{n-1}(I_{22} - \theta N_{22})^i = \sum_{j=0}^{\min(\iota-1,n-1)} \binom{n}{j+1} (-\theta N_{22})^j = O(n^{\iota}), \tag{3.77}$$

where the order symbol $O(\cdot)$ is to be interpreted for each element of a matrix. Hence if $f^0 \in \mathcal{R}(K^{\iota-1})$, we see that

$$\|g - Kf^n\| = \|g_N + (g_R - Kf^n)\| \leq \|g_N\| + \|g_R - Kf^n\| \sim O(n^{\iota})\|g_N\|. \tag{3.78}$$

If the iteration is terminated early, g_N is sufficiently small, and ι is not too large, then it will often be the case that f^n will be a near-solution when the iteration is terminated: even though, ultimately, $f^n \to \infty$. A similar argument can be made for the case where f^0 has a small component not in $\mathcal{R}(K^{\iota-1})$. In other words, Richardson's algorithm is somewhat robust to violation of the requirements that $g \in \mathcal{R}(K^{\iota})$ and $f^0 \in \mathcal{R}(K^{\iota-1})$. This is reassuring, since for discretizations of integral equations of the first kind g_N is likely to be small, but nonzero.

On the other hand, if $\rho(I_{11} - \theta J_{11}) = \rho_0 > 1$ then $a_1^n + a_2^n$ will diverge at the exponential rate ρ_0^n. The situation where $\rho_0 = 1$ is complicated, since whether a_1^n converges to zero, and whether a_2^n converges at all, depends on the particular value of f^0 and g, respectively. For a given vector z, $(I_{11} - \theta J_{11})^n z$ can converge to zero, converge to a nonzero vector, or else not converge at all, depending on the choice of z and the subspace of $\mathcal{R}(I_{11} - \theta J_{11})$ which has eigenvalues with moduli greater than or equal to one. It is difficult to make a general statement about the $\rho_0 = 1$ case, but this situation is not likely to be important in numerical practice.

Conditions on θ for Which $\rho(I - \theta K) < 1$

The condition $\rho(I_{11} - \theta J_{11}) < 1$ involves both the nonzero eigenvalues of K and the constant θ. We will assume, without loss of generality, that $\theta > 0$. Let the nonzero eigenvalues of K be denoted $\{\lambda_i\}_{i=1}^s$, and let $\mu_i \equiv 1 - \theta\lambda_i$. Conditions on $\{\lambda_i\}_{i=1}^s$ and θ which lead to $\max_i |\mu_i| < 1$ are given in the following lemma:

Lemma 3.1.2 *Let $\{\lambda_i\}_{i=1}^s$ be a set of complex numbers, and let $\mu_i \equiv 1 - \theta\lambda_i$, for $i = 1, \ldots, s$. The following conditions together imply that*

$$\max_i |\mu_i| < 1 : \tag{3.79}$$

1. *For $i = 1, \ldots, s$, $\Re\lambda_i > 0$, and*

2. $$0 < \theta < \min_i \frac{2\Re\lambda_i}{|\lambda_i|^2}. \tag{3.80}$$

Proof: Assume $\Re\lambda_i > 0$ for each i. The condition that the $\{\mu_i\}_{i=1}^{s}$ be in the interior of the unit circle is equivalent to

$$|\mu_i|^2 < 1 \iff [1 - \theta(\Re\lambda_i)]^2 + \theta^2(\Im\lambda_i)^2 < 1,$$

or

$$0 < \theta < \frac{2\Re\lambda_i}{|\lambda_i|^2}. \tag{3.81}$$

Since (3.81) must hold for all i, we have the condition (3.80). ∎

The Nullspace of $G = I - \theta K$ When $\theta K > 0$ and $\rho(\theta K) = 1$

We consider next the special case of Richardson's algorithm (3.3) applied to a matrix equation $Kf = g$ for which K is positive. A positive matrix is a matrix for which all of the elements are positive, and we write $K > 0$. From Lemma 3.1.2 and Theorem 3.1.2, it is clear that it is very desirable for the nonzero eigenvalues of K to have positive real parts, since if this is not the case, then there exists no θ for which the sufficient conditions of Theorem 3.1.2 and the necessary and sufficient conditions of Theorem 3.1.1 and Corollary 3.1.1 for Richardson's algorithm will be satisfied. So we will assume that in addition to K being positive, all of the nonzero eigenvalues of K have positive real parts.

By the Perron-Frobenius theorem (Theorem 2.1.3), the largest eigenvalue of K in magnitude is positive and equal to $\rho(K)$, all other eigenvalues have modulus less than $\rho(K)$, and the corresponding Perron-Frobenius eigenvector is a positive vector z. By selecting

$$\theta = 1/\rho(K), \tag{3.82}$$

we have $\rho(\theta K) = 1$. Then $G = I - \theta K$ has one eigenvalue equal to zero and all other eigenvalues of G have positive real parts.

The matrix θK^T is also positive, with $\rho(\theta K^T) = 1$. The Perron-Frobenius theorem implies that there exists a positive vector y which is an eigenvector of θK^T corresponding to the eigenvalue one. Since

$$y^T \theta K = y^T, \tag{3.83}$$

it is customary to refer to y^T as a *left eigenvector* of K and to z as a *right eigenvector* of K, both corresponding to the same eigenvalue. When there is no risk of confusion, we will continue to refer to right eigenvectors simply as eigenvectors.

The main result of this subsection is clarification of the role of the positive eigenvalue and corresponding left and right eigenvectors in Richardson's algorithm. In order to establish this result, we need to build on the geometry of Lemma 3.1.1, and we do this next for a general square matrix. Later we will specialize to positive K.

Assume that K is an $m \times m$ matrix and that the Jordan form of K consists of $r \leq m$ Jordan blocks J_{ii}. Let

$$J = \operatorname{diag}(J_{11}, J_{22}, \ldots, J_{rr}), \tag{3.84}$$

where we order these blocks so that the corresponding eigenvalues,

$$|\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_r| \geq 0, \tag{3.85}$$

are in order of decreasing modulus.

Using notation similar to that of Section 3.1.2, we can represent K as

$$K = B^{-1} J B, \tag{3.86}$$

where

$$B^{-1} = \left[B^{\cdot 1} B^{\cdot 2} \ldots B^{\cdot r}\right] \tag{3.87}$$

and

$$B = \begin{bmatrix} B_{1\cdot} \\ B_{2\cdot} \\ \cdot \\ B_{r\cdot} \end{bmatrix}. \tag{3.88}$$

Then (3.86) becomes

$$K = \sum_{i=1}^{r} B^{\cdot i} J_{ii} B_{i\cdot}. \tag{3.89}$$

Because

$$BB^{-1} = B^{-1}B = I, \tag{3.90}$$

we have the following relations among the components of the partitioned matrices (3.87) and (3.88):

$$\sum_{i=1}^{r} B^{\cdot i} B_{i\cdot} = I, \tag{3.91}$$

$$B_{i\cdot} B^{\cdot i} = I_{ii}, \tag{3.92}$$

and, for $i \neq j$,

$$B_{i\cdot} B^{\cdot j} = 0_{ij}, \tag{3.93}$$

where the I_{ii} and 0_{ij} are identity and zero matrices, respectively, of the appropriate dimensions. Because of these relations, we can easily generalize Lemma 3.1.1 to consider projections onto the r subspaces corresponding to the Jordan representation (3.86). In particular, we have the r projection matrices

$$V_i \equiv B^{\cdot i} B_{i\cdot}, \tag{3.94}$$

for $i = 1, \ldots, r$, where $V_i^2 = V_i$ and, for $i \neq j$, $V_i V_j = 0_{ij}$. For any vector $x \in C^m$, we have

$$x = \sum_{i=1}^{r} V_i x \equiv \sum_{i=1}^{r} x_i, \tag{3.95}$$

where for each i, the vector x_i is the projection of x onto $\mathcal{R}(V_i)$. It is important to note that these projections are, in general, *not* orthogonal.

Now assume that K is positive, and that all nonzero eigenvalues of K have positive real parts. Because of the Perron-Frobenius theorem, $\lambda_1 = 1/\theta$ is larger in modulus than all other eigenvalues of K, and λ_1 has algebraic multiplicity one. Hence V_1 is a matrix of rank one. In fact, it is not difficult to show that

$$V_1 \propto ozy^T. \tag{3.96}$$

Since

$$G = I - \theta K = B^{-1}(I - \theta J)B, \tag{3.97}$$

the subspace onto which V_1 projects corresponds to an eigenvalue of G which equals $1 - \theta/\theta = 0$, and this eigenvalue has multiplicity one.

Next, let f be a solution to the consistent matrix equation $Kf = g$, and assume that Richardson's algorithm (3.3) converges to f. Let the discrepancy be

$$u^n \equiv f - f^n, \tag{3.98}$$

and write Richardson's iteration (3.3) in the form

$$u^{n+1} = Gu^n, \tag{3.99}$$

where

$$G = I - \theta K \tag{3.100}$$

and $Gz = 0$.

We have the following decomposition of u^0 in terms of the subspaces $\{\mathcal{R}(V_i)\}_{i=1}^r$ corresponding to G:

$$u^0 = \sum_{i=1}^r V_i u^0 = \sum_{i=1}^r B^{\cdot i} B_{i\cdot} u^0. \tag{3.101}$$

Since V_1 corresponds to the zero eigenvalue of G, we have that

$$\begin{aligned} u^n = G^n u^0 &= \sum_{i=1}^r B^{\cdot i}(I - \theta J_{ii})^n B_{i\cdot} B^{\cdot i} B_{i\cdot} u^0 \\ &= \sum_{i=2}^r B^{\cdot i}(I - \theta J_{ii})^n B_{i\cdot} u^0. \end{aligned} \tag{3.102}$$

Hence, for all $n > 0$,

$$V_1 u^n = zy^T u^n = 0. \tag{3.103}$$

Since $z > 0$, this implies that $y^T u^n = 0$ for all $n > 0$.

What does this tell us? A weighted sum of the components of $f - f^n$ equals zero for each $n > 0$, with the weights corresponding to the positive *left* eigenvector of K. We can easily calculate this vector for a given problem, and this might lead to insight into how well Richardson's iteration can be expected to perform. However, to calculate the left Perron-Frobenius eigenvector is of roughly the same order of difficulty as the iteration itself.

If θK is *stochastic* (recall that a stochastic matrix is a nonnegative matrix for which the elements in each row sum to one) then θK is the transition matrix of some Markov chain, where the left (positive) eigenvector of θK is proportional to the stationary distribution of this chain, and the right eigenvector is positive and constant (e.g., Horn and Johnson, 1985, 487-489). We have shown above that, for all $n > 0$,

$$E(u^n) = 0, \tag{3.104}$$

which implies that

$$E(u^{n+1} - u^n) = E(\delta^n) = 0, \tag{3.105}$$

where the expectations are with respect to the stationary distribution of the Markov chain corresponding to θK.

If θK is *symmetric and stochastic*, then then both z and y equal a constant vector, so the sum of the components of $f - f^n$ will be zero for all positive n. In many situations, when the sum of the components of u^n equals zero, we will have $\|u^n\|$ small.

3.2 Stochastic Preconditioning and the Conditional Expectation Algorithm

When an ill-posed integral equation is discretized, the matrix equation which results will have many eigenvalues with small absolute values. Because of this, $G = I - \theta K$ will have many eigenvalues at, or near, one in the complex plane. We have seen in the previous section that the convergence of Richardson's algorithm (3.3) in the direction of an eigenvector corresponding to an eigenvalue λ of G for which $|\lambda| < 1$ and $|\lambda| \approx 1$ will be slow, since the convergence rate in the direction of this eigenvector is governed by the powers of λ.

This ultimate slow convergence is both an advantage and a disadvantage. It is advantageous to not rapidly approach a 'solution' which, because of noise, is neither smooth nor near any solution to the corresponding integral equation. Of course, it is also advantageous for the iterates to not diverge rapidly if the matrix equation is inconsistent. But it is disadvantageous to use an iteration for which the convergence becomes very slow when the distance $\|g - Kf^n\|$ is still unacceptably large.

At the beginning of this chapter, we mentioned the notion of *preconditioning* so as to accelerate convergence. The idea is to choose a nonsingular matrix D so that the iteration (3.2), repeated here for convenient reference,

$$f^{n+1} = f^n + \theta D(g - Kf^n), \quad \text{for } n = 0, 1, 2, \ldots, \tag{3.106}$$

converges rapidly, at least initially. We will restrict attention, for the most part, to nonsingular diagonal preconditioning matrices and to square matrices K with positive elements. We will provide several motivations for choosing D so that if K is positive, then DK is stochastic. We will refer to this form of preconditioning as *stochastic preconditioning* and to the algorithm which results, along with its nonlinear generalizations, as the *Conditional Expectation* algorithm. The effectiveness of this approach will be illustrated through examples in this and subsequent chapters. Stochastic matrices are relevant in the theory of Markov chains and stochastic processes, so the presence of a stochastic matrix here is a hint that a natural probabilistic interpretation of this preconditioned algorithm should be possible.

3.2.1 A Property of Positive Definite Preconditioning Matrices

Lemma 3.1.2 implies that if the nonzero eigenvalues of a matrix K have positive real parts, and if the positive constant θ is sufficiently small, then the eigenvalues of $I - \theta K$ which are not equal to one will be in the interior of the unit circle. It is therefore a desirable property of a preconditioning matrix D that if all of the eigenvalues of K have nonnegative real parts, then the eigenvalues of DK have nonnegative real parts as well. We demonstrate below that positive definite preconditioning matrices have this 'nonnegative real part preserving' property.

Theorem 3.2.1 *Let K be a square matrix, and assume that all of the eigenvalues of K have nonnegative real parts. Let D be positive definite. Then all of the eigenvalues of DK also have nonnegative real parts.*

Proof: Write K in the form

$$K = (K + K^*)/2 + i(K - K^*)/(2i) \equiv K_1 + iK_2. \tag{3.107}$$

Let λ be an arbitrary eigenvalue of K, and let x be a corresponding normalized eigenvector. Then

$$\lambda = x^*Kx = x^*K_1x + ix^*K_2x. \tag{3.108}$$

The matrices K_1 and K_2 are Hermitian, and so x^*K_1x and x^*K_2x are both real numbers. It follows that x^*K_1x and x^*K_2x are equal to $\Re\lambda \geq 0$ and $\Im\lambda$, respectively.

Since D is positive definite (hence, Hermitian), D has a positive definite Hermitian square root (Strang, 1976, p. 241). For $i = 1, 2$,

$$DK_i = D^{1/2}\left[D^{1/2}K_iD^{1/2}\right]D^{-1/2}, \tag{3.109}$$

and so DK_i is similar to $D^{1/2}K_iD^{1/2}$, which is congruent to K_i. Thus DK_i has the same eigenvalues as $D^{1/2}K_iD^{1/2}$. By congruence (Lemma 2.1.1), the eigenvalues of $D^{1/2}K_1D^{1/2}$ are nonnegative. Because $D^{1/2}K_2D^{1/2}$ is Hermitian, it has real eigenvalues. Thus, the eigenvalues of DK_1 are nonnegative and those of DK_2 are real. It follows that the eigenvalues of DK have nonnegative real parts. ∎

3.2.2 Positive, Bounded Kernels and Stochastic Matrices

Consider the integral equation of the first kind

$$\int_0^1 k(x,y)f(y)dy = g(x), \tag{3.110}$$

where we assume that the kernel, $k(x,y)$, is positive and bounded. We discretize this equation as discussed in Appendix A. This gives a matrix equation $Kf = g$. Let D be the diagonal matrix corresponding to stochastic preconditioning, that is assume that

$$\tilde{K} \equiv DK \tag{3.111}$$

is stochastic.

If the integral, in y, of the kernel of (3.110) were equal to one for each x, then this integral equation, once discretized, would lead to a matrix equation having a *nearly* stochastic matrix. The reason why the matrix might not be exactly stochastic is that the row sums for the discretized problem are numerical *approximations* to integrals. We transform (3.110) into a new equation, having the same solution, as follows:

$$\int_0^1 \tilde{k}(x,y)f(y)dy = \tilde{g}(x), \tag{3.112}$$

where

$$\tilde{k}(x,y) \equiv \frac{k(x,y)}{\int_0^1 k(x,y)dy}, \tag{3.113}$$

and

$$\tilde{g}(x) \equiv \frac{g(x)}{\int_0^1 k(x,y)dy}. \tag{3.114}$$

There are two slightly different approaches to applying Richardson's algorithm (3.106) with stochastic preconditioning. One way is to normalize the equation as in (3.112), and to discretize this transformed equation. Richardson's algorithm, (3.3), could then be applied, with $\theta = 1$. The other approach is to discretize (3.110), and then find the matrix D satisfying (3.111). The preconditioned Richardson algorithm, (3.106), could then be applied, for this choice of D and with $\theta = 1$. The second approach is the more desirable one for two reasons: the normalized kernel (3.113) need only be determined numerically, and the matrix of the resulting discretized equation is *exactly* stochastic.

3.2.3 Some Heuristic Motivations for Stochastic Preconditioning

We will make a case for stochastic preconditioning through several heuristic motivations, and we will illustrate this form of preconditioning in a later section, and in later chapters, with several examples. At this time, a complete understanding of why and when this form of preconditioning works well is not available. The heuristic motivations below suggest directions one might follow in order to attempt to answer these questions. For now, the real justification for our choice of preconditioning comes not from theory, but from the study of examples.

A Motivation Provided by Condition Numbers

From the point of view of numerical analysis, scaling a positive matrix so that it becomes a stochastic matrix tends to make the matrix better conditioned. The following is a special case of a theorem proved by Van der Sluis (1969, p.18):

Theorem 3.2.2 *Let K be a nonsingular positive matrix, and let $\|\cdot\|_*$ be either the l_2 or the l_∞ norm. Let D be a nonsingular diagonal matrix. Then the following measures of the condition of DK are minimized when the rows of DK each sum to one:*

1. $\chi_1(DK) \equiv \|DK\|_\infty \|(DK)^{-1}\|_*$, *and*
2. $\chi_2(DK) \equiv \|DK\|_\infty / \|DK\|_*$.

Although χ_1 and χ_2 each differs from the usual condition number based on the spectral norm, $\kappa \equiv \sigma_1/\sigma_m$, all three quantities are reasonable measures of the condition of a matrix. A preconditioning which minimizes χ_1 and χ_2 can be expected to usually reduce κ as well. In fact, if $\|\cdot\|_\infty$ is chosen for $\|\cdot\|_*$ in χ_1, then χ_1 becomes the condition number of a matrix, with respect to the infinity norm. In Section 2.2.3, we showed how a condition number relates changes in a right hand side to corresponding changes in a solution of a matrix equation. It follows from this that the smaller a condition number is, the smaller the change in the solution will be for a given change in right hand side, and so one would expect Richardson's algorithm to converge more rapidly for matrices with relatively small condition numbers.

A Taylor Series Motivation

Assume that (3.110) has a solution f, that $k(x,y)$ has a peak with location on a smooth, monotone curve $y = v(x)$ in the unit square, where $v(0) = 0$ and $v(1) = 1$. Let f^n be an approximation to f at the nth iteration, let $u^n = f - f^n$, and note that

$$\int_0^1 k(x,y)[f^n(y) + u^n(y)]dy = g(x), \tag{3.115}$$

where u^n is now the unknown. Assume that u^n has a Taylor series expansion for all y. Expand u^n about $v(x)$, keeping only the first term:

$$u^n[v(x)] \int_0^1 k(x,y)dy \approx g(x) - \int_0^1 k(x,y)f^n(y)dy, \tag{3.116}$$

or

$$\tilde{u}^n[v(x)] = \frac{g(x) - \int_0^1 k(x,y)f^n(y)dy}{\int_0^1 k(x,y)dy}. \tag{3.117}$$

If we let

$$f^{n+1}(x) \equiv f^n(x) + \bar{u}^n(x), \tag{3.118}$$

then, in the special case where $v(x) = x$, we have the L_2 version of Richardson's algorithm with stochastic preconditioning. If $v(x)$ is not the identity, then a change of variable in x reduces the problem to the special case.

A Probabilistic Motivation

A simple probabilistic argument provides another motivation for stochastic preconditioning. Since k is bounded and positive, it is proportional to the joint density of two random variables, say X and Y. We write this as

$$\pi_{X,Y}(x,y) \equiv ck(x,y), \tag{3.119}$$

where the constant c is

$$c = \left[\int_0^1 \int_0^1 k(x,y)dxdy\right]^{-1}. \tag{3.120}$$

The normalized kernel (3.113) is exactly the *conditional density* of the random variable Y given the random variable X:

$$\pi_{Y|X}(y|x) = \frac{\pi_{X,Y}(x,y)}{\int_0^1 \pi_{X,Y}(x,y)dy} = \tilde{k}(x,y). \tag{3.121}$$

Richardson's algorithm applied to (3.112) with $\theta = 1$ is

$$f^{n+1}(x) = f^n(x) + \int_0^1 \tilde{k}(x,y)(f(y) - f^n(y))dy. \tag{3.122}$$

Since the integral on the right hand side of (3.122) can be interpreted as the conditional expectation of the difference $f - f^n$, we can rewrite (3.122) (in terms of the *random variables X and Y*) as

$$f^{n+1}(X) - f^n(X) = E\left[f(Y) - f^n(Y)\,|X\right]. \tag{3.123}$$

In words: the nth step in this Richardson algorithm with stochastic preconditioning is the *conditional expectation of the difference between the solution and the approximation* f^n. Because of this, we will sometimes refer to Richardson's algorithm with stochastic preconditioning as the *Conditional Expectation* algorithm.

This probabilistic interpretation suggests that this preconditioned Richardson algorithm will converge rapidly when the conditional expectation, with respect to the density (3.121), of $f - f^n$ is nearly equal to $f - f^n$. This will occur when $Y \approx X$. For these random variables to be nearly equal, the original kernel $k(x,y)$ must be peaked about the line $y = x$. The more this kernel is peaked, the more rapidly convergent this preconditioned Richardson algorithm will be.

In fact, if $Y \approx h(X)$, for some monotone function h, then by defining $Z \equiv h(X)$, we have $Z \approx Y$, and so we have reduced the problem to the case considered in the previous paragraph.

A Motivation Based on Convolution Kernels

Consider the Fredholm integral equation of the first kind

$$\int_{-\infty}^{\infty} k(x-y)f(y)dy = g(x), \tag{3.124}$$

where the *convolution kernel* $k(x-y)$ is positive, bounded, and

$$\int_{-\infty}^{\infty} k(x-y)y^s dy < \infty, \tag{3.125}$$

for all $s \geq 0$.

Let $\pi_t(x)$ represent a polynomial of degree at most t. For any integer t, we have that

$$\begin{aligned} \int_{-\infty}^{\infty} k(x-y)y^t dy &= \int_{-\infty}^{\infty} k(y)(x-y)^t dy \qquad (3.126) \\ &= x^t \int_{-\infty}^{\infty} k(y)dy + \pi_{t-1}(x). \end{aligned}$$

If we transform the equation (3.124) so that the kernel of the transformed equation is

$$\tilde{k}(x-y) \equiv \frac{k(x-y)}{\int_{-\infty}^{\infty} k(y)dy}, \tag{3.127}$$

we have that

$$\int_{-\infty}^{\infty} \tilde{k}(x-y)y^t dy = x^t + \pi_{t-1}(x). \tag{3.128}$$

It is easy to see that, if $u^n = f - f^n$ is a polynomial of degree t, then the preconditioned Richardson algorithm (3.106), applied to the convolution equation (3.124) with stochastic preconditioning, and with $\theta = 1$, will *exactly* converge in at most t iterations, reducing the degree of u^n by at least one with each successive iteration.

To the extent that there is a function u^n for a specific problem which is well approximated by a low order polynomial, and to the extent that the normalized kernel for a specific problem is well approximated by a convolution, one would expect Richardson's algorithm with stochastic preconditioning to converge rapidly.

3.3 Richardson's Algorithm and Iterative Regularization

Integration tends to smooth. It is clear, therefore, that the Richardson iterations (3.3) and (3.2) will tend to produce smooth iterates when applied to integral equations of the first kind having smooth kernels. There is regularization *implicit* in using Richardson's algorithm, and it is the purpose of this section to examine the nature of this regularization for the special case of the Richardson algorithm (3.3), applied to matrix equations with matrices having a *diagonalizable nullspace*. We have in mind matrix equations which arise from the discretization of ill-posed integral equations of the first kind, so that the more oscillatory eigenvectors correspond to the many small eigenvalues of the matrix.

3.3.1 Regularization Methods

One approach to 'solving' a matrix equation $Kf = g$ which is the discretization of an ill-posed linear operator equation is the *method of regularization* of Tikhonov (1962) and Phillips (1963) (see also Tikhonov and Arsenin, 1977, and Groetsch 1984). The basic idea is very simple. We do not want to solve any discretized version of an ill-posed equation exactly. Instead, we minimize the quadratic form

$$\tilde{U}(z) \equiv (Kz - g)^*(Kz - g) + \gamma z^* Lz, \tag{3.129}$$

where L is positive definite, and is chosen so that $z^{\mathrm{T}}Lz$ will tend to be large when z is not smooth. A positive constant, γ, determines the relative importance of the first (*least-squares*) and second (*penalty*) terms of the functional $\tilde{U}(z)$. When γ is zero, minimizing $\tilde{U}(z)$ is equivalent to minimizing $\|Kz - g\|$. As γ is increased, increasing weight is put on the smoothness of the solution, and less on 'fidelity' to the equation.

The quadratic form (3.129) is usually associated with the method of regularization. We will instead be concerned with the functional

$$U(z) \equiv (z - f)^*(z - f) + \gamma z^* Lz. \tag{3.130}$$

Although we will not require L to be positive definite, our motivation for choosing L is the same as in (3.129). Minimizing U, like minimizing $\tilde{U}$, involves a compromise between fidelity to the equation and smoothness of the solution. The difference is that the first term in $\tilde{U}$ measures how close the *right hand side* corresponding to an approximate solution is to g, while the first term of U compares the approximate solution to a solution vector f.

3.3.2 Regularization Implicit in Richardson's Algorithm

Consider the Richardson iteration (3.3) applied to equations $Kf = g$ for which K is an $m \times m$ matrix with a *diagonalizable nullspace* (Section 2.1.4). Assume that the iteration converges for a particular choice of θ. We have shown (Corollary 3.1.1) that since (3.3) converges, it must converge for any f^0, so we can choose f^0 arbitrarily, and denote the corresponding solution by f. We will show that, under these conditions,

$$\delta^n \equiv f^{n+1} - f^n = \theta(g - Kf^n) \tag{3.131}$$

is a stationary point of the quadratic form

$$Q(z) \equiv Q_{\mathrm{LS}}(z) + Q_{\mathrm{P}}(z) \equiv (u^n - z)^*(u^n - z) + z^*(K^{\#}/\theta - I)z, \tag{3.132}$$

where $K^{\#}$ is the group inverse of K, which exists and is unique since K has a diagonalizable nullspace, and

$$u^n \equiv f - f^n. \tag{3.133}$$

Differentiating $Q(z)$ with respect to z and setting this derivative equal to zero, we note that a stationary point z must satisfy the linear relationship

$$K^{\#}z - \theta u^n = 0. \tag{3.134}$$

The matrix K has index $\iota = 1$, and hence there exists a nonsingular matrix B such that

$$K = B^{-1}\begin{bmatrix} J & 0 \\ 0 & 0 \end{bmatrix}B, \tag{3.135}$$

where J is a Jordan form matrix of blocks corresponding to nonzero eigenvalues of K. The group inverse $K^{\#}$ is then

$$K^{\#} = B^{-1} \begin{bmatrix} J^{-1} & 0 \\ 0 & 0 \end{bmatrix} B, \tag{3.136}$$

and we have, in the notation of Section 3.1.2 and Lemma 3.1.1, that

$$K^{\#}K = KK^{\#} = B^{\cdot 1}JB_{1\cdot}B^{\cdot 1}J^{-1}B_{1\cdot} = B^{\cdot 1}B_{1\cdot} \equiv V. \tag{3.137}$$

Substitute δ^n for z in (3.134) and use Lemma 3.1.1 (d) to get

$$d \equiv K^{\#}z - \theta u^n = -\theta(I - K^{\#}K)u^n = -\theta(I - V)u^n \in \mathcal{N}(K). \tag{3.138}$$

We will show next that $u^n \in \mathcal{R}(K)$, so that, using once again Lemma 3.1.1, $u^n = Vh^n$ for some vector h^n, and hence $d = 0$.

The vector $f^n - f^0$ can be expressed as a sum of steps δ^i

$$f^n - f^0 = \textstyle\sum_{i=0}^{n-1} \delta^i \quad \text{for } n > 0, \tag{3.139}$$

where $\delta^i = \theta K u^i \in \mathcal{R}(K)$ for every i. Therefore $f^n - f^0 \in \mathcal{R}(K)$ for every n. Since

$$\lim_{n \to \infty} (f^n - f^0) = f - f^0, \tag{3.140}$$

$f - f^0 \in \mathcal{R}(K)$. Hence

$$u^n = f - f^n = (f - f^0) + (f^0 - f^n) \in \mathcal{R}(K), \tag{3.141}$$

which completes the proof that δ^n is a stationary point of (3.132).

Lemma 3.1.1 (c) implies that we can express f and f^0 as

$$f = Vf + (I - V)f \equiv f_R + f_N \tag{3.142}$$

and

$$f^0 = Vf^0 + (I - V)f^0 \equiv f_R^0 + f_N^0, \tag{3.143}$$

where V and $(I - V)$ project onto $\mathcal{R}(K)$ and $\mathcal{N}(K)$, respectively. Since $f^n - f^0 \in \mathcal{R}(K)$ we have, for all $n \geq 0$,

$$f_N^0 = (I - V)f^n = f_N. \tag{3.144}$$

We can obtain a simple expression for (3.132) evaluated at δ^n in terms of u^n and u^{n+1}. Since

$$u^n - \delta^n = (f - f^n) - (f^{n+1} - f^n) = u^{n+1}, \tag{3.145}$$

we see that

$$Q_{LS}(\delta^n) = \|u^{n+1}\|^2. \tag{3.146}$$

Using basic properties of the group inverse (Section 2.1.4), straightforward (though somewhat tedious) algebra leads to

$$Q(\delta^n) = (u^{n+1})^* u^n. \tag{3.147}$$

Note the similarity between (3.130) and (3.132). We have shown that each step (3.131) corresponds to solving a *penalized least squares* problem, where the penalty term Q_P is determined by the matrix K, and the 'least squares' term Q_{LS} is $\|u^n - \delta^n\|$, where $u^n = f - f^n$. Further discussion of the relationship between linear smoothers and penalized least squares can be found in Buja, et. al. (1989). The notion that there can be regularization implicit in iterative algorithms is apparently due to Bakushinskii (1967).

3.3.3 Positive Definite K

Although (3.2) does not make explicit use of regularization, at each iteration regularization is *implicit* in this algorithm and the character of this regularization is determined by the matrix K. To see how the second term in (3.132) can penalize 'rough' iterates, we consider the simple special case of K $m \times m$ positive definite, though with many small eigenvalues. Let the (positive) eigenvalues of K be $\{\lambda_i\}_{i=1}^m$ and let the corresponding orthonormal eigenvectors be $\{t_i\}_{i=1}^m$. By the spectral theorem (Theorem 2.1.1),

$$K = \sum_{i=1}^{m} \lambda_i t_i t_i^*, \tag{3.148}$$

where

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m > 0.$$

Assume that $f = K^{-1}g$, and $0 < \theta < 2/\lambda_1$, so that $f^n \to f$ for all f^0. Let the expansions of δ^n in terms of the eigenvectors of K be

$$\delta^n = \sum_{i=1}^{m} \beta_i^n t_i. \tag{3.149}$$

In terms of the spectral decomposition (3.148) of K, the penalty term at the minimum becomes

$$Q_P(\delta^n) = \delta^{n*}(K^{-1}/\theta - I)\delta^n = \sum_{i=1}^{m}(\beta_i^n)^2 \left[(\theta\lambda_i)^{-1} - 1\right]. \tag{3.150}$$

Since the matrix K is a discretization of a smooth function, the more oscillatory eigenvectors will correspond to small eigenvalues. Components of δ^n in the directions of these highly oscillatory eigenvectors will have a large contribution to the penalty term, hence the minimum of Q will tend to occur at a vector δ^n which has small components in the direction of the 'rougher' eigenvectors - that is, δ^n will tend to be smooth if K is smooth.

3.4 Examples

We illustrate the ideas of this chapter by considering linear Fredholm and Volterra examples.

3.4.1 A Fredholm Example

In Chapter 1, we introduced a Fredholm integral equation of the first kind,

$$\int_0^1 k_1(x,y)f(y)dy = g(x), \tag{3.151}$$

with kernel

$$k_1(x,y) \equiv \begin{cases} y(1-x) & y \leq x \\ x(1-y) & y > x \end{cases} \tag{3.152}$$

We continue the discussion of this example which began in that chapter.

An Eigenfunction Analysis

For the very simple example (3.151), we can determine the eigenfunctions and eigenvalues of both the kernel (3.152) and the preconditioned kernel

$$\tilde{k}_1(x,y) \equiv \frac{2k_1(x,y)}{x(1-x)} = \begin{cases} y/x & y \leq x \\ (1-y)/(1-x) & y > x \end{cases} \tag{3.153}$$

Here we are transforming the kernel, and then comparing the eigenfunctions of the transformed kernel (3.153) with the kernel (3.152). In numerical examples we will, as discussed in Section 3.2.2, discretize (3.152) to get a matrix equation, and then premultiply both sides of this equation by the appropriate matrix D, so that the matrix becomes stochastic.

Since (3.152) is the Green's function for the differential equation

$$\frac{d^2g}{dx^2} + f = 0, \tag{3.154}$$

subject to the boundary conditions

$$g(0) = g(1) = 0, \tag{3.155}$$

the eigenfunctions of (3.151) are the same as those of the differential equation (3.154), subject to the boundary conditions (3.155). That is, the tth eigenfunction of (3.151) is

$$\phi_t(x) = \sin(t\pi x), \tag{3.156}$$

and the corresponding eigenvalue is

$$\lambda_t = \frac{1}{\pi^2 t^2}. \tag{3.157}$$

The key to the eigenfunction analysis of the preconditioned Fredholm operator with kernel (3.153), is to note that

$$\int_0^1 \tilde{k}_1(x,y)y^t dy = \frac{2\sum_{i=0}^t x^i}{(t+1)(t+2)}. \tag{3.158}$$

It follows that tth degree polynomials are transformed into tth degree polynomials by the integral equation with kernel (3.153). It turns out that the eigenfunctions are polynomials,

$$\tilde{\phi}_t(x) = \sum_{i=0}^t \alpha_{t,i} x^i. \tag{3.159}$$

The eigenvalues are

$$\tilde{\lambda}_t = \frac{2}{t(t+1)}, \tag{3.160}$$

and the coefficients in (3.159) can be determined recursively from the formulas

$$\begin{aligned} \alpha_{t,t} &\equiv 1, \\ \alpha_{t,i} &= \frac{\alpha_{t,i+1}}{1-\tilde{\lambda}_i/\tilde{\lambda}_t}. \end{aligned} \tag{3.161}$$

After scaling (3.152) by multiplying by π^2, so that both (3.152) and (3.153) have largest eigenvalue one, we note that the eigenvalues corresponding to the preconditioned equation are substantially larger than the eigenvalues corresponding to the kernel (3.152), particularly for moderate t. In the numerical examples to follow, we will iteratively solve a matrix equation having a stochastic matrix, consequently the largest eigenvalue of this matrix will equal one.

Another thing to note from this example is that the 'character' of the eigenfunctions is completely changed - from trigonometric functions (all of which equal zero at the endpoints) to polynomials - by the stochastic preconditioning.

A Numerical Investigation

In this subsection, we describe some numerical results on the equation (3.151). The computations were performed using the S programming language (Becker, Chambers and Wilks, 1988), and a software listing is in Appendix B.

Let the right hand side of (3.151) be

$$g_1(x) = x^3(1-x)^2. \tag{3.162}$$

We discretize the integral equation (3.151) with kernel (3.152) and right hand side (3.162) using 50 point Gauss-Legendre quadrature as discussed in Appendix A. Let the matrix of this discretized equation be denoted K_1, and let the corresponding preconditioned matrix be $\tilde{K}_1$. The matrix $\tilde{K}_1$ is formed by discretizing the kernel (3.152) as in Appendix A, and then normalizing the rows of this matrix to each sum to one (see Section 3.2.2). The largest eigenvalue of K_1 is .1013913, which is approximately equal to π^{-2}, the largest eigenvalue of the corresponding integral equation. For the Richardson iteration *without* preconditioning (3.3), we take θ to equal the reciprocal of the largest eigenvalue, i.e. $\theta \approx 9.863$, so that the largest eigenvalue of θK_1 is (very nearly) equal to one. For the Conditional Expectation algorithm (Richardson's algorithm (3.2) with stochastic preconditioning) the largest eigenvalue is equal to one, so we let $\theta = 1$. We choose the initial iterate $f^0 = 0$ for now; we will consider the important role of f^0 for the algorithm without preconditioning below. Fifty iterations of both methods are compared in Figure 3.1. The preconditioned method gives an approximation very near the solution

$$f(x) = -20x^3 + 24x^2 - 6x \tag{3.163}$$

before the convergence rate begins to decrease dramatically. The method without preconditioning is still far from the solution at the 50th iteration, and, since by the 50th iteration the steps taken at each iteration are very small, it will take many iterations to get appreciably closer to the solution.

Another way of seeing the dramatic effect preconditioning has had on the convergence rate is to examine the distance, in l_2 norm, to the discretized solution as a function of the iteration index. This comparison is made in Figure 3.2a. In Figure 3.2b, we have plotted the residual norms $\|Kf^n - g\|$ (for the discretized functions, using the l_2 norm).

The eigenvectors of (3.152) are $\sin(l\pi x)$, which equal zero, for all l, at $x = 0$ and $x = 1$. However, $f(1) = -2$, so contributions from eigenvectors of K_1 corresponding to very small eigenvalues are required in order for Richardson's algorithm without preconditioning to closely approximate $f(x)$ near $x = 1$. The eigenfunctions (3.159, 3.161) corresponding to the Conditional Expectation algorithm are polynomials, and they do not all go to zero at the endpoints of [0, 1].

One might argue that the comparison in Figures 3.1 and 3.2 is unfair, since the eigenfunctions of (3.152) are ill-suited for approximating (3.163), at least when $f^0 = 0$. One way to compare the two algorithms on a more even footing is to use the starting function

$$f^0(x) = -2x, \tag{3.164}$$

so that $f^0(0) = f(0)$ and $f^0(1) = f(1)$. (Of course, in practice one usually does not know the value of the unknown function at the endpoints.) Fifty iterations of both algorithms, begining with the starting iterate (3.164), are displayed in Figure 3.3. The distance from the solution and residual norm, as functions of the iteration index, are given in Figure 3.4. The methods both perform reasonably well, with Richardson initially doing better, but with the Conditional Expectation algorithm 'catching up' after 30 or 40 iterations.

These two numerical examples each illustrate the notion of 'near-convergence' and 'near-solution'. The Conditional Expectation algorithm is able to provide smooth approximate solutions which are close to the solutions of the continuous problem (Figures 3.2a and 3.4a), and for which the corresponding residuals $\|Kf^n - g\|$ are small (Figures 3.2b and 3.4b). The Richardson algorithm also provided smooth iterates, although in the first example the Richardson approximations are very slowly convergent near $x = 1$.

Both the Richardson and the Conditional Expectation algorithms produce smooth approximate solutions even with the inevitable error in the right hand side. This is an instance of the idea of iterative regularization discussed in Section 3.3.1. However, eventually the approximations may become less smooth, as the components of the right hand side in the directions of eigenvectors corresponding to smaller eigenvalues begin to contribute. Since the right hand side for this example is smooth, and since preconditioning has reduced the condition number substantially (from 156261 to 810.34), it would take many iterations to observe the approximations depart from the true solution, and even then the deviation would be slight. In order to see an effect in a reasonable number of iterations, we add a component, with coefficient .01, in the direction of the 25th singular vector of the matrix $\tilde{K}_1$ to the right hand side (3.162). This leads to a perturbed right hand side, the Fourier coefficients of which are presented in Figure 3.5a, and a plot of which is given in Figure 3.5b. In Figure 3.6a, we display 50 iterations of the Conditional Expectation algorithm with this perturbed right hand side, and in Figure 3.6b, we give the solution of the matrix equation obtained by matrix inversion. Some obvious points to be made here include the oscillatory nature of the 25th singular vector, as reflected in the 'noisy' right hand side in Figure 3.5b, and the unpleasant solution in Figure 3.6b. In Figure 3.6a, we see a dramatic illustration of near-convergence, as the Conditional Expectation approximations stay reasonably close to the discretized solution to the (unperturbed) continuous problem. With some smoothing of the steps $f^{n+1} - f^n$ (as discussed in Chapter 7), much of the roughness of the approximations in Figure 3.6a can be eliminated. The distances of the approximate solutions from both the perturbed and unperturbed right hand sides are given in Figure 3.7a, and the corresponding residual norms are in Figure 3.7b. Notice that the approximations are closest in norm to this solution at the 8th iteration, and that the corresponding residual norm at the 8th iteration is fairly small. From that point on, the iterations move further away from the solution which corresponds to the *unperturbed* right hand side as they approach the exact solution, which corresponds to the *perturbed* right hand side. However, the residual norm with respect to the unperturbed right hand side continues to slowly decrease until about the 30th iteration.

3.4.2 A Volterra Example

As an example of a Volterra equation,

$$\int_0^x k_2(x,y)f(y)dy = g_2(x), \tag{3.165}$$

we take the differentiation problem, with kernel

$$k_2(x,y) \equiv \begin{cases} (x-y)^\alpha & \text{for } y \le x \\ 0 & \text{for } y > x \end{cases}, \tag{3.166}$$

for $\alpha > -1$, and with right hand side given by the power series

$$g_2(x) = \sum_{s=0}^{\infty} a_s x^s. \tag{3.167}$$

If α is a nonnegative integer, then the solution to this equation is

$$f(x) = g^{(\alpha+1)}(x), \tag{3.168}$$

where

$$g(0) = g'(0) = \cdots = g^{(\alpha)}(0) = 0. \tag{3.169}$$

This example is useful because it is easy to examine the Conditional Expectation algorithm analytically.

To precondition the kernel, we divide (3.166) by

$$h(x) \equiv \int_0^1 k_2(x,y)dy = \int_0^x (x-y)^\alpha dy = \frac{x^{\alpha+1}}{\alpha+1}, \tag{3.170}$$

and we denote the quotient $\tilde{k}_2(x,y)$. For any $t > 0$

$$\int_0^x \tilde{k}_2(x,y)y^t dy = \frac{\Gamma(\alpha+2)\Gamma(t+1)}{\Gamma(\alpha+t+2)}x^t, \tag{3.171}$$

hence the eigenfunctions of the preconditioned kernel are the powers

$$\psi_t(x) = x^t \tag{3.172}$$

for $t = 0, 1, \ldots$, and the corresponding eigenvalues are

$$\nu_t = \frac{\Gamma(\alpha+2)\Gamma(t+1)}{\Gamma(\alpha+t+2)}. \tag{3.173}$$

For example, let $\alpha = 0$. A little algebra shows that, if $g(x) = x^{s+1}/(s+1)$, then the corresponding f^n are given by

$$f^n(x) = [1 - (1 - 1/(s+1))^n]x^s. \tag{3.174}$$

Without preconditioning, it is easy to show that the Richardson iteration does not converge for this example, regardless of θ. From the linearity of the Volterra integral operator and (3.174) we see that, for the right hand side (3.167),

$$f^n(x) = \sum_{s=1}^{\infty} s a_s [1 - (1 - 1/s)^n] x^{s-1}. \tag{3.175}$$

If g is a smooth function plus noise, then f^n will reflect the smooth components initially, since these will correspond to fairly small values of s. Eventually, the solution will become rougher, but only when $(1 - 1/s)^n$ becomes small for fairly large s.

Numerical experimentation suggests that, for reasonably smooth right hand sides, the iterative algorithm outlined in this section can be useful for numerical differentiation. We will discuss the numerical solution of an equation related to the Volterra equation with kernel $k_2(x,y)$ with $\alpha = 1/2$ in Chapter 7.

Figure 3.1a: Fifty Iterations of Richardson's Algorithm Without Preconditioning

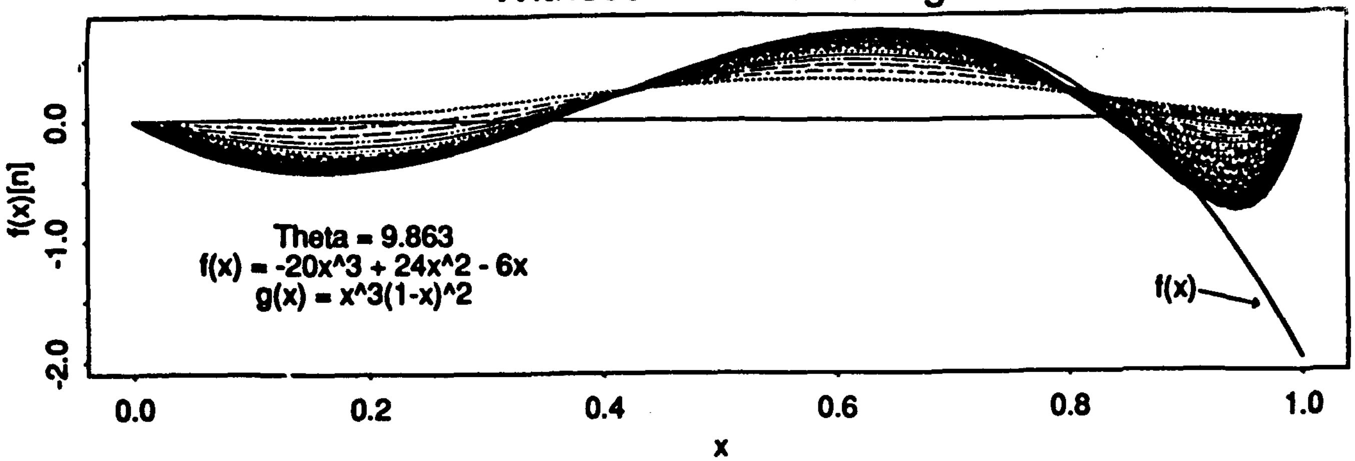


Figure 3.1b: Fifty Iterations of The Conditional Expectation Algorithm

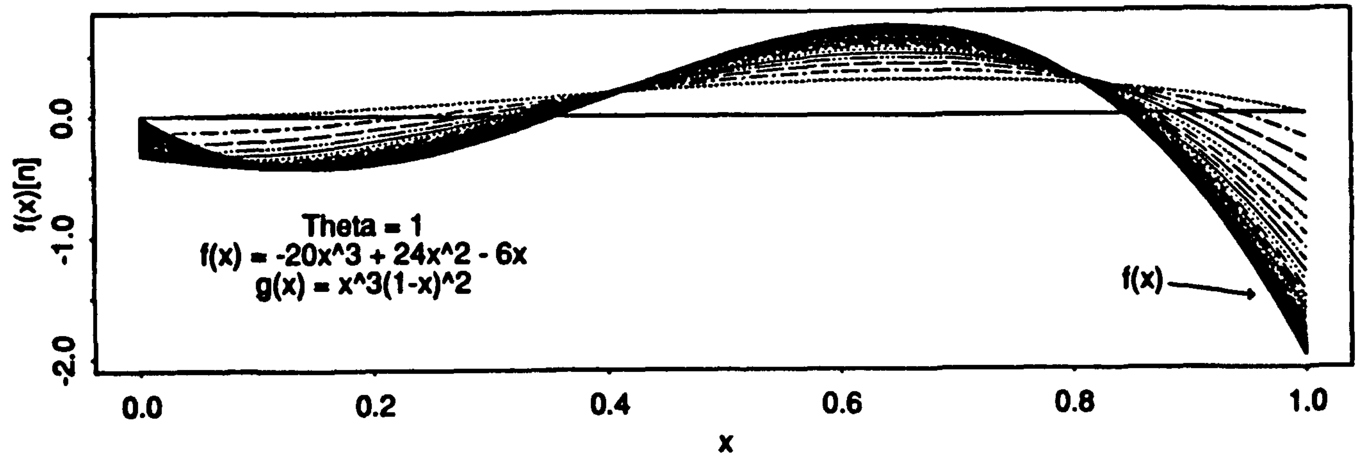


Figure 3.2a: Convergence to Solution for Richardson and Conditional Expectation Algorithms

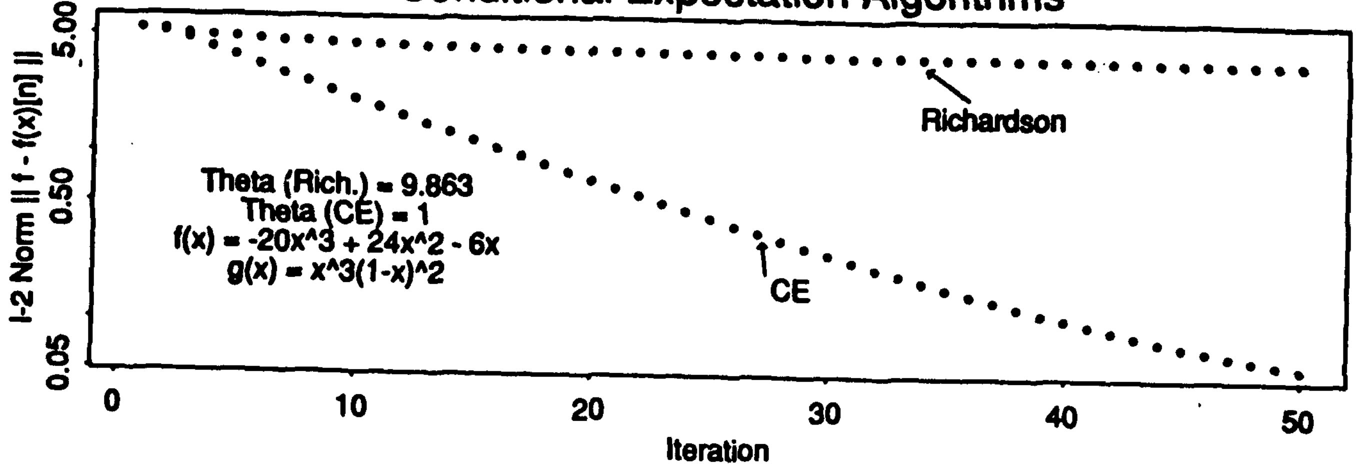


Figure 3.2b: Convergence to Right Hand Side for Richardson and Conditional Expectation Algorithms

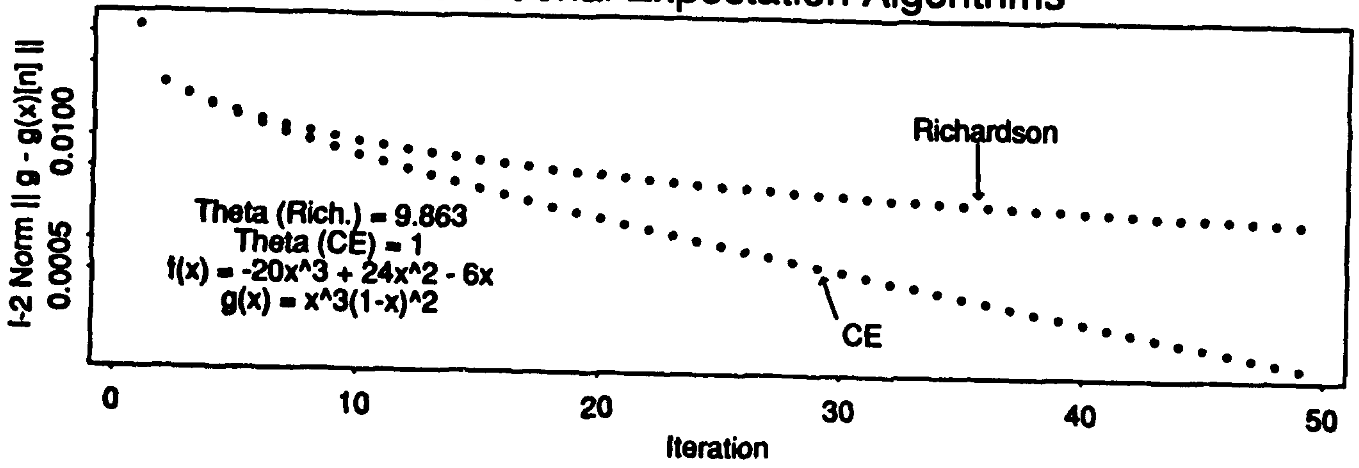


Figure 3.3a: Fifty Iterations of Richardson's Algorithm Without Preconditioning (f(x)[0] = -2x)

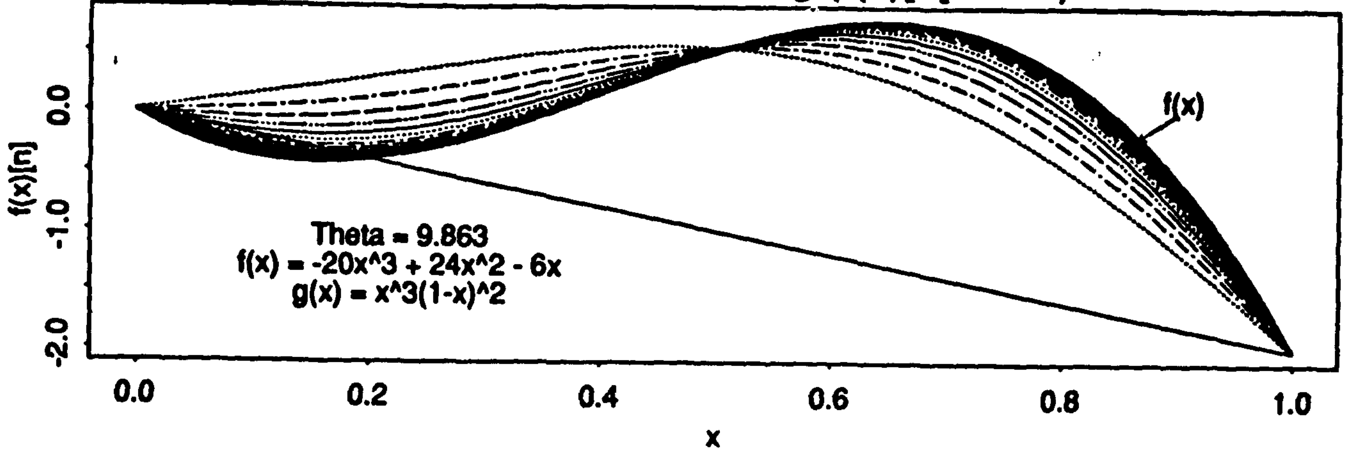


Figure 3.3b: Fifty Iterations of the Conditional Expectation Algorithm (f(x)[0] = -2x)

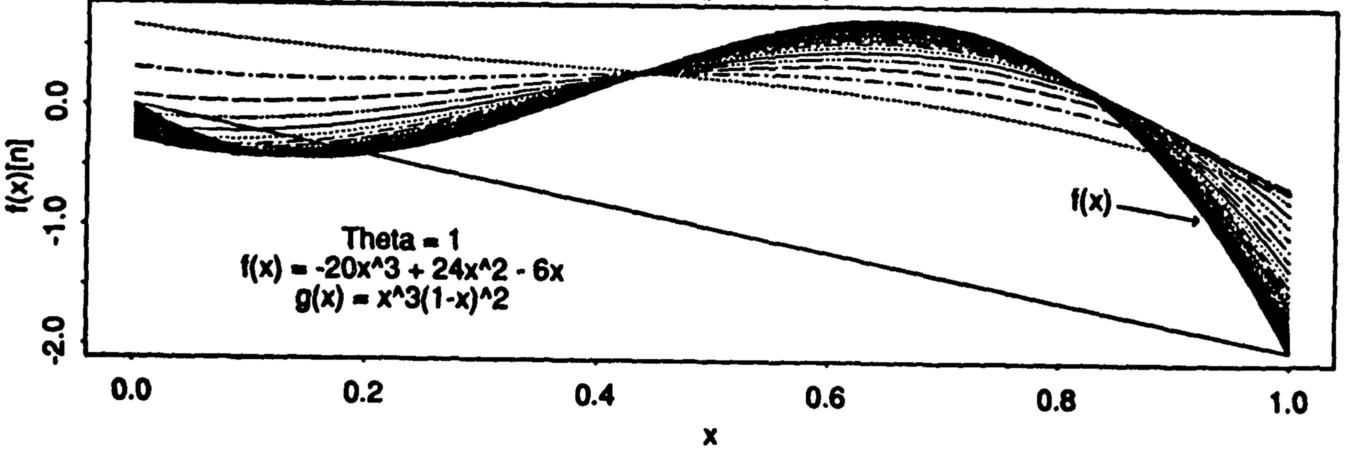


Figure 3.4a: Convergence to Solution for Richardson and Conditional Expectation Algorithms (f(x)[0] = -2x)

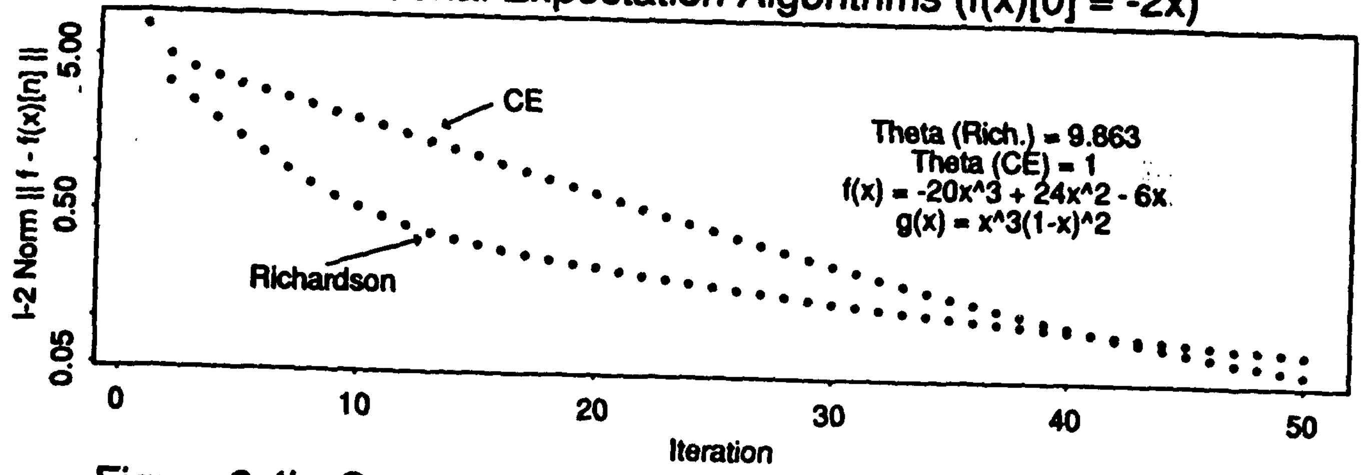


Figure 3.4b: Convergence to Right Hand Side for Richardson and Conditional Expectation Algorithms (f(x)[0] = -2x)

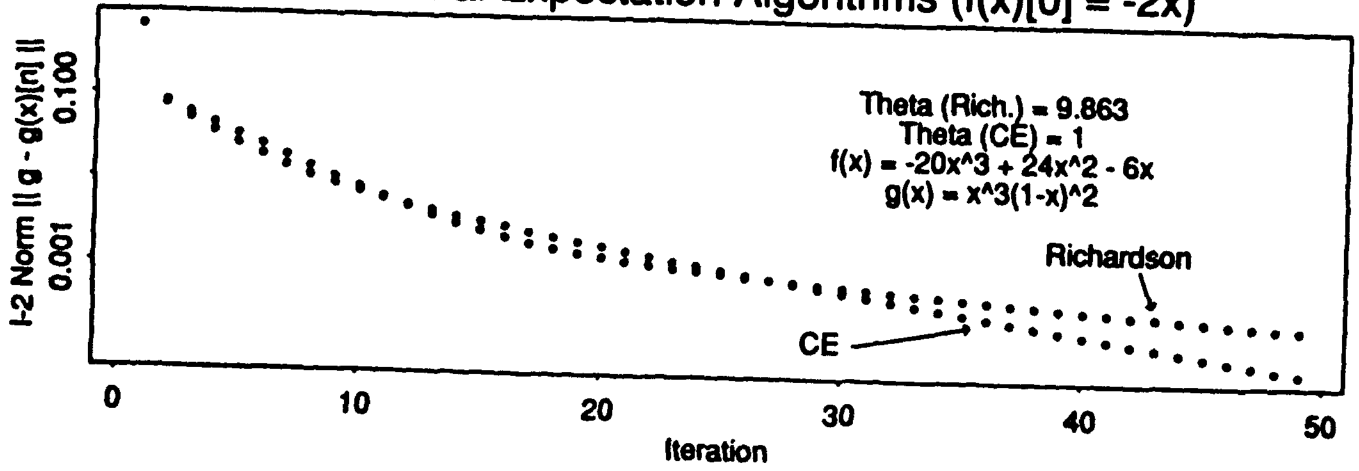


Figure 3.5a: Fourier Coefficients of Perturbed Right Hand Side

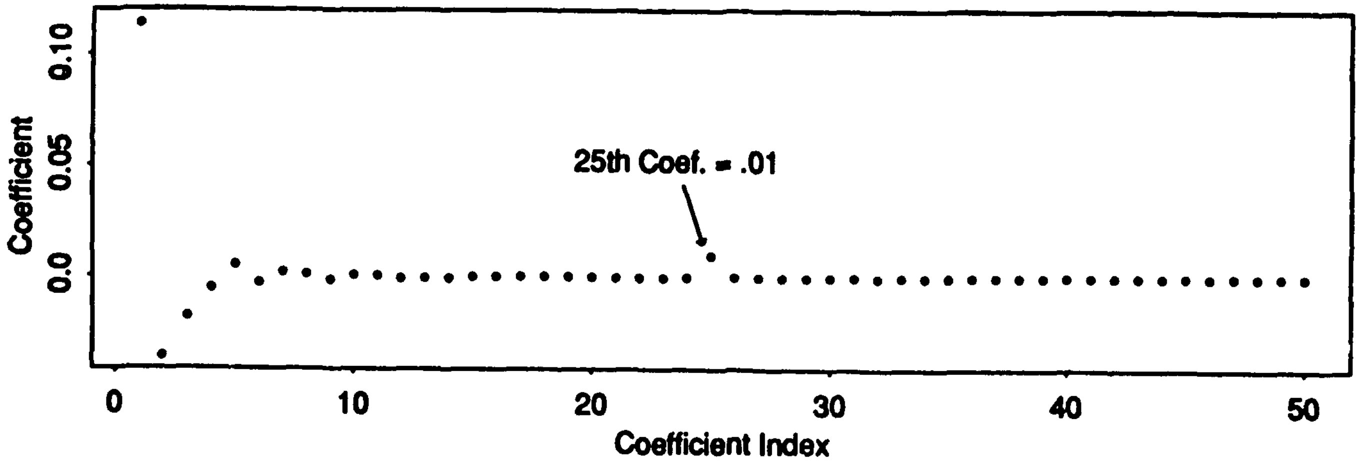


Figure 3.5b: Perturbed Right Hand Side

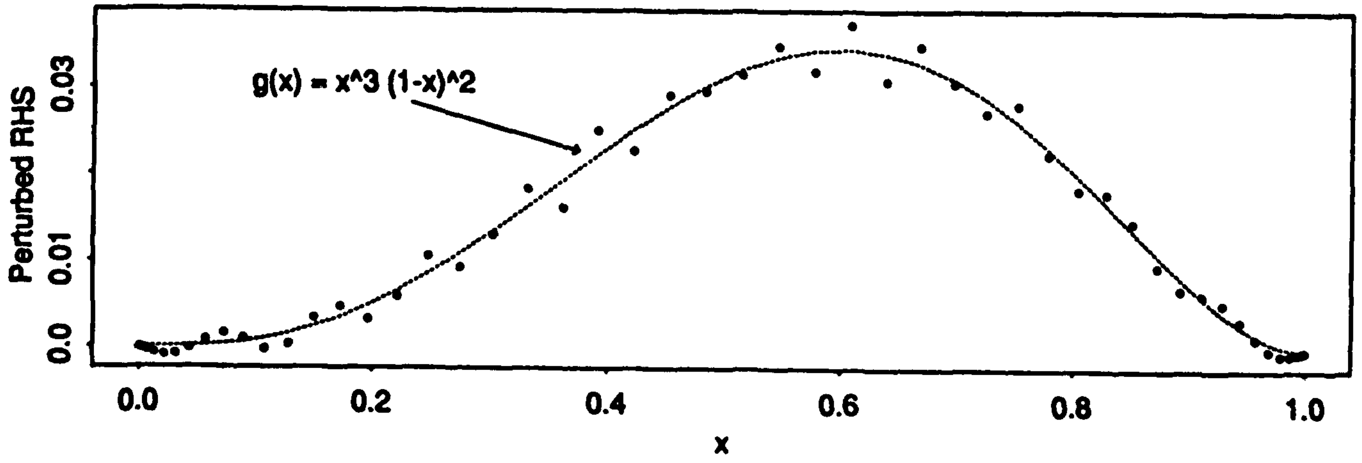


Figure 3.6a: Fifty Iterations of Conditional Expectation Algorithm
(Perturbed RHS)

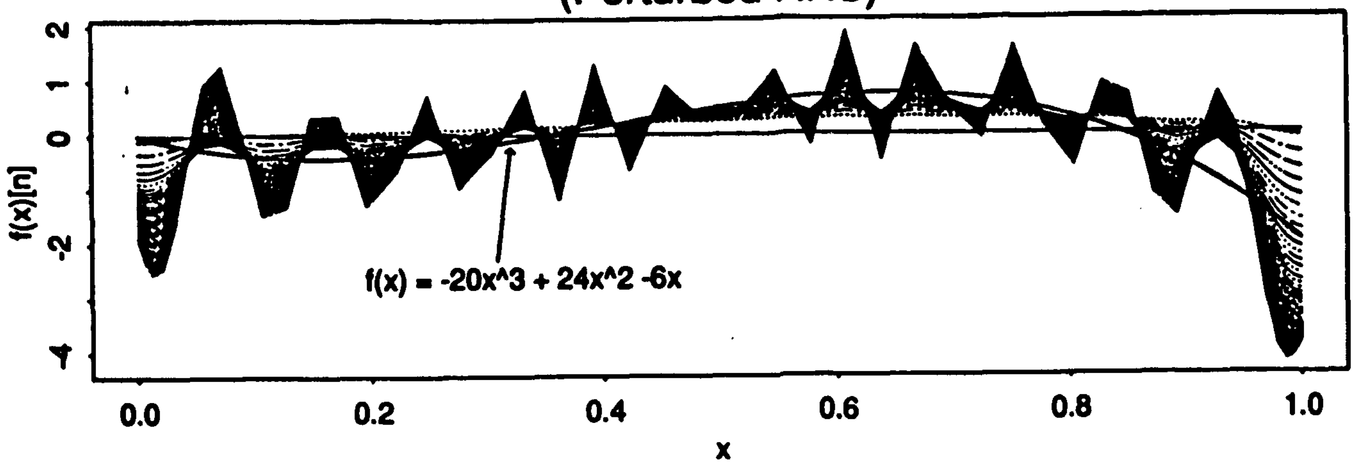


Figure 3.6b: Solution by Matrix Inversion
(Perturbed RHS)

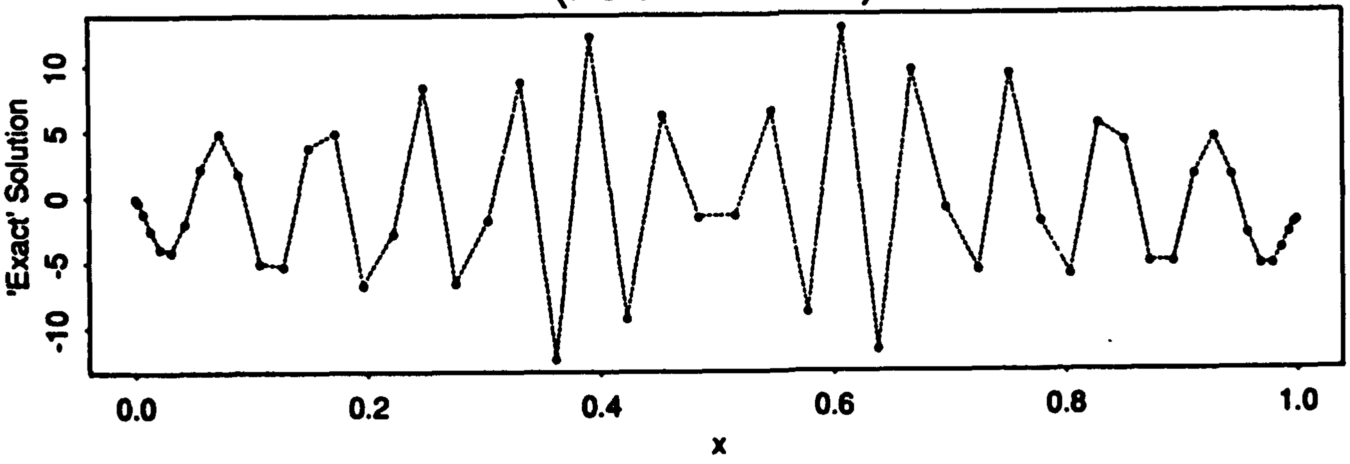


Figure 3.7a: Distances to Solutions for Perturbed and Unperturbed RHS

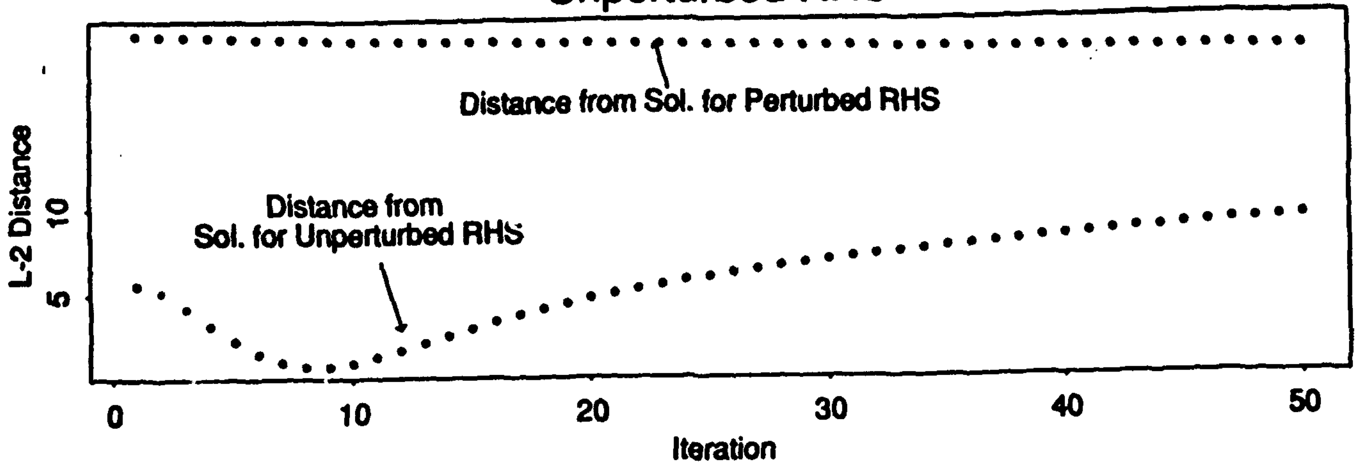
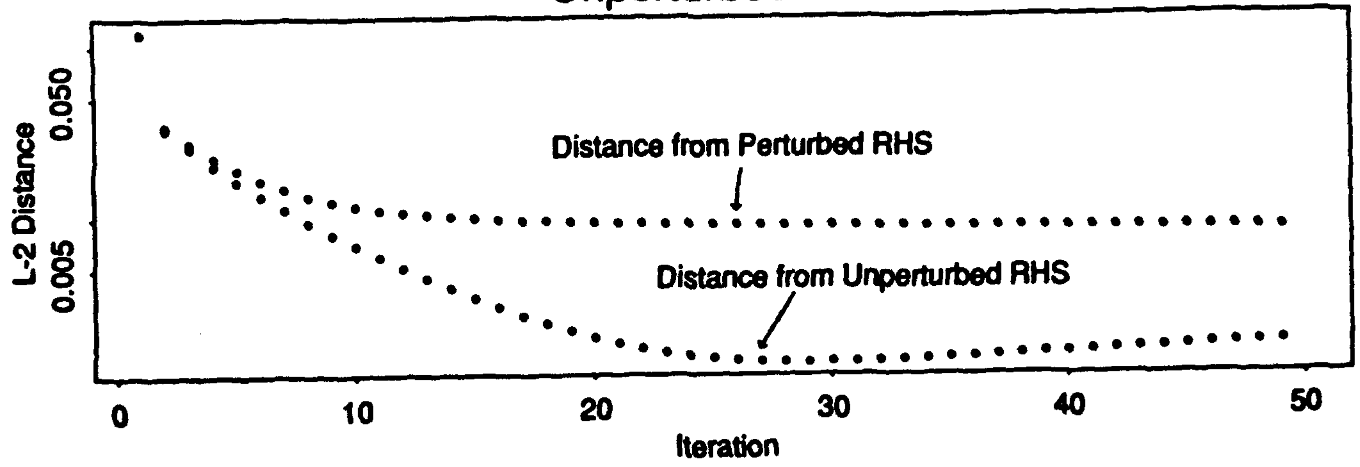


Figure 3.7b: Residual Norms for Perturbed and Unperturbed RHS



Chapter 4

The Conditional Expectation Algorithm for Nonlinear Integral Equations with Peaked Kernels

4.1 A Nonlinear Equation

All of the integral equations which we will consider can be expressed as integral equations of the first kind of the form

$$\int_0^1 k\{x, y, f[\phi(x,y)]\}dy = g(x), \tag{4.1}$$

where the kernel k and the function $\phi : [0,1] \times [0,1] \rightarrow [0,1]$ are known, and k is nonnegative and bounded. To fix ideas, we will restrict attention, for the most part, to kernels defined on the unit square. However, this restriction is not essential.

We will often need to refer to the kernel of (4.1) and to the derivative with respect to its third argument, so we introduce the following notation to save writing:

$$k(x,y,z)|_{z=f[\phi(x,y)]} \equiv k(x,y,f) \tag{4.2}$$

and

$$\left.\frac{\partial k(x,y,z)}{\partial z}\right|_{z=f[\phi(x,y)]} \equiv k'(x,y,f). \tag{4.3}$$

In addition to requiring that k be nonnegative and bounded, we also require k' to be nonnegative and bounded. The reason for this is that we will introduce an iterative algorithm, based on applying the Conditional Expectation algorithm of Section 3.2.3 to linearizations of (4.1), which can be motivated by considering k' to be proportional to a bivariate probability density.

4.1.1 Peaked Kernels

We will restrict attention to integral equations with kernels having certain features, which we summarize here. A kernel k will be said to be peaked if k is nonnegative and bounded, and k' has the following properties

1. $k'(x,y,f) \geq 0$ for all x and y,

2. $\int_0^1 k'(x,y,f)dy < \infty$ for all x, and

3. There exists a monotone function $t : [0,1] \rightarrow [0,1]$, such that, for all x,

$$\max_{y\in[0,1]} k'(x,y,f) = k'(x,t(x),z)\big|_{z=f[\phi(x,t(x))]}.$$

A nonlinear kernel can be peaked for some values of f and not for others. When we refer to a kernel as being peaked, we intend, somewhat imprecisely, for this to mean that this kernel is peaked for functions f of interest.

A simple (and important) example of a peaked kernel is the kernel,

$$k(x,y,f) \equiv w(x,y)f(y), \tag{4.4}$$

of a *linear* Fredholm equation, where w is nonnegative, bounded, and peaked along the line $x = y$. The kernel (4.4) is linear in f and, as discussed in Section 3.2.3, its derivative

$$k'(x,y,f) = w(x,y) \tag{4.5}$$

is proportional to the joint density of two random variables X and Y. To the extent that w is peaked along $x = y$, we can say that $Y \approx X$. The definition of a peaked kernel attempts to generalize this idea to kernels with derivatives with respect to f which can be regarded as bivariate probability densities for which $Y \approx t(X)$, for some monotone function $y = t(x)$ in the unit square. We will consider a simple example for which $t(x)$ is not the identity in Section 4.3.2.

Note the slight difference in terminology between Chapter 3 and Chapter 4. If we are restricting attention to linear equations, it is natural to refer to $w(x,y)$ as the kernel, since the relationship between $w(x,y)$ and $k(x,y,f)$ is the same for any linear problem. However, when we regard a linear equation as merely a special case in a class of *nonlinear* equations, then we will refer to $k(x,y,f)$ as the kernel, where $k'(x,y,f) = w(x,y)$.

4.2 Newton's Method

When solving a system of nonlinear equations the method of choice is often Newton's method. Newton's method is known to converge for a 'good enough' starting value and to converge quadratically in most cases. The sufficient conditions for convergence and the rate of convergence of Newton's method are provided by the Newton-Kantorovich theorem, (e.g., Ortega, 1972) and this theorem is proved in Banach space. In particular, Newton's method is a useful algorithm for nonlinear integral equations in L_2.

4.2.1 The Fréchet Derivative

In order to extend the definition of Newton's method to functional (in particular, integral) equations, a concept of functional derivative is necessary. We introduce here one such derivative, the Fréchet derivative. We give here the definition, in Banach space, following Debnath and Mikusiński 1990, p.416).

Definition 4.2.1 (Fréchet Derivative) *Let B_1 and B_2 be Banach spaces, and let $x \in B_1$ be fixed. A continuous linear operator $A : B_1 \rightarrow B_2$ is called the* **Fréchet derivative** *of an operator $T : B_1 \rightarrow B_2$ at x if*

$$T(x+h) - T(x) = Ah + \Phi(x,h),$$

and

$$\lim_{\|h\|\to 0} \frac{\|\Phi(x,h)\|}{\|h\|} = 0.$$

The Fréchet derivative at x of T will be denoted $T'(x)$.

It can be easily shown that if the Fréchet derivative exists, then it is unique (Debnath and Mikusiński (1990, p.417).

Define $T : L_2 \to L_2$ by

$$T(\tilde{f}) \equiv \int_0^1 k(x,y,\tilde{f})dy - g. \tag{4.6}$$

The Fréchet derivative of T is

$$T'(\tilde{f})h = \int_0^1 k'(x,y,\tilde{f})h(y)dy. \tag{4.7}$$

Now that we've extended the notion of derivative to integral operators of the form (4.1), we can state what Newton's method is for this equation.

4.2.2 The Newton-Step Equation

Let T be an operator between Hilbert spaces, and let f^n be a point in the domain at which T is Fréchet differentiable. We would like to approximately determine a function f such that $T(f) = 0$ (where here '0' denotes the *function* which is identically zero), and we assume that f^n is 'near' f. Expand T about f^n to first order in a Taylor series, giving

$$T(f) - T(f^n) \approx T'(f^n)(f - f^n). \tag{4.8}$$

But $T(f) = 0$, so we have the approximate linear equation

$$T(f^n) \approx -T'(f^n)(f - f^n) \tag{4.9}$$

relating f^n to f. Let f^{n+1} be a value of f which makes (4.9) an equality. Solving the *linear* operator equation (4.9) for f^{n+1} constitutes one step of *Newton's method.*

For (4.1), let

$$\tilde{h}^n \equiv f^{n+1} - f^n. \tag{4.10}$$

The function $\tilde{h}^n$ is a solution to the following *Newton-step equation*:

$$g - \int_0^1 k(x,y,f^n)dy = \int_0^1 k'(x,y,f^n)\tilde{h}^n dy. \tag{4.11}$$

Solving the linear integral equation (4.11) for $\tilde{h}^n$ is in general difficult. We will instead investigate a *quasi-Newton* iterative algorithm, in which we use one or several steps of the Conditional Expectation algorithm of Section 3.2 as an easily determined *approximate* Newton-step. By doing this, we replace a quadratically convergent algorithm with a linearly convergent algorithm, but since the steps of the quasi-Newton algorithm are, by design, very easy to calculate, we are often better off using the linearly convergent method.

There is another reason to want to use an approximate Newton-step. Recall from the discussion of Section 2.4.2, that the *exact* solution of any numerical representation of an ill-posed integral equation of the first kind is likely to be either nonexistent, or else quite different from *any* solution to the integral equation. For an ill-posed nonlinear problem, Newton's method involves the exact solution of an ill-posed linear equation *at each step.*

4.3 The Conditional Expectation Algorithm

In Section 3.2, we motivated a specific preconditioned Richardson algorithm for a linear integral equation of the first kind. We now propose extending this algorithm in order to iteratively approximate solutions of nonlinear integral equations. Because of the probabilistic motivation of Section 3.2.3, we will refer to this algorithm, whether applied to linear or to nonlinear equations, as the *Conditional Expectation algorithm.* To illustrate this algorithm, we first consider the special case of (4.1) where $\phi(x,y) = y$, k is peaked, and $t(x) \approx x$. Following this, we suggest how the method can be extended to some more general problems.

4.3.1 A Simple Case

Let

$$\int_0^1 k[x,y,f(y)]dy = g(x), \tag{4.12}$$

where k is a peaked kernel with $t(x) \approx x$. We propose attempting to solve (4.12) using a nested iteration, in which the outer iteration is an approximate Newton method, with the approximate Newton step provided by the inner iteration. Since the Newton step equation is linear, we can use Richardson's algorithm with stochastic preconditioning (Section 3.2) in order to approximately determine the Newton steps. We call this nested algorithm the Conditional Expectation algorithm, and we note that it reduces to the algorithm (of the same name) discussed in Section 3.2 when (4.12) is linear in f.

Let the outer iteration be indexed by n, and let the inner iteration be indexed by s, for $s = 1,\ldots,l$. Actually, the inner iteration limit can depend on n, but we will not state the algorithm in this much generality in order to keep the notation as simple as possible. Write the Newton-step equation, at the nth outer iteration, as

$$r^n \equiv g - \int_0^1 k(x,y,f^n)dy = \int_0^1 k'(x,y,f^n)\tilde{h}^n(y)dy. \tag{4.13}$$

Approximate $\tilde{h}^n$ by $h^{n,l}$ where, since $t(x) = x$,

$$h^{n,s+1}(y) = h^{n,s}(y) + \delta^s(y) \approx h^{n,s}(y) + \delta^s(x), \tag{4.14}$$

$h^{n,0}$ is arbitrary, and

$$\delta^s(x) \equiv \frac{r^n(x) - \int_0^1 k'(x,y,f^n)h^{n,s}(y)dy}{\int_0^1 k'(x,y,f^n)dy}. \tag{4.15}$$

A simple case, which is often useful, is to let $l = 1$, and to take $h^{n,0} = 0$ for all n. The Conditional Expectation iteration is then

$$f^{n+1} = f^n + \frac{r^n}{\int_0^1 k'(x,y,f^n)dy}, \tag{4.16}$$

since

$$h^{n,l} = h^{n,1} = h^{n,0} + \delta^0 = r^n. \tag{4.17}$$

Another possibility, only slightly more complicated, is to let $l = 1$, $h^{0,0} = 0$, and

$$h^{n,0} = h^{n-1,l}; \tag{4.18}$$

the reasoning behind this being that if $\tilde{h}^{n+1} \approx \tilde{h}^n$, the $(n-1)$st approximate Newton step might provide a better initial approximation to the nth step than the zero vector. This choice of an inner iteration leads to

$$f^{n+1} = f^n + \frac{r^n - \int_0^1 k'(x,y,f^n)[f^n - f^{n-1}]dy}{\int_0^1 k'(x,y,f^n)dy}. \tag{4.19}$$

The iteration (4.19) makes clear the relationship of the nonlinear algorithm of this chapter to the linear Conditional Expectation algorithm of Chapter 3.

If the integral equation of interest is ill-posed, then Newton's method will almost certainly either diverge, or else converge to a solution of the *discretized* equation that is far from any solution to the original equation. Because of this, it is probably a good idea to keep l small; one exception being when the derivative of the kernel is expensive to compute. Newton's method has been used to motivate the iterative algorithm of this chapter, which is an iteration in its own right. Hence, one should not regard the closeness with which one can approximate Newton steps as an overriding consideration in using the Conditional Expectation algorithm.

4.3.2 An Example for Which $t(x) \neq x$ and $\phi(x,y) \neq y$

Consider the integral equation

$$\int_0^1 k_\alpha(x,y)f(xy)dy = g(x), \tag{4.20}$$

where

$$k_\alpha(x,y) \equiv \begin{cases} y(1-x^\alpha) & \text{if } y < x^\alpha \\ x^\alpha(1-y) & \text{if } y \geq x^\alpha, \end{cases}, \tag{4.21}$$

and $\alpha \neq 1$ is a parameter. For this example $\phi(x,y) = xy$ and $t(x) = x^\alpha$. The Newton-step equation is

$$g(x) - \int_0^1 k_\alpha(x,y)f^n(xy)dy = \int_0^1 k_\alpha(x,y)\tilde{h}^n(xy)dy. \tag{4.22}$$

We approximate the unknown $\tilde{h}^n(xy)$ by a function which is constant in y by replacing y with

$$y_* = t(x) = x^\alpha, \tag{4.23}$$

which is the location of the peak in y for each x. This leads to the Conditional Expectation algorithm step

$$h^n(x^{\alpha+1}) \equiv \frac{g(x) - \int_0^1 k_\alpha(x,y)f^n(xy)dy}{\int_0^1 k_\alpha(x,y)dy}, \tag{4.24}$$

or alternatively,

$$h^n(x) \equiv \frac{g(x^{1/(\alpha+1)}) - \int_0^1 k_{\alpha/(\alpha+1)}(x,y)f^n(x^{1/(\alpha+1)}y)dy}{\int_0^1 k_{\alpha/(\alpha+1)}(x,y)dy}. \tag{4.25}$$

4.3.3 The General Case

The example of the previous subsection motivates the following generalization of the Conditional Expectation algorithm. For the general case, we must approximately solve (4.11) at each iteration, and we rewrite this equation as

$$r^n(x) = \int_0^1 k'(x,y,f^n[\phi(x,y)])\tilde{h}^n[\phi(x,y)]dy. \tag{4.26}$$

Assume that $k'(x,y,f^n)$ has a peak in y, for given x, with location $y_* = t(x)$. A reasonable approximation to $\tilde{h}^n[\phi(x,y)]$ might be $h^n[\phi(x,u(x))]$, where $u(x) \approx t(x)$, and

$$h^n[\phi(x,u(x))] \equiv h^n(z) = \frac{r^n(x)}{\int_0^1 k'\{x,y,f^n[\phi(x,y)])\}dy}. \tag{4.27}$$

In order to provide a useful approximation, it seems reasonable to require that $u(x)$ have the following two properties:

1. $k'\{x,u(x),f^n[\phi(x,u(x))]\}$ is 'approximately' equal to $\max_{y\in[0,1]} k'\{x,y,f^n[\phi(x,y)]\}$ for all x.
2. $\phi[x,u(x)] : [0,1] \rightarrow [0,1]$ is monotone increasing, with $\phi(0,u(0)) = 0$ and $\phi(1,u(1)) = 1$.

If $t(x) = x$, then $u(x) = x$ exactly satisfies both of the above conditions. In general, considerable experimentation may be required in order to determine a useful function u. In Chapter 6, we discuss, in some detail, an example for which

$$\phi(x,y) \propto \left(\frac{1}{1-x}\right)\left(\frac{y}{1-y}\right) \tag{4.28}$$

and $t(x)$ is approximately a constant function over most of the range of x.

Of the two conditions that we have imposed on u, the requirement that ϕ be a monotone increasing function mapping the unit interval into itself is important; the other condition is merely heuristic. There is no guarantee that the 'best' choice of u exactly maximizes $k'(x,u(x),f^n)$. A practical approach might be to first try simple choices of u, however crude, and see what happens.

4.4 A Simple Numerical Nonlinear Example

We conclude this chapter by illustrating the Conditional Expectation algorithm applied to a nonlinear problem. Let v be a given differentiable function with a positive derivative, and consider the following integral equation:

$$\int_0^1 w(x,y)v[f(y)]dy = g(x), \tag{4.29}$$

for $w(x,y)$ given by the Green's function

$$w(x,y) = \begin{cases} y(1-x) & \text{if } y < x \\ x(1-y) & \text{if } y \geq x, \end{cases} \tag{4.30}$$

which we used in the examples of Section 3.4.1. The equation (4.29), though linear in v, is nonlinear in the unknown function f, and of the form (4.1). The derivative of the kernel at f is

$$k'(x,y,f) = w(x,y)v'(f). \tag{4.31}$$

The peak of $w(x,y)$ is at $y = x$, thus it is not unreasonable to assume that the peak of (4.31) in y for fixed x is near the line $t(x) = x$. For the Conditional Expectation method, we choose $l = 1$, and $h^{n,0} = 0$ for all n, so that the Conditional Expectation step is

$$h^{n,1}(x) = \frac{g(x) - \int_0^1 w(x,y)v[f^n(y)]dy}{\int_0^1 w(x,y)v'[f^n(y)]dy,} \tag{4.32}$$

and

$$f^{n+1}(y) = f^n(y) + h^{n,1}(y). \tag{4.33}$$

For a numerical example, we choose

$$v(x) = e^x, \tag{4.34}$$

and

$$f(x) = 2x - 1, \tag{4.35}$$

so that

$$g(x) = \frac{1 + x(e^2 - 1) - e^{2x}}{4e}. \tag{4.36}$$

Note that there is a nontrivial distinction between the nonlinear iteration with $v(f) = e^f$, and the linear case with $v(f) = f$. For this example, with $v(f) = e^f$, the solution f and *all approximate solutions* f^n must be everywhere positive.

We take $f^0 \equiv 1$ as a starting value, and discretize as in Appendix A, using 50-point Gauss-Legendre quadrature. The first 50 approximations to the solution are displayed in Figure 4.1. Initially, the algorithm converges rapidly, although eventually the convergence rate becomes very slow. From the plot, in Figure 4.2, of the L_2 distance from the solution as a function of the iteration index, it is clear that the rate of convergence begins to decrease substantially after only a few iterations. However, convergence is sufficiently rapid initially that approximations are near the solution before the iteration becomes slowly convergent.

Fgure 4.1: Fifty Iterations of the Conditional Expectation Algorithm for a Nonlinear Problem

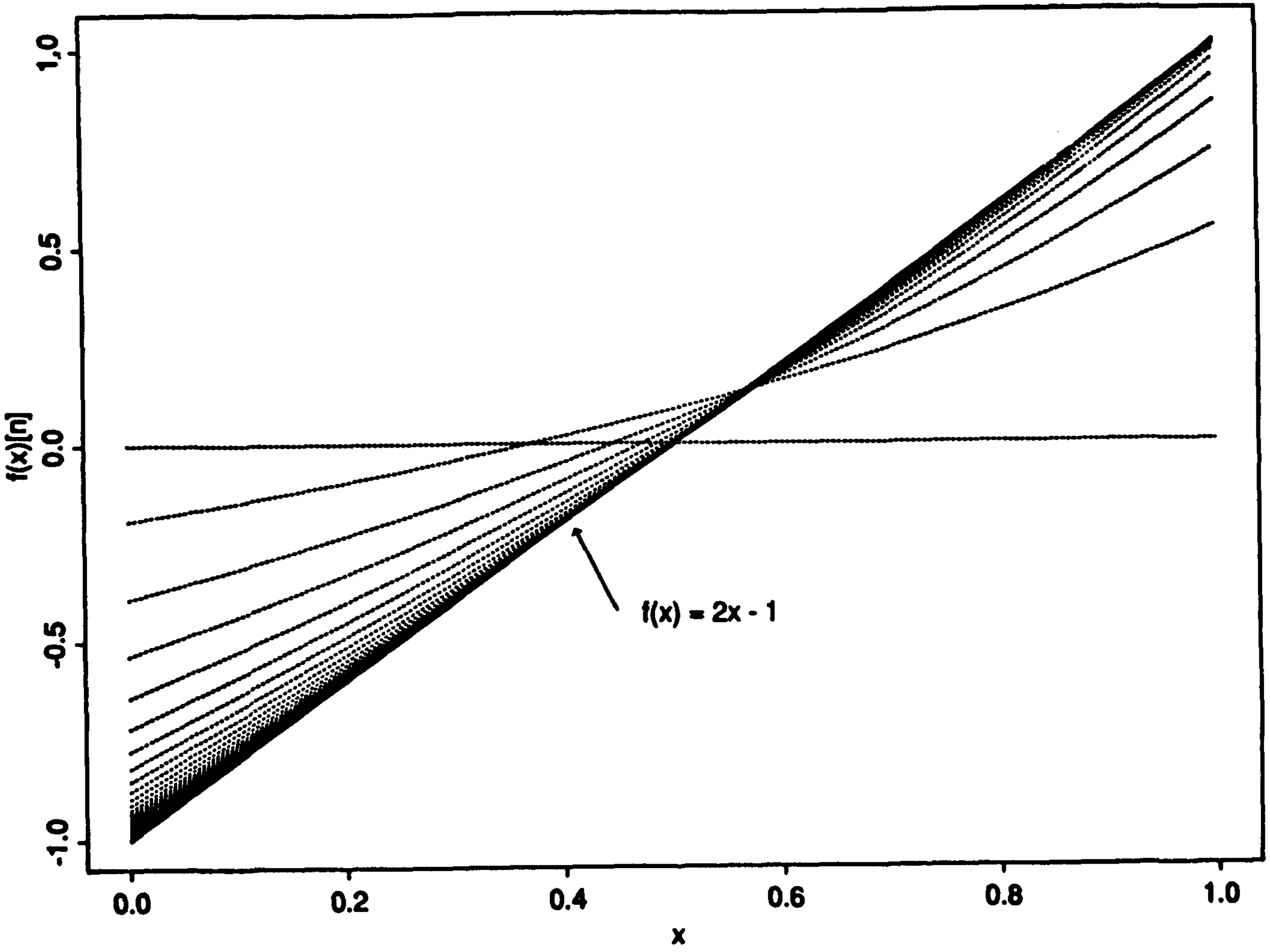
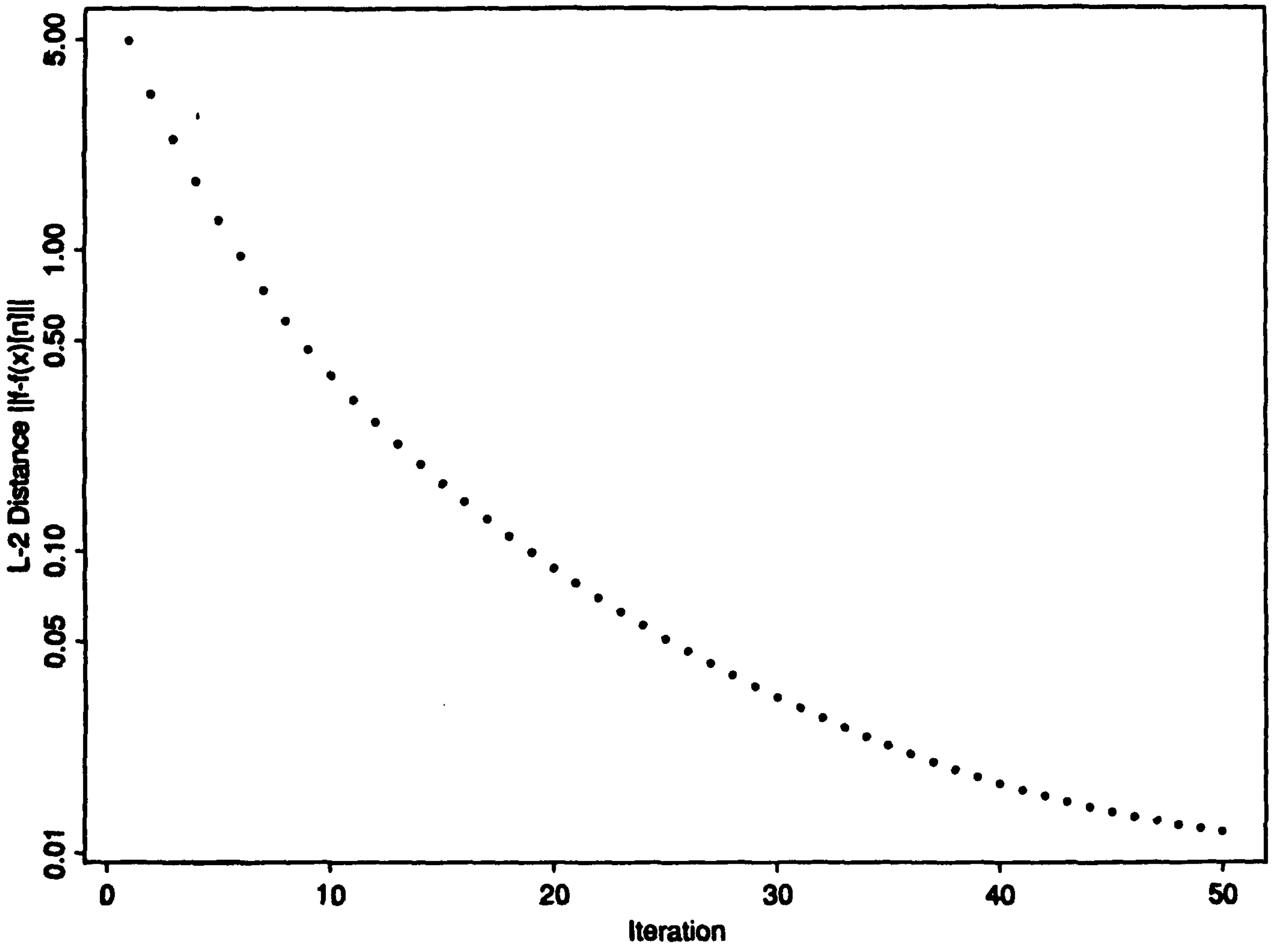


Figure 4.2: Convergence Rate of the Conditional Expectation Algorithm for a Nonlinear Problem



Chapter 5

The Behrens-Fisher Problem

5.1 Historical Background

The Behrens-Fisher problem is the problem of comparing the means of two normal populations with no assumptions about the variances. This problem has received much attention, and caused much controversy, because it is the simplest example of any practical importance where the fiducial (and noninformative-prior Bayesian) and Neyman-Pearson approaches arrive at substantially different answers.

Let $X_i, i = 1, \ldots, n_1$ be a random sample from a $N(\mu_1, \sigma_1^2)$ population, and let $Y_i, i = 1, \ldots, n_2$ be a random sample from a $N(\mu_2, \sigma_2^2)$ population, where μ_1, μ_2, σ_1^2, and σ_2^2 are unknown. Let $\bar{X}_j$ and S_j^2, for $j = 1, 2$, be the usual sample estimates of the means and variances (defined in Section 2.6). We wish to test the composite hypothesis

$$H_0 : \mu_1 = \mu_2 \tag{5.1}$$

against the alternative

$$H_1 : \mu_1 > \mu_2, \tag{5.2}$$

where σ_1^2/σ_2^2 or, equivalently,

$$\theta = \frac{\sigma_1^2/n_1}{\sigma_1^2/n_1 + \sigma_2^2/n_2} \tag{5.3}$$

is a nuisance parameter of special importance.

For both the fiducial and the frequentist approaches the test (of size α) is of the form 'Reject H_0 if U exceeds a critical value $c_\alpha(S_1^2, S_2^2)$', where

$$U = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}; \tag{5.4}$$

however the critical values for the tests differ for the two approaches.

A fiducial argument suggests determining critical values for the test from the inverse of the distribution of a certain linear combination of two Student t random variables. The distribution of this linear combination is known as the Behrens-Fisher distribution. A Bayesian analysis with a noninformative prior leads to the same result. The Behrens-Fisher distribution can be evaluated numerically to provide a test which has size α, from the standpoint of fiducial probability, for all θ; but this test is not accepted in the Neyman-Pearson framework (e.g., Wallace, 1980).

In the Neyman-Pearson framework, a test of

$$H_0 : \omega \in \Omega_0$$

which achieves a nominal size α for all $\omega \in \Omega_0$ is said to be *similar*. A similar test for the Behrens-Fisher problem must achieve a fixed size α for all values of the nuisance parameter θ, and such a test does not exist. To be precise, a critical value statistic which results in a similar test does not exist if n_1 and n_2 are of the same parity, and any critical value statistic for sample sizes of opposite parity must be a function with infinitely many discontinuities. This result was proved by Linnik and others in the 1960s (see Pfanzagl, 1974), and it was suspected to be true by many for years before. However, by the time the Linnik results became available, much progress had been made toward a practical solution from a frequentist perspective (e.g., Kendall and Stuart, 1977, Vol. 2, Chapter 21).

5.2 The Trickett-Welch Approach

Welch (1947) and Aspin (1948) tacitly assume the existence of a continuous critical value statistic $v_\alpha(R)$ such that

$$P(U \le v_\alpha(R)) = 1 - \alpha \tag{5.5}$$

for all θ, where

$$R = \frac{S_1^2/n_1}{S_1^2/n_1 + S_2^2/n_2} \tag{5.6}$$

is a sample estimate of θ, and the probability is determined assuming that the null hypothesis is true. They proceed to calculate an asymptotic series for v_α including terms of $O(1/(n_i - 1)^4)$. This series provides a test which is *very* nearly similar for all but very small sample sizes. Of course, in the light of Linnik's results, it should not be surprising that no bound was given by Welch and Aspin for the distance between their series approximation and a solution to (5.5).

For n_1 and n_2 less then about seven, this asymptotic series is not adequate, and so Trickett and Welch (1954) consider an alternative numerical approach. Equation (5.5) can be written as an integral equation, and Trickett and Welch do so by conditioning on the variance estimates and averaging over their distributions. We will derive this integral equation next, using a somewhat simpler approach.

5.2.1 The Trickett-Welch Equation

The sample means can be written as

$$\bar{X}_j = \mu_j + Z_j\sigma_j/\sqrt{n_j}, \tag{5.7}$$

where the Z_j are *iid* N(0,1), and $j = 1,2$. So,

$$\bar{X}_1 - \bar{X}_2 = (\mu_1 - \mu_2) + Z_3\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}, \tag{5.8}$$

for $Z_3 \sim$ N(0,1), and

$$Z \equiv \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim \mathrm{N}(\delta, 1), \tag{5.9}$$

with

$$\delta \equiv \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \tag{5.10}$$

would be a natural test statistic to use if we knew the variances σ_1^2 and σ_2^2. Not knowing these, we use the estimate

$$U \equiv \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}, \tag{5.11}$$

where

$$V_1 = \frac{\nu_1 S_1^2}{\sigma_1^2} \sim \chi^2_{\nu_1} \tag{5.12}$$

and

$$V_2 = \frac{\nu_2 S_2^2}{\sigma_2^2} \sim \chi^2_{\nu_2}, \tag{5.13}$$

with $\nu_j \equiv n_j - 1$, are independent of each other and of Z. Let

$$W \equiv \frac{\nu_1 S_1^2}{\sigma_1^2} + \frac{\nu_2 S_2^2}{\sigma_2^2} = V_1 + V_2, \tag{5.14}$$

and

$$Y \equiv \frac{\nu_1 S_1^2/\sigma_1^2}{W}, \tag{5.15}$$

and note that

$$\begin{aligned} U &= \frac{(\bar{X}_1 - \bar{X}_2)/\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}{\sqrt{\sigma_1^2 V_1/(n_1\nu_1) + \sigma_2^2 V_2/(n_2\nu_2)}/\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \\ &= Z\left(\frac{\theta V_1}{\nu_1} + \frac{(1-\theta)V_2}{\nu_2}\right)^{-1/2} \\ &= Z\left\{W\left[\theta Y/\nu_1 + (1-\theta)(1-Y)/\nu_2\right]\right\}^{-1/2}, \\ &= \frac{Z}{\sqrt{W/(\nu_1+\nu_2)}}\left\{(\nu_1+\nu_2)\left[\frac{Y\theta}{\nu_1} + \frac{(1-Y)(1-\theta)}{\nu_2}\right]\right\}^{-1/2}. \end{aligned} \tag{5.16}$$

where Z, V_1, and V_2 are independent random variables with distributions independent of the parameters.

It is well known, and easy to show, that

$$W \sim \chi^2_{\nu_1+\nu_2} \tag{5.17}$$

is independent of

$$Y \sim \text{Beta}(\nu_1/2, \nu_2/2). \tag{5.18}$$

We can see from (5.9) and (5.16) that under H_0 the distribution of U depends on the parameters μ_1, μ_2, σ_1, and σ_2 only through the nuisance parameter θ, so it is natural to use a test of the form: '*Reject* H_0 *if* $U \geq v_\alpha(R)$', where

$$R \equiv \frac{S_1^2/n_1}{S_1^2/n_1 + S_2^2/n_2} = \frac{Y\theta/\nu_1}{Y\theta/\nu_1 + (1-Y)(1-\theta)/\nu_2} \equiv \phi(\theta, Y) \tag{5.19}$$

is an estimator of θ. If H_0 is true, then

$$T = \frac{Z}{\sqrt{W/(\nu_1+\nu_2)}} \sim T_{\nu_1+\nu_2}; \tag{5.20}$$

that is, T has a t distribution with $\nu_1+\nu_2$ degrees of freedom. Hence the events

$$\{U \le v_\alpha(R)\} \tag{5.21}$$

and

$$\left\{T \le v_\alpha(R)\left\{(\nu_1+\nu_2)\left[\frac{Y\theta}{\nu_1}+\frac{(1-Y)(1-\theta)}{\nu_2}\right]\right\}^{1/2}\right\} \tag{5.22}$$

are equivalent.

Therefore we want to find $v_\alpha(R)$ so that, for all θ,

$$P_\theta[U \le v_\alpha(R)] = E\left\{T_{\nu_1+\nu_2}\left[v_\alpha(R)\left\{(\nu_1+\nu_2)\left[\frac{Y\theta}{\nu_1}+\frac{(1-Y)(1-\theta)}{\nu_2}\right]\right\}^{1/2}\right]\right\} = 1-\alpha, \tag{5.23}$$

where $E\{\cdot\}$ denotes the expectation, under the null hypothesis, of an expression depending only on one random variable Y, which has a Beta distribution, and where $T_{\nu_1+\nu_2}$ is the cumulative of the t distribution with $\nu_1+\nu_2$ degrees of freedom. Thus our expectation can be represented as an integral of the form

$$\int_0^1 k\{\theta, y, v_\alpha[\phi(\theta,y)]\}dy. \tag{5.24}$$

We write the Trickett-Welch equation in full as

$$\int_0^1 k\{\theta, y, v_\alpha[\phi(\theta,y)]\}dy = g(\theta) \equiv 1-\alpha, \tag{5.25}$$

where

$$\phi(\theta,y) = \frac{y\theta/\nu_1}{y\theta/\nu_1+(1-y)(1-\theta)/\nu_2}, \tag{5.26}$$

$$\begin{aligned} k(\theta,y,v_\alpha) &= \text{Beta}\,(y;\nu_1/2,\nu_2/2) \qquad (5.27)\\ &\quad \cdot\mathrm{T}_{\nu_1+\nu_2}\left\{v_\alpha[\phi(\theta,y)]\sqrt{(\nu_1+\nu_2)\left[\frac{y\theta}{\nu_1}+\frac{(1-y)(1-\theta)}{\nu_2}\right]}\right\},\\ \text{Beta}\,(y;\nu_1/2,\nu_2/2) &\equiv \frac{\Gamma\left[(\nu_1+\nu_2)/2\right]}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)}y^{\nu_1/2-1}(1-y)^{\nu_2/2-1}, \end{aligned}$$

$$(5.28)$$

and

$$T_\nu(t) \equiv \int_{-\infty}^{t} \frac{\Gamma[(\nu+1)/2]}{\Gamma(\nu/2)\sqrt{\nu\pi}}(1+x^2/\nu)^{-(\nu+1)/2}dx. \tag{5.29}$$

Note that $v_\alpha(r)$ is the same function of the deterministic argument r that $v_\alpha(R)$ is of the random variable R.

We will sometimes use functional notation and write (5.23) as

$$F(v_\alpha) = 1-\alpha. \tag{5.30}$$

This is equivalent to the integral equation (5.25), which Trickett and Welch solve numerically.

5.3 Quasi-Newton Methods and the Trickett-Welch Algorithm

Trickett and Welch approximate a solution to (5.23) by using a quasi-Newton iterative algorithm. In Chapter 4, we introduced iterative algorithms for nonlinear integral equations in general. In this section, we discuss, in the context of the Behrens-Fisher problem, the quasi-Newton algorithm which Trickett and Welch used, as well as a Conditional Expectation algorithm.

5.3.1 Newton's Method

We will begin examining the application of iterative algorithms to (5.23) by considering what Newton's method is for this problem. Assume that v_α solves (5.23), and expand to first order about an approximate solution v_α^0. Using functional notation, we have

$$F(v_\alpha) = 1 - \alpha \approx F(v_\alpha^0) + F'(v_\alpha^0)(v_\alpha - v_\alpha^0), \tag{5.31}$$

where $F'(v_\alpha^0)$ is the Fréchet derivative of F evaluated at v_α^0.

If we regard (5.31) as an equality, and solve

$$F'(v_\alpha^0)\tilde{h} \equiv F'(v_\alpha^0)(\tilde{v}_\alpha - v_\alpha^0) = 1 - \alpha - F(v_\alpha^0) \tag{5.32}$$

for $\tilde{h}$, then we will be able to take a Newton step. Equation (5.32) is equivalent to

$$\begin{aligned} &E\left\{\tilde{h}(R)\sqrt{(\nu_1+\nu_2)\left(\frac{\theta Y}{\nu_1} + \frac{(1-\theta)(1-Y)}{\nu_2}\right)}\right. \\ &\left.\cdot\, T'_{\nu_1+\nu_2}\left[v_\alpha^0(R)\sqrt{(\nu_1+\nu_2)\left(\frac{\theta Y}{\nu_1} + \frac{(1-\theta)(1-Y)}{\nu_2}\right)}\right]\right\} = \\ &1 - \alpha - F(v_\alpha^0), \end{aligned} \tag{5.33}$$

where T'_ν denotes the t density with ν degrees of freedom.

Since Newton's method is quadratically convergent (when it does converge) in Banach space, for v_α^0 'close enough' to a solution, one might think that Newton's method would be a good choice for this problem. However, as discussed in Chapter 4, we must keep in mind that, although (5.33) is a linear integral equation, it is ill-posed and difficult to solve. Also, the Behrens-Fisher problem has either none or else only pathological solutions, so we have no reason to expect that Newton's method will work well, even when applied to a discretized problem. It turns out that more conservative, linearly convergent iterative algorithms perform quite well for this problem.

5.3.2 Quasi-Newton Procedures

We will suggest two simple algorithms based on approximating the Newton-step equation (5.33). The first approximation is used by Trickett and Welch and is adequate for the Behrens-Fisher problem. The second approximation is a form of the Conditional Expectation algorithm of Chapters 3 and 4. There are heuristic reasons (Section 3.2) to suspect the Conditional Expectation algorithm to be an improvement over the original Trickett-Welch procedure, however, for the Behrens-Fisher problem, there is virtually no difference between results obtained using the two procedures. We have used the Conditional Expectation algorithm in the calculations below, but the simpler Trickett-Welch algorithm results in the same iterates to several significant figures.

The Trickett-Welch Algorithm

Figure 5.1 show a contour plot of a typical kernel for the Newton step equation (5.33). To be specific, we have taken $n_1 = 20$, $n_2 = 10$, v_α equal to the constant 1.65, and $\alpha = .05$. We refer to these contours as typical since the shape of the kernel does not depend strongly on v_α. The effect of n_1 and n_2 on the kernel is primarily limited to the location and sharpness of the peak – the contours remain nearly straight vertical lines over a wide range of n_1 and n_2. Also, for most applications, it is sufficient to consider α in the range $.01 \le \alpha \le .10$, and, over this range, the shape of the kernel remains qualitatively similar.

For the example with $n_1 = 20$ and $n_2 = 10$, the variance of Y is small, so the kernel is sharply peaked in y (and the location of this peak is almost independent of θ). The mean of Y, $\nu_1/(\nu_1+\nu_2)$, is near the mode of the density of Y and is indicated by the broken line. The effect on the kernel of changes in v_α of the magnitude which occur in practice does not significantly effect the conclusion that the kernel is generally sharply peaked near $y = \nu_1/(\nu_1+\nu_2)$. Trickett and Welch note this fact, and use it to motivate a quasi-Newton procedure.

Since the kernel has a peak in y which does not depend very much on θ, Trickett and Welch replace the argument of the expectation in (5.33) by the value that this function assumes when $Y = \nu_1/(\nu_1+\nu_2)$. If we let $y_* \equiv \nu_1/(\nu_1+\nu_2)$, then the Trickett-Welch approximation is

$$\begin{aligned} & \tilde{h}[\phi(\theta,Y)]\sqrt{(\nu_1+\nu_2)\left(\frac{\theta Y}{\nu_1}+\frac{(1-\theta)(1-Y)}{\nu_2}\right)} \\ & \cdot T'_{\nu_1+\nu_2}\left[v^0_\alpha[\phi(\theta,Y)]\sqrt{(\nu_1+\nu_2)\left(\frac{\theta Y}{\nu_1}+\frac{(1-\theta)(1-Y)}{\nu_2}\right)}\right]\Bigg\} \\ \approx\ & \tilde{h}[\phi(\theta,y_*)]\sqrt{(\nu_1+\nu_2)\left(\frac{\theta y_*}{\nu_1}+\frac{(1-\theta)(1-y_*)}{\nu_2}\right)} \\ & \cdot T'_{\nu_1+\nu_2}\left[v^0_\alpha[\phi(\theta,y_*)]\sqrt{(\nu_1+\nu_2)\left(\frac{\theta y_*}{\nu_1}+\frac{(1-\theta)(1-y_*)}{\nu_2}\right)}\right] \\ =\ & \tilde{h}(\theta)T'_{\nu_1+\nu_2}[v^0_\alpha(\theta)], \end{aligned} \tag{5.34}$$

where we have used

$$\phi(\theta,y_*) = \theta, \tag{5.35}$$

and

$$\sqrt{(\nu_1+\nu_2)\left(\frac{\theta y_*}{\nu_1}+\frac{(1-\theta)(1-y_*)}{\nu_2}\right)} = 1. \tag{5.36}$$

Since $\tilde{h}(\theta)T'_{\nu_1+\nu_2}[v^0_\alpha(\theta)]$ does not depend on the Beta random variable Y, it can be taken out of the expectation in (5.33). Equivalently, since $\tilde{h}(\theta)T'_{\nu_1+\nu_2}[v^0_\alpha(\theta)]$ does not depend on the variable of integration y, it can be taken out of the integrand if (5.33) is written explicitly as an integral equation. Having made this approximation in (5.32), we solve for an $h(\theta)$ which approximates the Newton step $\tilde{h}(\theta)$:

$$\tilde{h}(\theta) \approx h(\theta) = \frac{1 - \alpha - F(v_\alpha^0)}{T'_{\nu_1+\nu_2}[v_\alpha^0(\theta)]}. \tag{5.37}$$

Given v_α^0, we can calculate the right hand side of (5.37) numerically for any values of θ that we choose, and thereby determine the approximate Newton step $h(\theta)$ at as many points as we like. Since θ takes the role of a dummy variable in (5.37), by determining $h(\theta)$ we also determine $h(r)$ for the same values of the independent variable as $h(\theta)$. We let the next approximation to v_α be

$$v_\alpha^1(r) = v_\alpha^0(r) + h(r), \tag{5.38}$$

where we use an interpolation rule in order to get functions for all $r \in [0,1]$.

The Conditional Expectation Algorithm

We now apply the Conditional Expectation algorithm in the form (4.16), that is, by taking one inner iteration, and by using the zero function as the initial iterate for the inner iteration. Using the notation of Chapter 4 for the kernel in the Newton step equation, we write (5.33) as

$$1 - \alpha - F(v_\alpha^0) = \int_0^1 k'\{\theta, y, v_\alpha^0[\phi(\theta, y)]\}\tilde{h}[\phi(\theta, y)]dy \tag{5.39}$$

We know that $k'\{\theta, y, v_\alpha^0[\phi(\theta, y)]\}$ has a peak in y at approximately $y_* = \nu_1/(\nu_1 + \nu_2)$ for all θ, so we approximate $\tilde{h}[\phi(\theta, y)]$ by

$$\tilde{h}[\phi(\theta, y_*)] = \tilde{h}(\theta), \tag{5.40}$$

which leads to the Conditional Expectation method quasi-Newton step

$$\tilde{h}(\theta) \approx h(\theta) = \frac{1 - \alpha - F(v_\alpha^0)}{\int_0^1 k'\{\theta, y, v_\alpha^0[\phi(\theta, y_*)]\}dy} \tag{5.41}$$

In the next section, we illustrate the Conditional Expectation algorithm with quasi-Newton step (5.41) by means of a numerical example.

5.4 A Numerical Example

In Figure 5.2 we present the result of applying the Conditional Expectation algorithm for the case of $n_1 = 20$ and $n_2 = 10$. We have chosen values of the nuisance parameter to be

$$\theta_i = \tfrac{i-1}{m} \quad \text{for } i = 1, \ldots, m+1, \tag{5.42}$$

where $m = 24$. Integration is by 25 point Gauss-Legendre quadrature, and the function v_α^n is interpolated using a linear spline in order to evaluate $F(v_\alpha^n)$ numerically. Appendix C consists of the function, written in the S programming language (Becker, Chambers and Wilks, 1988), which was used to interactively perform the calculations.

The successive approximations v_α^n are displayed in Figure 5.2, and the successive calculations of the actual size (as a function of θ) for a nominal size of $\alpha = .05$ is presented in Figure 5.3. The actual size is calculated at the 25 nuisance parameter values chosen for the discretization. In practice, the true nuisance parameter will be between two of the values used for the discretization, so the numerical demonstration of near similarity in Figure 5.3 is a bit deceiving. However, when the actual size is evaluated for other nuisance parameter values, the actual size is found to be still virtually equal to the nominal size.

We have added to Figures 5.2 and 5.3 the critical value and actual size from the commonly used *Welch's approximate t* method (Welch, 1937; Bickel and Doksum, 1977, p. 219), that is

$$\hat{v}(R)_\alpha \equiv T_{\nu(R)}^{-1}(1-\alpha), \tag{5.43}$$

where T_ν^{-1} is the inverse of the t cumulative, and the degrees of freedom ν is given by

$$\nu(R) \equiv \left[\frac{R^2}{n_1-1} + \frac{(1-R)^2}{n_2-1}\right]^{-1}, \tag{5.44}$$

and R is the nuisance parameter estimate (5.6). Although the Conditional Expectation results are outstanding, the simple approximate t method also provides a nearly similar test.

Welch's approximate t is certainly easier to use than the method which results from 'solving' the Trickett-Welch integral equation, and for most applications the approximate t provides a test that is as near to being similar as is necessary. But the Conditional Expectation method can serve another purpose, even if it is not used very often in practice. Ad-hoc approximations such as Welch's t are often compared on the basis of their nearness to similarity, and also power (see, e.g., Best and Rayner, 1987). Measuring the distances between an ad-hoc critical function and a critical function for an 'exactly' similar test provides information on how close a proposed *confidence interval* comes to 'the best possible' result.

No detail can be obtained from Figure 5.3 except for the first few iterations because the convergence to similarity is so rapid. Also, it is difficult to infer much about the rate of convergence from Figure 5.3. In Figure 5.4 we see the distance, in the L_∞ norm (maximum absolute deviation) , from the nominal size in a semilog plot against the iteration number.

We can see from Figure 5.4 that the rate of convergence is rapid at first, and then eventually decreases to the point where, after twenty or so steps, it hardly seems worthwhile to continue. Intuitively, this is consistent with our previous discussion of the convergence of Richardson's algorithm. After the components of the initial approximation in the directions of the dominant eigenfunctions decay, the 'less important' eigenfunctions remain, and these decay much more slowly since they correspond to smaller eigenvalues.

In summary, these results are quite spectacular. Although there is no exact solution to the Behrens-Fisher problem, we are able to easily determine a *smooth* critical function for which the L_∞ distance of the right hand side from $1-\alpha$ is less than 10^{-6}!

5.5 The Power of Tests for the Behrens-Fisher Problem

If the null hypothesis is *not* true, then $P(U < v_\alpha(R))$ is the power function for a test using the critical value $v_\alpha(R)$. We can easily obtain an expression for the power, using the same approach as in Section 5.2. Since the null hypothesis is not true, we obtain

$$U = \frac{Z+\delta}{\sqrt{W/(n_1+n_2)}} \sim T_{n_1+n_2}(\delta) \tag{5.45}$$

where $T_{n_1+n_2}(\delta)$ denotes the *noncentral t* cumulative with $n_1 + n_2$ degrees of freedom and noncentrality parameter δ given by (5.10) instead of the expression

$$\frac{Z}{\sqrt{W/(n_1+n_2)}} \sim T_{n_1+n_2}, \tag{5.46}$$

which appears in Section 5.2. The power is

$$\pi(\delta)_\theta = 1 - E\left\{T_{\nu_1+\nu_2}\left[v_\alpha(R)\left\{(n_1+n_2)\left[\frac{Y\theta}{\nu_1} + \frac{(1-Y)(1-\theta)}{\nu_2}\right]\right\}^{1/2}, \delta\right]\right\}, \tag{5.47}$$

where the notation $\pi(\delta)_\theta$ indicates that the power is a family of functions of δ, indexed by the nuisance parameter value θ.

We calculate (5.47) numerically next as a continuation of the numerical example of the previous section. As an example of a power calculation, we compare the Conditional Expectation procedure with Welch's approximate t. It only makes sense to compare the pcwer of tests which have the same size, so we begin by examining Figure 5.2 in order to determine a θ value for which the sizes of the two methods are nearly the same. We thereby choose $\theta = .35$ for the value of the nuisance parameter, and calculate the size of the Conditional Expectation method to be .05000073, and the size of Welch's t to be .049928. The power function for the Conditional Expectation method is displayed in Figure 5.5. The power curve for Welch's t is not graphed in Figure 5.5 since it would not be discernible from the other power function. The difference in the two power functions (times 1000) is displayed in Figure 5.6. Note that the maximum difference between the powers is not much larger than the (very small) difference between the sizes. It seems that the test derived by the *Conditional Expectation algorithm* achieves near similarity without sacrificing power relative to Welch's approximate t test.

Figure 5.1: Derivative of Kernel for Behrens-Fisher Problem (n1=20, n2=10)

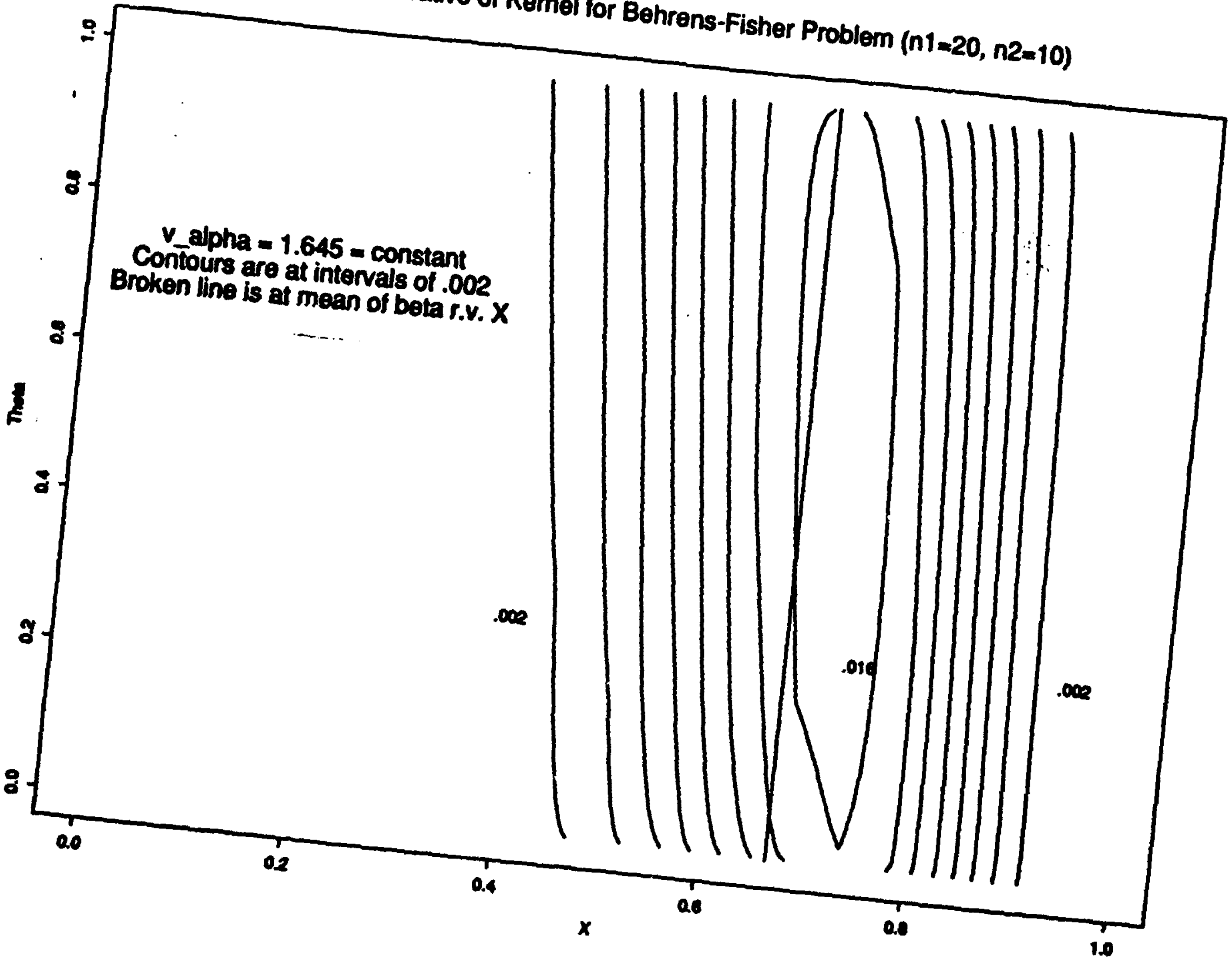


Figure 5.2: The Trickett-Welch Approximations to the Critical Function

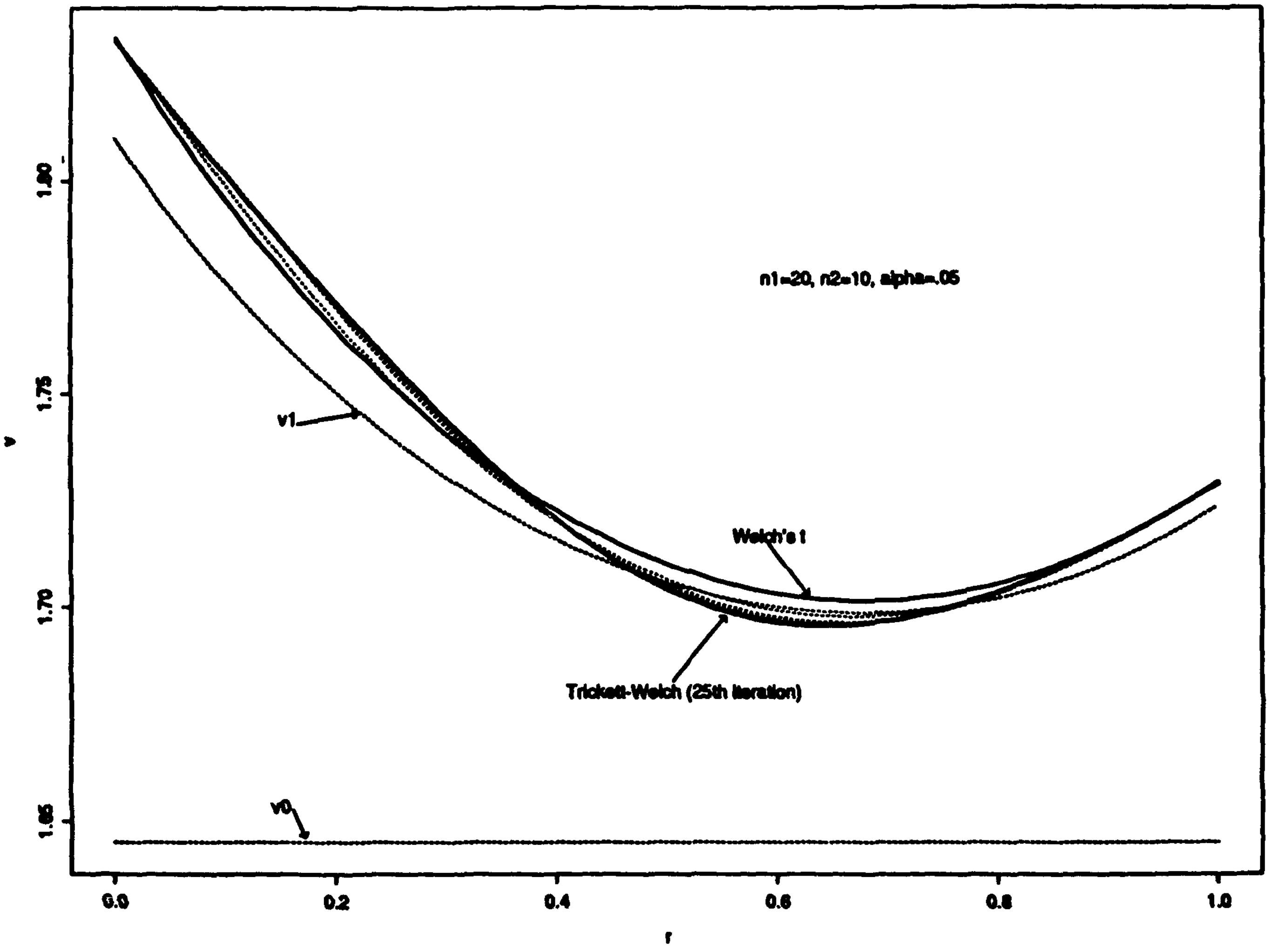


Figure 5.3: Size for Trickett-Welch Critical Value Function Approximations

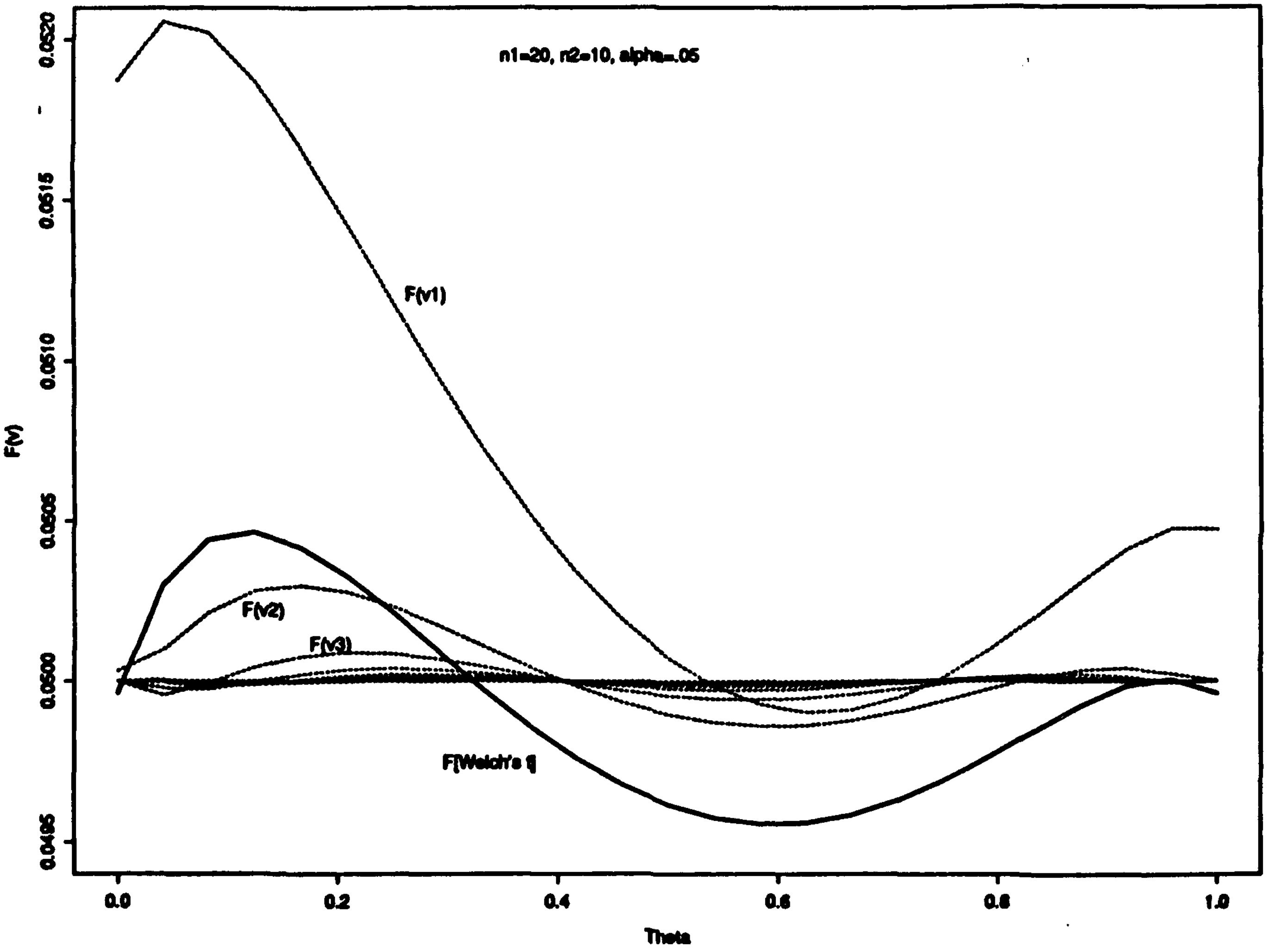


Figure 5.4: Convergence to Similarity of the Trickett-Welch Iterates

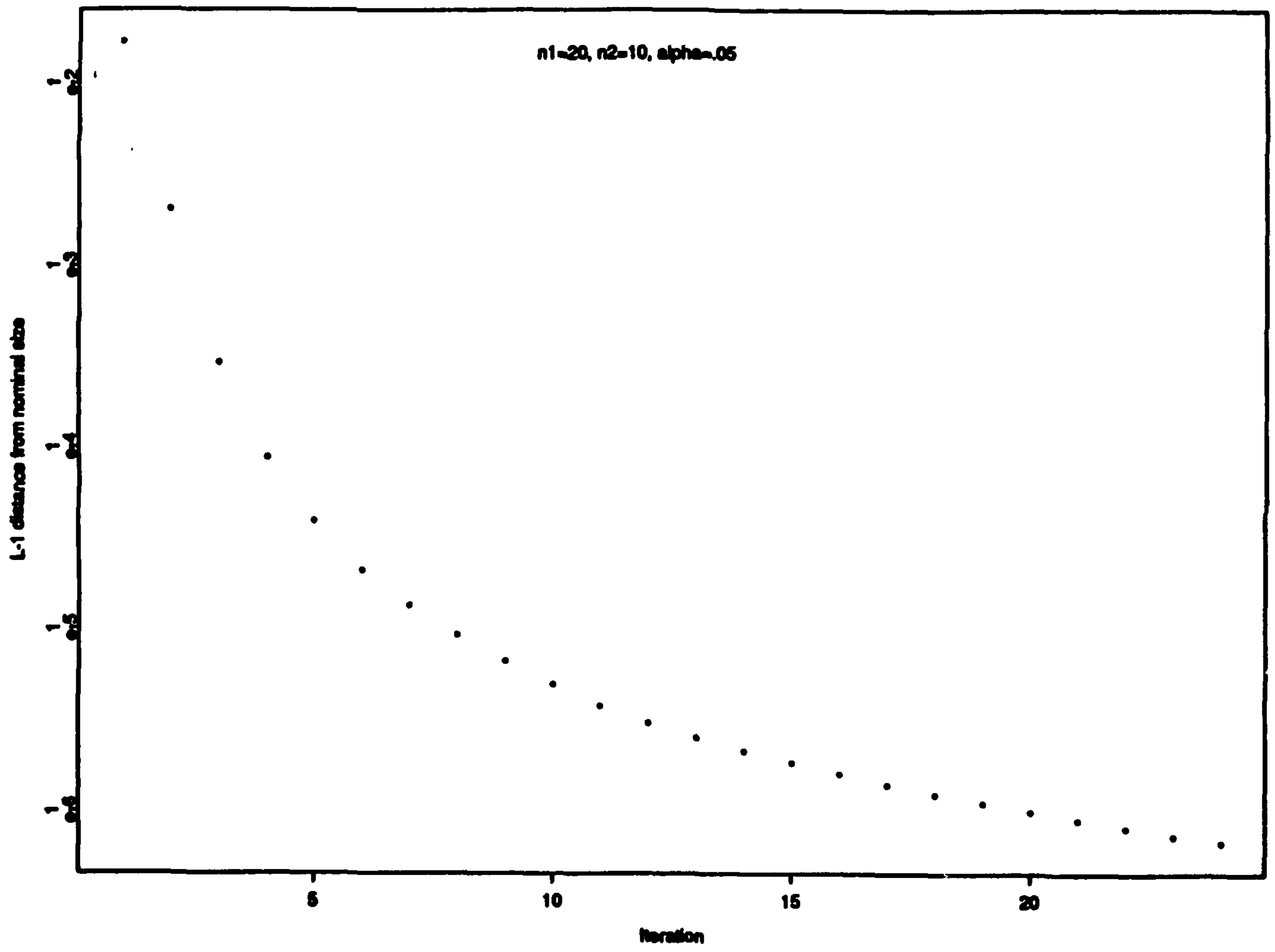


Figure 5.5: Power Function for Trickett-Welch Test

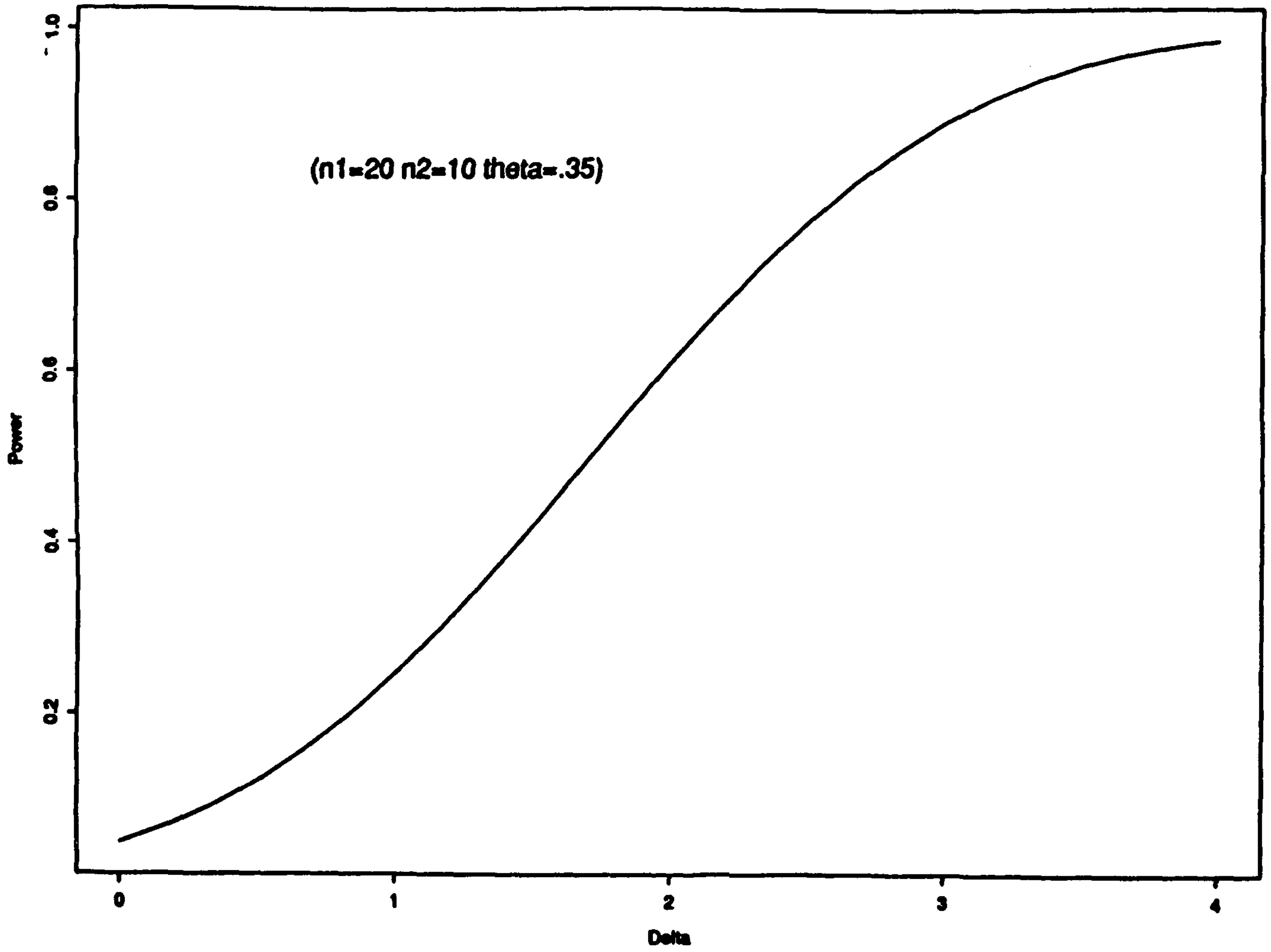
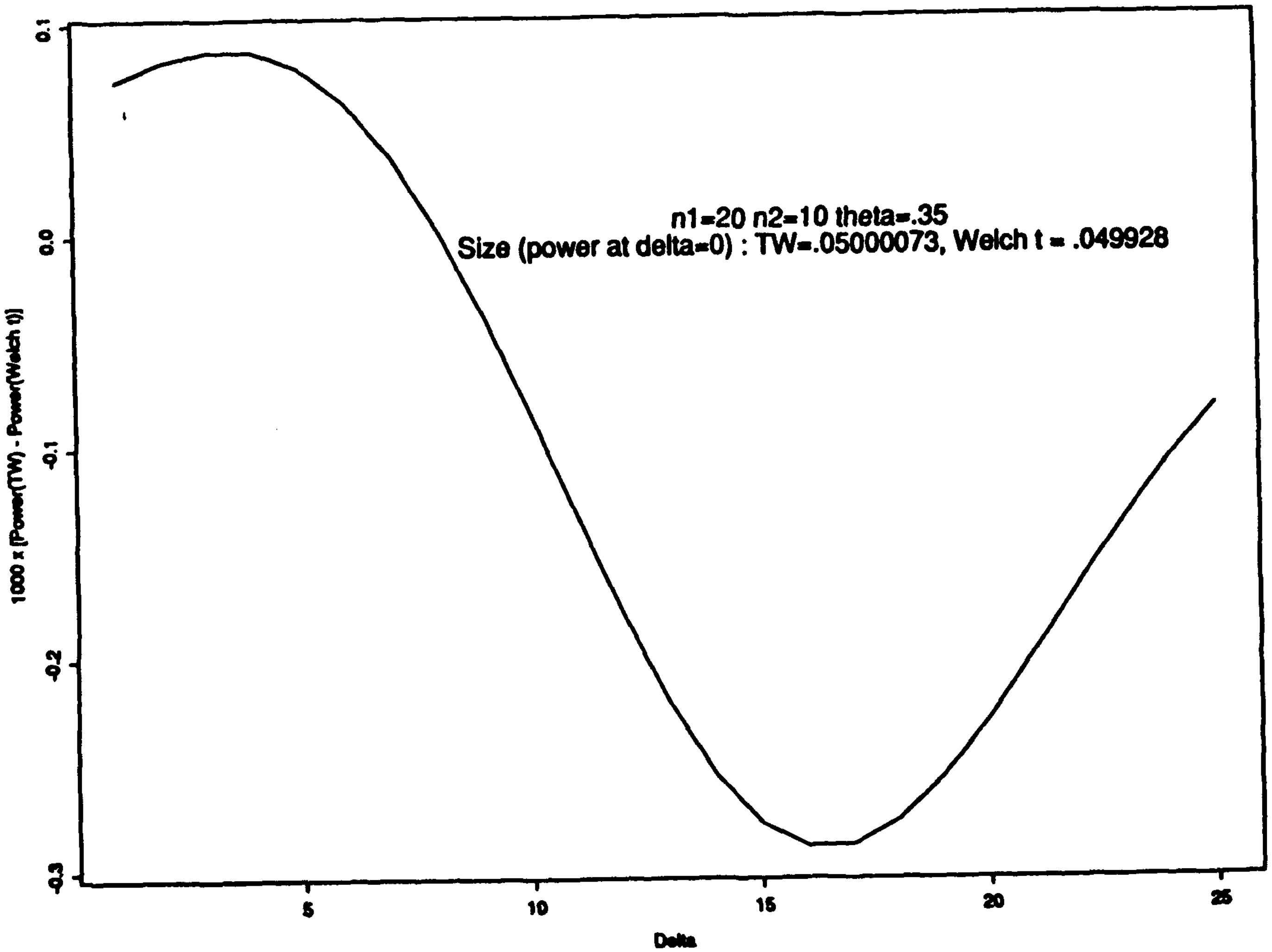


Figure 5.6: . Difference in Power Functions -- T.W. vs. Welch t



Chapter 6

One-Sided Tolerance Limits for a One-Way Balanced Random-Effects ANOVA Model

6.1 Other Applications of Iterative Algorithms

The Behrens-Fisher problem is only one example of a normal-theory problem with a nuisance parameter. Other examples include: confidence intervals for the common mean of two normal populations and one-sided prediction intervals for a one-way balanced random effects model. The (unmodified) Trickett-Welch algorithm has been applied successfully to these problems by Maric and Graybill (1979) and Wang (1988), respectively. In fact, Maric and Graybill apparently independently discovered the Trickett-Welch algorithm.

We will discuss another problem, one which has some importance for applications and which was the starting point for this thesis. This problem concerns one-sided confidence intervals for a quantile of a normal population with two components of variance estimated using data from a one-way balanced random-effects ANOVA model.

These tolerance limits are important for characterizing the strength of composite materials (Mil-HDBK-17C, 1992), and there was concern over the conservatism of an approximate procedure for this problem due to Mee and Owen (1983). Attempts to reduce this conservatism led to the application of the ideas first of Welch (1947) and later of Trickett and Welch (1954). The integral equation for this tolerance limit problem is substantially more complicated than the Trickett-Welch equation (5.25) of Chapter 5. Consequently only first order terms of the Welch-Aspin type asymptotic expansion are tractable, and the Trickett-Welch algorithm does not work at all. However, the Conditional Expectation algorithm is very effective on this problem. In addition to this thesis, this work is reported on in Vangel (1987, 1990, 1992).

6.2 The Tolerance Limit Problem

Let X be a normally distributed random variable with mean μ and variance $\sigma^2 = \sigma_b^2 + \sigma_e^2$. A lower confidence limit for a quantile of this population (i.e., a lower tolerance limit) is to be determined using data from a one-way balanced random effects ANOVA sample with between-group and within-group variances σ_b^2 and σ_e^2 respectively.

For example, let X represent the strength of a randomly selected specimen of a material manufactured in a batch which can be considered to be randomly selected from a population of batches. A quantity of interest to aircraft designers is the 'B-basis value', which is a 95 percent lower confidence limit on the tenth percentile of the distribution of X. For this situation, it is important that nearly the nominal coverage probability be attained whatever the unknown population variance ratio. It is also very desirable that the calculated limit be as large as possible, since unnecessarily low values cause undue conservatism in design.

We discuss below techniques for determining one-sided tolerance limits for X based on a random sample of J items from each of I batches. A (β,γ) ***lower tolerance limit*** is a statistic T such that at least a proportion β of the population is covered by the interval (T,∞) with probability at least γ. The methods developed here for lower tolerance limits can be adapted in an obvious way to upper limits. We will refer to β as the *coverage* and γ as the *confidence.*

This problem was first considered by Lemon (1977) who proposed an approximate solution too conservative for most applications. Mee and Owen (1983) greatly improved on Lemon's results by using a Satterthwaite (1947) approximation. Seeger and Thorsson (1972) proposed the same approximation for the corresponding two-sided problem. The Mee-Owen method is reviewed in Vangel (1990) and will not be described here. Instead, we will regard this problem as a typical normal-theory inverse problem, requiring the solution of an integral equation, and apply the Conditional Expectation algorithm.

First we shall consider the case where the nuisance parameters are known. Then we shall develop a Welch-Aspin type of expansion. The latter can serve as an initial approximation for the Conditional Expectation algorithm.

6.3 The One-Way Balanced Random-Effects Model

Let X_{ij} denote the jth of J observations from the ith of I batches. If X_{ij} follows a one-way balanced random-effects model, then

$$X_{ij} = \mu + b_i + e_{ij}, \tag{6.1}$$

where μ denotes the population mean, $\mu + b_i$ denotes the mean of the ith batch, and e_{ij} is the error term. The b_i's and the e_{ij}'s are assumed to be independently distributed normal with mean zero and variance σ_b^2 and σ_e^2 respectively. An observation X from this pcpulation is thus normally distributed with mean μ and variance

$$\sigma_X^2 = \sigma_b^2 + \sigma_e^2. \tag{6.2}$$

Let $n = IJ$ denote the sample size. The parameters μ, σ_e^2 and σ_b^2 of the random effects model can be estimated by the pooled mean $\hat{\mu}$, the within batch mean square MS$_e$, and a linear combination of MS$_e$ with the between batch mean square MS$_b$ where:

$$\hat{\mu} = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{X_{ij}}{IJ}, \tag{6.3}$$

$$\bar{X}_i = \sum_{j=1}^{J} X_{ij}/J, \tag{6.4}$$

$$\mathrm{MS}_b = J \sum_{i=1}^{I} \frac{(\hat{\mu} - \bar{X}_i)^2}{I - 1}, \tag{6.5}$$

and

$$\mathrm{MS}_e = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(X_{ij} - \bar{X}_i)^2}{I(J - 1)}. \tag{6.6}$$

An unbiased estimator of the population variance σ_X^2 is

$$\hat{\sigma}_X^2 = \mathrm{MS}_b/J + (1 - 1/J)\mathrm{MS}_e. \tag{6.7}$$

For $0 < \beta < 1$, let z_β be the β quantile of the standard normal distribution, i.e

$$\beta = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_\beta} e^{-t^2/2} dt. \tag{6.8}$$

A (β, γ) lower tolerance limit is a 100γ percent lower confidence bound for

$$\mu - z_\beta \sigma_X. \tag{6.9}$$

By analogy with the single sample case (see, for example, Owen (1968)), we seek an estimator of the form

$$\hat{\mu} - k\hat{\sigma}_X, \tag{6.10}$$

where k is chosen to satisfy, for all σ_b^2 and σ_e^2,

$$P(\hat{\mu} - k\hat{\sigma}_X \leq \mu - z_\beta \sigma_X) = \gamma. \tag{6.11}$$

Since $\hat{\mu}$ has a normal distribution with mean μ and variance

$$\sigma_{\hat{\mu}}{}^2 = (J\sigma_b^2 + \sigma_e^2)/n, \tag{6.12}$$

we can rewrite (6.11) as

$$P\left(\frac{Z + \sqrt{n} z_\beta b}{\hat{\sigma}_X/\sigma_X} \leq \sqrt{n} k b\right) = \gamma, \tag{6.13}$$

where

$$Z \equiv \frac{\hat{\mu} - \mu}{\sigma_{\hat{\mu}}}, \tag{6.14}$$

$$b \equiv \sqrt{\frac{r + 1}{Jr + 1}}, \tag{6.15}$$

and

$$r \equiv \sigma_b^2/\sigma_e^2. \tag{6.16}$$

The random variable (6.14) has a N(0,1) distribution, and is independent of (6.7), whose component terms are independent.

6.4 An Exact Solution for Known r

For a simple random sample, a solution to the one sided tolerance limit problem is readily obtained in terms of the noncentral t distribution (see, e.g., Owen 1968). If one assumes that the variance ratio r is known, then the corresponding problem for a sample from a balanced random effects model can be solved almost as easily. What is required is the distribution of a 'generalized noncentral t' random variable, a generalization of the noncentral t to a random variable with the square root of a linear combination of two χ^2 random variables in the denominator. In this section, we derive this distribution, and then we show how it can be used to solve the tolerance limit problem for known r.

Let the random variables Z, Y_1, and Y_2 have the following distributions:

$$Z \sim N(0,1), \tag{6.17}$$

and

$$Y_j \sim \chi^2_{n_j}, \tag{6.18}$$

for $j = 1,2$, where Z, Y_1, and Y_2 are mutually independent. We will call A a *generalized noncentral t* random variable if A has the form

$$A = (n_1 + n_2)^{1/2} \frac{Z+\delta}{\sqrt{d_1Y_1 + d_2Y_2}}, \tag{6.19}$$

where d_1, d_2, and δ are constants with d_1 and d_2 positive.

We will find the distribution of (6.19) by a technique very similar to the approach used in Section 5.2.1. We express A as the product of a *noncentral t random variable* times an expression involving only known constants and a beta random variable. Since the two terms in this product are independent, conditioning on the beta random variable and integrating yields the distribution of A.

The random variable A is easily seen to be equal to

$$\begin{aligned} A &= \frac{Z+\delta}{\sqrt{(Y_1+Y_2)/(n_1+n_2)}} \left[\frac{Y_1+Y_2}{d_1Y_1+d_2Y_2}\right]^{1/2} \\ &= T\left[\frac{d_1Y_1}{Y_1+Y_2} + \frac{d_2Y_2}{Y_1+Y_2}\right]^{-1/2} \\ &= \frac{T}{\sqrt{d_1Y + d_2(1-Y)}}, \end{aligned} \tag{6.20}$$

where T has the noncentral t distribution with degrees of freedom n_1+n_2 and noncentrality parameter δ, denoted $T_{n_1+n_2}(\delta)$, and Y has the beta distribution, with parameters $n_1/2$ and $n_2/2$, denoted Beta $(n_1/2, n_2/2)$. It is well known (e.g., Fleiss, 1971) that Y has the claimed beta distribution, and that $L \equiv Y_1 + Y_2$ is independent of Y. Since Z is independent of Y_j, hence of Y, by assumption, it follows that T and Y are independent.

By conditioning on Y, we see that

$$\begin{aligned} F_A(t) \equiv P(A \le t) &= P\left[T \le t\sqrt{d_1Y + d_2(1-Y)}\right] \\ &= E\left[T_{n_1+n_2}\left(t\sqrt{d_1Y + d_2(1-Y)}, \delta\right)\right], \end{aligned} \tag{6.21}$$

where $T_f(t,\delta)$ denotes the noncentral t cumulative distribution with f degrees of freedom and noncentrality parameter δ, that is

$$\begin{aligned} T_f(t,\delta) &= P\left[Z \le t\sqrt{\frac{Y}{f}} - \delta\right] \qquad (6.22) \\ &= \int_0^\infty \Phi\left(t\sqrt{\frac{y}{f}} - \delta\right) C_f(y)\,dy, \end{aligned}$$

where C_f denotes the χ^2 density with f degrees of freedom and $\Phi(\cdot)$ is the standard normal distribution. Thus, $F_A(t)$ can be expressed as an integral of a function of y, i.e. the argument of T in (6.21) times the beta density. F_A is a distribution with argument t and implicit parameters n_1, n_2, d_1, d_2, and δ.

For the tolerance limit problem, we have using (6.13) that,

$$P\left(\frac{Z + \sqrt{n}z_\beta b}{\hat{\sigma}_X/\sigma_X} \le \sqrt{n}kb\right) = P\left(\frac{Z + \sqrt{n}z_\beta b}{\sqrt{n}b\hat{\sigma}_X/\sigma_X} \le k\right) = \gamma. \qquad (6.23)$$

Let $n_1 = I - 1$ and $n_2 = I(J-1)$: the between-group and within-group degrees of freedom, respectively. Since the mean squares MS_b and MS_e are proportional to χ^2 random variables, we see that

$$\text{MS}_b = (J\sigma_b^2 + \sigma_e^2)Y_1/n_1, \qquad (6.24)$$

$$\text{MS}_e = \sigma_e^2 Y_2/n_2, \qquad (6.25)$$

and

$$\hat{\sigma}_X^2 = (\sigma_b^2 + \sigma_e^2/J)Y_1/n_1 + (1 - 1/J)\sigma_e^2 Y_2/n_2. \qquad (6.26)$$

Simple algebra now leads to

$$\frac{\hat{\sigma}_X^2}{\sigma_X^2} nb^2 = \frac{IY_1}{I-1} + \frac{Y_2}{Jr+1}, \qquad (6.27)$$

where r is the variance ratio, $r \equiv \sigma_b^2/\sigma_e^2$.

If we let

$$d_1 = \frac{(n_1+n_2)I}{I-1}, \qquad (6.28)$$

and

$$d_2 = \frac{n_1+n_2}{Jr+1}, \qquad (6.29)$$

then, for these specific values of n_1, n_2, d_1, and d_2, with r known, we have that

$$P(\hat{\mu} - k\hat{\sigma}_X \le \mu - z_\beta\sigma_X) = F_A(k) \qquad (6.30)$$

and

$$\delta = z_\beta b\sqrt{n} = z_\beta\sqrt{\frac{n(r+1)}{Jr+1}}. \qquad (6.31)$$

A constant or function, such as the constant k in (6.30), which leads to a tolerance limit is called a *tolerance limit factor*. The value $k(r)$ of k such that $F_A(k) = \gamma$ thus provides an exact solution to the problem for known variance ratio r, where $F_A(k)$ depends on σ_b^2 and σ_e^2 only through r.

Later we will consider the case where the constant k is replaced by a function

$$c = c(\mathrm{MS}_b, \mathrm{MS}_e).$$

In that case, t in (6.21) can be replaced by c as long as c can be represented as a function of the *mean square ratio*:

$$Q \equiv \frac{\mathrm{MS}_b}{\mathrm{MS}_e} = (Jr+1)\frac{n_2 Y_1}{n_1 Y_2}. \tag{6.32}$$

The result is an integral equation depending on σ_b^2 and σ_e^2 only through r.

6.5 The Solution for Unknown r: A Welch-Aspin Type Asymptotic Expansion

For unknown variance ratio, the tolerance limit problem is closely related to the Behrens-Fisher problem. Since it is well known that there is no 'well behaved' solution to the Behrens-Fisher problem, it is likely that a tolerance limit factor for which the corresponding tolerance limit has exactly the nominal confidence for all r does not exist. However, we can proceed as if a tolerance limit factor does exist and attempt to approximate it. Following the work of Welch (1947), Aspin (1948), and Trickett and Welch (1954), we will propose three tolerance limit factors, which we will sometimes refer to as 'solutions'.

The first solution discussed is based on an asymptotic expansion (for large I and J) of the type considered by Welch and Aspin; we will call this solution the *Asymptotic Expansion tolerance limit factor.* While computationally simple, the first order approximation presented here is anticonservative and may only be suitable for many batches.

We could improve this procedure by taking higher order approximations. However, this becomes very tedious to carry out. Instead, we propose an ad-hoc modification to the Asymptotic Expansion tolerance limit factor which is *very* easy to use and which is adequate for most applications. We will refer to this result as the *Modified Asymptotic Expansion tolerance limit factor.* In Section 6.6, the tolerance limit factor as a function of the mean square ratio will be obtained approximately as a solution of an integral equation by means of the Conditional Expectation algorithm. The *Conditional Expectation tolerance limit factor* which results provides confidence *extremely* close to the nominal level for all values of the nuisance parameter: even for very small sample sizes.

To simplify the notation in what follows, let S_i^2 be the mean squares, σ_i^2 their expected values, and n_i the associated degrees of freedom for $i = 1, 2$, i.e. :

$$\begin{aligned} S_1^2 &= \mathrm{MS}_b, & \sigma_1^2 &= J\sigma_b^2 + \sigma_e^2, & n_1 &= I - 1, \\ S_2^2 &= \mathrm{MS}_e, & \sigma_2^2 &= \sigma_e^2, & n_2 &= I(J-1). \end{aligned}$$

The pooled sample size is $n = IJ$ and the population variance is denoted by

$$\sigma^2 = \sigma_X^2 = \sigma_b^2 + \sigma_e^2 = \sigma_1^2/J + \sigma_2^2(1 - 1/J), \tag{6.33}$$

and estimated by

$$S^2 = \hat{\sigma}_X^2 = S_1^2/J + S_2^2(1 - 1/J). \tag{6.34}$$

The subscript X for the population variance and the estimate of this variance will be omitted for the remainder of this section.

We will consider tolerance limits of the form $\hat{\mu} - k\hat{\sigma}_X$. If the variance ratio r were known, then the factor $k(r)$ determined from the generalized noncentral t distribution in Section 6.4 would be appropriate. Since r is not known, but can be estimated as a function of S_1^2 and S_2^2, we will replace k by $c(S_1^2, S_2^2)$. We will call c the tolerance limit factor, and we define $h(S_1^2, S_2^2)$ to be $c\hat{\sigma}$.

The tolerance limit corresponding to this factor c can be expressed as an expectation with respect to the distributions of the mean squares in terms of the standard normal distribution, so that (6.11) becomes

$$\begin{aligned}\gamma &= P(\hat{\mu} - c\hat{\sigma} \leq \mu - z_\beta \sigma) \\ &= E\left[\Phi\left(\frac{c\hat{\sigma}}{\sigma_1/\sqrt{n}} - \delta\right)\right] \\ &= E\left[\Phi\left(\frac{h\left(S_1^2, S_2^2\right)}{\sigma_1/\sqrt{n}} - \delta\right)\right], \end{aligned} \tag{6.35}$$

where as above

$$\delta = z_\beta \sqrt{\frac{n(r+1)}{Jr+1}} = \frac{z_\beta \sigma}{\sigma_1/\sqrt{n}}. \tag{6.36}$$

The problem is to determine a function $h(S_1^2, S_2^2)$ so that (6.35) is approximately satisfied for all σ_1^2 and σ_2^2. If tolerance limits on the median are desired, then $\delta = 0$ and the results of Welch (1947) and Aspin (1948) can be used directly. If δ is not zero, the idea behind the Welch-Aspin derivation can still be applied, although the algebra is considerably messier.

The Welch-Aspin approach makes use of differential operators in order to develop an asymptotic expansion for h (for large I and J). The same approach can be used here, but for first order calculations the algebraic simplifications which result are insufficient to justify the additional formalism which the operator technique requires. Hence, the discussion below consists of a straightforward Taylor series derivation. Of course, both methods must give the same answer, and this has been used to provide a check on the calculations.

We begin by rewriting (6.35) as

$$E\left[\Phi\left(z_\gamma + U\right)\right] = \gamma, \tag{6.37}$$

where

$$z_\gamma + U = \frac{h\left(S_1^2, S_2^2\right)}{\sigma_1/\sqrt{n}} - \delta. \tag{6.38}$$

We then expand h in a series of inverse half-powers of n_i:

$$h = h_0 + h_1 + o(I^{-1/2} + J^{-1/2}). \tag{6.39}$$

Up to terms of second order in h we have that

$$U = \frac{h_1(S_1^2, S_2^2)}{\sigma_1/\sqrt{n}} + \frac{h_0(S_1^2, S_2^2)}{\sigma_1/\sqrt{n}} - \frac{z_\beta \sigma}{\sigma_1/\sqrt{n}} - z_\gamma, \tag{6.40}$$

where we have substituted (6.36) for δ.

For the zeroth order approximation we approximate $h(S_1^2, S_2^2)$ by

$$h_0(S_1^2, S_2^2) \approx h_0(\sigma_1^2, \sigma_2^2)$$

and we have, $U = 0$ and

$$\frac{h_0(\sigma_1^2, \sigma_2^2) - z_\beta \sigma}{\sigma_1/\sqrt{n}} - z_\gamma = 0 \tag{6.41}$$

or

$$h_0(S_1^2, S_2^2) = \frac{z_\gamma S_1}{\sqrt{n}} + z_\beta S. \tag{6.42}$$

For the first order expression, we approximate h by

$$\begin{aligned} h_0(S_1^2, S_2^2) + h_1(S_1^2, S_2^2) \approx h_0(S_1^2, S_2^2) + h_1(\sigma_1^2, \sigma_2^2) = \\ z_\beta \sigma \left[1 + \left(\frac{S}{\sigma} - 1\right)\right] + \frac{z_\gamma \sigma_1}{\sqrt{n}} \left[1 + \left(\frac{S_1}{\sigma_1} - 1\right)\right] + h_1(\sigma_1^2, \sigma_2^2) \end{aligned} \tag{6.43}$$

then

$$U = z_\gamma U_1 + z_\beta U_2 + \frac{h_1(\sigma_1^2, \sigma_2^2)}{\sigma_1/\sqrt{n}}, \tag{6.44}$$

where

$$U_1 = \frac{S_1}{\sigma_1} - 1 \tag{6.45}$$

and

$$U_2 = \frac{\sigma}{\sigma_1/\sqrt{n}} \left(\frac{S}{\sigma} - 1\right). \tag{6.46}$$

Let Y_i denote a χ^2 random variable with n_i degrees of freedom and define

$$V_{n_i} \equiv \frac{Y_i}{n_i} - 1 \tag{6.47}$$

for $i = 1, 2$. The U_i can be expressed in terms of the V_{n_i} as follows:

$$\begin{aligned} U_1 &= (1 + V_{n_1})^{1/2} - 1, & (6.48) \\ U_2 &= \frac{\sigma}{\sigma_1/\sqrt{n}} \left[\left(\frac{\sigma_2^2 (J-1)}{\sigma^2 J} (1 + V_{n_2}) + \frac{\sigma_1^2}{\sigma^2 J} (1 + V_{n_1}) \right)^{1/2} - 1 \right]. & (6.49) \end{aligned}$$

After expanding the square roots in (6.48) and (6.49) in power series, one can readily obtain approximations to the first two moments of the U_i suitable for first order calculations:

$$\begin{aligned} E(U_1) &\approx -\frac{1}{4n_1}, & (6.50) \\ E(U_1^2) &\approx \frac{1}{2n_1}, & (6.51) \\ E(U_2) &\approx -\frac{1}{4}\left[\frac{a_1}{n_1} + \frac{a_2}{n_2}\right], & (6.52) \\ E(U_2^2) &\approx \frac{\sigma}{2\sigma_1/\sqrt{n}}\left[\frac{a_1}{n_1} + \frac{a_2}{n_2}\right], & (6.53) \end{aligned}$$

and

$$E(U_1 U_2) \approx \frac{\sigma_1}{J\sigma/\sqrt{n}} \frac{1}{2n_1}, \tag{6.54}$$

where

$$a_1 \equiv \frac{\sigma_1^3}{\sigma^3}\frac{n^{1/2}}{J^2} \tag{6.55}$$

and

$$a_2 \equiv \frac{\sigma_2^4}{\sigma_1\sigma^3}(1 - J^{-1})^2 n^{1/2}. \tag{6.56}$$

The next step is to expand the normal cdf about z_γ, so that (6.37) can be replaced by the following approximation:

$$\begin{aligned} E\left[\Phi\left(z_\gamma + U\right)\right] = \gamma \approx \Phi(z_\gamma) \\ +\phi(z_\gamma)E(U) - z_\gamma\phi(z_\gamma)E(U^2)/2, \end{aligned} \tag{6.57}$$

where $\phi(\cdot)$ denotes the standard normal density. The expectation of U can be determined immediately from (6.44), (6.50) amd (6.52). Since

$$E(U^2) \approx z_\gamma^2 E(U_1^2) + z_\beta^2 E(U_2^2) + 2z_\beta z_\gamma E(U_1 U_2), \tag{6.58}$$

we need only substitute (6.51), (6.53) and (6.54) into (6.58) in order to complete the evaluation of (6.57).

To complete these calculations, solve (6.57) for $h_1(\sigma_1^2, \sigma_2^2)$ (note that h_1 appears through $E(U)$), replace each occurrence of σ_i^2 or σ^2 with S_i^2 or S^2 respectively ($i = 1, 2$) and divide $h_1(S_1^2, S_2^2)$ by S to finally obtain the tolerance limit factor c. The terms of $c = (h_0 + h_1)/S$ can then be rearranged to reveal their structure. The following expression for the *Asymptotic Expansion* tolerance limit factor c is one possibility:

$$\begin{aligned} c = {} & z_\beta + \frac{z_\gamma W}{\sqrt{I}} + \frac{W}{4\sqrt{I}}\left[\frac{z_\gamma(z_\gamma^2 + 1)}{n_1}\right. \\ & + \frac{2z_\beta z_\gamma^2\sqrt{I}W}{n_1} + \frac{z_\beta^2 z_\gamma I W^2}{n_1} \\ & + \frac{z_\beta\sqrt{I}W^3}{n_1} + \frac{z_\beta^2 z_\gamma I(J-1)^2 W^2}{n_2 Q^2} \\ & \left. + \frac{z_\beta(J-1)^2\sqrt{I}W^3}{n_2 Q^2}\right], \end{aligned} \tag{6.59}$$

where

$$W \equiv (1 + (J-1)/Q)^{-1/2} \tag{6.60}$$

and

$$Q = \frac{S_1^2}{S_2^2}. \tag{6.61}$$

The confidence for the above approximation as a function of the population variance ratio is plotted in Figure 6.1 for a (.90, .95) tolerance limit and $J = 5$. Note that for many batches this solution performs well, though for few batches it is anticonservative.

Table 6.1: *Range in Actual Confidence for Approximate Tolerance Limit c^**

		(β,γ)					
I	J	(.90..95)		(.99,.95)		(.99,.99)	
3	2	.929	.962	.931	.962	.970	.993
3	5	.921	.962	.914	.962	.954	.992
3	10	.927	.962	.922	.962	.956	.992
5	2	.942	.960	.940	.960	.981	.993
5	5	.944	.962	.945	.962	.980	.993
5	10	.950	.962	.950	.963	.982	.993
10	2	.950	.958	.950	.958	.989	.992
10	5	.950	.960	.950	.960	.990	.993
10	10	.950	.964	.950	.971	.990	.994

6.5.1 A Simple, Accurate Tolerance Limit Factor Based on an Asymptotic Expansion

The following two steps lead to an improved tolerance limit factor based on equation (6.59). First, omit the terms in (6.59) which are proportional to $1/Q^2$, since these are singular at $Q = 0$ and are very small for moderate to large Q. What remains is a polynomial in W. The random variable $z_\beta\sqrt{I}/W$ estimates the noncentrality parameter δ defined in (6.36), so a polynomial in W is a polynomial in powers of estimates of the reciprocal of δ.

There are many ways to choose the coefficients and terms of a polynomial in W so as to provide approximate tolerance limit factors with good properties. The following approximation performs remarkably well, considering its extreme simplicity:

$$c^* \equiv \begin{cases} [u_{IJ} - u_I/\sqrt{J} + (u_I - u_{IJ})W]/(1 - 1/\sqrt{J}) & \text{for } Q > 1 \\ u_{IJ} & \text{for } Q \leq 1 \end{cases}, \tag{6.62}$$

where u_l denotes the corresponding tolerance limit factor for a simple random sample of size l. We will refer to c^* as the *Modified Asymptotic Expansion* tolerance limit factor.

As $Q \to \infty$, $W \to 1$ and $c^* \to u_I$. If $Q = 1$, then the variance estimate (6.12) is equal to the pooled sample variance. Since $c^* = u_{IJ}$ when $Q = 1$, the approximate tolerance limit for the random effects model, using c^* with $Q = 1$, will exactly equal the corresponding simple random sample tolerance limit factor for the pooled data. If $Q < 1$, then we take c^* to equal u_{IJ} so that the random effects tolerance limit factor will never be less than the tolerance limit factor corresponding to a simple random sample of size IJ. Truncating Q in this way is reasonable since Q estimates $Jr + 1$, which cannot be less than one.

For any sample size, therefore, c^* will provide a tolerance limit which is exact in the limit of large r and conservative (because of the requirement that c^* not exceed u_{IJ}) for r near zero. For intermediate r, this tolerance limit can be anticonservative, although the anticonservatism is not prohibitive, except possibly for very few batches. In Table 6.1, the range in the actual confidence of the tolerance limit factor (6.62) is given for selected values of β, γ, I and J.

6.6 The Conditional Expectation Tolerance Limit Factor

For small samples, the first order approximation developed above may not be adequate, and higher order calculations are clearly prohibitive. An alternative approach to be discussed next is to formulate the problem as an integral equation, and iteratively improve on the first order approximation numerically.

It is convenient to transform from the parameter r to

$$\tau \equiv Jr + 1 = \frac{\sigma_1^2}{\sigma_2^2}. \tag{6.63}$$

Then

$$P(\hat{\mu} - k\hat{\sigma}_X \leq \mu - z_\beta \sigma_X) = \tag{6.64}$$

$$E\left[T_{n_1+n_2}\left(k(\tau)(n_1+n_2)^{1/2}\sqrt{\frac{YI}{I-1}+\frac{1-Y}{\tau}}, \delta(\tau)\right)\right] = \gamma,$$

where

$$\delta(\tau) = \sqrt{n} z_\beta b = z_\beta \sqrt{I\left(1 + \frac{J-1}{\tau}\right)}, \tag{6.65}$$

Y is a beta random variable with parameters $n_1/2$ and $n_2/2$, and b is defined in (6.15). The parameter τ can be estimated by the sample variance ratio (6.32):

$$Q = \frac{S_1^2}{S_2^2} = \tau F_{n_1,n_2} = \frac{\tau n_2 Y}{n_1(1-Y)}, \tag{6.66}$$

where we use F_{n_1,n_2} to denote a random variable having an F distribution with n_1 and n_2 degrees of freedom.

If we seek a tolerance limit factor of the form

$$c(S_1^2, S_2^2) = v(Q), \tag{6.67}$$

the remark at the end of Section 6.4 indicates that we seek a solution $v(Q)$ of the integral equation

$$V_\tau(v) \equiv E\left[T_{n_1+n_2}\left(v(Q)(n_1+n_2)^{1/2}\sqrt{\frac{YI}{I-1}+\frac{1-Y}{\tau}}, \delta(\tau)\right)\right] = \gamma, \tag{6.68}$$

where the expectation is with respect to the beta density of Y.

In Section 6.5, we derived two approximations to $v(Q)$, either of which we label here $v^0(Q)$. We will improve on this approximation by using the Conditional Expectation algorithm. Let the approximation at the nth iteration be denoted $v^n(Q)$ and define the iteration

$$v^{n+1} \equiv v^n + \psi^n, \tag{6.69}$$

where the quasi-Newton step ψ^n is an approximation to the solution $\tilde{\psi}^n$ of the Newton-step equation

$$\gamma - V_\tau(v^n) =$$

$$E\left[\tilde{\psi}^n(Q)(n_1+n_2)^{1/2}\sqrt{\frac{YI}{I-1}+\frac{1-Y}{\tau}}\right.$$

$$\left.\cdot T'_{n_1+n_2}\left(v^n(Q)(n_1+n_2)^{1/2}\sqrt{\frac{YI}{I-1}+\frac{1-Y}{\tau}}, \delta\right)\right], \tag{6.70}$$

where $T'_{n_1+n_2}(\cdot,\cdot)$ denotes the noncentral t density and $V_\tau(\cdot)$ is given in (6.68).

The noncentral t density with f degrees of freedom and noncentrality parameter δ can be calculated by means of the following formula (Odeh and Owen, 1980, p. 272):

$$T'_f(x,\delta) = \frac{f}{x}\left[T_{f+2}\left(x\sqrt{\frac{f+2}{f}},\delta\right) - T_f(x,\delta)\right]. \tag{6.71}$$

Since there are computer subroutines available for determining the noncentral t cdf (see, e.g., Griffiths and Hill, 1985), (6.71) is very useful for computation.

Using the shorthand notation of Chapter 4, we write (6.68) as

$$\int_0^1 k\{\tau,y,v^n[\phi(\tau,y)]\}dy = g(\tau) \equiv \gamma, \tag{6.72}$$

where

$$\phi(\tau,y) \equiv \frac{\tau n_2 y}{n_1(1-y)}. \tag{6.73}$$

We also rewrite the Newton-step equation (6.70) as

$$\gamma - \int_0^1 k\{\tau,y,v^n[\phi(\tau,y)]\}dy = \int_0^1 k'\{\tau,y,v^n[\phi(\tau,y)]\}\tilde{\psi}^n[\phi(\tau,y)]dy. \tag{6.74}$$

The kernel of the integral equation (6.72) and the derivative of this kernel with respect to its third argument are given by

$$\begin{aligned} k\{\tau,y,v^n[\phi(\tau,y)]\} \equiv\ & \text{Beta}\,(y;n_1/2,n_2/2) \\ & \cdot T_{n_1+n_2}\left(v^n[\phi(\tau,y)](n_1+n_2)^{1/2}\sqrt{\frac{yI}{I-1}+\frac{1-y}{\tau}},\delta(\tau)\right) \end{aligned} \tag{6.75}$$

and

$$\begin{aligned} k'\{\tau,y,v^n[\phi(\tau,y)]\} \equiv\ & \text{Beta}\,(y;n_1/2,n_2/2) \\ & \cdot(n_1+n_2)^{1/2}\sqrt{\frac{yI}{I-1}+\frac{1-y}{\tau}}\,T'_{n_1+n_2}\left(v^n[\phi(\tau,y)](n_1+n_2)^{1/2}\sqrt{\frac{yI}{I-1}+\frac{1-y}{\tau}},\delta\right)\Bigg], \end{aligned} \tag{6.76}$$

respectively, where

$$\text{Beta}\,(y;n_1/2,n_2/2) \equiv \frac{\Gamma\left[(n_1+n_2)/2\right]}{\Gamma\left(n_1/2\right)\Gamma\left(n_2/2\right)}y^{n_1/2-1}(1-y)^{n_2/2-1}. \tag{6.77}$$

For any fixed τ, we can numerically determine the location $y_*(\tau)$ of the peak of the kernel, and define

$$q_*(\tau) \equiv \frac{\tau n_2 y_*(\tau)}{n_1(1-y_*(\tau))}. \tag{6.78}$$

We propose doing this for many values of τ. Inspection of the kernel shows that the peaks fall on a nearly straight ridge. Qualitatively, the contours of this kernel look much like Figure 5.1 of Chapter 5. We make the approximation

$$\tilde{\psi}^n[\phi(\tau,y)] \approx \psi^n[\phi(\tau,y_*)] = \psi^n[q_*(\tau)], \tag{6.79}$$

and we note that $\psi^n[q_*(\tau)]$ is not a function of y and so can be removed from the integrand. We have the following approximate Newton step

$$\psi^n[q_*(\tau)] = \frac{\gamma - \int_0^1 k\{\tau, y, v^n[\phi(\tau, y)]\}dy}{\int_0^1 k'\{\tau, y, v^n[\phi(\tau, y)]\}dy}, \tag{6.80}$$

which is in the form of a Conditional Expectation step with a single step for the inner iteration, and with the initial iterate for each inner iteration identically zero; that is, we have applied the Conditional Expectation algorithm in the form (4.16).

It is fortunate that in our tolerance limit problem, the function $y_*(\tau)$ is nearly independent of τ. Thus, $\psi^n[\phi(\tau, y)]$ can be evaluated at or very nearly at a specified grid of q_* values by adjusting τ after the nearly constant value $\hat{y}_*$ of $y_*(\tau)$ is approximated for a typical τ value.

One difficulty with the above proposal arises from the fact that, strictly speaking, τ should only be taken to be greater than one, in which case the range of q_* values is from $n_2\hat{y}_*/[n_1(1-\hat{y}_*)]$ to ∞ instead of from 0 to ∞ as is required for the numerical integration. Since $n_2\hat{y}_*/[n_1(1-\hat{y}_*)]$ turns out to be relatively small we translate the value of q_* by this amount, so that the range of q values will be 0 to ∞. In other words, we replace $v^n(q_*)$ in the approximation

$$\gamma \approx \int_0^1 k\{\tau, y, v^n(q_*)\}dy + \int_0^1 k'\{\tau, y, v^n(q_*)\}\psi^n[\phi(\tau, y)]dy. \tag{6.81}$$

by $v^n\{q_* - n_2\hat{y}_*/[n_1(1-\hat{y}_*)]\}$. After this approximation is carried out, the method can be iterated using

$$v^{n+1}\left[q_* - \frac{n_2\hat{y}_*}{n_1(1-\hat{y}_*)}\right] = v^n\left[q_* - \frac{n_2\hat{y}_*}{n_1(1-\hat{y}_*)}\right] + \psi(q_*) \tag{6.82}$$

to replace v^n.

With each iteration the value of the constant $\hat{y}_*$ is likely to change and should be recalculated.

The above simple improvement of the approximation underlying the Conditional Expectation approach enables one to calculate tolerance limit factors which provide very nearly the nominal confidence even for few batches and small batch size. Tolerance limit factors determined by means of the Conditional Expectation algorithm will be referred to as *Conditional Expectation tolerance limits.*

The simple Conditional Expectation iteration outlined in this section is easily implemented, and works astonishingly well for this difficult (unsolvable?) nonlinear problem. Ten or twenty iterations will usually provide a *smooth* tolerance limit factor which provides almost exactly the nominal size for all values of the nuisance parameter.

In fact, the calculations are simple enough to be performed interactively, and functions written in *S* for doing this are provided in Appendix D.

6.6.1 Polynomial Approximations to the Integral Equation Solutions

The Conditional Expectation tolerance limit factors are, for many situations, well approximated by polynomials in W, where W is defined in (6.60). For the combinations of β, γ, I, and J most important for aircraft design allowable applications, the cubic polynomial

$$\hat{v} = a + bW + cW^2 + dW^3 \tag{6.83}$$

was fit, by least squares, to the approximate numerical solutions to (6.68). Since the numerical method of this section is not useful for the case of $I = 2$, we only consider $I > 2$. The approximate tolerance limit factor $\hat{v}$, obtained using the coefficients in Tables 6.2 and 6.3, provides very nearly the nominal confidence for all values of r.

6.7 The Distributions of the Tolerance Limits

Once the function v of Section 6.6 has been determined it is straightforward to calculate the cumulative distribution function of the corresponding tolerance limit. It is obviously preferable to compare distributions of confidence bounds rather than merely confidence levels, and we make such a comparison in this section.

Using the notation of Section (6.6), the tolerance limit cdf is a function $H(t)$ given by

$$H(t;\beta,\tau) \equiv P(\bar{X} - v(Q)S \leq \mu - t\sigma). \tag{6.84}$$

For given $v(Q)$, we would like $H(t;\beta,\tau)$ to be less than γ for $t < \mu - z_\beta\sigma$, greater than γ for $t > \mu - z_\beta\sigma$, and equal to γ for $t = \mu - z_\beta\sigma$. This cdf does not depend on μ, and it depends on σ_b^2 and σ_e^2 only through τ. For our procedure β is fixed, so we let $H(t;\tau) = H(t;\tau,\beta)$ and see how well we do compared to the ideal case of known r. Since this is just the function $V_\tau(v^n)$ of (6.68) with v^n replaced by v and z_β replaced by t, we are able to examine the entire distribution of the tolerance limit with little more effort than is required to calculate the tolerance limit factor.

In Figure 6.2, the cumulative distributions for (.90, .95) Conditional Expectation lower tolerance limits with $I = J = 5$ are presented for various values of the intraclass correlation $\rho \equiv r/(r+1)$.

Note that all of the curves pass very nearly through $(x_\beta, .95)$, where $x_\beta \equiv \mu - z_\beta\sigma_X$, indicating the striking success that we have had at removing the nuisance parameter, even for as few as five batches. As the intraclass correlation is increased the random effects sample goes from behaving essentially like a single sample of size $n = IJ$ when $\rho = 0$ to being equivalent to a single batch of size I when $\rho = 1$.

In Figure 6.3 three cdfs are plotted, corresponding to the Mee-Owen method, the Conditional Expectation tolerance limit and the solution for known $r = \rho = 0$. The intraclass correlation is taken to equal zero and the sample size is again $I = J = 5$. Note that the Conditional Expectation tolerance limit is clearly preferable to the Mee-Owen solution and doesn't fare too badly when compared to the known-r solution.

6.8 Discussion

The situation of primary interest to the aircraft industry, (.90, .95) lower tolerance limits, is used here for illustration. The methods presented in this chapter include a *Modified Asymptotic Expansion* tolerance limit based on the Welch-Aspin expansion (6.62), and the *Conditional Expectation* tolerance limit based on the numerical solution of an integral equation (Section 6.6). The confidence for these two methods and for the Mee-Owen method as a function of the intraclass correlation is presented in Figure 6.4 for five batches each of size five.

The various proposed tolerance limit factors, along with the factor of Mee and Owen (1983), are displayed in Figure 6.5. The Mee-Owen tolerance limit factor is discontinuous

Table 6.2: *Coefficients of $\hat{v}$ for (.90, .95) Lower Tolerance Limits*

Sample Size		Coefficients			
I	*J*	*a*	*b*	*c*	*d*
3	2	1.783	8.360	-10.762	6.773
3	3	1.355	2.839	2.725	-0.763
3	4	1.369	1.499	5.960	-2.672
3	5	1.403	1.051	6.880	-3.179
3	6	1.444	0.843	7.118	-3.250
3	7	1.450	0.925	6.843	-3.063
3	8	1.442	0.995	6.714	-2.995
3	9	1.443	0.981	6.748	-3.016
3	10	1.426	1.195	6.275	-2.741
3	∞	1.255	1.960	5.233	-2.293
4	2	1.820	-1.036	5.548	-2.170
4	3	1.604	-0.389	4.887	-1.940
4	4	1.559	-0.286	4.848	-1.960
4	5	1.550	-0.307	4.946	-2.028
4	6	1.542	-0.305	4.986	-2.061
4	7	1.531	-0.275	4.964	-2.059
4	8	1.520	-0.241	4.934	-2.051
4	9	1.508	-0.190	4.868	-2.024
4	10	1.484	-0.077	4.702	-1.947
4	∞	1.281	0.940	3.148	-1.208
5	2	1.860	-1.878	5.814	-2.389
5	3	1.710	-1.042	4.462	-1.723
5	4	1.635	-0.743	4.074	-1.559
5	5	1.598	-0.638	3.984	-1.537
5	6	1.574	-0.575	3.939	-1.531
5	7	1.555	-0.516	3.884	-1.516
5	8	1.539	-0.464	3.833	-1.501
5	9	1.525	-0.419	3.786	-1.485
5	10	1.502	-0.317	3.645	-1.423
5	∞	1.286	0.707	2.125	-0.712
6	2	1.861	-2.064	5.431	-2.222
6	3	1.721	-1.024	3.607	-1.298
6	4	1.644	-0.747	3.271	-1.162
6	5	1.604	-0.654	3.219	-1.163
6	6	1.577	-0.591	3.186	-1.165
6	7	1.555	-0.533	3.145	-1.160
6	8	1.537	-0.486	3.107	-1.152
6	9	1.523	-0.447	3.077	-1.147
6	10	1.507	-0.387	3.005	-1.119
6	∞	1.287	0.613	1.564	-0.457

Table 6.2: *Coefficients of $\hat{v}$ for (.90, .95) Lower Tolerance Limits*

Sample Size		Coefficients			
I	*J*	*a*	*b*	*c*	*d*
7	2	1.845	-1.974	4.808	-1.924
7	3	1.711	-0.911	2.926	-0.970
7	4	1.637	-0.682	2.682	-0.881
7	5	1.598	-0.609	2.671	-0.905
7	6	1.569	-0.553	2.660	-0.920
7	7	1.547	-0.504	2.637	-0.924
	8	1.530	-0.465	2.622	-0.931
7	9	1.515	-0.426	2.589	-0.923
7	10	1.498	-0.358	2.529	-0.914
7	∞	1.287	0.558	1.222	-0.311
8	2	1.749	-1.136	2.979	-1.010
8	3	1.660	-0.668	2.260	-0.670
8	4	1.607	-0.554	2.197	-0.668
8	5	1.578	-0.530	2.254	-0.721
8	6	1.555	-0.501	2.281	-0.753
8	7	1.536	-0.466	2.279	-0.767
8	8	1.520	-0.431	2.263	-0.771
8	9	1.506	-0.394	2.236	-0.766
8	10	1.485	-0.315	2.138	-0.727
8	∞	1.286	0.520	0.996	-0.220
9	2	1.740	-1.068	2.640	-0.859
9	3	1.651	-0.611	1.943	-0.529
9	4	1.599	-0.511	1.905	-0.539
9	5	1.569	-0.490	1.970	-0.596
9	6	1.546	-0.463	2.001	-0.631
9	7	1.527	-0.429	2.003	-0.647
9	8	1.510	-0.395	1.990	-0.652
9	9	1.494	-0.347	1.949	-0.642
9	10	1.480	-0.308	1.912	-0.631
9	∞	1.286	0.490	0.837	-0.159
10	2	1.730	-0.992	2.343	-0.727
10	3	1.640	-0.556	1.689	-0.418
10	4	1.590	-0.471	1.676	-0.440
10	5	1.560	-0.452	1.748	-0.501
10	6	1.536	-0.426	1.782	-0.537
10	7	1.517	-0.395	1.789	-0.556
10	8	1.501	-0.363	1.779	-0.563
10	9	1.486	-0.322	1.749	-0.557
10	10	1.475	-0.301	1.740	-0.560
10	∞	1.285	0.466	0.720	-0.117

Table 6.3: *Coefficients of $\hat{v}$ for (.99, .95) Lower Tolerance Limits*

Sample Size		Coefficients			
I	*J*	*a*	*b*	*c*	*d*
3	2	3.105	4.815	2.357	0.276
3	3	2.554	2.311	9.725	-4.038
3	4	2.543	2.021	10.472	-4.484
3	5	2.552	1.857	10.843	-4.699
3	6	2.558	1.743	11.104	-4.852
3	7	2.555	1.719	11.184	-4.904
3	8	2.550	1.717	11.214	-4.928
3	9	2.501	1.948	10.883	-4.778
3	10	2.468	2.480	9.619	-4.014
3	∞	2.269	3.024	9.363	-4.104
4	2	2.933	-0.544	7.263	-2.610
4	3	2.608	0.125	6.989	-2.680
4	4	2.613	-0.082	7.464	-2.952
4	5	2.648	-0.362	7.984	-3.228
4	6	2.671	-0.543	8.324	-3.410
4	7	2.681	-0.646	8.529	-3.523
4	8	2.622	-0.391	8.201	-3.390
4	9	2.622	-0.074	7.463	-2.970
4	10	2.554	-0.062	7.712	-3.162
4	∞	2.310	1.071	6.064	-2.403
5	2	2.919	-1.441	6.702	-2.439
5	3	2.608	-0.129	4.949	-1.687
5	4	2.615	-0.310	5.339	-1.902
5	5	2.655	-0.617	5.890	-2.187
5	6	2.682	-0.830	6.283	-2.393
5	7	2.696	-0.961	6.535	-2.529
5	8	2.673	-0.895	6.477	-2.514
5	9	2.606	-0.314	5.540	-2.090
5	10	2.542	0.192	4.248	-1.241
5	∞	2.323	0.634	4.351	-1.567
6	2	2.905	-1.421	5.433	-1.856
6	3	2.587	0.036	3.356	-0.917
6	4	2.601	-0.198	3.827	-1.168
6	5	2.637	-0.505	4.387	-1.457
6	6	2.662	-0.723	4.794	-1.671
6	7	2.676	-0.868	5.075	-1.822
6	8	2.556	-0.075	3.805	-1.224
6	9	2.592	-0.350	4.271	-1.451
6	10	2.616	-0.695	4.918	-1.777
6	∞	2.328	0.484	3.356	-1.106

Table 6.3: *Coefficients of $\hat{v}$ for (.99, .95) Lower Tolerance Limits*

Sample Size		Coefficients			
I	*J*	*a*	*b*	*c*	*d*
7	2	2.833	-0.830	3.638	-1.000
7	3	2.568	0.265	2.138	-0.330
7	4	2.588	-0.042	2.743	-0.647
7	5	2.616	-0.336	3.300	-0.939
7	6	2.635	-0.542	3.701	-1.152
7	7	2.647	-0.686	3.989	-1.307
7	8	2.575	-0.041	2.770	-0.663
7	9	2.575	-0.421	3.664	-1.177
7	10	2.562	-0.265	3.317	-0.972
7	∞	2.330	0.416	2.720	-0.824
8	2	2.728	-0.005	1.697	-0.067
8	3	2.557	0.465	1.219	0.113
8	4	2.576	0.103	1.941	-0.267
8	5	2.595	-0.174	2.493	-0.561
8	6	2.608	-0.365	2.881	-0.770
8	7	2.617	-0.503	3.165	-0.925
8	8	2.585	-0.380	3.006	-0.857
8	9	2.624	-0.671	3.528	-1.127
8	10	2.506	0.535	0.822	0.492
8	∞	2.330	0.378	2.284	-0.638
9	2	2.642	0.758	-0.035	0.778
9	3	2.551	0.620	0.527	0.446
9	4	2.566	0.224	1.338	0.014
9	5	2.578	-0.038	1.883	-0.280
9	6	2.585	-0.213	2.256	-0.484
9	7	2.573	-0.264	2.416	-0.583
9	8	2.592	-0.435	2.737	-0.751
9	9	2.558	-0.218	2.390	-0.588
9	10	2.546	-0.323	2.664	-0.744
9	∞	2.330	0.354	1.968	-0.509
10	2	2.593	1.357	-1.453	1.485
10	3	2.549	0.735	0.001	0.697
10	4	2.558	0.320	0.878	0.225
10	5	2.562	0.075	1.412	-0.068
10	6	2.563	-0.085	1.769	-0.266
10	7	2.549	-0.073	1.792	-0.287
10	8	2.545	-0.168	2.031	-0.427
10	9	2.544	-0.171	2.090	-0.482
10	10	2.516	-0.158	2.123	-0.500
10	∞	2.330	0.336	1.730	-0.415

because these authors recommend pooling the data if $Q < 1$. For the most part, the differences in the tolerance limit factors are not large.

The integral equation approach virtually removes the nuisance parameter from the problem. The Mee-Owen method has the disadvantage of being substantially conservative when the variance ratio is small.

From the rescaled plot of the coverage probability function for the integral equation solution (Figure 6.6) it can be seen that for $r > 1$ the actual coverage probability differs from .95 by no more than ±.001. This small difference can be attributed to the limited accuracy of the numerical integration. For $r < 1$, however, the difference in the actual and nominal coverage probability increases substantially, but never does it reach a magnitude that warrants concern for applications.

Figure 6.7 illustrates the convergence of the Conditional Expectation algorithm for various values of the intraclass correlation. Note that for practical purposes ten iterations is adequate, although some slight improvement can result from considering more iterations.

6.9 Examples

We consider two examples in this section. The first example is a situation where there is considerable between-batch variability, and the second is a case where the true between-batch variance is zero, since the 'batches' are artificially constructed from a simple random sample.

A manufacturer of aircraft components always performs certain mechanical tests on specimens from each batch of composite material. The data in Table 6.4 are coded tensile strength measurements made on five consecutive batches (R. Zabora, personal communication, 1988). The results of an analysis using the Mee-Owen method and the methods of this chapter are also presented in Table 6.4.

All of the tolerance limit methods give nearly the same answer. These three methods will always agree in the limit of large between-batch variability.

To see how much these methods differ when the between-batch variability is minimal, we begin with a simple random sample of 180 composite tensile strength measurements (Reese and Sorem, 1981). The normal distribution fits these data reasonably well, especially in the tails, so we proceed to choose 25 specimens at random (with replacement) from this set and to divide these into five 'batches' of size five. These data are given, along with tolerance limit calculations, in Table 6.5. Note the difference between the Conditional Expectation solution and the other results. Although this difference is a fraction of a standard deviation, it might be large enough to be of engineering importance for some applications.

Since the 'batches' in this second example were artificially created, it is interesting to compare the above random effects tolerance limits with the pooled sample tolerance limit: $209.93 - 1.838(18.38) = 176.15$.

Table 6.4: *Example # 1 : Coded Strength Measurements From Five Batches*

Batch	Coded Strength Measurements				
1	379	357	390	376	376
2	363	367	382	381	359
3	401	402	407	402	396
4	402	387	392	395	394
5	415	405	396	390	395

$$\bar{X} = 388.36 \quad S_1^2 = 1040.84 \quad S_2^2 = 78.92 \quad \hat{\sigma}_X^2 = 271.30$$

$$k_{mo} = 3.072 \quad \bar{X} - k_{mo}S = 337.76$$
$$k_{tw} = 3.063 \quad \bar{X} - k_{ce}S = 337.90$$
$$k_{mae} = 3.055 \quad \bar{X} - k_{mae}S = 338.04$$

NOTE: $\beta = .9$, $\gamma = .95$. The subscripts *mo, ce*, and *mae* denote the Mee-Owen, Conditional Expectation (6.83), and Modified Asymptotic Expansion (6.62) tolerance limit factors, respectively.

Table 6.5: *Example # 2 : Artificially Batched Data From a Simple Random Sample*

'Batch'	Tensile Strength in 1000 psi				
1	203.41	209.58	213.35	218.56	242.76
2	185.97	190.67	207.88	210.80	231.46
3	184.41	200.73	206.51	209.84	212.15
4	160.44	180.95	201.95	204.60	219.51
5	174.63	185.34	205.59	212.00	225.25

$$\bar{X} = 203.93 \quad S_1^2 = 386.04 \quad S_2^2 = 325.56 \quad \hat{\sigma}_X^2 = 337.65$$

$$k_{mo} = 2.12 \quad \bar{X} - k_{mo}S = 164.98$$
$$k_{ce} = 2.04 \quad \bar{X} - k_{ce}S = 166.45$$
$$k_{mae} = 1.93 \quad \bar{X} - k_{mae}S = 168.47$$

NOTE: $\beta = .9$, $\gamma = .95$. The subscripts *mo, ce*, and *mae* denote the Mee-Owen, Conditional Expectation (6.83), and Modified Asymptotic Expansion (6.62) tolerance limit factors, respectively.

Figure 6.1: Confidence for Asymptotic Expansion Tolerance Limits

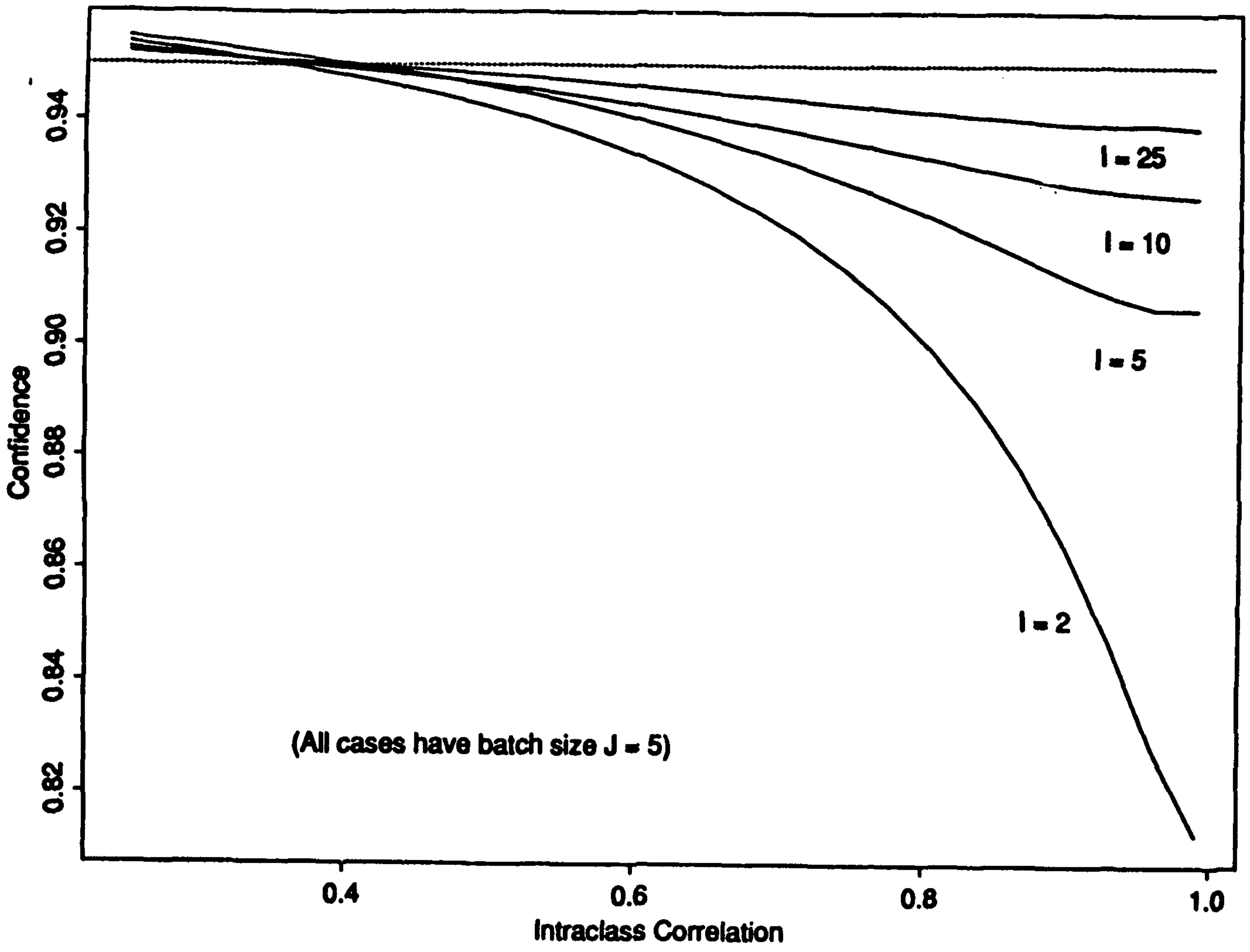


Figure 6.2: Distributions of (.90,.95) Conditional Expectation Tolerance Limits.

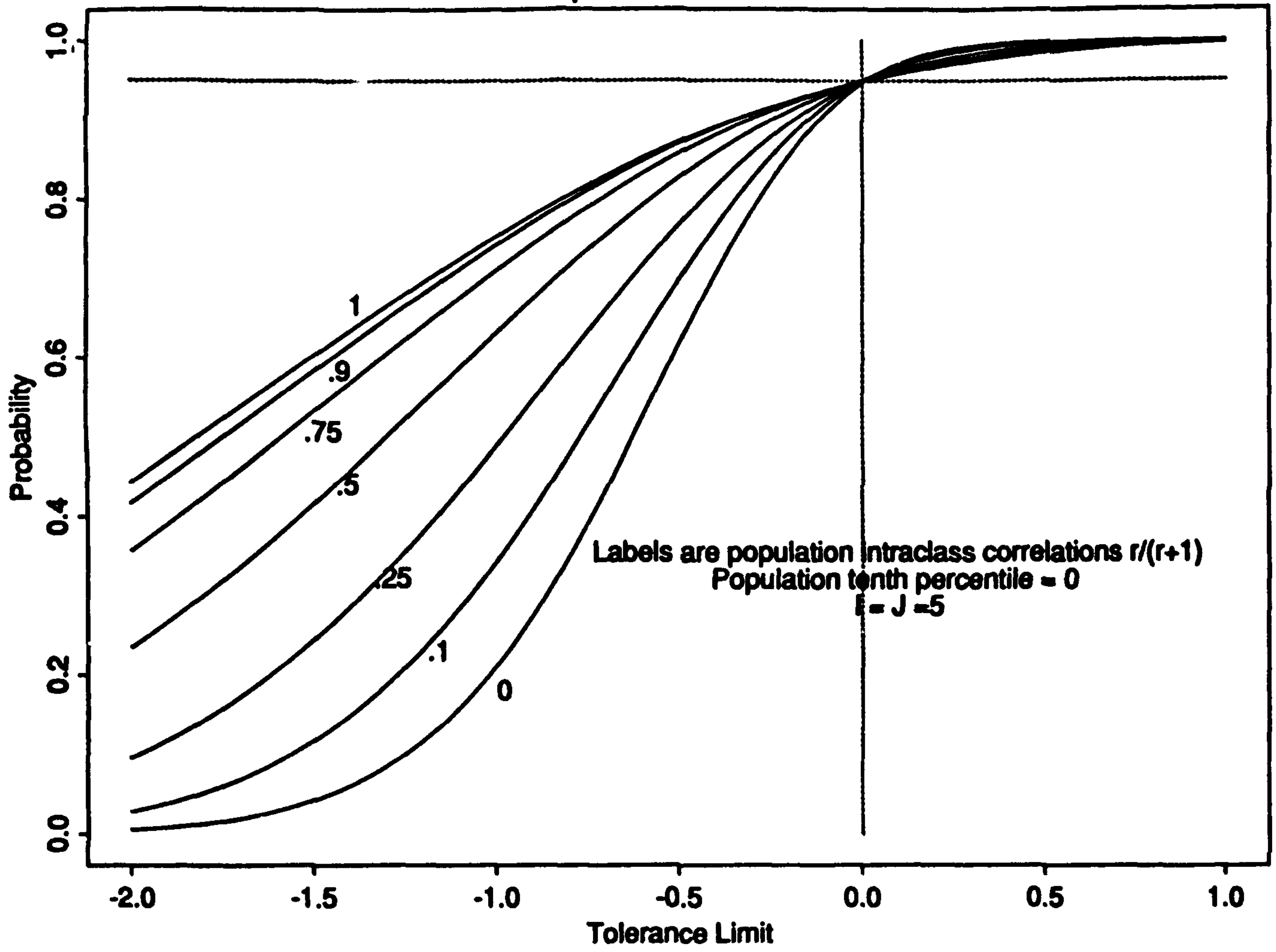


Figure 6.3: Distributions of Tolerance Limits for r = 0.

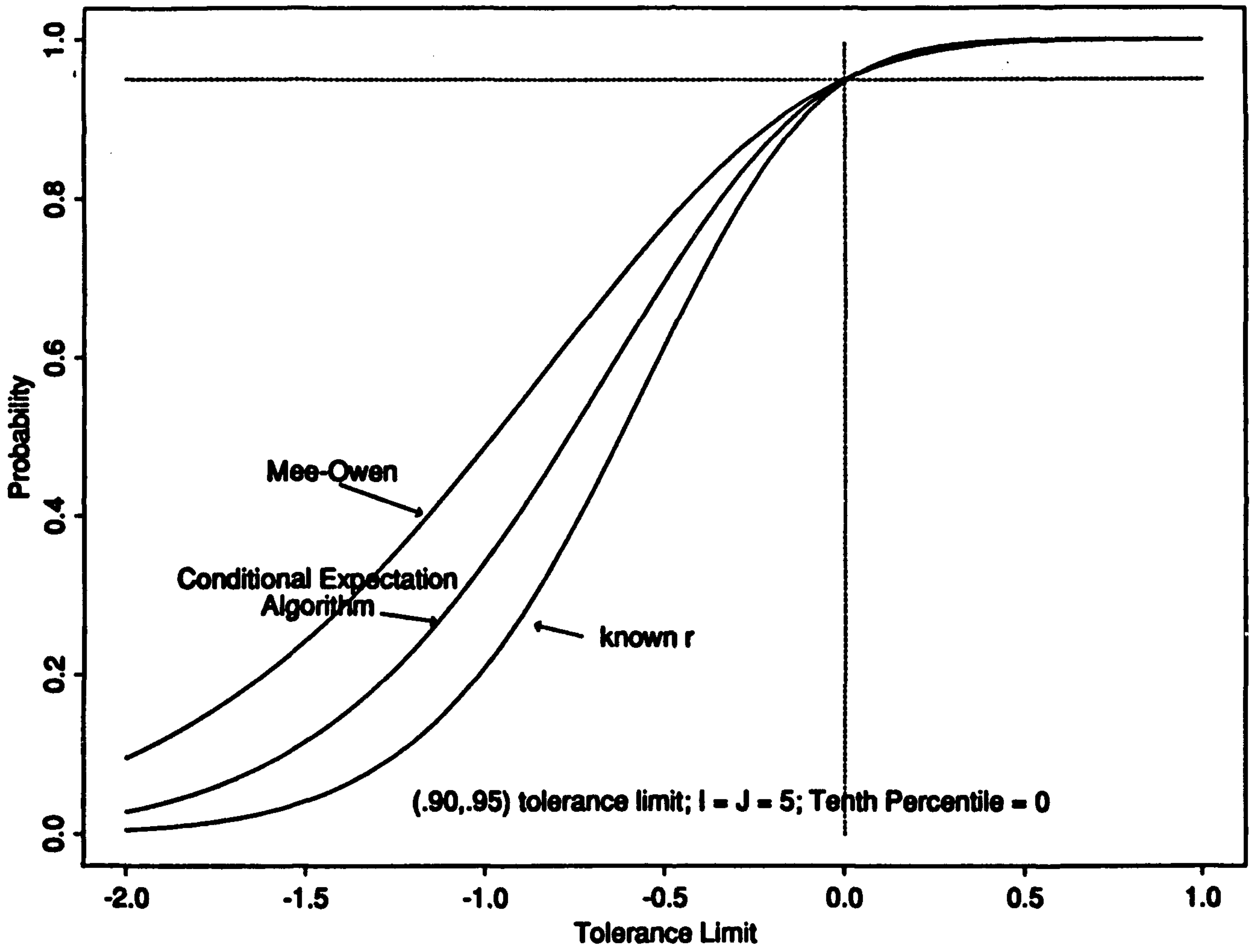


Figure 6.4: Comparison of Confidence as Functions of the Population Intraclass Correlation.

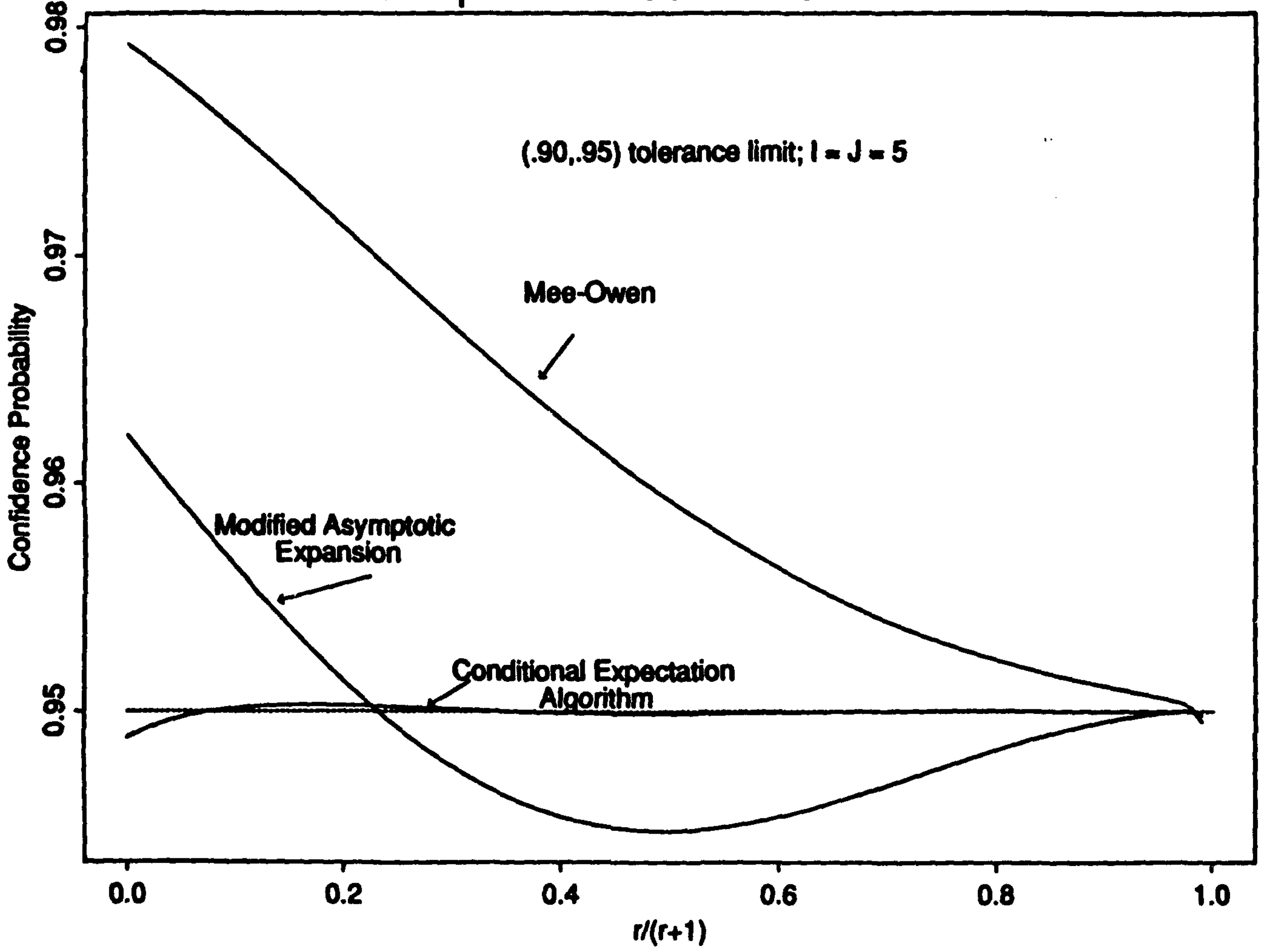


Figure 6.5: A Comparison of Various Tolerance Limit Factors

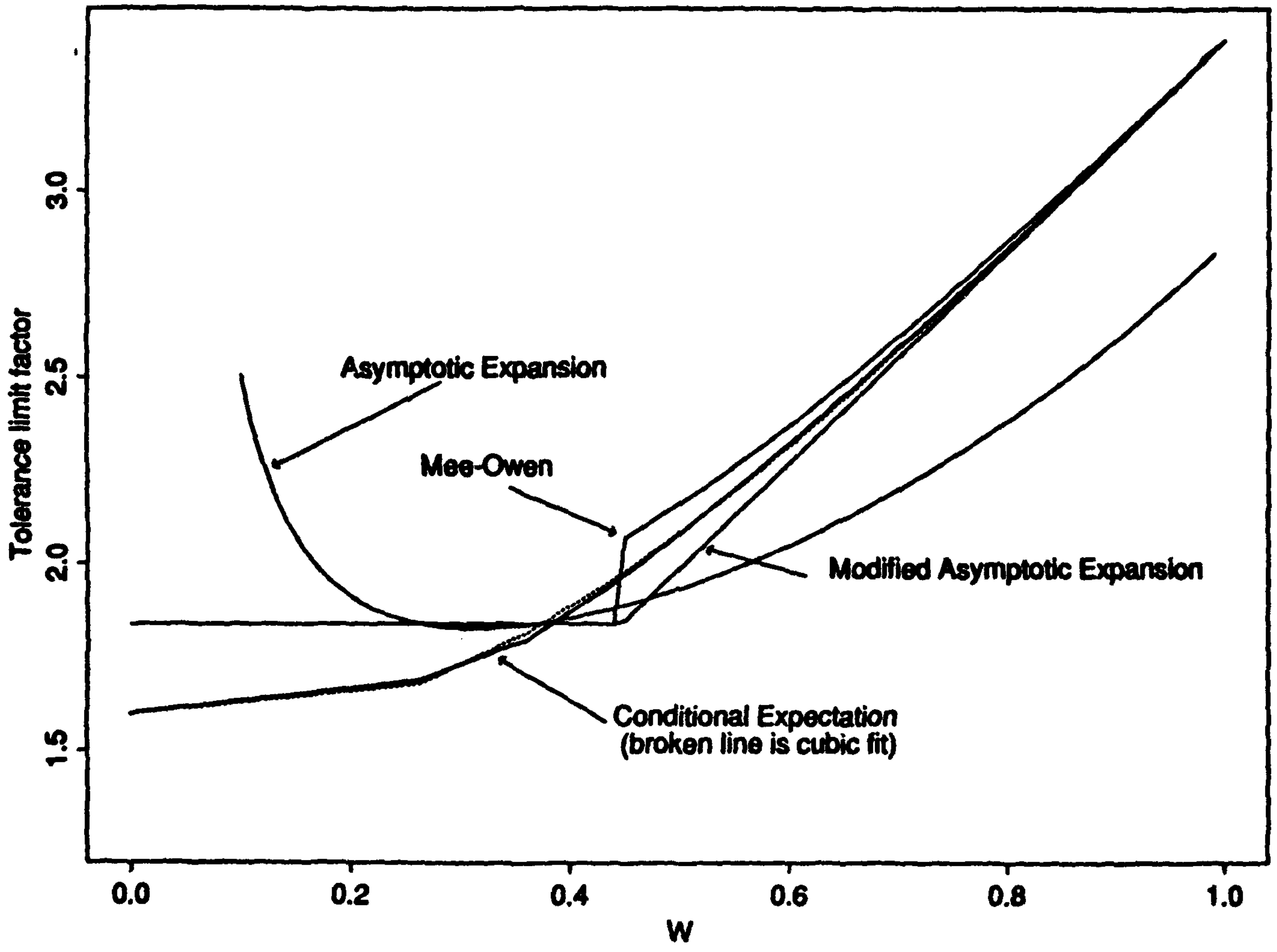


Figure 6.6: Coverage Probability for Conditional Expectation Algorithm Tolerance Limit as a Function of the Intraclass Correlation.

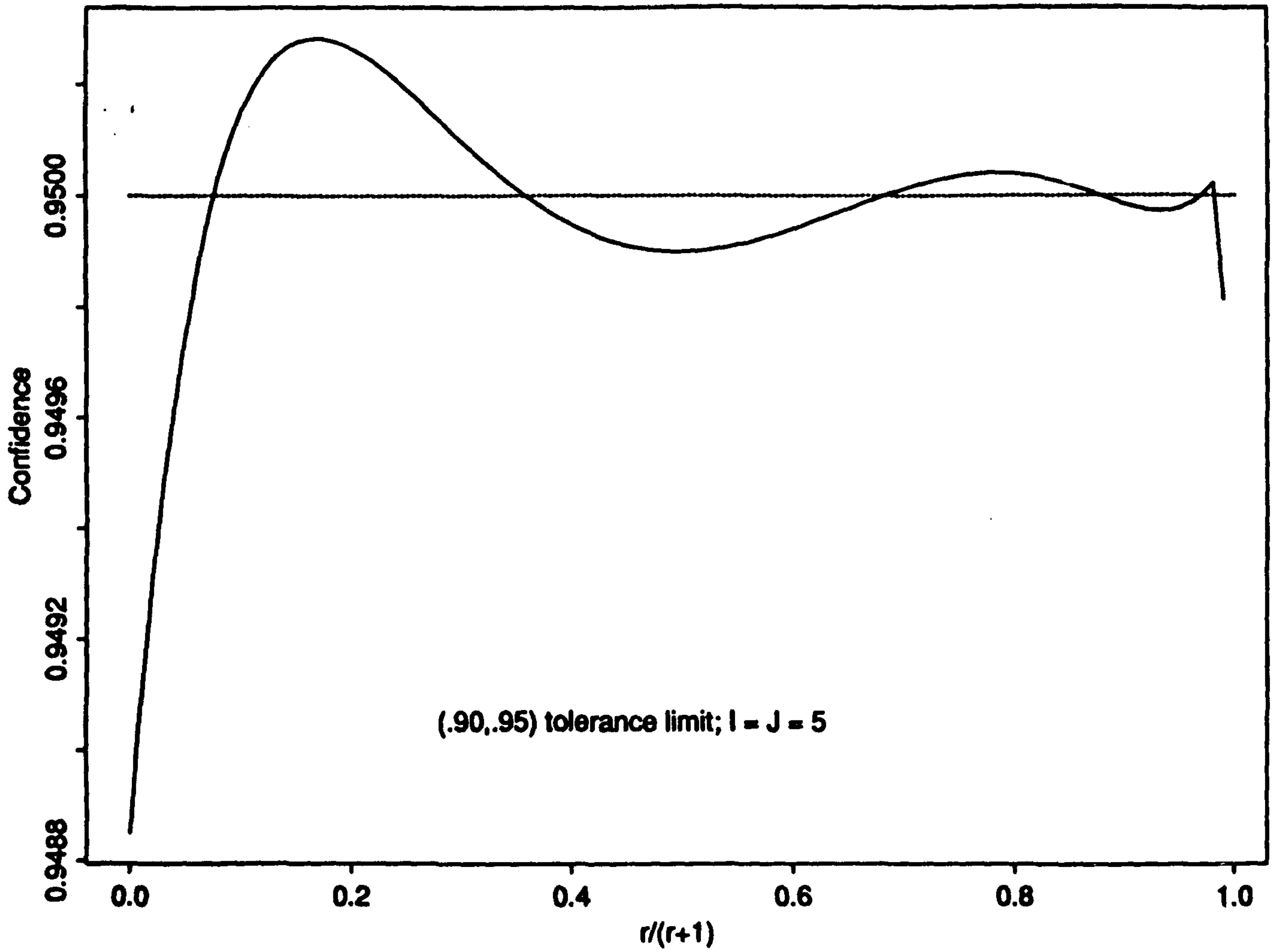
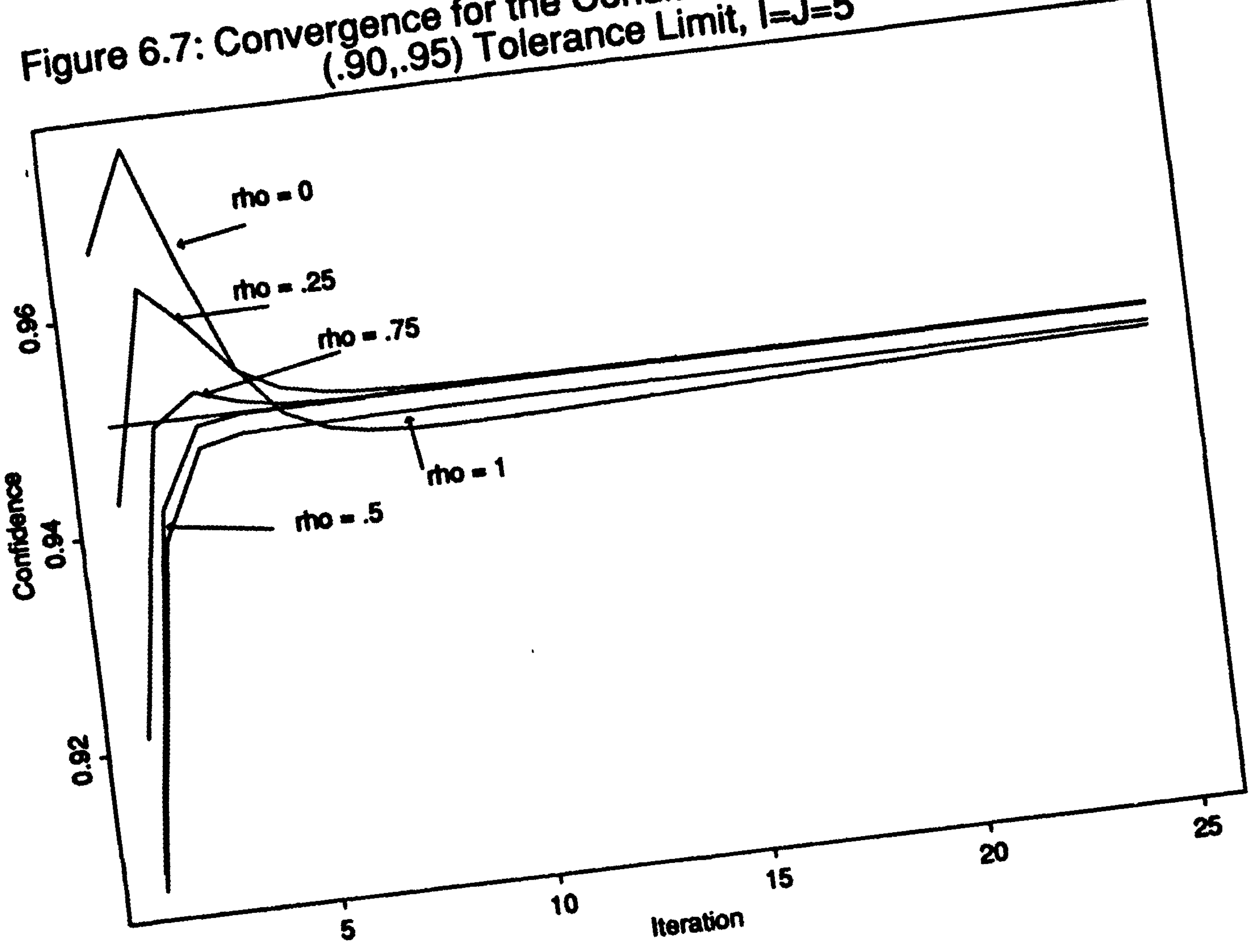


Figure 6.7: Convergence for the Conditional Expectation Algorithm (.90,.95) Tolerance Limit, I=J=5



Chapter 7

An Ill-Posed Inverse Problem in Stereology

7.1 Ill-Posed Inverse Problems in Applied Science

This thesis has been concerned thus far with describing and applying the Conditional Expectation algorithm to the solution of integral equations of the first kind, where the known functions are given without error. Problems of this sort are examples of ill-posed inverse problems. The study of ill-posed inverse problems in applications, where the right hand side is observed with error, or where the right hand size is an estimate of a probability density, is receiving increasing attention in statistics (O'Sullivan, 1986). There are many examples of inverse problems, in such diverse areas as geophysics, tomography, water resource management, and stereology.

Problems involving integral equations of the first kind generally arise when a quantity is indirectly observed. For a linear problem, we have

$$\int_0^1 k(x,y)f(y)dy = g(x), \tag{7.1}$$

where $g(x)$ is a function, observed with error, which acts as a proxy for the unobservable $f(y)$. The kernel, $k(x,y)$, relates the observable, g, to the quantity of interest, f. The kernel k, which we will always assume to be known, is often a model for the response of a measuring instrument.

The Conditional Expectation algorithm is rapidly convergent for a fairly wide class of problems and produces smooth near-solutions. Because this algorithm produces smooth near-solutions (see Section 1.2.3), it may be useful for certain inverse problems in applied science as well. We consider next one such problem, the classical random sphere problem of stereology.

7.2 The Random Sphere Problem

Often investigators in medicine, materials science, and astronomy, among other fields, are faced with the following situation. Observations are made on a two-phase material where the first phase consists of spheres of random radius, and these spheres are randomly distributed in a second phase. Examples include stars in a globular cluster (Wicksell, 1926), tumor cell nuclei in a mouse liver (Keiding et. al., 1972), and air bubbles in polystyrene (Meisner, 1967). The distribution of the radii of the spheres is desired, but data are available only on the radii either of circular projections or else of sections of these spheres: for example, circular cross sections of tumors measured from a thin slice of a dissected organ.

This problem was apparently first correctly modeled by Wicksell (1925). A large literature has its origin with this Wicksell article, including a wide variety of solution techniques. The interested reader can begin with the reviews of Anderssen and Jakeman (1974), Jakeman and Anderssen (1974), Cruz-Orive (1983), and Colman (1989). For purposes of practical stereology, the Wicksell problem has been largely solved, but 'its very simple structure makes it a perfect vehicle for testing numerical and statistical procedures' (Coleman, 1989, p. 244) . So this problem is a natural one to consider, and we begin by introducing some of the theory for a class of integral equations to which the random sphere equation belongs.

7.3 Singular Integral Equations of Abel Type

Abel's integral equation, in its simplest form, is

$$\int_0^x \frac{f(y)}{(x-y)^{1/2}} dy = g(x). \tag{7.2}$$

This is a weakly singular Volterra equation. An equation is said to be *singular* if either the kernel is singular, the range of integration is unbounded, or both (Porter and Stirling, 1990, Chapter 9). A *weak* singularity is of the form $(x-y)^{-\alpha}$ for $0 < \alpha < 1$. The Abel equation appears in the solution of the brachistochrone problem with which the calculus of variations began (e.g., Weinstock, 1974, pp. 19, 28-29), and so it is of considerable historical, as well as practical, importance.

To solve the equation (7.2) analytically, we apply the operator

$$Kf \equiv \int_0^x \frac{f(y)}{(x-y)^{1/2}} dy \tag{7.3}$$

to both sides, giving

$$\int_0^x \frac{ds}{(x-s)^{1/2}} \int_0^s \frac{f(t)dt}{(s-t)^{1/2}} = \int_0^x \frac{g(s)ds}{(x-s)^{1/2}}. \tag{7.4}$$

Interchanging the order of integration on the left hand side of (7.4), we have

$$\int_0^x f(t)dt \int_t^x \frac{ds}{(x-s)^{1/2}(s-t)^{1/2}} = \int_0^x \frac{g(s)ds}{(x-s)^{1/2}}. \tag{7.5}$$

The change of variable

$$s = x\sin^2\theta + t\cos^2\theta \tag{7.6}$$

gives

$$\int_t^x \frac{ds}{(x-s)^{1/2}(s-t)^{1/2}} = \pi. \tag{7.7}$$

The solution to (7.2) can now be seen to be

$$f(x) = \frac{1}{\pi}\frac{d}{dx}\int_0^x \frac{g(t)dt}{(x-t)^{1/2}}. \tag{7.8}$$

Various generalizations of (7.2) are possible, the most important being replacing the exponent 1/2 in the denominator of the integrand of (7.2) with any $\alpha \in (0,1)$, for which we have the inversion formula (Porter and Stirling, 1990, p. 293)

$$f(x) = \frac{\sin(\alpha\pi)}{\pi}\frac{d}{dx}\int_0^x \frac{g(t)dt}{(x-t)^{(1-\alpha)}}. \tag{7.9}$$

To use either of the inversion formulas (7.8) or (7.9) numerically, one must perform numerical differentiation. Algorithms for solving Abel integral equations numerically by means of the inversion formula use devices such as spectral differentiation and smoothing to deal with the well-known difficulties inherent in numerical differentiation. Iterative algorithms, on the other hand, exploit the smoothing capability of the kernel itself, and do not require explicit inversion formulas.

7.4 The Wicksell Solution to the Random Sphere Problem

An argument in geometric probability leads to an Abel equation for the random sphere problem. We follow here the presentation of the conditioning argument given by Nychka et. al. (1984).

Consider a single sphere of radius R, where R is a random variable with density $f(r)$. Condition on the radius $R = r$, and let this sphere be cut at random by a plane, i.e. let the distance U from the cutting plane to the center of the sphere be uniform on $[0, r]$, and let the radius of a cross-sectional circle (a *profile radius*) be denoted by the random variable X. Let E denote the event that a sphere cut by the plane has radius r. It is easy to see that

$$\begin{aligned} G_{X|E}(x) &\equiv P(X \le x|E) = P(U \ge \sqrt{r^2-x^2}|E) \\ &= \begin{cases} 1 - \sqrt{r^2-x^2}/r & \text{for } 0 \le x \le r \\ 1 & \text{for } x > r \\ 0 & \text{for } x < 0 \end{cases} . \end{aligned} \tag{7.10}$$

The probability that a sphere will be cut by a given plane depends on its radius, r. Thus the conditional density of the radii of spheres given that they are cut by a specified plane changes from $f(r)$ to $l(r)$, which is proportional to $rf(r)$. To see this, consider an infinite population of spheres intersecting a given plane. Replace each cut sphere by the diameter which is orthogonal to the intersecting plane. The probability density that a cut sphere has radius r is clearly the ratio of the *total length* of diameters of length $2r$ to the total length of all diameters, i.e.

$$l(r) = \frac{2rf(r)}{2\int_0^{r_0} zf(z)dz} = \frac{rf(r)}{\int_0^{r_0} zf(z)dz} \equiv \frac{rf(r)}{\mu}, \tag{7.11}$$

where μ is defined to be mean sphere radius, and r_0 is the largest observable profile radius.

Let F and G denote the cumulative distributions of sphere and profile radii, respectively, and let $I_{\{A\}}(t)$ denote the indicator function (2.66) for the set A. We have

$$\begin{aligned} G(x) &= \frac{1}{\mu}\int_0^{r_0}[1 - I_{\{r \ge x\}}(r)\sqrt{r^2-x^2}/r]rf(r)dr \\ &= 1 - \frac{1}{\mu}\int_x^{r_0}\sqrt{r^2-x^2}f(r)dr. \end{aligned} \tag{7.12}$$

Differentiating both sides of (7.12) with respect to x gives an Abel equation relating the density of profile radii to the density of sphere radii:

$$g(x) = \frac{1}{\mu}\int_x^{r_0} \frac{xf(r)dr}{\sqrt{r^2 - x^2}}. \tag{7.13}$$

Because μ is the mean of a random variable (the sphere radius) having density $f(x)$, the equation (7.13) is actually nonlinear. We will see below, though, that this nonlinearity does not introduce any serious difficulties.

7.5 The Conditional Expectation Algorithm for the Random Sphere Problem

We now apply the Conditional Expectation algorithm (4.16) to (7.13), and consider several numerical examples. First we discuss how to transform (7.13) into an equation for which the limits of integration are constant, and then we define the modified algorithm. We begin the iteration with $f^0 = g(x)$.

Let the mean of the sphere radius density $f^n(x)$ be

$$\mu^n \equiv \int_0^1 xf^n(x)dx. \tag{7.14}$$

We now replace the functional μ in (7.13) with the *constant* μ^n. The result is a linear integral equation, and a Conditional Expectation algorithm step for this equation is

$$\begin{aligned} h^n(x) &\equiv \left[\int_x^{r_0} \frac{xdr}{\sqrt{r^2 - x^2}}\right]^{-1} \left[\mu^n g(x) - \int_x^{r_0} \frac{xf^n(r)dr}{\sqrt{r^2-x^2}}\right] \\ &= \left\{x\left[\log\left(\sqrt{r_0^2 - x^2} + r_0\right) - \log(x)\right]\right\}^{-1} \left[\mu^n g(x) - \int_x^{r_0} \frac{xf^n(r)dr}{\sqrt{r^2-x^2}}\right]. \end{aligned} \tag{7.15}$$

Following the approach in Appendix A for discretizing Volterra equations (A.3), we simplify the computation by changing the variable of integration so that the quadrature points can be chosen independently of x. To do this, we make the change of variable to w where

$$r = (r_0 - x)w + x. \tag{7.16}$$

If the integral of the kernel is denoted

$$q(x) \equiv x\left[\log\left(\sqrt{r_0^2 - x^2} + r_0\right) - \log(x)\right], \tag{7.17}$$

then, after the change of variable (7.16),

$$h^n(x) = \left[\mu^n g(x) - \int_0^{r_0} \frac{x(r_0 - x)f^n((r_0 - x)w + x)dw}{\sqrt{(r_0 - x)^2 w^2 + 2wx(r_0 - x)}}\right] \Big/ q(x). \tag{7.18}$$

We have transformed the Volterra equation (7.13) into a *series* of integral equations corresponding to $\{\mu^n\}$, all with kernel

$$k(x, w) = \frac{x(r_0 - x)}{q(x)\sqrt{(r_0 - x)^2 w^2 + 2wx(r_0 - x)}}, \tag{7.19}$$

for $(x,w) \in [0,1] \times [0,1]$. This kernel has a singularity along the line $\{(x,w)|w = 0\}$. Figure 7.1 is a plot of (7.19). Note that $k(x,w)$ drops off very rapidly with increasing w for each x and that, if h^n is evaluated along the singular line, then

$$h^n((r_0 - x)w + x)|_{w=0} = h^n(x). \tag{7.20}$$

We can, without loss of generality, let $r_0 = 1$, since this is equivalent to choosing a suitable unit of length.

We now outline an approach, based on the Conditional Expectation algorithm, for simultaneously approximating f and μ. Let f^n be a given sphere radius density, which we will regard as an approximation to a solution f to (7.13). Corresponding to this f^n, there is a mean sphere radius μ^n and a profile radius density g^n. We can easily determine the product $\mu^n g^n$:

$$\tilde{g}^n(x) \equiv \mu^n g^n(x) = \int_x^1 \frac{x f^n(r) dr}{\sqrt{r^2 - x^2}} dy. \tag{7.21}$$

We know that $g^n(x)$, a probability density, must integrate to one. The mean radius μ^n is therefore the normalizing constant:

$$\mu^n = \int_0^1 \tilde{g}^n(x) dx = \int_0^1 \left[\int_x^1 \frac{x f^n(r) dr}{\sqrt{r^2 - x^2}} dr \right] dx. \tag{7.22}$$

From f^n we determine, successively, $\tilde{g}^n$, μ^n, g^n, and h^n (where h^n is given by (7.18)). Now we can calculate $\tilde{f}^{n+1} = f^n + h^n$. Assume, for the moment, that $\tilde{f}^{n+1}$ is positive. Since $\tilde{f}^{n+1}$ need not be a probability density, we let

$$f^{n+1}(x) \equiv \frac{\tilde{f}^{n+1}(x)}{\int_0^1 \tilde{f}^{n+1}(y) dy}, \tag{7.23}$$

and continue the iteration.

If $\tilde{f}^{n+1}(x) < 0$ for some x values, the simplest thing to do is to replace $\tilde{f}^{n+1}$ with $\max(\tilde{f}^{n+1}, 0)$ before performing the normalization (7.23). This approach is usually adequate, and it has been followed in the examples of this chapter.

Another approach is to replace (7.13) with the nonlinear equation

$$g(x) = \frac{1}{\mu} \int_x^1 \frac{x e^{f(r)} dr}{\sqrt{r^2 - x^2}}. \tag{7.24}$$

The Newton-step equation corresponding to (7.24) is easy to determine, and the corresponding Conditional Expectation quasi-Newton step is

$$u^n(x) = \frac{\mu^n g(x)}{\int_x^1 \frac{x e^{f^n(r)} dr}{\sqrt{r^2 - x^2}}}, \tag{7.25}$$

which cannot be negative for any x.

The nonlinear iteration based on (7.25) is closely related to the EM algorithm iteration for this problem (Silverman, et. al., 1990), and to an iterative algorithm recently proposed by Vardi (1992). However, the iterates from this nonlinear algorithm tend to be less smooth than those from the Conditional Expectation algorithm, particularly where the denominator in (7.25) is small.

7.5.1 Density Estimation Issues

In practice, we are almost never given a density $g(x)$. Instead, we have profile radius measurements, and the first order of business is to estimate their density. For those situations where raw radius data is available, Taylor (1982) recommends using a (variable bandwidth) Rosenblatt kernel estimator.

If only a histogram of profile radii is available, then this histogram, once normalized, can be used as a piecewise constant estimate of $g(x)$. Alternatively, one can interpolate between the points with abscissas at the midpoints of the histogram intervals and ordinates given by the normalized counts in the corresponding cells.

We discuss in some detail the use of a piecewise constant estimate of g. Let the endpoints of the ith cell of the normalized histogram be $l_i < u_i$, for $i = 1, \ldots, m$, and denote the estimate of $g(x)$ for $x \in [l_i, u_i)$ by $\hat{g}_i$. Then, from (7.13) we have, for each i, that

$$\begin{aligned} \hat{g}_i &\approx \int_{l_i}^{u_i} g(x) = \frac{1}{\mu}\int_{l_i}^{u_i}\int_x^1 \frac{xf(r)dr}{\sqrt{r^2-x^2}}drdx \\ &= \int_x^1 k_i(r)dr, \end{aligned} \tag{7.26}$$

where

$$\hat{k}_i(r) = \begin{cases} 0 & 0 \le r < l_i \\ \sqrt{r^2 - l_i^2} & l_i \le r < u_i \\ \sqrt{r^2 - l_i^2} - \sqrt{r^2 - u_i^2} & u_i \le r < 1 \end{cases}, \tag{7.27}$$

where we have set $r_0 = 1$. The kernel (7.27) is continuous in r but discrete in x. Although this kernel is not singular, it does have a peak where $r = u_i$ and the Conditional Expectation algorithm can still be successfully applied.

7.6 Numerical Examples

We begin by considering a simple example for the solution is known:

$$f_R(r) \equiv f(r) = 6r(1-r), \tag{7.28}$$

where R is the sphere radius, and $\mu = E(R) = 1/2$. The corresponding profile radius density, $g(x)$, is (Anderssen and Jakeman, 1974, p. 136)

$$g(x) = 6\left\{x\sqrt{1-x^2} - x^3\log[x^{-1} + \sqrt{x^{-2}-1}]\right\}. \tag{7.29}$$

We will consider the cases where g is observed with and without error; and for the case where g is noisy, we will examine the effects of smoothing. We will also consider the case where instead of the right hand side being a function observed with error, we are given a random sample of profile radii from the density (7.29). Following this, we will attempt to invert (7.13) for a real data set on cross sections of liver cell nuclei (Keiding, 1972). The computations were programmed in S (Becker, Chambers and Wilks, 1988), and the code is included in Appendix E.

7.6.1 Sphere Radius Density $f(r) = 6r(1-r)$

Let $f(r)$ be given by (7.28) and let $g(x)$ be given by (7.29), where $g(x)$ might be observed with error. We discretize this problem as in Appendix A by evaluating x and r each at 25 Gauss-Legendre quadrature points and then we apply the Conditional Expectation algorithm (4.16) (see the discussion in Section 7.5) for ten iterations.

We consider first the case where g is observed without error (except for computer roundoff and discretization error), and no smoothing is performed. The results of ten iterations of the algorithm for this situation is presented in Figure 7.2a-b. The heavy line in the Figure 7. 2a is the true solution, and the heavy line in Figure 7.2b is the true right hand side. Note that the algorithm converges rapidly to the solution to the problem, and that the computations have apparently not been substantially effected by roundoff error.

Next, we introduce error in g. It turns out that, unless the noise level is low, some smoothing of the profile radii is helpful. Within the S package, it is convenient to use the function '*smooth*', which is an implementation of the '4(3RSR)2H twice' smoother (Velleman and Hoaglin, 1981). Of course, for a real application, careful consideration must be given to the density estimation process. However, it is our intention in this section to demonstrate that the Conditional Expectation algorithm is *potentially* useful for certain inverse problems in applied science, so we will not be concerned much with the details of density estimation.

Let $Z_i, i = 1, \ldots, 25$ be *iid* standard normal random variables. At each value $g(x_i)$ of $g(x)$ in the discretized problem, we introduce *relative* error by the relationship

$$\tilde{g}(x_i) \equiv g(x_i)(1 + \epsilon Z_i). \tag{7.30}$$

We sometimes smooth the $\tilde{g}(x_i)$ by one pass of the '4(3RSR)2H twice' smoother. The Conditional Expectation algorithm is then applied with no further smoothing.

In Figures 7.3 and 7.4 we 'roughen' the profile radius measurements by using equation (7.30) with $\epsilon = .001$ and $\epsilon = .01$ respectively. No smoothing was done, and it is obvious from the results that no smoothing was necessary. Although noise levels within the range considered here may seem small (in fact, the perturbed $g(x)$ looks quite smooth to the eye), it is worth noting that we have demonstrated that the inversion algorithm is useful with only *two significant digits* of accuracy.

In Figures 7.5 and 7.6 we show the results from $\epsilon = .1$, both with and without smoothing of g. In Figures 7.7 and 7.8 we examine the extreme case of $\epsilon = .25$, again with and without smoothing. It is significant that even for these noisy examples, where the perturbed $g(x)$ is visibly rough, the algorithm performs reasonably well.

7.6.2 Sphere Radius Density $f(r) = 6r(1-r)$: Sampling from $g(x)$

It is straightforward to randomly sample from the density $g(x)$. To do so, begin by choosing a sphere radius at random from the density $f(r)$; that is, from a Beta (2, 2) density. Let this selected sphere radius be r_j. Next, choose a center for this sphere from the uniform density on [0,1]; let the chosen center be c_j. Let the plane of the profile sections be at 1. If $c_j + r_j > 1$, then the jth sphere has been cut, and we can proceed to select a profile radius. If $c_j + r_j \leq 1$, then the jth sphere was too far from the plane to be sectioned, and the selected r_j does not provide a profile radius. Let $d_j \equiv c_j + r_j - 1$. If $d_j > 0$, then simple geometry shows that the desired profile radius is $q_j = \sqrt{r_j^2 - (r_j - d_j)^2}$.

For a numerical example, we selected 1000 sphere radii, of which 515 were cut by the sectioning plane, resulting in 515 random draws from the density (7.29). A 50-cell histogram of these 515 profile radii is given in Figure 7.9. We will explain below the solid and broken lines in this figure. As a check, we have that the average of these 515 radii is $\bar{x} = .4773$ with a standard error of .0092, which is less than one standard deviation greater than

$$\int_0^1 xg(x)dx = \int_0^1 6x\left\{x\sqrt{1-x^2} - x^3\log[x^{-1} + \sqrt{x^{-2}-1}]\right\}dx \doteq .4712. \qquad (7.31)$$

We now use this histogram to estimate the density (7.29) as follows. First we form two vectors: an abscissa vector of the midpoints of the histogram cells in Figure 7.9, and an ordinate vector of the cell counts. These two vectors determine a piecewise linear function which we take to be our right hand side g. Next, we evaluate this piecewise linear interpolant at the quadrature points for 25-point Gauss-Legendre quadrature. We integrate the resulting function, and normalize it so as to provide a density estimate. The estimate which results is superposed, suitably scaled, on the histogram as a solid piecewise linear function. The broken line results from applying one pass of the smoother '4(3RSR)2H twice' to the solid line. We will refer to the broken line as a 'smoothed density estimate'.

In Figure 7.10a, successive approximations to the solution are displayed for the smoothed density estimate in Figure 7.9. The results are disappointing; there is an unwanted peak near $x = 0$. This difficulty does not occur in the real data example of the following subsection, but it is not yet understood. Silverman et. al. (1990) comment on the same phenomenon when they treat this random sphere problem using a 'smoothed EM' approach. In Figure 7.10b we see evidence of the ill-posed nature of our problem: the unreasonable approximation with the unwanted peak still gives a right hand side close to the smoothed profile density estimate.

7.6.3 A Real Data Example: Liver Cell Nuclei

We now consider a real example, taken from Keiding et. al. (1972). The function $g(x)$ consists of smoothed midpoints of a histogram of liver cell nuclei profile radii (Keiding, 1972, p. 823).

In Figure 7.11, this histogram is displayed along with the estimate of $f(r)$ which the Conditional Expectation algorithm provides. The sphere density estimate becomes slightly negative for small r; it is truncated to zero in the figure. Of course, an estimate of a profile radius density obtained from real data need not correspond to *any* sphere density under our idealized model. Density estimates which are negative in places are to be expected with real data from any algorithm unless the algorithm constrains the solution to remain positive. Overall, though, the sphere radius density estimate looks reasonable, and it compares favorably with estimates for this (and other) datasets in the stereology literature.

Figure 7.1: Random Sphere Kernel After Change of Variable

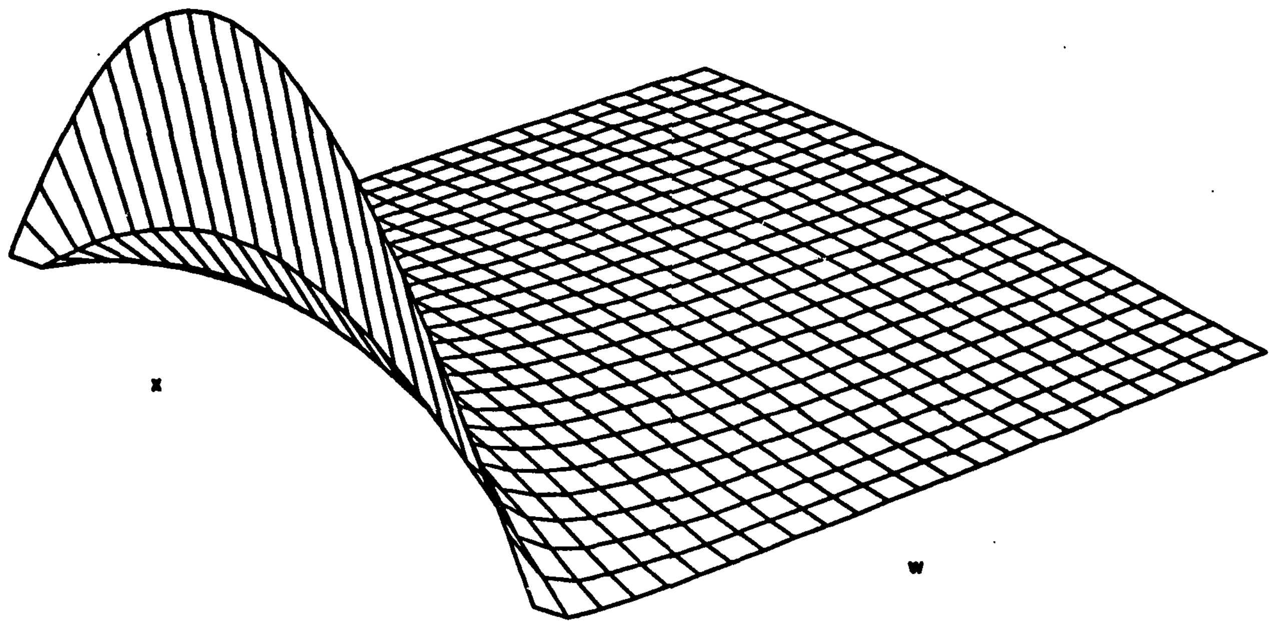


Figure 7.2a: Approximations to Solution (epsilon=0)

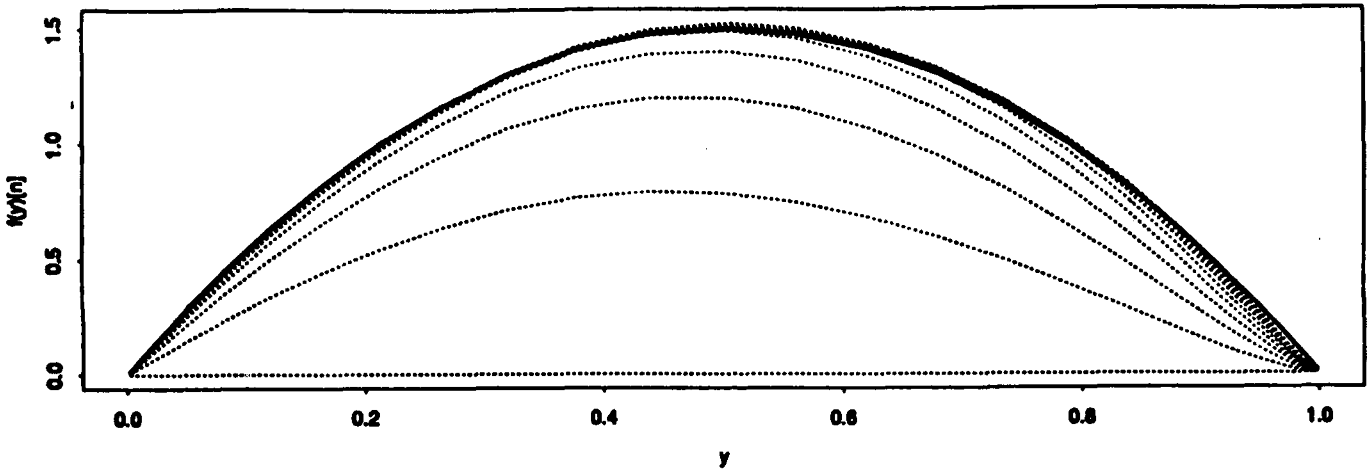


Figure 7.2b: Approximations to g(x) (epsilon=0)

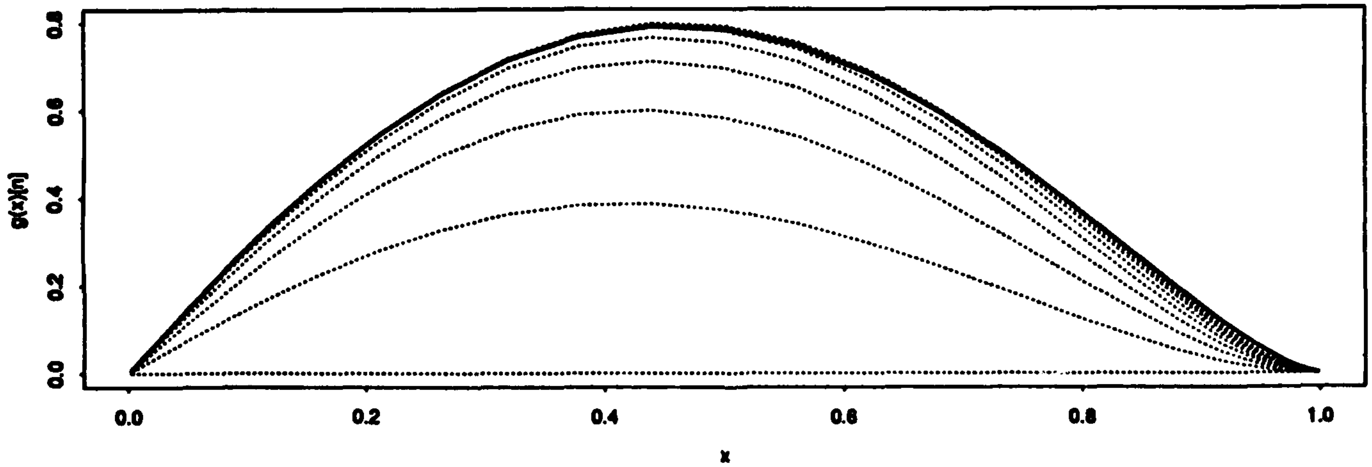


Figure 7.3a: Approximations to Solution (epsilon=.001, no smoothing)

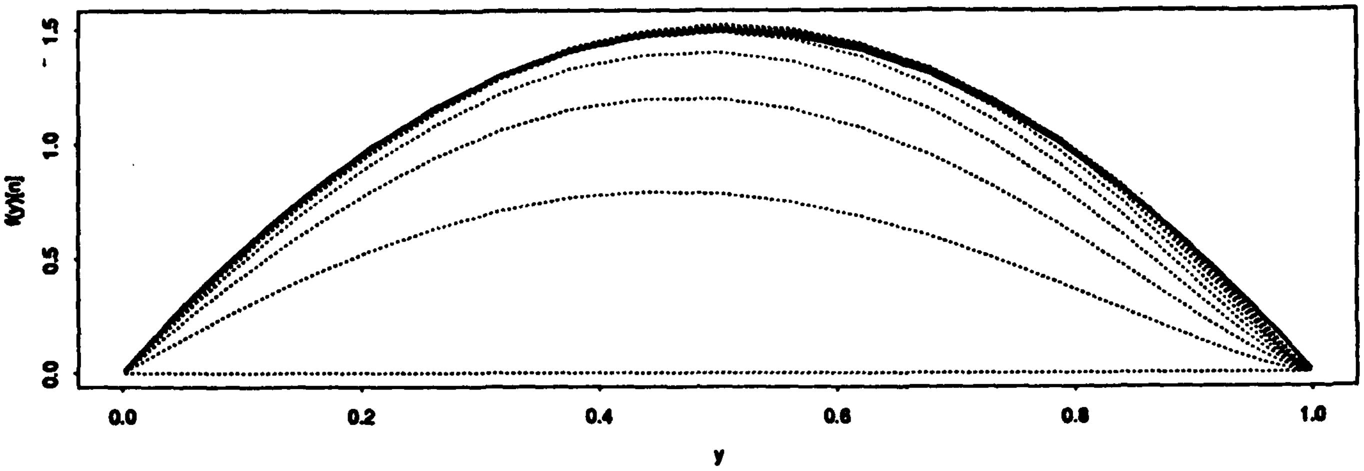


Figure 7.3b: . Approximations to g(x) (epsilon=.001, no smoothing)

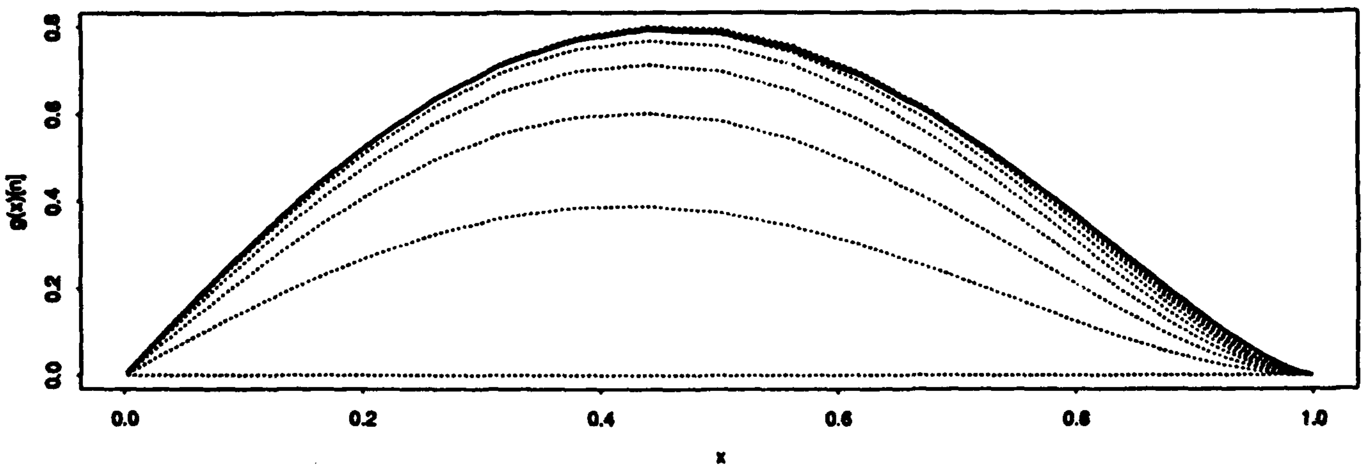


Figure 7.4a: Approximations to Solution (epsilon=.01, no smoothing)

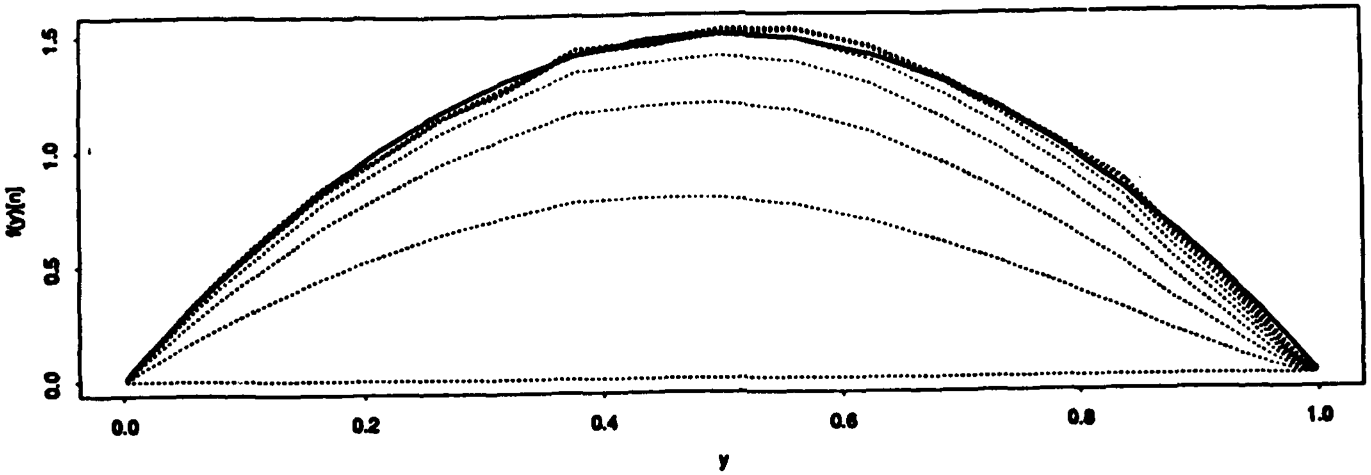


Figure 7.4b: Approximations to g(x) (epsilon=.01, no smoothing)

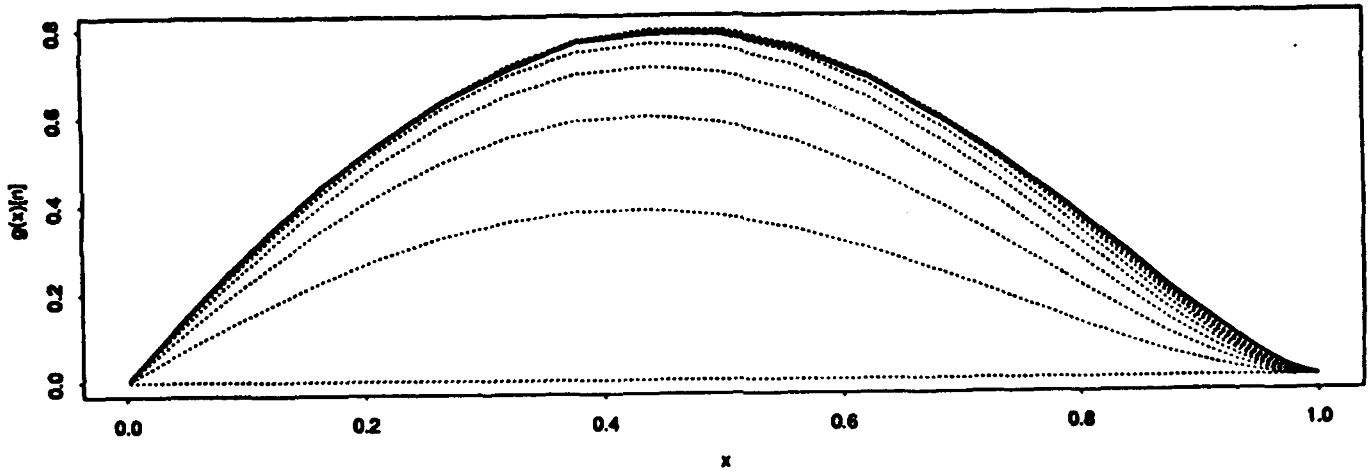


Figure 7.5a: Approximations to Solution (epsilon=.1, no smoothing)

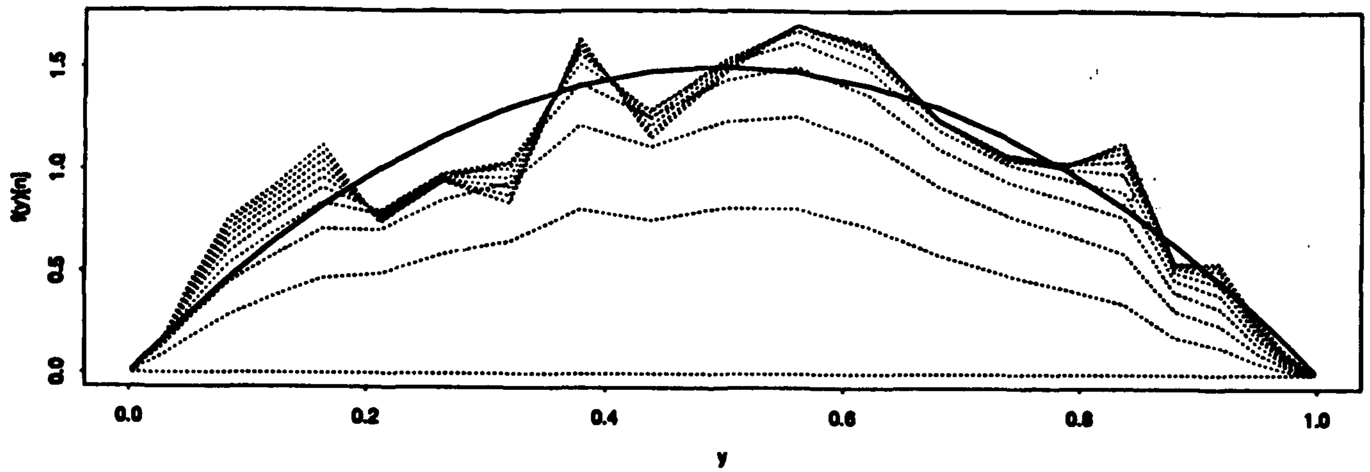


Figure 7.5b: Approximations to g(x) (epsilon=.1, no smoothing)

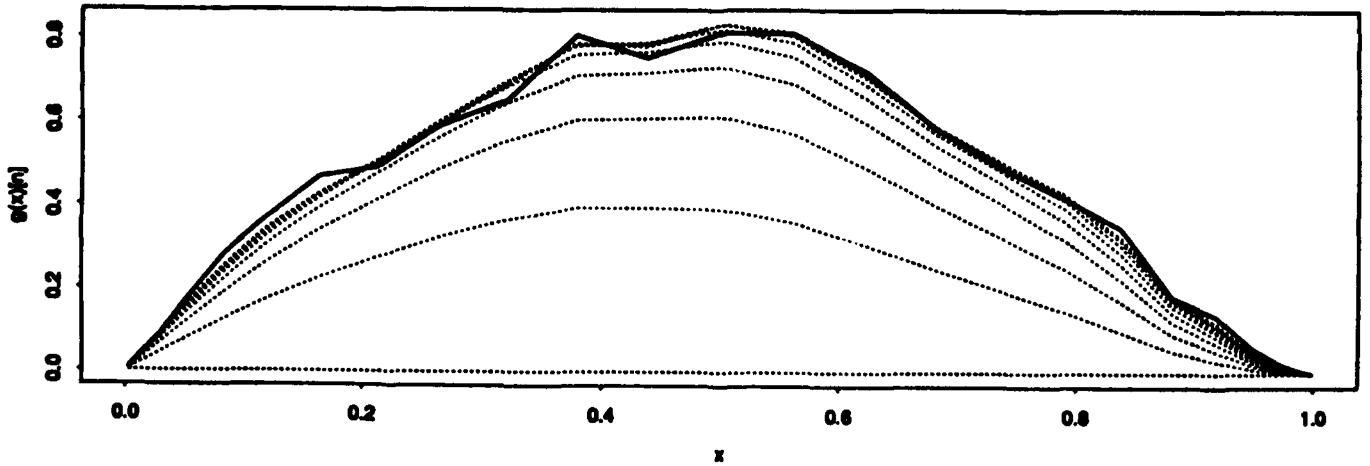


Figure 7.6a: Approximations to Solution (epsilon=.1, smoothing)

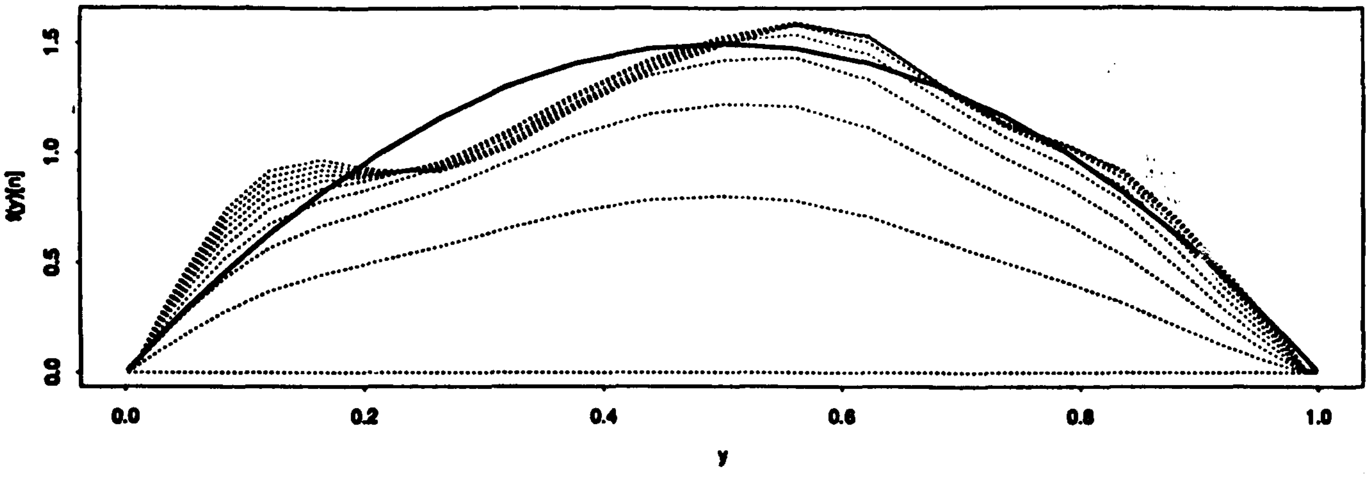


Figure 7.6b: Approximations to g(x) (epsilon=.1, smoothing)

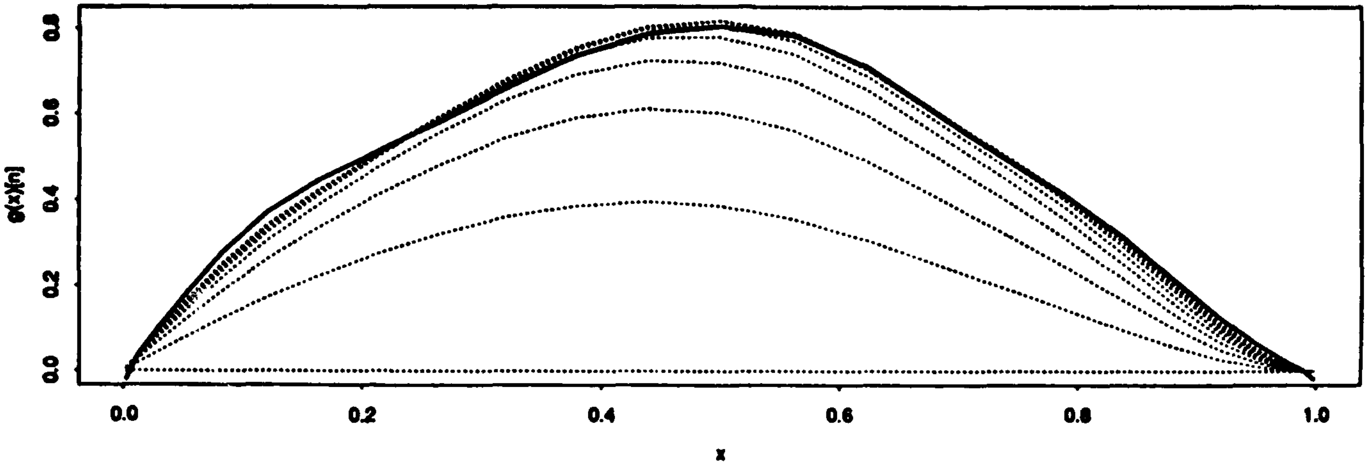


Figure 7.7a: Approximations to Solution (epsilon=.25, no smoothing)

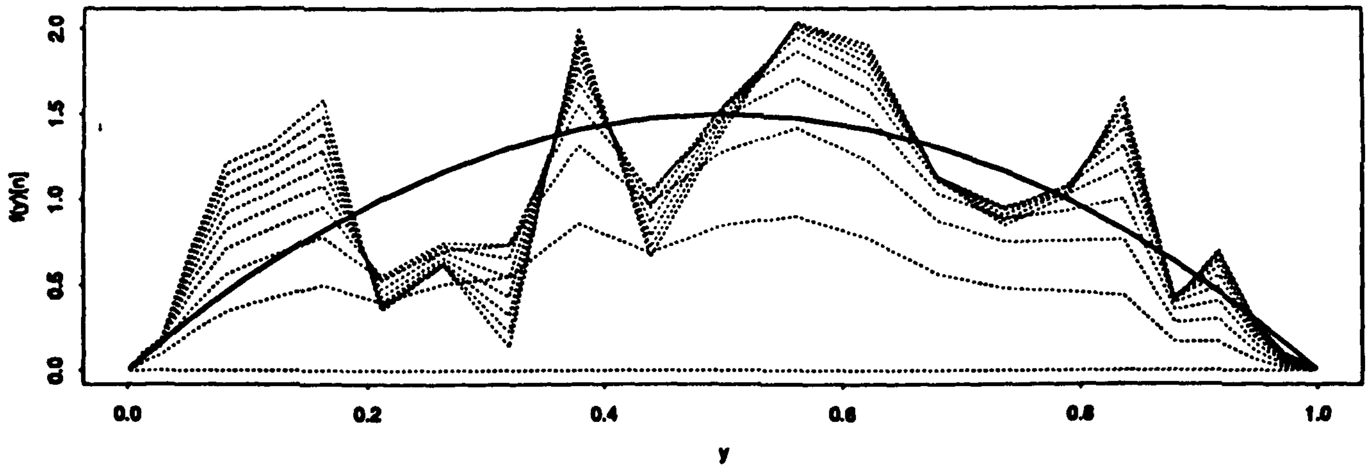


Figure 7.7b: Approximations to g(x) (epsilon=.25, no smoothing)

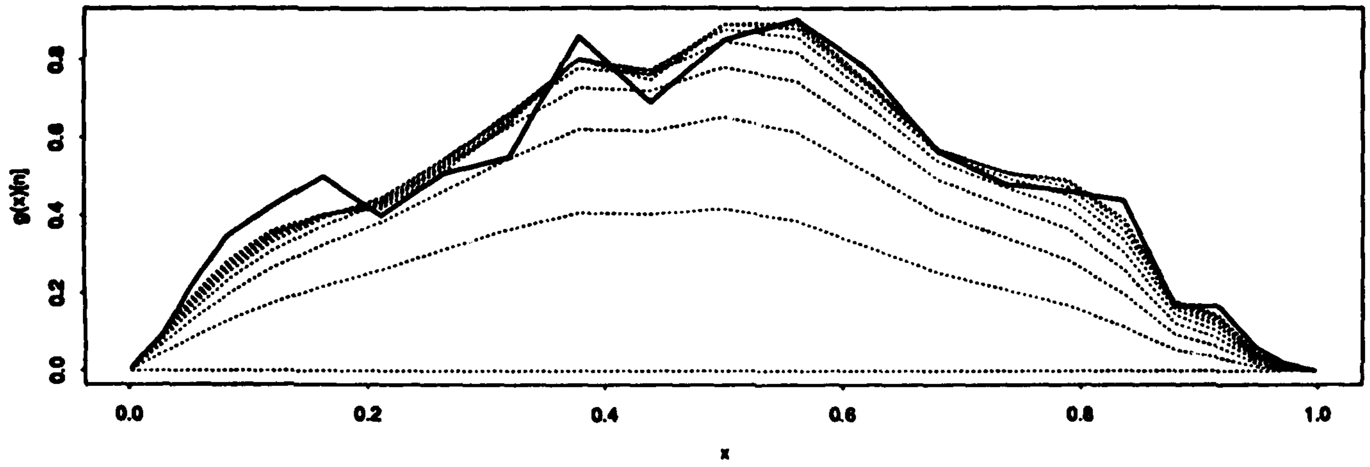


Figure 7.8a: Approximations to Solution (epsilon=.25, smoothing)

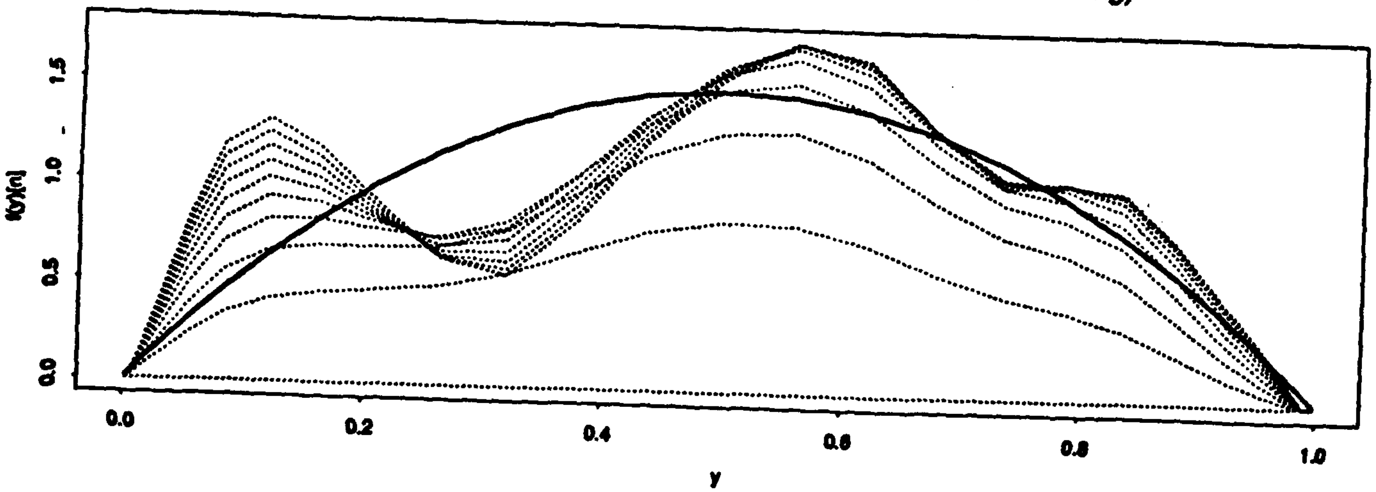


Figure 7.8b: Approximations to g(x) (epsilon=.25, smoothing)

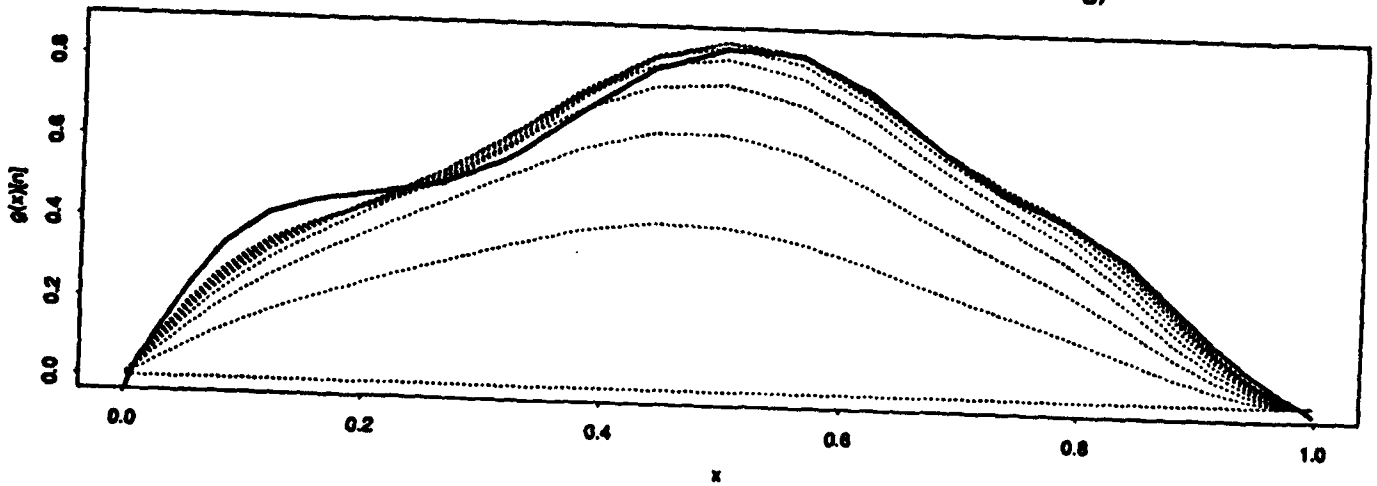


Figure 7.9: A Sample from a Profile Radius Density

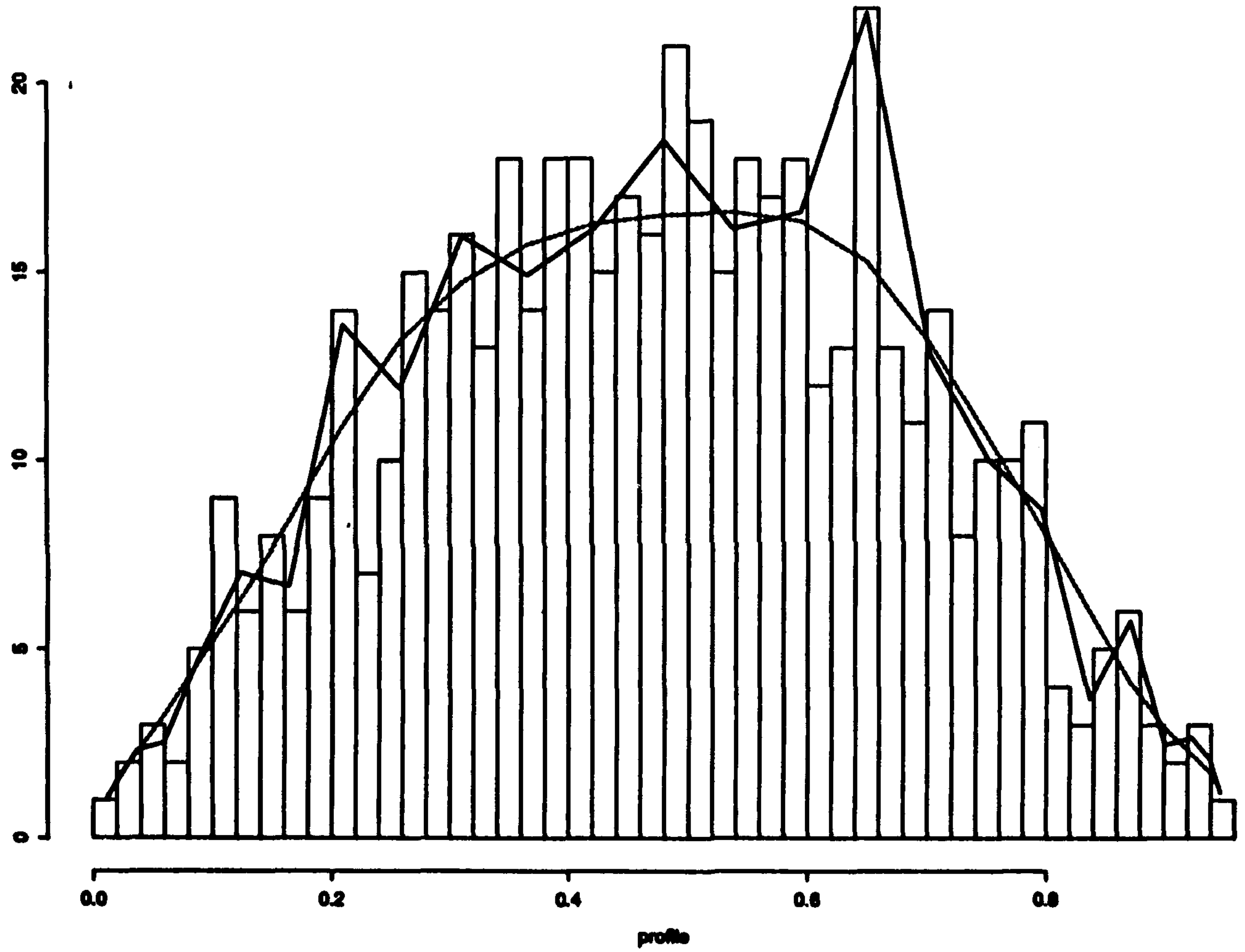


Figure 7.10a: Approximations to Solution for Density Estimate RHS

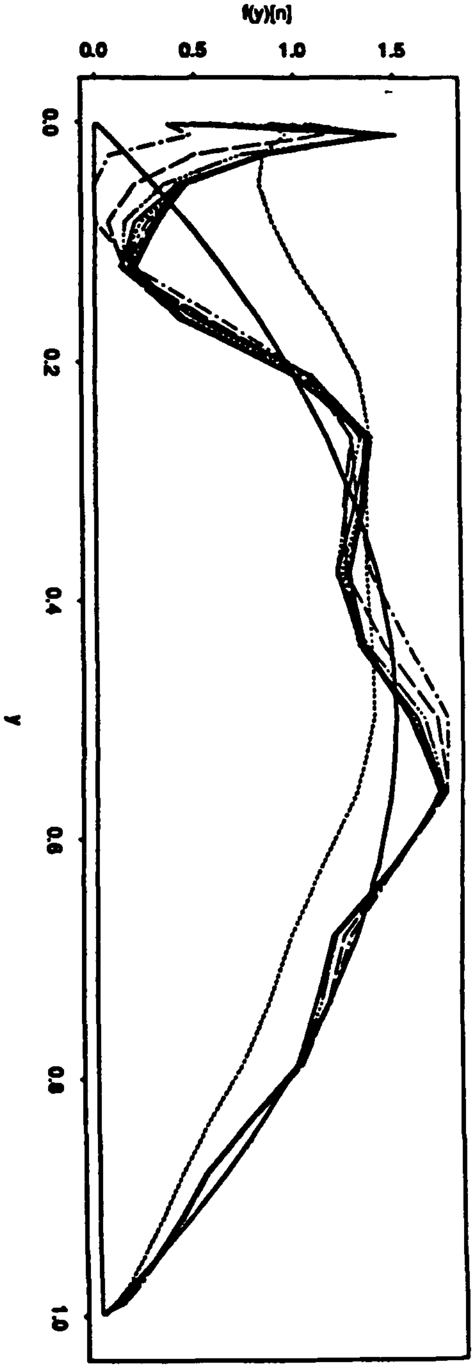


Figure 7.10b: Approximations to Density and Smoothed Density Estimate

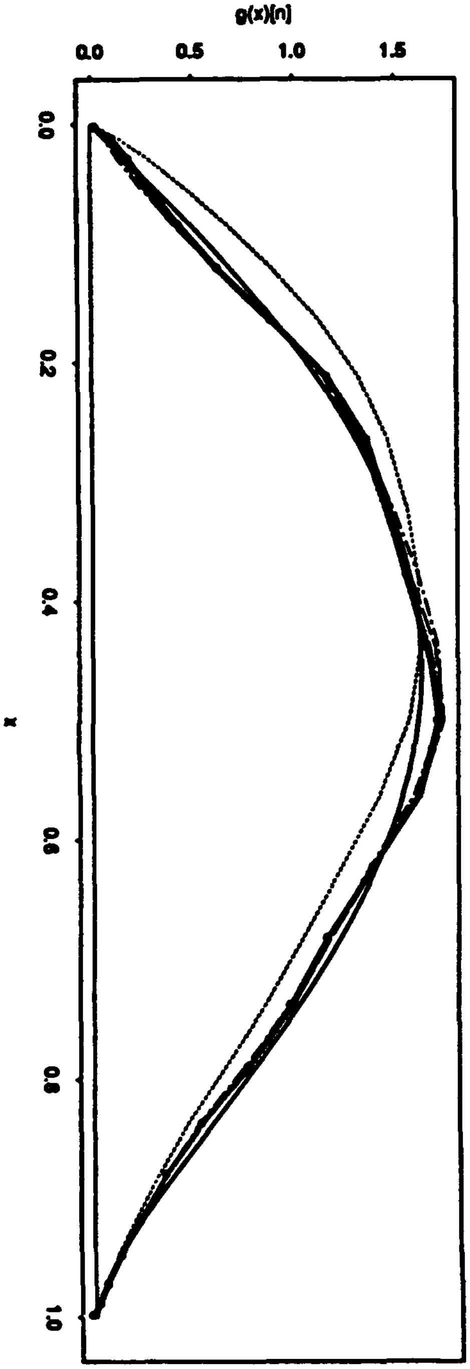
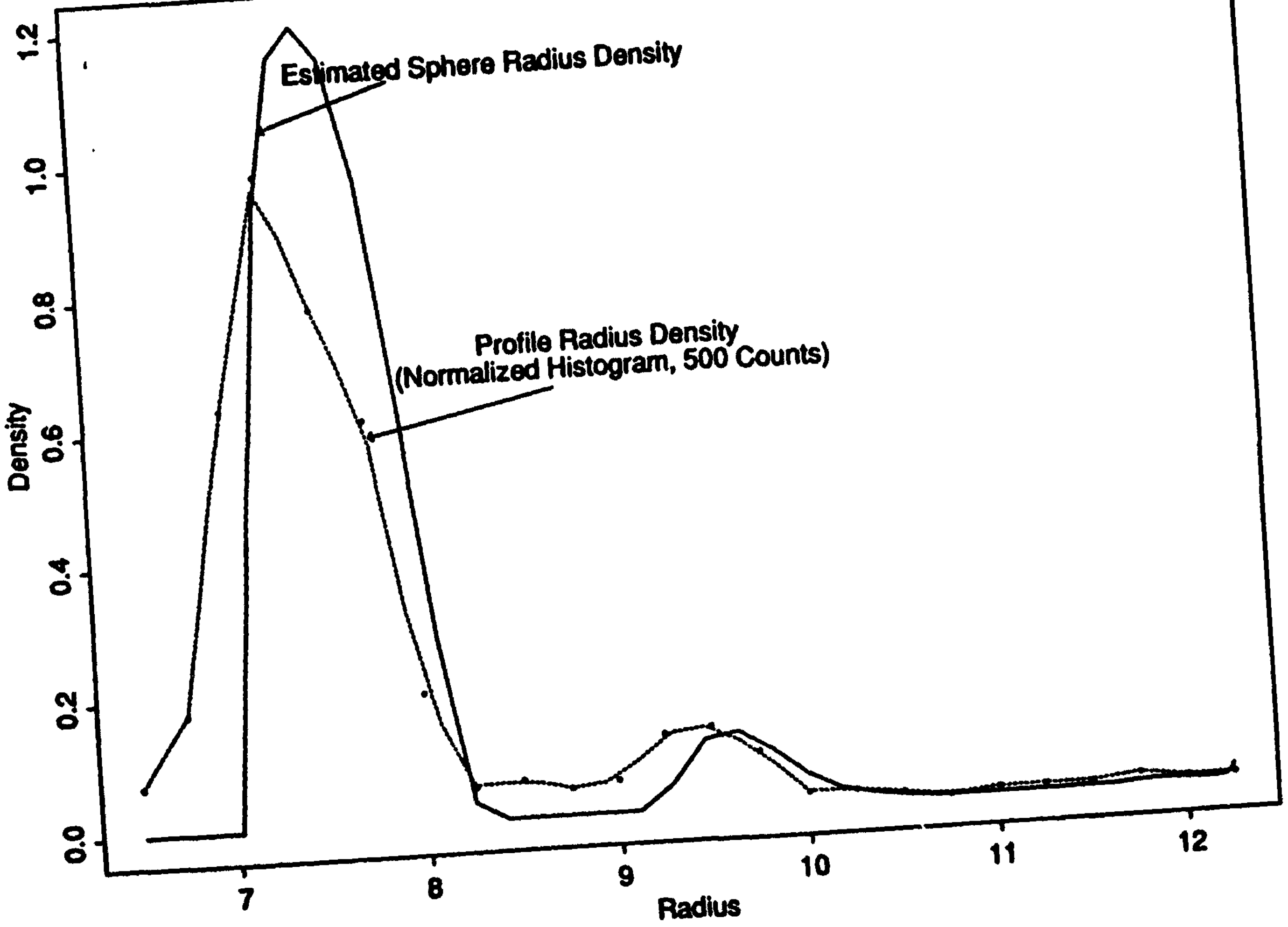


Figure 7.11: Keiding, et. al., Biometrics (1972), Patient 2003



Chapter 8

Conclusions

The focus of this thesis has been on a simple iterative algorithm for integral equations of the first kind with positive kernels, which we have called the *Conditional Expectation Algorithm.* A study of a numerical algorithm for solving deterministic equations might be regarded more as work in numerical analysis than as statistics. However there are numerous connections with this work to statistics, some of which are:

1. The first instance of an approximation to the Conditional Expectation algorithm appears in the attempt of Trickett and Welch (1954) to find a similar test for the Behrens-Fisher problem. The Trickett-Welch algorithm apparently converged to a ⁻ᵐooth 'solution' for a problem which has either none or only pathological solutions.

2. The Conditional Expectation algorithm has been applied successfully to a difficult problem in one-sided β-content tolerance limits for a balanced one-way ANOVA model, a problem which is of some importance in engineering statistics. The result is a new method (Vangel, 1992) which provides tolerance limits with confidence level virtually independent of nuisance parameters. Like the Behrens-Fisher problem, this problem most likely has, at best, pathological exact solutions.

3. Many other applications to problems in mathematical statistics which can be formulated as integral equations of the first kind are clearly possible.

4. The Conditional Expectation algorithm has been shown to be potentially useful for certain inverse problems of indirect measurement. This usefulness has been demonstrated by means of a classical inverse problem of stereology. The statistical analysis of inverse problems is an area of considerable interest to statistics.

5. The name 'Conditional Expectation algorithm' was chosen to emphasize the probabilistic motivation for the method. We have introduced the notion of *stochastic preconditioning*, a process which transforms any positive, bounded kernel into a conditional density. Each step of the proposed algorithm then constitutes the conditional expectation of the true discrepancy of a solution from the current iterate, with respect to this density.

In Chapters 5-7 of this thesis we consider, in succession, the Behrens-Fisher problem, a random effects tolerance limit problem, and an inverse problem in stereology. All of these problems involve solving nonlinear ill-posed integral equations of the first kind, all are of statistical interest, and all are successfully treated by the Conditional Expectation algorithm. This algorithm has also been useful in several other examples, but we have chosen not to report on them here. Instead, we have attempted to explain why the remarkably simple algorithm which we have proposed often works so well on problems formulated as ill-posed integral equations.

In order to attempt to answer this question we have made a long detour into numerical and functional analysis, with some interesting results:

1. Sufficient and, in some cases, necessary conditions for the convergence of Richardson's algorithm for singular matrix equations have been established for singular and nonsingular matrix equations. These general theorems provide insight into why such algorithms can still perform well when applied to inconsistent equations.

2. Several motivations for the stochastic preconditioning which leads to the Conditional Expectation algorithm have been developed. One of these heuristic motivations suggested the name for the algorithm.

3. One peculiarity of iterative algorithms applied to ill-posed integral equations is that these algorithms can produce smooth near-solutions without requiring explicit regularization. There must be regularization implicit in iteration, and this regularization arises because the kernel of the equation tends to smooth. We have made this idea precise by showing that each step in a linear iteration minimizes a quadratic form, which is a sum of two terms. The first term measures how close the present iterate is to the vector to which it is converging, and the second term (for discretizations of integral equations with smooth kernels) tends to penalize 'rough' iterates.

We conclude by suggesting three directions for future research. The first is into the numerical aspects of the Conditional Expectation algorithm, an area in which our results have been incomplete. This is work for a numerical analyst. The second two areas are of more statistical interest:

1. Apply the Conditional Expectation algorithm to other problems in statistical methodology. These problems include similar tests and confidence intervals in normal theory problems (such as, for example, confidence intervals for linear combinations of variance components). Another promising class of problems in empirical Bayes methodology concerns estimating a prior density on a parameter given an estimate of the marginal density of the data.

2. Explore the role of the Conditional Expectation algorithm in other applied inverse problems, perhaps in image processing.

This thesis had its origin in 1985, with a chance encounter with Trickett and Welch (1954), an article virtually ignored in the statistical literature. Some points of interest along the path followed since then are summarized in this document. There is much more to be seen; a fundamental understanding of how the Trickett-Welch and related algorithms work their magic is still lacking. Perhaps others will take up the trail.

Bibliography

[1] Anderssen, R. S. and A. J. Jakeman (1974), "Abel Type Integral Equations in Stereology. II. Computational Methods of Solution and the Random Sphere Approximation", *Journal of Microscopy*, 105, 135-153.

[2] Aspin, A. A. (1948), "An Examination a..d Further Development of a Formula Arising in the Problem of Comparing Two Mean Values", *Biometrika* 35, 88-96.

[3] Bakushinskii, A. B. (1967), "A General Merhod of Constructing Regularizing Algorithms for a Linear Ill-Posed Equation in Hilbert Space", *U. S. S. R. Computational Mathematics and Mathematical Physics*, 7, 3, 279-286.

[4] Becker, R. A., Chambers, J. M., and Wilks, A. R. (1988),*The New S Programming Language: A Programming Environment for Data Analysis and Graphics*, Wadsworth & Brooks/Cole, Pacific Grove, California.

[5] Berger, J. O. (1985), *Statistical Decision Theory and Bayesian Analysis*, Second Edition, Springer-Verlag, New York.

[6] Best, D. J., and Rayner, J. C. W. (1987), "Welch's Approximate Solution for the Behren's-Fisher Problem", *Technometrics* 29, 205.

[7] Bialy, H. (1959), "Iterative Behandlung Linearer Functionalgleichungen", *Arch. Rat. Mech. Anal.*, 4, 166-176.

[8] Bickel, P. J., and Doksum, K. A. (1977), *Mathematical Statistics: 556 Basic Ideas and Selected Topics*, Holden-Day, Oakland, California.

[9] Buja, A., Hastie, T., and Tibshirani, R. (1989), "Linear Smoothers and Additive Models", *Annals of Statistics*, 17, 453-555.

[10] Campbell, S. L. and Meyer, C. D. (1979), *Generalized Inverses of Linear Transformations*, Pitman, London. (Reprinted by Dover Publications, New York, 1990).

[11] Chernoff, H and Moses, L. E. (1959), *Elementary Decision Theory*, Reprinted by Dover Publications, New York, 1985).

[12] Chung, K.L. (1974), *A Course in Probability Theory*, Second Edition, Academic Press, New York.

[13] Coleman, R. (1987), "Inverse Problems", *Journal of Microscopy*, 153, 233-248.

[14] Cruze-Orive, L. M. (1983), "Distribution-free Estimation of Sphere Size Distributions From Slabs Showing Overprojection and Truncation, With a Review of Previous Methods", *Journal of Microscopy*, 131, 265-290.

[15] Debnath, L. and Mikusiński, P. (1990), *Introduction to Hilbert Spaces With Applications*, Academic Press, New York.

[16] Ferguson, T. S. (1967), Mathematical Statistics: A Decision Theoretic Approach, Academic Press, New York.

[17] Fleiss, J. L. (1971), "On the Distribution of a Linear Combination of Independent Chi Squares", *Journal of the American Statistical Association*, 66, 142-144.

[18] Griffiths, P. and Hill, I. D. (1985), *Applied Statistics Algorithms*, New York: John Wiley.

[19] Groetsch, C. W. (1980), *Elements of Applicable Functional Analysis*, Marcel Dekker, New York.

[20] Groetsch, C. W. (1984), *The Theory of Tikhonov Regularization for Fredholm Equations of the First Kind*, Pitman, Marshfield, Massachusetts.

[21] Horn, R. A. and Johnson, C. R. (1985), *Matrix Analysis*, Cambridge University Press, Cambridge.

[22] Horn, R. A. and Johnson, C. R. (1990), *Topics in Matrix Analysis*, Cambridge University Press, Cambridge.

[23] Jakeman, A. J. and R. S. Anderssen (1974), "Abel Type Integral Equations in Stereology. II. General Discussion", *Journal of Microscopy*, 105, 121-133.

[24] Keiding, N. and S. Tolver Jensen (1972), "Maximum Likelihood Estimation of the Size Distribution of Liver Cell Nuclei From the Observed Distribution in a Plane Section", *Biometrics*, 28, 813-829.

[25] Kendall, M., and Stuart, A. (1977), *The Advanced Theory of Statistics: Volume 2, Classical Inference and Relationships*,Oxford University Press, Oxford.

[26] Kreyszig, E. (1978), *Introductory Functional Analysis with Applications*, John Wiley and Sons, New York.

[27] Kress, R. (1989), em Linear Integral Equations, Springer-Verlag, Berlin.

[28] Landweber, L (1951), "An Iteration Formula for Fredholm Integral Equations of the First Kind", *Amer. J. of Math.*, 73, 615-624.

[29] Lehmann, E. L. (1986), *Testing Statistical Hypotheses*, 2nd ed., John Wiley and Sons, New York.

[30] Lemon, G. H. (1977) "Factors for One-Sided Tolerance Limits for Balanced One-Way ANOVA Random-Effects Model", *Journal of the American Statistical Association*, 72, 676-680.

[31] Lenth, Russell V. (1988), "Cumulative Distribution Function for the Non-central *t* Distribution", Algorithm AS-243, *Applied Statistics*, 38, 185-189.

[32] Linnik, Y. V. (1968), *Statistical Problems with Nuisance Parameters*, Translations of Mathematical Monographs Volume 20, Providence, RI: American Mathematical Society.

[33] Maric, N. and Graybill, F. A. (1979), "Small Sample Confidence Intervals on Common Mean of Two Normal Distributions With Unequal Variances", *Communications in Statistics: Theory and Methods*, A8(13), 1255-1269.

[34] Mee, R. W. and Owen, D. B. (1983) "Improved Factors for One-Sided Tolerance Limits for Balanced One-Way ANOVA Random Model", *Journal of the American Statistical Association*, 78, 901-905.

[35] Meisner, J. (1967), "Estimation of the Distribution of Diameters of Spherical Particles From a Given Grouped Distribution of Diameters of Observed Circles Formed by a Plane Section", *Statistica Neerlandica*, 21, 11-30.

[36] Mil Handbook 5E (1987), *Metallic Components for Aircraft Structures*, Naval Publications and Forms Center, Philadelphia.

[37] Mil Handbook 17C (1992), *Polymer Matrix Composites, Volume I: Guidelines*, Naval Publications and Forms Center, Philadelphia.

[38] Nychka, D., Wahba, G., Goldfarb, S. and T. Pugh (1984),"Cross-Validated Spline Methods for the Estimation of Three-Dimensional Tumor Size Distributions From Observations on Two-Dimensional Cross Sections", *J. of the Am. Stat. Assn.*, 79, 832-846.

[39] Odeh, R. E. and Owen, D. B. (1980), *Tables of Normal Tolerance Limits, Sampling Plans and Screening*, Marcel Dekker.

[40] Ortega, J. M. (1972), *Numerical Analysis, a Second Course*, Academic Press, New York (Reprinted by SIAM, Philadelphia, 1990).

[41] Ortega, J. M., and Rheinboldt, W. C., (1970), *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York.

[42] O'Sullivan, F. (1986), "A Statistical Perspective on Ill-posed Inverse Problems", *Statistical Science*, 1, 502-527.

[43] Owen, D. B. (1968), "A Survey of Properties and Applications of the Noncentral t-Distribution", *Technometrics*, 10, 445-478.

[44] Patterson, W. M. (1974), *Iterative Methods for the Solution of a Linear Operator Equation in Hilbert Space - A Survey*, Springer Lecture Notes in Mathematics #394, Springer-Verlag, New York.

[45] Pfanzagl, J. (1974), "On the Behrens-Fisher Problem", *Biometrika*, 61, 39-47.

[46] Phillips, D. L. (1962), "A Technique for the Numerical Solution of Certain Integral Equations of the First Kind", *J. of the Association of Computing Machinery*, 9, 84-97.

[47] Picard, E. (1910), "Sur un Théorème Générale Relatif aux Equations Integrales de Première Espèce et sur Quelques Problèmes de Physique Mathématique, *Rend. Circ. Math. Palermo*, 29, 79-97.

[48] Porter, D. and Stirling D. S. G. (1990), *Integral Equations*, Cambridge University Press.

[49] Press, W. H., et. al. (1986), *Numerical Recipes*, Cambridge University Press, Cambridge.

[50] Reese, C. and Sorem, J. (1981), "Statistical Distribution of Mechanical Properties for Three Graphite-Epoxy Material Systems", Contractor Report CR-16736, National Aeronautics and Space Administration.

[51] Richardson, L. F. (1910), "The Approximate Arithmetical Solution by Finite Differences of Physical Problems Involving Differential Equations With an Application to Stresses in a Masonry Dam", *Phil. Trans. Roy. Soc. Lond.*, A, 210, 307-357.

[52] Satterthwaite, F. E. (1946), "An Approximate Distribution of Estimates of Variance Components", *Biometrics Bulletin*, 2, 110-114.

[53] Seeger, P. and Thorsson, U. (1972), "Two Sided Tolerance Limits With Two-Stage Sampling From Normal Populations - Monte Carlo Studies of the Distributions of Coverages", *Applied Statistics*, 22, 292-300.

[54] Silverman, B. W., Jones, M. C., Wilson, J. D. and D. W. Nychka (1990), "A Smoothed EM Approach to Indirect Estimation Problems With Particular Reference to Stereology and Emission Tomography", *J. R. Statist. Soc. B*, 52, 271-324.

[55] Smithies, F (1958), *Integral Equations*, Cambridge University Press, London

[56] Stoer, J. and Bulirsch, R. (1980), *Introduction to Numerical Analysis*, Springer-Verlag, New York.

[57] Strand, O. N. (1974), "Theory and Methods Related to Singular Function Expansions and Landweber's Iteration for Integral Equations of the First Kind", SIAM J. Numer. Anal., 11, 798-825.

[58] Strang, G. (1976), *Linear Algebra an its Applications*, Academic Press, New York.

[59] Taylor, C. C. (1982), "A New Method for Unfolding Sphere Size Distributions", *Journal of Microscopy*, 29, 57-66.

[60] Tikhonov, A. N. (1963), "Regularization of Incorrectly Posed Problems", *Soviet Math. Doklady*, 4, 1624-1627.

[61] Tikhonov, A. N. and Arsenin, V. Y. (1977), *Solutions of Ill-Posed Problems*, Wiley, New York.

[62] Trickett, W. H. and Welch, B. L. (1954), "On the Comparison of Two Means: Further Discussion of Iterative Methods for Calculating Tables", *Biometrika*, 41, 361-374.

[63] Tricomi, F. G. (1957), *Integral Equations*, Interscience Publishers, New York (reprinted by Dover Publications, New York, 1985).

[64] Tucker H. G. (1967), *A Graduate Course in Probability*, Academic Press, New York.

[65] Van der Sluis, A (1969), "Condition Numbers and Equilibration of Matrices", *Numer. Math.*, 14, 14-23.

[66] Vangel, M. G. (1987), "An Exact Method for One-Sided Tolerance Limits in the Presence of Batch-to-Batch Variation", in *Proceedings of the Thirty-Second Conference on the Design of Experiments in Army Research, Development, and Testing*, U. S. Army Research Office Report ARO 87-2, 77-102.

[67] Vangel, M. G. (1990), "The Trickett-Welch 'Solution' to the Behrens-Fisher Problem Applied to One-Sided Tolerance Limits for Random Effects Models", Technical Report ONR-C-5, Harvard University.

[68] Vangel, M. G. (1992), "New Methods for One-Sided Tolerance Limits for a One-Way Balanced Random Effects ANOVA Model", to appear in *Technometrics.*

[69] Varga, R. S. (1962), *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, New Jersey.

[70] Vardi, Y. (1992), "From Image Deblurring to Optimal Investments: Using Statistics to Solve an Integral Equation", preprint.

[71] Velleman P. F. and D. C. Hoaglin (1981), *Applications, Basics and Computing of Exploratory Data Analysis*, Duxbury, Boston.

[72] Wallace, D. L. (1980), "The Behrens-Fisher and Fieller-Creasy Problems", in *R. A. Fisher: An Appreciation*, S. E. Fienberg and D. V. Hinkley, eds., Springer-Verlag, New York, 119-147.

[73] Wang, C. M. (1988), "β-Expectation Tolerance Limits for Balanced One-Way One-Way Random-Effects Model", in *Probability and Statistics: Essays in Honor in Franklin A. Graybill*, 285-295.

[74] Weinstock, R (1974), *Calculus of Variations: With Applications to Physics and Engineering*, Dover Publications, New York.

[75] Welch, B. L. (1937), "The Significance of the Difference Between Two Means When the Population Variances are Unequal", *Biometrika*, 29, 350-362.

[76] Welch, B. L. (1947), "The Generalization of Student's Problem When Several Different Population Variances are Involved", *Biometrika*, 34, 28-35.

[77] Wicksell, S. D. (1925), "The Corpuscle Problem", *Biometrika*, 17, 84-99.

[78] Wicksell, S. D. (1926), "The Corpuscle Problem II", *Biometrika*, 18, 151.

[79] Young, N (1988), *A Introduction to Hilbert Space*, Cambridge University Press, Cambridge.

Appendix A

A Setup for Numerical Problems

To iteratively approximate the solution of an integral equation numerically, we replace the integral equation by an approximating system of algebraic equations. We then apply a version of Richardson's algorithm to the solution of this system of equations.

All integral equations which we will consider are special cases of

$$\int_0^1 k\{x, y, f[\phi(x,y)]\}dy = g(x), \tag{A.1}$$

where k, g, and $\phi : [0,1] \times [0,1] \to [0,1]$ are known functions. By choosing a mesh of x and y values, a quadrature rule, and an interpolation rule for f, equation (A.1) can be replaced with a system of nonlinear equations.

Note that (A.1) allows for the possibility that the kernel, k, is nonlinear in the unknown, f. When this is the case, we will linearize this equation, thus reducing the numerical problem to one of approximately solving a linear integral equation, with a different kernel, at each iteration.

We discuss first a numerical setup suitable for the general equation (A.1), and then we consider important special cases. Although we present this discretization scheme for an integral equation on the unit square, extension to other regions is straightforward.

A.1 The General Setup

Let $\{y_j\}_{j=1}^c$ be quadrature abscissas, and let $\{w_j\}_{j=1}^c$ denote the quadrature weights. We choose r points at which the unknown function, f, is to be determined, and we let these points be $\{x_i\}_{i=1}^r$. We define the x_i and y_j values to be ordered:

$$0 \leq x_1 < x_2 < \cdots < x_r \leq 1$$

and

$$0 \leq y_1 < y_2 < \cdots < y_c \leq 1.$$

We will use iterative algorithms to solve r equations in r unknowns. Unless $\phi(x,y) = y$, $c = r$, and $x_i = y_i$ for $i = 1, \ldots, r$, it is necessary to specify a function $\tilde{f}$ which approximates f by interpolation, and possibly also by extrapolation. We will give the details of $\tilde{f}$ defined by linear interpolation, and we will use this definition exclusively in numerical examples.

For any point $z^* = \phi(x_i, y_j)$ such that $x_s < z^* < x_{s+1}$, let

$$\tilde{f}(z^*) \equiv f(x_s) + \frac{f(x_{s+1}) - f(x_s)}{x_{s+1} - x_s}(z^* - x_s). \tag{A.2}$$

If $z^* < x_1$ or $z^* > x_r$, then it is necessary to specify an extrapolation rule. This will usually be required if $y_1 < x_1$ or $y_c > x_r$. Let $x_0 \equiv 0 \leq x_1$ and $x_{r+1} \equiv 1 \geq x_r$. If $x_0 \neq x_1$, assume $f(0)$ is known. Similarly, if $x_{r+1} \neq x_r$, assume $f(1)$ is known. This situation does not arise in this thesis; if it did one could choose a quadrature rule which did not require $f(0)$ or $f(1)$.

For $z^* < x_1$, define

$$\bar{f}(z^*) \equiv f(x_1) + \frac{f(x_1) - f(x_0)}{x_1}(z^* - x_1), \tag{A.3}$$

and for $z^* > x_r$, let

$$\bar{f}(z^*) \equiv f(x_r) + \frac{f(x_{r+1}) - f(x_r)}{1 - x_r}(z^* - x_r). \tag{A.4}$$

When discussing discrete approximations to integral equations, we will let f and g denote the r-dimensional vectors

$$f \equiv \begin{bmatrix} f(x_1) \\ f(x_2) \\ \cdot \\ \cdot \\ \cdot \\ f(x_r) \end{bmatrix} \tag{A.5}$$

and

$$g \equiv \begin{bmatrix} g(x_1) \\ g(x_2) \\ \cdot \\ \cdot \\ \cdot \\ g(x_r) \end{bmatrix}. \tag{A.6}$$

We define $\bar{f}$ to be the rc-dimensional vector

$$\bar{f} \equiv \begin{bmatrix} \bar{f}[\phi(x_1, y_1)] \\ \bar{f}[\phi(x_1, y_2)] \\ \cdot \\ \cdot \\ \cdot \\ \bar{f}[\phi(x_r, y_c)] \end{bmatrix}, \tag{A.7}$$

where the *function* $\bar{f}(z)$ is given by the formulas (A.2)-(A.4) in terms of the elements of the *vector* f. All of the elements of $\bar{f}$ need not be distinct. For example, if $\phi(x, y) = y$ and $r = c$, then there will only be r distinct elements in $\bar{f}$.

There exists an $rc \times r$ matrix M, which is implicitly defined by (A.2)-(A.4), such that

$$\bar{f} = Mf, \tag{A.8}$$

but we will not make any explicit use of this matrix.

We can now replace the equation (A.1) by the approximating system of equations

$$\hat{g}_i(f) \equiv \sum_{j=1}^{c} k\{x_i, y_j, \bar{f}[\phi(x_i, y_j)]\} w_j = g_i, \tag{A.9}$$

for $i = 1, \ldots, r$.

In numerical examples, we will restrict attention to the case $r = c$, $y_i = x_i$, and Gauss-Legendre quadrature, wherever possible.

A preconditioned Richardson algorithm for the system of equations (A.9) is

$$f^{n+1}_{r\times 1} = f^{n}_{r\times 1} + \theta D^{-1}_{r\times r}\left[g_{r\times 1} - \hat{g}(f^n)_{r\times 1}\right], \tag{A.10}$$

where the dimensions of the matrices are as indicated, θ is a positive constant, and D is a nonsingular matrix. We would like to choose D to accelerate convergence.

The reason why the matrix M in (A.8) is never needed should now be clear. The difference $f^{n+1} - f^n$ involves f^n only through the *residual* $g - \hat{g}(f^n)$.

If the elements of $\hat{g}(f)$ are linear in the elements of f, that is, if the interpolation rule is linear and if the kernel, $k(x, y, f)$, is linear in the *function* f, then it is always possible to write

$$\hat{g}(f) = Kf, \tag{A.11}$$

for some $r \times r$ matrix K, although it can, in general, be tedious to determine the elements of K. If $\hat{g}$ is *nonlinear* in f, then we will linearize the system of equations (A.9) at each iteration.

Important special cases of this general setup are discretizations for linear Fredholm and linear Volterra equations. We discuss next the specific discretizations used for all linear Fredholm and Volterra numerical examples in this thesis.

A.2 A Linear Fredholm Example

Consider the Fredholm equation of the first kind

$$\int_0^1 k(x,y)f(y)dy = g(x). \tag{A.12}$$

We will use the setup of Section A.1, with $r = c$ and $x_i = y_i$ for $i = 1, \ldots, r$. Let $\{x_i\}_{i=1}^r$ be the abscissa values for Gauss-Legendre quadrature, and let $\{w_j\}_{j=1}^r$ be the corresponding weights. Define the typical elements of the r-dimensional vectors f and g by

$$f_i \equiv f(x_i) \tag{A.13}$$

and

$$g_i \equiv g(x_i), \tag{A.14}$$

respectively. Note that $\tilde{f}$ has at most r distinct values. The ith element of $\hat{g}(f)$ is

$$\hat{g}_i(f) = \sum_{j=1}^{c} k(x_i, y_j)\tilde{f}_{(i-1)c+j}w_j = \sum_{j=1}^{c} k(x_i, y_j)f_j w_j, \tag{A.15}$$

for $i = 1, \ldots, r$. We can write, for this example, $Kf = g$, where the typical element of K is

$$\kappa_{ij} \equiv k(x_i, y_j)w_j. \tag{A.16}$$

A preconditioned Richardson algorithm, for this discretization of a linear Fredholm equation, is then of the form (3.3).

A.3 A Linear Volterra Example

Consider the Volterra equation of the first kind

$$\int_0^x \hat{k}(x,u)f(u)du = g(x). \tag{A.17}$$

Since the upper limit of integration of (A.17) is not constant, (A.17) is not in the general form of (A.1). However, the change of variable

$$u = xy \tag{A.18}$$

results in the following integral equation with fixed limits of integration:

$$\int_0^1 x\hat{k}(x, xy)f(xy)dy = g(x). \tag{A.19}$$

If we let

$$\phi(x, y) \equiv xy \tag{A.20}$$

and

$$k\{x, y, f[\phi(x, y)]\} \equiv x\hat{k}(x, xy)f(xy), \tag{A.21}$$

then (A.19) is a special case of (A.1).

As with the Fredholm example, we will use the setup of Section A.1, with $r = c$ and $x_i = y_i$ for $i = 1, \ldots, r$. Let $\{x_i\}_{i=1}^r$ be the abscissa values for Gauss-Legendre quadrature, and let $\{w_j\}_{j=1}^r$ be the corresponding weights. Define the ith element of the r-dimensional vector g by

$$g_i \equiv g(x_i). \tag{A.22}$$

We will assume $f(0)$ is known, and define the $r^2 \times 1$ vector $\bar{f}$ by equations (A.2)-(A.4). The elements of $\hat{g}(f)$ are given by

$$\hat{g}_i(f) = \sum_{j=1}^{c} x_i\hat{k}(x_i, x_iy_j)\tilde{f}(x_iy_j)w_j, \tag{A.23}$$

for $i = 1, \ldots, r$. Although $\{\hat{g}(f)_i = g_i\}_{i=1}^r$, where $\hat{g}$ is given by (A.23), is a system of r linear equations in r unknowns, it is not easy, and certainly not necessary, to find a matrix K so that this system can be written in the form $Kf = g$. Richardson's algorithm can be used directly, in the form (A.10).

The approach to the discretization of Volterra equations presented in this subsection will be used exclusively for the Volterra numerical examples in this thesis. However, it is important to note that by transforming a Volterra kernel (on a *triangular* domain) into a kernel on the unit square, we have violated a notion of *causality* inherent in the Volterra equation. The value of the right hand size $g(x)$ depends only on $\{f(y)|y \leq x\}$, and conversely, the solution $f(y)$ depends only on $\{g(x)|x \leq y\}$. Our numerical treatment does not preserve this property. In this sense what we are doing is unconventional. Although it has worked well for the examples considered, we are not in a position, at this time, to advocate it as a general approach.

Appendix B

S Code for Conditional Expectation Algorithm and Richardson Algorithm for the Green's Function Kernel

The following function, written in the *S* programming language, (Becker, Chambers and Wilks, 1988), can be used for solving the integral equation with kernel (4.30). The function *gauleg* calls a FORTRAN routine to determine Gauss-Legendre quadrature abscissas and weights, and is documented in Appendix C.

Note that to solve other linear integral equations on the unit square, only the functions *kernel* and *rhs* need to be modified. This is therefore a very useful function for exploring iterative algorithms for linear integral equations.

```
#
#  Mark Vangel Sept 1990
#
# Richardson and Conditional Expectation algorithms for Green's function
# example.
#
 inteqn_function(niter=10, npt=25, xl=0, xh=1, norm=T, fct=1,
              jac=F, land=F){

#
#  -- Matrices of successive corrections, sums of corrections,
#     and approximate r.h.s
   h_matrix (0, niter, npt)
   f_h
   v_h
#
#  -- Gauss-Legendre abscissas and \weights
   gl_gauleg (npt, xl, xh)
#
#  -- Arrays of x and y values at which functions are evaluated
   x_matrix (gl$x, npt, npt, byrow=T)
#  x_(-6*x**3+9*x**2-x)/2
   y_matrix (gl$x, npt, npt)
#
#  -- Array of weights for numerically integrating kernel w.r.t. y
```

```
   wy_matrix (gl$w, npt, npt)
#
#  -- Kernel, integral of kernel w.r.t. y, and rhs
   k_kernel (x, y) *wy
   u_1 /apply  (k, 2, 'sum')
   if (norm == F) u_fct*u/u
   if (jac  == T) u_1/diag(kernel(x,y))
   g_rhs    (x[1,])
   if (land == T){
       g_ t(k) %*% g
       k_ (t(k) %*% kernel(x,y)) *wy}
#
#  -- First approximation to solution
   f[1,]_g
#
#  -- Now calculate the successive corrections
   for (i in 1:(niter-1)) {
       v[i,]  _t(k) %*% f[i,]
       h[i,]  _u * (g -v[i,])
       f[i+1,]_f[i,]+h[i,]
     }
   inteqn_list (h, f, v,  g, kernel(x,y))
   names(inteqn)_c('h', 'f', 'g', 'rhs', 'k')
   return (inteqn)
   }
#
#  Calculate kernel at gauss points
#
 kernel_function (x, y) {
   i_y<x
   kernel_ i *y *(1-x) +(1-i) *x *(1-y)
   return(kernel)
   }
#
#   Right hand side of equation
#
 rhs_function(x) {
    rhs_x**3*(1-x)**2
    return (rhs)
   }
```

Appendix C

S Code for The Trickett-Welch Algorithm and Conditional Expectation Algorithms for the Behrens-Fisher Problem

C.1 The Trickett-Welch Algorithm

The following function is written in the *S* programming language (Becker, Chambers and Wilks, 1988). The only required call to an external FORTRAN routine is the function *gauleg* which calculates Gauss-Legendre quadrature points and weights. This FORTRAN routine was obtained from Press, et. al. (1986, pp. 125-126), and is not reproduced here. The FORTRAN routines *mydbeta*, *mydt* and *mypt* are double precision versions of the single precision *S* functions *dbeta*, *dt* and *pt* respectively. These FORTRAN functions were obtained from Griffiths and Hill (1985), but satisfactory results for most applications can probably be obtained from the corresponding *S* functions.

```
#
#  Mark Vangel May 1991
#
# Solve Behrens-Fisher integral equation
# (Algorithm as in Trickett-Welch (1954))
#
 bftw_function(niter=10, npt=25, nquad=25,  nspl=100,
             conf=.95, initf=qnorm(conf), n=c(5,5)) {

#  -- Matrices of successive corrections; sums of corrections;
#     and approximate r.h.s.
   h _matrix (0, niter, npt)
   f _h
   v _h
   r _h
#
#  -- d.f. of variance estimates; r.h.s for equation; initial guess for
#     critical value.
   df   _n-1
   g    _matrix(conf, npt, 1)
   f[1,]_initf
   h[1,]_0
#
#  -- Gauss-Legendre abscissas and weights; beta density evaluated at
#     abscissas.
   gl_gauleg (nquad, 0, 1)
   b _mydbeta (gl$x, df[1]/2, df[2]/2)
#
#  -- Arrays of x, y and beta values at which functions are evaluated.
#     The beta values are weighted by the quadrature weights.
   x_matrix ((0:(npt-1))/(npt-1), nquad, npt, byrow=T)
   b_matrix (b, nquad, npt) *matrix (gl$w, nquad, npt)
   y_matrix (gl$x, nquad, npt)
#
#  -- Argument for t density and cdf.
   arg_sqrt((df[1]+df[2]) *(x*y/df[1]+(1-x)*(1-y)/df[2]))
#
#  -- Argument for critical value statistic.
    z_x*y/df[1] /(x*y/df[1] +(1-x)*(1-y)/df[2])
#
#  -- Evaluate the gradient at the mean of the beta density
   xpeak_ df[1]/(df[1]+df[2])
   ipeak_ sum (gl$x < xpeak)
   w    _ gl$x[ipeak+1]-xpeak
#
#  -- Now calculate the successive corrections
   for (i in 1:(niter-1)) {
       sp     _spline(x[1,], f[i,], n=nspl)
       mf     _approx(sp$x, sp$y, c(z))$y
       mf     _matrix(mf, nquad, npt)
       k0     _mypt (arg *mf, df[1]+df[2]) *b
       k1     _mydt (arg *mf, df[1]+df[2])
       drv    _k1[ipeak,]*(1-w) +k1[ipeak+1,]*w
       v[i,]  _apply (k0, 2, 'sum')
       h[i+1,]_-(v[i,]-g)/drv
       f[i+1,]_f[i,] +h[i+1,]
        }
       r_g -v[niter-1,]
    bf_list (x[1,], h, f, v, k1, r)
    names(bf)_c('x', 'h', 'f', 'g', 'k1', 'r')
    return (bf)
   }

   gauleg_function (n, xlow=-1, xhgh=1) {
#
#   Mark Vangel, Sept 1990
#
#   Calculate Gauss-Legendre abscissas and weights
```

```
#
#
   x_matrix(0,n,1)
   w_x
   z_.Fortran("gauleg",as.double (xlow),
                        as.double (xhgh),
                        x = as.double (x),
                        w = as.double (w),
                        as.integer (n))
   z_list(z$x, z$w)
   names (z)_ c('x','w')
   gauleg_z
          }
```

C.2 The Conditional Expectation Algorithm

The following function is written in the *S* programming language (Becker, Chambers and Wilks, 1988). For information on the use of external FORTRAN routines, see the comments preceeding the code in Appendix C.1.

```
#
#  Mark Vangel, Jan 1991
#
# Solve Behrens-Fisher integral equation
#
 bf_function(niter=10, npt=25, nquad=25,  nspl=100,
             conf=.95, initf=qnorm(conf), n=c(5,5)) {
#
#  -- Matrices of successive corrections; sums of corrections;
#     and approximate r.h.s.
  h _matrix (0, niter, npt)
  f _h
  v _h
  r _h
#
#  -- d.f. of variance estimates; r.h.s for equation; initial guess for
#     critical value.
  df   _n-1
  g    _matrix(conf, npt, 1)
  f[1,]_initf
  h[1,]_0
#
#  -- Gauss-Legendre abscissas and weights; beta density evaluated at
#     abscissas.
  gl_gauleg (nquad, 0, 1)
  b _mydbeta (gl$x, df[1]/2, df[2]/2)
#
#  -- Arrays of x, y and beta values at which functions are evaluated.
#     The beta values are weighted by the quadrature weights.
  x_matrix ((0:(npt-1))/(npt-1), nquad, npt, byrow=T)
  b_matrix (b, nquad, npt) *matrix (gl$w, nquad, npt)
```

```
   y_matrix (gl$x, nquad, npt)
#
#  -- Argument for t density and cdf.
   arg_sqrt((df[1]+df[2]) *(x*y/df[1]+(1-x)*(1-y)/df[2]))
#
#  -- Argument for critical value statistic.
    z_x*y/df[1] /(x*y/df[1] +(1-x)*(1-y)/df[2])
#
#  -- Now calculate the successive corrections
   for (i in 1:(niter-1)) {
       sp     _spline(x[1,], f[i,], n=nspl)
       mf     _approx(sp$x, sp$y, c(z))$y
       mf     _matrix(mf, nquad, npt)
       k0     _mypt (arg *mf, df[1]+df[2]) *b
       k1     _mydt (arg *mf, df[1]+df[2]) *b *arg
       drv    _apply (k1, 2, 'sum')
       v[i,]  _apply (k0, 2, 'sum')
       h[i+1,]_-(v[i,]-g)/drv
       f[i+1,]_f[i,] +h[i+1,]
        }
       r_g -v[niter-1,]
    bf_list (x[1,], h, f, v, k1, r)
    names(bf)_c('x', 'h', 'f', 'g', 'k1', 'r')
    return (bf)
   }
```

Appendix D

S Code for The Coditional Expectation Algorithm for the Tolerance Limit Problem

The following function is written in the *S* programming language (Becker, Chambers and Wilks, 1988). The function *gauleg* calls a FORTRAN routine as documented in Appendix C. The function *dtnc* calls a FORTRAN subroutine to determine the noncentral-t density and cummulative (Lenth, 1988)

```
#
#    Mark Vangel, September 1990
#
#   Function to determine critical values for tolerance limit problem.
#
tw_function (niter=10, i=5, j=5, npt=25, nquad=25, p=.9, g=.95,
             k0=welch (r, p, g, i, j, accel=F),
             kfact=3.406632){
#
#    -- Degrees of freedom between, within, total
     df1_i-1
     df2_i*(j-1)
     df _df1+df2
#
#  -- Matrices of successive corrections; sums of corrections;
#     and approximate r.h.s.
   h _matrix (0, niter, npt)
   f _h
   v _h
#
#    -- Normal quantile, beta function.
   z  _qnorm (p)
   con_lgamma((df1+df2)/2) -lgamma(df1/2) -lgamma(df2/2)
#
#    -- Gauss-Legendre abscissas and weights
   gpt_gauleg(nquad, 0, 1)
```

```
  x  _matrix(gpt$x, npt, nquad, byrow=T)
#
#   -- Nuisance parameter values
  r  _gl$x/(1-gl$x)
  tau_matrix(r, npt, nquad) +1
#
#  -- Limiting value for f
  klim _kfact
  f[1,]_k0
#
#  -- Iterate quasi-Newton algorithm
  for (i in 1:niter) {
#
#  -- Determine maximum value of argument of f
    xm   _x[1,npt]
    tm   _tau[npt,1]
    rm   _tm *df2 *xm /(df1 *(1-xm))
    rlim _rm
#
#  -- Interpolate f
    msr _tau *df2 *x /(df1 *(1-x))
    k   _array(approx (c(r,rlim),
                 c(f[i,],klim),
        c(msr), rule=2)$y, dim(msr))
#
#  -- Evaluate V_1 and find the peak
    arg _k *sqrt ((df1+df2)*(i*x/(i-1) +(1-x)/tau))
    ncp _z *sqrt (i*(1 +(j-1)/tau))
    u   _dtnc(arg, ncp, df1+df2)
    beta_exp(con +(df1/2-1)*log(x) +(df2/2-1)*log(1-x))
    browser()
    v1  _ beta *u$dens *arg /k
    v[i,] _ (beta *u$cdf) %*% gpt$w
    vmax_apply(v1,1,'max')
    vmax_apply(v1<=vmax, 1, 'sum')
    xmax_x[10,vmax[10]]
    browser()
#
#  -- Transform the nuisance parameter estimate
    tau2_tau *df1 *(1-xmax)/(df2 *xmax)
#
#  -- Determine new maximum value of argument of f
    xm   _x[1,npt]
    tm   _tau2[npt,1]
    rm   _tm *df2 *xm /(df1 *(1-xm))
    rlim _rm
#
#  -- Interpolate f again
    msr _tau2 *df2 *x /(df1 *(1-x))
    k   _array(approx (c(r,rlim),
                 c(f[i,],klim),
        c(msr), rule=2)$y, dim(msr))
#
#  -- Calculate next step
```

```
      arg _k *sqrt ((df1+df2)*(i*x/(i-1) +(1-x)/tau2))
      ncp _z *sqrt (i*(1 +(j-1)/tau2))
      u   _dtnc(arg, ncp, df1+df2)
      beta_exp(con +(df1/2-1)*log(x) +(df2/2-1)*log(1-x))
      v1  _ (beta *u$dens *arg /k) %*% gpt$w
      v0  _ (beta *u$cdf) %*% gpt$w
#
      h[i,]  _(g -v0) /v1
      f[i+1,]_f[i,] +h[i,]
      }
#
#    -- Return arg of k, old k, new k, F(k)
      tw_list (h, f, v)
      names(tw)_c('h', 'f', 'g')
      return(tw)
  }

   welch_function(msra, p=.9, g=.95, k=5, l=5) {
#
#  Welch-Aspin series critical value for tolerance limit problem
#
     xkp _qnorm(p)
     xkg _qnorm(g)
     len _length(msra)
     msr _array (msra, len)
     n   _k*l
     t1  _sqrt (1/(1+(l-1)/msr))
     t2  _sqrt (1/(msr^2 +(l-1)*msr))
     rtk _sqrt (k)
     rtn _sqrt (n)
     xl1 _1/l^2
     xl2 _((l-1)/l) ^2
#
      xk _xkp +t1/rtn *(xkg +1/(4*(k-1)) *(
                xkg *(xkg*xkg +1)          +xkp*xkp*xkg *n *t1*t1 *xl1
               +xkp *rtn *t1*t1*t1 *xl1     +xkp*xkg*xkg *rtn *t1 /l)
                             +1/(4*k*(l-1)) *(
               +xkp*xkp* xkg *n *t2*t2 *xl2  +xkp *rtn *t2*t2*t2 *xl2))
#
      idx_sum((1:length(xk))* (xk==min(xk)))
      if (idx > 1) xk[1:idx-1]_xk[idx]
      if (sum (dim (msra)) != 0) xk_array (xk, dim(msra))
      return (xk)
      }
   dtnc_function (ta, ncpa, df)
    {
#
#  Mark Vangel, Sept. 1990
#
#  Noncentral-t density and cdf
#
   n    _length(ta)
   tnc  _array(0,    n)
   fault_array(0,    n)
   t    _array(ta,   n)
   ncp  _array(ncpa, n)
#
   z1_.Fortran("tnc1", as.double(t*sqrt((df+2)/df)),
                      as.double(df+2),
                      as.double(ncp),
                      fault=as.integer(fault),
                      tnc=as.double(tnc),
                      as.integer(n))
#
   z2_.Fortran("tnc1", as.double(t),
                      as.double(df),
                      as.double(ncp),
                      fault=as.integer(fault),
                      tnc=as.double(tnc),
                      as.integer(n))
#
  p_df/t *(z1$tnc-z2$tnc)
  if (sum(dim(ta)) != 0) n_dim(ta)
  p     _array(p, n)
  tnc   _array(z2$tnc,n)
  fault1_array(z1$fault,n)
  fault2_array(z2$fault,n)
  z_list(tnc, fault)
  z_list(p, tnc, fault1+fault2)
  names(z)_c('dens', 'cdf', 'fault')
  return (z)
#
    }
```

Appendix E

S Code for Conditional Expectation Algorithm for Random Sphere Problem

The following function, written in the *S* programming language, (Becker, Chambers and Wilks, 1988), can be used for solving the integral equation (7.13). The function ***gauleg*** calls a FORTRAN routine to determine Gauss-Legendre quadrature abscissas and weights, and is documented in Appendix C.

```
spheres_function(rhs, npt=25, niter=25, xl=0, xh=1) {
#
#   Mark Vangel Nov 1990
#
#      Stereology problem
#
    h _matrix (0, niter, npt)
    f _h
    v _h
    gl_gauleg (npt, xl, xh)
    x _matrix (gl$x, npt, npt, byrow=T)
    y _t(x)
    w _matrix (gl$w, npt, npt)
#
    k _x *(xh-x)/ sqrt (y**2*(xh-x)**2 +2*x*y *(xh-x))
#
    f[1,]_0
    fx   _c(xl, gl$x, xh)
    d    _gl$x *(log(xh +sqrt(xh -gl$x**2)) -log(gl$x))
    tp   _c ((1-x)*y +x)
#
    for (i in 1:(niter-1)) {
       fy     _c(0, f[i,], 0)
       fz     _approx (fx, fy, xout=tp, rule=2)
       v[i,]  _apply (k*w*matrix(fz$y, npt, npt), 1, 'sum')
       h[i,]  _(rhs-v[i,]) /d
       f[i+1,]_f[i,] +h[i,]
       f[i+1,]_f[i+1,]*(f[i+1,]>0)
    }
#
   spheres_list (h, f, v, rhs, k)
   names(spheres)_c('h', 'f', 'g', 'u', 'k')
   return (spheres)
   }
```