Modeling of Data Sets Using a Continuum of Beta Distributions

by

Sandip Roy

B.S., University of Illinois at Urbana-Champaign (1998)

Submitted to the Department of Electrical Engineering and Computer Science in partial fulfillment of the requirements for the degree of Master of Science in Electrical Engineering and Computer Science at the Massachusetts Institute of Technology December 1999

© 1999 Massachusetts Institute of Technology. All rights reserved.

Author

Department of Electrical Engineering and Computer Science

December 13, 1999

Certified by___

B. C. Lesieutre

Associate Professor of Electrical Engineering

Thesis Supervisor

Accepted by____

Arthur C. Smith

Chairman, Department Committee on Graduate Students
Modeling of Data Sets Using a Continuum of Beta Distributions

by

Sandip Roy

Submitted to the Department of Electrical Engineering and Computer Science on December 13, 1999 in partial fulfillment of the requirements for the degree of Master of Science in Electrical Engineering

Abstract

The primary objective is the development of models to describe a data set using a continuum of beta distributions. In order to estimate the set of distributions, a parametric description for the distributions is chosen. Several different parametric models are considered in order to adequately describe various types of data, including time series.

Three methods are then utilized to estimate the parameters of the distribution. First, a maximum likelihood estimate for the parameters is formulated. Unfortunately, the estimated parameters are significantly biased, so standard maximum likelihood estimation fails. However, altered versions of the maximum likelihood estimator are more accurate. Next, the mean and the second, third, and fourth central moments of the distributions are parametrized, and least squares regression analysis is used to determine each moment. The regression results for the higher moments are generally biased; a process for correcting this bias is derived. Also, a technique for estimating the set of distributions by minimizing the error between the real and estimated cumulative distribution functions is considered. A technique of hypothesis testing for excess parameters and a test for model sufficiency are considered in this thesis. Additionally, methods for analyzing the adequacy of a parametric model are discussed. A straightforward generalization of the model to describe data as a function of multiple independent parameters is described. We also briefly explore techniques for optimal estimation of the mean of a beta distribution and calculate relationships between beta distribution parameter and moment estimates.

Finally, we consider the question of interpolating a continuum of beta distributions from a few known distributions. Both band-limited and linear interpolation of moments and cumulative distribution functions are attempted.

Thesis Supervisor: B. C. Lesieutre
Title: Associate Professor of Electrical Engineering
Acknowledgements

Most of all, I would like to thank my advisor, Dr. Bernie Lesieutre, for his advice and support throughout the project, as well as his dedication to my research. The part of the project that I enjoyed most was the many long afternoons that Dr. Lesieutre and I spent contemplating and discussing every aspect of my research and many, many other interesting things.

I would like to thank Mr. Thomas Neff of Daimler-Benz for his help in Discussing the various aspects of the research and for his many constructive suggestions in regards to this project. I thank him also for formulating the problem that led to this thesis topic and hope that these research results are useful to him for his Ph.D. thesis. Thanks also to Mr. Neff, as well as Dr. Michael Kokes and Ms. Heike Voss, for organizing a trip to Frankfurt, Germany for me, and allowing me to discuss the research results with the group there.

Next, I would like to thank the other members in the power systems group in LEES, for their friendship, help, and interest in an amazing variety of research.

A big thanks to Ms. Vivian Misuzo, the secretary in LEES, for her help with everything that I needed, and for helping to introduce me to MIT.

I also thank a close friend, David Shear, for helping out whenever I needed, showing me around Boston when I was new here, and making life here interesting and exciting for me.

Finally, thanks to my wonderful family: my parents, Dilip and Sikha Roy, and my brother Kaushik, for their support and love. This work was only possible because of everything that I have learned from the three of you.

Dedication

To all the teachers that I have ever had.
## Contents

1. **Introduction** 11  
   1.1. Motivation 11  
   1.2. Scope of the Research 12  
   1.3. Implementation of the Research Results 14  
   1.4. Mathematical Description of the Problem 15  
   1.5. Outline of the Thesis 16

2. **Background-Probability Theory** 19  
   2.1. Attributes of Random Variables 20  
      2.1.1. Distribution Functions for Random Variables 20  
      2.1.1. Moments of Random Variables 21  
   2.2. Attributes of the Beta Random Variable 22  
      2.2.1. Distributions and Moments of the Beta Random Variable 23  
      2.2.1. The Beta Distribution as a Pearson Curve 25  
   2.3. An Introduction to Estimation Theory 28  
      2.3.1. The Definition of an Estimator 28  
      2.3.2. Estimation of Parameters 28  
      2.3.3. Accuracy of Parameter Estimates 29

3. **Maximum Likelihood Estimation** 31  
   3.1. Background on Maximum Likelihood Estimation 31  
      3.1.1. Rationale for Maximum Likelihood Estimation 32  
      3.1.2. Background on Maximum Likelihood Estimation of Beta Distributions 32  
   3.2. Maximum Likelihood Estimation of a Continuum of Beta Distributions 33  
      3.2.1. Two Beta Distribution Models 34  
      3.2.2. Maximum Likelihood Parameter Estimates for the Models 36  
         3.2.2.1. Analytical Simplification of the Maximum Likelihood Function 37  
         3.2.2.2. Numerical Solutions to the Maximum Likelihood Equations 38  
      3.2.3. Cramer-Rao Bounds 39  
      3.2.4. Maximum Likelihood Estimation Fails 40  
      3.2.5. Asymptotic Properties of the Maximum Likelihood Estimator 42  
   3.3. Altered Maximum Likelihood Techniques 42  
      3.3.1. Local Maxima as the Maximum Likelihood Estimate 42  
      3.3.2. Stability Constrained Maximum Likelihood 43  
      3.3.3. Endpoint Constrained Maximum Likelihood 47  
         3.3.3.1. Unbiased Estimation of Endpoints 48  
         3.3.3.2. Constrained Maximum Likelihood Based on Unbiased Estimation of Endpoints 51  
      3.3.4. Maximum Likelihood with Probability Truncation 53
3.3.5. Maximum Likelihood Using the Edgeworth Approximation

3.3.5.1. The Edgeworth Approximation to a Distribution

3.3.5.2. Maximum Likelihood with a Fourth-Order Edgeworth Approximation

Appendix 3.1

4. Regression Analysis of the Moments

4.1. A Synopsis of Linear Regression

4.2. Regression of the Moments Based on a Single Independent Parameter

4.2.1. Model Formulation and Conceptual Overview

4.2.2. Regression of the Mean: Polynomial Regression Based Solely on the Independent Parameter

4.2.3. Regression of the Variance: Polynomial Regression Based Solely on the Independent Parameter

4.2.4. Regression of the Skewness: Polynomial Regression Based Solely on the Independent Parameter

4.2.5. Regression of the Kurtosis: Polynomial Regression Based Solely on the Independent Parameter

4.2.6. Regression of Moments for the Estimation of Beta Distributions

4.3. Regression of Moments Based on an Independent Parameter and on Other Data Samples.

4.3.1. Why Are Time Series Different?

4.3.1.1. A Rationale for a Different Method-Introduction

4.3.1.2. A Rationale for a Different Model-Two Reasons

4.3.1.3. The Philosophical Difference Between Time Series and Other Data

4.3.2. Two Models for Time Series Analysis

4.3.2.1. Removal of Independent Parameter-Dependent Information from the Time Series

4.3.2.2. Regression Models for the Transformed Time Series: Introduction

4.3.2.3. Analysis of the First Model

4.3.2.4. Analysis of the Second Model

4.3.2.5. Estimation of Silicon Prices

4.3.2.6. Interpolation using the Two Time Series Models

Appendix 4.1

Appendix 4.2

5. Regression Using Cumulative Distribution Function Fitting

5.1. Determination of a Single Cumulative Distribution Function

5.1.1. Model Formulation and Cumulative Distribution Function Fitting Technique

5.1.2. Formulation of Error Bounds

5.1.3. Asymptotic Properties of the Error Bounds
5.1.4. General Properties of the Cumulative Distribution Function Matching Estimator 124
5.1.5. A Computationally Simpler Estimator 127
5.1.6. Bounds on Beta Distribution Parameters 129
5.1.7. Maximum Likelihood Cumulative Distribution Estimation 131
5.2. Beta Distribution Modeling of Data Through Multivariate Cumulative Distribution Estimation. 133
5.2.1. Description of the Technique 133
5.2.2. Models for Estimating Beta Distribution Continua 135

6. Interpolation of Beta Distributions 141
6.1. Interpolation of the Cumulative Distribution Function 141
   6.1.1. Band-limited Interpolation of Cumulative Distribution Functions 142
   6.1.2. Linear Interpolation of Cumulative Distribution Functions 147
6.2. Interpolation of Distribution Parameters 152
   6.2.1. Interpolation of Beta Distributions using Moment Samples 153

7. Improvements and Generalizations in Estimation 159
7.1. Model Selection Using the Phase Spectrum 159
7.2. Data Estimation as a Function of Several Independent Parameters 161
   7.2.1. Choice of Basis Functions for the Multivariate Problem: An Overview 162
   7.2.2. Some Thoughts on Estimation and Estimates in the Multivariate Problem 164
7.3. Relationships between Parameter Estimate Variances 166
   7.3.1. Linearization of the Moment Equations: Calculation of Moment Variances 167
   7.3.2. Linearization of the Beta Parameter Equations: Calculation of Parameter Variances 168
7.4. An Improvement of the Mean Estimate 170
   7.4.1. Likelihood Mean Based Estimator 171
   7.4.2. Application of the Estimator for Improvement of the Mean Estimate 172

8. Conclusions and Future Work 173
8.1. Summary and Conclusions 173
8.2. Future Work 175
   8.2.1. Future Work in Model Selection 175
   8.2.2. Future Work in Estimation 176
   8.2.3. Other Possibilities for Future Work 177
8.3. Final Thoughts 177
List of Figures

Figure 1.1.
The approach used to model data sets with a set of distributions is presented. 13

Figure 2.1.
This figure depicts the probability density functions and associated cumulative distributions for three beta random variables. 26

Figure 2.2.
Each normalized skewness and normalized kurtosis pair correspond to a specific Pearson curve. The beta distribution is the Type 1 Pearson curve. 27

Figure 3.1.
The possibility for arbitrarily probability density functions results in the failure of the maximum likelihood estimation technique. 41

Figure 3.2.
Local maximization of the likelihood function provides reasonable estimates for sufficiently large data sets. 44

Figure 3.3.
By scaling a beta distribution to the unit interval, the expected distance between the smallest and largest points and the endpoints can be determined. 49

Figure 3.4.
Endpoint constrained maximum likelihood is successful in estimating distributions for data with statistics that do not depend on the independent parameter. 53

Figure 3.5.
This figure highlights the truncation of a probability distribution. In this case, the distribution is truncated at a probability value of 0.4. 55

Figure 3.6.
This plot is an example of an Edgeworth approximation based likelihood function. 57

Figures 4.1, 4.2.
These two figures demonstrate polynomial regressions to determine the mean of 100 and 20 point data sets, respectively, which were generated from the same model. 70, 71
Figure 4.3.
Regression of bias-corrected sample variances provides an estimate of the variance of the data. 81

Figure 4.4.
An extrapolated estimate for the normalized skewness is shown for the two example data sets. 87

Figure 4.5.
An extrapolated estimate for the normalized kurtosis is shown for the two example data sets. 89

Figure 4.6.
Comparison of probability and cumulative distribution functions at x=0 for the regression result (100 point example). 91

Figure 4.7.
Estimation of silicon prices using the first model. 105

Figure 4.9.
Estimation of silicon prices using the second model. 106

Figure 4.10.
This figure compares extrapolation results of silicon prices using the second model with two different filter lengths. 108

Figure 5.1.
Data generated from a uniform distribution on the interval [0,1] is plotted along with the corresponding bounds on the actual cumulative distribution based on the observed data. 119

Figure 5.2.
This figure is a diagram of the example that shows the possibility for arbitrarily large bias using the cumulative distribution function-based estimator. 125

Figure 5.3.
Figure 5.3a, 5.3b, and 5.3c demonstrate beta cumulative distribution fitting to a 20 point sample cumulative distribution. 127

Figure 5.4.
This figure illustrates estimation using multivariate cumulative distribution function fitting. 138
Figure 6.1.
This figure demonstrates the possible failure of band-limited cumulative distribution function fitting given an insufficient number of samples. 144

Figure 6.2.
Band-limited cumulative distribution function interpolation. 147

Figure 6.3.
This figure depicts linear cumulative distribution function interpolation. 148

Figure 6.4.
Linear Interpolation of Cumulative Distribution Functions. 149

Figure 6.5.
Possible paths of interpolation of the higher moments are shown. 157

Figure 6.6.
Interpolation of distributions using moment regression is shown. 158

Figure 7.1.
Plots of silicon price data and associated phase spectra before and after independent parameter-based information has been removed. 160

Figure 7.2.
The curve of best fit for the mean of a set of data that varies with two parameters is shown. Additionally, a regression curve for the mean is shown for a model that uses only one of the two parameters to describe the data. 166
Chapter 1

Introduction

1.1. Motivation

The specific motivation for this research is the development of beta distribution models to describe a variety of different data sets in which the data vary as a function of one or more independent parameters and contain a random component. For example, models are formulated to describe the price of a motor as a function of the power generated by the motor or to describe the price of copper on the metal exchange as a function of time. The objective of the study is the description of the dependent variable, such as price, at each value of the independent parameter with a beta random variable. These models will be incorporated in a system for estimating production prices and profits. The system, developed at DaimlerChrysler, is meant to facilitate prediction of costs. Specifically, beta random variables are used to describe the costs of elementary components, such as the price of individual raw materials or the salary of an employee. These beta random variable models, which sometimes must be formulated from data sets, are then combined to estimate the total costs or profits of a project. The system will be used to estimate the financial effect of project management decisions in the early stages of the project and so to facilitate the decision-making and design processes.

More generally, the motivation for the project is the need for developing techniques that allow adequate estimation of arbitrary random quantities. A cohesive theory exists for the estimation of certain classes of tractable random variables, including
Gaussian random variables. However, the estimation of parameters in other probabilistic
descriptions of data is much less developed. Additionally, because both the predictable
and the random component of the data can vary with an independent parameter, this
study focuses on developing probabilistic models that vary as a function of one or more
parameters.

1.2. Scope of the Research

Based on the motivation for the project, our primary objective is the development
and analysis of models that describe data sets, which may vary with one or more
independent parameters, using a continuum of beta distributions. The scope of the
project includes the development of these models, the analysis of methods for
determining the unknown quantities in the model, and the formulation of techniques for
testing the validity of the model (see Figure 1.1).

We subdivide the modeling problem based on techniques used to estimate the
parameters of the model. We consider maximum likelihood estimation of beta
distribution parameters, regression of the moments, and cumulative distribution function
fitting as estimation methods and also briefly discuss other procedures that are useful for
improving parameter estimates. Some of the proposed parametric models can be
estimated using several estimation methods, while the parameters of other models can
only be determined based on one of the estimation techniques.
Figure 1.1.
The approach used to model data sets with a set of distributions is presented.

The parametric models are designed to adequately describe a variety of data sets. Some of the models describe the data points using a set of independent beta random variables that are solely functions of the independent parameter. Other models do not attempt to describe the deterministic component of the data set solely as a function of the independent parameter value and instead also account for correlation between data samples. The possibility of using these models for interpolation and extrapolation with respect to the independent parameter is also within the scope of the project.

An important component of the research is the development of techniques for determining the adequacy of the developed models. Several techniques are utilized to determine whether the chosen model sufficiently describes the data and whether the model captures the predictable component of the data.
In addition to the primary task of developing and analyzing models to describe data sets, a potpourri of other problems are discussed. We briefly consider techniques for modeling data as a function of multiple independent parameters. We also consider the interpolation of a set of known distributions. For example, if a beta distribution model is known for the data at two independent parameter values, we consider the problem of determining beta distribution models for intermediate parameter values.

More philosophically, the events and processes surrounding us are partially predictable based on known information and are partially random. This thesis explores a few specific methods for extracting the predictable information in real processes and events.

1.3. Implementation of the Research Results

The completed project work also includes the development of Matlab software which implements the modeling and parameter estimation described in this thesis. In developing the software, we specifically attempt to utilize general models for describing data sets. Further, the parameter estimation techniques that are programmed are chosen not only for their accuracy but also for their intuitive simplicity. We have developed a complete program for the modeling of data sets based on the regression analysis method for parameter estimation, and we have also developed demonstration procedures for the other parameter estimation methods. Further, a Matlab program that compares several methods for the interpolation of a set of known beta distributions is produced.
1.4. Mathematical Description of the Problem

We consider beta random variable models that describe the set of data points $(x_1, y_1)\ldots(x_N, y_N)$. Specifically, we describe the dependent observations $y_j$ as an outcome of a four parameter beta random variable $Y_i$ that depends on $x_i$ and $y_j$, $j \neq i$.

More mathematically, we model the data using the $N$ random variables

$$Y_i = \beta(x_i, y_j; \vec{p}), j \neq i,$$

where $\vec{p}$ is a vector of parameters that specifies the model for a specific set of data. In order to estimate the parameter vector $\vec{p}$, we utilize and develop a variety of different random process parameter estimation techniques, including maximum likelihood estimation, regression analysis of the moments of the data, and estimation through cumulative distribution function matching, as well as more general minimum variance unbiased estimation processes. These techniques utilize the set of observations $y_1\ldots y_N$ and the corresponding independent parameter values $x_1\ldots x_N$ to compute optimal values for the parameters. The efficacy of an estimator is tested by determining the difference between the parameter estimates and the actual parameters based on data generated from a known parametric model.

Next, the utility of the parametric model must be determined. A parametric model is useful only if parameters can be chosen so that both the predictable and the random information in a data set can accurately be captured using the model. Various types of hypothesis testing are utilized to verify model sufficiency. The chosen models are also appropriate if they can accurately describe the distribution of the dependent variable for $x \not\in x_1\ldots x_N$. In other words, the accuracy of the interpolated and extrapolated
distributions generated by a certain parametric model is an important characteristic of that model.

Several extensions of the modeling and estimation problem are briefly considered. We generalize some of the models and their respective estimation techniques to multivariate problems. Specifically, we consider models of the form

$$Y_i = \beta(\bar{x}_i, y_j; \bar{p}), \ j \neq i$$  \hspace{1cm} (1.2)

and develop techniques for their estimation. We also consider the interpolation of known distributions. For example, if the random variables $Y_i$ and $Y_j$, where $x_i < x_j$, are exactly known, an estimate for the distribution of the random variable $Y_k$, $x_i < x_k < x_j$ is formulated. More generally, we attempt to determine a continuum of distributions based on a set of $N$ known distributions.

1.5. Outline of the Thesis

The thesis is organized to match the order in which research was completed. We do not necessarily present the most general models and successful estimation techniques first; instead, we develop simple models and estimation techniques, analyze these models and techniques, and then suggest improvements.

Chapter 2 is an introductory chapter. We describe the properties of the beta random variable, introduce the concept of estimation, and state other necessary probabilistic concepts.
Chapter 3 focuses on maximum likelihood estimation. We formulate two beta random variable models for describing data sets and then estimate the unknown parameters of the model using the maximum likelihood technique. We describe the shortcomings of the maximum likelihood estimation of beta distributions and propose several adjustments to the estimation process. Simulation results and some analytical justification for the improvements are provided. In the context of maximum likelihood estimation, we also discuss the Cramer-Rao bounds for parameter estimates.

Chapter 4 is a description of the moment regression technique for parameter estimation. We utilize moment regression to calculate the parameters of one of the models described in the previous chapter. We also formulate two new models that are especially adept in describing time series data and apply moment regression to determine the parameters of the model. The correction of bias in the higher central moment estimates is a significant problem analyzed in the chapter. We also consider general methods for hypothesis testing of the chosen model and explore interpolation and extrapolation using the chosen models.

Chapter 5 focuses on the cumulative distribution function matching method for estimation. The first half of the chapter describes the estimation of a single beta random variable; previous work in the estimation of cumulative distribution functions is presented, and some new properties of cumulative distribution function matching estimation are also suggested. The second half of the chapter focuses on the estimation
of the parameters of one of the models presented in Chapter 3 using a multivariate cumulative distribution approximation approach.

Chapter 6 discusses interpolation between two exactly known beta random variables. We develop and analyze techniques based on band-limited and linear interpolation of cumulative distribution functions. We also consider interpolation of beta distributions based on interpolation of the first four moments of the distribution. Both analytical and experimental comparisons between the methods are given.

Chapter 7 incorporates a variety of different concepts. We outline the problem of modeling data as a function of several independent parameters and briefly describe techniques for estimating the parameters of these models. Next, we compare the estimation techniques described in the previous chapters using both numerical and analytical methods. A compelling estimator that provides minimum variance unbiased estimates for some problems is presented. Finally, we discuss hypothesis testing in a more general context.

Chapter 8 concludes the thesis by summarizing the findings and results of the research and proposing further work in the estimation of beta random variable continua.
Chapter 2

Background-Probability Theory

Throughout this study, we characterize sets of data using four parameter beta distributions. In order to determine techniques for estimating this continuum of distributions, a detailed understanding of the beta distribution is useful. Because the analysis of random variables and specifically of the beta random variable is central to the work, this chapter is devoted to the description of random variables and probability distributions. First, important characteristics of arbitrary random variables are defined. Next, the probability distribution of the beta random variable is given, and the specific properties of the beta random variable are highlighted. Finally, the chapter explores the beta distribution within the context of the Pearson family of curves. Much of this background information is also stated in the thesis by Trefz [1], who developed methods for estimating functions of beta random variables with other beta random variables.

Additionally, several general concepts in estimation theory are discussed in this chapter. The techniques for estimation are not discussed here and instead are presented in Chapters 3, 4, and 5. However, a general definition for an estimator is given, and general measures for determining the accuracy of an estimate are considered.
2.1. Attributes of Random Variables

The outcome of a process is often uncertain rather than deterministic. A random variable is the mathematical construct that is used to describe this uncertain outcome [2]. A random variable assumes certain real values, and a probability is assigned to each value that the variable can attain.

2.1.1. Distribution Functions for Random Variables

The cumulative distribution function of a random variable $Y$ is a non-negative, monotonically non-decreasing function $F(y)$. The cumulative distribution function represents the probability that the random variable $Y$ is less than or equal to $y$:

$$F(y) = P(Y \leq y).$$  \hfill (2.1)

The probability density function for $Y$ is defined as the generalized derivative of $F(y)$:

$$p_Y(y) = \frac{dF(y)}{dy}. \hfill (2.2)$$

A generalized derivative is necessary because the probability density function may include delta functions corresponding to steps in the cumulative distribution. The probability density function $f(y)$ is always non-negative for all $y$, and the integral of the function over all $y$ is unity. If the random variable only assumes a discrete set of values $y$, the probability mass function, defined as

$$g_Y(y) = P(Y = y) \hfill (2.3)$$

is a simpler representation of the random variable than the probability density function.
2.1.2. Moments of Random Variables

The expectation of a function of a random variable, \( h(Y) \), is defined by [2]

\[
E[h(Y)] = \int_{-\infty}^{\infty} h(y) p_Y(y) dy.
\]  

(2.4)

The moments of the random variable around zero are expectations of powers of the random variable [2]:

\[
u_r = E[Y^r] = \int_{-\infty}^{\infty} y^r p_Y(y) dy, \quad r \geq 1.
\]  

(2.5)

The first moment around zero is labeled as the mean.

Next, moments of the random variable around the mean are often useful measures of the random variable. The moments around the mean, labeled central moments, are

\[
u_r = E[(Y-u_1)^r] = \int_{-\infty}^{\infty} (y-u_1)^r p_Y(y) dy, \quad r \geq 1.
\]  

(2.6)

The central moment corresponding to \( r=1 \) is zero. The second, third, and fourth central moments are labeled as the variance, skewness, and kurtosis, respectively. Incidentally, the definitions of skewness and kurtosis vary throughout the literature [3]. Specifically, some authors label \( \frac{u_3}{u_2^{\frac{3}{2}}} \) and \( \frac{u_4}{u_2^{\frac{2}{2}}} \) as the skewness and kurtosis, respectively. To avoid confusion, we label

\[
\sqrt{\beta_1} = \frac{u_3}{u_2^{\frac{3}{2}}}
\]  

(2.7a)

and

\[
\beta_2 = \frac{u_4}{u_2^{\frac{2}{2}}}
\]  

(2.7b)
as the normalized skewness and normalized kurtosis, respectively, according to the symbolic notation given in [3].

Further, the central moments of a random variable are easily described as functions of the moments about zero of that random variable. The first central moment of a random variable is always zero. Expressions for the second, third, and fourth central moments as functions of the moments about zero are given:

\[ u_2 = u_2 - u_1^2 \]  \hspace{1cm} (2.8a)
\[ u_3 = u_3 - 3u_2u_1 + 2u_1^3 \]  \hspace{1cm} (2.8b)
\[ u_4 = u_4 - 4u_3u_1 + 6u_2u_1^2 - 3u_1^4 \]  \hspace{1cm} (2.8c)

2.2. Attributes of the Beta Random Variable

Random variables are characterized by the structures of their probability density functions. In this work, the probability density functions that describe the studied data sets are unknown and must be deciphered from the data. Although arbitrary probability density functions may be chosen to describe unknown random variables, a parametric form for the probability density functions facilitates their estimation. We utilize a four parameter distribution known as the beta distribution to describe the random variables that are estimated.

The beta distribution is chosen for its versatility. The probability density function is non-zero in only a finite interval, an appropriate description for many physical processes. Further, the probability density function may be either unimodal or bimodal. Also, the set of all beta random variables encompasses or is limited by many other
random variables, including the uniform random variable, the exponential random variable, the Gaussian random variable, and the gamma random variable [4].

2.2.1. Distributions and Moments of the Beta Random Variable

Two parametric forms are generally used to represent the probability density function of a beta random variable. First, the beta probability density function is uniquely defined using two endpoint parameters and two shape parameters. Specifically, the probability density function for a beta random variable $Y$ is

$$p_Y(y; \alpha, \beta, a, b) = \frac{\Gamma(\alpha + \beta)}{(b-a)^{\alpha + \beta - 1} \Gamma(\alpha) \Gamma(\beta)} (y-a)^{\alpha-1} (b-y)^{\beta-1}, \quad (2.9)$$

for $a < y < b$ and zero otherwise. The function $\Gamma(x)$ is defined as

$$\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} dt, \quad x > 0. \quad (2.10)$$

The parameters $a$ and $b$ represent the endpoints of the probability density, while $\alpha$ and $\beta$ are strictly positive shape parameters.

Alternately, a beta random variable is uniquely specified by four moments. McNichols [5] describes the relationship between the moments and the endpoint and shape parameters. The relationship of the mean, variance, skewness, and kurtosis to the endpoint and shape parameters is given by

$$u_1 = a + (b-a) \frac{\alpha}{\alpha + \beta} \quad (2.11a)$$

$$u_2 = (b-a)^2 \left[ \frac{\alpha \beta}{(\alpha + \beta + 1)(\alpha + \beta)^2} \right] \quad (2.11b)$$
Additionally, the shape and endpoint parameters of a beta distribution with known moments can be determined [6]. First, the shape parameters are determined as

\[
\alpha, \beta = \frac{1}{2} r \{ 1 \pm (r + 2) \sqrt{\beta_1 \{(r + 2)^2 \beta_1 + 16(r + 1)\}^{-1}} \} \tag{2.12a}
\]

where

\[
r = \frac{6(\beta_2 - \beta_1 - 1)}{6 + 3\beta_1 - 2\beta_2} \tag{2.12b}
\]

and \( \beta > \alpha \) if \( u_3 > 0 \), \( \beta \leq \alpha \) otherwise.

Next, the difference between the two endpoints can be determined as

\[
c = b - a = \sqrt{u_2 \frac{(\alpha + \beta + 1)(\alpha + \beta)^2}{\alpha + \beta}}. \tag{2.12c}
\]

Finally, the two endpoints are determined as

\[
a = u_i - c \frac{\alpha}{\alpha + \beta}. \tag{2.12d}
\]

and

\[
b = a + c. \tag{2.12e}
\]

The expression for the probability density of a beta random variable can be written directly as a function of four moments using the described relationships between the moments and the shape and endpoint parameters. However, the expression based on the moments is cumbersome and does not highlight the shape of the distribution. Instead, the representation of the beta random variable based on its moments is important because
estimation of the mean and higher central moments from data is sometimes straightforward and so allows a unique estimation of a beta random variable.

The values of the shape parameters, correspond to certain attributes of a beta random variable. If either of the shape parameters is less than one, then the probability density function becomes unbounded for at least one value of \( y \) (Figure 2.1a, 2.1c). Also, if exactly one of the two shape parameters is less than one, then the probability density function is unimodal, with the mode corresponding to one of the endpoints (Figure 2.1a). If both shape parameters are less than one, then the probability density is bimodal (Figure 2.1c). If both shape parameters equal one, the probability density function is uniform, and if both shape parameters are greater than or equal to one, with at least one parameter greater than one, the distribution is unimodal (Figure 2.1e).

Unfortunately, no closed form expression for the cumulative distribution function exists, but the probability density function can be numerically integrated to approximate the cumulative distribution. Graphs of cumulative distribution functions for certain parameter values are shown (Figure 2.1b,d,f).

2.2.2. The Beta Distribution as a Pearson Curve

The beta distribution is a member of the family of curves developed by Karl Pearson [5]. The Pearson curves are typically subdivided into twelve sets of distributions. Each Pearson curve is uniquely defined using four moments, and conversely each valid set of four moments corresponds to a Pearson curve. The beta distribution is labeled as the Type I Pearson curve.
Figure 2.1.
This figure depicts the probability density functions and associated cumulative distributions for three beta random variables.

The description of the beta distribution as a member of the Pearson curves is important in this study because the formulation of the Pearson curves constrains the moments of the beta distribution. The mean of a beta random variable may be any real number, and the variance can equal any positive real number, so that these two moments
are not constrained for the beta distribution. However, only certain values for the higher moments are valid for beta distributions, while other values for the higher moments correspond to other Pearson curves. This constraint in the higher moments is most easily described as a constraint on $\beta_2$ in terms of $\beta_1$:

$$1 + \beta_1 < \beta_2 < 3 + \frac{3}{2} \beta_1. \quad (2.13)$$

The regions of the $\beta_1 - \beta_2$ plane that correspond to the various Pearson curves is illustrative [3].

![Figure 2.2.](image_url)

Each normalized skewness and normalized kurtosis pair correspond to a specific Pearson curve. The beta distribution is the Type 1 Pearson curve [3].
2.3 An Introduction to Estimation Theory

Estimation theory is an intricate subject that has been studied extensively by many noted scientists and engineers. Before delving into the methods for estimating a continuum of beta distributions from data samples, we briefly discuss the general concepts of estimation theory to provide a context for the techniques described later.

2.3.1. The Definition of an Estimator

An estimator is a function that is utilized to determine an unknown quantity of interest based on another known quantity [6]. A variety of different quantities are estimated, including unknown constants and functions, as well as unknown probability density functions. The quantity that is estimated must somehow influence the known quantity that is used for estimation. The technique for estimation uses this correlation between the two quantities to generate an accurate estimate.

2.3.2. Estimation of Parameters

Often, the probability density of the known quantity is expressed as a function of the unknown values, or parameters, that will be estimated [6]. More mathematically, consider a set of parameters \( \vec{p} = [p_1...p_k]^T \) that must be estimated, as well as a set of observations that are instances of the random variables \( \vec{Y} = [Y_1...Y_N]^T \). A mathematical model is then formulated that describes the probability density of an arbitrary set of observations \( \vec{y} = [y_1...y_N]^T \):

\[
p_{\vec{p}} (\vec{y}; \vec{p}) = \Pr[Y_1 = y_1,...,Y_N = y_N; p_1,...,p_k].
\] (2.14)
The designed estimator then maps each set of observations to an optimal set of values for $p_1 \ldots p_k$. The estimate for the parameters is denoted as $\tilde{p}(\tilde{y})$.\hfill (2.15)

### 2.3.3. Accuracy of Parameter Estimates

Several standards are utilized to gauge the accuracy of parameter estimates, and these standards lead to the development of several optimal estimators [6]. Before discussing measures for the accuracy of parameter estimates, a distinction between two possible types of parameter estimation is important. In some problems, the parameters can naturally be classified as random variables with a quantifiable distribution function, and so the joint distribution of $\tilde{y}$ and $\tilde{p}$ can be used for the estimation of the parameters. In other problems, the parameters are regarded as non-random values with no prior probabilities.

If the parameters are considered random quantities with a specified distribution form, an overall measure of the accuracy of the estimator can be formulated. One element of the estimate's accuracy is bias vector, which is defined as

$$\vec{b}_{\text{est}} = E[\tilde{p}(\tilde{y}) - \bar{p}],$$  \hfill (2.16)

where the expectation is calculated over both the parameters and the observed variables $y$. In the case of random parameter estimation, the bias can simply be removed to generate an unbiased estimator, and so a good estimate for a parameter is always unbiased. Next, the covariance matrix of the parameter estimates is defined as

$$\Lambda_{\text{est}} = E[(\tilde{p}(\tilde{y}) - p)(\tilde{p}(\tilde{y}) - p)^T],$$  \hfill (2.17)

where the estimate is again calculated with respect to both the parameter and the observation variables. A variety of metrics based on the estimate covariance matrix and
on other expectations that depend on the estimate are used to determine the accuracy of an estimator and to select optimal estimators. The sum of the main diagonal elements of $\Lambda_{est}$, which represent the variances of each estimate, is often minimized. The elements in the expectation vector

$$\bar{\lambda}_{MAE} = E[\tilde{p}(\bar{y}) - p]$$

(2.18)

are also often used to determine the accuracy of the estimator. Other metrics for estimator performance will be discussed in the context of specific estimators.

The non-random parameter estimation problem is less straightforward. Since the parameters do not have an associated probability density function, the bias and variance are calculated as a function of the parameters. In this case, the bias is

$$\bar{b}_{est}(\bar{p}) = E[\tilde{p}(\bar{y}) - \bar{p}]$$

(2.19)

where the expectation is calculated only in terms of the observation variables. Similarly, the covariance matrix is actually a matrix of functions:

$$\Lambda_{est}(\bar{p}) = E[(\tilde{p}(\bar{y}) - p)(\tilde{p}(\bar{y}) - p)^T].$$

(2.20)

The development of unbiased estimators is much more difficult for the non-random parameter estimation problem. One possible standard for an optimal unbiased estimator is minimum variance at all parameter values, but no general technique is known for developing minimum variance unbiased estimators. Again, minimum variance is defined as the minimum possible sum of the diagonal elements of $\Lambda_{est}(\bar{p})$. Several other standards for the accuracy of the estimate can also be developed.

The concept of estimation and the standards for gauging the accuracy of an estimate are important in developing techniques to model data with a continuum of beta distributions.
Chapter 3

Maximum Likelihood Estimation

Maximum likelihood estimation is used to model a set of data with a continuum of beta distributions. At the beginning of this chapter, a brief overview of maximum likelihood estimation is provided. Next, two general parametric models are formulated to describe the sets of data considered in this study. The technique for determining the parameters of this model using maximum likelihood estimation is described. The lower bound on the variance of an estimator, the Cramer-Rao bound, is connected to the theory of maximum likelihood estimation, so these bounds are also calculated for the two described models. Unfortunately, maximum likelihood estimation of beta distributions is biased; several novel estimators that are similar to the maximum likelihood estimator are considered in an attempt to formulate an unbiased estimator.

3.1. Background on Maximum Likelihood Estimation

The set of parameter values that maximize the probability of the observation vector in a given parametric model is the maximum likelihood estimate [7]. Formally, the maximum likelihood estimate of the parameter vector for each observation \( \tilde{y} \) is

\[
\tilde{p}(\tilde{y}) = \arg\max_\tilde{p} \left( p_{\tilde{p}}(\tilde{y}; \tilde{p}) \right).
\] (3.1)

Alternately, the maximum likelihood estimate can be determined as

\[
\tilde{\tilde{p}}(\tilde{y}) = \arg\max_\tilde{\tilde{p}} \left( g(p_{\tilde{p}}(\tilde{y}; \tilde{p})) \right),
\] (3.2)
where \( g \) is any monotonically increasing function. For example, the natural logarithm of the probability density function is often used for the maximization. In this context, the model for the data, \( p_y(y; \bar{p}) \), is labeled as the likelihood function.

3.1.1. Rationale for Maximum Likelihood Estimation

The maximum likelihood estimator is widely used in non-random parameter estimation for several reasons. Intuitively, if no prior probabilities are known for the parameters, the selection of parameters that maximize the probability of the observed data set is sensible. Also, for non-random parameter estimation problems, the maximum likelihood estimate is the only estimate that can achieve the Cramer-Rao bound, which describes the minimum variance for any unbiased estimate [7]. If the maximum likelihood estimate achieves the Cramer-Rao bound, the estimate is termed an efficient estimate.

Though many maximum likelihood estimates are not efficient, the maximum likelihood technique is often used for inefficient estimation because of the simplicity of implementing the estimation procedure. Even if an analytical result for the maximum likelihood estimator cannot be derived for a parametric model, a numerical implementation of the maximization is possible. A thorough description of maximum likelihood estimation and the Cramer-Rao bounds is given in [7].

3.1.2. Background on Maximum Likelihood Estimation of Beta Distributions

Several researchers have considered maximum likelihood estimation of single four-parameter beta distributions and other similar distributions [8, 9]. Specifically, they
consider the estimation of the four beta distribution parameters based on observations of
N independent, identically distributed beta random variables with these parameters.

Carnahan [8] summarizes work completed in the estimation of four parameter beta
distributions and of other similar distributions. Much of the analytical work in the
estimation of beta distributions has focused on asymptotic properties of the estimator, or
the behavior of the estimator as the number of observations becomes large. According to
[8], discrepancies in the analysis of asymptotic bias and variance among several
researchers have not been resolved. This article also provides the Fisher information
matrix and the associated Cramer-Rao bounds for the estimation of four parameter beta
distributions. The studies of beta distribution parameter estimation typically constrain
the shape parameters to exceed one in order to eliminate distributions with an infinite
probability point. For estimation based on a finite number of observations and for shape
parameters constrained to be larger than one, simulations demonstrate that the maximum
likelihood estimate has some bias [8].

3.2. Maximum Likelihood Estimation of a Continuum of Beta
Distributions

The maximum likelihood estimation technique is used to model a set of data with
a continuum of beta distributions. In order to formulate the estimate, we first decide a
parametric form for the probability density of the data. In this chapter, we consider two
parametric forms for describing the data. For one of the parametric forms, equations that
can be solved for the maximum likelihood parameter estimates are presented. The
associated Fisher information matrix and Cramer-Rao bounds are generated.

Unfortunately, the estimates cannot be derived analytically, and so numerical procedures
for determining the parameters are discussed. Both analytical and simulation-based procedures for determining the accuracy of the estimate are presented; these procedures highlight a significant flaw with the maximum likelihood estimator.

### 3.2.1. Two Beta Distribution Models

The two beta distribution models presented for maximum likelihood estimation are useful for describing a variety of different data sets that vary as a function of an independent parameter. In this case, the vector of observations \( \vec{y}(x) \) consists of \( N \) coordinate pairs \((x_1, y_1) \ldots (x_N, y_N)\) rather than \( N \) scalars. Further, we expect that the \( y \) coordinate \( y_i \) may be approximated as a function \( f(x_i) \) of the \( x \) coordinate in each coordinate pair, and that the residuals \( y_i - f(x_i) \) for each coordinate pair are approximately independent random variables.

In the first model, the endpoint and shape parameters of the beta distributions are written as functions of the independent parameter \( x \). Each function is linearly scaled by a parameter that will be estimated using maximum likelihood. Specifically, the probability for a set of observations \( \vec{y}(x) \) is modeled as

\[
p_{\vec{y}(x)}(\vec{y}(x); \vec{p}) = \prod_{i=1}^{N} K(x_i)(y_i - a(x_i))^{\alpha(x_i)-1}(b(x_i) - y_i)^{\beta(x_i)-1},
\]

where

\[
K(x_i) = \frac{1}{(b(x_i) - a(x_i))^{\alpha(x_i)+\beta(x_i)-1} \beta(\alpha(x_i), \beta(x_i))},
\]

\[
\alpha(x_i) = \alpha(x_i; \vec{p}) = p_{a0} f_{a0}(x_i) + \ldots + p_{ak} f_{ak}(x_i),
\]

\[
\beta(x_i) = \beta(x_i; \vec{p}) = p_{b0} f_{b0}(x_i) + \ldots + p_{bk} f_{bk}(x_i),
\]
and the functions are appropriately chosen to describe the data. Although arbitrary sets of functions can be used for the maximum likelihood formulation, we choose a polynomial basis for the parameters, so that

\[
\alpha(x_i) = \alpha(x_i; \bar{p}) = p_{a0} + \ldots + p_{ak_i} x_i^{k_i}, \tag{3.5a}
\]

\[
\beta(x_i) = \beta(x_i; \bar{p}) = p_{b0} + \ldots + p_{bk_i} x_i^{k_i}, \tag{3.5b}
\]

\[
a(x_i) = a(x_i; \bar{p}) = p_{a0} + \ldots + p_{a_k} x_i^k, \tag{3.5c}
\]

\[
b(x_i) = b(x_i; \bar{p}) = p_{b0} + \ldots + p_{b_k} x_i^k. \tag{3.5d}
\]

In total,

\[
k = k_\alpha + k_\beta + k_a + k_b + 4 \tag{3.6}
\]

parameters must be estimated from \(N\) data samples. Polynomials are chosen to describe the endpoint parameters because many data sets are bounded by linear and other low order polynomial functions. Polynomial models for the shape parameters are more difficult to rationalize. In many situations, if the shape of the distribution changes, a linear change in the normalized skewness and kurtosis is a reasonable model for this change. The effect of such a change on the shape parameters is complicated but can be approximated by a polynomial model.

Alternately, the mean, variance, skewness, and kurtosis can be modeled as functions of the independent parameter \(x\), with the unknown parameters serving as coefficients to the functions. Again, the probability of a set of observations is modeled as a product of the probability of each individual observation. The expression for the
probability is not given because of the cumbersome transformation between moments and beta distribution parameters. The expressions relating the mean and the higher central moments to the independent parameter are analogous to the expressions used to describe the endpoint and shape parameters in the earlier model and so are not repeated. Again, arbitrary function bases relating the independent parameter to the \( y \) values are possible, but polynomial basis functions are again explored in detail.

A comparison and contrast between the two models is difficult. The two models can only be contrasted by generating data from each model and then calculating the bias and variance in the estimate using both the correct and the incorrect model to reconstruct the data sets. These calculations are analytically intractable and can only be approximated through Monte Carlo simulations.

As a side-note, the described models are insufficient for describing data with little functional dependence on the independent parameter, such as time series data. For these data sets, \( y_i \) must be defined as a function of \( x_i \) and of other variables, including other \( y \) coordinates, in order to produce independent residuals. These more general models are presented in the context of regression estimation, but parameter estimation through maximum likelihood techniques is also possible.

### 3.2.2. Maximum Likelihood Parameter Estimates for the Models

The unknown non-random parameters in the two models can be estimated using the maximum likelihood method. The maximum likelihood estimates cannot analytically be calculated for either model, and numerical methods are instead necessary. However, the maximum likelihood problem can be simplified to a system of \( k \) equations involving
the k unknown parameters. This further analysis of the maximum likelihood problem is shown for the first model, and the numerical procedures used to find the maximum likelihood estimate are described for both models.

3.2.2.1. Analytical Simplification of the Maximum Likelihood Function

The problem of maximizing the likelihood of the data observations over the set of parameters can be solved by taking the derivative of $p_{\tilde{y}(x)}(\tilde{y}(x); \tilde{p})$ or of a monotonically increasing function of $p_{\tilde{y}(x)}(\tilde{y}(x); \tilde{p})$ with respect to each parameter $p$ and then setting the results equal to zero. The resulting values for the parameters will include the global minimum of $p_{\tilde{y}(x)}(\tilde{y}(x); \tilde{p})$ if the parameters are unconstrained. If the parameters are constrained, the value of $p_{\tilde{y}(x)}(\tilde{y}(x); \tilde{p})$ at the limiting parameter values should also be calculated. The maximization of $\ln(p_{\tilde{y}(x)}(\tilde{y}(x); \tilde{p}))$ with respect to the parameter vector leads to the following set of simultaneous equations:

$$\sum_{i=1}^{N} \left[ (-\ln(b(x_i) - a(x_i)) + \Psi(\alpha(x_i) + \beta(x_i)) - \Psi(\alpha(x_i)) + \ln(y(x_i) - a(x_i)))x_i^r \right] = 0, \quad 0 \leq r \leq k_{a}$$

(3.7a)

$$\sum_{i=1}^{N} \left[ (-\ln(b(x_i) - a(x_i)) + \Psi(\alpha(x_i) + \beta(x_i)) - \Psi(\beta(x_i)) + \ln(b(x_i) - y(x_i)))x_i^r \right] = 0, \quad 0 \leq r \leq k_{b}$$

(3.7b)

$$\sum_{i=1}^{N} \left[ \Psi'\left(\frac{\alpha(x_i) + \beta(x_i) - 1}{b(x_i) - a(x_i)} + \frac{\alpha(x_i) - 1}{\alpha(x_i) - y(x_i)}\right) \right] x_i^r = 0, \quad 0 \leq r \leq k_{s}$$

(3.7c)

$$\sum_{i=1}^{N} \left[ \frac{\alpha(x_i) + \beta(x_i) - 1}{b(x_i) - a(x_i)} + \frac{1 - \beta(x_i)}{b(x_i) - y(x_i)} \right] x_i^r = 0, \quad 0 \leq r \leq k_{b},$$

(3.7d)
where

\[ \Psi(x) = \frac{d \ln \Gamma(x)}{dx}. \]  \hspace{1cm} (3.8)

No analytical solutions could be obtained for this system of equations, and further, parameter endpoint values must be checked to determine the global maximum of the likelihood function, so numerical methods are used instead. These equations are a generalization of the maximum likelihood estimation equations for the beta distribution parameters given in [8].

3.2.2.2. Numerical Solutions to the Maximum Likelihood Equations

In order to numerically calculate the parameters that maximize the likelihood function, the simplex algorithm is utilized to search for the global minimum of

\[ -\ln(p_{\tilde{y}}(\bar{y}(x); \tilde{p})). \]  Specifically, the simplex algorithm is capable of finding local minima in the negative log-likelihood surface. Repeated application of the simplex algorithm is then utilized to determine the global minimum. The simplex algorithm is given initial values for the parameters and an initial step size. Based on the initial values and step size, the algorithm calculates the negative log-likelihood function for \( k+l \) sets of parameter estimates. The algorithm then chooses a new set of parameter estimates that reduces the negative log-likelihood function, and this set of parameter estimates replaces the set of parameter estimates with the largest negative log-likelihood value. This iterative process continues until the algorithm has determined a local minimum in the negative log-likelihood function. This local minimum and a new large step size can then be utilized to search for another local minimum in the log-likelihood function. The
Simplex algorithm was implemented in the Matlab programming language [10] based on the description of the process given in [11].

Simulations demonstrate that the model in which the shape and endpoint parameters are described as polynomials is more suited to this numerical optimization. The negative log-likelihood surface seems to contain fewer local minima for this model, simplifying the analysis. Also, if the Simplex algorithm is applied to the second model, one of the moments of the tested points will sometimes exceed the bounds for that moment, leading to improper optimization.

3.2.3. Cramer-Rao Bounds

The \( k \times k \) Fisher information matrix \( I_{\tilde{f}(x)}(\tilde{p}) \) is calculated and then inverted to determine the covariance matrix of the parameter estimates. As described in [6], the term at the \( r \)-th row and \( s \)-th column in the Fisher information matrix is given by

\[
- E\left[ \frac{\partial^2}{\partial p_r \partial p_s} \ln(p_{\tilde{f}(x)}(\tilde{y}(x); \tilde{p})) \right].
\]

In total, \( \frac{k(k + 1)}{2} \) different expectations must be calculated, and the expressions for these expectations are cumbersome in this problem. Consequently, the entries in the Fisher Information matrix are relegated to Appendix 3.1.

Next, the Cramer-Rao bounds on the minimum variance of any unbiased estimate for the parameters are the diagonal elements of the inverse of the Fisher Information matrix:

\[
\Lambda_{est} = [I_{\tilde{f}(x)}(\tilde{p})]^{-1}.
\]
3.2.4. **Maximum Likelihood Estimation Fails**

Standard maximum likelihood estimation of a continuum of beta distributions fails for the two described models. Specifically, the maximum likelihood parameter estimates are highly biased, and the shape of the estimated probability densities or cumulative distribution functions are not accurate in a mean-squared sense. The inaccuracy resulting from maximum likelihood estimation is easily explained analytically, though we first encountered the flaw in the maximum likelihood technique through simulation.

Because the probability density function of the beta distribution is unbounded if one of the shape parameters is less than one, the likelihood of a set of points can be maximized by setting one of the endpoint curves to equal one of the data samples at a specific value for the independent parameter and then setting the corresponding shape parameter to a value less than one. In this case, $p_{\tilde{y}(x)}(\tilde{y}(x); \tilde{p})$ is infinite as long as the endpoint curves have also been chosen to include the other points. If polynomial bases are used for the endpoint curves, the set of parameters that determine the endpoints can clearly be chosen to pass through one of the points and include the remaining points. Consequently, some of the unknown parameters defining the shape parameters and one of the two endpoints will always be indeterministic since the probability will remain infinite for a range of values for these parameters. An example of the distribution estimate resulting from maximum likelihood estimation is shown for a data set that is not a function of the independent parameter (Figure 3.1).

Because the maximum likelihood technique is implemented on a computer using numerical methods, $p_{\tilde{y}(x)}(\tilde{y}(x); \tilde{p})$ does not become infinite but simply becomes very large before the simplex algorithm converges. However, the maximum likelihood result
is still biased and is indeterministic. Sometimes, one of the local minima obtained through the Simplex algorithm corresponds to a much more accurate estimate for the parameters than the global maximum, but we cannot rationalize a technique for choosing the correct local minimum. Instead, we discuss several other estimators that modify the maximum likelihood technique.

The flaw in the maximum likelihood technique in estimating the four parameter beta distribution also occurs in the estimation of several other unbounded distributions. Several authors [12,13] discuss problems in maximum likelihood estimation of a variety

\begin{figure}[h]
\centering
\begin{subfigure}{.5\textwidth}
\centering
\includegraphics[width=\textwidth]{figure3a.png}
\caption{Beta Distribution Endpoints}
\end{subfigure} \hfill
\begin{subfigure}{.5\textwidth}
\centering
\includegraphics[width=\textwidth]{figure3b.png}
\caption{Probability Density and CDF}
\end{subfigure}
\end{figure}

\begin{figure}[h]
\centering
\begin{subfigure}{.5\textwidth}
\centering
\includegraphics[width=\textwidth]{figure3c.png}
\caption{Original and Estimated PDFs}
\end{subfigure} \hfill
\begin{subfigure}{.5\textwidth}
\centering
\includegraphics[width=\textwidth]{figure3d.png}
\caption{Original and Estimated CDFs}
\end{subfigure}
\end{figure}

\textbf{Figure 3.1.}

The possibility for arbitrarily probability density functions results in the failure of the maximum likelihood estimation technique. 3.1a. shows the estimated endpoints, and 3.1b. shows the estimated probability distribution. 3.1c. and 3.1d compare these estimates to the distributions used to generate the data.
of distributions, including the four parameter beta distribution, and also propose different approaches for correcting the bias. These techniques, as well as others that we propose, are discussed in the context of adjusting the maximum likelihood estimator.

3.2.5. Asymptotic Properties of the Maximum Likelihood Estimator

The asymptotic behavior of the maximum likelihood estimator may be formulated by considering the estimate for any finite number of observations. For any finite number of observations, several sets of parameter estimates lead to an infinite likelihood function. Since the maximum likelihood estimate is not unique for any set of N points, the problem of determining the limit of the maximum likelihood estimate as the number of points approaches infinity in not well-posed and cannot easily be solved.

3.3. Altered Maximum Likelihood Techniques

We consider several altered maximum likelihood estimators that mitigate the problem of infinite probability evident in standard maximum likelihood estimation. Although we mostly present newly developed techniques, we compare these techniques with some previously developed adjustments to the maximum likelihood estimator. Unfortunately, analytical justification of the improved estimators is difficult, but some conceptual or analytical basis for the success of each estimator is given.

3.3.1. Local Maxima as the Maximum Likelihood Estimate

As we briefly described before, a local maximum in the likelihood function
sometimes provides a better estimate for the parameters than the global maximum in the described models. Previous research in the estimation of non-regular distributions, including the beta distribution, has quantified the possibility for estimation using local maxima in the likelihood function. Cheng and Amin [14] prove that a local maximum can be found such that the parameter estimates are consistent only if the shape parameters are both greater than or equal to one. Unfortunately, we are also concerned with beta distribution models with \( \alpha \) or \( \beta \) less than one. Also, even for \( \alpha, \beta \geq 1 \), the existence of the local maximum is only guaranteed asymptotically, leading to highly biased estimates for small data sets. Successful local maximization of the probability density function is shown for a 200 point data set that does not vary with the independent parameter.

3.3.2. Stability Constrained Maximum Likelihood

In order to develop improved maximum-likelihood type estimators, we must first highlight the flaws in using the standard maximum likelihood estimate. One primary characteristic of the maximum likelihood estimator is the calculation of parameters such that one of the beta distribution endpoints coincides exactly with a data point for at least one independent parameter value, resulting in infinite probability. One problem with this estimate is the significant variation in \( p_{f(x)}(\tilde{y}(x); \tilde{p}) \) due to the change in one of the parameters or one of the data points. In order to prevent this fluctuation in the likelihood of the points, we can constrain the maximum likelihood result by penalizing drastic changes in the likelihood due to differential changes in either the parameters or the points.
Figure 3.2.

Local maximization of the likelihood function provides reasonable estimates for sufficiently large data sets. 3.1a. shows the estimated endpoints, and 3.1b. shows the estimated probability distribution. 3.1c. and 3.1d. compare these estimates to the distributions used to generate the data.

The following estimator form is used to force stability in the likelihood value with respect to differential changes in the parameters:

$$\tilde{p}(\tilde{y}) = \arg \max_\tilde{p} \left[ g(p_{\tilde{y}}(\tilde{y}; \tilde{p})) - f(\frac{d}{dp} g(p_{\tilde{y}}(\tilde{y}; \tilde{p}))) \right],$$

(3.11)

where $g()$ is again a monotonically increasing function, $\tilde{y} = \tilde{y}(x)$ may more generally represent a dependent variable that is connected with another independent parameter, $\frac{d}{dp} ()$ is the vector of partial derivatives with respect to each element of $\tilde{p}$, and $f()$ is an
adequately chosen function that maps the partial derivatives to a penalty value. The function $f$ must be chosen to become infinite as \( \frac{d}{dp} g(p_{\varphi}(\bar{y}; \bar{p})) \) becomes large. In this case, the penalty for choosing a beta distribution endpoint that passes through a data point will be infinite also. Unfortunately, the function $f$ is difficult to choose. Though the penalty theoretically corrects for the instability of the distribution, the likelihood expression is still undefined for certain parameter choices. Specifically,

\[
g(p_{\varphi}(\bar{y}; \bar{p})) - f\left(\frac{d}{dp} g(p_{\varphi}(\bar{y}; \bar{p}))\right)
\]

is the difference between two undefined or infinite quantities for certain sets of parameters. Even though this difference will not be infinite in a practical implementation of the maximization, the function $f$ must be chosen so that the estimate is accurate in some sense. We do not know how to accurately choose the penalty function.

An estimator based on the stability of the likelihood to variations in the points is more useful and easier to justify. Instead of considering a penalty-based constraint, we assign uncertainties to the locations of the points. For example, assume that each observation is not an exact value but instead lies in a range. The actual observation is most likely similar to the stated observation but may also differ from the stated observation with lower probability. This idea of uncertain observations, or simply the constraint of a stable estimator, leads to an adjusted maximum likelihood estimator for certain models. This estimator is presented for the specific case of the first model described in this chapter. The standard likelihood function is replaced by

\[
p^* = \prod_{i=1}^{N} \int_{-\infty}^{\infty} K(x_i)(y - a(x_i))^{\alpha(x_i)-1} (b(x_i) - y)^{\beta(x_i)-1} f(y; y_i) \, dy,
\]  

(3.12)
where \( f(y; y_i) \) are unit area functions that represent the range of interest or stability for each observation. If the functions are chosen as delta functions located at the observation points, then \( p^* \) simplifies to the maximum likelihood estimator. Meanwhile, if the functions \( f(y; y_i) \) are chosen as rectangular functions, we maximize the joint probability that the points lie in certain ranges rather than the joint probability density of the points. Cheng and Iles [15] also rationalize estimators based on interval rather than point probabilities.

To provide some analytical justification for the estimator, we briefly consider the simpler model

\[
p^* = \prod_{i=1}^{N} \int K(y-a)^{a-1}(b-y)^{b-1} f(y; y_i) dy.
\]

(3.13a)

where

\[
K = \frac{1}{(b-a)^{a+b-1}} \frac{1}{\beta(\alpha, \beta)}.
\]

(3.13b)

Next, without loss of generality, order the observations as \( y_1 < y_2 < \ldots < y_N \) and then choose the functions \( f(y; y_i) \) so that the estimator reduces to

\[
p^* = \prod_{i=2}^{N-1} K(y_i-a)^{a-1}(b-y_i)^{b-1} (\int_{y_i}^{y_{i+1}} K(y-a)^{a-1}(b-y)^{b-1} dy)
\]

\[
(\int_{y_N-h}^{y_N} K(y-a)^{a-1}(b-y)^{b-1} dy)
\]

(3.14)

Cheng and Amin [14] show that this estimator is an asymptotically consistent estimator for all four parameters for appropriately chosen \( h \). Unfortunately, the work by Cheng and
Amin [14] could not be found, but the outline of an analogous proof and the conditions on \( h \) are given in Cheng and Iles [15] for the Weibull distribution.

Since the actual data sets analyzed in this study are small, we must choose functions \( f \) such that the bias in the estimator for small samples. A slight modification in equation (3.14) should significantly improve the non-asymptotic behavior of the estimator:

\[
P^* = \left( \prod_{i=2}^{N-1} K(y_i - a)^{a-1}(b - y_i)^{b-1} \right) \left( \int_{y_{i-h}}^{y_{i+h}} K(y - a)^{a-1}(b - y)^{b-1} dy \right) \]

where \( h \) is again appropriately chosen based on the number of points. Specifically, this estimator will prevent the estimation of \( a \) as the smallest data sample and the estimation of \( b \) as the largest data samples. We do not explore choices for the function \( f \) that improve estimation more generally. We feel that many of these stability-constrained estimators will be asymptotically consistent and will also provide adequate estimation for the parameters based on finite data sets.

3.3.3. Endpoint Constrained Maximum Likelihood

In the previous section, we described the standard maximum likelihood estimate as flawed because of the variation in the likelihood function due to differential changes in the points or the parameters. Alternately, we may describe the failure of the maximum likelihood estimate as resulting from the improper estimation of the endpoints of the distribution. Specifically, if the endpoints of the beta distribution are known, the maximum likelihood estimate for the shape parameters is accurate [16]. In contrast,
however, if the shape parameters are known, maximum likelihood estimation of the endpoints are still clearly biased for shape parameters that are less than one. Based on the failure of the maximum likelihood estimate in predicting endpoints, we propose an estimator that constrains the endpoints as a function of the shape parameters and then generates a maximum likelihood estimate for only the shape parameters.

Unfortunately, the endpoint constrained maximum likelihood estimator is difficult to develop for arbitrary data sets and so we limit the development the following simple model:

\[
\hat{p}_f(\bar{y}; \hat{p}) = \prod_{i=1}^{N} K(y_i - a)^{\alpha-1}(b - y_i)^{\beta-1}, \tag{3.16a}
\]

where

\[
K = \frac{1}{(b-a)^{\alpha+\beta-1}} \frac{1}{\beta(\alpha, \beta)}, \tag{3.16b}
\]

and \( \alpha, \beta, a, \) and \( b \) are the four parameters in the model. Again, we order the observations used to determine the model parameters and label them as \( y_1 < y_2 < \ldots < y_N \). We realize that this model is insufficient in describing most of the data sets studied in this work but hope that the technique can be generalized in the future.

### 3.3.3.1. Unbiased Estimation of Endpoints

We commence the discussion of endpoint constrained maximum likelihood by first describing an unbiased estimate for the endpoints given a set of \( N \) samples and given the values of the two shape parameters \( \alpha, \beta \). In this thesis, the proof given for the unbiasedness of the estimator focuses on the conceptual explanation rather than a
rigorous derivation of the zero bias. The following estimates for the endpoints are unbiased:

\[ \tilde{a} = y_1 - \frac{q_1}{1 - q_1 - q_2} (y_N - y_1), \quad (3.17a) \]

\[ \tilde{b} = y_N + \frac{q_2}{1 - q_1 - q_2} (y_N - y_1), \quad (3.17b) \]

where \( q_1 \) is the expected value of the smallest of \( N \) points generated by a beta distribution with identical shape parameters \( \alpha, \beta \) and endpoints of 0 and 1. Similarly, \( q_2 \) is the expected distance between the largest point and the larger endpoint, 1, for this distribution.

---

**Figure 3.3.**

By scaling a beta distribution to the unit interval, the expected distance between the smallest and largest points and the endpoints can be determined.
Conceptually, the estimator is unbiased because the expected distance between $y_1$ and the endpoint $a$ and between $y_N$ and the endpoint $b$ are calculated using a scaled version of the distribution (Figure 3.3). Based on Figure 3.3, the following equality should hold:

$$b - a = \frac{E[y_N - y_1]}{(1 - q_2) - q_1}.$$  \hfill (3.18)

Further, we expect that

$$q_1 = \frac{E[y_1] - a}{b - E[y_N]}, \quad q_2 = \frac{E[y_N] - E[y_1]}{(1 - q_2) - q_1}.$$  \hfill (3.19)

These expressions can be rearranged to show that

$$a = E[y_1] - \frac{q_1}{1 - q_1 - q_2} (E[y_N] - E[y_1]),$$  \hfill (3.20a)

$$b = E[y_N] + \frac{q_2}{1 - q_1 - q_2} (E[y_N] - E[y_1]).$$  \hfill (3.20b)

Consequently, the estimators (3.17a) and (3.17b) are unbiased.

Next, the expectations $q_1$ and $q_2$ can be determined. We consider $N$ data samples generated from a beta random variable with shape parameters $\alpha, \beta$ and endpoints of 0 and 1; these samples are labeled $y_1^* < y_2^* < \ldots < y_N^*$. To calculate $q_1$, we note that the probability that $y_1^* > y$ is

$$P(y_1^* > y) = (1 - F(y))^N,$$  \hfill (3.21)

where $F(y)$ is the cumulative distribution function of the beta random variable with shape parameters $\alpha, \beta$ and endpoints of 0 and 1. Since $y_1^*$ is a positive random variable, the expected value of $y_1^*$ can be calculated as
\[ q_1 = \int_0^1 P(y_1 > y)dy = \int_0^1 (1 - F(y))^N dy. \]  

(3.22a)

Similarly,

\[ q_2 = \int_0^1 (F(y))^N dy. \]  

(3.22b)

Since the two expectations for the distances of the smallest and largest points from the distribution endpoints can be calculated for the distribution with endpoints of 0 and 1, unbiased estimation of the endpoints of any distribution with known shape parameters is possible.

### 3.3.3.2. Constrained Maximum Likelihood Based on Unbiased Estimation of Endpoints

Next, we propose an estimator for all four parameters of an unknown beta distributions. The estimator is optimized only in the space of the two shape parameters; the endpoint parameters are constrained according to the unbiased estimation of endpoints presented in the previous section. Specifically, the estimates for the shape parameters are calculated based on the following likelihood function:

\[ p_f(\tilde{y}; \alpha, \beta) = \prod_{i=1}^N K(y_i - a(\alpha, \beta))^{\alpha-1}(b(\alpha, \beta) - y_i)^{\beta-1}, \]  

(3.23a)

where

\[ K = \frac{1}{(b-a)^{\alpha+\beta-1}} - \frac{1}{\beta(\alpha, \beta)} \]  

(3.23b)

and \( a(\alpha, \beta) \) and \( b(\alpha, \beta) \) are unbiased estimates for endpoints given the shape parameters.

A conceptual rationale for this estimator is straightforward. Among all estimators with specific shape parameters \( \alpha, \beta \), the only reasonable estimates for the endpoints are
near the unbiased estimates predicted for given shape parameters. Unfortunately, a more mathematical statement of the validity of the constraint is difficult. If a sufficient number of points are used, the following argument shows why the constraint is valid. Assume that a consistent estimator can be formulated for the shape parameters. In this case, as the number of points becomes large, the estimates for the shape parameters approach their real values. Consequently, the endpoints must approach the value predicted by their proper estimator given the shape parameters in order to achieve unbiased estimation. Therefore, the unbiased estimate of the four parameters is contained within the set of estimates with variable shape parameters and endpoints constrained based on these shape parameters.

Monte Carlo trials are utilized to determine the mean and variance of the estimator for various finite data sets. Examples of distribution estimates are shown for \( N = 20 \) and \( 50 \) data samples (Figure 3.4). Repeated trials demonstrate that the estimates are not unbiased, but the bias is small. Unfortunately, the variances of the shape parameters are quite large for small numbers of points. The estimates for the endpoints are much more accurate for the completed trials, although this accuracy will decrease if larger shape parameters are chosen for the original distribution.
Figure 3.4.

Endpoint constrained maximum likelihood is successful in estimating distributions for data with statistics that do not depend on the independent parameter. 3.4a and 3.4b compare source and estimated cumulative distribution functions based on 20 samples. 3.4c and 3.4d compare source and estimated cumulative distribution functions based on 50 samples.

3.3.4. Maximum Likelihood with Probability Truncation

The primary difficulty with maximum likelihood estimation of beta distributions is the tendency of the likelihood function to approach infinity for some parameter values. Consequently, we propose an estimator that limits the likelihood function by truncating the beta density function for the maximum likelihood calculation. The following likelihood function is used for estimation:
\[ p_{\hat{p}(x)}(\tilde{y}(x); \tilde{p}, \delta) = \prod_{i=1}^{N} \gamma(K(x_i)(y_i - a(x_i))^{\alpha(x_i)}(b(x_i) - y_i)^{\beta(x_i)}), \]  

(3.24a)

where

\[ K(x_i) = \frac{1}{(b(x_i) - a(x_i))^{\alpha(x_i)+\beta(x_i)-1}} \frac{1}{\beta(\alpha(x_i), \beta(x_i))}, \]  

(3.24b)

and

\[ \gamma(r) = \begin{cases} r, & r < \delta \\ \delta, & r > \delta \end{cases} \]  

(3.24c)

The proper choice of the threshold depends upon the size of the data set used for the estimation.

The estimate provided by probability truncation-based maximum likelihood is necessarily biased for some sets of parameters. Specifically, the estimator will never choose certain parameter sets. For example, consider a data set with no dependence on the independence parameter. For any given probability threshold, the estimated shape parameters have a minimum value greater than zero, since the new likelihood function will decrease regardless of the data set for sufficiently small shape parameter estimates (Figure 3.5). To prevent the existence of parameter values that cannot be achieved by the estimator, a variable probability threshold that is raised if a sufficient number of the points attain the original threshold can be used.

Unfortunately, a systematic method for choosing the threshold or varying the threshold in order to accurately estimate a variety of distributions was not developed. The idea of using truncated probability distributions for maximum likelihood estimation is interesting and should perhaps be explored more rigorously.
Figure 3.5.
This figure highlights the truncation of a probability distribution. In this case, the distribution is truncated at a probability value of 0.4.

3.3.5. Maximum Likelihood Using the Edgeworth Approximation

The adjusted maximum likelihood estimators presented so far improve the accuracy of the estimate by altering the calculation of the likelihood function. In this final version of the maximum likelihood estimation, we consider using a probability representation that approximates the beta distribution and then calculating the parameters of this new model using the maximum likelihood technique.

3.3.5.1. The Edgeworth Approximation to a Distribution

Knowledge of all the moments of a random variable is sufficient information for determining the probability density function of that random variable. If only a finite number of the moments are known, the Edgeworth approximation can be used to estimate the probability density function [17]. We consider the first order Edgeworth approximation to a probability distribution based on the first four moments of that distribution. We disregard both the higher-order third and fourth moment contributions.
to the distribution and the contributions of higher moments. The following expression is
the Edgeworth approximation to a distribution using the mean and the second, third, and
fourth central moments [17]:

\[ p_T(y) = N(y; u_1, u_2)[1 - \frac{u_3}{u_2^{3/2}} (3y - y^3) + \frac{u_4 - 3u_2^2}{24u_2^2} (3 - 6y^2 + y^4)] \tag{3.25a} \]

where \( N(y; u_1, u_2) \) is the Gaussian distribution:

\[ N(y; u_1, u_2) = \frac{1}{\sqrt{2\pi u_2}} e^{-\frac{(y-u_1)^2}{2u_2}}. \tag{3.25b} \]

3.3.5.2. Maximum Likelihood With a Fourth-Order Edgeworth Approximation

Once the expression for the Edgeworth approximation to a beta distribution has
been established, application of the maximum likelihood procedure is straightforward.
Again, we model the moments as polynomial functions of the independent parameter and
attempt to determine the coefficients in the polynomial models. The following likelihood
is then maximized over the set of parameters:

\[ p_T(x_1) = \prod_{i=1}^{N} p_T(y_i, x_i; \bar{p}), \tag{3.26a} \]

where \( p_T(y_i, x_i; \bar{p}) \) is calculated using the Edgeworth approximation. The moment
estimates obtained using maximum likelihood then are assumed to be the moments of the
beta distribution estimate for the data.

Simulations of estimation using the Edgeworth approximation highlight some of
the advantages of this maximum likelihood technique. Most importantly, the Edgeworth
approximation to a distribution is bounded for finite parameter values, preventing an
infinite likelihood. Further, the estimation technique usually correctly determines
whether the shape parameters of the beta distributions that describe the data are greater or
less than one.

Unfortunately, the Edgeworth approximation for a beta distribution based on only
four moments is not accurate. Consequently, the variance of the parameter estimates is
large. Also, we cannot justify the use of the Edgeworth approximation for the estimation
of beta distributions. The Edgeworth approximation likelihood function is given as a
function of one unknown parameter, skewness, for data generated from an exponential
random variable (Figure 3.6).

Figure 3.6.
This plot is an example of an Edgeworth approximation based likelihood function.
Appendix 3.1.

The entries in the Fisher information matrix are labeled according to the two parameters that are present in the derivative expression 

\[-E\left[ \frac{\partial^2}{\partial p_i \partial p_j} (\ln(p_{\tilde{y}(x)}(\tilde{y}(x); \bar{p})) \right].\]

First, the elements of the Fisher Information matrix for which both parameters originate from the same beta distribution parameter are calculated:

\[I(p_{b_k}, p_{b_l}) = -\sum_{i=1}^{N} [x_i^{k+l} (\Psi'(\alpha(x_i) + \beta(x_i)) - \Psi'(\beta(x_i))],\]

\[I(p_{a_k}, p_{a_l}) = -\sum_{i=1}^{N} [x_i^{k+l} (\Psi'(\alpha(x_i) + \beta(x_i)) - \Psi'(\alpha(x_i))],\]

\[I(p_{b_k}, p_{b_l}) = -\sum_{i=1}^{N} [x_i^{k+l} (\alpha(x_i) + \beta(x_i) - 1) - E]\left[ \frac{\beta(x_i) - 1}{(b(x_i) - a(x_i))^2} \right],\]

and

\[I(p_{a_k}, p_{a_l}) = -\sum_{i=1}^{N} [x_i^{k+l} (\alpha(x_i) + \beta(x_i) - 1) - E]\left[ \frac{\alpha(x_i) - 1}{(Y_i - a(x_i))^2} \right],\]

where

\[\Psi'(x) = \frac{d\Psi(x)}{dx}.\]

The diagonal elements of the Fisher Information matrix correspond to \(k=l\) in the previous expressions.
Next, the terms in the Fisher Information matrix that require derivatives in terms of parameters from two different beta distribution parameter expressions are given:

\[
I(p_{b_k}, p_{al}) = -\sum_{i=1}^{N} \left[ x_i^{k+1} \Psi'(\alpha(x_i) + \beta(x_i)) \right],
\]

\[
I(p_{b_k}, p_{al}) = -\sum_{i=1}^{N} \left[ x_i^{k+1} \left( \frac{1}{a(x_i) - b(x_i)} + E\left[ \frac{1}{b(x_i) - Y_i} \right] \right) \right],
\]

\[
I(p_{b_k}, p_{al}) = -\sum_{i=1}^{N} \left[ x_i^{k+1} \left( \frac{1}{b(x_i) - a(x_i)} \right) \right],
\]

\[
I(p_{b_k}, p_{al}) = -\sum_{i=1}^{N} \left[ x_i^{k+1} \left( \frac{1}{b(x_i) - a(x_i)} + E\left[ \frac{1}{Y_i - a(x_i)} \right] \right) \right],
\]

and

\[
I(p_{b_k}, p_{al}) = -\sum_{i=1}^{N} \left[ x_i^{k+1} \left( \frac{1 - \alpha(x_i) - \beta(x_i)}{(b(x_i) - a(x_i))^2} \right) \right].
\]
Chapter 4

Regression Analysis of the Moments

In this chapter, we consider estimation of a continuum of beta distributions from a data set using least squares regression analysis of the first four moments of the distribution. We consider three different models: one model attempts to describe the data solely as a function of the independent parameter, while the other two models describe a data point as a function of known data. The model that most accurately captures the deterministic information in the data set in the mean estimate, and so minimizes the correlation between the random components of each data point, is chosen to describe a data set. Next, the parameters of the chosen model are calculated using linear regression techniques. Appropriate bias corrections for the parameter estimates are implemented, and the accuracy of the parameter estimates is also determined. Finally, hypothesis testing is utilized to eliminate unnecessary parameters from the model.

In the chapter, we first introduce the technique of least squares linear regression. We then formulate and discuss a parametric model that attempts to describe data sets based solely on the independent parameter. We apply linear regression to determine the unknown parameters of the model, calculate the error in the parameter estimates, and consider methods for testing the number of parameters needed in the model. We also consider interpolation and extrapolation using the chosen model. Next, we formulate the two models that describe data based on other known data samples. Again, we describe the method for determining the model parameters using regression analysis and consider interpolation and extrapolation using the model.
4.1. A Synopsis of Linear Regression

A brief review of linear regression is presented to facilitate the development of the process for regression analysis of moments [18]. In linear regression, the dependent variable of a set of data is estimated as a linear combination of functions of the independent variables. Consider a set of \( N \) data points with coordinates 
\((x_1, y_1) \ldots (x_N, y_N)\), where \( x \) is a vector of independent variables. Further, assume that the samples originate from \( N \) random variables labeled \( Y(x_1) \ldots Y(x_N) \). Next, assume that \( y \) will be estimated using \( k+1 \) basis functions \( f_0(x) \ldots f_k(x) \), where \( k < N \). Further assume that these \( k+1 \) basis functions are independent, so that no basis function can be written as a linear combination of the other basis functions. The estimate for \( y \), labeled \( e(x) \), can then be written as

\[
e(x) = \sum_{i=1}^{k} p_i f_i(x),
\]  

(4.1)

and the objective of the regression is the estimation of the parameters \( p_0 \ldots p_k \). A metric for determining the optimal set of parameters is necessary; for this analysis, the least-squares metric is utilized. Specifically, the vector \( p_0 \ldots p_k \) is chosen so that

\[
\sum_{j=1}^{N} (y_j - e(x_j))^2 = \sum_{j=1}^{N} (y_j - \sum_{i=0}^{k} p_i f_i(x_j))^2
\]

(4.2)

is minimized. The parameters that minimize the sum may be calculated by setting the derivative of the sum with respect to each parameter to zero. The resulting set of \( k+1 \) equations is easily represented in matrix form:
For notational simplicity, the equation is rewritten as

\[ R = Q \tilde{p}, \]

and the solution for the vector of parameters is

\[ \tilde{p} = Q^{-1} R. \]

In vector notation, the value of \( e(\bar{x}) \) for an arbitrary vector \( \bar{x} \) is

\[ e(\bar{x}) = \bar{f}^T \tilde{p} = \bar{f}^T Q^{-1} R, \]

where

\[ \bar{f}^T = [f_0(\bar{x}) \ldots f_k(\bar{x})]. \]

The estimate \( e(\bar{x}) \) is a linear combination of \( y_1 \ldots y_N \), and so the variance of the estimate can be calculated based on the covariance matrix of the data. This variance calculation assumes that the parametric form describing the data is correct and only accounts for the possible error in parameter calculation within the context of the parametric form. Specifically,

\[ \text{var}[e(\bar{x})] = \bar{f}^T (G^T \Lambda^{-1} G)^{-1} \bar{f}, \]

where

\[ G^T = \begin{bmatrix}
  f_0(\bar{x}_0) & \ldots & f_0(\bar{x}_N) \\
  \vdots & \ddots & \vdots \\
  f_k(\bar{x}_0) & \ldots & f_k(\bar{x}_N)
\end{bmatrix} \]
and $\Lambda_y$ is the NxN covariance matrix of $Y(\tilde{x}_i)\ldots Y(\tilde{x}_N)$. The covariance matrix $\Lambda_y$ is generally unknown but can be estimated from the data. Also, the covariance matrix of the parameters, a quantity that is useful in testing the validity of the model, is

$$\Lambda_p = (G^T \Lambda_y^{-1} G)^{-1}. \quad (4.8)$$

In the formulation of linear regression, we have considered estimating a dependent variable as a function of an independent variable. More generally, however, variables can be labeled arbitrarily as independent or dependent, and the regression process simply attempts to describe one member of a coordinate set as a function of other members of that coordinate set. For example, the same sample $y_i$ may be the dependent, or estimated, variable in one coordinate set and also one of the independent parameter values in another coordinate set. This modeling of random process samples as functions of other samples in the same process is especially useful in describing time series. The value of a time series often demonstrates a strong correlation to previous values of the time series and is not functionally correlated to the parameter time, so that a recursive regression model is necessary.

### 4.2. Regression of the Moments Based on a Single Independent Parameter

We model each of the moments as a linear combination of functions of the independent parameter. The parameters in the model are the coefficients of these basis functions; regression analysis is utilized to determine the parameters. We assume that the deterministic information in the data can be modeled solely using the independent parameter, so that the beta random variables describing each data sample are
independent. Although the model used for the regression technique is identical to the second model used for maximum likelihood analysis, we briefly describe the model again and provide an overview of the application of regression analysis to the model. Next, we describe the regression of each moment in detail and then consider interpolation and extrapolation of the model result.

4.2.1. Model Formulation and Conceptual Overview

A function regression method is developed for estimating a continuum of beta distributions based on data generated from that set of distributions. A random variable parametrized by an independent variable, labeled $Y[x]$, is the source of the data. $Y[x]$ is assumed to be a beta random variable and so is uniquely described by the mean $\mu(x)$ and the second, third, and fourth central moments, $\mu_2(x), \mu_3(x), \mu_4(x)$. Estimates for $\mu(x), \mu_2(x), \mu_3(x), \mu_4(x)$ are formulated based on outcomes of independent experiments of $Y[x]$ for $x \in \{x_1, \ldots, x_N\}$. Specifically, function regressions of the outcomes of these experiments, $y_1, \ldots, y_N$, provide the estimates for the moments of the distribution. First, the mean of the data is estimated using an appropriate regression of the data. The higher order moments are then estimated based on the estimate for the mean, $\mu(x)$, and the data.

The first query in the regression analysis problem is the choice of the functions that are fitted to the mean and higher central moments of the data. For a general formulation of the problem, the set of possible functions that can fit the mean is unfortunately not constrained at all. For example, consider a situation in which the means of the data at $x \in \{x_1, \ldots, x_N\}$ are determined by $N$ independent trials of some other
zero mean Gaussian random variable. In this case, \( u'(x) \) may equal an arbitrary zero mean random process. In this situation, however, no regression method can approximate the mean. Instead, a parametric form for each of the four moments is assumed. In this case, \( u'(x) \), \( u_2(x) \), \( u_3(x) \), and \( u_4(x) \) are each described by a functional form and a set of parameters \( p_0 \ldots p_k \). For example, the mean of the data might be described by a \( k \)-th order polynomial of the independent parameter \( x \) with \( k+1 \) unknown coefficients.

Estimation of one of the moments is then analogous to the determination of the set of parameters \( p_0 \ldots p_k \).

More philosophically, although the estimates for each moment are calculated using data at \( x \in \{x_1 \ldots x_N\} \), the resulting functions that describe each moment are utilized to calculate beta distributions for a continuum of independent parameter values. In other words, we assume that each moment is properly modeled with a specific parametric form. The choice of a proper parametric form, which corresponds to choosing an adequate set of basis functions to model the moments of the data, is an important and difficult question. We discuss the determination and evaluation of parametric forms for the data in a general context, as well as non-parametric estimation techniques, in the last section of this chapter. In this section, only the possible errors in the parameter choices and the resulting variability in the interpolation and extrapolation results for a specific parametric form are calculated.

In this section, only parametric descriptions that are linear with respect to the parameters are considered. The estimate for a specific moment is formulated as

\[
e(x) = \sum_{i=0}^{4} p_i f_i(x),
\]  

(4.9)
where $f_0(x)...f_k(x)$ are appropriately chosen basis functions. Two sets of basis functions, polynomials and sinusoids, are used for the described regression analyses.

The chosen parametric form for the moments leads to several other questions. Specifically, although we specify a possible set of basis functions as the parametric form for each moment, only a subset of these basis functions is actually necessary for the regression. The actual set of basis functions used for the regression must be chosen and so the number of parameters must be decided. Unfortunately, even when a finite subset of basis functions is used, the proper set of bases cannot always be deciphered. For example, if the mean of $Y[x]$ is an $N$th order polynomial and the higher order moments are non-zero, then the mean of the data values $y_1...y_N$ cannot ever be properly modeled with the correct $N$th order polynomial. Further, the same data $y_1...y_N$ can be generated from a process $Y[x]$ with a lower order polynomial model for the mean, so the number of necessary parameters is unclear. Two standards are utilized to determine whether the chosen form for the parametric model and the number of parameters used for the model are appropriate. One significant assumption in the regression problem formulation is the independence of the random variables used to describe $y_1...y_N$. If the data generated from the chosen parametric model violates this assumption, which implies that the deterministic information in the data has not been modeled as a function of the independent parameter, then the parametric model is improper. Next, a hypothesis test is utilized to determine whether individual parameters in the formulated model are necessary.

A final note in attempting the regression analysis is the existence of constraints on the regression result due to the assumption that the modeled data originates from a set of
independent beta distributions. The even moments of any random variable are constrained to be positive, and the higher order even moments are further limited. The moments of a beta distribution are even more constrained; these constraints suggest optimal choices for the basis functions used in the regression.

The analytical development of the model is interspersed with applications of the model to actual data sets.

4.2.2. Regression of the Mean: Polynomial Regression Based Solely on the Independent Parameter

The model is formulated so that the data values are uncorrelated; equivalently, we wish to capture the deterministic information of the signal in the mean, while the higher order moments will describe the possible deviation in the result at a given value for the independent parameter. In this case, the vector of basis functions

\[ f^T = [f_0(x) = 1 \ldots f_k(x) = x^k], \]  

(4.10)

where \( x \) represents the independent parameter, are used to describe the data. The best choices for the coefficients of each basis are determined by direct application of the linear regression method. Specifically, for known given set of data \((x_1, y_1) \ldots (x_N, y_N)\), the regression equation for the mean using a k-th degree polynomial is

\[ \bar{u}(x) = \hat{f}^T \hat{p} = \tilde{f}^T \tilde{Q}^{-1} \tilde{R}, \]  

(4.11a)

where
Q = \begin{bmatrix}
\sum_{j=1}^{N} 1 & \sum_{j=1}^{N} x_j & \cdots & \sum_{j=1}^{N} x_j^k \\
\sum_{j=1}^{N} x_j & \cdots & \sum_{j=1}^{N} x_j^k \\
\sum_{j=1}^{N} x_j^2 & \sum_{j=1}^{N} x_j^2 & \cdots & \sum_{j=1}^{N} x_j^{2k}
\end{bmatrix} \quad (4.11b)

and

R = \begin{bmatrix}
\sum_{j=1}^{N} y_j \\
\sum_{j=1}^{N} y_j x_j \\
\sum_{j=1}^{N} y_j x_j^2 \\
\sum_{j=1}^{N} y_j x_j^{2k}
\end{bmatrix} \quad (4.11c)

Two examples are given for the regression of the mean (Figures 4.1a and 4.2a).

In both cases, a second order polynomial is used to estimate the mean of the data. In both examples, the data is generated from a known beta distribution model with a mean 

\[ u'(x) = x + x^2, \]

with \( N = 20 \) and \( 100 \) for the two examples. Both of the analyses are completed assuming that the statistical model underlying the random process is unknown.

In addition to the estimation of \( u'(x) \), two other analyses are important. In order to decide the validity of the model, a process for hypothesis testing of the model must be instituted. Secondly, the accuracy of the mean estimate must be determined.

Two types of hypothesis testing together can validate a certain model for the mean of the data. One test is useful for eliminating extra parameters that probabilistically are indeterministic; the second test determines the sufficiency of the model by determining whether the sequence of deviations of the data points from the mean is a white process.

In order to test for unnecessary parameters, an estimate for the covariance matrix of the parameter estimates is useful. The covariance matrix of the parameters is
Figures 4.1, 4.2
These two figures demonstrate polynomial regressions to determine the mean of 100 and 20 point data sets, respectively, which were generated from the same model. The mean line, autocorrelation of the residuals, and extrapolated mean line and mean line variance are shown.
\[ \Lambda_p = (G^T \Lambda_p^T G)^{-1}, \]  
(4.12a)

where

\[ G^T = \begin{bmatrix} 1 & \ldots & 1 \\ \vdots & \ddots & \vdots \\ x_1^k & \ldots & x_N^k \end{bmatrix}. \]  
(4.12b)

The covariance matrix of the parameters can only be determined if the covariance matrix of the data is known. Since the model is chosen to concentrate the data correlation in the mean, the covariance matrix of the data should be a diagonal matrix. The estimate for the variance of the data is utilized to generate the diagonal entries of the matrix. Once the covariance matrix of the data has been estimated, the covariance matrix of the parameters is determined. Next, for most sets of data, each of the parameters is a linear combination of independent random variables with similar variances and so is approximately Gaussian according to the central limit theorem. Since the approximate distribution of each parameter is known, a suitable hypothesis test can be formulated.

One possible hypothesis test rejects a basis function with a parameter that can switch sign with a probability larger than some threshold. Formally, a parameter \( p_k \) should be retained if and only if

\[ \min\{\Pr[p_k > 0], \Pr[p_k < 0]\} < q, \]  
(4.13)

where 0<q<0.5. If q is appropriately chosen, the basis function with parameter \( p_k \) is only retained when the parameter exceeds the variance of the parameter. To test the accuracy of the hypothesis test, cubic polynomials are fitted to the multiple samples of data generated from the known distribution with mean

\[ u^*(x) = x + x^2. \]  
(4.14)
The success of the hypothesis test in excluding unnecessary parameters is determined (Table 4.1). Further, polynomial regressions with various numbers of parameters are completed for motor price data, and the hypothesis test is utilized to eliminate unnecessary parameters in the model. A second order polynomial model is found to accurately describe the price of motors as a function of the independent parameter, the motor power.

Next, the hypothesis test for the sufficiency of a model for the mean is based on the autocorrelation of the noise signal resulting from the subtraction of the estimated mean. The necessary autocorrelation is not known but can be estimated based on the data and the estimated mean function. The sample autocorrelation is easily calculated for time series data since the outcomes of the random process are known at discrete, evenly spaced intervals. Autocorrelations of the noise signal remaining after a second-order polynomial fitting of the mean are shown for two examples with evenly spaced data (Figures 4.1b and 4.2b). The test for correlation is more difficult with non-evenly spaced data. The following test is proposed:

$$L = \sum_{j=1}^{N} \sum_{i<j}^{N} \frac{y_i y_j}{x_j - x_i}$$

(4.15)

is calculated for the data. If the data are uncorrelated, the expected value of the sum is zero, and the variance of the sum is

$$\text{var}(L) = \sum_{j=1}^{N} \sum_{i=j+1}^{N} \frac{\text{var}(y_i) \text{var}(y_j)}{(x_j - x_i)^2}.$$  

(4.16)

If the actual sum exceeds twice the standard deviation, then the data can be rejected as non-white. The caveat in using this test is the possibility that certain correlated sequences may also be less than twice the standard deviation of the white
sequence with high probability. However, the weighting of the sample product by the distance between samples ensures that data where the correlation is significant between nearby elements will be rejected.

A polynomial estimate for the mean can easily be extrapolated. However, to gauge the accuracy of the estimate and the extrapolation of the estimate, a calculation of the variance of the mean estimate is useful. If the variance of the mean estimate is much smaller than the variance of the data, then the estimated distribution based on the mean estimate will be accurate. The variance of the mean is

$$\text{var}[\bar{u}(x)] = f^T \Lambda_p f = f^T (G^T \Lambda^{-1}_y G)^{-1} f = f^T Q^{-1} G^T \Lambda_y G Q^{-1} f, \quad (4.17)$$

where the last expression can directly be derived by calculating the variance of

$$\bar{u}(x) = \tilde{f}^T \tilde{p}. \quad (4.18)$$

The variance of the mean line provides a standard for bounding the mean estimate; however, this bound can be misleading if the polynomial model for the mean is inaccurate. An extrapolation of the mean is shown for three examples, and the standard deviation of the mean line is used to bound the mean line (Figures 4.1c and 4.2c).
Source Model: \( u'(x) = x + x^2 \).

Estimation Model: \( \tilde{u}(x) = p_1 x^3 + p_2 x^2 + p_3 x + p_4 \)

A parameter was rejected if the standard deviation of the parameter exceeded the absolute value of the parameter. The results from one hundred trials are tabulated.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Parameter Accepted</th>
<th>Parameter Rejected</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p_1 )</td>
<td>5 trials</td>
<td>95 trials</td>
</tr>
<tr>
<td>( p_2 )</td>
<td>97 trials</td>
<td>3 trials</td>
</tr>
<tr>
<td>( p_3 )</td>
<td>100 trials</td>
<td>0 trials</td>
</tr>
<tr>
<td>( p_4 )</td>
<td>27 trials</td>
<td>73 trials</td>
</tr>
</tbody>
</table>

Table 4.1

This figure demonstrates hypothesis testing for extraneous parameters based on the specified source model and estimation model. Twenty data points were used in each trial, and the variance, skewness, and kurtosis of the data were 1, 0, and 2.4, respectively.
4.2.3. Regression of the Variance: Polynomial Regression Based Solely on the Independent Parameter

Polynomial regression of the variance of the data is more intricate than regression of the mean because samples values of the deviation from the mean are not exactly known. Specifically, let

\[ Z[x] = (Y[x] - u'(x))^2, \]  

so that

\[ u_2(x) = \text{var}[Y[x]] = E[Z[x]]. \]  

(4.20)

The estimate \( \tilde{u}_2(x) \) cannot be based, however, on independent trails of the random variable \( Z[x] \); instead, data generated from the random variable

\[ \tilde{Z}[x] = (Y[x] - \tilde{u}(x))^2 \]  

(4.21)

must be used. Consequently, differences between the random variables \( Z[x] \) and \( \tilde{Z}[x] \) must be considered in estimating the variance.

In order to estimate the variance, samples of \( \tilde{Z}[x] \) are first calculated. Next, polynomial regression provides an estimate for the expected value of \( \tilde{Z}[x] \). A calculation of the bias between \( \tilde{Z}[x] \) and \( Z[x] \), which is dependent on an estimate of the expected value of \( Z[x] \), then allows calculation of samples of \( Z[x] \). A second polynomial regression then provides a more accurate estimate for the expected value of \( Z[x] \). In some cases, the procedure may be recursively applied to improve the estimate of the variance. Each step in the procedure for estimating the variance of the data is described in detail.
First, independent outcomes of the random process $Z[x]$ are calculated. The samples of the process $z_1...z_N$ are calculated as

$$z_i = (y_i - \bar{u}(x_i))^2, \quad i=1...N. \quad (4.22)$$

These samples are the squared deviations from the mean estimate rather than the squared deviations from the actual mean. Next, a polynomial regression based on the independent parameter is used to calculate the expected value of $\tilde{Z}[x]$. Before completing the polynomial regression, the order of the regression must be chosen. Although the expected value of $\tilde{Z}[x]$ does not generally have the same order as the expected value of $Z[x]$, a model is chosen for $Z[x]$, and the same model is used to estimate $\tilde{Z}[x]$. This model is sufficient to allow the unbiased estimation of the expected value of $Z[x]$, which is the eventual goal of the estimation process.

The regression of $E[\tilde{Z}[x]]$ is analogous to the regression of the mean of $Y[x]$. Again, consider a k-th order regression model with a parameter vector $\tilde{p}$. Note that the parameter vector and the order k are different from the parameters and order of the regression for the mean; identical symbols are reused in order to simplify the discussion. The regression results in an estimate for $E[\tilde{Z}[x]]$, labeled $\tilde{u}_z^*(x)$. The regression estimate is

$$\tilde{u}_z^*(x) = \tilde{f}^T \tilde{p} = \tilde{f}^T Q^{-1} R_2, \quad (4.23a)$$

where

$$\tilde{f}^T = [f_0(x) = 1...f_k(x) = x^k], \quad (4.23b)$$
\[ Q = \begin{bmatrix}
N \sum_{j=1}^{N} x_j & \ldots & N \sum_{j=1}^{N} x_j^k \\
\vdots & \ddots & \vdots \\
N \sum_{j=1}^{N} x_j^k & \ldots & N \sum_{j=1}^{N} x_j^{2k}
\end{bmatrix}, \quad (4.23c) \]

and

\[ R_2 = \begin{bmatrix}
N \sum_{j=1}^{N} z_j \\
\vdots \\
N \sum_{j=1}^{N} z_j x_j^k
\end{bmatrix}. \quad (4.23d) \]

Next, the bias inherent to modeling \( Z[x] \) as \( \tilde{Z}[x] \) is calculated. The bias is

\[ E[Z[x] - \tilde{Z}[x]] = E[(Y[x] - u'(x))^2 - (Y[x] - \tilde{u}(x))^2]. \quad (4.24) \]

Factoring the term within the expectation,

\[ E[Z[x] - \tilde{Z}[x]] = E[(\tilde{u}(x) - u'(x))(2Y[x] - \tilde{u}(x) - u'(x))]. \quad (4.25) \]

Next, the expectation can be written as

\[ E[Z[x] - \tilde{Z}[x]] = E[(\tilde{u}(x) - E[\tilde{u}(x)] + E[\tilde{u}(x)] - u'(x))(2Y[x] - \tilde{u}(x) - E[\tilde{u}(x)] + E[\tilde{u}(x)] - u'(x))]. \quad (4.26) \]

\[ E[Z[x] - \tilde{Z}[x]] = E[(\tilde{u}(x) - E[\tilde{u}(x)])(2Y[x] - \tilde{u}(x) - E[\tilde{u}(x)])] \]

\[ -E[(E[\tilde{u}(x)] - u'(x))^2]. \quad (4.27) \]

\[ E[Z[x] - \tilde{Z}[x]] = \text{cov}((\tilde{u}(x), 2Y[x] - \tilde{u}(x)) - E[(E[\tilde{u}(x)] - u'(x))^2]. \quad (4.28) \]

The two terms in the bias equation correspond to bias within the context of a correct regression of the mean and bias resulting from an error in the model of the mean. The second term cannot be calculated without prior knowledge of the mean of the data, but the expectation is minimized in a summed mean squared sense amongst the set of
possible estimators for the mean. The proof that \( \sum_{i=1}^{N} (E[\tilde{u}(x_i)] - \tilde{u}(x_i))^2 \) is minimized is equivalent to proving that the expectation of the line of best fit is identical to the line of best fit of the expectation (Appendix 4.1).

Next, the bias is calculated based on the assumption that the expected value of the mean estimate is the mean. As long as the mean can be written in terms of the basis functions used in the regression, the expected value of the mean estimate will be the mean. In this case,

\[
E[Z[x] - \tilde{Z}[x]] = \text{cov}(\tilde{u}(x), 2Y[x] - \tilde{u}(x))
\]

\[
= 2\text{cov}(\tilde{u}(x), Y[x]) - \text{var}(\tilde{u}(x)).
\]

The variance of the mean estimate has already been calculated, and the necessary covariance can easily be calculated since the mean estimate is a linear function of the data. Specifically,

\[
\text{cov}(\tilde{u}(x), Y[x]) = \text{cov}(\vec{f}^T Q^{-1} G^T \vec{y}, Y[x]),
\]

where \( \vec{y} \) is a column vector of \( y_1 \ldots y_N \), and the other quantities are defined as before.

Therefore, the covariance is

\[
\text{cov}(\tilde{u}(x), Y[x]) = \begin{cases} 
(\vec{f}^T Q^{-1} \vec{f})(\text{var}(Y[x])), & x \in x_1 \ldots x_N \\
0, & \text{otherwise}
\end{cases}
\]

In words, the mean is positively correlated with a data value if and only if that data value has been utilized to calculate the mean. Since all data values are used for both the mean and the variance estimate, however, only the covariance for these values of the independent parameter are important. In turn, the bias \( E[Z[x] - \tilde{Z}[x]] \) can be expressed as
Next, note that $E[Z[x]] = \text{var}(Y[x])$ and that $f^T (G^T \Lambda_y^{-1} G)^{-1} f$ is a linear combination of $\text{var}(Y[x_1])...\text{var}(Y[x_N])$; in consequence, the bias equation leads to $N$ linear equations relating $E[Z[x]], x \in x_1...x_N$, and $\text{var}(Y[x_1])...\text{var}(Y[x_N])$. Three different approaches may then be utilized to obtain the curve of best fit for the variances. First, the regression result for $\tilde{\mu}_2^*(x)$ may be used as an estimate for $E[Z[x]], x \in x_1...x_N$, and $\text{var}(Y[x_1])...\text{var}(Y[x_N])$ can be calculated from the set of linear equations. Another regression on these variance estimates constitutes the curve of best fit for the variance, labeled $\tilde{\mu}_2(x)$. Secondly, the data values $z_1...z_N$ can directly be used as an estimate for $E[Z[x]], x \in x_1...x_N$; the other steps in the estimation procedure are then repeated. Alternately, the formula

$$E[Z[x] - \tilde{Z}[x]] = \frac{2(f^T Q^{-1} f)(\text{var}(Y[x]))}{-f^T (G^T \Lambda_y^{-1} G)^{-1} f}, \ x \in x_1...x_N . \quad (4.33)$$

may be used to recursively compute the curve of best fit for the variance. Specifically, $\tilde{\mu}_2^*(x)$ can be used as a first estimate for $\text{var}(y_1)...\text{var}(y_N)$ for the right side of the equation, allowing an approximate calculation of the bias, $E[Z[x] - \tilde{Z}[x]]$. Next, the bias estimate is added to the samples $z_1...z_N$, and an improved regression for $\tilde{\mu}_2^*(x)$ is completed using the new data points. The procedure is repeated until $\tilde{\mu}_2^*(x)$ converges to $\tilde{\mu}_2(x)$. The second technique always produces an unbiased estimate for the variance as long as the proper regression model for the variance is chosen (Justification in Appendix 4.2). The other two techniques also seem to converge to the proper result, though no
proof of their bias is given. Plots of the bias corrected variance samples and the associated lines of best fit are shown for the two example data sets (Figure 4.3).

In addition to the development of an unbiased estimator for the variance, a method for choosing the proper regression basis and a measure of the estimator's

**Figure 4.3.** Regression of bias-corrected sample variances provides an estimate of the variance of the data. The bias-corrected points and regression are shown for the 20 and 100 point examples, with an extrapolation of the results also shown for the 100 point example.
accuracy is important. Again, a hypothesis test for the sufficiency of the model and another test for the evaluation of parameters are necessary. The hypothesis tests that validate the model for the mean are also applicable to the variance estimate model. First, if the sign of a parameter is probabilistically indeterministic, then that parameter and its associated basis are considered unnecessary. In order to test the hypothesis that the sign of a parameter is inaccurate, a probabilistic description of the parameters is necessary. The calculation of this probabilistic description and the subsequent formulation of the hypothesis test are identical in technique to the hypothesis test for parameters describing the mean. However, the variance of the parameters now equals

$$\Lambda_{zp} = (G^T \Lambda_z^{-1} G)^{-1} , \quad (4.34)$$

where $\Lambda_z$ is the covariance matrix for $Z[x]$. The covariance matrix for $Z[x]$ can be approximated as a diagonal matrix, and the diagonal elements can be estimated from the approximation for the fourth order statistics of the data. Specifically, the diagonal elements of $\Lambda_z$ are calculated as

$$\text{var}(z_i) = \bar{u}_4(x_i) - (\bar{u}_2(x_i))^2 \quad (4.35)$$

Secondly, the error between the estimate of the variance and the samples of $Z[x]$ are formulated as white noise and so should be tested for uncorrelation as a hypothesis test for a model’s sufficiency. The formulation of the error as a white noise is a useful construct for modeling higher order moments and calculating error bounds on the variance estimate rather than an intrinsic description of the random process; however, this limitation on the noise is useful for choosing regression bases. Again, we wish to utilize the second order moment estimate to describe any second order deterministic information.
remaining in the signal. The autocorrelations of the variance errors are shown for the two data sets (Figure 4.3)

Additionally, a measure of the accuracy of the variance estimate is useful. The covariance matrix of the parameter estimates that comprise \( \tilde{u}_2(x) \) leads to the variance of the estimate \( \tilde{u}_2(x) \). The variance of \( \tilde{u}_2(x) \) is

\[
\text{var}[\tilde{u}_2(x)] = f^T \Lambda \rho f = f^T (G^T \Lambda_\gamma^{-1} G)^{-1} f = f^T G^{-1} G^T \Lambda G G^{-1} f.
\]  
(4.36)

As long as the variance of \( \tilde{u}_2(x) \) is negligible in comparison to \( \tilde{u}_2(x) \), the estimate is a meaningful model for the variance of a beta distribution. The standard deviation in the variance curve is also included with the plots of the variance best fit curves (Figure 4.3).

4.2.4. Regression of the Skewness: Polynomial Regression Based Solely on the Independent Parameter

A polynomial regression is also utilized to describe the skewness of the data, or the cube of the deviation of the data from the mean. However, the polynomial model is assumed to describe the normalized skewness, defined as

\[
u_3^n(x) = E[Q[x]] = \frac{E[(Y[x] - \mu(x))^3]}{(\mu(x))^{3/2}}.
\]  
(4.37)

The rationale for fitting a polynomial model to the normalized skewness, rather than the skewness, is based on the interpretation of the skewness as a shape parameter. If only the mean, variance, and skewness are utilized to describe a distribution, the shape of the distribution depends solely on the normalized skewness. All distributions with identical values for the normalized skewness are scaled and shifted versions of the same distribution, so that the normalized skewness is the critical parameter for describing the shape of the distribution. A constant normalized skewness, which can be estimated with
one parameter using a polynomial model, describes a simple class of distributions. Consequently, a polynomial model for the normalized skewness is used for estimation of the skewness.

The mean and the variance of the data are not known, and the calculated mean and variance are used instead. Specifically, a polynomial model is fit to the random variables

\[ \tilde{Q}[x_i] = \frac{(Y[x_i] - \bar{u}(x_i))^3}{(u_2(x_i))^{3/2}}, \quad i \in 1\ldots N. \]  

(4.38)

The bias of the estimate is then calculated, and the estimate is appropriately corrected. In calculating the bias of the estimate, the polynomial models for the mean and variance are assumed to be correct, so that the estimates for these quantities are unbiased. We now calculate the bias

\[ E[Q[x_i] - \tilde{Q}[x_i]] = E\left[ \frac{(Y[x_i] - u'(x_i))^3}{(u_2(x_i))^{3/2}} \right. \]

\[ - \frac{(Y[x_i] - \bar{u}(x_i))^3}{(u_2(x_i))^{3/2}} \] \quad for \: i \in 1\ldots N \]  

(4.39)

Next, note that \( \bar{u}_2(x) \) can be written as \( u_2(x) + \nu(x) \), where \( \nu(x) \) is a zero mean random variable. Next, a small signal approximation for \( (\bar{u}_2(x))^{3/2} \) is \( (u_2(x))^{3/2} + \frac{3}{2} u_2(x) \nu(x) \).

Consequently, the bias is

\[ E[Q[x_i] - \tilde{Q}[x_i]] = E\left[ \frac{(Y[x_i] - u'(x_i))^3}{(u_2(x_i))^{3/2}} \right. \]

\[ - \frac{(Y[x_i] - \bar{u}(x_i))^3}{(u_2(x_i))^{3/2} + \frac{3}{2} u_2(x_i) \nu(x_i)} \]. \]  

(4.40)

86
The second term in the bias can further be approximated as

\[
\frac{(Y[x_i] - \bar{u}(x_i))^3}{(u_2(x_i))^{3/2}} - \frac{3(Y[x_i] - \bar{u}(x_i))^3 v(x_i)}{2(u_2(x_i))^2}.
\] (4.41)

A reasonable calculation for the expectation of the second term seems difficult and tedious. For now, the term is neglected and the bias is calculated as

\[
E[\tilde{Q}[x_i] - \tilde{Q}[x_i]] = E[\frac{(Y[x_i] - u'(x_i))^3}{(u_2(x_i))^{3/2}} - \frac{(Y[x_i] - \bar{u}(x_i))^3}{(u_2(x_i))^{3/2}}].
\] (4.42)

Rewriting the expression,

\[
E[\tilde{Q}[x_i] - \tilde{Q}[x_i]] = \frac{1}{(u_2(x_i))^{3/2}} E[(\tilde{u}(x_i) - u(x_i))^3 - 3(\tilde{u}(x_i) - u(x_i))^2(Y[x_i] - u'(x_i)) + 3(\tilde{u}(x_i) - u'(x_i))(Y[x_i] - u'(x_i))^2].
\] (4.43)

Noting that each term in the expectation is a covariance, the expectation can be rewritten as

\[
\text{cov}(\tilde{u}(x_i) - u'(x_i))^2, \tilde{u}(x_i))
\]

\[
E[\tilde{Q}[x_i] - \tilde{Q}[x_i]] = -3 \text{cov}(\tilde{u}(x_i) - u'(x_i))^2, Y[x_i]).
\]

\[+ 3 \text{cov}(\tilde{u}(x_i), (Y[x_i] - u'(x_i))^2)
\] (4.44)

Each covariance is then evaluated separately. The third covariance,

\[3 \text{cov}(\tilde{u}(x_i), (Y[x_i] - u'(x_i))^2),
\]

is easily determined to equal \[3(f^T Q^{-1} f)(u_3(x_i)).\] The other two terms in the covariance expression are typically much smaller, but all three covariances can be exactly calculated. The calculations are tedious but straightforward since the mean and the square of the mean are easily expressed as linear operators on a matrix of data. The second term in the bias expression is
\[-3 \text{cov}((\bar{\mu}(x_i) - u'(x_i))^2, Y[x_i]) = -3c_2(x_i) = \]
\[-3(\bar{f}^TQ^{-1}G^T\Lambda_s GQ^{-1}\bar{f} - 2(\bar{f}^TQ^{-1}\bar{f})u_2(y_j)u'(y_j))' \]  
(4.45)

where
\[
\Lambda_s = \begin{bmatrix}
0 & 0 & u'(y_1)u_2(y_1) & 0 \\
0 & 0 & \ldots & 0 \\
u'(x_1)u_2(x_1) & \ldots & 2u'(x_1)u_2(x_1) + u_3(x_1) & \ldots & u'(x_N)u_2(x_1) \\
0 & 0 & u'(x_N)u_2(x_1) & 0
\end{bmatrix}. \quad (4.46)
\]

Next, \(\text{cov}((\bar{\mu}(x_i) - u'(x_i))^2, \bar{\mu}(x_i))\) is a linear combination of the covariances \(c_2(x_i), i = 1...N\) and specifically equals \(\bar{f}^TQ^{-1}G^T\bar{c}_2\), where \(\bar{c}_2 = [c_2(x_1)...c_2(x_N)]\).

The following procedure is then utilized to generate an approximately unbiased estimate for the normalized skewness. The data points
\[
\tilde{q}_i = \frac{(y_i - \bar{\mu}(x_i))^3}{(\bar{u}_2(x_i))^{3/2}}, i \in 1...N \quad (4.47)
\]
are calculated. Next, a linear regression is completed on these data points, and a biased set of parameter estimates is determined. The regression result is utilized to calculate the bias in each data sample, and the bias is removed. Another regression is then used to generate an unbiased estimate for the normalized skewness. Finally, the skewness estimate is calculated as
\[
\bar{u}_3(x) = \bar{u}_3^n(x)(\bar{u}_2(x))^{3/2}. \quad (4.48)
\]

Again, the variance in the normalized skewness estimate can be calculated, and so hypothesis testing is possible. One significant difference in the estimate variance calculation should be noted. To estimate the variance of the normalized skewness, a
model for the sixth moment of the data is necessary. For simplicity, we assume that the variance in the data samples \( \bar{q}_i \) is a constant for all \( i \). This constant can be directly calculated by averaging the squared deviation of the \( \bar{q}_i \) from their best fit curve. The variance in the parameters and the skewness estimate can then be calculated. The skewness estimate and associated error in the estimate are shown for the two example data sets (Figure 4.4).

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{image1.png}
\caption{Extrapolated Normalized Skewness Curve and Variance}
\end{figure}

An extrapolated estimate for the normalized skewness is shown for the two example data sets.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{image2.png}
\caption{Extrapolated Normalized Skewness Curve and Variance}
\end{figure}
4.2.5. Regression of the Kurtosis: Polynomial Regression Based Solely on the Independent Parameter

The analysis of the bias in the skewness estimate was quite complicated, and we do not attempt a similar analysis for the regression of the kurtosis. Instead, we present an estimator for the kurtosis without an explicit calculation of the bias in the estimator. Conceptually, the estimator for the kurtosis uses the corrected sample variances, and so the estimator bias should be small.

A polynomial model is utilized to describe the normalized kurtosis, which is denoted as

\[ u_4^*(x) = E[Q(x)] = \frac{E[(Y[x] - u'(x))^4]}{(u_2(x))^2}. \] (4.49)

The normalized kurtosis rather than the kurtosis is modeled with a polynomial because the normalization automatically accounts for the shift and spread of the set of random variables. Consequently, a model with a smaller order may be sufficient for describing the normalized kurtosis.

Since the mean and the variance are not exactly known, we must use our estimates for these quantities. Instead of using the sample mean and variance and then correcting for the bias, however, we use the corrected variance samples. The resulting estimate is still biased, but Monte Carlo trials demonstrate that the bias is small.

Specifically, the data points

\[ r_i = \frac{(z_i + E[Z(x_i)] - \bar{Z}(x_i))^2}{(\bar{u}_2(x_i))^2}, \quad i \in 1...N \] (4.50)
are calculated as samples of the kurtosis. A polynomial regression is then completed to estimate the normalized kurtosis. Finally, the normalized kurtosis estimate is scaled by the square of the variance estimate to determine $\tilde{u}_4(x)$.

Again, the accuracy of the parameter estimates can be determined. The variance in the normalized kurtosis is assumed to be a constant and so this variance is calculated. We can then determine the variance of the kurtosis parameter estimates and so the variance in $\tilde{u}_4(x)$. Examples of the kurtosis regression are again illustrative (Figure 4.5).

**Figure 4.5.**
An extrapolated estimate for the normalized kurtosis is shown for the two example data sets.
4.2.6. Regression of Moments for the Estimation of Beta Distributions

Though the developed regression technique has focused on the adequate estimation of the four central moments of the data as a function of the independent parameter, the final objective of the regression is the estimation of a continuum of the beta distributions. Based on the four moments calculated at each independent parameter value, the beta distribution that represents the probabilistic description of the data at each independent parameter value can directly be calculated.

An attempt to calculate the beta distribution as a function of the independent parameter highlights some possible caveats in the estimation technique. Most importantly, even if the model for describing the data is correctly chosen, the regression results may produce moment values that do not correspond to any beta random variable. The probability of the occurrence of an estimate that is not a beta distribution can be calculated based on the calculated variance of the moment estimate. More importantly, however, a remedy for these estimates must be suggested. We propose an adjustment to kurtosis of the estimated distribution to force a proper beta distribution estimate. The rationale for adjusting the kurtosis is based on the observation that the variance in the kurtosis estimate tends to be larger than the variances of the other estimate, so that a significant variation in the kurtosis estimate is probabilistically likely. Alternately, both the normalized skewness and kurtosis can be adjusted to produce a beta distribution; the adjustment is smaller in a mean squared sense than the adjustment of just the normalized kurtosis, but the correction is more difficult to implement.

The high variance in the estimates for the normalized skewness and kurtosis and the resulting uncertainty in the shape of the beta distribution estimate is another important
observation. Typically, for data sets with less than 50 points, the relationship between the independent parameter and skewness and kurtosis is difficult to determine. Consequently, we recommend modeling the normalized skewness and normalized kurtosis as constants for data sets with less than 50 data points. For data sets with less than approximately 15 points, the data often cannot be significantly differentiated from data generated from a Gaussian, so that regression of the first two moments may be sufficient to adequately model the data. The estimated distribution in the 100 point example is compared to the source distribution of the data at one independent parameter value (Figure 4.6).

Figure 4.6.
Comparison of probability and cumulative distribution functions at x=0 for the regression result (100 point example).
4.3. Regression of Moments Based on an Independent Parameter and on Other Data Samples.

We consider models for data that depend on the independent parameter but also allow for correlation between data samples. Although the analysis is simply a generalization of the independent parameter based regression presented before, the implications of these models are significant.

The presentation in this section is rather different from the development of the solely independent parameter based model. Specifically, the development focuses on the analysis of time series, which require correlation sensitive models, rather than on the general analysis of the procedure. Also, instead of analyzing the regression procedure in detail, we focus on distinct aspects of the new models, including differences in interpolation and extrapolation.

4.3.1. Why Are Time Series Different?

Certain models describe types of data sets more accurately than other models. In this study, we seek models that capture any deterministic information in a data set in the descriptions of the mean and central moments. In some problems, the moments can be accurately modeled as polynomial or other functions of an independent parameter. Sometimes, however, correlation between data samples that does not depend functionally on the independent parameter is present, and the proper model must attempt to capture this useful information. The rationale for using models that capture correlation effects is provided for the specific problem of time series analysis.
4.3.1.1. A Rationale for a Different Method-Introduction

We believe that the analysis of time series data requires a different methodology from the analysis of other data sets that vary as a function of another independent parameter. In both problems, we wish to describe the data with a continuum of beta distributions in order to evaluate the possible range of a dependent variable at specific values of the independent variable. In both problems, we will use a set of data points to generate the model. The two problems are fundamentally different, however, because of differences in the known set of data samples and because of the different objectives in the two analyses. The use of two different models for the analysis of time series and the analysis of data that varies with another independent parameter improves the results obtained by the analysis.

4.3.1.2. A Rationale for a Different Model-Two Reasons

First, relationship of the known set of data to the independent parameter varies between the two problems. In the analysis of non-temporal data, the method of generating a data point may be envisioned as a two-step process. First, a value of the independent parameter is chosen, and then a data point is chosen based on the possible distribution of the dependent variable at that value of the independent parameter. Consequently, we expect that the important information contained in the data points corresponds to a fixed set of beta distributions. In turn, we hope that the moments of this set of beta distributions can be described with groups of basis functions, such as polynomials. We believe that these basis functions describe the possible distribution of data points not only at the values of the independent parameter with known data but also
at intermediate and extrapolated values for the independent parameter. These basis functions constitute a sort of underlying principle that determines the value of the dependent value in terms of the independent parameter. In contrast, time series data, such as metals market prices, are difficult to describe as originating from a fixed set of distributions. Instead, the origin of the data may be envisioned as a distribution that depends on the extent of the known data. For example, we would be more capable of guessing the price of copper in August 2000 in July 2000 than in July 1999, so that the distribution of copper prices for that month would change. In modeling time series, only a small fraction of information in the data is contained in underlying basis functions, such as a linear increase in copper prices with large fluctuations around that overall mean. The remainder of the information is dependent on the extent of known data. Consequently, a beta distribution model for copper prices is formulated by relating the moments of the distribution at a certain time to the known value of the data at previous times. The set of data is used to determine the relationship between data points and the moments of future time series values.

The difference between the time series analysis and the modeling of data as a function of another independent parameter may also be interpreted as a difference in the objectives of the two studies. For non-temporal analysis, we seek a set of rules that determine the possible range of the dependent variable at a certain value for the independent variable. We assume that this set of rules is not altered because we have more data, even though the increased data may allow us to more accurately determine the model. In the case of time-series analysis, we do not seek a set of rules that specifies a constant distribution for future time series values. With the exception of some general
long term trends, such as a linear increase in copper prices with a large variance about this average, the rules for estimating times series values are based on other data points rather than the independent parameter. As the number of data points increases, a new distribution for future time series values must be calculated. Consequently, we seek to generate a distribution of the dependent variable that accounts for the known samples of the random process as other times. In fact, we might argue that the time series data points at past instances are not guided by distributions and are instead deterministic values; future values of the time series data depend probabilistically on these past values.

4.3.1.3. The Philosophical Difference Between Time Series and Other Data

Most generally, the analysis of both types of data becomes a philosophical question. For example, we might wonder if the values of a time series at certain times are in fact generated by a specific distribution that creates the data values with high probability. In fact, is it possible that the value of a time series at any time is a deterministic value that we simply do not know? This query may also be considered for data that varies as a function of another independent parameter. Similarly, we might wonder if non-time series data actually depends on other data samples, in addition to the independent parameter. Regardless of the answers to these questions, our techniques for modeling this data attempt to extract the information that we believe to be pertinent from the data. Two different models are most efficient in describing time-series data and other data.
4.3.2. Two Models for Time Series Analysis

We have considered several different models to describe correlated data and specifically to model time series. The linear regression technique allows modeling both based on the independent parameter and based on existing data samples. However, this general method of regression requires many parameters and is rather cumbersome, so we focus on modeling stationary processes; we additionally describe heuristic procedures for converting time series into stationary processes. More general procedures for separating independent parameter-based data and correlation information are discussed in a later section. The developed models closely resemble standard autoregressive models but are generalized to allow modeling of four moments.

4.3.2.1. Removal of Independent Parameter-Dependent Information from the Time Series

The following heuristic procedure is used to convert a time series into an approximately stationary process. The mean of the time series is described as a function of the independent using a low order polynomial. This independent parameter dependent mean is subtracted from the time series. Based on the time series that we observe, we then divide the time series by the absolute value of the mean line to produce a stationary process. The variance in many time series seems to increase with the mean, so that normalization by the mean produces a stationary time series. More general processes for determining the independent parameter-dependent information in data sets are described in Chapter 7.
4.3.2.2. Regression Models for the Transformed Time Series: Introduction

Two different models are utilized to describe the transformed time series data. Both models are similar in that they provide a beta distribution description for the unknown values of the time series given a set of known values for the time series at other times. In contrast to the modeling of data that does not vary with time, the resulting distributions as each time instant are only valid for a specific set of known temporal data. If data is gathered for other time instances, the set of distributions will change because they depend data samples in addition to the independent parameter. In both formulations of the regression model, a discrete-time data series is first considered to simplify the analysis and the analogous continuous time formulation is considered later.

For both models, assume that the value of the time series is known for \( x \in \{ x_1 = 1 \ldots x_N = N \} \), where \( x \) is the independent parameter, time. The values of the dependent variable are known at these time instances and are labeled \( y[1] \ldots y[N] \) for convenience. These data points are assumed to originate from a zero-mean random process since any dependence of the data mean on the independent parameter has been removed. The resulting process is not necessarily stationary, however, since the higher moments of the data or the relationship between data samples may also depend on the independent parameter. For now, the heuristic process for transforming the data sequence to a stationary process discussed in the previous section is completed. The discussed techniques for generating beta distribution models for random processes are also valid for non-stationary processes but require a more complicated model. The following discussion assumes that \( y[1] \ldots y[N] \) originate from a zero mean, strict sense stationary process.
4.3.2.3. Analysis of the First Model

The first model for the time series data is only applicable for a constrained set of time series but provides a clear conceptual description of the random process. In this model, the mean of the data given previous samples of the time series is a linear function of these previous data samples. Meanwhile, the higher order central moments of the random process given previous samples of the data are modeled as constants. The parameter estimation for the model is a straightforward extension of the autoregressive modeling technique, while the estimation of future time series values using the developed technique is an extension of Wiener filtering [7]. The mean of the data at time \( n \) given the data values at \( k \) previous instances is modeled as

\[
u'(x_n, y[n-1]...y[n-k]) = E[y[n]|y[n-1]...y[n-k]] = \bar{p}^T \bar{y},
\]

where

\[
\bar{p}^T = [p_1,...,p_k]
\]

is a vector of parameters and

\[
\bar{y}^T = [y[n-1]...y[n-k]]
\]

is a vector of the previous \( k \) data points. Additionally, the variance is modeled as

\[
u_2(x_n, y[n-1]...y[n-k]) =
E[(y[n] - \nu'(x_n, y[n-1]...y[n-k]))^2 | y[n-1]...y[n-k]] = p_{s,1},
\]

the skewness is modeled as

\[
u_3(x_n, y[n-1]...y[n-k]) =
E[(y[n] - \nu'(x_n, y[n-1]...y[n-k]))^3 | y[n-1]...y[n-k]] = p_{s,2},
\]

and the kurtosis is modeled as
\[ u_n(x_n \mid y[n-1]...y[n-k]) = \]
\[ E[(y[n]-u'(x_n \mid y[n-1]...y[n-k]))^4 \mid y[n-1]...y[n-k]] = p_{k+3} \]

Since a linear model for the conditional moments has been formulated, a regression based on the known data allows calculation of the parameters of the model. Specifically, the parameters \( \hat{p}^T = [p_1,...p_k] \) can be estimated as

\[
\hat{p} = Q_{p1}^{-1} R_{p1} = \left[ \begin{array}{cccc}
\sum_{j=k+1}^{N} (y[j-1])^2 & \sum_{j=k+1}^{N} y[j-2]y[j-1] & \cdots & \sum_{j=k+1}^{N} y[j-k]y[j-1]
\end{array} \right]^{-1} \left[ \begin{array}{c}
\sum_{j=k+1}^{N} y[j]y[j-1]
\end{array} \right].
\]

The estimate for the mean given \( k \) previous samples then is

\[ \bar{u}(x_n \mid y[n-1]...y[n-k]) = \hat{p}^T \bar{y}. \]

Next, the estimates for the parameters describing the higher order moments are calculated as

\[
p_{k+1} = \frac{1}{N-k} \sum_{j=k+1}^{N} (y[j]-\bar{u}(x_n \mid y[n-1]...y[n-k]))^2
\]

\[
p_{k+2} = \frac{1}{N-k} \sum_{j=k+1}^{N} (y[j]-\bar{u}(x_n \mid y[n-1]...y[n-k]))^3
\]

\[
p_{k+3} = \frac{1}{N-k} \sum_{j=k+1}^{N} (y[j]-\bar{u}(x_n \mid y[n-1]...y[n-k]))^4
\]

for the variance, skewness, and kurtosis, respectively. These estimates are biased in general. In order to remove the bias, the techniques developed for independent parameter-based regression can be repeated for this regression. However, for most of the time series described, the number of available samples is large, so that the bias in the estimated conditional moments will be small if the model is correctly chosen.
Although the parameters of the conditional moments have been determined, the distribution of the sample at arbitrary times beyond the known data should be quantified. A recursive calculation of the time series moments for $x > N$ is described. To generate the mean estimate, note that the unconditional mean at an arbitrary integer time $x$ may be estimated as

$$
E[y[x]] = \bar{u}(x) = \int_{-\infty}^{\infty} \bar{u}(x \mid y[n-1]...y[n-k]) f_\varphi(\bar{y}) d\bar{y}.
$$

(4.61)

Substituting the parametric form for the conditional mean estimate, the unconditional mean estimate becomes

$$
\bar{u}(x) = \int_{-\infty}^{\infty} \bar{p}^T \bar{y} f_\varphi(\bar{y}) d\bar{y} = \bar{p}^T E[\bar{y}] \approx \bar{p}^T \bar{u}(\bar{x}),
$$

(4.62)

where

$$
\bar{u}(\bar{x}) = [\bar{u}(x-1)...
\bar{u}(x-k)].
$$

(4.63)

The expression for the mean estimate is a $k$-th order difference equation. Further, $y[N-1]...y[N-k]$ constitute $k$ initial conditions for the difference equation, so that the mean estimate can be calculated for all integer $x$ greater than $N$.

For this formulation, the recursions describing the higher order moments become cumbersome for $k > 1$. We will complete the more general formulation of the model in the future but concentrate on the analysis for $k = 1$ in this study. The recursion describing the variance at each value of the independent parameter is described in detail; the recursions of the higher order moments are more tedious to derive, so the results are stated without derivation. The procedure for the higher order moments is similar to the analysis of the second central moment. Specifically, the unconditional variance at an integer $x$ can be estimated as
\[ \bar{u}_2(x) = E[(y[x])^2] - (\bar{u}(x))^2. \]  
\hspace{1cm} (4.64)

Next,
\[ E[(y[x])^2] = \int_{-\infty}^{\infty} E[(y[x])^2 \mid y[x-1]] f_{y[x-1]}(y_{x-1})dy_{x-1}. \]  
\hspace{1cm} (4.65)

Since
\[ u_2(x, y[n-1], \ldots, y[n-k]) = p_{k+1} = \bar{p}_{k+1}, \]  
\hspace{1cm} (4.66)
\[ E[(y[x])^2 \mid y[x-1]] \bar{p}_{k+1} + (E[y[x] \mid y[x-1]])^2. \]  
\hspace{1cm} (4.67)

Substituting the conditional expectation,
\[ E[(y[x])^2] = \bar{p}_{k+1} + p_1^2 E[y[x-1]^2]. \]  
\hspace{1cm} (4.68)

Consequently, the estimate for the variance becomes
\[ \bar{u}_2(x) = \bar{p}_{k+1} + \bar{p}_1^2 \bar{u}_2(x-1). \]  
\hspace{1cm} (4.69)

Similarly, the third and fourth central moments are estimated as
\[ \bar{u}_3(x) = \bar{p}_{k+2} + \bar{p}_1^3 \bar{u}_3(x-1) \]  
\hspace{1cm} (4.70)

and
\[ \bar{u}_4(x) = \bar{p}_{k+3} + \bar{p}_1^4 \bar{u}_4(x-1) + 6 \bar{p}_1^2 \bar{p}_{k+1} \bar{u}_2(x-1). \]  
\hspace{1cm} (4.71)

Next, since \(y[N-1]\) is known,
\[ E[y[N-1]] = y[N-1], \]  
\hspace{1cm} (4.72)

while the higher central moments of \(y[N-1]\) are zero. Consequently, initial conditions are known for each moment recursion, and the moments can be calculated for future instances of the time series.
4.3.2.4. Analysis of the Second Model

The second model is less intuitive but is capable of representing a larger class of stochastic processes. Also, the calculation of distributions for future values of the time series is simplified regardless of the number of known data points utilized for the estimation. However, a larger number of parameters is typically necessary for the model, and hypothesis testing of parameters is sometimes necessary.

In contrast to the first model, which parametrizes the central moments of the data, the second model describes each non-central moment as a quartic polynomial of previous data samples. Specifically, the mean is described as

\[
u'(x_n \mid y[n-1]...y[n-k]) = E[y[n] \mid y[n-1]...y[n-k]] = \tilde{p}_1^T \tilde{y}, \tag{4.73}
\]

where

\[
\tilde{y}_q = [1, y[n-1],...y[n-k],(y[n-1])^2,...(y[n-k])^2,...(y[n-k])^4]
\]

and \(\tilde{p}_1^T\) is a vector consisting of 5k+1 parameters. Similarly, the second moment is modeled as

\[
u'(x_n \mid y[n-1]...y[n-k]) = E[(y[n])^2 \mid y[n-1]...y[n-k]] = \tilde{p}_2^T \tilde{y}, \tag{4.75}
\]

the third conditional moment is modeled as

\[
u'(x_n \mid y[n-1]...y[n-k]) = E[(y[n])^3 \mid y[n-1]...y[n-k]] = \tilde{p}_3^T \tilde{y}, \tag{4.76}
\]

and the fourth moment is modeled as

\[
u'(x_n \mid y[n-1]...y[n-k]) = E[(y[n])^4 \mid y[n-1]...y[n-k]] = \tilde{p}_4^T \tilde{y}. \tag{4.77}
\]

For many problems, only a linear dependence of the moments on the data samples \(y[n-2]...y[n-k]\) is sufficient, and occasionally a linear dependent of the moments on \(y[N-1]\) is also sufficient. Alternately, the more general model may be used at first, and
unnecessary parameters can be set to zero based on hypothesis testing of the parameters. Another analysis of the model using the smaller parameter vector is then possible.

Next, linear regression provides an estimate of the parameter vectors. Each conditional moment is independently estimated using the standard regression technique. For notational simplicity, the resulting parameter vector estimates are combined into a matrix:

\[
P = \begin{bmatrix}
  P_1^T \\
  P_2^T \\
  P_3^T \\
  P_4^T
\end{bmatrix}
\]

After estimates for the parameters have been determined, distributions that estimate future values of the time series can be calculated. Specifically, unconditioned estimates for each moment are recursively calculated. The four unconditioned moments may be written as

\[
E[(y[x])^q] = \tilde{u}_q(x) = \int_{\mathbb{R}} \tilde{u}_q(x | y[n-1]...y[n-k]) f_y(\bar{y}) \, d\bar{y}, q \in \{1,...,4\}
\]

Since each conditional expectation is a fourth order function of \(y[n-1]...y[n-k]\), \(E[(y[x])^q]\) is a function of the four moments of \(y[x-1]...y[x-k]\). Specifically, the vector of parameters

\[
\tilde{u}_q(x) = \begin{bmatrix}
  \tilde{u}_1(x) \\
  \tilde{u}_2(x) \\
  \tilde{u}_3(x) \\
  \tilde{u}_4(x)
\end{bmatrix}
\]

is recursively estimated as \(P\tilde{u}_{x-1}\), where

\[
\tilde{u}_{x-1}^T = [1, \tilde{u}(x-1), \tilde{u}(x-k), (\tilde{u}(x-1))^2, ..., (\tilde{u}(x-k))^2, ..., (\tilde{u}(x-k))^4].
\]
Y[N]...y[N-k+1] provide the initials condition for this set of difference equations, and the moment estimates can subsequently be calculated for each future time instance. Further, estimates of the central moments can easily be calculated from the estimates of the moments around zero. Surprisingly, the moment estimates are unbiased, but the calculated central moments will again be biased.

The second model differs from the first model in that multiple previous samples can easily be used to describe the data at each independent parameter value. Several examples of estimation with the second model demonstrate, however, that hypothesis testing to remove excess parameters is extremely important to ensure the stability of the estimate and to prevent negative variance and kurtosis extrapolations.

4.3.2.5. Estimation of Silicon Prices

The actual behavior of the two models in extrapolating time series data is most easily demonstrated by applying the models to extrapolate real data sets. We consider the estimation of average monthly silicon prices for the forty-eight months beginning in January 1995 based on silicon price data for the years 1987-1994. The moment estimates as a function of time are plotted for each data set, and the distribution of silicon prices in January 1995 and December 1996 are shown for each model (Figures 4.7, 4.8).
Figure 4.7.
Estimation of silicon prices using the first model. Estimates of the mean, standard deviation, and higher moments are shown (a,b,c,d,e). The predicted probability distributions at times of 2 months and 24 months are also shown (f,g). The time axis is scaled to vary from 0 to 1 for the known data, which spans eight years.
Figure 4.8.

Estimation of silicon prices using the second model. Estimates of the mean, standard deviation, and higher moments are shown (a,b,c,d,e). The predicted probability distributions at times of 2 months and 24 months are also shown (f,g). The time axis is scaled to vary from 0 to 1 for the known data, which spans eight years.
The predicted probability distributions from the two models are rather different for the first month after the final known data sample but are similar for predictions beyond several months. The difference in the estimated probability distribution in the first few predictions seems to originate from a tendency of the two models to stress different characteristics of the data. Specifically, the first model suggests a small variance in the first prediction but admits the possibility of a data point far from the mean because of the skewness of the distribution. Meanwhile, the second model allows for a larger variance but concentrates the points near the mean with a nearly Gaussian distribution. A comparison of estimation with the second model using two different lengths of data for regression is illustrated (Figure 4.9). As data samples are modeled as a function of many previous data points, the estimated moments exhibit sinusoidal fluctuations.
Figure 4.9.
This figure compares extrapolation results of silicon prices using the second model with two different filter lengths. 4.10a and 4.10b describe each silicon price based on 3 previous prices, while 4.10c and 4.10d are results of an eight element filter model.
4.3.2.6. Interpolation using the Two Time Series Models

Because the time series models describe the correlation between adjacent data samples, they are also useful for interpolating between data samples. Although prediction is usually the most important question in analyzing time series, interpolation may sometimes be useful. For example, historical climatological data may be known for certain time periods but may not be known in other periods. An interpolation procedure can be utilized to determine distributions for intermediate time series values. Unfortunately, the problem of estimating intermediate data values from intervals of known data is not a simple question. Specifically, the procedure for solving the prediction problem using the two described models cannot easily be used to solve the interpolation problem. However, the first two moments of the interpolated distributions can be determined using the first model based on the assumption that the data is modeled properly with a Gaussian distribution. Although the procedure for interpolating time series was developed and implemented in Matlab software, the topic is not quantitatively discussed in this thesis.
Appendix 4.1

We prove that \( \sum_{i=1}^{N} (E[\tilde{u}(x_i)] - u'(x_i))^2 \) is minimized among the set of estimators.

First, we note that \( E[\sum_{i=1}^{N} (\tilde{u}(x_i) - u'(x_i))^2] \) is minimized iff \( \tilde{u}(x_i) \) is the regression estimate for the given set of data: \( \tilde{u}(x) = \tilde{f}^T \tilde{\beta} = \tilde{f}^T Q^{-1} R = \tilde{f}^T Q^{-1} G^T \tilde{y} \). Taking the expectation of each side of the expression, we find that \( E[\tilde{u}(x)] = \tilde{f}^T Q^{-1} G^T E[\tilde{y}] \). Consequently, \( E[\tilde{u}(x)] \) is the regression estimate for \( E[u'(x)] = u'(x) \), and \( \sum_{i=1}^{N} (E[\tilde{u}(x_i)] - u'(x_i))^2 \) is minimized amongst the set of estimators.

Appendix 4.2

For notational simplicity, we define a column vector composed of \( z_1 ... z_N \) as \( \lambda \).

Similarly, \( \text{var}(Y[x_1]) ... \text{var}(Y[x_N]) \) is combined into the vector \( \nu \). Based on equation 4.33, \( \nu = LE[\lambda] \), where \( L \) is a linear operator. The exact expression for \( L \) is easily determined but is not necessary for the proof.

Next, the vector of parameters determined using the first bias correction procedure is calculated. Specifically, the operator \( L \) is utilized on the vector \( \lambda \), and linear regression using these corrected observations is completed. The parameter vector obtained through regression is given by \( \tilde{\beta} = Q^{-1} G^T L \lambda \), and the expected value of the parameters are \( E[\tilde{\beta}] = E[Q^{-1} G^T L \lambda] = Q^{-1} G^T LE[\lambda] = Q^{-1} G^T \nu \). Next, note that \( E[\tilde{\beta}] \) is equivalent to the parameters that describe the line of best fit of the variance vector \( \nu \).
Therefore, as long as the set of estimators includes the function that correctly fits the variances, this estimate is unbiased.
Chapter 5

Regression Using Cumulative Distribution Function Fitting

In this chapter, the joint sample cumulative distribution function of a data set defined by two parameters is fitted with an appropriate continuum of beta cumulative distribution functions. A parametric model is selected to describe the data set, and the parameters of the model are determined by minimizing the squared difference between the sample cumulative distribution and the parametrized model.

The technique for determining a probabilistic model with a cumulative distribution function that closely resembles the sample cumulative distribution, or the fraction of data points that are smaller than a certain value, is based on the concept that the sample cumulative distribution will approximate the cumulative distribution of the probability function producing that data. Consequently, the valid cumulative distribution which achieves the minimal mean squared error to the cumulative distribution function of the data is chosen as the proper probabilistic description of the data. This technique is attractive because bounds can be determined for the range of cumulative distributions that can describe a data set, because the convergence of the result to the proper cumulative and probability distribution is guaranteed asymptotically, and because the technique is independent of the type of distribution used for the model. Further, the technique provides a natural rubrik for the extrapolation of the probabilistic model and facilitates hypothesis testing of the form of the parametric model describing the data.
Although the primary objective is the calculation of a continuum of distributions to describe a data set, a simpler problem of the determination of a single distribution based on multiple data points generated using this distribution is first analyzed. Specifically, we formulate the problem of a single cumulative distribution using the sample cumulative distribution; further, we determine bounds on the error in the estimated distribution, consider the asymptotic properties of the estimator, and describe some modifications of the estimator that weight the importance of parts of the estimate differently. This analysis provides a method for describing the distribution of the independent parameter values in the original problem. Further, the technique can be generalized to calculate the joint cumulative distribution of the data as a function of the independent parameter and the dependent parameter. Finally, the use of the cumulative distribution function for extrapolation and hypothesis testing is described.

5.1. Determination of a Single Cumulative Distribution Function

5.1.1. Model Formulation and Cumulative Distribution Function Fitting Technique

Consider N independent, identically distributed continuous random variables \( Y_1, \ldots, Y_N \) with an arbitrary probability density function \( f_Y(y) \). The cumulative distribution function of \( Y_1, \ldots, Y_N \), or the probability that \( Y_i \) is less than some index variable \( x \) is given by

\[
F_Y(y) = \int_{-\infty}^{x} f_Y(a) da
\]  

(5.1)

Next, consider realizations of the N random variables, labeled \( y_1, \ldots, y_N \). A cumulative distribution function of the data is then defined using these realizations of the random
variable. Specifically, the data or sample cumulative distribution function \( \tilde{F}_y(y) \) is defined as

\[
\tilde{F}_y(y) = \frac{1}{N} \sum_{i=1}^{N} u(y - y_i),
\]

where

\[
u(a) = \begin{cases} 
0, & a < 0 \\
1, & a \geq 0
\end{cases} \tag{5.3}
\]

The problem of interest is the estimation of \( F_y(y) \) from \( \tilde{F}_y(y) \). Next, some constraints on the form of the estimate for \( F_y(y) \), labeled \( \hat{F}_y(y) \), are necessary. Specifically, assume that

\[
\hat{F}_y(y) = G(p_1, \ldots, p_k, y), \tag{5.4}
\]

where \( G(p_1, \ldots, p_k, y) \) is the form of the cumulative distribution function determined by \( k \) parameters and \( \tilde{p}_1, \ldots, \tilde{p}_k \) are appropriately chosen parameter values. For example, \( G(p_1, \ldots, p_k, y) \) may be the cumulative distribution function of a beta random variable.

The method used to determine \( \tilde{p}_1, \ldots, \tilde{p}_k \) is

\[
\tilde{p}_1, \ldots, \tilde{p}_k = \arg \min_{p_1, \ldots, p_k} \int \left[ G(p_1, \ldots, p_k, y) - \tilde{F}_y(y) \right]^2 dy. \tag{5.5}
\]

A rationale for this method for parameter determination is provided in analyzing the accuracy of the cumulative distribution function method. Considerable research on fitting cumulative distribution functions using several measures, including the minimization of the squared error and the minimization of the maximum deviation, have been completed [19,20]. We use the results of this research and also extend some of the results.
Next, bounds are formulated for the possible cumulative distribution functions that can generate a given set of data. If the estimated cumulative distribution is contained within the bounds, the form of $G(p_1,\ldots, p_k, y)$ is deemed appropriate, and the chosen estimate is a reasonable estimate for $F_y(y)$. However, if the estimated cumulative distribution is not contained within the bounds, a new form for the function $G()$ is necessary, and the distribution fitting procedure must be repeated.

The analysis of the cumulative distribution fitting technique will first entail the determination of bounds on a valid cumulative distribution and on the mean squared error between the estimated and actual cumulative distribution functions. Secondly, the asymptotic properties of the estimated cumulative distribution function will be considered. Thirdly, the bias and variance of the estimator will be determined.

5.1.2. Formulation of Error Bounds

The possible range of cumulative distributions that sufficiently describe a data set is calculated. First, the possible range of sample cumulative distributions for a given cumulative distribution $F_y(y)$ is determined. Next, the set of all distributions for which a specific data cumulative distribution is within the valid range is calculated; subsequently, bounds on the distributions that are capable of generating the data cumulative distribution can be generated.

First, we bound the cumulative distribution of the data for the random variable $Y$. The cumulative distribution of the data, $\tilde{F}_y(y)$, is a discrete random variable which equals a value in the set $\{0, \frac{1}{N}, \frac{2}{N}, \ldots, \frac{N-1}{N}, 1\}$ with non-zero probability. The probability
mass function for $\tilde{F}_y(y)$ can then be calculated. The probability that $\tilde{F}_y(y)$ equals $\frac{k}{N}$, where $k$ is an integer between 0 and $N$, is the probability that any $k$ of the $N$ points generated by the distribution are less than $y$. Therefore, at each $y$, $\tilde{F}_y(y)$ is a binomial random variable, and

$$\Pr[\tilde{F}_y(y) = \frac{k}{N}] = \binom{N}{k} (F(y))^k (1 - F(y))^{N-k}. \quad (5.6)$$

Further, the mean and variance of $\tilde{F}_y(y)$ are

$$E[\tilde{F}_y(y)] = F_y(y) \quad (5.7a)$$

and

$$\text{var}(\tilde{F}_y(y)) = \frac{1}{N} (F_y(y))(1 - F_y(y)). \quad (5.7b)$$

Since the probability mass function for $\tilde{F}_y(y)$ is exactly known, a variety of bounds for $\tilde{F}_y(y)$ can be formulated for a given random variable $Y$. A commonly used bound is

$$F_y(y) - 2\sqrt{\text{var}(\tilde{F}_y(y))} \leq \tilde{F}_y(y) \leq F_y(y) + 2\sqrt{\text{var}(\tilde{F}_y(y))}. \quad (5.8)$$

The further constraint that

$$0 \leq \tilde{F}_y(y) \leq 1 \quad (5.9)$$

is useful since $F_y(y) - 2\sqrt{\text{var}(\tilde{F}_y(y))}$ and $F_y(y) + 2\sqrt{\text{var}(\tilde{F}_y(y))}$ can be less than zero or greater than one, respectively, while the sample cumulative distribution is bounded between zero and one. More generally, a bound may be found such that

$$F_y(y) - \alpha(y) \leq \tilde{F}_y(y) \leq F_y(y) + \gamma(y), \quad (5.10)$$

where $\alpha(y)$ and $\gamma(y)$ are chosen so that
\[ \Pr[\tilde{F}_y(y) - F_y(y) > \gamma(y)] = \Pr[F_y(y) - \tilde{F}_y(y) > \alpha(y)] = q, \] (5.11)

where \( q \) is some small probability. We will employ the bound based on the standard deviation rather than the more general bound.

Before determining bounds on the set of cumulative distribution functions that can generate a known data cumulative distribution, one further observation is expedient. Subject to the additional constraint that \( 0 \leq \tilde{F}_y(y) \leq 1 \), the upper and lower bounds on \( \tilde{F}_y(y) \) are both non-decreasing functions of \( F_y(y) \). Consequently, for a given sample cumulative distribution, only cumulative distributions in a continuous range around the data cumulative distribution will generate that data cumulative distribution with high probability. The endpoints of this continuous range are considered bounds on \( F_y(y) \).

By solving the standard deviation-based bound equation for \( F_y(y) \), we find that the cumulative distribution function is constrained as

\[ r_1(y) < F_y(y) < r_2(y), \] (5.12)

where \( r_1(y) \) and \( r_2(y) \) are the two roots of the equation

\[ g(r) = (4 + N)r^2 - (2NFFFFF_y(y) + 4)r + N(FFFFF_y(y))^2. \] (5.13)

The bounds on \( F_y(y) \) are shown for data sets of 10, 100, and 1000 points (Figure 5.1).

In this case, the data were generated from a uniform distribution with probability density 1 for \( 0 < y < 1 \). The large variety of cumulative distribution functions which may fit within the bounds immediately highlights the need for constraints on the distribution form chosen to fit the data. For example, the cumulative distribution of the data is a valid choice for \( F_y(y) \), yet this distribution does not represent a continuous random variable.
Figure 5.1.

Data generated from a uniform distribution on the interval [0,1] is plotted along with the corresponding bounds on the actual cumulative distribution based on the observed data. The three examples correspond to data sets of 10, 100, and 1000 points. The bounds are based on twice the standard deviation of the sample cumulative distribution for a given actual cumulative distribution.
and is inaccurate in many circumstances, necessitating a parametric cumulative distribution form.

Since the possible cumulative functions $F_r(y)$ capable of producing a known data set are bounded, the mean squared error between $F_r(y)$ and $\hat{F}_r(y)$, the estimate for $F_r(y)$ based on the sample cumulative distribution function, can sometimes be bounded. First, consider a set of possible estimators $G(p_1,...,p_k,y)$ that include the correct cumulative distribution $F_r(y)$. Next, note that

$$\int_{-\infty}^{\infty} [F_r(y) - \tilde{F}_r(y)]^2 dy = \int_{-\infty}^{\infty} [E[\tilde{F}_r(y)] - \tilde{F}_r(y)]^2 dy$$  \hspace{1cm} (5.14)$$

is bounded by

$$\int_{-\infty}^{\infty} [2\sqrt{\text{var}(\tilde{F}_r(y))}]^2 dy$$  \hspace{1cm} (5.15)$$
in a two standard deviation sense. Since $F_r(y)$ is contained in the set of functions $G(p_1,...,p_k,y)$ and the squared difference between $G(p_1,...,p_k,y)$ and $\tilde{F}_r(y)$ is minimized to determine the estimate $\hat{F}_r(y)$, the following inequality holds:

$$\int_{-\infty}^{\infty} [\hat{F}_r(y) - \tilde{F}_r(y)]^2 dy \leq \int_{-\infty}^{\infty} [F_r(y) - \tilde{F}_r(y)]^2 dy.$$  \hspace{1cm} (5.16)$$

Next,

$$\int_{-\infty}^{\infty} [F_r(y) - \hat{F}_r(y)]^2 dy = \int_{-\infty}^{\infty} [F_r(y) - \tilde{F}_r(y)]^2 dy$$

$$+ 2 \int_{-\infty}^{\infty} [\hat{F}_r(y) - \tilde{F}_r(y)][\tilde{F}_r(y) - F_r(y)] dy + \int_{-\infty}^{\infty} [\hat{F}_r(y) - \tilde{F}_r(y)]^2 dy$$  \hspace{1cm} (5.17)$$

so
\[
\int_{-\infty}^{-}[F_{Y}(y) - \tilde{F}_{Y}(y)]^2 dy \leq 2 \int_{-\infty}^{-}[F_{Y}(y) - \tilde{F}_{Y}(y)]^2 dy + 2 \int_{-\infty}^{-}[	ilde{F}_{Y}(y) - \tilde{F}_{Y}(y)]^2 dy
\]

However, each quantity on the right side of the expression is bounded, and

\[
\int_{-\infty}^{-}[F_{Y}(y) - \tilde{F}_{Y}(y)]^2 dy \leq 16 \int_{-\infty}^{-}\text{var}(\tilde{F}(x)) dx.
\] (5.19)

Further, consider a set of possible estimators \(G(p_{1...p_{k}}, y)\) that does not necessarily contain \(F_{Y}(y)\) but contains at least one function that is contained entirely within the bounds on \(F_{Y}(y)\). In this case, \(\int_{-\infty}^{-}[	ilde{F}_{Y}(y) - \tilde{F}_{Y}(y)]^2 dy\) is directly bounded by

\[
\int_{-\infty}^{-}[2\sqrt{\text{var}(\tilde{F}_{Y}(y))}]^2 dy, \text{ and the bound } \int_{-\infty}^{-}[F_{Y}(y) - \tilde{F}_{Y}(y)]^2 dy \leq 16 \int_{-\infty}^{-}\text{var}(\tilde{F}(x)) dx \text{ is again valid. If the set } G(p_{1...p_{k}}, y) \text{ does not contain any function within the bounds on the cumulative distribution, then no bound exists for } \int_{-\infty}^{-}[	ilde{F}_{Y}(y) - \tilde{F}_{Y}(y)]^2 dy.
\]

5.1.3. Asymptotic Properties of the Error Bounds

The asymptotic properties of the error bounds provides a justification for the use of the cumulative distribution function fitting technique for the estimation of random variable probability functions. We prove the convergence of the estimate to the correct cumulative distribution if \(F_{Y}(y)\) is contained in the set of \(G(p_{1...p_{k}}, y)\). The description of the asymptotic convergence of the estimator, follows closely the work completed by Ozturk and Hettmansperger [20].
First, Chebyshev’s inequality states that

\[ \Pr[|F_y(y) - \tilde{F}_y(y)| > \varepsilon] < \frac{\text{var}(\tilde{F}_y(y))}{N\varepsilon^2}. \tag{5.20} \]

Substituting the expression for the variance of the data cumulative distribution function,

\[ \Pr[|F_y(y) - \tilde{F}_y(y)| > \varepsilon] < \frac{(F_y(y))(1 - F_y(y))}{N\varepsilon^2}. \tag{5.21} \]

Since

\[ 0 \leq F_y(y) \leq 1 \tag{5.22} \]

this probability approaches 0 as N approaches infinity for any \( \varepsilon \), and the data cumulative distribution function converges to \( F_y(y) \). Further, the result guarantees that the range of functions \( F_y(y) \) that are likely to generate a specific data cumulative distribution function converges to that data cumulative distribution function as the number of points approaches infinity.

A more interesting query is the mean squared error between \( F_y(y) \) and its estimate based on the sample cumulative distribution. If the set of \( G(p_1...p_k, y) \) contains \( F_y(y) \), the previously discussed bound on \( \int [F_y(y) - \tilde{F}_y(y)]^2 dy \) approaches zero as the number of points approaches infinity. However, this bound only proves that a certain fraction of the errors \( \int [F_y(y) - \tilde{F}_y(y)]^2 dy \) are bounded arbitrarily close to zero.

However, the stronger claim that \( \int [F_y(y) - \tilde{F}_y(y)]^2 dy \) approaches zero with probability one is also valid. To prove this claim, first, note that
\[ \text{Pr}[(F_y(y) - \hat{F}_y(y))^2 > \varepsilon^2(F_y(y))(1 - F_y(y))] \]
\[ \leq \frac{\text{var}(\hat{F}_y(y))}{\varepsilon^2(F_y(y))(1 - F_y(y))} = \frac{1}{N\varepsilon^2} \quad (5.23) \]

Next, assume that \( \int (F_y(y) - \hat{F}_y(y)) dy \) is finite in order to simplify the limit calculation; this condition is valid for all finite variance random variables. Given the additional constraint, \( \varepsilon^2 \int (F_y(y) - \hat{F}_y(y)) dy \) is also finite, and

\[ \text{Pr}\left[ \int (F_y(y) - \hat{F}_y(y))^2 dy > \varepsilon^2 \int (F_y(y))(1 - F_y(y))dy \right] \leq \frac{1}{N\varepsilon^2}, \quad (5.24) \]

since \( (F_y(y) - \hat{F}_y(y))^2 \) is then greater than or equal to \( \varepsilon^2(F_y(y))(1 - F_y(y)) \) for at least one \( y \). Next, from previous analysis,

\[ \int (F_y(y) - \hat{F}_y(y))^2 dy \leq \int (F_y(y) - \hat{F}_y(y))^2 dy \quad (5.25) \]

if the set of \( G(p_1, \ldots, p_k, y) \) contains \( F_y(y) \). Consequently,

\[ \text{Pr}\left[ \int (F_y(y) - \hat{F}_y(y))^2 dy > 4\varepsilon^2 \int (F_y(y))(1 - F_y(y))dy \right] \leq \frac{1}{N\varepsilon^2}. \quad (5.26) \]

As \( N \to \infty \),

\[ \text{Pr}\left[ \int (F_y(y) - \hat{F}_y(y))^2 dy > \tau \right] \leq \frac{4\int (F_y(y))(1 - F_y(y))dy}{N\tau} \to 0 \quad (5.27) \]

for arbitrarily small \( \tau \), and the estimate approaches \( F_y(y) \) in a mean squared sense.

If \( F_y(y) \) is not contained in the set of \( G(p_1, \ldots, p_k, y) \), \( \hat{F}_y(y) \) does not approach \( F_y(y) \). However, \( \tilde{F}_y(y) \) still approaches \( F_y(y) \) in a mean squared sense since
Pr\[\int_{-\infty}^{\infty} (F(y) - \hat{F}_y(y))^2 dy > \epsilon^2 \int_{-\infty}^{\infty} (F(y))(1 - F(y))] \leq \frac{1}{Ne^2}\] is still valid. Consequently, in the limit, the estimate $\hat{F}_y(y)$ is the member of the set of functions $G(p_1, ..., p_k, y)$ that most closely approximates $F_y(y)$ in the mean squared sense.

5.1.4. General Properties of the Cumulative Distribution Function Matching Estimator

Although asymptotic results for the behavior of the cumulative distribution function are promising, an analysis of the bias and variance of the estimator are necessary. First, the bias of the estimator must be defined. For example, should the bias of the estimator represent the average error in estimating $F_y(y)$ as a function of $y$ or the error in estimating the mean of the random variable $Y$. We consider two definitions for the bias of the estimate for $F_y(y)$. In turn, the second of these bias measures can be connected with the bias of the estimate for the random variable $Y$. The first measure for the bias is

$$b(y) = E[\hat{F}_y(y) - F_y(y)] = E[\hat{F}_y(y) - \tilde{F}_y(y)].$$  \hspace{1cm} (5.28)

This measure provides a measure of the accuracy of the cumulative distribution function estimate at each value of $y$. Next,

$$b_2 = \int_{-\infty}^{\infty} E[\hat{F}_y(y) - F_y(y)]dy = \int_{-\infty}^{\infty} E[\hat{F}_y(y) - \tilde{F}_y(y)]dy$$  \hspace{1cm} (5.29)

constrains the difference in the area under the original and estimated cumulative distribution functions. Throughout the described bias calculations, the set of estimators $G(p_1, ..., p_k, y)$ is assumed to include the correct estimate $F_y(y)$.
The bias \( b(y) \) can be bounded at each \( y \). Conceptually, a bound exists because a certain \( \hat{F}_y(y) - \tilde{F}_y(y) \) implies a minimum value for \( \int [\hat{F}_y(y) - \tilde{F}_y(y)]^2 dy \), which is bounded. A weak bound generated by this limitation is

\[
1 b(y) \leq \frac{4 \sqrt{\int F_y(y)(1 - F_y(y))dy}}{\sqrt{N} |y_i - y|},
\]

(5.30)

where \( y_i \) is the location of the step in \( \hat{F}_y(y) \) that is nearest to \( y \).

---

**Figure 5.2.**

This figure is a diagram of the example that shows the possibility for arbitrarily large bias using the cumulative distribution function-based estimator.

Unfortunately, \( b_2 \) can be arbitrarily large for certain sets of possible estimators \( G(p_1, \ldots, p_k, y) \). For example, consider the estimation problem in which \( \hat{F}_y(y) \) equals either \( F_y(y) \) or \( c(y) \) (Figure 5.2). Next, consider a sample cumulative distribution

127
\( \tilde{F}_y(y) \) resulting from the random generation of one point. Regardless of the value for \( q \) (see Figure), the probability that \( c(y) \) is chosen as the estimate \( \tilde{F}_y(y) \) is greater than zero, since \( q \) does not affect the integrated mean squared error between \( c(y) \) and \( \tilde{F}_y(y) \) if the step in \( \tilde{F}_y(y) \) occurs at \( y>p \), and clearly some such \( \tilde{F}_y(y) \) are closer to \( c(y) \) than \( F_y(y) \) in a mean squared sense. Meanwhile, \( b_2 \) approaches infinity as \( q \) approaches infinity.

For any finite \( N>1 \), the probability that \( c(y) \) is chosen as \( \tilde{F}_y(y) \) decreases, but the expected bias approaches infinity for sufficiently large \( q \).

If the set of estimators consists of all beta distributions, we believe that the bias in the estimator will be much smaller. We have not yet been able to calculate or bound this bias. However, Monte Carlo trials demonstrate that the estimator bias is small. We show an example of beta distribution estimation to highlight the accuracy of the estimator graphically (Figure 5.3). In both examples, we generate data from a beta distribution and then plot the sample cumulative distribution and the estimate. As the number of points increases, the estimate becomes almost indistinguishable from the sample cumulative distribution. For the example with fewer points, we also plot the source and estimated probability density function in order to gauge the accuracy of the probability density estimate.

An interesting alternative to the original estimator guarantees unbiased estimation in the integrated sense. Specifically, if we restrict the set of possible estimators \( G(p_1, \ldots, p_k, y) \) to only those estimators for which

\[
\int_{-\infty}^{\infty} [G(p_1, \ldots, p_k, y) - \tilde{F}_y(y)] dy = 0 \quad (5.31)
\]
then the bias in the estimator is also zero.

Figure 5.3.
Figure 5.3a, 5.3b, and 5.3c demonstrate beta cumulative distribution fitting to a 20 point sample cumulative distribution.

5.1.5. A Computationally Simpler Estimator

A computationally simpler metric for the estimation of the cumulative distribution function for the estimation of the cumulative distribution function of a random variable $Y$ is considered. The new metric for cumulative distribution estimation is also useful because the range of values of the cumulative likelihood function can be calculated for this metric, and because the new estimator leads to a method for extrapolating the results of the estimation in multi-dimension problems.
Specifically, instead of minimizing $\int [G(p_1, ..., p_k, y) - \tilde{F}_y(y)]^2 \, dy$, we find

$$\tilde{p}_1, ..., \tilde{p}_k = \arg\min_{p_1, ..., p_k} \sum_{i=1}^{N} [G(p_1, ..., p_k, y_i) - \tilde{F}_y(y_i)]^2$$

(5.32)

Generally, the various bounds developed for the original metric are also valid for the new metric at the points $y_1, ..., y_N$. Some of these bounds are important for the extrapolation question and so are considered for the multidimensional problem. However, the bound for $\sum_{i=1}^{N} [\hat{F}_y(y_i) - \tilde{F}_y(y_i)]^2$ will be presented to facilitate the analysis of the cumulative maximum likelihood function. Utilizing an analogous analysis to the integral estimator, the bound is given by

$$\Pr[\sum_{i=1}^{N} (F_y(y_i)) - \hat{F}_y(y_i))^2 > \tau] \leq \frac{4 \sum_{i=1}^{N} (F_y(y_i))(1 - F_y(y_i))}{N \tau}$$

(5.33)

In one respect, the new estimator is significantly different from the original metric. The metric automatically weights the data cumulative distribution so that regions of the cumulative distribution with a large number of points are more significant in the calculation than regions with few points. If the number of points reflects the amount of information contained in a region of the sample cumulative distribution, then the new metric will result in a more accurate estimate than the original metric. Conversely, we might interpret the number of points in a region as a measure of the slope of the cumulative distribution function and not a measure of the information in the region, so that the original metric will be more accurate.
5.1.6. Bounds on Beta Distribution Parameters

If the set of possible estimate functions \( G(p_1, \ldots, p_k, y) \) is parametrized, approximate bounds on the parameters can also be developed. The estimation of the four parameter generalized beta distribution is important to this study, so bounds on beta distribution parameters are formulated.

Bounds for individual parameters highlight the sensitivity of the cumulative distribution estimate to a change in a single parameter. The difference between \( F_y(y; \alpha, \beta, a, b) \) and \( \hat{F}_y(y; \alpha, \beta, a, b) \) due to an incremental variation in the alpha parameter, labeled \( \Delta \alpha \), is

\[
F_y(y; \alpha, \beta, a, b) - \hat{F}_y(y; \alpha, \beta, a, b) = \frac{\partial}{\partial \alpha} (F_y(y; \alpha, \beta, a, b)) (\Delta \alpha). \tag{5.34}
\]

Consequently,

\[
\int (F_y(y; \alpha, \beta, a, b) - \hat{F}_y(y; \alpha, \beta, a, b))^2 \, dy
= (\Delta \alpha)^2 \int \frac{\partial}{\partial \alpha} F_y(y; \alpha, \beta, a, b)^2 \, dy \tag{5.35}
\]

Since

\[
\Pr\left[ \int (F_y(y; \alpha, \beta, a, b) - \hat{F}_y(y; \alpha, \beta, a, b))^2 \, dy > \tau \right]
\leq \frac{4 \int (F_y(y; \alpha, \beta, a, b))(1 - F_y(y; \alpha, \beta, a, b))dy}{N \tau}. \tag{5.36}
\]

Therefore,
for small enough \( \Delta \alpha \). Upon some algebraic manipulation, a bound for \( \Delta \alpha \) is determined as

\[
\Pr[| \Delta \alpha | > q] \leq \frac{4 \int (F_y(y; \alpha, \beta, a, b))(1 - F_y(y; \alpha, \beta, a, b))dy}{Nq^2 \int (\frac{d}{d\alpha} F_y(y; \alpha, \beta, a, b))^2 dy}
\]

Again, the bound cannot be calculated without knowledge of \( F_y(y; \alpha, \beta, a, b) \). However,

\[
\Pr[| \Delta \alpha | > q] \leq \frac{4 \int (\hat{F}_y(y; \alpha, \beta, a, b))(1 - \hat{F}_y(y; \alpha, \beta, a, b))dy}{Nq^2 \int (\frac{d}{d\alpha} \hat{F}_y(y; \alpha, \beta, a, b))^2 dy}
\]

is an adequate estimate for the bound for most \( F_y(y; \alpha, \beta, a, b) \) and \( \hat{F}_y(y; \alpha, \beta, a, b) \).

Similar bounds for the other shape parameter and two endpoint parameters can be calculated. These single parameter bounds are only accurate if the other distribution parameters are known, however, and therefore are not useful in defining limits on actual parameter estimates. The bounds are useful because they provide a measure for the possible variations in the shape and extent of the estimated distributions and because they allow a graphical depiction for the possible range of estimates.

Next, a bound can be developed for the probability that all four parameters deviate from their actual value by \( \Delta \alpha \), \( \Delta \beta \), \( \Delta a \), and \( \Delta b \). Specifically,
\[
\Pr \left[ \int \left\{ \frac{\partial}{\partial \alpha} F_y(y; \alpha, \beta, a, b)(\Delta \alpha) + \frac{\partial}{\partial \beta} F_y(y; \alpha, \beta, a, b)(\Delta \beta) \\
- \frac{\partial}{\partial a} F_y(y; \alpha, \beta, a, b)(\Delta a) + \frac{\partial}{\partial b} F_y(y; \alpha, \beta, a, b)(\Delta b) \right\} dy > \tau \right]
\]

Unfortunately, the bound involves a quadratic expression in \( \Delta \alpha \), \( \Delta \beta \), \( \Delta a \), and \( \Delta b \) and cannot be simplified further. Nevertheless, the bound can approximately be calculated for a beta distribution estimate and so provides an analysis of the range of possible parameter values.

5.1.7. Maximum Likelihood Cumulative Distribution Estimation

An estimate for a cumulative distribution function is formulated by determining the function which maximizes the probability of a set of samples of the data cumulative distribution. Specifically, consider \( M \) samples of the data cumulative distribution, \( \tilde{F}_y(y_1), \ldots, \tilde{F}_y(y_M) \), and consider the functions \( G(p_1 \ldots p_k, y) \) which comprise the set of possible estimators. Then the cumulative maximum likelihood estimate is defined as

\[
\tilde{p}_1 \ldots \tilde{p}_k = \arg \max_{p_1 \ldots p_k} \Pr[\tilde{F}_y(y_1), \ldots, \tilde{F}_y(y_M) | F_y(y) = G(p_1 \ldots p_k, y)]
\] (5.41)

The values of the sample cumulative distribution for a given cumulative distribution are given by a multinomial distribution, so that the joint probability of the samples in the maximization equation is
\[
\text{Pr} \left[ \overline{F}_y(y_1) = \frac{k_1}{N} \ldots \overline{F}_y(y_M) = \frac{k_M}{N} \mid F_y(y) = G(p_1 \ldots p_k, y) \right] \\
= \frac{N!}{k_1!(k_2 - k_1)!(N - k_M)!} (F_y(y_1))^{k_1} \\
(F_y(y_2) - F_y(y_1))^{k_2 - k_1} \ldots \left(1 - F_y(y_k) \right)^{N_k - k_M} 
\]
(5.42)
5.2. **Beta Distribution Modeling of Data Through Multivariate Cumulative Distribution Estimation.**

5.2.1. **Description of the Technique**

In order to describe a set of data that may vary as a function of an independent parameter \( x \), the joint probability density of a continuum of random variables \( Y[x] \) must be determined. We have previously applied maximum likelihood and linear regression methods to determine the parameters in several different beta distribution-based models. We attempt to determine the parameters of another model describing the data using the cumulative distribution function estimation method. If the random variables \( Y[x] \) are modeled to depend only on the independent parameter \( x \), the problem of describing data that varies with an independent parameter may be rephrased. Specifically, the location of the \( x \) coordinates of the data may be described by a random variable \( X \). \( Y[x] \) is then the random variable for the \( y \) coordinate given that \( X=x \). If this general model is sufficient to describe the data, then the determination of the joint cumulative distribution function of \( X \) and \( Y \) is sufficient for the analysis of the random variables \( Y[x] \).

Based on this general problem formulation, the data, which consists of \( N \) coordinate pairs \((x_1, y_1), \ldots, (x_N, y_N)\), is assumed to be instances of \( N \) independent, identically distributed random variable pairs labeled \((X_1, Y_1), \ldots, (X_N, Y_N)\). The objective of the fitting procedure is the accurate estimation of the joint cumulative distribution function \( F_{X,Y}(x, y) \) that is common to the random variable pairs. The sample cumulative distribution, which is used for estimation, is defined as

\[
\bar{F}_{X,Y}(x, y) = \frac{1}{N} \sum_{i=1}^{N} u(x - x_i)u(y - y_i). \tag{5.43}
\]
Two different formulations for the estimation of $F_{X,Y}(x, y)$ from $\hat{F}_{X,Y}(x, y)$ are useful. Most generally, assume that the estimate for $F_{X,Y}(x, y)$, labeled $\hat{F}_{X,Y}(x, y)$, is contained among a parametrized set of functions $G(p_1 \ldots p_k, x, y)$. Then the parameters of the estimate $\hat{F}_{X,Y}(x, y)$ are chosen as

$$\tilde{p}_1 \ldots \tilde{p}_k = \arg \min_{p_1 \ldots p_k} \int [G(p_1 \ldots p_k, x, y) - \hat{F}_Y(x, y)]^2 dy \, dx. \quad (5.44)$$

Alternately, since the goal of the analysis is the determination of the cumulative distribution function of $Y[x]$, rather than joint cumulative distribution of $X$ and $Y$, we can consider functions $G(p_1 \ldots p_k, x, y)$ with the following form:

$$G(p_1^\ast \ldots p_k^\ast, p_1^\prime \ldots p_k^\prime, x, y) = \int [g_X(p_1^\ast \ldots p_k^\ast, a)G_{Y|X}(p_1^\prime \ldots p_k^\prime, y|a)] da. \quad (5.45)$$

An alternate method for estimation is suggested by this form for $G(p_1 \ldots p_k, x, y)$. The parameters $(p_1^\ast \ldots p_k^\ast)$ can be estimated first as a single variable cumulative distribution function problem; the remaining parameters can then be estimated separately based on the joint distribution.

In the following section, we describe the specific models used to describe data with a continuum of beta distributions. The asymptotic properties of the multi-dimensional cumulative distribution fitting process are identical to the asymptotic properties of the single cumulative distribution estimation process, so these properties are not rederived. An important question is the accuracy of the estimated conditional distribution at each independent parameter value. Unfortunately, the development of
bounds for these individual distributions proved difficult, and the topic is not explored further in this thesis.

5.2.2. Models for Estimating Beta Distribution Continua

The models that are used to describe data using a continuum of beta distributions are similar to models described previously in the context of both maximum likelihood and linear regression estimation. However, the parametric models that are analyzed using cumulative distribution function fitting must also provide a structure for the location of the independent parameter values of the points. A natural model for the data is one that describes the independent parameter with one parametrized distribution and then describes the dependent parameter conditioned on the independent parameter as a beta random variable. In other words, we consider a set of estimation functions of the form given in (5.45) and separately determine appropriate forms for \( g_{X}(p_{1}^{*}\ldots p_{j}^{*}, a) \) and \( G_{Y|X}(p_{1}^{*}\ldots p_{j}^{*}, y | X = a) \). An estimate for the cumulative distribution function \( F_{X}(a) \) can be determined using a one variable estimation; finally, a two variable cumulative distribution fitting is utilized to determine the best estimate for \( F_{Y|X}(y | X = a) \).

The proper form for the probability density function \( g_{X}(p_{1}^{*}\ldots p_{j}^{*}, a) \) and hence its corresponding cumulative distribution \( G_{X}(p_{1}^{*}\ldots p_{j}^{*}, a) \), is not constrained in general and instead depends on the specific problem and data set. In many data sets, a uniform distribution will adequately represent the independent parameter values of the known data points. Alternately, Gaussian and exponential distributions or a variety of other
distributions may accurately describe $X$. In fact, a beta distribution for $X$ is often appropriate because of the variety of different distributions that can be approximated with a beta distribution. In other problems, the values of the independent parameter are necessarily discrete, and so a discrete random variable is an accurate representation for $X$. A discrete uniform random variable and a Poisson random variable are two obvious possibilities for describing the independent parameter. Alternately, if the independent parameter values are necessarily discrete and also limited to finite number of points, the sample cumulative distribution of $X$ may be an accurate representation for actual cumulative distribution.

Although one generic form does not adequately represent all possible probabilistic representations of the independent parameter, we can easily determine whether or not a specific form can describe the independent parameter. For any form, we first calculate the estimate for the cumulative distribution function, labeled $\tilde{F}_X(x)$. Next, if the correct distribution $F_X(x)$ is contained in the set of estimators, then $\int_{-\infty}^{\infty} [\tilde{F}_X(x) - \tilde{F}_X(x)]^2 dx$ is bounded probabilistically. If the actual value for $\int_{-\infty}^{\infty} [\tilde{F}_X(x) - \tilde{F}_X(x)]^2 dx$ occurs with small probability, then $F_X(a)$ is likely not contained in the set of estimators, and a different parametrization of the estimator should be considered. Alternately, if the actual value for $\int_{-\infty}^{\infty} [\tilde{F}_X(x) - \tilde{F}_X(x)]^2 dx$ is smaller than the bound and satisfies any other constraints on the distribution form, such as continuity of the probability density function,
then the chosen estimator is reasonable. Formally, once the proper model has been chosen, the estimates for \( p_i^* \ldots p_j^* \), labeled \( p_i^{**} \ldots p_j^{**} \), are

\[
\begin{align*}
\hat{p}_i^{**} \ldots \hat{p}_j^{**} &= \arg \min_{p_i^{*} \ldots p_j^{*}} \int \left[ G_X(x; p_i^{*} \ldots p_j^{*}) - \tilde{F}_X(x) \right]^2 dx,
\end{align*}
\]

(5.46)

where

\[
G_X(x; p_i^{*} \ldots p_j^{*}) = \int g_X(a; p_i^{*} \ldots p_j^{*}) da.
\]

(5.47)

The determined parameters then allow the exact calculation of the cumulative distribution estimate, \( \hat{F}_X(x) \), and the corresponding probability density estimate, \( \hat{f}_X(x) \).

Once the parametric representation for \( X \) and the corresponding optimal estimate have been determined, the parameters of the model \( G_{\{i\ldots j\}}(p_i^{*}, \ldots p_j^{*}, y | X = a) \) can be determined by fitting the two-dimensional cumulative distribution function. First, we must determine appropriate function forms \( G_{\{i\ldots j\}}(p_i^{*}, \ldots p_j^{*}, y | X = a) \). We can again use the models described for maximum likelihood and regression estimation. Specifically, \( G_{\{i\ldots j\}}(p_i^{*}, \ldots p_j^{*}, y | X = a) \) is beta cumulative distribution function for each \( a \), and either the moments or the standard beta distribution parameters are described as linear combinations of basis functions that change with \( a \). The coefficients of each basis function are the parameters in the model.

Once a parametric form \( G_{\{i\ldots j\}}(p_i^{*}, \ldots p_j^{*}, y | X = a) \) has been chosen such that the cumulative distribution function at each \( a \) is a beta distribution, the parameters of the model can be determined. For each possible set of parameters \( p_i^{*}, \ldots p_j^{*} \), we first determine the representation for the joint cumulative distribution function of \( X \) and \( Y \):
\[
G(p^*_1, \ldots, p^*_t, x, y) = \int [f_x(a)G_{y|x}(p^*_1, \ldots, p^*_t, y | x, a)] da 
\]

(5.48)

The parameters \( p^*_1, \ldots, p^*_t \) that minimize the mean squared error between the two variable cumulative distribution and the corresponding sample distribution are then chosen as the parameter estimates:

\[
p^*_1, \ldots, p^*_t = \arg \min_{p_1, \ldots, p_t} \int \int [G(p^*_1, \ldots, p^*_t, x, y) - \tilde{F}_{x,y}(x, y)]^2 dy dx.
\]

(5.49)

We have now determined the estimate for the joint cumulative distribution of \( X \) and \( Y \), labeled \( F_{x,y}(x, y) \), as well as the estimate for the conditional probability of \( Y \) given \( X \), labeled \( F_{y|x}(y | x) \).

Estimation of a joint cumulative distribution function in order to describe a set of data with a continuum of beta distributions is demonstrated in Figure 5.4. In this case, we assume that the independent parameter is distributed uniformly between zero and one. Except for the scaling of the independent parameter axis, the modeled data set is identical to the 100 point data set estimated using moment regression techniques.

![Figure 5.4](image)

**Figure 5.4.**

This figure illustrates estimation using multivariate cumulative distribution function fitting.
The model for estimating a continuum of beta distributions is also useful because it provides a natural measure for the possibility for extrapolation. Specifically, the distribution fitted to the independent parameter values can be interpreted as a measure of the information present at each independent parameter value. As the probability density for a sample point at an independent parameter value decreases beyond the end of the data set, the accuracy of any extrapolation of the estimated distribution continuum should also decrease. Although a quantitative formulation of this measure of extrapolation is not attempted here, the use of the independent parameter model to gauge extrapolation is interesting and should be studied further.
Chapter 6

Interpolation of Beta Distributions

Often, the behavior of a random process at one or more values of the independent parameter in the process is well known, but the remainder of the random process cannot easily be described. Specifically, consider a continuum of independent random variables $Y[x]$ that vary in probability distribution with the parameter $x$. The probability density function for $Y[x]$ may exactly be known for the independent parameter values $x_1...x_N$. Based on the probability density functions, $f_Y(y)[x_1]...f_Y(y)[x_N]$, an estimate for the continuum of distributions $f_Y(y)[x]$ may be formulated.

We consider several techniques for extrapolating the continuum of distributions and highlight the characteristics of these techniques. Both band-limited and linear interpolation of the cumulative distribution functions $F_Y(y)[x_1]...F_Y(y)[x_N]$ are considered. Also, interpolation of the moments of the distribution is attempted. These techniques for interpolating the moments are applied to the interpolation of beta distributions, and the four parameters of the beta distribution are also directly interpolated.

6.1. Interpolation of the Cumulative Distribution Function

The cumulative distribution function of a continuum of distributions, labeled $F_Y(y)[x]$, is a surface that is defined for all $(x,y)$. The methods for determining $F_Y(y)[x]$
from a set of sample distributions are based on interpolation of slices of \( F_y(y)[x] \) that are parallel to the x-y plane. For each slice,

\[
F_y(y)[x]=q,
\]

(6.1)

the value of the probability variable \( y \), labeled \( y_q[x] \), is known for \( x \in x_1 \ldots x_N \). These \( N \) coordinate pairs, \((x_1, y_q[x_1]) \ldots (x_N, y_q[x_N])\), are used to determine \( y_q[x] \). The procedure can then be repeated for all \( q \in [0,1] \). If an appropriate reconstruction method is chosen, the resulting function \( F_y(y)[x] \) will be a valid cumulative distribution function for each \( x \). Two techniques for interpolation, band-limited interpolation and linear interpolation, are discussed. Practical methods for implementing the interpolation methods and determining the cumulative distribution arguments for all \( q \in [0,1] \), rather than a discrete set of \( q \), are considered.

6.1.1. Band-limited Interpolation of Cumulative Distribution Functions

First, the concept of a band-limited continuum of cumulative distribution function is necessary. We define a continuum of cumulative distributions as band-limited if and only if \( y_q[x] \) is band-limited for all \( q \in (0,1) \). A band-limit on a set of distributions limits the variability of the argument of the distributions with the independent parameter rather than the variability of the distribution at one value of the independent parameter. Mathematically, the defined band-limit of a set of distributions is \( \omega_0 \) if and only if \( \omega_0 \) is the smallest value such that

\[
Y_q(\omega) = \int y_q[x]e^{-j\omega x}dx = 0, \omega > \omega_0
\]

(6.2)
for all \( q \in (0,1) \). The band-limit of the set of distributions is equal to the largest band-limit of the set of functions \( y_q[x] \).

Given a band-limited set of distributions, the continuum of cumulative distribution functions can be exactly derived from a sufficient set of sample cumulative distributions. In order to achieve exact reconstruction, the sampling period must be less than the Nyquist period,

\[
T_N = \frac{\pi}{\omega_0},
\]

and an infinite number of samples is necessary. These samples do not need to be evenly spaced; in case of evenly spaced samples, however, an expression for the reconstructed continuum of distributions is easily formulated. Specifically, assume that \( F_y(y)[x] \) is exactly known for \( x \in \{0, \pm T_N, \pm 2T_N, \ldots \} \). If the cumulative distribution is known for \( x \in \{0, \pm T_N, \pm 2T_N, \ldots \} \), then \( y_q[x] \) is also known for the same discrete \( x \) values. Next, a reconstruction for \( y_q[x] \) at each \( q \) is

\[
y_q[x] = \sum_{n=-\infty}^{\infty} y_q(x - n T_N) \text{sinc}(\pi(n - \frac{x}{T_N})),
\]

where

\[
\text{sinc}(z) = \frac{\sin(z)}{z}.
\]

Consequently, \( F_y(y)[x] \) can be determined as the \( q \) such that \( y = y_q[x] \).

Typically, a perfect band-limited reconstruction is impossible because the cumulative distributions are not band-limited or because a sufficient set of distributions is not known. In fact, the reconstruction requires an infinite number of cumulative
distribution samples, while only a small set of cumulative distributions will be known in most problems. Several practical considerations become important in using this reconstruction technique if the distributions are under-sampled or if the samples do not extend to $x = \pm \infty$.

First, the set of functions $F_{y}(y)[x]$ resulting from the band-limited reconstruction may no longer be a set of cumulative distributions. Specifically, $y_{q}[x]$ at a specific independent parameter value may no longer be a monotonic function of $q$, so that $F_{y}(y)[x]$ is no longer monotonically increasing in $y$ and is in fact not a function. The plots of two sample cumulative distributions and of two functions $y_{q}[x]$ that result in an improper $F_{y}(y)[x]$ are shown (Figure 6.1). In general, a band-limited interpolation of a set of cumulative distribution functions does not necessarily produce another cumulative distribution function.

![Cumulative Distribution](image)

**Figure 6.1.**

This figure demonstrates the possible failure of band-limited cumulative distribution function fitting given an insufficient number of samples. A smaller $y$ value corresponds to a larger cumulative distribution value so that the resulting function at intermediate independent parameter values is not a valid cumulative distribution.
If the rate of sampling is slower than the Nyquist rate for the set of distributions, then at least one of the functions $y_q[x]$ will be improperly reconstructed. The amount of error in the reconstructed $y_q[x]$ depends upon the component of the function that is lost because of the sinc function reconstruction. For sampling below the Nyquist rate, the process of sampling followed by reconstruction is equivalent to a low pass filtering of the signal $y_q[x]$. An important question is the effect of this low pass filtering on the resulting reconstruction of $F_y(y)[x]$. When will this resulting function be a proper set of cumulative distributions? If the result is a cumulative distribution, how much will the filtered version differ from the original set of distributions?

If distribution samples are known only for a finite range of the independent parameter, the reconstruction method must be modified slightly to generate reasonable reconstructions of the cumulative distribution functions. An interpolation of a function based solely on the known samples, which implicitly assumes that all other samples of the function are zero, will differ significantly from the original signal near the endpoints of the known samples. Further, the sinc reconstruction will lead to ringing in the interpolated result near the endpoints of the known samples; this ringing is usually an unnatural description of the function. Instead, a reasonable reconstruction guesses the characteristics of the signal and so the sample values outside the range of known sample values. For this application, we reverse the known samples of $y_q[x]$ with respect to $x$ and append this reversed signal to the original signal. We repeat the signal several times and then complete a band-limited interpolation of this signal to determine the functions $y_q[x]$. 
Additionally, the process of band-limited reconstruction must be adjusted because $y_q[x]$ cannot be determined for all $q$ and instead must be determined for a discrete set $q_1...q_M$. For a given independent parameter value, the cumulative distribution of the dependent variable is known only at $M$ discrete points. Although an arbitrarily large number of samples $M$ of the cumulative distribution may be chosen, the reconstruction becomes computationally more intensive as $M$ increases, and the cumulative distribution function is not completely known regardless of the number of samples of each distribution. To determine the entire cumulative distribution at each independent parameter value, a band-limited approximation to the cumulative distribution may be completed, or the sampled cumulative distribution may be approximated with a parametrized distribution. Specifically, assume that $y_q[x]$ has been determined at $M$ points for a certain independent parameter $x_0$, so that $M$ samples of $F_y(y)[x_0]$, labeled $F_y(y_1)[x_0]...F_y(y_M)[x_0]$, are also available. Further, the cumulative distribution will equal zero for sufficiently small $y$ and will equal one for sufficiently large $y$, so that an infinite number of samples of $F_y(y)[x_0]$ are known. In many situations, the cumulative distribution at each $x_0$ will be band-limited or approximately band-limited, and a band-limited reconstruction will be accurate for sufficiently large $M$. The samples of the distribution are not evenly spaced, however, and so the expression for the bandlimited reconstruction is more complicated. Alternately, a parametric form is chosen for the distribution for $F_y(y)[x_0]$, and the parametrized distribution is fitted to the samples of the cumulative distribution function using a least squares metric. Because a four-parameter beta distribution is a good approximation for a wide variety of probability
densities, this distribution is used to approximate the cumulative distribution samples.

Specifically, a beta distribution representation for $F_y(y|x_0)$ is labeled

$$\beta_y(y; \alpha, \beta, a, b)[x_0],$$

and the approximate cumulative distribution function is

$$\arg \min_{\alpha, \beta, a, b} \sum_{i=1}^{M} (F_y(y_i|x_0) - \beta_y(y_i; \alpha, \beta, a, b)[x_0])^2. \quad (6.6)$$

An example of band-limited cumulative distribution function interpolation is shown (Figure 6.2). The two distributions with smallest and largest mean in the diagram represent the two known endpoint distributions, and the intermediate distributions are the results of the interpolation.

![BL CDF Interpolation](image)

**Figure 6.2.**
Band-limited cumulative distribution function interpolation.

---

### 6.1.2. Linear Interpolation of Cumulative Distribution Functions

Linear interpolation of the arguments of a set of cumulative distribution functions is a compelling alternative to band-limited interpolation (Figure 6.3). For a band-limited set of cumulative distribution functions, linear interpolation is a reasonable
approximation to the band-limited result. Further, linear interpolation will always produce valid cumulative distributions regardless of the spectral characteristics of the distributions and does not require an infinite number of sample distributions. Linear interpolation is also computationally less intensive than band-limited interpolation, and the moments of the distributions resulting from linear interpolation can be bounded.

---

**Figure 6.3.**

This figure depicts linear cumulative distribution function interpolation. Given the cumulative distribution functions at two independent parameter values, intermediate cumulative distribution functions can be generated by linearly interpolating the argument of the cumulative distribution function.

---

Linear interpolation is shown for a pair of cumulative distribution functions (Figure 6.4). The graphical interpretation for linear interpolation suggests that cumulative distribution function interpolation is a natural method for estimating a continuum of distributions from several sample distributions. Specifically, the argument of the interpolated distribution is an average of the argument of the known distribution, a reasonable method for interpolation.
Figure 6.4.
Linear Interpolation of Cumulative Distribution Functions.

The mathematical description for linear interpolation is straightforward. If

$F_y(y)[x]$ is known for $x \in x_1 \ldots x_N$, $x_1 < x_2 \ldots < x_N$, then $y_q[x]$ is also known for $x \in x_1 \ldots x_N$. The linear interpolation of each $y_q[x]$ is possible only for $x_1 < x < x_N$. For a specific $x$ in this range, two adjacent known independent parameter values, $x_a$ and $x_{a+1}$, are chosen so that $x_a < x < x_{a+1}$. Next,

$$y_q[x] = y_q(x_a) + \frac{x-x_a}{x_{a+1}-x_a}(y_q(x_{a+1}) - y_q(x_a)). \quad (6.7)$$

Once $y_q[x]$, $x_1 < x < x_N$, has been determined for all $q \in [0,1]$, $F_y(y)[x]$ can also be determined for $x_1 < x < x_N$.

The result of linear interpolation will always be a valid cumulative distribution function since $y_q[x]$ is necessarily a monotonically non-decreasing function of $q$.

However, the cumulative distribution function at each $x$ can again only be determined for
a discrete set of values for \( q \). Again, band-limited interpolation of the single cumulative distribution or least squares fitting of a parametrized distribution is required.

One benefit of the linear interpolation of cumulative distributions is the possibility for quantifying the moments of the interpolated distributions. The mean of the random variables with distributions that are produced by linear interpolation of the set of sample distributions is equal to the linear interpolation of the mean of these sample distributions. To prove this result, the mean of a random variable is written in terms of its cumulative distribution function. The mean of a continuous random variable \( Y \) equals

\[
E[Y] = \int_{-\infty}^{\infty} y F_Y(y) \, dy.
\] (6.8)

Consequently, the mean of the two random variables \( Y[x_a] \) and \( Y[x_{a+1}] \) are

\[
E[Y[x_a]] = \int_{-\infty}^{\infty} y q[x_a] \, dq
\] (6.9)

and

\[
E[Y[x_{a+1}]] = \int_{-\infty}^{\infty} y q[x_{a+1}] \, dq,
\] (6.10)

respectively. However, these means can be more conveniently written as

\[
E[Y[x_a]] = \int_{0}^{1} y q[x_a] \, dq
\] (6.11)

and

\[
E[Y[x_{a+1}]] = \int_{0}^{1} y q[x_{a+1}] \, dq.
\] (6.12)

Next, the mean of the random variable \( Y[x] \), \( x_a < x < x_{a+1} \), is
\[ E[Y[x]] = \int_{0}^{1} y_q[x] dq = \int_{0}^{1} [y_q(x_a) + \frac{x - x_a}{x_{a+1} - x_a} (y_q(x_{a+1}) - y_q(x_a))] dq. \quad (6.13) \]

Rewriting the expression,

\[ E[Y[x]] = E[Y[x_a]] + \frac{x - x_a}{x_{a+1} - x_a} (E[Y[x_{a+1}]] - E[Y[x_a]]). \quad (6.14) \]

Consequently, the linear interpolation of the cumulative distribution arguments corresponds to a linear interpolation of the mean.

Analytical expressions for the higher order moments are less straightforward. However, these analyses provide some important information about the relationship between distribution shapes and moments and suggest similarities and differences between direct interpolation of the moments and linear interpolation of cumulative distributions. Specifically, the expression for the variance of the interpolation result elucidates some of the properties of linear cumulative distribution interpolation. To simplify the derivation, we consider only zero mean distributions. The result generalizes easily to interpolation between non-zero mean distributions. First, for notational simplicity, we rewrite the argument of the cumulative distribution upon linear cumulative distribution interpolation as

\[ y_q[x] = r y_q(x_a) + (1 - r)(y_q(x_{a+1})) \], \quad (6.15)\]

where

\[ r = \frac{x_{a+1} - x}{x_{a+1} - x_a} \] \quad (6.16)\]

Next, note that the variance of the zero mean random variable \( Y[x] \) is simply given by

\[ E[Y^2[x]] = \int_{0}^{1} y_q^2[x] dq = \int_{0}^{1} [r y_q(x_a) + (1 - r)(y_q(x_{a+1})))^2 dq \quad (6.17) \]
Rewriting the expression,
\[
E[Y^2[x]] = r^2 \int_0^1 y^2(x_a) dq + (1-r)^2 \int_0^1 y^2(x_{a+1}) dq \\
+ 2r(1-r) \int_0^1 y(x_a) y(x_{a+1}) dq
\]

(6.18)

Surprisingly, regardless of the actual cumulative distributions that are interpolated, the variances of the intermediate cumulative distributions are rather strictly bounded. We state the bounds:
\[
E[Y^2[x]] > r^2 \int_0^1 y^2(x_a) dq + (1-r)^2 \int_0^1 y^2(x_{a+1}) dq \]
\[
< (r \sqrt{\int_0^1 y^2(x_a) dq} + (1-r) \sqrt{\int_0^1 y^2(x_{a+1}) dq})^2
\]

(6.19)

The bounds are easily verified using the Cauchy-Schwarz inequality. Consequently, we find that variance of the intermediate distribution is bounded, but not exactly determined, by the variances of the two endpoint distributions.

Similar expressions bounding the higher moments of distributions generated through linear cumulative distribution function estimation could not be calculated.

### 6.2. Interpolation of Distribution Parameters

In contrast to interpolation of sample cumulative distribution functions, procedures for interpolating the parameters of a distribution require little computation and produce distributions that automatically match a parametric model. However, the parametric model also limits the possible shape for the interpolated distribution, so that a comparison between cumulative distribution interpolation and parametric interpolation is important. Many distributions, including beta distributions, are parametrized by a finite
set of moments; consequently, a procedure for interpolating sample moments from
known distributions is described. Further, because the study focuses on beta distribution
representations for random variables, a direct method for estimating the beta distribution
parameters is considered. These methods for interpolating known sample distributions
are compared with direct interpolation of cumulative distribution function arguments.

6.2.1. Interpolation of Beta Distributions using Moment Samples

Knowledge of the cumulative distribution functions \( F_y(y)[x_1]...F_y(y)[x_N] \)
implies knowledge of the moments and central moments of the distributions at
\( x_1, x_2, ..., x_N \). Because many of the interpolation problems involve beta distributions,
methods for interpolating the mean and the second, third, and fourth central moments of a
set of distributions is discussed. The formulated method is especially useful for
interpolating beta distributions because the results of the interpolation are also necessarily
beta distributions.

First, assume that the mean of the random variable \( Y[x] \), labeled \( u'(x) \), is known
for \( x \in x_1, x_2, ..., x_N \). Either a band-limited or linear interpolation of the mean may be
completed to determine \( u'(x) \) for each \( x \). As shown in the analysis of the cumulative
distribution extrapolation, this extrapolated mean is identical to the mean of the
distribution derived from the respective band-limited or linear cumulative distribution
function extrapolation. The equivalence of the two mean curves justifies the linear or
band-limited interpolation of the mean.

If \( u'(x) \) is known for an evenly spaced set of independent parameter values,
\( x \in \{0, \pm T_N, \pm 2T_N, ...\} \), then the band-limited reconstruction is
\[ u'(x) = \sum_{n=-\infty}^{\infty} u'(x - nT_N) \text{sinc}(\pi(n \frac{x}{T_N})). \quad (6.20) \]

As before, the mean is typically only known for a finite number of samples, so that an exact band-limited reconstruction is impossible. Consequently, the known samples are reversed with respect to the independent parameter, and the new set of samples is appended to the original sample set. The band-limited reconstruction of the mean is then completed on the appended set of samples.

The linear interpolation of the mean is even more straightforward. For \( x_a < x < x_{a+1} \), where \( x_a \) and \( x_{a+1} \) are the independent parameter values for two adjacent mean samples, the linearly interpolated mean is

\[ u'(x) = u'(x_a) + \frac{x - x_a}{x_{a+1} - x_a} (u'(x_{a+1}) - u'(x_a)). \quad (6.21) \]

The calculation of the variance for the linear interpolation of cumulative distribution functions suggests that linear interpolation of the square root of the variance is an appropriate procedure for directly interpolating the variance. Specifically, if two sample distributions are identical in shape but differ in variance, a linear interpolation of the cumulative distribution function corresponds to a linear interpolation of the square root of the variance. Consequently, linear interpolation of the square root of the variance is used to deduce the variance interpolation. For \( x_a < x < x_{a+1} \), where \( x_a \) and \( x_{a+1} \) are the independent parameter values for two adjacent mean samples, the interpolated variance is

\[ u_2(x) = \left[ \sqrt{u_2(x_a)} + \frac{x - x_a}{x_{a+1} - x_a} (\sqrt{u_2(x_{a+1})} - \sqrt{u_2(x_a)}) \right]^2. \quad (6.22) \]

The chosen interpolation method for the third and fourth central moments is especially useful for the interpolation of beta distributions because the resulting
interpolated distributions are also beta distributions. Additionally, simulations suggest that the interpolated results are comparable to the moments resulting from linear cumulative distribution interpolation. Specifically, linear interpolation of the squared normalized skewness and the normalized kurtosis is the method chosen for interpolating the skewness and kurtosis. The squared normalized skewness is defined as

\[
\beta_1(x) = \frac{(u_3(x))^2}{(u_2(x))^3},
\]

while the normalized kurtosis is

\[
\beta_2(x) = \frac{(u_4(x))}{(u_2(x))^2}.
\]

Next, a plot of \( \beta_2(x) \) as a function of \( \beta_1(x) \text{sgn}(u_3(x)) \), with \( x \) considered a parameter, is helpful for visualizing the extrapolation technique (Figure 6.5). Also, the boundaries for the beta distribution in the plane defined by \( \beta_1(x) \text{sgn}(u_3(x)) \) and \( \beta_2(x) \), as well as the coordinates of several possible sample points for independent parameter values \( x_1, x_2, \) and \( x_3 \), are shown. For two distributions with the same direction of skewness, or equivalently the same sign for the skewness, an interpolation of the moments that follows a straight line in the \( \beta_1(x) \text{sgn}(u_3(x)) - \beta_2(x) \) plane will always have moments corresponding to a beta distributions. In this case, linear interpolations of \( \beta_1(x) \) and \( \beta_2(x) \) with respect to \( x \) are sufficient, though not necessary, to guarantee that the resulting moments correspond to beta distributions. A variety of other interpolation schemes follow straight lines in the \( \beta_1(x) \text{sgn}(u_3(x)) - \beta_2(x) \) plane, but the higher moments generated by linear interpolation of \( \beta_1(x) \) and \( \beta_2(x) \) are similar to the higher
moments resulting from cumulative distribution interpolation. Mathematically, for the example points in the figure with independent parameter values $x_1$ and $x_2$,

$$\beta_1(x) = \beta_1(x_1) + \frac{x - x_1}{x_2 - x_1} (\beta_1(x_2) - \beta_1(x_1))$$

(6.25)

and

$$\beta_2(x) = \beta_2(x_1) + \frac{x - x_1}{x_2 - x_1} (\beta_2(x_2) - \beta_2(x_1)).$$

(6.26)

The kurtosis is determined by multiplying $\beta_2(x)$ by the square of the variance, while the skewness is determined by multiplying $\beta_1(x)$ by the cube of the variance and then taking the appropriate square root. If the signs of the skewness at two adjacent sample points differ, the interpolation is more complicated. If the straight line connecting the two points in the $\beta_1(x) \text{ sgn}(\mu_3(x)) - \beta_2(x)$ plane remains within the domain of a beta distribution, then linear interpolation of $\beta_1(x) \text{ sgn}(\mu_3(x))$ and $\beta_2(x)$ suffices for higher moment interpolation. However, if this line escapes the boundaries of the beta distribution, interpolation of the higher moments becomes more complicated. Several possible paths for interpolation in the higher normalized moment plane are shown (Figure 6.5). The simplest interpolation that produces beta distribution moments involves choosing the Gaussian distribution ($\beta_1 = 0$ and $\beta_2 = 3$) as an intermediate distribution.

The independent parameter value for this intermediate distribution must be chosen so that the rate of movement along the interpolation curve in the $\beta_1(x) \text{ sgn}(\mu_3(x)) - \beta_2(x)$ plane remains constant for the interpolation. The interpolation of the higher moments for two sample distributions may readily be extended to problems with multiple sample
distributions since linear interpolation is utilized. Band-limited interpolation is also possible, though the interpolated distributions may then no longer be beta distributions.

Figure 6.5.
Possible paths of interpolation of the higher moments are shown. The interpolation sometimes is not a straight line with respect to the higher moments if the sign of the skewness changes.

Again, an example of distribution interpolation is shown. In this case, both the cumulative distribution functions and the probability density functions of the interpolated random variables are shown (Figure 6.6). Also, two three dimensional graphics showing the continuum of intermediate distributions are provided. For convenience in display, the intermediate distributions are plotted in even intervals between their endpoints and not as a function of an absolute scale. The two plots are simply different viewpoints to the continuum of distributions.
Distribution Interpolation using Four Moments

Value of the Random Variable

Figure 6.6.
Interpolation of distributions using moment regression is shown. 6.6a. and 6.6b. show the interpolated distributions for several independent parameter values. Three dimensional plots of the continuum of interpolated distributions is also shown from two different angles.
Chapter 7

Improvements and Generalizations in Estimation

7.1. Model Selection Using the Phase Spectrum

For some data sets, modeling based both on the independent parameter and based on other samples may be necessary. In the context of modeling time series, we described several methods for incorporating independent parameter-dependent information into the model. A more systematic approach, however, includes first using linear regression-based models or other models to determine the effect of the independent parameter on the mean and variance of the data set. Once this dependence has been determined, the data set can be corrected to eliminate the dependence on the independent parameter. The data can then be modeled based on other data samples using autoregressive methods. An important unanswered question, however, is the choice of model order for the independent parameter-based model. The standard hypothesis testing procedures may not work, since the data remains correlated even after removal of the independent parameter-dependent information.

Instead, we propose a test that determines if a data set contains any independent parameter dependent information. Specifically, we believe that correlation in the phase spectrum of a data set indicates that the data set contains independent parameter-based information. Consequently, once a model has been used to remove independent
parameter-based information, correlation in the phase spectrum can be tested to verify model sufficiency.

A proof that an uncorrelated phase spectrum corresponds to a data set that does not depend on the independent parameter is not attempted, and instead phase spectra are plotted for an example data set (Figure 7.1). Much more research in using spectra to test for independent parameter-based information is necessary.

Figure 7.1.
Plots of silicon price data and associated phase spectra before and after independent parameter-based information has been removed.
7.2. Data Estimation as a Function of Several Independent Parameters

In many problems, a set of data can be modeled as a function of several known parameters, rather than one parameter. The technique for modeling data as a function of several parameters is analogous to the process of modeling data as a function of one independent parameter. Specifically, a set of basis functions that may describe the describe the moments or probability density function parameters of the data is formulated, and weighted linear combinations of these basis functions are used to describe the data set. The appropriate weights for the basis functions are determined one of the described estimation techniques.

More mathematically, we attempt to model a multivariate continuum of random variables $Y(X)$ based on $N$ instances of the set of random variables, labeled $y_1(\bar{x})...y_N(\bar{x}_N)$. The continuum of random variables is modeled with a set of beta distributions at each independent parameter vector value, which is equivalent to modeling the four moments as a function of the independent parameter vector. The significant new task in multivariate modeling is the determination of appropriate basis functions for the data. Once a sufficient set of basis functions has been chosen, the estimation of parameters is identical to the scalar modeling problem. Either maximum likelihood techniques or regression of the moments can directly be used to estimate the parameters. If the cumulative distribution function fitting method is used, the procedure is slightly different in that a multidimensional cumulative distribution must be used to describe the independent parameter, but the remainder of the procedure is identical to before.
7.2.1. Choice of Basis Functions for the Multivariate Problem: An Overview

Throughout the literature on linear regression, various different models are used to describe characteristics of a data set. We do not attempt to cover the breadth of possible multivariate models but simply highlight a couple model forms that were used in modeling multivariate data sets for this study. Since least squares regression of the moments, rather than maximum likelihood estimation or CDF fitting, is the primary technique chosen to estimate parameters, this section focuses on models of the four central moments that can be determined directly through regression.

Most generally, the first four moments of the beta distribution can be modeled as a linear combination of arbitrary basis functions of the independent parameter vector:

\[ \tilde{u}(\bar{x}) = \sum_{i=1}^{k} p_{iu} \phi_i(\bar{x}) , \]  

\[ \tilde{u}_2(\bar{x}) = \sum_{i=1}^{k} p_{iua2} \phi_{i2}(\bar{x}) , \]  

\[ \tilde{u}_3(\bar{x}) = \sum_{i=1}^{k} p_{iua3} \phi_{i3}(\bar{x}) , \]  

\[ \tilde{u}_4(\bar{x}) = \sum_{i=1}^{k} p_{iua4} \phi_{i4}(\bar{x}) , \]

where \( \tilde{u}(\bar{x}), \tilde{u}_2(\bar{x}), \tilde{u}_3(\bar{x}), \) and \( \tilde{u}_4(\bar{x}) \) represent the estimates for the mean and the second, third, and fourth central moments. As with the scalar independent parameter problem, the normalized skewness and normalized kurtosis, rather than the third and fourth central moments, are instead modeled as linear combinations of basis functions to reduce the number of necessary parameters:
\begin{align*}
\tilde{\beta}_1(\bar{x}) &= \sum_{i=1}^{k_{p1}} p_{i\beta_1} f_i(\bar{x}), \\
\tilde{\beta}_2(\bar{x}) &= \sum_{i=1}^{k_{p2}} p_{i\beta_2} f_i(\bar{x}).
\end{align*}

When data are modeled as a function of one independent parameter, the basis functions are chosen as monomials. In the vector case, the polynomial model for the moments is slightly more complicated. In addition to monomials of individual independent parameters, cross terms involving several independent parameters are also used to describe the data. Most generally, a polynomial basis function in the multivariate problem may assume the following form:

\[ f(\bar{x}) = f(x_1 \ldots x_m) = x_1^{k_1} \ldots x_m^{k_m}. \]

The mean, variance, normalized skewness, and normalized kurtosis can then be described as polynomial functions of basis functions of this form. As the number of independent parameters increases, the number of possible basis function increases greatly, so that the proper number of basis functions used in the regression must be chosen carefully.

Typically, we complete the regression with all basis functions such that \( k_1 + \ldots + k_m \leq k \) for several values of \( k \) until the data are sufficiently described; additionally, we remove any basis functions that fail an appropriate hypothesis testing criterion. The number of basis functions such that \( k_1 + \ldots + k_m \leq k \) is given by

\[ n(k, m) = \frac{(k + m + 1)!}{k!(m + 1)!}. \]

Polynomial models can be utilized to describe the mean, variance, and normalized higher moments. Typically, the polynomials chosen for the higher moments will have lower order than the polynomial chosen for the mean.
7.2.2. Some Thoughts on Estimation and Estimates in the Multivariate Problem

The procedures for regression and maximum likelihood estimation of a multivariate continuum of distributions are identical to the procedures in the single independent parameter problem. When regression techniques are utilized, autocorrelation and spectral techniques can first be utilized to determine the sufficiency of the model. If the model sufficiently describes the data, variances in the model parameters can be calculated, and so the variance in the moment estimates can be calculated. These procedures are analogous to the procedures in the univariate case and so are not repeated. Note that an autocorrelation cannot be directly calculated, but a measure of the data correlation can nevertheless be obtained easily by calculating the sample distance normalized correlation. Again, bias correction for the higher moments may be necessary.

The multivariate problem differs from the univariate problem in some respects. For example, the idea of extrapolation of the result becomes less lucid. In the single independent parameter problem, the endpoints of the known data sample were clearly defined, so that the results beyond these data samples might be considered an extrapolation of the result. In the multiple independent parameter problem, the distinction between interpolation and extrapolation are fuzzy. Because the model formulations do not distinguish between interpolation and extrapolation, however, the validity of the determined model is no different for interpolated and extrapolated points and so the distinction is not really important.

Unfortunately, we did not study the connection between multiple independent parameter and single independent parameter based regression systematically. Two
examples elucidate some of the features of multivariate distribution estimation, however. In one example, we generate data that depends on two parameters. We then perform a regression of the data based on only one of the parameters and also based on both parameters (Figure 7.2). The figure demonstrates that the model can much more accurately specify the value of the dependent variable when both independent parameters are used for estimation. In the second example, we generate 50 data samples from a univariate continuum of distributions and estimate the mean of this continuum of distributions. As a contrast, we also generate 50 data samples from an analogous multivariate continuum of distributions and estimate the parameters in the mean model. The variances in the parameter estimates of the multivariate model are larger the variances of the parameter estimates in the univariate model, suggesting that more points are necessary for estimation. However, the difference in the two variances is not large; based on several examples, we believe that the number of points necessary for regression based on two parameters is approximately two times the number of points necessary for an analogous univariate regression.
Figure 7.2.
The curve of best fit for the mean of a set of data that varies with two parameters is shown. Additionally, a regression curve for the mean is shown for a model that uses only one of the two parameters to describe the data. 7.2c. and 7.2d. also show the estimated probability and cumulative distribution at a specific pair of independent parameter values. The estimated variance of the data is much larger if only one parameter is used to model the mean.

7.3. **Relationships between Parameter Estimate Variances**

The representation of an estimated continuum of distributions can be formed either in terms of the moments of the distributions or in terms of the four distribution parameters. Regression estimation of a distribution continuum leads to estimates for the moments and corresponding variances in these estimates. A calculation of the
corresponding variances in the four beta distribution parameters is useful for gauging the accuracy of the estimate. Similarly, calculation of moment variances given parameter variances is useful for understanding the estimation of beta distributions.

7.3.1. Linearization of the Moment Equations: Calculation of Moment Variances

In order to approximately calculate the variances in moments given the variances of the distribution parameters, we approximate small changes in the moments as a linear function of small changes in the beta distribution parameters. In this section, we label $\Delta x$ as an incremental change in the variable $x$. Using this notation, a small change in the mean of a beta distribution due to changes in the distribution parameters can be approximated as

$$
\Delta u_1' = \frac{\beta}{\alpha + \beta} \Delta a + \frac{\alpha}{\alpha + \beta} \Delta b + (b-a) \frac{\beta}{(\alpha + \beta)^2} \Delta \alpha \\
+ (b-a) \frac{-\alpha}{(\alpha + \beta)^2} \Delta \beta
$$

(7.5)

Similarly, a differential change in the variance due to beta parameter variations can be calculated as

$$
\Delta u_2 = -2(b-a) \frac{\alpha \beta}{(\alpha + \beta + 1)(\alpha + \beta)^3} \Delta a + 2(b-a) \frac{\alpha \beta}{(\alpha + \beta + 1)(\alpha + \beta)^2} \Delta b \\
+ (b-a)^2 \frac{(\alpha + \beta + 1)(\beta - 2\alpha) + \alpha}{(\alpha + \beta)^3(\alpha + \beta + 1)^2} \Delta \alpha + \frac{(\alpha + \beta + 1)(\alpha - 2\beta) + \beta}{(\alpha + \beta)^3(\alpha + \beta + 1)^2} \Delta \beta
$$

(7.6)

Similar expressions can be formulated for the third and fourth central moments, but they are complicated and not elucidative.

In an estimation problem, the parameter estimates can usually accurately be modeled as jointly Gaussian random variables with the parameters as the mean of each variable and the covariance matrix of the parameters as the variance structure.
Consequently, $\Delta a, \Delta b, \Delta \alpha$, and $\Delta \beta$ can be modeled as zero-mean jointly Gaussian random variables, and the consequently $\Delta u_1$ and $\Delta u_2$ are approximately Gaussian random variables as long as the parameter variances are small. The variance of the mean and variance estimates can be calculated given the covariance matrix of the parameter estimates.

One important observation from this linearized model is the insensitivity of the mean and variance estimates to variations in the shape parameter estimates when the shape parameters become large. When the shape parameters are large, a differential change in one of the parameters only slightly affects the shape of the distribution and so does not affect the mean and variance significantly.

7.3.2. Linearization of the Beta Parameter Equations: Calculation of Parameter Variances

The shape parameters of the beta distribution depend solely on the higher normalized central moments of the distribution, so the differential change in the shape parameters due to differential changes in the normalized skewness and kurtosis are calculated. In this development, we assume that the normalized skewness is positive. If the normalized skewness is smaller than zero, the result for the two shape parameters is simply reversed. The expressions for the differential changes in each shape parameter are given:

$$
\Delta \beta = \left[ \frac{1}{2} \frac{\partial r}{\partial \beta_1} + (r + 1) \frac{\partial r}{\partial \beta_1} + \frac{1}{2} r(r + 2)Q_1 \right] \Delta \beta_1 + \left[ \frac{1}{2} \frac{\partial r}{\partial \beta_2} + (r + 1) \frac{\partial r}{\partial \beta_2} \right] \Delta \beta_2 + \frac{1}{2} r(r + 2)Q_2 \Delta \beta_2
$$

(7.7a)

where
\[ Q_1 = \left( \frac{1}{2} \sqrt{\frac{(r+2)^2 \beta_1 + 16(r+1)}{\beta_1}} \right) \]

\[ (r+2)^2 \beta_1 + 16(r+1) - \beta_1 \left( (2\beta_1(r+2) + 16) \frac{\partial r}{\partial \beta_1} + (r+2)^2 \right), \]  

\[ \left( \frac{1}{((r+2)^2 \beta_1 + 16(r+1))^2} \right) \]  

\[ Q_2 = \frac{1}{2} \left( \sqrt{\frac{(r+2)^2 \beta_1 + 16(r+1)}{\beta_1}} \right)^2 \left( \frac{\beta_1 (2\beta_1(r+2) + 16) \frac{\partial r}{\partial \beta_1}}{((r+2)^2 \beta_1 + 16(r+1))^2} \right), \]  

\[ r = \frac{6(\beta_2 - \beta_1 - 1)}{6 + 3\beta_1 - 2\beta_2}, \]  

\[ \frac{\partial r}{\partial \beta_1} = \frac{-18 - 6\beta_2}{(6 + 3\beta_1 - 2\beta_2)^2}, \]  

and

\[ \frac{\partial r}{\partial \beta_2} = \frac{24 + 6\beta_1}{(6 + 3\beta_1 - 2\beta_2)^2}. \]  

Similarly,

\[ \Delta \alpha = \left[ \frac{1}{2} \frac{\partial r}{\partial \beta_1} - (r+1) \frac{\partial r}{\partial \beta_2} - \frac{1}{2} r(r+2)Q_1 \right] \Delta \beta_1 + \left[ \frac{1}{2} \frac{\partial r}{\partial \beta_2} - (r+1) \frac{\partial r}{\partial \beta_2} \right] \Delta \beta_2 \]  

\[ - \frac{1}{2} r(r+2)Q_2 \Delta \beta_2. \]  

Again, if the estimates for the normalized skewness and kurtosis are considered jointly Gaussian random variables, a reasonable assumption if the estimates originate for a regression analysis, the estimates for the shape parameters will also approximately be Gaussian random variables with an easily calculated mean and variance.

Unfortunately, the expressions for the differential change in the shape parameters are complicated enough that simple insights about the effect of the normalized moments
on the distribution parameters are not obvious. However, insights about the relationship between errors in the moments and errors in the shape parameters can be determined for constrained examples, such as distributions without any skewness or distributions with a normalized kurtosis of three, which corresponds to the Gaussian distribution when the skewness is also zero. We present results for the problem where the normalized skewness is constrained to be zero and is also assumed not to vary differentially:

\[
\Delta \alpha = \Delta \beta = \frac{1}{2} \frac{\partial r}{\partial \beta_2} \Delta \beta_2 = \frac{3}{(3 - \beta_2)^2} \Delta \beta_2. \tag{7.9}
\]

The expression suggests that a small change in the normalized kurtosis achieves a large change in shape parameters whenever the distribution approaches a Gaussian distribution. Consequently, as the estimated distribution approaches the Gaussian distribution, estimates for the shape parameters become less and less accurate in an absolute sense given the same variance in the normalized moments.

### 7.4. An Improvement of the Mean Estimate

In this section, we discuss another technique for estimation based on the likelihood function of a set of data. Specifically, an improvement of the mean estimate for data originating from a single beta distribution is discussed. The suggested technique, which is based on finding the mean value of the likelihood function rather than the maximum of the function, can perhaps be used for the estimation of a continuum of varying distributions and not just a single mean. However, only an introduction to the topic is given in this thesis; hopefully, further research in using a likelihood mean based estimator will be completed.
7.4.1. Likelihood Mean Based Estimator

Consider a random variable \( Y \) parametrized by one unknown parameter, labeled \( p \).

Next, consider instances of \( N \) independent, identically distributed random variables \( y_1 \ldots y_N \) with the same probability distribution as \( Y \). Then the likelihood function is

\[
L(p) = \prod_{i=1}^{N} p_Y(y_i ; p). \tag{7.10}
\]

The proposed estimate for the parameter \( p \) is

\[
\hat{p} = \frac{\int pL(p)dp}{\int L(p)dp}. \tag{7.11}
\]

This estimate for the parameter is optimal in several respects. First, the estimator minimizes

\[
\int (\hat{p} - p)^2 L(p)dp \tag{7.12}
\]

with respect to \( \hat{p} \). Further, Pitman proves that the estimator is the optimal estimator in the least squares sense if the parameter \( p \) is a location parameter [21]. Specifically, if

\[
p_Y(y; p) = q(y - p), \tag{7.13}
\]

where \( q \) is an appropriate probability density function, then

\[
\int (\hat{p} - p)^2 L(p)d\bar{y} \tag{7.14}
\]

where \( \bar{y} \) is a set of \( N \) observations and the integral represents a set of \( N \) integrals.
7.4.2. Application of the Estimator for Improvement of the Mean Estimate.

For non-Gaussian probability density functions with a known form, the likelihood mean based estimator allows optimal estimation of the mean of the distribution. In the context of estimation of beta distributions, assume that another technique has been utilized to estimate the variance, skewness, and kurtosis of a beta distribution based on N data points generated by that beta distribution. If the three higher moments are known, the probability density of the distribution to be estimated can be written as

\[ p_y(y; p) = q(y - p) \]

where \( p \) is the mean of the beta distribution and \( q(y) \) is the zero mean beta distribution with the previously estimated values for variance, skewness, and kurtosis. The estimation technique outlined in the last section can then be utilized to estimate the mean of the beta distribution. If the variance, skewness, and kurtosis were exactly known, the mean estimate would be provably optimal. Monte Carlo trials demonstrate that the mean estimate improves using the likelihood function based estimator for some data sets.
Chapter 8

Conclusions and Future Work

8.1. Summary and Conclusions

The objective of this research was the development of techniques for modeling arbitrary data sets that vary as a function of one or more independent parameters with a continuum of beta distributions. The developed technique for modeling data sets can be subdivided into three processes: formulation of beta distribution based models, estimation of model parameters, and evaluation of the model form and the parameter estimates.

The first step in modeling data sets, the development of appropriate beta distribution models for data, is the least structured component of the modeling process. We formulated several different models; specifically, the probability distribution at each independent parameter value was modeled as an independent beta random variable for non-time series data, while conditional distributions were modeled as beta distributions for time series data. Further, the dependence of the beta distribution on the independent parameter was specified by writing the moments or beta distribution parameters as combinations of basis functions of the independent parameter. The developed models can appropriately model a variety of different real data sets. Each model formulated in the work was a parametric model, so that the problem of estimating an infinite number of probability density functions was reduced to a problem of estimating a finite set of parameters.

Much of the study focuses on the estimation of parameters for the given models.
Three techniques were used to estimate beta distribution parameters: maximum likelihood estimation of certain beta parameter-based models, least squares regression of the distribution moments, and least squares sample cumulative distribution function fitting. Each method was modified to accurately estimate beta distribution continua. Regression of the distribution moments was chosen as the primary estimation technique because of robustness of the estimate accuracy and because of the possibility for calculating the estimate variances. However, modified versions of all three techniques are useful for modeling some data sets. Also, the techniques can be applied to a variety of different beta distribution models, including many models that were not discussed in the context of the work.

Thirdly, several techniques for determining the accuracy of the chosen model and parameters are considered. The autocorrelation of the data, as well as spectral representations for the data, can be used to verify the sufficiency of the chosen model. Additionally, when regression estimation is used, the variance of the parameter estimates can be used to remove excess parameters from a model.

In addition to the modeling of arbitrary data sets, the research addresses several other questions. The question of interpolating known distributions is considered, and both moment and cumulative distribution function-based techniques are used to interpolate distributions. We also briefly explore possibilities for optimal estimation and relationships among the variances of beta random variable parameters, moments, and density functions.
8.2. Future Work

This study highlights the importance of developing techniques for modeling arbitrary data sets with a set of non-Gaussian distributions. Because the dependence on the model on the independent parameter is not very structured, appropriate model selection is a difficult task. Further, estimation of parameters in a non-stationary model with non-Gaussian probability density functions has not been studied generally and systematically. We highlight some of the specific questions that may lead to improvements in modeling arbitrary data sets probabilistically.

8.2.1. Future Work in Model Selection

Future work in model selection will focus on the formulation of models that can describe a wide variety of data sets. One possible improvement in model selection may result from the systematic study of basis functions for describing the moments of sets of data. We found the problem of finding basis functions that guarantee appropriate modeling of arbitrary data sets to be a difficult problem. A systematic study of possible basis functions for sets of data and the development of a technique for determining which sets of basis functions are optimal for describing data is important. Specifically, basis functions for modeling the mean of a set of data, but choosing basis functions for the higher moments is more difficult because of the constraints on these moments.

In addition to analysis of possible basis functions, future work in formulating other models is important. For example, non-parametric representations for the moments of a data set should be explored. These representations may allow modeling of data sets that are not simple functions of the independent parameters. Also, chaos theory based
models rather than stochastically based models for data sets, and specifically for times series data, may be useful [22]. Similarly, wavelet based models may provide improved models for the moments. Another interesting model formulation, briefly explored in the context of cumulative distribution function fitting, describes the independent parameter as a random variable.

Further work in modeling data as a function of several parameters is also necessary. One important topic is the modeling of data with time as one of several independent parameters.

8.2.2. Future Work in Estimation

Development of optimal estimation procedures is one important direction for future work. In this project, we primarily focused on producing estimates with sufficient, measurable accuracy characteristics rather than optimal estimators. Unfortunately, the development of optimal estimators for arbitrary non-stationary, non-Gaussian processes is an unsolved problem. The development of optimal estimators for specific models of this sort is an important direction for future research. A study of the work of Pitman on estimation may provide a starting point for the development of optimal estimators for non-Gaussian processes. The analysis of arbitrary time series is especially important because of its usefulness in prediction. Weiner, Kalman, and several other astute researchers developed procedures for analyzing stationary Gaussian time series. An extension of these techniques to non-stationary and non-Gaussian approach is an important, though perhaps difficult approach, to the analysis of data sets.
Also, each of the estimation techniques discussed in this thesis can be further analyzed. For example, each of the adjusted maximum likelihood estimators must be systematically analyzed, and the variance of the various estimators should be calculated. Similarly, the accuracy of estimates using multivariate cumulative distribution function fitting was not discussed and is a possibility for future research.

8.2.3. Other Possibilities for Future Work

Interpolation of distributions is one topic that deserves further analysis. Based on our literature survey, we believe that little work in the interpolation of distributions has been completed. Further research of distribution interpolation may prove to be useful and interesting. Future research may focus on possible applications of distribution interpolation; also, the band-limited interpolation of distributions relates to the idea of constraining or specifying the probabilistic description of a random process and can be studied in this context.

Several generalizations of the original problem are also important. Although the estimation techniques developed in the work are fairly general, the model formulation is specific to beta distribution-based models. Ideally, techniques will be developed that can be used to model data with arbitrary probability density functions.

8.3. Final Thoughts

One of the goals of science and engineering is understanding the behaviors of systems in nature and in human society. When we attempt to model these systems, we seek answers to a variety of questions: Based on the information that we have, how well
can the system be described? Is the actual behavior of the system, or the behavior of the observed system, deterministic or random? If some aspect of the system cannot be modeled deterministically, how can this randomness be modeled? What are the underlying principles that guide the behavior of the system, and how do these principles lead to proper models for the system?

Modeling of data sets is one approach for attempting to describe the behavior of some systems. Within the context of describing data sets, models that parametrize random, uncertain, or unknown behavior are sometimes important. We consider one method for modeling data sets with parametric models for uncertainty. Hopefully, the work in this thesis is a small contribution toward understanding the behavior of systems and answering the fundamental questions associated with these systems.
Bibliography


