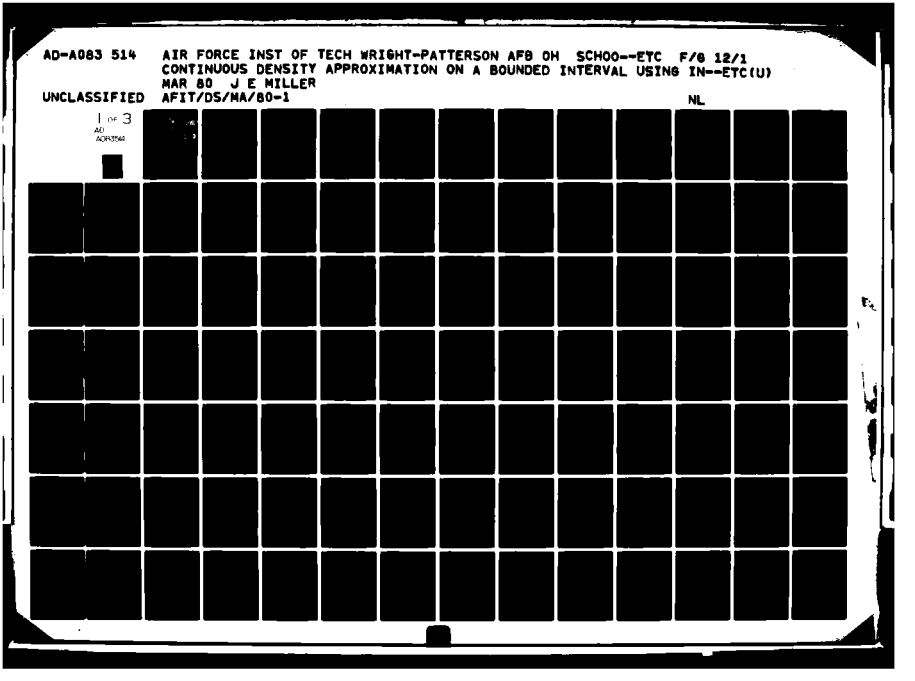
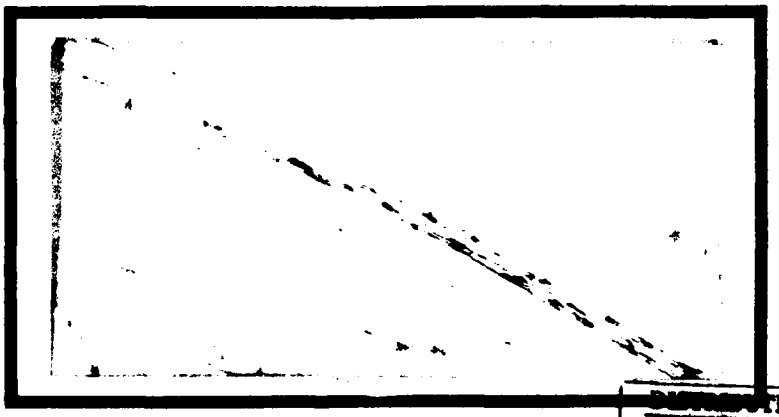
AD-A083 514 AIR FORCE INST OF TECH WRIGHT-PATTERSON AFB OH SCHOO--ETC F/G 12/1
CONTINUOUS DENSITY APPROXIMATION ON A BOUNDED INTERVAL USING IN--ETC(U)
MAR 80 J E MILLER
UNCLASSIFIED AFIT/DS/MA/80-1 NL





UNITED STATES AIR FORCE

AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT/DS/MA/80-1

CONTINUOUS DENSITY APPROXIMATION ON A BOUNDED INTERVAL USING INFORMATION THEORETIC CONCEPTS

DISSERTATION

AFIT/DS/MA/80-1 James E. Miller, Jr
Captain USAF

AFIT/DS/MA/80-1

CONTINUOUS DENSITY APPROXIMATION ON A BOUNDED INTERVAL USING INFORMATION THEORETIC CONCEPTS

by

James E. Miller, Jr., B.S., M.S.

Captain USAF

Approved:

Richard W. Trulp Chairman	22 Feb 1980
Albert H. Moore	22 Feb 1980
Kenneth Mahn	22 Feb 1980
George E. Ow	22 Feb 1980
[illegible]	[illegible]
Ronald Tummala	22 Feb 1980

Accepted:

J. S. Przemieniecki Dean, School of Engineering	26 Feb 1980

AFIT/DS/MA/80-1

CONTINUOUS DENSITY APPROXIMATION ON A BOUNDED INTERVAL USING INFORMATION THEORETIC CONCEPTS

DISSERTATION

Presented to the Faculty of the School of Engineering
of the Air Force Institute of Technology
Air University
In Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

by

James E. Miller, Jr., B.S., M.S.
Captain USAF

March 1980

ACKNOWLEDGEMENTS

This research was sponsored by the Air Force Flight Dynamics Laboratory (AFFDL), Vehicle Synthesis Branch. I wish to express thanks to AFFDL management for the excellent support and interest. A special word of thanks is due to Dr. Squire Brown for his continued involvement and consultation. Dr. Brown's expertise was very influential in the problem formulation and practical application phases of the research.

I am deeply indebted to the members of my advisory committee for their enthusiastic support and suggestions in all aspects of the research. I especially want to thank Major Richard Kulp and Captain George Orr. Captain Orr provided the initial impetus in suggesting the area of study and in structuring my formal education. Major Kulp, committee chairman, provided stimulating advice, encouragement and guidance throughout the course of this research effort. Association with this group of professionals was a rewarding experience.

To my wife, Leslie, and my daughter, Tammy, I wish to express loving gratitude for their patience, understanding and selfless cooperation through the graduate program.

Contents

List of Figures

List of Tables

Abstract

This report presents the theoretical development and numerical implementation of a procedure for approximating continuous probability density functions on a bounded interval. The work is applicable to Bayesian decision models in that available information is used to update or obtain the prior distribution. The procedure is based on the solution of a constrained entropy maximization problem and requires information in the form of expected values of "information functions." The approach involves three steps: estimation of expected (or average) values of "potential" information functions, selection of the "active" subset of functions to define the approximation family, and simultaneous solution of the constraints to select the specific approximating density for a given set of data.

A useful set of potential information functions is developed, and three numerical methods for active set selection are demonstrated. Numerical techniques for expected value computation are discussed, and a scheme for solution of the constraints is developed and implemented. Theoretical development includes theorems on form and uniqueness. Approximation accuracy is related to potential set definition and data accuracy. The procedure is applied to several known distributions to demonstrate applicability. Applications to computer simulation and interval arithmetic models are demonstrated with specific examples.

CONTINUOUS DENSITY APPROXIMATION ON A BOUNDED INTERVAL USING INFORMATION THEORETIC CONCEPTS

Chapter I. Introduction

This dissertation concerns the representation or approximation of unknown probability distributions. The proposed approximation method is based on the concept of maximum entropy and uses known or calculable information about the unknown distributions. The work was motivated by Bayesian decision models, as discussed by Tribus (Ref 82), in that available information is used to update a prior estimate of the unknown distribution. The prior estimate is assumed to be the uniform distribution or is represented by a random sample from the unknown distribution. The initial estimate is updated via information in the form of expected values of certain "information functions." Selection of the information functions determines the form and accuracy of the approximating distribution.

The word "characterize" carries special meaning, for purposes of this dissertation, in describing the accuracy of approximation. Our use of the word is here defined to preclude later misinterpretation, and because our use is different from the usual statistical meaning. Assume that the unknown distribution is generated by the analytic density function $f(x)$, and let $p(x)$ symbolize the approximating

density. When p(x) is reducible to the exact form of f(x) to include the correct parameter values, then we say that p(x) characterizes f(x). If p(x) is of a different form than f(x), then p(x) approximates f(x). Thus "characterize" is used to indicate an exact representation of the unknown analytic density. Given this definition, we wish to characterize or accurately approximate the unknown distribution.

An initial concern of the research was to provide a method to represent the output distribution of a computer simulation in the interest of error propagation studies. Although the resulting method has direct benefit to simulation, the method can be applied to more general characterization or approximation problems. The following chapters present the proposed method in detail, discuss computer implementation of the method to include efficient numerical techniques, and investigate potential applications.

Chapter II provides a background summary of the information theoretic concepts which form the foundation of the proposed method. Concepts such as information variation and maximum entropy are discussed as well as recent applications of these concepts. Chapters III through VIII discuss the proposed characterization method and numerical techniques for implementation. The method is applied to computer simulation in Chapter IX followed by discussion of method sensitivity in Chapter X. Additional applications are presented in Chapter XI. The paper concludes with a [illegible] future research in Chapter XII.

Chapter II. Background

Distribution Approximation with Maximum Uncertainty

The problem of interest concerns the characterization of an unknown distribution based on information that is provided, or information that one may obtain, concerning the distribution. We assume that the unknown distribution of random variable X is generated by an unknown probability density function, f(x), where X may be a vector and thus f(x) may be a multivariate distribution. We concentrate on providing an algebraic characterization or approximation, p(x), for the unknown density. We define "information" as anything that is known or assumed about the random variable or the distribution of the random variable. Clearly, the amount and nature of available or assumed information will greatly influence the resulting approximation, p(x). For example one may assume (or know) that the unknown density is normal, $N(\mu,\sigma^2)$, and obtain further information in terms of a random sample, x_i, i=1,2,...N. From the sample, one calculates

$$\bar{x} = \sum_{i=1}^{N} x_i/N \quad \text{and} \quad s^2 = \sum_{i=1}^{N} (x_i-\bar{x})^2/(N-1)$$

to approximate the mean and variance and, thus, determines the appropriate estimation of the unknown distribution,

i.e.,

$$p(x) = (2\pi s^2)^{-\frac{1}{2}} \exp[-(x-\bar{x})^2/(2s^2)]$$

If the form of the distribution is known, as in our example, then the problem reduces to one of parameter estimation. However, in many engineering or statistical problems, sufficient information to determine the form of an unknown distribution is not available.

Assuming a particular density form without evidence to support such a form will unnecessarily bias the distribution approximation and produce potential bias or error in the ultimate solution of the problem. Consequently, the method of distribution characterization that is proposed in this paper derives a density form that utilizes only the available information while maintaining "maximum uncertainty" with respect to other, unspecified information. As discussed in later sections, the method assumes that the specified information can be provided in terms of average values of certain functions of the random variable X. The method is based on a specific measure of uncertainty for a distribution, the distribution entropy.

Entropy and Information Variation

The entropy function, S, was defined by Claude Shannon (Refs 71; 72; 82) in 1948 as a measure of the uncertainty of a given answer to a well defined question. Shannon's work centered on communications theory, but

provided a basis for E. T. Jaynes' extension of entropy in statistical mechanics (Refs 42; 43; 44). Myron Tribus (Ref 82) consolidates the work of Shannon and Jaynes and provides a thorough discussion of the concept of maximum entropy. Tribus (Ref 82:111-117) presents a derivation of the entropy measure and several examples of entropy as a measure of uncertainty. We consider a simplified example for illustration and definition. For a complete discussion of the development of the entropy measure, see E. T. Jaynes' 1979 article (Ref 45:15-118).

Consider a set of N possible events with the probability of occurrence of each event known. The probabilities p_i, i=1,2,...N are known, but no further information is available concerning which event will occur. Then $S(p_1 p_2 \ldots p_N)$, defined by Shannon as $S(p_1, p_2 \ldots p_N) = -\sum_{i=1}^{N} p_i \ln p_i$, is a measure of how much "choice" is involved in the selection of a single event or how "uncertain" one is of the outcome of event selection. As an indication of this uncertainty measure, consider N equally likely events, that is $p_1=p_2=\ldots=p_N=1/N$. One's uncertainty as to which event will occur increases as N increases, i.e., as the number of possible events increases. In a similar manner, the value of the uncertainty measure or system entropy, $S(p_1, p_2, \ldots p_N) = -\sum_{i=1}^{N} p_i \ln p_i = \ln N$, increases as N increases. Consequently, to paraphrase Shannon/Jaynes/

Tribus, if one wishes to construct a minimally prejudiced probability distribution (a distribution which maximizes uncertainty) based on information about that distribution, one must maximize the entropy subject to constraints which are specified by the given information.

Before proceeding with the maximum entropy characterization method, we must extend the entropy measure to include continuous probability density functions. Several of the references discuss the continuous case (Refs 14; 17; 32; 89; 90), but Silviu Guiasu (Ref 33) provides the most satisfactory treatment. Shannon's work tells us that the entropy

$$S(p_1, p_2 \ldots p_N) = - \sum_{i=1}^{N} p_i \ln p_i \qquad (2.1)$$

provides a measure of uncertainty for the finite, N dimensional probability space where p_i = probability of the i^{th} event; $p_i \geq 0$, $i=1,2,\ldots N$; and $\sum_{i=1}^{N} p_i = 1$. Guiasu considers entropy, in a comparable fashion, as the "amount of information" conveyed by the given distribution. We now consider a continuous probability density, p(x), on a bounded interval [a,b] such that $p(x) \geq 0$ and $\int_a^b p(x)\, dx = 1$. One might erroneously assume that equation (2.1) is logically extended, in the limit, to the integrable case in the form

$$S(p(x)) = - \int_a^b p(x) \ln p(x)\, dx \qquad (2.2)$$

Equation (2.2), in fact, represents the Boltzman H-function from classical thermodynamics which was defined as early as 1896. The H-function measures the disorder of a physical system (Ref 33:14) and inspired Shannon to study, by analogy, the discrete entropy $S(p_1,p_2 \ldots p_N)$. However, equation (2.2) is not the limiting case of equation (2.1). Consider the uniform probability density on [a,b];

$$p(x) = \begin{cases} 1/(b-a) & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases} \qquad (2.3)$$

Then $S(p(x)) = - \int_a^b 1/(b-a) \ln [1/(b-a)] \, dx = \ln(b-a)$. However, $S(p_1,p_2 \ldots p_N) = \ln N$ as previously stated for $p_1=p_2=\ldots p_N = 1/N$, i.e., the discrete uniform equivalent. Clearly, $\lim_{N \to \infty} S(p_1,p_2 \ldots p_N) \neq S(p(x))$. Thus the question remains of how to relate "continuous entropy" to a measure of uncertainty while forcing consistency between the discrete and continuous cases. S. Kullback and R. A. Leibler (Ref 50) provided insight with a measure of information variation.

The Kullback-Leibler information discrimination measure provides a means of comparing or measuring the information that is lost or gained when one probability measure replaces a second probability measure. Kullback (Ref 51) and Guiasu (Ref 33) both offer excellent development of the information discrimination measure which is based on the well known Radon-Nikodym theorem (Ref 35). The development will not be repeated here, but we consider

only definition and relationship to entropy. Consider sample space X and the sigma algebra, L, of measurable sets of elements of X. We define probability measures U_1, U_2 on L to denote probability spaces (X, L, U_i), i=1,2. Probability measures U_1 and U_2 are assumed to be absolutely continuous with respect to each other; that is, for every set E in L if $U_1(E)=0$ then $U_2(E)=0$ or if $U_2(E)=0$ then $U_1(E)=0$. Tnen the "variation of information" when we pass from initial probability measure U_1 to the new probability measure U_2, absolutely continuous with respect to U_1, is the integral

$$I(U_2,U_1) = \int_X \phi(x) \ln \phi(x) \, dU_1(x)$$
$$= \int_X \ln \phi(x) \, dU_2 \qquad (2.4)$$

where $\phi(x) = dU_2(x)/dU_1(x)$ is the Radon-Nikodym derivative. If we now associate cumulative distribution functions $P_1(x)$ and $P_2(x)$ with measures $U_1(x)$ and $U_2(x)$ (Ref 66:261) where $p_1(x)$ and $p_2(x)$ represent respective density functions, we may reduce (2.4) to a measure of information variation between two continuous probability density functions:

$$I(p_2(x),p_1(x)) = \int p_2(x) \ln [p_2(x)/p_1(x)] \, dx \qquad (2.5)$$

The variation of information function, $I(U_2,U_1)$, appears frequently in the literature (Refs 18; 26; 29; 73; 77; 79). Both Guiasu and Kullback offer thorough presentations of the properties of $I(U_2,U_1)$.

Following the thrust of Guiasu's development, we may now relate entropy to a variation of information. We use equation (2.5), let $X=[a,b]$, and let the initial distribution, $p_1(x)$, be the uniform distribution on [a,b]. The uniform density is given in equation (2.3). Then equation (2.5) reduces to

$$I(p_2(x),p_1(x)) = \int_a^b p_2(x)\,[\ln p_2(x) + \ln(b-a)]\,dx$$

$$= \ln(b-a)\int_a^b p_2(x)\,dx + \int_a^b p_2(x)\ln p_2(x)\,dx$$

$$I(p_2(x),p_1(x)) = \ln(b-a) - S(p_2(x))$$

Therefore, as Guiasu states, the continuous entropy $S(p(x))$ may be interpreted (up to an additive constant) as the variation of information in passing from the uniform probability distribution on [a,b] to the new probability measure defined by $p(x)$ on [a,b]. A similar development follows for Shannon entropy in the discrete case. Given $p_i \geq 0$, $i=1,2,\ldots N$ and $\sum_{i=1}^{N} p_i = 1$, then $S(p_1,p_2\ldots p_N) = \ln N - I(p_1, p_2\ldots p_N, q_1, q_2\ldots q_N)$ where $q_i = 1/N$, $i=1,2,\ldots N$. Thus both Boltzman's continuous entropy and Shannon's discrete entropy serve as a measure of the variation of information when we pass from the initial uniform distribution to the corresponding probability density of interest. With this confirmation, we proceed to investigate the maximum entropy concept.

Maximum Entropy Formalism

Tribus (Ref 82) formalizes the maximum entropy concept for practical application. The goal is to approximate the unknown distribution of random variable X with a "minimally prejudiced" probability density function, p(x), based on known or calculable information about the unknown distribution. The basic underlying principle, as originally put forward by E. T. Jaynes, is here repeated; "The minimally prejudiced probability distribution is that which maximizes the entropy subject to constraints supplied by the given information" (Ref 82:120). An adaptation of the Jaynes/Tribus formalism is presented in the following steps:

1. Define the density structure, i.e., discrete or continuous. If a discrete density is involved then this step includes definition of possible outcomes; the entropy formalism assumes that the possible outcomes are known and that we desire an approximation to the probability of each outcome. For a continuous density, we require definition of the set X of possible outcomes, i.e., the interval of integration [a,b] in equation (2.2). Notice that [a,b] may be infinite. In practical application, the interval may be determined (or approximated) via random sample from the unknown distribution where x_i, i=1,2,...N, is the random sample and [a,b]=[min x_i, max x_i]. Notice also that the random variable X may be vector valued or multivariate.

2. Constrain the density approximation, $p(x)$, to satisfy the given information. We here consider continuous densities although a parallel development holds for the discrete case. The given information is assumed to consist of expected values (or average value approximations) of functions $g_j(x)$, $j=1,2,...K$, of the random variable X. In the work that follows, we call these functions "information functions" to indicate their significance in providing information about the unknown distribution. The formalism assumes that the information functions, $g_j(x)$, and the expected values, $\langle g_j(x) \rangle$, are known or specified. Thus, constraints on $p(x)$ take the following form:

$$\langle g_j(x) \rangle = \int_a^b g_j(x) p(x)\, dx, \quad j=1,2,...K$$

The selection of specific $g_j(x)$ and calculation of $\langle g_j(x) \rangle$, $j=1,2,...K$, in fact determine the form of the resulting approximation, $p(x)$. Consequently, much of our effort pertains to an intelligent selection of information functions. We notice that the formalism (or an adaptation of the formalism) may still be applied if the available information takes a form other than expected (average) values of specified functions. One such example is discussed in the applications chapter, Chapter XI.

3. Finally, maximize the entropy subject to the given constraints.

Application of the above formalism, for the bounded, integrable case, produces a constrained maximization problem:

$$\max S(p(x)) = \max \left(- \int_a^b p(x) \ln p(x)\, dx\right)$$

$$\text{subject to;} \quad \int_a^b p(x)\, dx = 1, \tag{2.6}$$

$$\int_a^b g_j(x)\, p(x)\, dx = \langle g_j(x) \rangle, \quad j=1,2,\ldots K$$

with $p(x)$ unknown, the value of $\langle g_j(x) \rangle$ and the form of $g_j(x)$, $j=1,2,\ldots K$, given. Tribus solves this problem in the discrete case using the Lagrange method of undetermined coefficients. The Lagrange method also applies to the integrable case (Refs 27; 56). The analytical form of the solution is

$$p(x) = \exp[-\lambda_0 - \lambda_1 g_1(x) - \ldots \lambda_K g_K(x)] \tag{2.7}$$

where λ_j, $j=1,2,\ldots K$, are the Lagrange multipliers. Equation (2.7) represents the form of the minimally prejudiced, maximum entropy distribution. We show in Chapter IV, Theorem 4.1, that equation (2.7) is the [illegible] form for the entropy density. In a similar sense, equation (2.7) represents a family of distributions where the specific distribution of interest is selected through appropriate selection of the Lagrange multiplier vector $\Lambda = [\lambda_0, \lambda_1, \ldots \lambda_K]^T$. The Lagrange multipliers are unknown at this point, but the form of each information function, $g_j(x)$,

is predefined by the analyst. The $g_j(x)$, $j=1,2,...K$, may take any form such that the expected values are known or calculable.

Entropy Applications in the Literature

Several authors have discussed application of the maximum entropy formalism as indicated in the list of references. Applications to spectral and time series analysis, economic problems, decision theory and pattern recognition problems, and physics and thermodynamics problems are examples of the available literature (Refs 7; 9; 12; 58; 74; 84). D. V. Gokhale (Ref 31) provides excellent supportive discussion for entropy characterization based on known expected values of certain functions of a random variable. However, most of the available literature concentrates on application in the discrete density case. The discrete entropy maximization problem (the continuous case is represented in equations (2.6)) represents a set of simultaneous linear equations. Solution of the k+1 constraint equations for the appropriate λ is thus somewhat simpler in the discrete case as compared to the nonlinear continuous problem. Agmon (Ref 3) presents an algorithm for computer solution of the discrete problem. Gokhale (Ref 30) presents a second approach to the discrete case and the list of references provides several applications to specific discrete problems (Refs 22; 52; 83; 86; 87).

Application of the entropy method to continuous density approximation results in a system of nonlinear constraint equations, equations (2.6), that must be solved simultaneously to find Λ. Although solution of the constraint equations is, in general, quite difficult, a few specific continuous distributions are well known. For example, if all that is known about the distribution of random variable X on $(-\infty,\infty)$ is the values of $\langle x \rangle$ and $\langle x^2 \rangle$, then the resulting maximum entropy distribution is the normal distribution. To see this, consider the known form of the entropy density:

$$p(x) = \exp[-\lambda_0 - \lambda_1 x - \lambda_2 x^2]$$

The normal density function f(x) with mean μ and variance σ^2 follows:

$$\begin{aligned} f(x) &= (2\pi\sigma^2)^{-\frac{1}{2}} \exp[-(x-\mu)^2/(2\sigma^2)] \\ &= (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left[-(2\sigma^2)^{-1}x^2 + (\mu/\sigma^2)x - (\mu^2/2\sigma^2)\right] \\ &= \exp[(C-(\mu^2/2\sigma^2)) + (\mu/\sigma^2)x - (2\sigma^2)^{-1}x^2] \end{aligned}$$

where $C = \ln(1/\sqrt{2\pi\sigma^2})$. Thus

$$\lambda_0 = (\mu^2/2\sigma^2) - C, \quad \lambda_1 = -\mu/\sigma^2, \quad \lambda_2 = 1/2\sigma^2$$

will produce a normal distribution with mean μ and variance σ^2. Theoretical development by Guiasu and Tribus (Refs 33:299; 82:131) point to the above results; experimentation, with a method of approximation that is detailed

in this dissertation, substantiates the expected results. Other examples include the uniform distribution if no information (except [a,b]) is known, and the exponential distribution if only <x> (on $[0,\infty)$) is known. Much of the literature on continuous entropy centers on application of these known entropy forms. For example, Dudewicz and van der Meulen (Ref 24) utilize known entropy forms to develop the concept of "entropy-distinguishability" and entropy-based tests of hypothesis. Other examples may be found in the list of references.

Two separate approaches to continuous density approximation based on entropy concepts are found in the literature. The primary difference between the approaches is the choice of what we have called "information functions," i.e., $g_j(x)$, $j=1,2,\ldots K$. B. R. Crain (Refs 15; 16; 17) selects the Legendre polynomials as information functions and restricts $p(x)$ to be an element of $L^2[-1,1]$, i.e., square integrable functions over $[-1,1]$. The Legendre polynomials form a complete orthonormal basis of $L^2[-1,1]$ which leads to theoretically sound convergence properties for selected approximation densities. Wilson and Wragg (Ref 89) and Wragg and Dowson (Ref 90) provide good theoretical development for a similar approach on $[0,\infty)$ where the information functions are taken as moments. Practical application of either method is restricted by the need to answer the following questions:

1. How does one determine the number of moments or Legendre polynomials that are needed for a particular approximation?

2. How does one find the necessary Lagrange multiplier vector Λ?

3. How useful is the resulting algebraic/analytic expression for p(x), i.e., as the number of required information functions increases, does p(x) become computationally and conceptually cumbersome?

Collins and Wragg (Ref 14) took a logical step toward reducing the computational difficulty of an approximation based on moments. They reduced the continuous problem to a discrete, and thus linear, problem by resorting to frequency histograms. The histogram method is computationally appealing but does not provide an algebraic representation of p(x). Young and Coraluppi (Ref 91) present an interesting approach to a reduced problem. They present an algorithm for the approximation of a probability density function by a mixture of normal density functions with unknown means and variances. Their approach is also based on minimizing an information criterion.

The following chapters present a practical method for approximating an unknown probability density function based on information in the form of average or expected values of information functions, $g_j(x)$, $j=1,2,\ldots K$. The heart of the method is intelligent selection of the

information functions from a large set of potential information functions. The method builds on the maximum entropy formalism as outlined in this chapter. The theoretical development includes theorems on the form and uniqueness of the approximating density for a given set of information (Ref Chapter IV). The resulting method is computationally feasible and efficient, and numerical techniques for implementation are demonstrated.

Chapter III. Entropy Approximation

Introduction

The distribution approximation problem, as introduced in Chapter II, is refined in the first section of this chapter. The second section presents the general approximation procedure which results from application of maximum entropy concepts. Subsequent chapters explore the detailed steps of the general procedure and specific applications.

Problem Refinement

The problem of interest concerns approximation of the unknown distribution of random variable X based on information that is provided, or information that one may obtain, concerning the unknown distribution. We are particularly interested in approximating the output distributions of computer simulations. The goal of this research is to produce an approximation method that is theoretically sound, suitable for practical application, and specifically adaptable to computer implementation. With the goal in mind, we may apply the entropy formalism of Chapter II.

Our previous adaptation of the entropy formalism includes three steps: define the density structure, constrain the density to given information, and select the specific density that maximizes the entropy. First we define the density structure. This paper will restrict

investigation to continuous densities on the bounded interval [a,b]. The investigation centers on characterization of univariate distributions where we assume that the distribution is generated by an underlying, unknown density. We seek a representation or approximation for the unknown density. While we concentrate on univariate distributions, notice that nothing in our conceptual or theoretical development precludes extension of the method to multivariate distributions, i.e., where random variable X is vector valued. The density structure defined, we now proceed with the entropy formalism.

Approximation Procedure

As previously discussed, the unknown density function for random variable X will be approximated by a maximum entropy function of the form

$$p(x) = \exp\left[-\lambda_0-\lambda_1 g_1(x)-\ldots\lambda_K g_K(x)\right] \qquad (3.1)$$

where the $g_i(x)$, i=1,2,...K are "information functions," and the λ_i, i=0,1,2,...K are Lagrange multipliers. The key to providing an accurate representation of an unknown density, thus the key to our approximation procedure, rests in the ability to select the proper information functions and the appropriate Lagrange multipliers. The approximation procedure is thus composed of three basic steps: select the appropriate information functions; calculate the

expected or average values, $\langle g_i(x) \rangle$, i=1,2,...K; and solve the entropy maximization problem for the Lagrange multipliers. We now consider the steps in more detail.

Information Function Selection. The form of the provided or calculated information, i.e., the forms of the information functions, and the amount of information, i.e., the number of information functions, determine the form of the resulting entropy density as shown in equation 3.1. Clearly, specifying the wrong information functions or too little information may lead to an unacceptable approximation. Moments (Ref 90) and orthogonal polynomials (Ref 15) are examples of possible information functions. Our procedure allows great flexibility in definition of information functions.

A two-phased approach is used in specifying the information functions that best approximate a particular continuous, univariate density on [a,b]. The first phase includes specifying a large, general class of "potential" information functions that have particular conceptual or theoretic value. For example, all moments of a random variable provide an extensive amount of information and would comprise a feasible potential set. For reasons indicated in Chapter II and expanded in Chapter V, moments do not provide a practical set of functions. A more useful potential set is discussed in Chapter V. The potential set should

be large to allow consideration of a wide range of useful functions. Use of the entire potential set to specify the entropy density, equation 3.1, would lead to a numerically intractable problem in solving for the Lagrange multiplier (just as using all moments) and a conceptually dissatisfying form for the approximation, p(x). In fact, much of the information may be redundant or unneeded when approximating a particular density. Thus, in phase two, we seek the minimum subset of the large potential set that will acceptably approximate the density of interest. The minimum subset will be called the "active set" of information functions. Thus, phase one is definition of a large class of potential functions that will serve in a wide variety of characterization or approximation problems, while phase two is selection of the active set of information functions that pertain to a specific approximation or characterization problem.

A point of clarification is needed. We will frequently interchange the terms characterization and approximation when referring to the entropy procedure. As we will see in Chapter V and subsequent chapters, the entropy procedure will exactly characterize the unknown density (given computational accuracy) if the potential information function set contains the correct functions. Chapter II provided such an example for the normal distribution with functions x and x^2. If the correct functions are not present,

then the procedure provides an approximation to the unknown distribution. Thus, assuming a broad potential set, interchange of the two words is permissible.

Generation of Expected or Average Values. Our entropy approximation procedure requires that the information be given in terms of expected values of the information functions, $\langle g_i(x) \rangle$, i=1,2,...K, or average value approximations to the expected values. The method used to obtain the $\langle g_i(x) \rangle$, i=1,2,...K, is transparent to the characterization procedure; in fact, alternate methods exist. For example, the analyst may possess information about the unknown distribution which was accumulated through years of experience or repeated trials. If this information is available in the form of average values, then the entropy method may be applied directly. For a more standard approach, we assume that a random sample of size N is available for random variable X, i.e., x_j, j=1,2,...N. The expected values are then approximated by average values:

$$\langle g_i(x) \rangle \approx \sum_{j=1}^{N} g_i(x_j)/N, \quad i=1,2,...K \qquad (3.2)$$

The accuracy of the approximation in equation 3.2 is dependent on the sample size N. A third method which is heavily used in this research is numerical quadrature.

Numerical quadrature, or numerical integration, provides an effective means of computer integration. Quadrature plays a key role in application of our characterization procedure to computer simulation, is a required tool for the numerical scheme which we employ to find the Lagrange multipliers, and can be used to calculate expected values. A detailed discussion of one quadrature form, Gauss-Legendre quadrature, may be found in Appendix A. The general quadrature form follows:

$$\int_a^b q(x)\,dx \approx \frac{b-a}{2} \sum_{j=1}^{m} W_j\, q(x_j) \tag{3.3}$$

where $a \leq x_1 \leq x_2 \leq \ldots \leq x_m < b$ and the W_j and x_j are defined in Appendix A. Now consider the expected value equation:

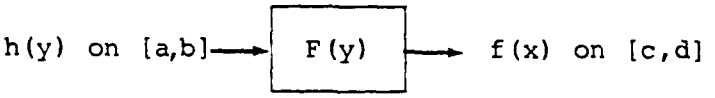
$$\langle g_i(x) \rangle = \int_a^b g_i(x)\, f(x)\,dx \tag{3.4}$$

where $f(x)$ represents the unknown density that we wish to approximate. If the values of the unknown density can be approximated at the points x_j, $j=1,2,\ldots m$, then

$$\langle g_i(x) \rangle \approx \frac{b-a}{2} \sum_{j=1}^{m} W_j\, g_i(x_j)\, f(x_j)$$

The values $f(x_j)$, $j=1,2,\ldots m$, might be reasonably approximated by a frequency table or even numerical differentiation of a sample cumulative distribution.

A very important application of our characterization method is to computer simulation. The key role of quadrature in this application is introduced at this point for consistency and is expanded in Chapter IX. Consider the simplified simulation model:



where h(y) is the known probability density function of input random variable Y on [a,b], F(y) is a mathematical transformation representing the simulation, and f(x) is the unknown probability density function for random variable X that we wish to approximate. We apply basic transformation of variables techniques (Ref 39:127) to equation 3.4 to obtain the following:

$$\langle g_i(x) \rangle = \int_c^d g_i(x)\, f(x)\, dx = \int_a^b g_i(F(y))\, h(y)\, dy;$$

and applying equation (3.3)

$$\langle g_i(x) \rangle \approx \frac{b-a}{2} \sum_{j=1}^{m} W_j\, g_i(F(y_j))\, h(y_j) \qquad (3.5)$$

$$i=1,2,\ldots K.$$

Thus we may calculate the expected values of information functions for a computer simulation by sampling from the

simulation at m predefined points. The benefits of this result are pursued in Chapter IX.

Lagrange Multipliers. Once the active set of information functions has been selected and the expected values have been calculated or approximated, the constraint equations must be solved to find the K+1 Lagrange multipliers, λ_i, i=0,1,...K, where K is the number of functions in the active set. The simultaneous solution of K+1 nonlinear equations is a difficult task analytically and numerically. Acton, Fox, Luenberger, and Saaty and Bram (Refs 2; 27; 57; 67) describe various numerical approaches. The method of choice in this paper is the Newton method. A computer program to solve for the K+1 lambdas, given K expected values and the forms of the information functions, using the Newton method has been implemented. See reference 3 for a similar approach to the discrete density problem. Existence and uniqueness properties, and a numerical scheme for general solution of the constraints are discussed in the next chapter.

Resulting Density Function. The form of the maximum entropy approximation is known: $p(x) = \exp[-\lambda_0-\lambda_1 g_1(x) -...\lambda_K g_K(x)]$. The specific entropy density, p(x), that will approximate or characterize the unknown density, f(x), is selected through application of the above procedural steps. We summarize the procedure. The active information

function set is selected to include the form and number of information functions. Average or expected values of the information functions are generated. Finally, the Lagrange multipliers are calculated via the Newton method, and p(x) is completely defined.

Application of this method has produced excellent approximations in numerous test cases. The specifics of the above procedure, to include test examples and applications, are discussed in the following chapters.

Chapter IV. Solution of the Constraint Equations

Introduction

The theoretical backbone of our entropy characterization method is presented in the first section of this chapter. The entropy approach is based on the solution of a constrained optimization problem. We present the problem and derive Theorem 4.1 which defines the form of the solution density,

$$p(x) = \exp[-\lambda_0,\lambda_1 g_1(x) - \ldots \lambda_k g_k(x)] \qquad (4.1)$$

We then address solution of the constraint equations for the lambda vector, $\Lambda=(\lambda_0,\lambda_1 \ldots \lambda_k)^T$, which equates to selection of a particular density from the family represented in equation 4.1. Two theorems pertaining to uniqueness of solution are presented. Theorem 4.2 shows that, given existence, there is only one solution vector that maximizes the entropy. However, iterative solution of the constraints may lead to a local optimum as shown in Theorem 4.3. The first section concludes with Theorem 4.4 which shows that the average values of our information functions are complete sufficient statistics for selection of a specific p(x). The second section of the chapter presents a numerical scheme to apply the theory. Performance of the scheme and numerical sensitivities are discussed.

Theoretical Development

Form of the Solution. We assume that the forms and number of "active" information functions are known, and that expected values are given; thus, $g_j(x)$ and $\langle g_j(x) \rangle$, $j=1,2,\ldots k$, are known. We wish to find the continuous density, $p(x)$, on $[a,b]$ that will satisfy the given information, i.e., produce the given expected values, while maintaining maximum uncertainty with respect to other, unspecified information. The mathematical statement of this problem is repeated from equations 2.6:

$$\max S(p(x)) = \max \left(-\int_a^b p(x) \ln p(x)\, dx\right)$$

subject to

$$\int_a^b p(x)\, dx = 1 ,$$

$$\int_a^b g_j(x)\, p(x)\, dx = \langle g_j(x) \rangle, \quad j=1,2,\ldots k \qquad (4.2)$$

We assume that the $g_j(x)$, $j=1,2,\ldots k$, are continuous and bounded on $[a,b]$. In terms of a probability space, we consider probability space (X, L, U) where X is the interval $[a,b]$, L is the sigma algebra of Lebesgue measurable sets on X, and the probability measure U on L is defined by the probability density function

$$p(x) \geq 0, \quad a \leq x \leq b, \quad \int_a^b p(x)\, dx = 1 \qquad (4.3)$$

We apply the Lagrange method of undetermined coefficients to equations 4.2 to find the density in 4.3. The

Lagrangian, $L(p(x),\Lambda)$, follows:

$$L(p(x),\Lambda) = S(p(x)) - \lambda_0(\int p(x)\,dx-1) - \sum_{j=1}^{k} \lambda_j(\int g_j(x)p(x)\,dx-\langle g_j(x)\rangle)$$

$$= \int p(x)[\ln(1/p(x))-\lambda_0-\sum_{j=1}^{k}\lambda_j g_j(x)]\,dx + \lambda_0 + \sum_{j=1}^{k}\lambda_j\langle g_j(x)\rangle$$

$$= \int p(x)\{\ln[(1/p(x))\exp(-\lambda_0-\sum_{j=1}^{k}\lambda_j g_j(x))]\}dx + \lambda_0 + \sum_{j=1}^{k}\lambda_j\langle g_j(x)\rangle$$

We apply the knowledge that for all $x>0$

$$\ln(x) < x-1 \quad \text{if} \quad x\neq 1 \quad \text{and}$$

$$\ln(x) = x-1 \quad \text{if} \quad x=1 \quad \text{to get}$$

$$L(p(x),\Lambda) \leq \int p(x)[(1/p(x))\exp(-\lambda_0-\sum_{j=1}^{k}\lambda_j g_j(x)) - 1]\,dx + \lambda_0 + \sum_{j=1}^{k}\lambda_j\langle g_j(x)\rangle$$

Since we want to maximize $L(p(x),\Lambda)$, we seek equality in our last expression which occurs if and only if $p(x) = \exp[-\lambda_0 - \sum_{j=1}^{k}\lambda_j g_j(x)]$. The preceding result is well known and is mentioned in several references without proof for the continuous case (Ref 42; 82; 89). This derivation is given to enhance clarity. The derivation is a generalization of work presented by Guiasu (Ref 33:298-301) and attributed to Kampé de Fériet (Ref 48) and Ingarden and Kossakowski (Ref 41). The Guiasu presentation was concerned with the normal and Poisson distributions. We summarize with a theorem.

Theorem 4.1. Given probability space (X, L, U) where X is the interval [a,b], L is a sigma algebra of Lebesgue measurable sets, and U is defined in terms of p(x) as in equation 4.3, then the density function, p(x), which maximizes the entropy subject to constraints as represented in equations 4.2 is of the form

$$p(x) = \exp[-\lambda_o - \lambda_1 g_1(x) - \ldots - \lambda_k g_k(x)]$$

where the $g_j(x)$ are continuous, bounded functions on X.

Theorem 4.1 provides the form of the entropy characterization density. Given a specific set of expected values, $\langle g_j(x) \rangle$, j=1,2,...k, we solve the k+1 constraints for $\Lambda = (\lambda_0, \lambda_1, \ldots \lambda_k)^T$ to completely determine p(x). We now relate concepts from Tribus (Ref 82), Guiasu (Ref 33), and Kullback (Ref 51) with the special properties of our problem to discuss existence and uniqueness properties.

Existence. The existence of a solution to the constraints is not guaranteed. We may clearly specify a set of expected values which form an inconsistent set of constaints and for which no density exists. For example, consider k=2, $g_1(x)=x$, $g_2(x)=x^2$, $\langle g_1(x) \rangle = 10$, and $\langle g_2(x) \rangle = 99$. As discussed in Chapter II of this paper, the maximum entropy density given only $\langle x \rangle$ and $\langle x^2 \rangle$ is the normal density. Consider the variance of the density we have specified:

$$\text{Var} = \langle (x-\langle x \rangle)^2 \rangle = \langle x^2 \rangle - \langle x \rangle^2 = 99 - 100$$

or $\text{Var} = -1.0$

This example thus asks for a normal density with negative variance which is not possible. Collins and Wragg (Ref 14) state that, in the general case, the precise conditions on the expected values of moments for which a Λ vector will exist, where the Λ vector satisfies the constraints, do not seem to be known. Only for information functions $g_1(x)=x$ $(k=1)$ and $g_1(x)=x$, $g_2(x)=x^2$ $(k=2)$ are conditions known in any completeness. Wragg and Dowson (Ref 90), Widder (Ref 88), and Ahiezer and Krein (Ref 4) provide extended discussion of conditions for valid moment sequences.

We are, however, concerned with practical application and our problem is somewhat restricted. Our problem centers on approximating an "existing," though unknown, density. Samples from the unknown density are used to approximate the expected values via quadrature or sample averages. Thus, we have a consistent set of constraints provided that the expected value approximations are accurate. Inconsistencies that result from sampling or computational errors may be alleviated by increasing sample size, or increasing the number of quadrature points, and including computational checks to produce more accurate expected value approximations. We thus assume a consistent

set of constraints, i.e., the unknown density, f(x), satisfies the constraints. Assuming an intelligent choice of information functions, as discussed in later chapters, we will produce a p(x), Theorem 4.1, that acceptably approximates f(x). In this manner, the existence problem is conceptually translated to a problem of specifying the correct information functions. For purposes of this paper, we assume existence of a solution vector, $\Lambda = (\lambda_0,\lambda_1,\ldots\lambda_k)^T$, for a given expected (or average) value vector, $\langle G\rangle = (1,\langle g_1(x)\rangle,\langle g_2(x)\rangle,\ldots\langle g_k(x)\rangle)^T$.

Uniqueness. We wish to discuss uniqueness in two respects. First we show that if there exists a second density on [a,b], $\tilde{p}(x)$, where $\tilde{p}(x)$ may take any form such that $\tilde{p}(x)\geq 0$ and $\tilde{p}(x)>0$ almost everywhere (a.e.) on [a,b], $\tilde{p}(x)$ satisfies the constraints, and $\tilde{p}(x)$ maximizes the entropy, then $p(x)=\tilde{p}(x)$ a.e. Secondly, we show that the solution, Λ, which maximizes the entropy is a global solution, i.e., if there exists a $B = (\beta_0,\beta_1,\ldots\beta_k)$ such that

$$\tilde{p}(x) = \exp[-\beta_0-\beta_1 g_1(x) -\ldots\beta_k g_k(x)]$$

where $\tilde{p}(x)$ satisfies the constraints and maximizes the entropy, then $B=\Lambda$. We will combine these results in one theorem. The approach is to assume a second solution, $\tilde{p}(x)$, of any form such that $\tilde{p}(x)\geq 0$ for all x in [a,b] and $\tilde{p}(x)>0$ a.e. on [a,b]. We then consider the entropies of

our two solutions, $F=S(p(x))-S(\tilde{p}(x))$, and show $F \geq 0$ a.e. and $F=0$ if and only if $p(x)=\tilde{p}(x)$ a.e. We then assume the existence of a β vector and obtain our second result. The theorem and proof are motivated in a proof by Tribus of a discrete maximum (Ref 82:123) and a theorem on information discrimination by Kullback (Ref 51:14).

Prior to a statement of our theorem, we discuss the Kullback theorem. The Kullback-Leibler information discrimination measure was introduced in the background chapter of this report (equation 2.4). In keeping with Kullback, we use the probability spaces of Chapter II, (X, L, U_i), $i=1,2$, and define a third probability measure, U. Let U be absolutely continuous with respect to (w.r.t.) U_i and U_i be absolutely continuous w.r.t. U, $i=1,2$; for example, U may be U_1, or U_2, or $(U_1+U_2)/2$. The Radon-Nikodym derivatives are now defined in terms of U as $\Phi_i(x)=dU_i(x)/dU(x)$ where $\Phi_i(x)$, $i=1,2$, are functions, unique up to sets of measure (probability) zero in U, $0<\Phi_i(x)<\infty$ such that $U_i(E)=\int_E \Phi_i(x)\,dU(x)$, $i=1,2$, E an element of L. Using this nomenclature, we rewrite the discrimination measure of Chapter II, $I(U_2,U_1)$, to an equivalent form:

$$I(U_2,U_1) = \int \ln \frac{\Phi_2(x)}{\Phi_1(x)}\, dU_2(x) = \int \Phi_2(x) \ln \frac{\Phi_2(x)}{\Phi_1(x)}\, dU(x) \tag{4.4}$$

Kullback Theorem. $I(U_2,U_1)$ is almost positive definite; that is, $I(U_2,U_1) \geq 0$, with equality if and only if $\Phi_1(x)=\Phi_2(x)$ a.e. w.r.t. U. See Kullback (Ref 51:14) or Guiasu (Ref 33:22) for proof of this theorem.

Armed with the Kullback Theorem, we turn to the question of uniqueness of p(x) and uniqueness of the Lagrange multiplier vector, Λ. We have p(x) as shown in equation 4.1 and k+1 constraints:

$$\int_a^b g_j(x)\, p(x)\, dx = \langle g_j(x) \rangle, \quad j=0,1,\ldots k$$

where $g_0(x) \equiv 1$. For a specified set of expected values, $\langle G \rangle_0 = (1, \langle g_1(x) \rangle, \ldots \langle g_k(x) \rangle)^T_0$, we solve the constraints for $\Lambda_0 = (\lambda_0, \lambda_1 \ldots \lambda_k)^T_0$ to completely determine p(x). From Theorem 4.1 we know that p(x) maximizes the entropy, S(p(x)). Now assume that there exists density $\tilde{p}(x)$ that satisfies the constraints where $\tilde{p}(x)$ may take "any form" subject to two conditions: $\tilde{p}(x) \geq 0$, x in [a,b] and $\tilde{p}(x) > 0$ a.e. on [a,b]. The second condition is needed to insure that the probability measures associated with p(x) and $\tilde{p}(x)$ are absolutely continuous w.r.t. each other. We wish to determine if $\tilde{p}(x)$ is also of maximum entropy. We may represent our state of knowledge with two sets of equations.

$$S = -\int_a^b p(x) \ln p(x)\, dx, \qquad \tilde{S} = -\int_a^b \tilde{p}(x) \ln \tilde{p}(x)\, dx$$

with

$$\int_a^b g_j(x)\, p(x)\, dx = \langle g_j(x) \rangle, \qquad j=0,1,\ldots k$$

and

$$p(x) = \exp\left[-\sum_{j=0}^{k} g_j(x) \lambda_j\right]$$

with

$$\int_a^b g_j(x)\, \tilde{p}(x)\, dx = \langle g_j(x) \rangle, \qquad j=0,1,\ldots k$$

and

$$\tilde{p}(x) \geq 0,\ x \in [a,b],$$
$$p(x) > 0,\ \text{a.e. on } [a,b].$$

All integrals in the following derivation are over the interval [a,b] although the limits of integration will not be shown. Consider,

$$F = S - \tilde{S} = \int \tilde{p}(x) \ln \tilde{p}(x)\, dx - \int p(x) \ln p(x)\, dx$$

Now add and subtract $\int \tilde{p}(x) \ln p(x)\, dx$,

$$F = \int \tilde{p}(x) \ln (\tilde{p}(x)/p(x))\, dx + \int (\tilde{p}(x) - p(x)) \ln p(x)\, dx$$

We substitute for the known form of p(x) in the last integral:

$$F = \int \tilde{p}(x) \ln (\tilde{p}(x)/p(x))\, dx + \int (\tilde{p}(x) - p(x)) \sum_{j=0}^{k} \lambda_j g_j(x)\, dx$$

$$= \int \tilde{p}(x) \ln (\tilde{p}(x)/p(x))\, dx + \sum_{j=0}^{k} \lambda_j \left\{ \int g_j(x) \tilde{p}(x)\, dx - \int g_j(x)\, p(x)\, dx \right\}$$

Clearly, the last (k+1) terms cancel due to the requirement of constraint satisfaction and thus,

$$F = S - \tilde{S} = \int \tilde{p}(x) \ln (\tilde{p}(x)/p(x))\, dx \qquad (4.5)$$

Since $p(x)$ and $\tilde{p}(x)$ are probability density functions on $[a,b]$, we may define,

$$P(x) = \text{prob } (X \leq x) = \int_a^x p(y)\, dy \text{ and}$$

$$\tilde{P}(x) = \text{prob } (X \leq x) = \int_a^x \tilde{p}(y)\, dy, \text{ with}$$

$$P(x) = \tilde{P}(x) = 0 \text{ if } x<a, \text{ and}$$

$$P(x) = \tilde{P}(x) = 1 \text{ if } x>b,$$

as probability measures on $[a,b]$. We know $p(x)$ and $\tilde{p}(x)>0$ for all x in $[a,b]$, except for a set of measure zero, and by definition $p(x)=\tilde{p}(x)=0$ for all x not in $[a,b]$. Thus $P(x)$ is absolutely continuous w.r.t. $\tilde{P}(x)$ and vice versa.

We are now in a position to apply the Kullback Theorem. With the Kullback nomenclature, we let

$$U_1(x) = U(x) = P(x);\ U_2(x) = \tilde{P}(x);$$

$$\Phi_1(x) = dP(x)/dP(x);\ \Phi_2(x) = d\tilde{P}(x)/dP(x);$$

$$I(U_2,U_1) = \int \Phi_2(x) \ln (\Phi_2(x)/\Phi_1(x))\, dU(x)$$

$$= \int (d\tilde{P}(x)/dP(x)) \ln (d\tilde{P}(x)/dP(x))\, dP(x)$$

$$I(U_2,U_1) = \int \tilde{p}(x) \ln (\tilde{p}(x)/p(x))\, dx, \qquad (4.6)$$

We equate equations 4.5 and 4.6 to obtain

$$F = S - \tilde{S} = I(U_2,U_1).$$

Thus $S-\tilde{S}\geq 0$ with $S-\tilde{S}=0$ if and only if $\phi_1(x)=\phi_2(x)$ a.e. w.r.t. U, or $dP(x)/dP(x)=d\tilde{P}(x)/dP(x)$ a.e. w.r.t. P(x) which implies that $\tilde{p}(x)=p(x)$ a.e. w.r.t. P(x). Thus we obtain the important result that either p(x) is the only form of solution that maximizes the entropy or any other solution, $\tilde{p}(x)$, must satisfy $p(x)=\tilde{p}(x)$ a.e.

We take this development one step further by assuming the existence of a Lagrange multiplier vector $\beta\neq\Lambda_0$ such that $\tilde{p}(x)=\exp[-\sum_{j=0}^{k}\beta_j g_j(x)]$, $\tilde{p}(x)$ satisfies the constraints, and $\tilde{p}(x)$ has maximum entropy. Thus $S-\tilde{S}=0$ and $p(x)=\tilde{p}(x)$ a.e. which implies

$$\ln p(x) = \ln \tilde{p}(x) \text{ a.e., and}$$

$$-\sum_{j=0}^{k}\lambda_j g_j(x) = -\sum_{j=0}^{k}\beta_j g_j(x) \text{ a.e., or}$$

$$\sum_{j=0}^{k}(\beta_j-\lambda_j)\, g_j(x) = 0 \text{ a.e.}$$

If the $g_j(x)$ are linearly independent functions, then the only linear combination of $g_j(x)$ that equals zero a.e. is if $(\beta_j-\lambda_j)=0$ for all j. Thus $\beta_j=\lambda_j$ for j=0,1,...k. We summarize the above developments in a theorem.

Theorem 4.2. Let there exist $\Lambda_0=(\lambda_0,\lambda_1,\ldots\lambda_k)^T_0$ with Λ_0 an element of R^{k+1} such that $p(x)=\exp[-\sum_{j=0}^{k}\lambda_j g_j(x)]$,

$g_o(x) \equiv 1$, and $p(x)$ satisfies the constraints $\int_a^b g_j(x)p(x)\,dx = \langle g_j(x) \rangle$, $j=0,1,\ldots k$, with $S(p(x)) = -\int_a^b p(x) \ln p(x)\,dx$. If there exists a function $\tilde{p}(x)$ such that $\tilde{p}(x) \geq 0$ for all x in $[a,b]$ and $\tilde{p}(x) > 0$ a.e. on $[a,b]$, then $S(p(x)) \geq S(\tilde{p}(x))$ and $S(p(x)) = S(\tilde{p}(x))$ if and only if $p(x) = \tilde{p}(x)$ a.e. on [a.b.] If there exists a β in R^{k+1} such that $\tilde{p}(x) = \exp[-\sum_{j=0}^{k} \beta_j g_j(x)]$, with linearly independent $g_j(x)$, and $\tilde{p}(x)$ satisfies the constraints, then $S(p(x)) = S(\tilde{p}(x))$ if and only if $\lambda_j = \beta_j$, $j=0,1,\ldots k$.

We have established the form and uniqueness of solution for our optimization problem given that a solution exists. We now directly explore the constraint equations and the possibility of "local" optimum solutions. Only one $\Lambda = \Lambda_0$ exists for which $p(x)$ satisfies the constraints and $S(p(x))$ is maximum; however, there may exist a $\Lambda = \beta$ with corresponding $\tilde{p}(x)$ which satisfies the constraints, and where $S(\tilde{p}(x)) < S(p(x))$. If there exists a neighborhood of $\tilde{p}(x)$, where $p(x)$ is not an element of the neighborhood, such that $S(\tilde{p}(x)) \geq S(q(x))$ for all $q(x)$ in the neighborhood then $\tilde{p}(x)$ is a local optimum. Local optimums are of concern because the k+1 nonlinear constraint equations must be solved with iterative numerical techniques, and such techniques may converge to local optimums.

Given that Λ_0 is a solution vector for the nonlinear system of equations $F(\Lambda) = \langle G \rangle$ where

$$F(\Lambda) = \begin{bmatrix} \int g_0(x)\, p(x)\, dx \\ \vdots \\ \int g_k(x)\, p(x)\, dx \end{bmatrix} = \begin{bmatrix} f_0(\Lambda) \\ \\ f_k(\Lambda) \end{bmatrix} = \begin{bmatrix} \langle g_0(x) \rangle \\ \vdots \\ \langle g_k(x) \rangle \end{bmatrix} = \langle G \rangle,$$

$p(x) = \exp[-\lambda_0 g_0(x) - \lambda_1 g_1(x) - \ldots \lambda_k g_k(x)]$, $g_0(x) \equiv 1$, and all integrals are over the interval [a,b], then the following theorem addresses solution uniqueness.

Theorem 4.3. Let $g_0(x)$, $g_1(x), \ldots g_k(x)$, finite k, be continuous functions in $L^2[a,b]$ and $g_i(x)\, g_j(x)$ be in $L^2[a,b]$ for i,j=1,...k, $g_0(x) \equiv 1$. If Λ_0 is a solution of $F(\Lambda) = \langle G \rangle$ for a specific $\langle G \rangle_0$, and if the $g_j(x)$, j=0,1,...k, are linearly independent functions, then there exists a *neighborhood, W, of* Λ_0 *where* Λ_0 is the unique solution of $F(\Lambda) = \langle G \rangle_0$.

Proof. Consider $F(\Lambda)$ to be a function from some subset of R^{k+1} to R^{k+1}. By the fundamental Inverse Function Theorem (Ref 78:354), if $F(\Lambda)$ is continuously differentiable in some neighborhood of Λ_0 and if the linear transformation $F'(\Lambda_0)$ is invertible (nonsingular), then there exists neighborhoods W and V of Λ_0 and $F(\Lambda_0)$, respectively, such that F: $W \rightarrow V$ is a one-to-one, onto mapping. Thus, given solution vector Λ_0, for all $\beta \varepsilon W$, $F(\beta) = \langle G \rangle_0$ if and only if $\beta \equiv \Lambda_0$, or Λ_0 is the unique solution vector in neighborhood W. We will show that F is continuously differentiable in some neighborhood of Λ_0 and that $F'(\Lambda_0)$

is nonsingular in that neighborhood, and the proof will be complete.

1. Continuous differentiability. Continuous first partial derivatives are necessary and sufficient for continuous differentiability. Thus we consider

$$J = \begin{bmatrix} \partial f_0(\Lambda)/\partial\lambda_0, & \partial f_0(\Lambda)/\partial\lambda_1, & \ldots & \partial f_0(\Lambda)/\partial\lambda_k \\ & \vdots & & \\ \partial f_k(\Lambda)/\partial\lambda_0, & \partial f_k(\Lambda)/\partial\lambda_1, & \ldots & \partial f_k(\Lambda)/\partial\lambda_k \end{bmatrix}$$

where $\partial f_i(\Lambda)/\partial\lambda_j = \int g_j(x) g_i(x) p(x) dx$; $i,j=0,1,\ldots k$. Clearly $p(x)$ is a composite of the continuous $g_j(x)$ and is continuous. Integration is a continuous operation and the products of continuous functions are continuous. Thus, each element of J is continuous and $F(\Lambda)$ is continuously differentiable.

2. Nonsingularity of $F'(\Lambda)=J$. J is nonsingular at Λ_0 if the determinant of J, $|J|$, at Λ_0 is not zero. We claim that $|J|\neq 0$ if and only if the $g_j(x)$, $j=0,1,\ldots k$ are linearly independent. We prove the contrapositive of this claim, i.e., $|J|=0$ if and only if the $g_j(x)$ are linearly dependent.

(a) Assume the $g_j(x)$, $j=0,1,\ldots k$ are linearly dependent and show $|J|=0$. Based on this assumption, there exists constants, α_i, not all zero, such that $\sum_{j=0}^{k} \alpha_j g_j(x)=0$ for all x in $[a,b]$. We may write one $g_j(x)$ as a linear

combination of the others and rearrange the indexing such that

$$g_k(x) = -\sum_{j=0}^{k-1} (\alpha_j/\alpha_k)\, g_j(x) = -\sum_{j=0}^{k-1} \beta_j\, g_j(x)$$

Now consider the m^{th} row of J, i.e.,

$$[-\int g_m(x)\, g_0(x) p(x)\, dx, \ldots, -\int g_m(x)\, g_k(x) p(x)\, dx].$$

We substitute for the k^{th} element of this m^{th} row (i.e., the k^{th} column entry):

$$-\int g_m(x)\, g_k(x)\, p(x)\, dx = -\int g_m(x) \left(-\sum_{j=0}^{k-1} \beta_j g_j(x)\right) p(x)\, dx$$

$$= \sum_{j=0}^{k-1} \beta_j \int g_m(x)\, g_j(x)\, p(x)\, dx$$

The last summation equates to a linear combination of the other (k-1) columns. This procedure holds for all rows. Thus, the k^{th} column is written as a linear combination of the first (k-1) columns and $|J|=0$.

(b) Assume $|J|=0$ and show that the $g_j(x)$ must be linearly dependent. Since $|J|=0$, then the columns (or rows) of J are linearly dependent. Thus we have constants, α_j, not all zero, such that $-\sum_{j=0}^{k} \alpha_j \int g_m(x) g_j p(x)\, dx=0$. It follows that $-\int g_m(x) \left(\sum_{j=0}^{k} \alpha_j g_j(x)\right) p(x)\, dx = 0$ for all m (i.e., for all rows) and hence

$$-\sum_{m=0}^{k} \alpha_m \int g_m(x) \left(\sum_{j=0}^{k} \alpha_j g_j(x) \right) p(x)\, dx = 0, \text{ or}$$

$$-\int \left(\sum_{m=0}^{k} \alpha_m g_m(x) \right) \left(\sum_{j=0}^{k} \alpha_j g_j(x) \right) p(x)\, dx = 0 \text{ and}$$

$$-\int \left(\sum_{j=0}^{k} \alpha_j g_j(x) \right)^2 p(x)\, dx = 0.$$

We know that $p(x)=\exp[-\lambda_0 g_0(x)-\lambda_1 g_1(x)-\ldots-\lambda_k g_k(\lambda)]$ and thus $p(x) \neq 0$ for all x in [a,b] where $a \neq -\infty$. Thus $\left(\sum_{j=0}^{k} \alpha_j g_j(x) \right)^2 = 0$ a.e. for the last equation to hold or $\sum_{j=0}^{k} \alpha_j g_j(x) = 0$ a.e. which is a statement of linear dependence of the $g_j(x)$. We remember that two functions in $L^2[a,b]$ are considered equivalent if they are equal a.e. (Ref 66:112). Thus, $|J| \neq 0$ implies linear dependence of the $g_j(x)$.

We have shown that J is nonsingular and the conditions for application of the Inverse Function Theorem are satisfied. The proof of Theorem 4.3 is complete.

Sufficient Statistics. Our final theorem concerns the concepts of complete, sufficient statistics for solution of the λ vector. The theorem shows us that the average values of our information functions (which approximate expected values in our work) contain all the information of the random sample which was used to generate the

values, i.e., information is not lost. Further, the average values contain sufficient information for estimating vector Λ. The theorem is a special case of a theorem presented in Hogg and Craig (Ref 39:232) for the "regular exponential class" of probability densities. We restate the Hogg and Craig results, in our terminology for our special case, as Theorem 4.4.

Theorem 4.4. Let the entropy density be of the form $p(x)=\exp[-\sum_{i=0}^{k} \lambda_i g_i(x)]$ for x in [a,b] and $p(x)=0$ for all other x with $g_0(x)=1$ and linearly independent $g_i(x)$, $i=0,1,\ldots k$. Given a random sample $(x_1,x_2,\ldots x_N)$ with $N>k$, then the functions

$$Y_i = \sum_{j=1}^{N} g_i(x_j)/N, \quad i=0,1,\ldots k,$$

are complete sufficient joint statistics for determining the vector $\Lambda=(\lambda_o,\lambda_1,\ldots\lambda_k)^T$.

Section Summary. In this section we have derived the form of the density family which maximizes the entropy while satisfying given constraints. We have shown that only one density will maximize the entropy, if a solution density exists, although numerical solution of the constraints may lead to a local maximum. We related existence to the correct selection of information functions. Finally, we have shown that the average values of the

information functions provide all the information needed to select the one entropy density. That is, the average values comprise all available information that can be produced by a random sample $(x_1, x_2, \ldots x_N)$ of the unknown density. We now direct our attention to application of the above theory in terms of numerical solution of the constraints.

Numerical Solution Scheme

Key to a general implementation of the entropy approximation procedure is the ability to find the correct $\Lambda = (\lambda_0, \lambda_1, \ldots \lambda_k)$ for a given set of expected values, i.e., solution of the constraints. We restate the problem: find $\Lambda = (\lambda_0, \lambda_1, \ldots \lambda_k)^T$ such that

$$\begin{aligned} f_0(\Lambda) &= \int p(x)\,dx \qquad -1.0 \quad = 0 \\ f_1(\Lambda) &= \int g_1(x)\,p(x)\,dx \quad -\langle g_1(x)\rangle \quad = 0 \\ &\vdots \\ f_k(\Lambda) &= \int g_k(x)\,p(x)\,dx \quad -\langle g_k(x)\rangle = 0 \end{aligned} \tag{4.8}$$

where $p(x) = \exp[-\lambda_0 - \lambda_1 g_1(x) - \ldots - \lambda_k g_k(x)]$ with $\langle g_i(x)\rangle$ and the form of $g_i(x)$ known for $i=1,2,\ldots k$. The $(k+1)$ constraints are nonlinear and, except for a few restricted cases, cannot be solved directly for the Λ vector. As previously mentioned, several authors discuss iterative numerical schemes for simultaneous solution of a system of nonlinear equations (Refs 2; 27; 56; 67). For our

approach, we write equations 4.8, in vector notation, as a fixed point problem, $F(\Lambda)=0$ where $F(\Lambda)=[f_0(\Lambda),f_1(\Lambda),\ldots f_k(\Lambda)]^T$. We have implemented a computer program which successfully applies the Newton-Raphson successive approximation procedure to the above fixed point problem.

The Newton method is based on iterative solution of the following equation:

$$J * (\Lambda_n-\Lambda_{n+1}) = F(\Lambda_n) \tag{4.9}$$

where Λ_n is the Lagrange multiplier vector, Λ, for the n^{th} iteration and J is the Jacobian matrix for $F(\Lambda_n)$. An initial guess, Λ_0, is selected and equation 4.9 is solved for Λ_1. The scheme repeats for $\Lambda_2,\Lambda_3,\ldots\Lambda_n,\Lambda_{n+1}$, until the difference $(\Lambda_n-\Lambda_{n+1})$ is less than a predefined value, i.e., until convergence occurs. The actual convergence criteria for our program requires that the final value of each element of the $(\Lambda_n-\Lambda_{n+1})$ vector be less than a predefined epsilon. When convergence is obtained, Λ_n contains the solution values of λ_j, $j=0,1,\ldots k$. The program is written as a subroutine, subroutine ENTROP, which requires the user to specify the following items: the number of active information functions, k; the approximation bounds, [a,b]; a vector to identify the active information functions; and a vector which contains the average or expected values of the active functions. The potential set of information functions is provided as a set of numbered

external functions, i.e., F1,F2,...FM. The potential set is thus easily modified without access to the subroutine. The user identifies the active set for a particular problem by specifying the respective function numbers. The program is currently implemented with twelve potential functions and a maximum of six active functions (plus $g_0(x)=1$). Larger sets can be accommodated with simple program changes. Subroutine ENTROP solves equations 4.9 for vector $(\Lambda_n - \Lambda_{n+1})$ using matrix decomposition and two programs from the International Mathematical and Statistical Libraries (IMSL). All integrations for production of the Jacobian and $F(\Lambda_n)$ values are accomplished using a 32 point Gauss-Legendre quadrature program from a local library. Once convergence is reached, the Λ vector is returned. Subroutine ENTROP has been extensively tested with very positive results.

Convergence and rate of convergence of the Newton method are dependent on the initial guess, Λ_0. Theorems exist which address convergence of iterative schemes in general (Refs 13; 47) and the Newton method explicitly (Ref 67). The theorems usually specify a neighborhood about the solution wherein the scheme will converge if the initial guess is within that neighborhood. Acton (Ref 2), Collatz (Ref 13), and others present examples of divergence due to poor initial guesses. As expected, subroutine ENTROP is sensitive to the initial guess Λ_0, and a poor

initial guess will cause divergence or numerous iterations for convergence. For example, we consider the data presented in Table IV.I. The data pertains to the output of a computer simulation which will be discussed in Chapter IX. We wish to find the Λ vector for the four information functions, F1, F2, F5, F8, where the average values and bounds, [a,b], are computed from the data. Previous application of subroutine ENTROP for other combinations of information functions, using the same data, indicated that Λ_A was a feasible starting value for the Newton method. However, the large value of λ_{A0}=374.114 produced a terminal numerical error in ENTROP. A second attempt with initial guess Λ_B, where λ_{B0}=0.0 and other elements of Λ_B were equal to Λ_A, failed to converge in 35 iterations, although a terminal error did not result. A final attempt with Λ_C=(0,0,...0) converged in 20 iterations. Several schemes for intelligent selection of the initial vector, Λ_0, were evaluated throughout the research. The one scheme that converged for every test on [a,b], where a solution existed, was the initial vector of all zeroes. Conceptually, this tells us that the first iteration of the Newton method produces a Λ_1, $\Lambda_1=-J(\Lambda_0)^{-1}F(\Lambda_0)$, where Λ_1 is an element of a convergent neighborhood of the solution; other initial guesses run the risk of missing that neighborhood. (It is not known why this occurs.)

TABLE IV.I

SIMULATION DATA AND CONVERGENCE COMPARISON

Find $\Lambda = (\lambda_0, \lambda_1, \lambda_2, \lambda_5, \lambda_8)$ for

$p(x) = \exp[-\lambda_0 - \lambda_1 F1 - \lambda_2 F2 - \lambda_5 F5 - \lambda_8 F8]$ on

$[a,b] = [5104.12, 8262.58]$.

Symbol	Function	Expected Value
F1	$(x-\bar{x})/s$	.00457
F2	$(x-\bar{x})^2/s^2$	1.0
F5	ln (b-x)	7.4647
F8	$(x-\bar{x})^4/s^4$	3.0168
$\bar{x}$	sample mean	6492.26
s^2	variance	72050.87

Test Comparisons

Initial guess: Λ_A=(374.114, 0.0, -.348, 35.707, .0153)
Result: Terminal numerical error

Initial guess: Λ_B=(0.0, 0.0, -.348, 35.707, .0153)
Result: Failed to converge in 35 iterations

Initial guess: Λ_C=(0., 0., 0., 0., 0.)
Result: Convergence in 20 iterations
Λ = (-24.336, .636, .540, 4.125, .002)

Subroutine ENTROP is implemented for a maximum of six active information functions plus the normalizing function $g_0(x)=1$. The number of active functions is restricted for two reasons. First, the Newton method becomes computationally cumbersome as the number of constraints, i.e., the number of active information functions, increases. The primary numerical difficulty centers on the symmetric Jacobian matrix, J, which was discussed in the proof of Theorem 4.3:

$$J = [\partial f_j(\Lambda)/\partial \lambda_i] \quad j,i=0,1,\ldots k.$$

We demonstrate the potential numerical difficulty by relating the initial research which considered moments about zero as information functions, i.e., $g_i(x)=x^i$. The Jacobian in this case follows:

$$J = \begin{bmatrix} -\int p(x)\,dx & -\int xp(x)\,dx & \cdots & -\int x^k\,p(x)\,dx \\ -\int xp(x)\,dx & -\int x^2p(x)\,dx & \cdots & -\int x^{k+1}p(x)\,dx \\ \vdots & & & \\ -\int x^k p(x)\,dx & -\int x^{k+1}p(x)\,dx & \cdots & -\int x^{2k}p(x)\,dx \end{bmatrix}$$

As k increases, or as the interval of integration, [a,b], increases, the elements in the k^{th} column (or row) will be much larger than elements in the first column (or row). Couple this structure with numerical error and limited machine precision, and we have created a matrix which the

computer interprets to be "singular." Such ill-conditioning is not restricted to moments. The example of Table IV.I with large initial value for λ_0, λ_0=374.114, produced the same difficulty. Hornbeck (Ref 40) discusses ill-conditioning and suggests means of circumventing the effects of an ill-conditioned matrix. Scaling and the use of double precision computation will delay the impact of an ill-conditioned matrix. Ill-conditioning is controlled in the entropy procedure by restricting the number of active information functions and normalizing functions when necessary. For example, functions $g_i(x)=((x-\mu)/\sigma)^i$ or $g_i(x)=(x-\mu)^i$ are used in place of moments about zero, $g_i(x)=x^i$, where μ is the mean and σ is the standard deviation.

A second reason for limiting the number of active information functions is the desire to produce a meaningful and usable closed form for p(x), the approximation density. If the number of functions used in a specific representation of p(x) is large, i.e., greater than six, then the practicality of the entropy method is reduced. If six functions are not enough, then we should consider whether we have included the correct potential functions. Selection of potential and active information function sets is the subject of the following four chapters. The potential set defined in Chapter V has produced excellent results for a variety of sample distributions and has

always required five or fewer active functions. An upper bound of six active functions is conceptually reasonable, and experimentation confirms this limit.

Although the Newton method as implemented in ENTROP has succeeded in all tests on [a,b] with initial guess $\Lambda_0=(0,0,\ldots 0)^T$, convergence is not guaranteed. Other numerical schemes exist which can be applied when necessary. An effective though slower method for solving the constraint equations is a method which Acton (Ref 2) has named the "curve crawler." This approach, for three constraints, is initiated by solving the first constraint, i.e., $f_0(\Lambda)=0$ of equations 4.8, for vector Λ. Small steps are then taken along the surface of $f_0(\Lambda)=0$, in the negative gradient direction, while seeking the zero of the second constraint, i.e., $f_1(\Lambda)=0$. The method proceeds by staying close to the $f_0(\Lambda)=0$ and $f_1(\Lambda)=0$ curve and seeking the Λ which zeroes $f_2(\Lambda)=0$. The method was implemented by Orr (Ref 63) in work on the $[0,\infty)$ interval and is explained in detail by Acton. Our success with ENTROP and the Newton method made the development of a backup method unnecessary. The "curve crawler," a gradient descent approach, is suggested as an alternative should the Newton method fail.

Chapter Summary

In this chapter, we have presented the theoretical development of the entropy method to include existence and

uniqueness discussions. Solution of the constraints (equations 4.8) was discussed in both theoretical and applications settings. An effective subroutine to solve the constraints has been developed, tested, and briefly discussed. Using the information function set of Chapter V, the subroutine has been tested against sample densities of the following forms on interval [a,b]: normal, beta, gamma, exponential, uniform, Weibull, mixtures of the preceding densities, and unknown samples. The routine produced exact results, where results were known ahead of time, and statistically acceptable results for unknown distributions. The routine converged for every test with an initial guess of $\Lambda_0=(0,\ldots 0)$. This subroutine is the key element to machine implementation of the entropy characterization method. However, the accuracy of the characterization method in representing unknown distributions centers on selection of the correct information functions. The next four chapters address information function selection.

Chapter V. Potential Information Functions

Information Functions

Given random variable X, we wish to approximate the distribution of X based on available (or computable) information. The information will be collected in terms of average or expected values of certain functions of X; we call these functions "information functions." By specifying the expected values of k information functions, $g_i(x)$, i=1,2,...k, and applying the maximum entropy procedure of previous chapters, we obtain p(x) an approximation to the unknown density of X, where f(x) is the unknown density and

$$p(x) = \exp[-\lambda_0-\lambda_1 g_1(x)-\ldots\lambda_k g_k(x)] \qquad (5.1)$$

Clearly, the number and forms of information functions will impact the accuracy of approximation.

To demonstrate the importance of proper information function selection, we use the moments about zero as information functions. Consider the beta distribution on [0,1]:

$$f(x) = C\,x^{P-1}\,(1-x)^{Q-1} \qquad (5.2)$$

where $C=\Gamma(P+Q)/(\Gamma(P)\Gamma(Q))$, and P and Q are the beta parameters. We first use equation 5.1 with k=1; that is, our

total available information consists of the average values of $g_0(x)$ and $g_1(x)$ where $g_0(x)=1$ and $g_1(x)=x$. As previously discussed, the function $g_0(x)$ corresponds to the constraint that $p(x)$ be a density, i.e., $\int p(x)\,dx=1$. The resulting entropy approximation, given $\langle g_1(x)\rangle$, is denoted $p_1(x)$ where $p_1(x)=\exp[-\lambda_0-\lambda_1(x)]$ on $[0,1]$. Figure 5.1 displays $f(x)$ $(P=4, Q=2)$ and $p_1(x)$ to illustrate the error of approximation. Now consider collecting additional information in terms of $\langle g_2(x)\rangle=\langle x^2\rangle$ to find our second approximation, $p_2(x)=\exp[-\lambda_0-\lambda_1 x-\lambda_2 x^2]$. Notice that λ_i represents the i^{th} Lagrange multiplier in each entropy characterization; however, the Lagrange multipliers in one representation are not related to (and need not equal) the multipliers in subsequent characterizations. Figure 5.2 demonstrates that the increased information, i.e., the second moment, has improved our approximation. Additional information, in terms of additional moments, continues to improve the approximation (Figures 5.3, 5.4 and Table V.I). It can be shown (Ref 90) that $p_k(x)\to f(x)$ as $k\to\infty$. Table V.I shows that at $k=6$ we are approaching an acceptable numeric approximation.

However, the moment approach presents two significant problems. First, as k increases or as the interval of interest, $[a,b]$, becomes large, solution of the constraints for the Lagrange multipliers becomes numerically intractable (see Chapter IV). Some authors feel that "most" well-behaved distributions are "amply" described by

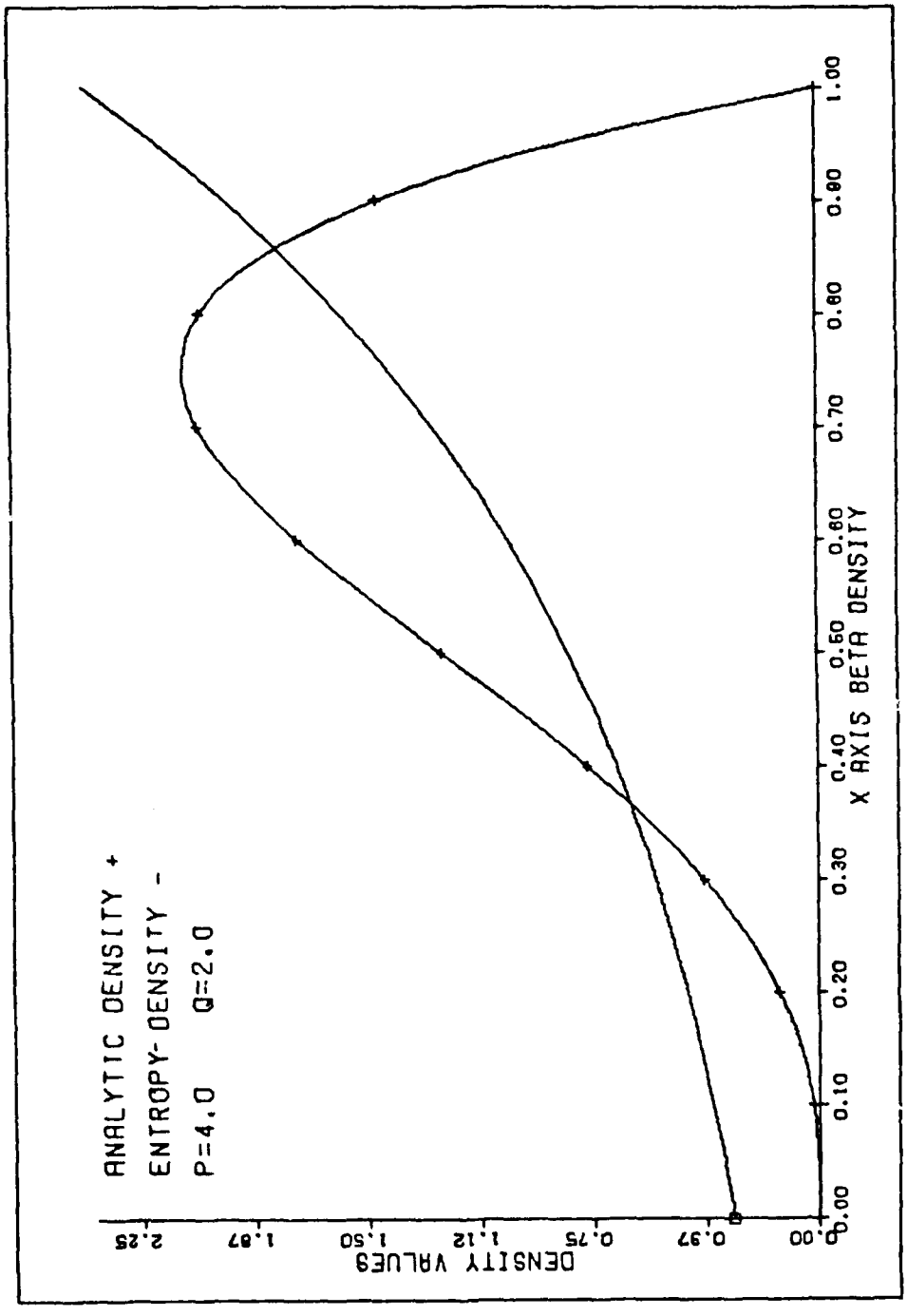


Fig. 5.1. Entropy Approximation with One Moment

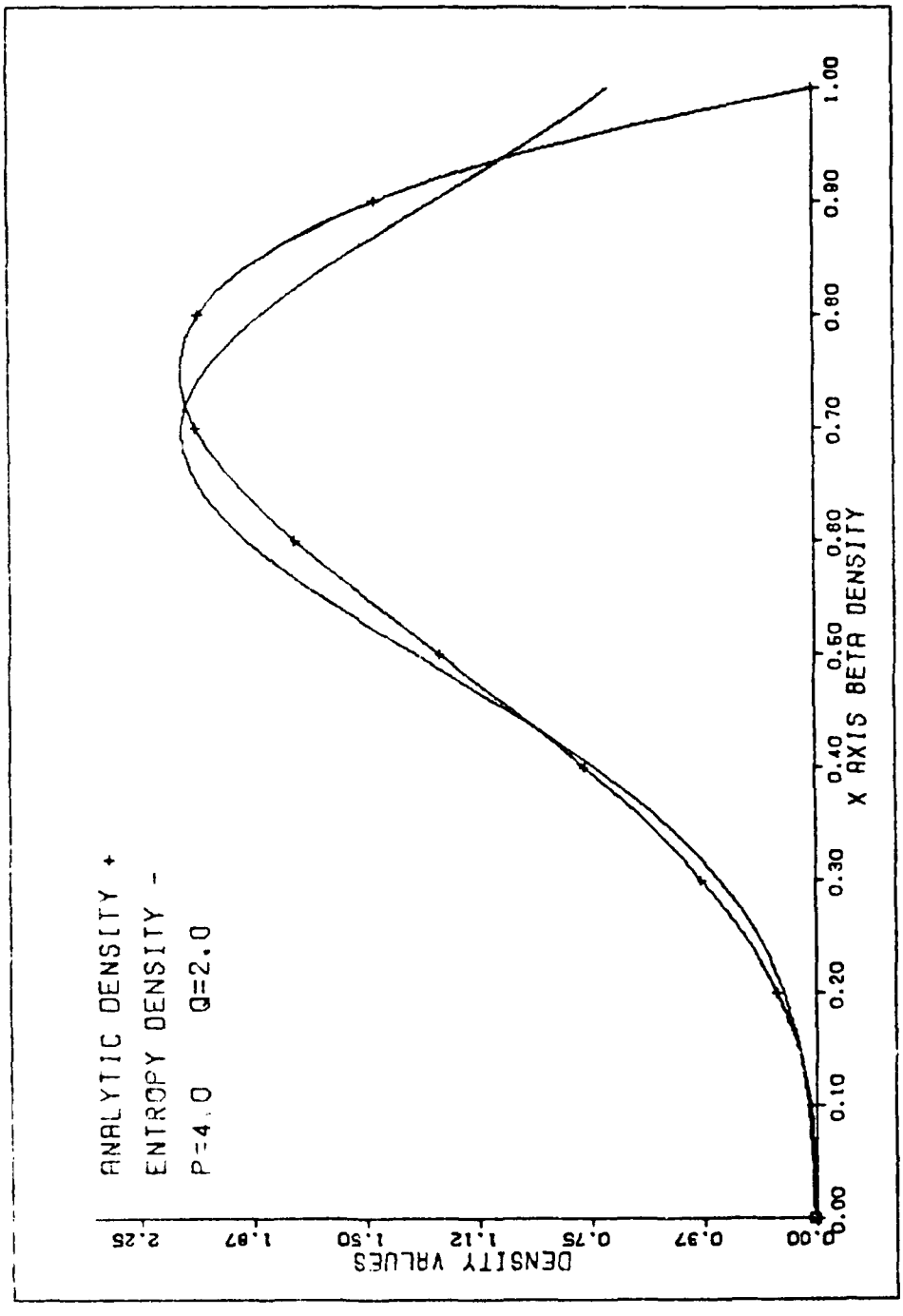


Fig. 5.2. Entropy Approximation with Two Moments

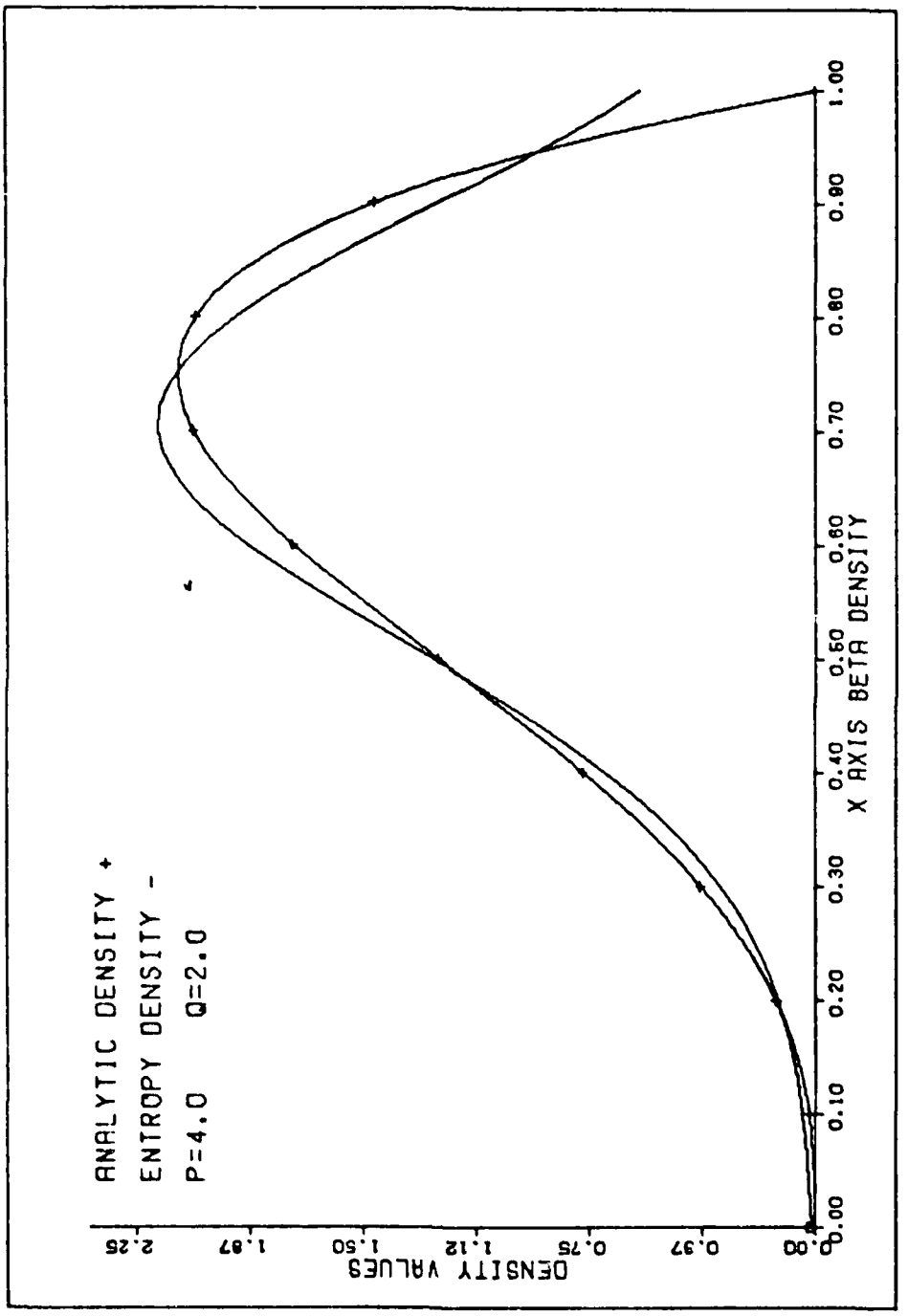


Fig. 5.3. Entropy Approximation with Three Moments

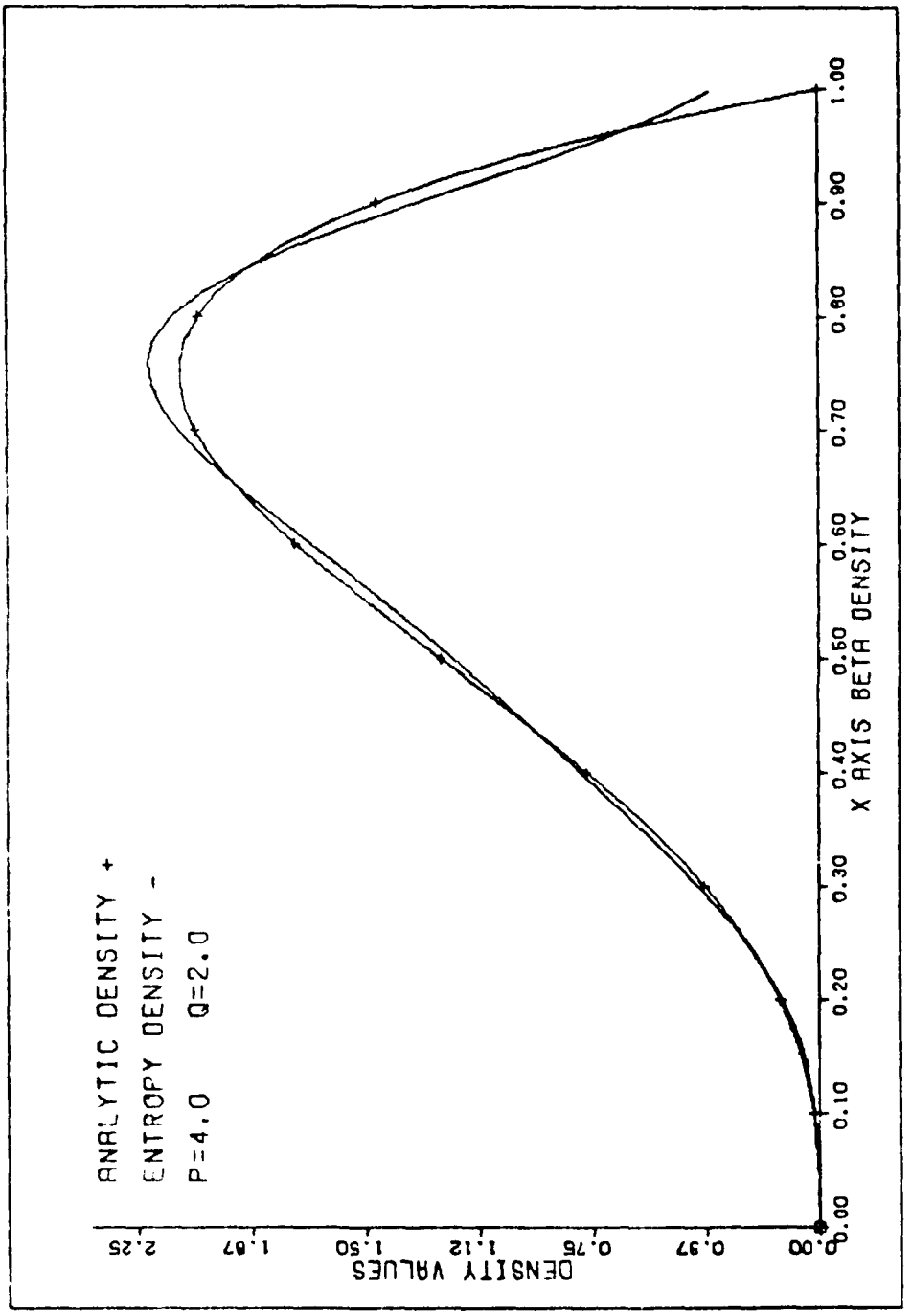


Fig. 5.4. Entropy Approximation with Four Moments

TABLE V.I

DENSITY VALUES FOR BETA AND APPROXIMATIONS

I	X(I)	Analytic Density	Entropy densities with following information functions: 2 Moments	4 Moments	6 Moments
1	.0	.0	.006	.4(-3)	.5(-4)
2	.1	.018	.029	.014	.016
3	.2	.128	.109	.121	.136
4	.3	.378	.319	.401	.373
5	.4	.768	.736	.787	.762
6	.5	1.250	1.331	1.209	1.272
7	.6	1.728	1.889	1.677	1.742
8	.7	2.058	2.104	2.114	2.010
9	.8	2.048	1.838	2.131	2.076
10	.9	1.458	1.260	1.321	1.462
11	1.0	4.3(-12)	.678	.338	.200

NOTE: The entropy density with information functions ln(x), ln(1-x) reproduces the analytic density to machine accuracy.

the first four moments (Ref 49). Pearson's "method of moments" classifies distributions based only on the first four moments (Ref 65). However, Wragg and Dowson (Ref 90) indicate, and our example with the beta distribution implies, that a general characterization scheme would require a larger number of moments. Given that a usable value of k can be distinguished and that the moments can be scaled to allow numerical solution of the constraints, we face a second problem. The second problem is that the algebraic form of p(x), with moments as information functions, tells the analyst very little about the unknown density, and p(x) may be computationally difficult to handle even with four moments. For example, consider the beta distribution of equation 5.2 *once more*. Let $g_1(x)=\ln(x)$ and $g_2(x)=\ln(1-x)$ and we obtain $p(x)=\exp[-\lambda_0-\lambda_1 \ln(x)-\lambda_2 \ln(1-x)]$. Application of our numerical scheme produces $\lambda_0=-2.9957323$, $\lambda_1=-3.0$, and $\lambda_2=-1.0$. We now consider the form of p(x):

$$p(x) = \exp[-\lambda_0] \exp[-\lambda_1 \ln(x)] \exp[-\lambda_2 \ln(1-x)], \text{ or}$$

$$p(x) = \exp[-\lambda_0] \, x^{-\lambda_1} \, (1-x)^{-\lambda_2}, \text{ and}$$

$$p(x) = 20.0 \, x^{3.0} \, (1-x) \, .$$

Thus p(x) exactly equals f(x), and we have only used two information functions. Further, the algebraic form of the entropy density tells us that we are working with a beta distribution. Our selected functions are clearly superior

to moments in this example. The above examples illustrate the importance of selecting the correct information functions and indicate the numerical and conceptual deficiencies of relying solely on moments. Orthogonal families of functions (Ref 15) present the same conceptual and numeric problems as observed with moments.

The procedure defined in Chapter III presents a viable alternative for approximation of unknown densities on a bounded interval, [a,b]. The information function selection step of the procedure includes two phases. In the first phase, we specify a large set of potential information functions; that is, linearly independent functions that may prove useful in representing distributions on [a,b]. A potential set that has proved extremely useful for a variety of unknown densities is defined in the next section of this chapter. The procedure is designed to allow flexibility in definition of the potential set, and this flexibility is also discussed. In the second phase, we select an "active set" of information functions from the potential set. A large number of active functions leads to more accuracy in the approximation. However, a large number of functions also leads to numerical difficulties and a loss of conceptual significance in the form of p(x). Selection of the active set is thus a compromise; we want the active set to be as small as possible within our accuracy

restrictions. Three different methods for selection are discussed in subsequent chapters.

A Potential Information Function Set

Our initial approach to specifying a potential set for use in a general approximation problem centers on an investigation of named distributions (Refs 37; 47; 55; 60). We consider the algebraic form of various well-known distributions and determine what information functions, if any, will produce an equivalent entropy density, p(x). In the example of equation 5.2 we saw that information functions ln(x) and ln(1-x) produced a beta distribution on [0,1]; that is, if we provide $\langle \ln(x) \rangle$, $\langle \ln(1-x) \rangle$, and apply the entropy procedure, then the resulting entropy density will be a beta. If we specify no information functions, i.e., only $g_0(x)=1$ on [a,b], then the resulting entropy density, $p(x)=\exp(-\lambda_0)$, equates to the uniform density. A list of the more well-known distributions and the resulting information functions is shown at Table V.II. We reason that many continuous distributions on [a,b] will be closely approximated by the listed distributions or some combination of these distributions. Using Table V.II, we select the most versatile distributions and eliminate redundant functions to produce the potential set shown in Table V.III. Notice that Table V.III includes the first four moments which we also represent as normalized central

TABLE V.II

INFORMATION FUNCTIONS FOR NAMED DISTRIBUTIONS

Distribution	Density		Information Functions
Uniform	$1/(b-a)$	$a \leq x \leq b$	None
Exponential	$\alpha \exp[-\alpha(x-a)]$	$0<\alpha$, $a<x$	x
Normal	$(2\pi\sigma^2)^{-\frac{1}{2}} \exp[-(x-\mu)^2/(2\sigma^2)]$	$0<\sigma$	x, x^2
Beta	$\frac{\Gamma(P+Q)}{\Gamma(P)\Gamma(Q)} \frac{(x-a)^{P-1}(b-x)^{Q-1}}{(b-a)^{P+Q-1}}$	$a \leq x \leq b$	$\ln(x-a)$, $\ln(b-x)$
Gamma	$(\beta^{\alpha}\Gamma(\alpha))^{-1} x^{\alpha-1} \exp[-x/\beta]$	$0<\alpha$, $0<\beta$, $0<x$	x, $\ln(x)$
Weibull	$c\alpha^{-c}(x-a)^{c-1}\exp[-(x-a)^c/\alpha^c]$	$a<x$, $0<c$, $0<\alpha$	$\ln(x-a)$, $(x-a)^c$
Log normal	$\alpha(2\pi)^{-\frac{1}{2}}(x-a)^{-1}\exp[-1/2(\alpha\ln(x-a))^2]$	$a<x$	$\ln(x-a)$, $[\ln(x-a)]^2$
Student t	$\frac{\Gamma((\nu+1)/2)}{(\pi\nu)^{\frac{1}{2}}\Gamma(\nu/2)} [(x^2/\nu)+1]^{-(\nu+1)/2}$	$\nu=1,2,\ldots$	$\ln((x^2/\nu)+1)$
Inverse Gaussian	$(\beta/(2\pi x^3))^{\frac{1}{2}} \exp[-\beta(x-\mu)^2/(2\upsilon x)]$	$0<x$, $0<\beta$, $0<\mu$	$\ln(x)$, x, $1/x$
Double Exponential	$1/2\alpha \exp[-\lvert x-\mu\rvert/\alpha]$	$0<\alpha$	$\lvert x-\mu\rvert$
hyperbeta (Ref 60)	$\frac{n\Gamma(b+c+1+1/n)}{\Gamma(b+1/n)\Gamma(c+1)} x^{nb}(1-x^n)^b$	$0 \leq x \leq 1$, $n=1,2,\ldots$	$\ln x$, $\ln(1-x^n)$

NOTE: Some of the listed densities will require a normalizing constant when restricted to a bounded interval, but the information functions will be unchanged.

TABLE V.III

A STARTING POTENTIAL SET

Symbol	Function	Symbol	Function
F1	x or $(x-\mu)/\sigma$	F6	$[\ln(x-a)]^2$
F2	x^2 or $[(x-\mu)/\sigma]^2$	F7	x^3 or $[(x-\mu)/\sigma]^3$
F3	$\ln(x)$	F8	x^4 or $[(x-\mu)/\sigma]^4$
F4	$\ln(x-a)$	F9	$\ln(x^2+1)$
F5	$\ln(b-x)$		

moments, i.e., $g_2(x) = (x-\mu)^2/\sigma^2$ where μ is the calculated mean and σ is the standard deviation. Normalization was needed to provide numerical stability for a specific simulation application on [5104.0, 8262.0]. Normalization is effective on large intervals, [a,b], but may produce the opposite result if $b-a \leq 1$. On small intervals normalized moments involve small values divided by small values which will lead to numerical instability. Thus, origin or central moments are more effective on small intervals.

The functions in Table V.III are not intended as the ultimate potential set but will serve as an excellent starting point for any characterization. Functions can and should be added to this set (or deleted) based on data analysis for a particular problem. As an example, we consider a distribution that was first investigated by Chanda and Kulp (Ref 11). The data consists of 2000 samples,

x_i, i=1,2,...2000, from an unknown distribution. We translate the data from [-.4894, .5028] to [.0106, 1.003] to preclude difficulty with the natural logarithms in Table V.III. We compute average values from the sample as explained in Chapter IV. Using the potential set of Table V.III, we select the active set with method three of Chapter VIII. The resulting "best" fit required six active functions and is shown in Figure 5.5. The sample density is also shown and was created by sorting the 2000 deviates, creating the cumulative at each sample point, CUM_i=i/2000, i=1,2,...2000, and numerically differentiating. The initial approximation missed the peaked structure of the sample which indicates that we failed to specify sufficient information. The peaked sample suggests the shape of a double exponential density, and we thus add information functions |x-.5| (the .5 accounts for translation of the data) to the potential set. Application of method three resulted in four active information functions with the excellent characterization in Figure 5.6. Thus the nature of the data suggested the addition of a function to the potential set, and that function was subsequently selected as active.

This example again illustrates the importance of proper information function selection but also highlights the flexibility of the procedure. The characterization procedure was designed as a tool for the analyst. Consequently, data analysis and an analyst's insight can be used

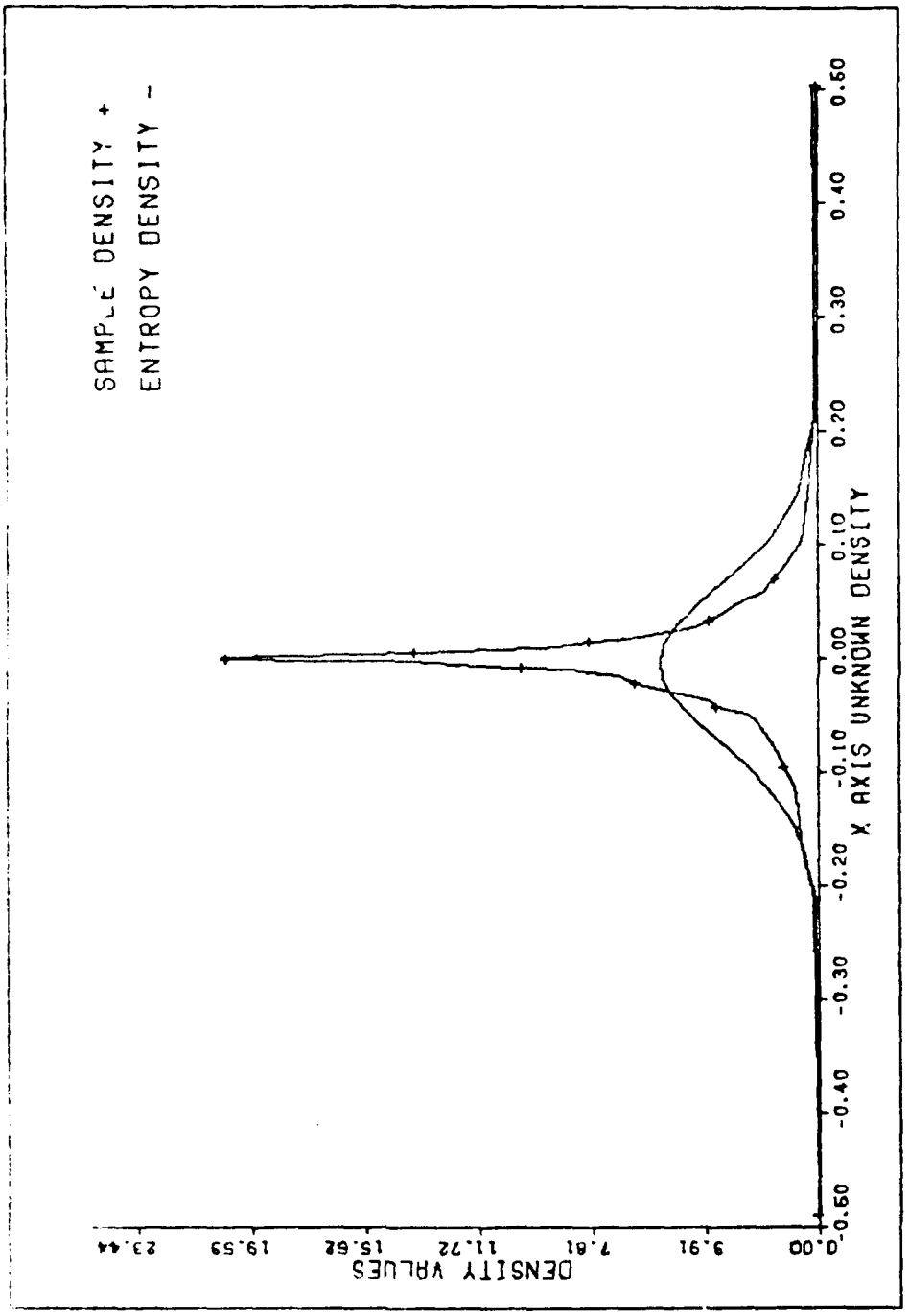


Fig. 5.5. Entropy Approximation with Standard Potential Set

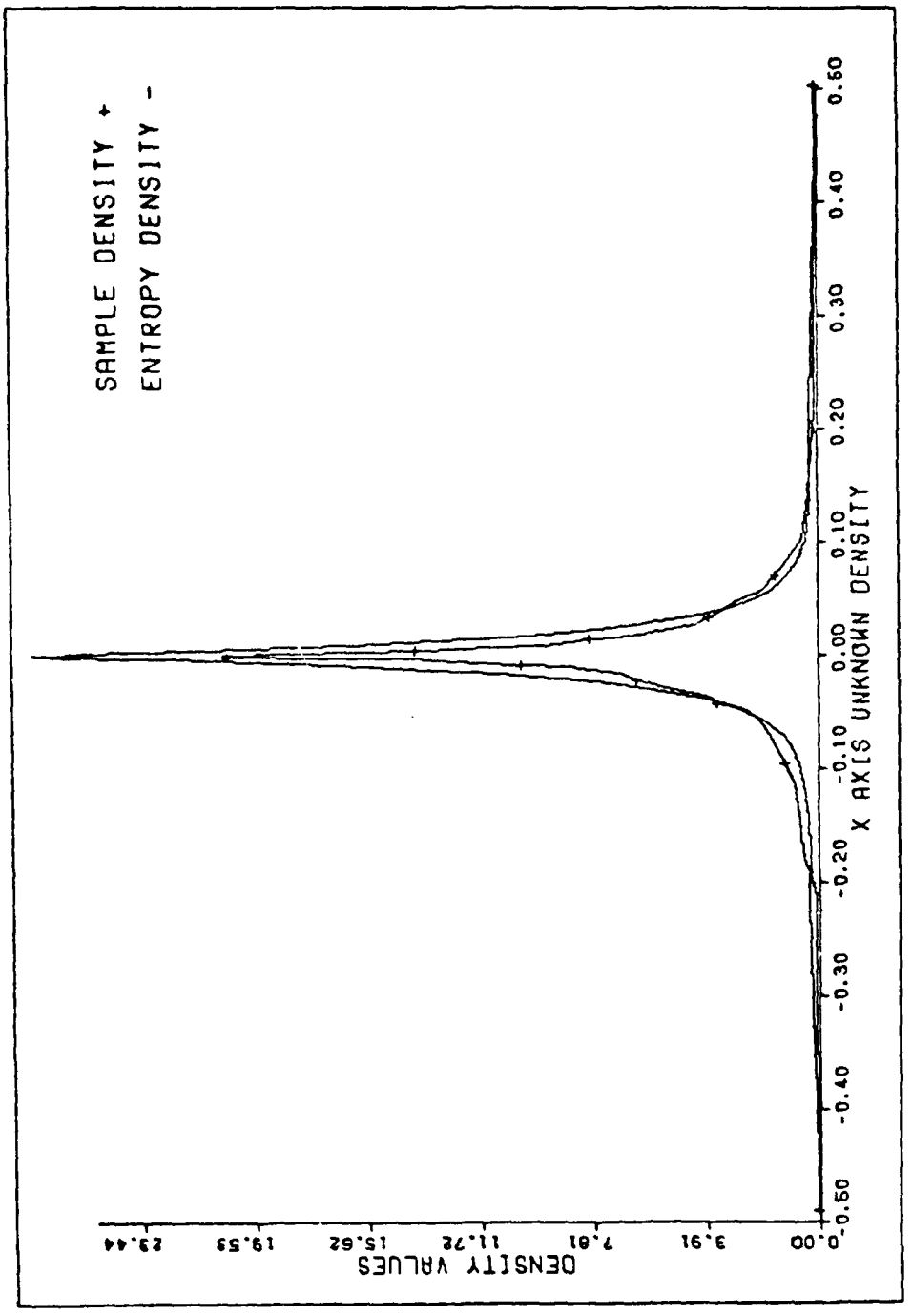


Fig. 5.6. Entropy Approximation with Modified Potential Set

to enhance the procedure, particularly in the information function selection phases. The potential set of Table V.III has been tested on a variety of unknown densities and produced excellent results. If the potential set contains a sufficient mixture of functions, then subsequent procedural steps will eliminate unnecessary functions and choose the useful functions, i.e., the active set.

Active Information Function Set

Selection of the active information functions from the potential set is the subject of the next three chapters. The active set was previously defined as that subset which is used in the entropy approximation for a specific set of data. The goal of the selection procedure is to pick the minimum subset which conveys enough information to provide an acceptable approximation to the unknown density, f(x). Selection of the active set depends on how one defines "acceptable approximation" and how one measures "closeness" of approximation. Three different approaches to the problem resulted in three viable methods, each with different qualities. The methods were tested by generating data from known distributions and evaluating the resulting approximations. If the potential set includes the correct information functions for a sample density, then the selection procedure should select those functions and produce an exact fit. For example, if the data is from a normal

density then the selection method should select x and x^2. For the beta distribution on [0,1], we expect ln(x) and ln(1-x) to be selected. If the correct information functions are not in the potential set, then we wish to select the best subset to approximate f(x). The three selection methods produce excellent approximations as will be demonstrated. The choice of method for a particular approximation problem will depend on the available information, accuracy of the information, and, in some cases, the analyst's preference.

Chapter VI. Active Set Selection-- Method One (Regression)

Introduction

The concepts of potential and active information function sets were discussed in Chapters III and V. One of three general methods to select the active set for a specific density approximation is presented in this chapter. The approach is based on linear regression and requires a random sample of the unknown distribution, x_i, i=1,2,...N. We first present the procedural steps for method one and follow with detailed discussion of a few of these steps. Sample applications are then presented to demonstrate method strengths and sensitivities. The excellent results of method one led to alternate methods as presented in subsequent chapters.

Method One Procedure

Method one includes five procedural steps:

1. The first step is generation of the sample cumulative distribution. The sample, x_i, i=1,2,...N, is sorted and the cumulative distribution is approximated at these N points; CUM_i=i/N, i=1,2,...N. Clearly, as N increases, the approximation becomes more accurate. The cumulative data is grouped, for large N, to produce better

results in subsequent steps; for example, with N=500 we use every tenth x_i to reduce the data from 500 points to 51 points with $CUM_j=(j-1)10/N$, $j=1,2,...51$.

2. Step two produces a numerical estimate for $f(x)$, the unknown density. We numerically differentiate the sample cumulative to produce a numerical density at M points, DEN_i, $i=1,2,...M$, $M \leq N$. Numerical differentiation is an ill-posed problem (Refs 36; 53) and requires care in application. Our work with differentiation techniques produced interesting results which are presented later in this chapter. The initial approach to numerical differentiation was central difference;

$$DEN_i = [CUM_{i+1}-CUM_{i-1}]/[x_{i+1}-x_{i-1}], \quad i=2,4,...(M-1).$$

3. The third step produces the natural logarithm of the numerical density; $\ln(DEN_i)$, $i=1,2,...M$. The purpose of this step is to establish a linear relationship between the numerical density and the entropy density. We seek the minimum set of information functions, $g_j(x)$, $j=1,2,...k$, which are essential for accurate approximation of $f(x)$. Let the entropy density include all the potential functions, i.e.,

$$p(x) = \exp[-\lambda_0-\lambda_1 g_1(x)-...\lambda_m g_m(x)]$$

where m is the number of functions in the potential set, g_m. We establish the linear relationship as follows:

$$\ln f(x) = \ln p(x) = -\lambda_0 - \lambda_1 g_1(x) - \ldots \lambda_m g_m(x) \qquad (6.1)$$

We now have a form which allows linear regression, and we have M data points ($\ln(DEN_i)$) to approximate $\ln(f(x))$.

4. The fourth step is to apply linear regression to identify several possible sets of active information functions. This regression step reduces the number of combinations of potential functions, i.e., subsets of the potential set, that will be considered in selection of the active set. We consider sets of five or fewer functions for reasons presented in previous chapters.

5. The final step is selection of the active set from the sets defined in step 4. A measure of "goodness of fit" is specified and the active set is the set whose corresponding entropy characterization provides the "best" fit to the data.

The above procedure has produced excellent results as will be shown. Currently, the procedure is implemented in two separate packages to allow maximum analyst involvement. The first package accomplishes steps 1 thru 4 and returns 10 candidate sets of functions. The analyst may use some or all of the 10 sets, or other combinations of functions, as input to the second program which accomplishes step 5. In our examples we will use only 5 of the 10 candidate sets. Steps 2, 4, and 5 are discussed in more detail in the following sections.

Numerical Differentiation

Method one is a straightforward application of well known numerical and statistical techniques; however, initial testing indicated a sensitivity to numerical and sampling errors. The ill-posed nature of numerical differentiation can result in exaggerated numerical error, thus various differentiation schemes were investigated to reduce this risk. The investigation resulted in a previously unpublished scheme, the "median" method, which generally outperformed other schemes for our application. A summary of the investigation follows.

Polynomial and Spline Approximations. International Mathematical and Statistical Libraries (IMSL) subroutines were used to fit a polynomial to the sample cumulative and, subsequently, to differentiate the polynomial. Polynomials of up to sixth degree were tested but produced poor results. Polynomial "wiggle" (Ref 40) caused negative density values. Spline approximation and spline interpolation (Ref 5) were attempted with existing IMSL software in an effort to reduce the polynomial "wiggle." Differentiation of the spline also produced negative density values at a few points and proved unsatisfactory.

Sliding Polynomial. A program was written which makes a least squares fit to five data points using a second degree polynomial. Beginning with the first five data

points (x_i and CUM_i, i=1,2,3,4,5), the program produces a second degree polynomial fit to the cumulative data and uses the polynomial coefficients to calculate the derivative at the middle point, i.e., at x_3. The program then advances the operative window by one data point and repeats the procedure for i=2,3,4,5,6 to find the derivative at x_4. The procedure continues in this fashion to produce DEN_i, i=3,4,...(N-2). Forward and backward difference formulas are used for the first two and last two data points. A second program was written to accomplish the identical procedure but using seven data points instead of five. The intent of using seven points is to provide more of a smoothing effect on the data. The seven-point formula did produce a more accurate derivative than the five-point formula, and both schemes generally outperformed the central difference approach.

Sliding Median. The median method is based upon a nonparametric regression parameter estimator which was first suggested by Theil (Ref 79). The distribution of this estimator was investigated by Sen (Ref 69) and Chanda and Kulp (Ref 11). To our knowledge, this scheme has not been previously used for numerical differentiation. As in the polynomial approach, we define an operating window about the first seven data sets, x_i and CUM_i, i=1,2,...7. We then consider all combinations of these seven distinct

data sets, taking two sets at a time, and calculate the value $[CUM_k-CUM_j]/[x_k-x_j]$ for each combination. This value equates to the central difference at the point midway between x_j and x_k. The values for the 21 combinations are then sorted, and DEN_4, the density at point x_4, is assigned the median value. The operating window is advanced one data point, and the procedure repeats to find DEN_5. We iterate for i=4,5,...(N-3). Forward difference is used to find the density at x_1, central difference for points x_2, x_3, x_{N-2}, x_{N-1}, and backward difference for point x_N. A similar program was created for an operating window of only 5 data points. As in the polynomial case, the 7 point formula performed better than the 5 point formula. The simplicity of the median method resulted in faster computation than the polynomial approach.

The listed methods were tested against sample cumulatives from known distributions and known densities, f(x). The sum of errors squared, $SE = \sum_{i=1}^{M} (f(x_i)-DEN_i)^2$, and mean squared error, SE/M, were calculated for comparison. Three example distributions are provided in Table VI.I. Each sample set in Table VI.I is composed of 500 deviates from the stated distribution. The sample distributions are further described in Table VI.II. The 500 deviates were grouped to M data points before differentiating. The first three comparisons demonstrate the effect of grouping data from a given sample set. The last

TABLE VI.I

NUMERICAL DIFFERENTIATION SCHEMES

Sample Set No.	Distribution	M	Scheme	Time (sec)	$(SE)^2$	$(SE)^2/M$
1	n(10,2)	500	Central Dif.	.035	230.13	.460
1	n(10,2)	500	Poly. 5	.506	49.9	.0998
1	n(10,2)	500	Poly. 7	.572	12.02	.0240
1	n(10,2)	500	Median 5	.211	20.51	.0410
1	n(10,2)	500	Median 7	.528	8.80	.0176
1	n(10,2)	101	Central Dif.	.002	.637	.00630
1	n(10,2)	101	Poly. 5	.098	.322	.00319
1	n(10,2)	101	Poly. 7	.113	.180	.00178
1	n(10,2)	101	Median 5	.041	.227	.00225
1	n(10,2)	101	Median 7	.096	.164	.00162
1	n(10,2)	51	Central Dif.	.0001	.1210	.00237
1	n(10,2)	51	Poly. 5	.046	.0664	.00130
1	n(10,2)	51	Poly. 7	.052	.0472	.000925
1	n(10,2)	51	Median 5	.022	.0560	.001098
1	n(10,2)	51	Median 7	.019	.0410	.000804
2	n(10,2)	51	Central Dif.	.028	.1021	.00201
2	n(10,2)	51	Poly. 5	.048	.0457	.000895
2	n(10,2)	51	Poly. 7	.055	.0255	.000500
2	n(10,2)	51	Median 5	.023	.0396	.000776
2	n(10,2)	51	Median 7	.039	.0245	.000481
3	beta P=4, Q=2	51	Central Dif.	.03	6.767	.1327
3	beta P=4, Q=2	51	Poly. 5	.049	3.579	.0702
3	beta P=4, Q=2	51	Poly. 7	.056	2.127	.0417
3	beta P=4, Q=2	51	Median 5	.019	3.581	.0702
3	beta P=4, Q=2	51	Median 7	.041	2.375	.0466

NOTE: Each sample set includes 500 data points which were grouped to M points before differentiation.

$$(SE)^2 = \sum_{i=1}^{M} (\text{Actual-Approx})^2.$$

TABLE VI.II

SAMPLE DISTRIBUTION CHARACTERISTICS

Sample Set No.	Distribution	$[x_1, x_N]$	Sample Mean/Variance	Analytic Mean/Variance	Generation Method
1	Normal	6.07, 13.35	10.05/1.859	10/2	Box-Muller (Ref 8)
2	Normal	5.54, 13.81	10.02/2.0142	10/2	Box-Muller
3	Beta	.1696, .995	.6643/.0313	.6667/.0317	IMSL Subroutine

NOTE: Each sample set contains 500 deviates.

two comparisons are representative examples from other samples. The point of significance is that the median method generally produced closer approximations to the known densities than either the polynomial or central difference methods. The median approach limits extreme values caused by numerical and sampling error, thus producing a closer fit to the true analytic density. The method does not eliminate differentiation "noise" but does control the magnitude of this noise. Figure 6.1 and Figure 6.2 provide examples of the densities produced by the median method for a normal distribution (mean 10 and variance 2) and a beta distribution (on [0,1] with P=4, Q=2). The sliding median method with 7 data points was used for all subsequent numerical differentiation in the research.

Regression

We use linear regression in step 4 to reduce the number of candidate active sets. Linear regression is a well known and well defined analytical tool. Drapper and Smith (Ref 23) and others (Refs 28; 37; 38) provide detailed explanation of regression procedures, regression statistics, and stopping criteria. Both "stepwise regression" and "regression by leaps and bounds" were researched to include available software for implementation; Statistical Package for the Social Sciences (SPSS) and IMSL subroutine libraries contain regression packages. The IMSL

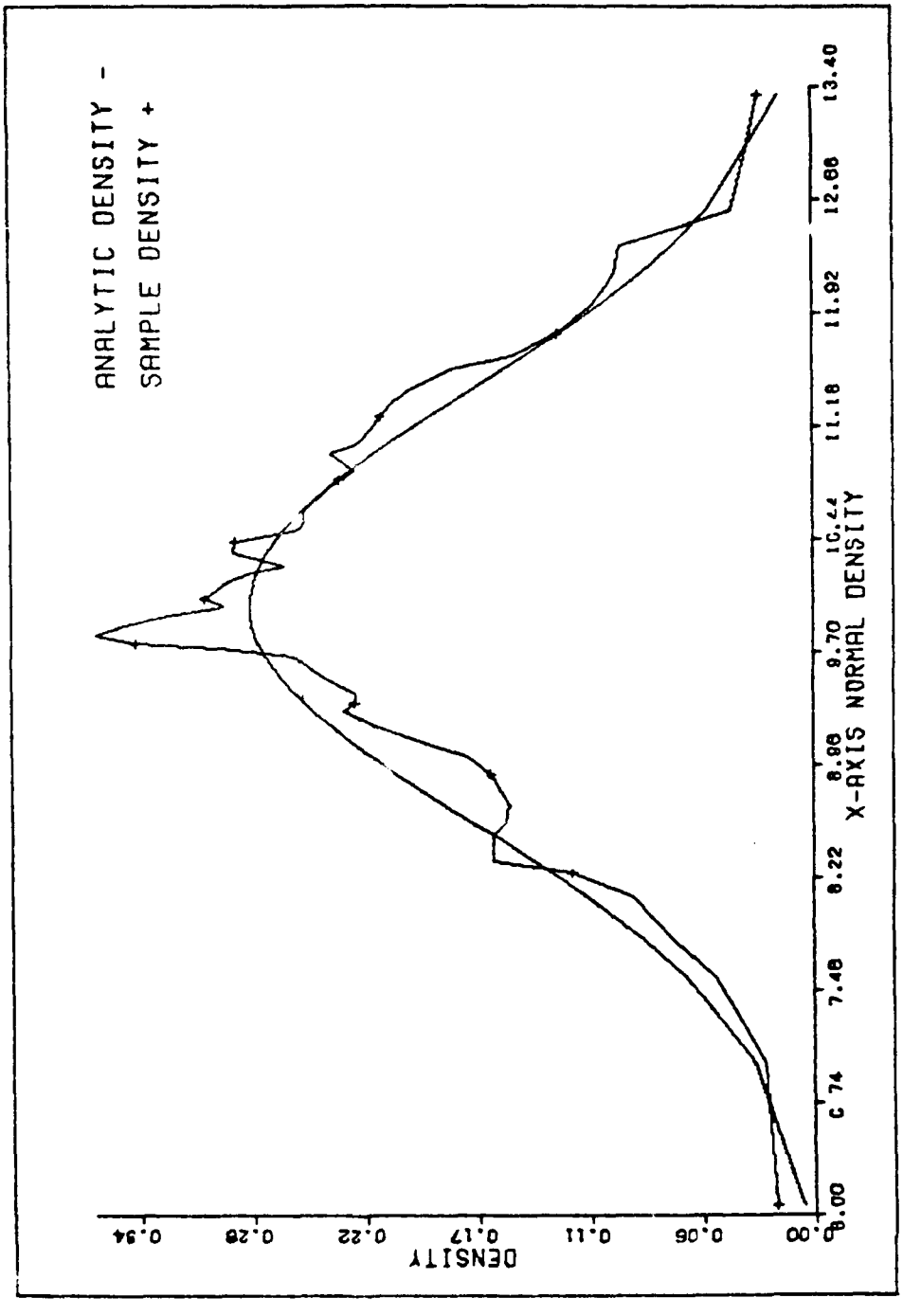


Fig. 6.1. Numerical Differentiation--Data Set 1

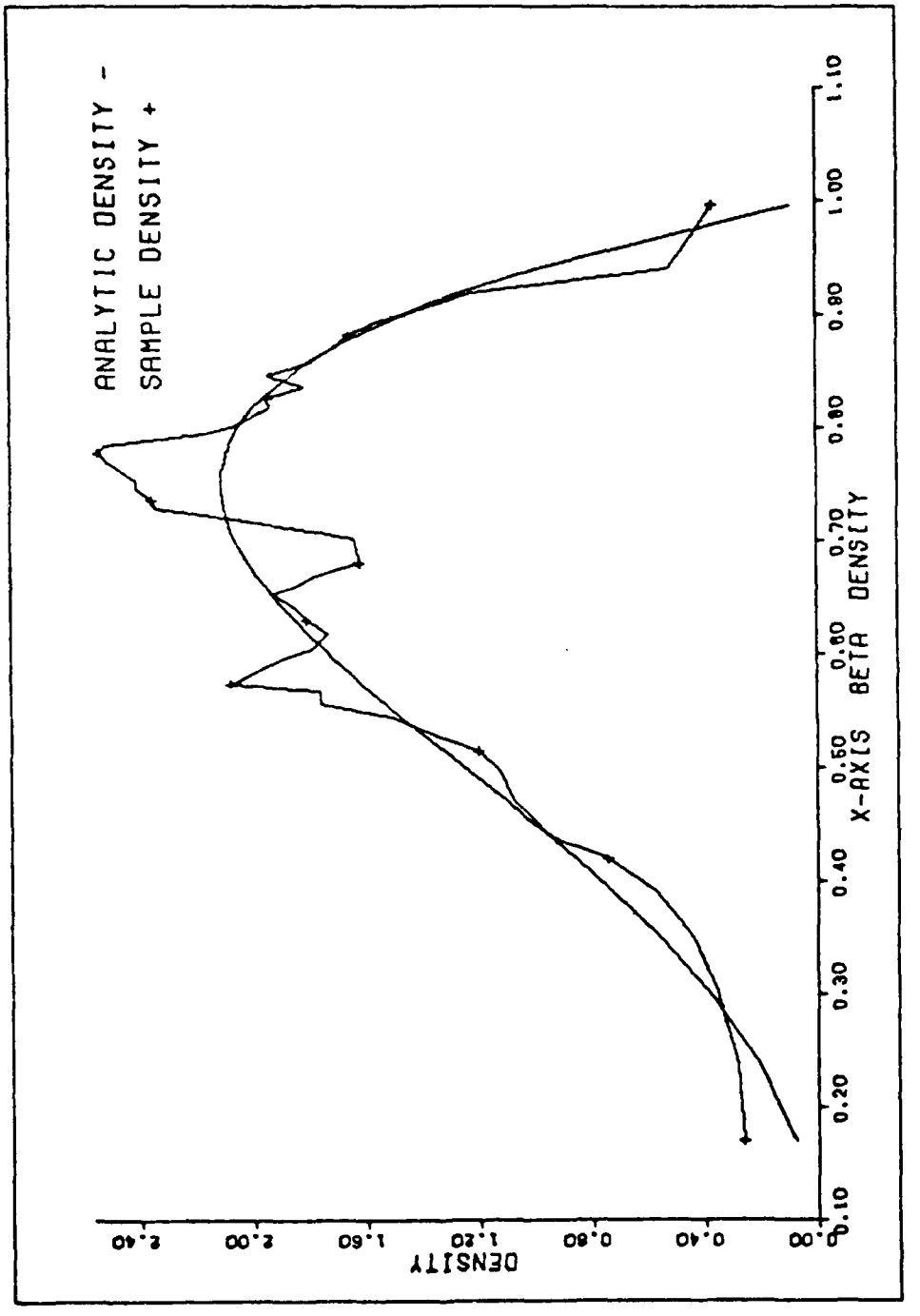


Fig. 6.2. Numerical Differentiation--Data Set 3

leaps and bounds package, RLEAP, was selected for our application with the adjusted R^2 statistic (Ref 38). We created a program which uses RLEAP to return ten candidate active sets with corresponding adjusted R^2 statistics; i.e., the two best sets with one active variable, through the two best sets with five active variables. "Best set" is defined as the set with the largest adjusted R^2 statistic.

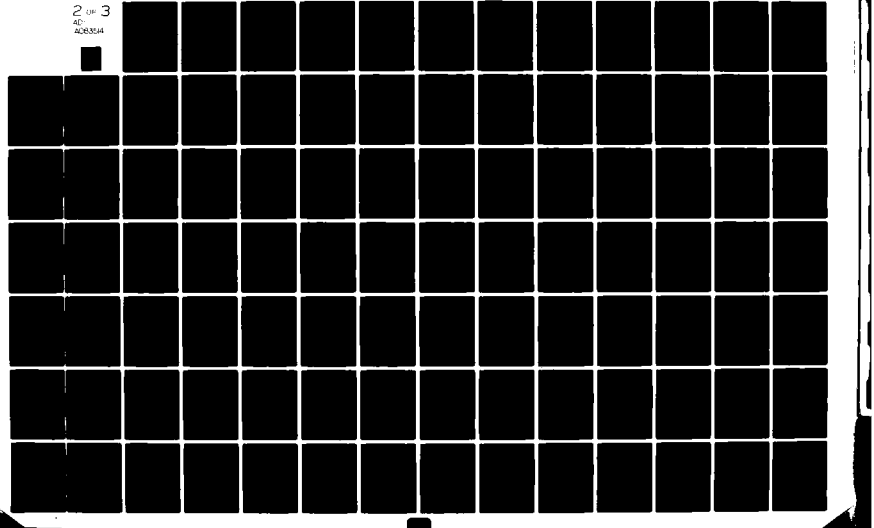
An examination of our regression procedure will identify a significant benefit of our entropy approximation method. We use regression to identify ten possible active sets. However, we could use the same regression packages to select the single active set that produces the best regression fit to the data by choosing the one set with the largest adjusted R^2. Regression will also produce the regression coefficients, λ_i, i=0,1,...k, for our selected set to completely specify p(x) as in equation (6.1). Thus, why not stop at the regression step? The reason for step 5 centers on the purpose of linear regression. Regression seeks a fit to the sample data and the p(x) produced in regression is not required to satisfy average value or density constraints (see Chapter IV). Thus, the regression p(x) need not be a density function at all and will be a maximum entropy density only by coincidence. The sole intent of the regression procedure, as we have defined it, is to identify information functions that

play a key role in describing the sample cumulative. Once regression has defined several candidate sets of functions, we return to the entropy approach to select the density (i.e., select λ_i, i=0,1,...k) which satisfies average value constraints. This entropy approach is thus a compromise between a fit to the sample data and constraint satisfaction. One may view constraint satisfaction as an attribute of the underlying analytic distribution versus a function of the sample distribution. Consequently, our approach provides a compromise between the unknown analytic distribution and the provided sample. The examples of the next section will demonstrate this quality.

Experimental Results

The strengths and sensitivities of method one are best demonstrated with examples. We consider the first four steps of method one in this section and subsequently discuss step five. Our goal is to select the minimum set of functions (active set) which produces an acceptable, closed form, entropy approximation, p(x), to the unknown density, f(x). To test method one, we generate random samples of size 500 from known distributions, i.e., normal, beta, and gamma. Thus the analytic cumulative and density functions are available for comparison to sample and entropy distributions. We will consider the three sample data sets of Table VI.II for purposes of illustration. The normal

AD-A083 514 AIR FORCE INST OF TECH WRIGHT-PATTERSON AFB OH SCHOO--ETC F/6 12/1
CONTINUOUS DENSITY APPROXIMATION ON A BOUNDED INTERVAL USING IN--ETC(U)
MAR 80 J E MILLER
UNCLASSIFIED AFIT/DS/MA/80-1 NL



samples were produced by the Box-Muller method (Ref 8) applied to uniform deviates. The beta deviates were produced via an IMSL subroutine.

We first discuss the sensitivity of method one to sampling and numerical differentiation errors. Experimentation with normal, beta, and gamma distributions has shown that if the actual analytic density is used in place of the numerical differentiation step (i.e., sampling and differentiation variations are not permitted), then the regression step will select the "correct" information functions with an exact fit to the data (i.e., adjusted R^2 equal to 100). The "correct" functions for a distribution are the information functions presented in Table V.II, i.e., x and x^2 for normal, ln(x) and ln(1-x) for beta, etc. However, when the complete procedure is applied, i.e., a sample cumulative is generated and differentiated, the regression step does not necessarily select the expected information functions. In fact, the randomness of sample data may cause selection of different function sets when multiple regression is applied to different samples from the same distribution. The interesting point is that whichever set of functions is selected, excellent approximations are obtained. Thus, method one demonstrates data dependence.

Samples one and two of Table VI.II for the normal distribution provide an example of data dependence. We

use the potential information functions of Table VI.III, where μ, σ^2, and [a,b] are 10, 2, $[x_1, x_N]$ respectively, and apply method one. Table VI.IV lists five candidate active sets that result for each sample, EI, I=1,2,...5, where I represents the number of active functions. Notice we have included only five candidate sets; the analyst may choose to consider a larger number of sets for other applications. For our potential set, $F2=((x-\mu)/\sigma)^2$ is the information function that will exactly characterize the normal distribution, and F2 is dominant for both data sets. If we next solve the constraint equations, given accurate expected values for the information functions, we will produce an exact fit to the analytic distribution for any set which contains F2. Figure 6.3 provides a comparison of sample and entropy cumulatives to the known analytic cumulative for active set E1. Entropy and analytic cumulative values were computed by numerical integration of respective densities. Differences in cumulatives, i.e., sample-analytic and entropy-analytic, are shown to facilitate comparison and because the distributions are very close. The entropy-analytic curve is identically zero because the entropy approximation provides a near perfect fit to the analytic distribution. Graphs were also produced for sets E2, E3, E4 from data set one and E1 through E5 for data set two. The graphs were nearly identical to Figure 6.3 because F2 (the correct information function) was part of

TABLE VI.III

POTENTIAL INFORMATION FUNCTION SET

$F1 = (x-\mu)/\sigma$	$F4 = \ln(x-a)$	$F7 = ((x-\mu)/\sigma)^3$
$F2 = ((x-\mu)/\sigma)^2$	$F5 = \ln(b-x)$	$F8 = ((x-\mu)/\sigma)^4$
$F3 = \ln(x)$	$F6 = (\ln(x-a))^2$	$F9 = \ln(x^2+1)$

NOTE: μ = mean; σ = standard deviation; [a,b] = bounds.

TABLE VI.IV

REGRESSION RESULTS FOR NORMAL SAMPLES

Candidate Set	Functions for Data Set #1	Adjusted R^2	Functions for Data Set #2	Adjusted R^2
E1	F2	91.76	F2	89.70
E2	F2,F6	95.75	F2,F8	95.70
E3	F2,F6,F8	96.97	F2,F6,F8	96.19
E4	F2,F4,F5,F6	97.07	F2,F3,F4,F5	97.37
E5	F1,F5,F6,F7,F9	97.12	F2,F4,F5,F8,F9	98.66

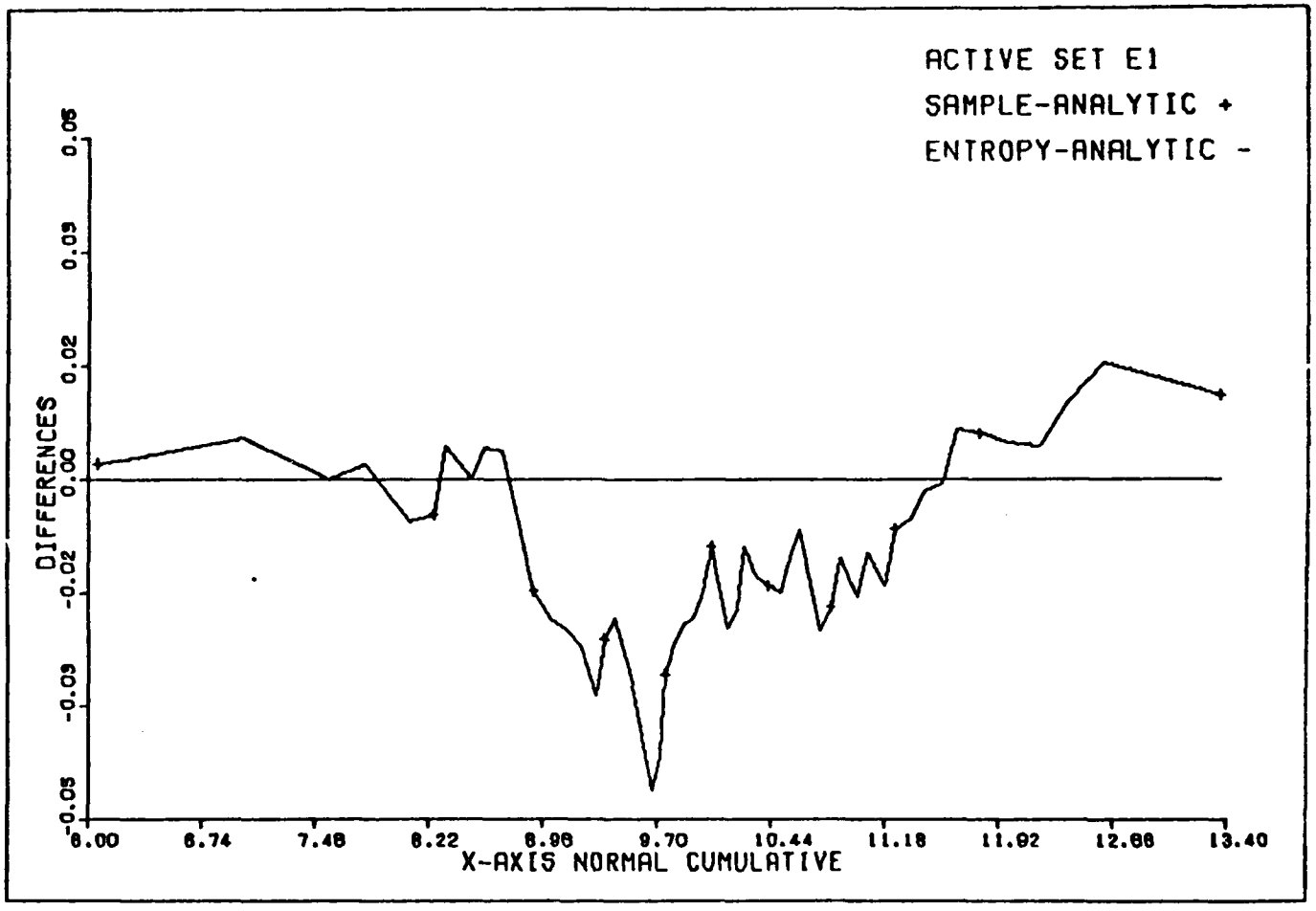


Fig. 6.3. Cumulative Difference--E1/Data Set 1

the entropy representation, and additional functions provided little usable information. Set E5 of data set one does not include F2 and yet provides an excellent regression fit to the data, i.e., R^2=97.12. The resulting entropy approximation also produces an excellent fit to the analytic and thus the sample as shown in Figures 6.4 and 6.5. Table VI.V provides further insight with a numerical interpretation of the entropy approximations produced by E1 through E5 for data set one.

The beta distribution on [0,1] provides a more revealing example. Because we are on the [0,1] interval, the normalized moments in our potential set (Table VI.III) are replaced with moments about zero. Table VI.VI represents the results of applying method one to data set three. The desired functions for a perfect fit to the analytic distribution are F3 and F5. Notice from Table VI.VI that only E4 includes F3 and F5. Figures 6.6 through 6.14 present a comparison of analytic, sample, and entropy densities and cumulatives. Notice again that for set E4 method one provides an exact fit to the analytic distribution (Figure 6.9 and 6.13). Sets E5 and E3 perform quite well even without the desired information functions. Table VI.VII provides a numerical evaluation of the entropy approximations.

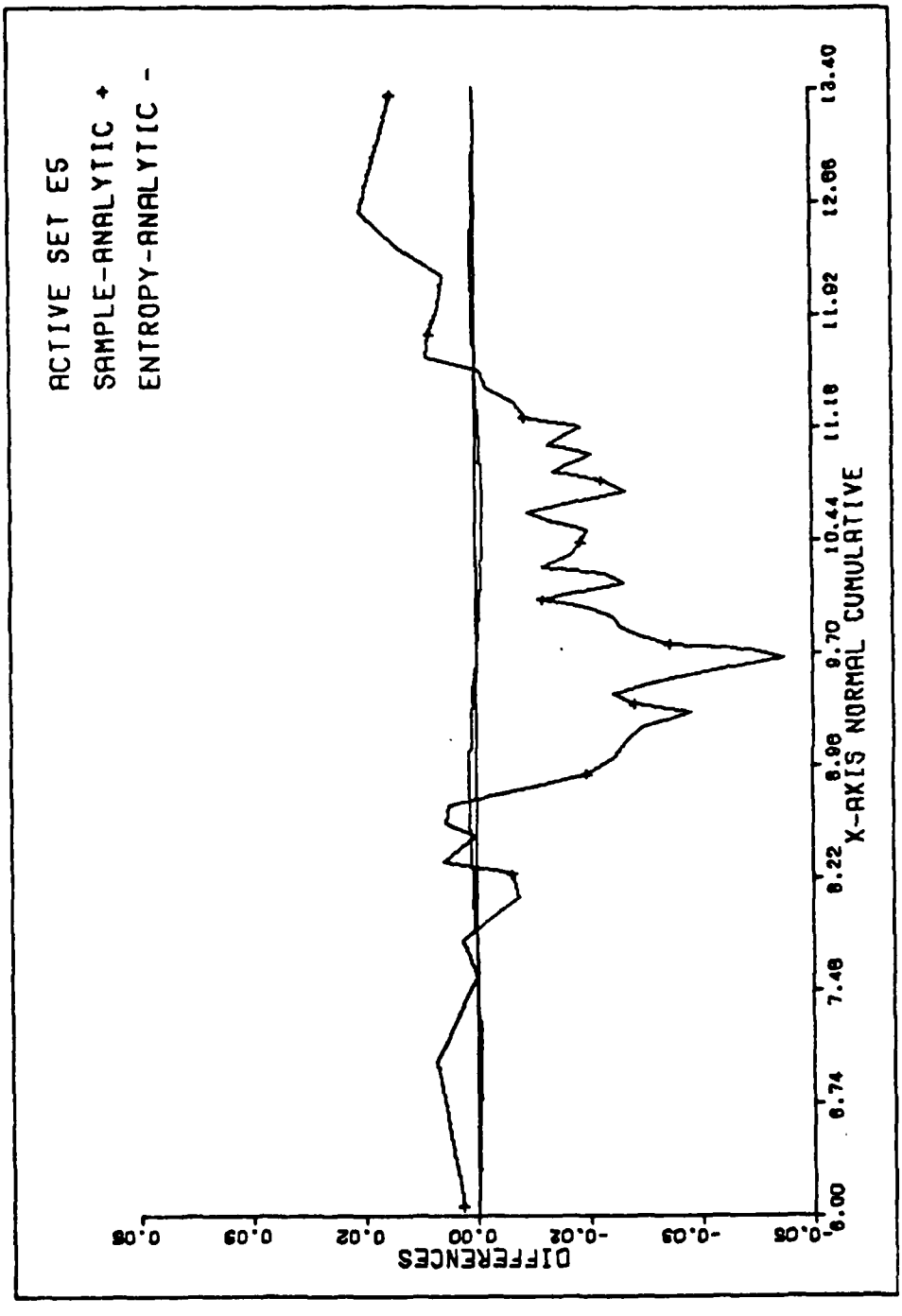


Fig. 6.4. Cumulative Differences--E5/Data Set 1

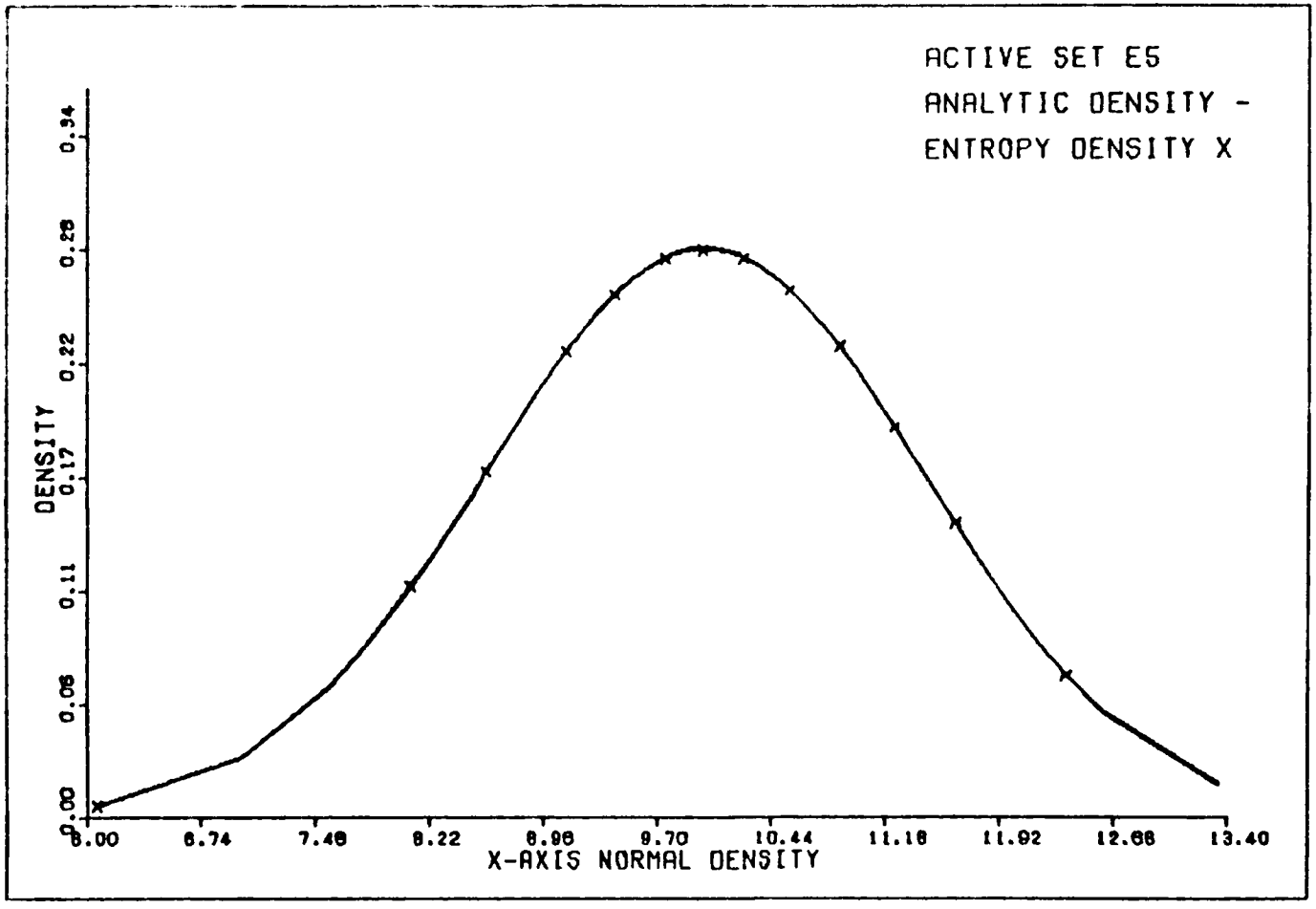


Fig. 6.5. Density Comparison--E5/Data Set 1

TABLE VI.V

NUMERICAL COMPARISON FOR NORMAL, DATA SET 1

	Max Error	Cumulative Max Error	Density Mean Error Sq.	Cumulative Mean Error Sq.
Analytic/Sample	.798 (-1)	.438 (-1)	.687 (-3)	.281 (-3)
E1 = F2				
Entropy/Sample	.798 (-1)	.438 (-1)	.687 (-3)	.281 (-3)
Entropy/Analytic	.533 (-14)	.142 (-13)	.429 (-29)	.271 (-28)
E2 through E4 are essentially equal to E1 thus not shown.				
E5 = F1,F5,F6,F7,F9				
Entropy/Sample	.814 (-1)	.439 (-1)	.693 (-3)	.277 (-3)
Entropy/Analytic	.165 (-2)	.919 (-3)	.104 (-5)	.333 (-6)
E1 through E5 for data set 2 are essentially equivalent to E1 above.				

TABLE VI.VI

REGRESSION RESULTS FOR BETA SAMPLE

Candidate Set	Information Functions	Adjusted R^2
E1	F6	56.74
E2	F7,F8	92.06
E3	F7,F8,F9	93.70
E4	F3,F5,F6,F8	93.99
E5	F1,F3,F7,F8,F9	94.70

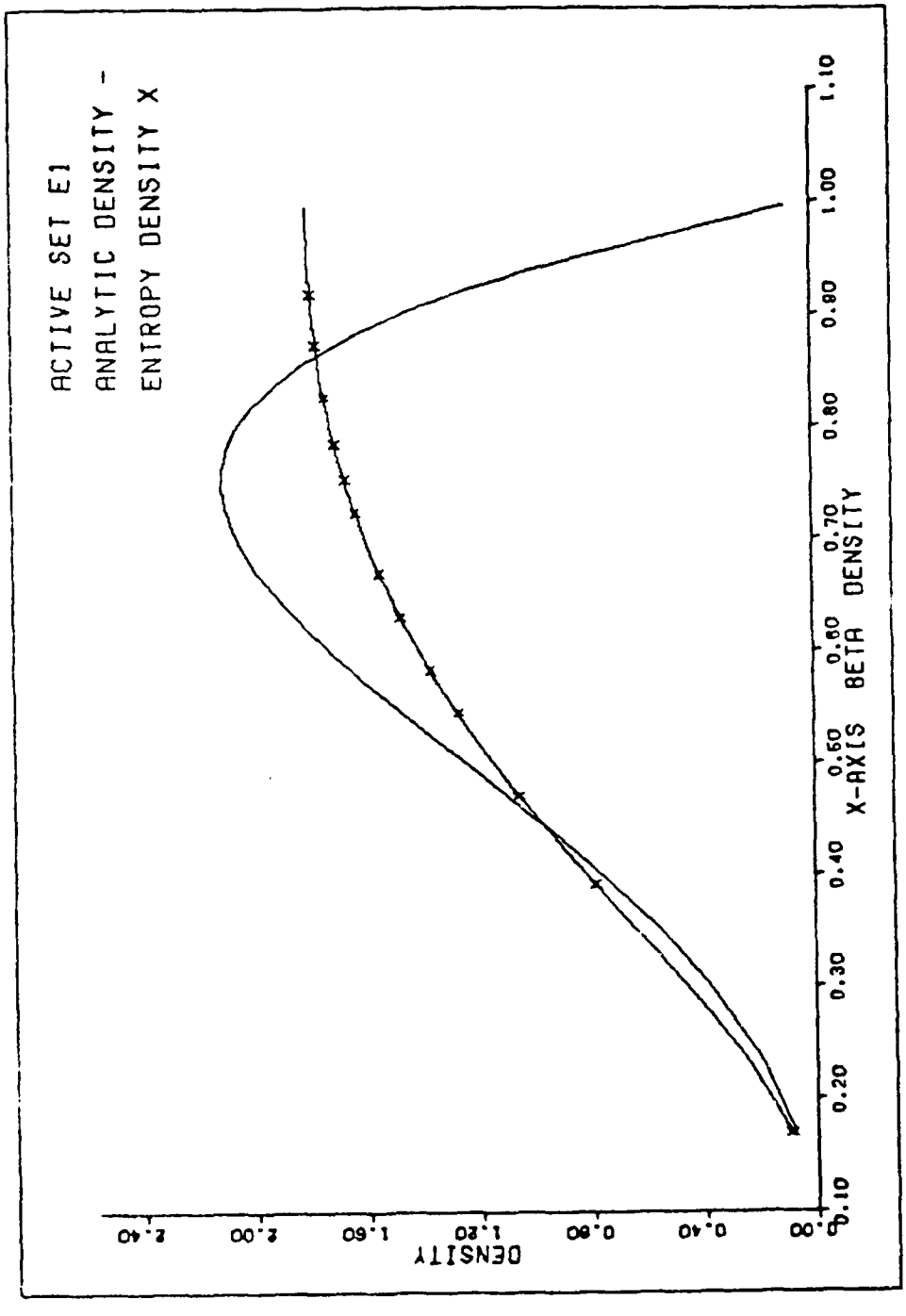


Fig. 6.6. Density Comparison--E1/Data Set 3

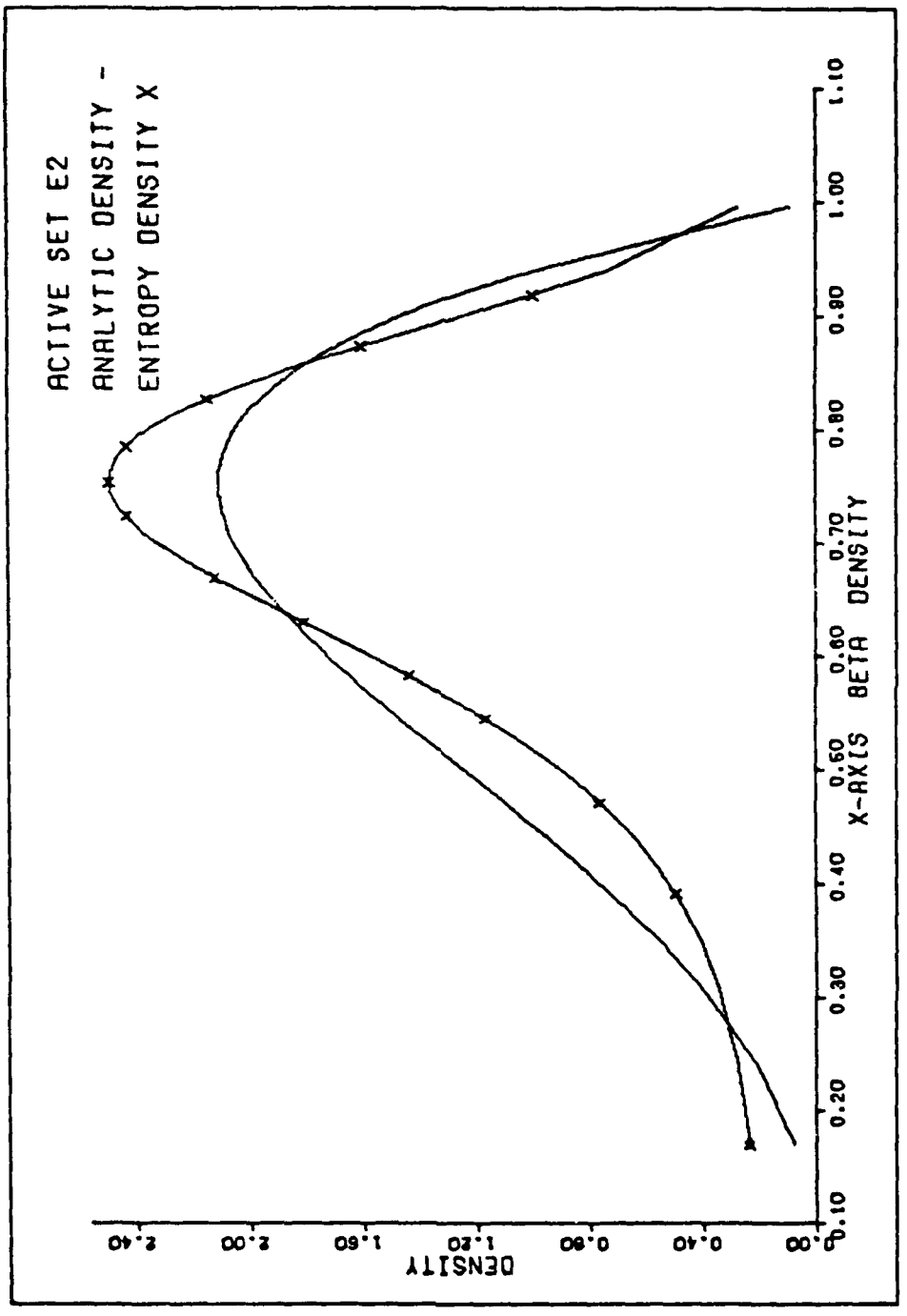


Fig. 6.7. Density Comparison--E2/Data Set 3

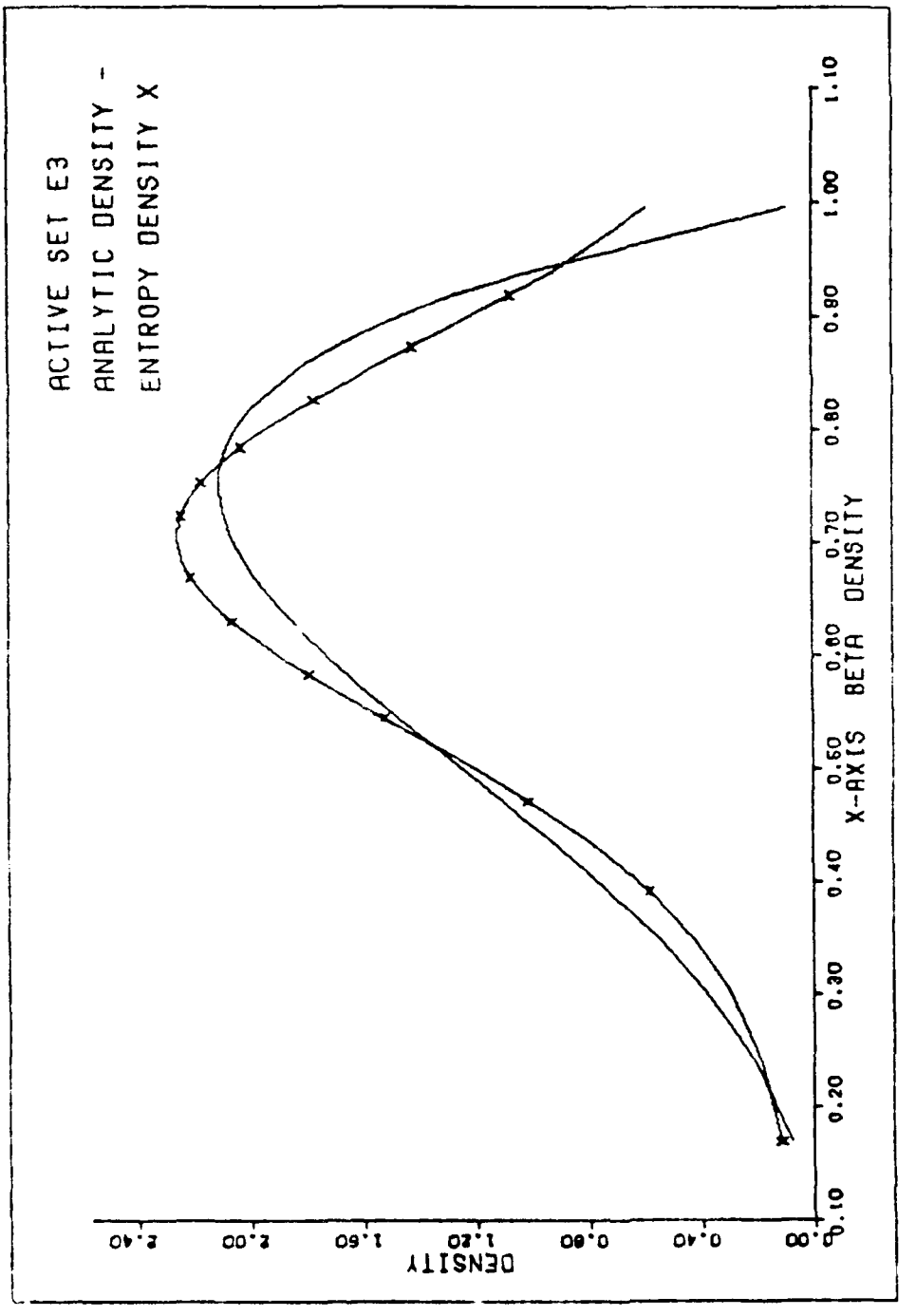


Fig. 6.8. Density Comparison--E3/Data Set 3

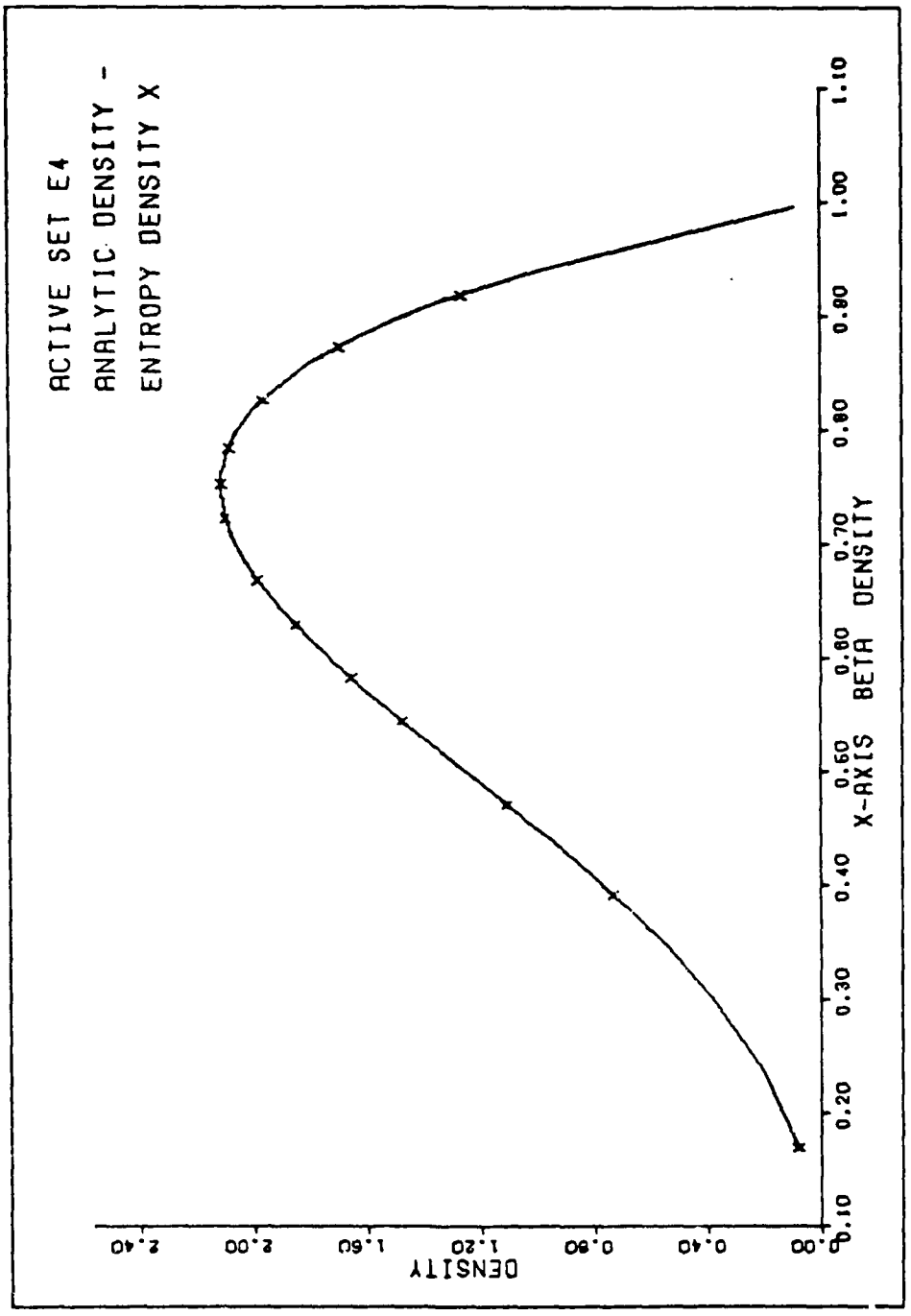


Fig. 6.9. Density Comparison--E4/Data Set 3

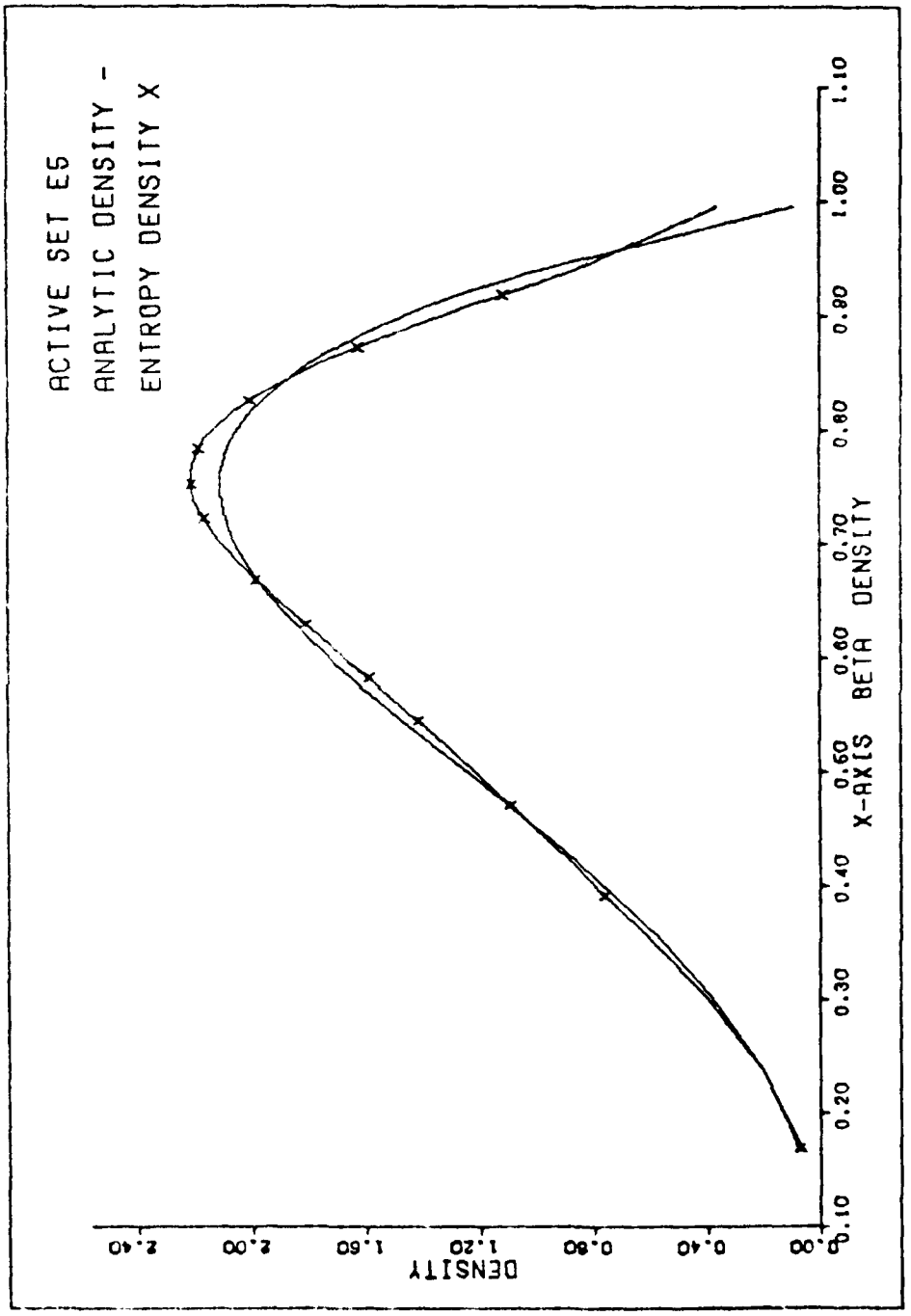


Fig. 6.10. Density Comparison--E5/Data Set 3

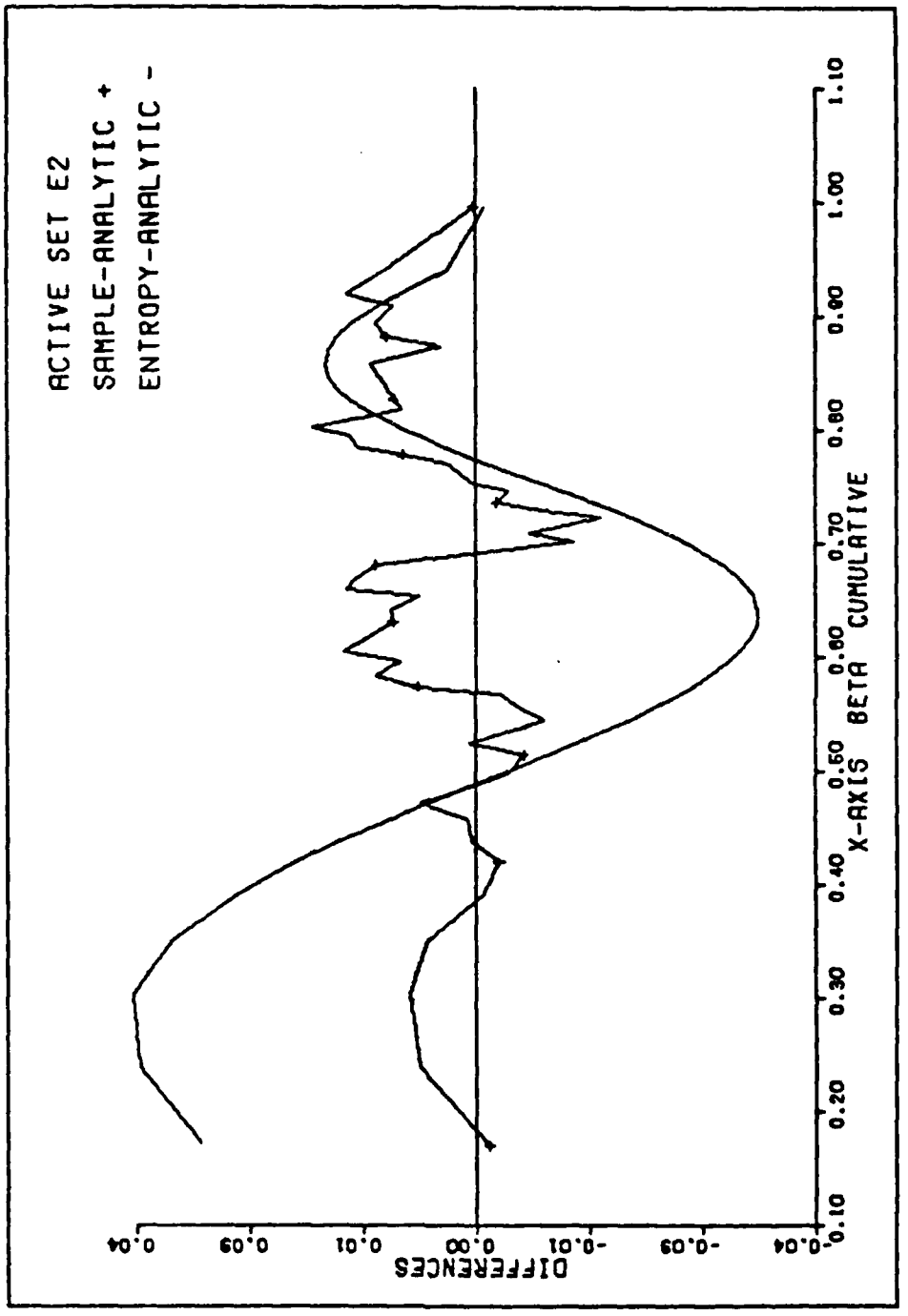


Fig. 6.11. Cumulative Differences--E2/Data Set 3

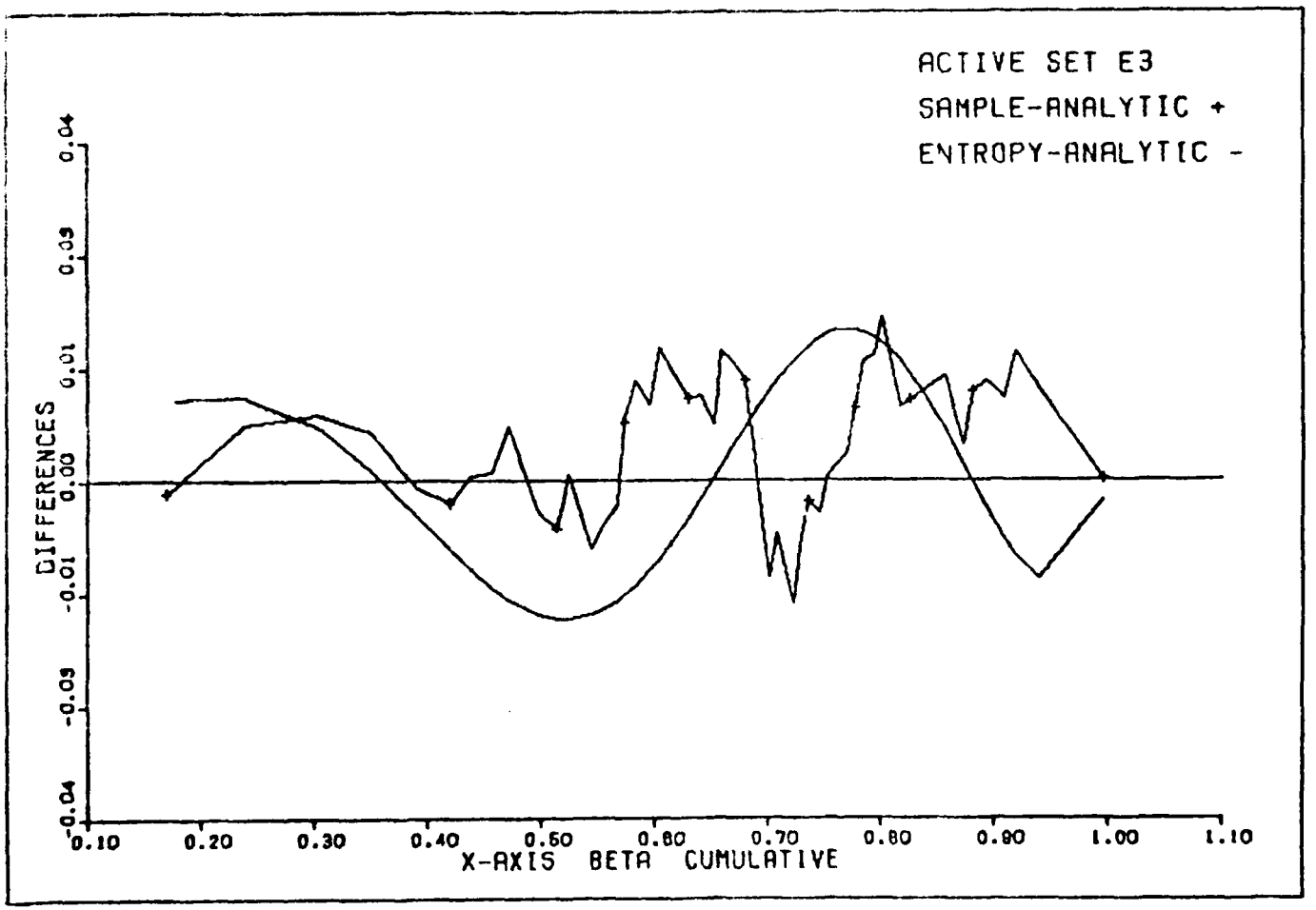


Fig. 6.12. Cumulative Differences--E3/Data Set 3

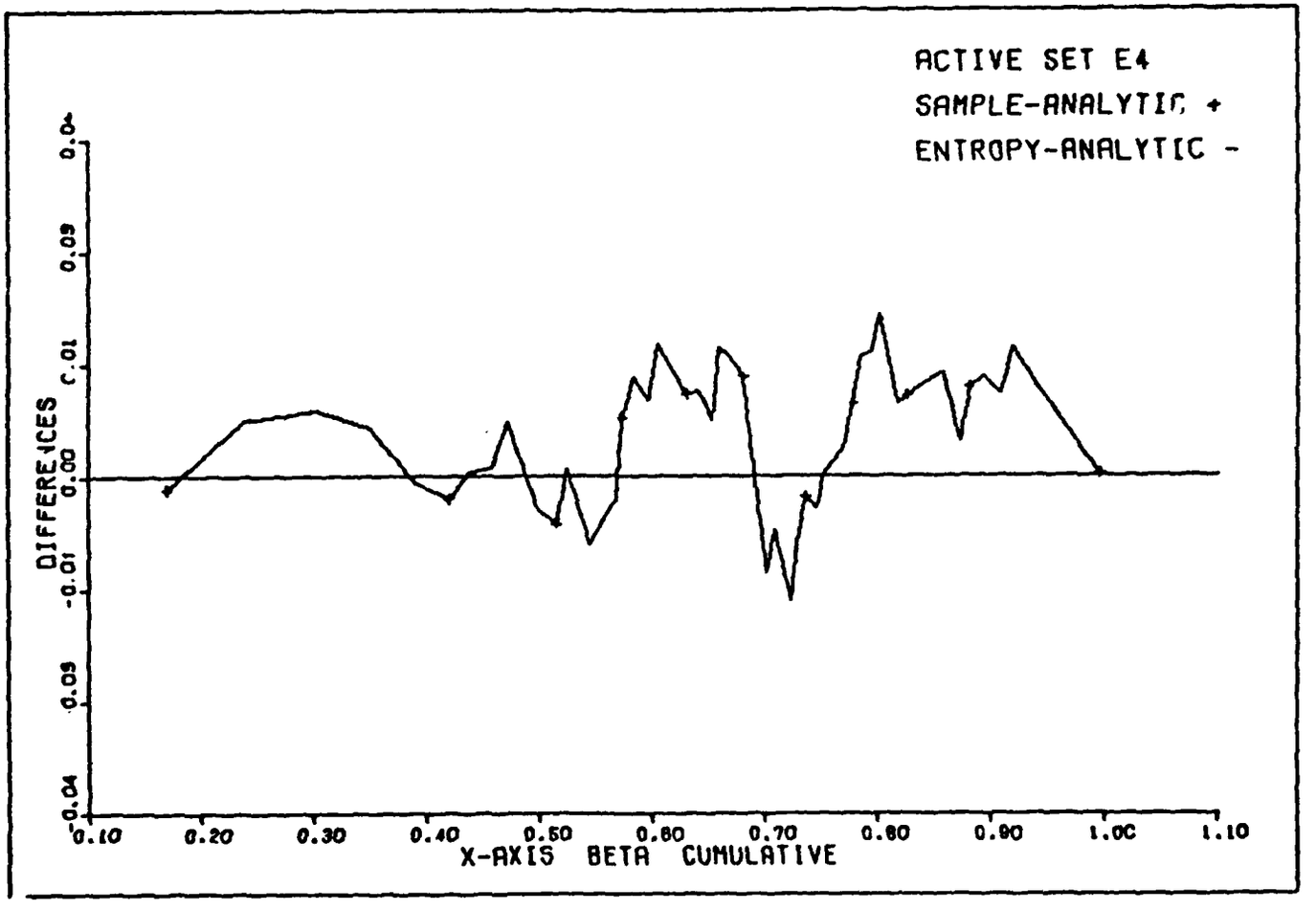


Fig. 6.13. Cumulative Differences--E4/Data Set 3

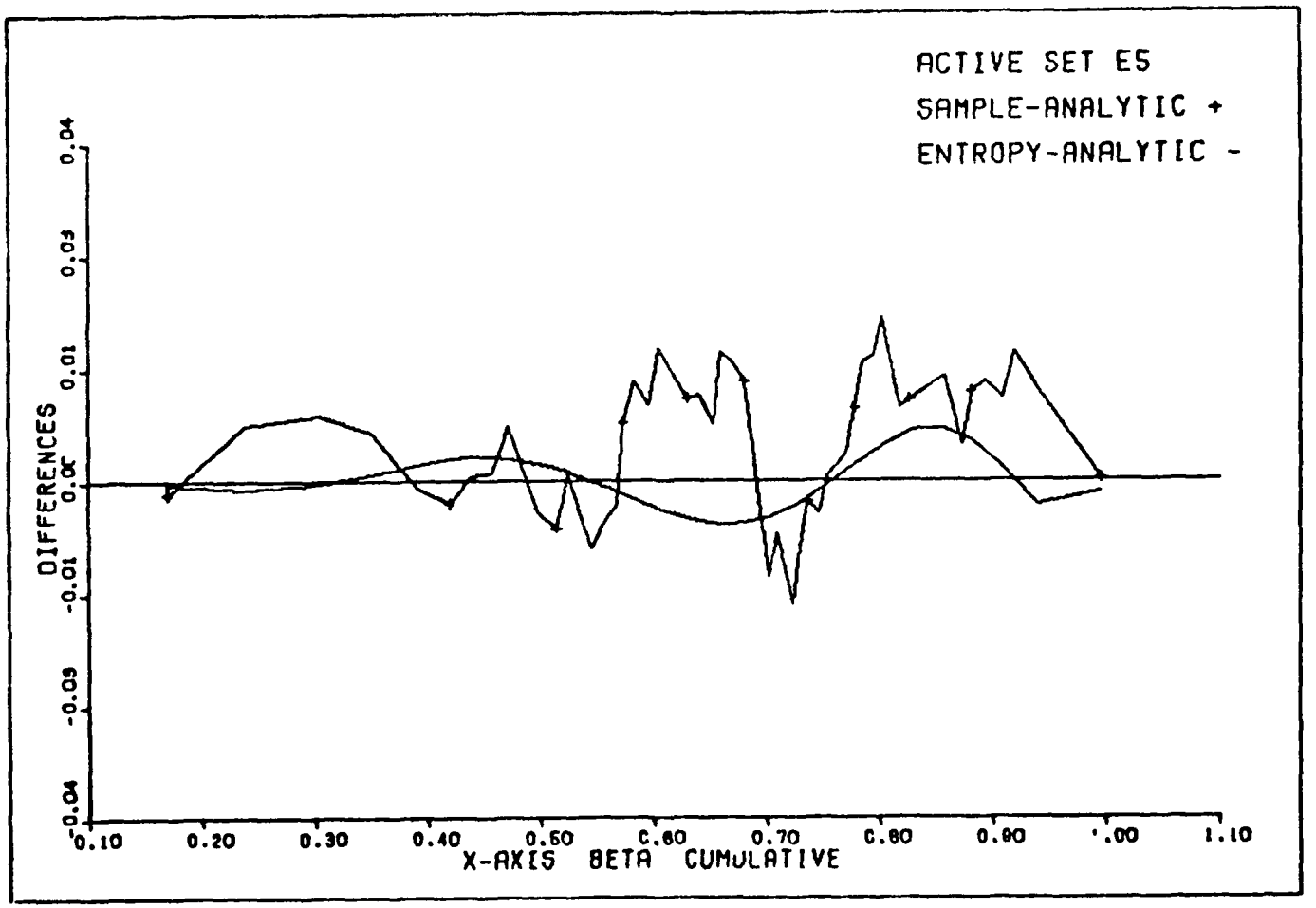


Fig. 6.14. Cumulative Differences--E5/Data Set 3

TABLE VI.VII

NUMERICAL COMPARISION FOR BETA, DATA SET 3

	Density Max Error	Cumulative Max Error	Density Mean Error Sq.	Cumulative Mean Error Sq.
Analytic/Sample	.4671	.2024 (-1)	.466 (-1)	.959 (-4)
E1 = F6				
Entropy/Sample	.144 (+1)	.1167	.242	.439 (-2)
Entropy/Analytic	.171 (+1)	.1037	.164	.362 (-2)
E2 = F7,F8				
Entropy/Sample	.706	.4995 (-1)	.791 (-1)	.605 (-3)
Entropy/Analytic	.387	.423 (-1)	.649 (-1)	.511 (-3)
E3 = F7,F8,F9				
Entropy/Sample	.617	.303 (-1)	.680 (-1)	.226 (-3)
Entropy/Analytic	.498	.187 (-1)	.295 (-1)	.153 (-3)
E4 = F3,F5,F6,F8				
Entropy/Sample	.4671	.202 (-1)	.466 (-1)	.959 (-4)
Entropy/Analytic	.1705 (-12)	.249 (-13)	.107 (-25)	.141 (-27)
E5 = F1,F3,F7,F8,F9				
Entropy/Sample	.529	.213 (-1)	.411 (-1)	.974 (-4)
Entropy/Analytic	.270	.647 (-2)	.596 (-2)	.123 (-4)

A review of Tables VI.V and VI.VII and previous figures will consolidate four significant points concerning method one:

1. The method is sensitive to sample data and may suggest candidate active sets that do not include the analytically correct functions. The reason is that the method provides a compromise between sample and analytic distributions and may require "other" functions to accomplish that compromise.

2. An acceptable approximation to the unknown distribution may be obtained even if the "correct" functions are not part of the active set. Set E5 for the beta example and set E5 for the normal demonstrate this quality.

3. Given accurate expected values, inclusion of the analytically correct functions in the active set will produce an exact fit.

4. Finally, the information functions from one sample will provide excellent approximations for subsequent samples. The two normal data sets exemplify this quality. Since our technique, in general, approximates the unknown analytic distribution, and the sample is an approximation to the analytic distribution, then one would expect an excellent fit to subsequent samples.

Before proceeding to final selection of the active set, we demonstrate the sensitivity of method one to errors in the calculation of expected values of information

functions. This subject is further pursued in Chapter X. These expected values determine the constraint equations and thus determine the final form of our approximation, p(x). Expected values may also be involved as parameters in the potential set such as μ and σ^2 in Table VI.III. The above examples used accurate expected values which were approximated by a 32 point quadrature formula. We might also approximate these values with averages from the random sample:

$$\langle g_j(x) \rangle \approx \sum_{i=1}^{500} g_j(x_i)/500.$$

Table VI.VIII lists the average and quadrature values for our three sample data sets.

As one would expect, use of averages in lieu of the more accurate quadrature values will produce a subsequent change in the entropy approximation. We demonstrate an interesting result by using the average values for the normal sample, data set one, to include mean and variance values in the information functions. The entropy procedure produced a less accurate fit to the analytic, as expected, but a more accurate approximation to the sample data. This is demonstrated in Figures 6.15 and 6.16 which graph sample, analytic, and entropy comparisons. Notice that the "entropy minus analytic" curve follows the trend of the "sample minus analytic" values. Comparison to Figures 6.3 and 6.4

TABLE VI.VIII

COMPARISON OF EXPECTED VALUE APPROXIMATION TECHNIQUES

	Normal			Beta	
<G>*	Quadrature	Averages Data Set 1	Averages Data Set 2	Quadrature	Averages Data Set 3
<F1>	-.01580	.02409	.00712	.666667	.66119
<F2>	.90808	1.0654	1.1813	-	.47188
<F3>	-	2.2947	2.2910	-.45000	-.46578
<F4>	1.2693	1.2520	1.3851	-.45000	-.46578
<F5>	1.0897	1.0470	1.2029	-1.2833	-1.3035
<F6>	1.8318	2.0222	2.3341	.30500	.34031
<F7>	-.10256	-.20452	-.23678	.357143	.35433
<F8>	2.2244	3.4442	4.7567	.277778	.27611
<F9>	4.5388	4.6001	4.5928	.377559	.37397

*F1,F2,F7,F8 are moments about zero for beta and are normalized moments about the mean for normal. Normal quadrature values represent integral of normal density over [a,b] and not over $[-\infty,\infty]$.

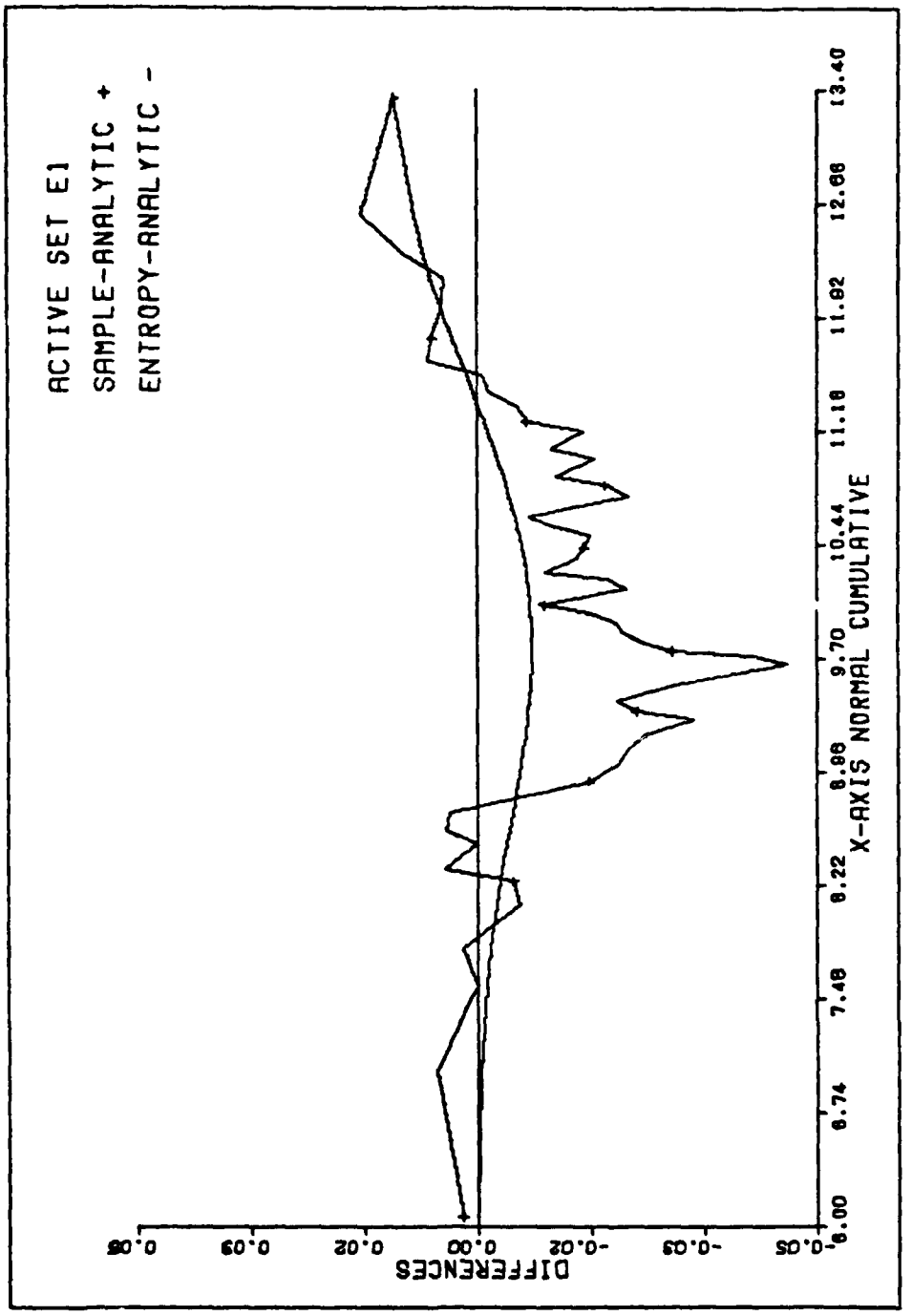


Fig. 6.15. Cumulative Differences--E1/Data Set 1 (Average Values)

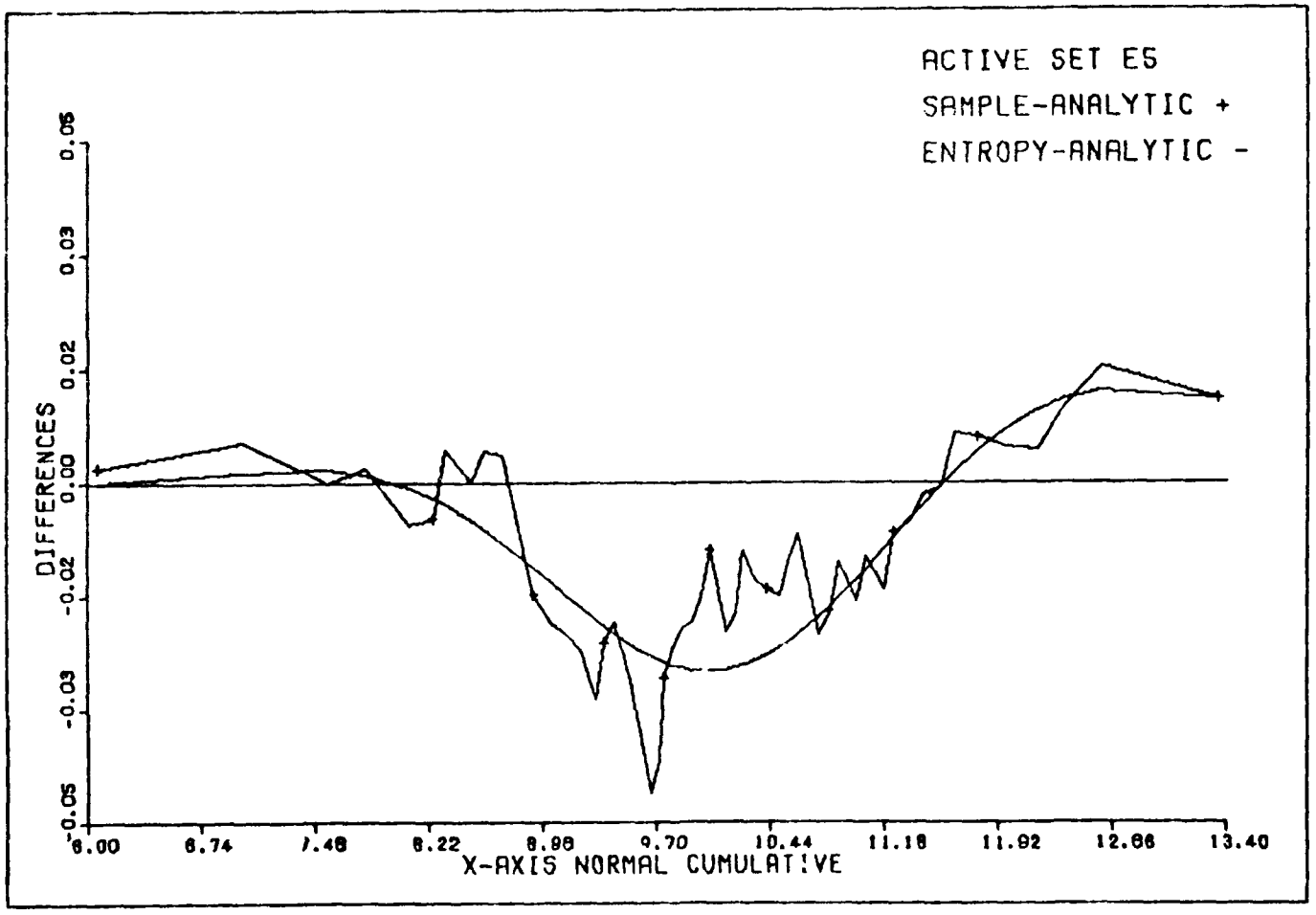


Fig. 6.16. Cumulative Differences--E5/Data Set 1 (Average Values)

(quadrature values) exemplifies this stronger correspondence to sample and weaker relationship to analytic. Similar results were obtained with beta, gamma, and simulation sample densities.

A word of caution is needed in reference to the figures. The figures were designed to accentuate the difference in distributions and not the closeness of approximation. Careful attention to the actual values of differences between distributions or review of Tables VI.V and VII will indicate that the entropy approximations are quite close to the sample and analytic distributions. In fact, the candidate sets provide such excellent approximations that the choice of the single best active set is difficult. Goodness of fit statistics are used in this final step of the procedure.

Goodness of Fit

Experimental results, in addition to the above examples, indicate that the regression procedure will produce adequate fits to the sample data and accurate, if not exact, fits to the underlying analytic distribution even when the selected information functions are not those expected. The question remains as to how one selects the best set of information functions from the several candidate sets. Since an accurate fit to the sample cumulative is desired, the active set will be the candidate set that

produces the smallest error between sample and entropy cumulatives. Again, the sample cumulative is available at N sample points, and the entropy cumulative may be calculated by integrating the entropy density.

Selection of a measure of error between sample and entropy cumulatives will impact selection of the active set. If we are only concerned with absolute error between the distributions, then we might use the errors squared measure of previous examples (Tables VI.V and VII). However, we would like to know more. Besides producing the "best" fit, we would like to know "how good" that fit is in a statistical sense. A goodness of fit statistic will provide this information. Step five of method one may now be stated explicitly:

5. Identify a goodness of fit statistic, SK, which is a function of sample and entropy distributions; calculate SK for each candidate set; select the set with minimum SK as the active set; and finally, specify the level of significance for the selected SK. We thus select the information function set that provides the best fit in the sense of our chosen statistic.

Several goodness of fit statistics are discussed in Appendix B. Each statistic has strengths and weaknesses as indicated in the appendix and references. Different statistics may result in different active sets. Two examples are presented in Table VI.IX for Anderson-Darling, A^2,

TABLE VI.IX

EXAMPLE STATISTICAL VALUES--ENTROPY VERSUS SAMPLE

Active Set	Anderson-Darling A^2	Kolmogorov-Smirnov D	Cramér-von Mises W^2	Cumulative Mean Error Sq.	Cumulative Max Error
Normal, Data Set 1					
E1=F2	.2077	.05086	.0158	.281 (-3)	.0438
E5=F1,F5,F6,F7,F9	.2067	.05093	.0155	.277 (-3)	.0439
E2 through E4 are equivalent to E1					
Beta, Data Set 3					
E1=F6	1.215	.1195	.2027	.440 (-2)	.1167
E2=F7,F8	.2463	.0613	.0328	.605 (-3)	.04995
E3=F7,F8,F9	.1242	.0413	.0161	.226 (-3)	.0303
E4=F3,F5,F6,F8	.1324	.0289	.00650	.959 (-4)	.0202
E5=F1,F3,F7,F8,F9	.1015	.0323	.00713	.974 (-4)	.0213
Critical Values					
α=.15	1.610	.0506	.2842	-	-
α=.10	1.933	.0545	.3471	-	-
α=.05	2.492	.0604	.4609	-	-
α=.01	3.857	.0724	.5806	-	-

NOTE: Critical values were calculated from table values in Ref 75.

Kolmogorov-Smirnov, D, and Cramér von Mises, W^2, statistics and the mean error squared, M^2. Consider the normal example of Table VI.IX. Since E1 through E4 produce identical results, our choice of active set is a decision between E1 and E5. Regardless of the analyst's choice of statistic, the statistical values for the two sets are very close. The analyst would probably select the smaller set, E1, in this case. Notice that use of the A^2 or W^2 statistics result in acceptance of the hypothesis of equal distributions at a critical value of $\alpha > .15$. The D statistic results in $\alpha > .15$.

The beta example of Table VI.IX better demonstrates the flexibility of method one and the importance of the choice of statistic. We see that E4 produces the smallest value of M^2, D, and W^2. This result is pleasing in that E4 contains the expected information functions F3=ln(x) and F5=ln(1-x). However, a concern for a fit to the tails of the unknown distribution may force the analyst to use the A^2 statistic (see Appendix B). Use of A^2 will result in selection of set E5. Notice that both E4 and E5 provide excellent approximations for all the listed measures.

Summary

Method one uses linear regression and "goodness of fit" principles to select the best active set for all possible combinations of the potential information functions.

The method involves five steps:

1. Generate the sample cumulative;

2. Obtain the sample density, f(x), via numerical differentiation;

3. Equate $\ln(f(x)) = \sum_{i=0}^{m} \lambda_i g_i(x)$ where m is the number of functions in the potential set;

4. Regress on the equation of step three to produce several candidate active sets;

5. Use statistical measures to select the best set.

Method one has been tested against various distributions in the normal, beta, and gamma families and against the simulation models of Chapter IX. An acceptable approximation resulted in every case using the potential set of Table VI.III. The method is based on proven analytic and statistical techniques, provides ample opportunity for analyst input and modification, and produces a compromise between the sample and the unknown analytic distributions. The excellent results of method one prompted further investigation of active set selection procedures. The next two chapters discuss alternate methods.

Chapter VII. Active Set Selection--
Method Two (Divergence)

Introduction

The linear regression approach of method one (Chapter VI) produces excellent results in selecting the "active" set of information functions, for a particular approximation, from the predefined "potential" set. Method one demonstrates strong sample dependence, sensitivity to numerical errors, and sensitivity to choice of goodness of fit statistic. The selected active set produces a distribution which adequately fits the sample and underlying analytic distributions but is generally a compromise between the two. Moreover, the active set need not include the desired analytic information functions. While a method which provides such an accurate approximation is certainly a useful tool, if the method can also identify the correct functions for the underlying analytic distribution then usefulness is increased. Additionally, our entropy approximation procedure is based on information theoretic concepts. A desire to improve method one while adhering to our information theoretic theme led to the information divergence measure and the development of method two. We discuss divergence and present method two with results.

Divergence

Experimentation with the regression method shows that different goodness of fit measures may result in different active sets. The goodness of fit measures test our fit to the sample data, i.e., to the sample cumulative or density. Although our goal is an accurate approximation to the sample, we now shift our concern to accurately approximating the "information content" of the data. Thus, we are changing our measure of fit to the data.

The Kullback-Leibler measure of information variation (Refs 50; 51) measures the information exchange when one probability measure is replaced by a second probability measure. As defined in Chapters II and IV, the information variation, i.e., the loss or gain of information, which occurs when density f(x) is replaced by density p(x) is defined

$$I(p(x),f(x)) = \int p(x) \ln [p(x)/f(x)] \, dx$$

Kullback, Guiasu, Jeffreys (Refs 51; 33; 46) and others discuss information variation and its properties.

Information variation seems like the perfect conceptual measure. We would like to minimize the information loss when the sample density, f(x), is replaced by the entropy density, p(x); thus we select the entropy density which minimizes I(p(x),f(x)). Unfortunately, information variation is not commutative, i.e., $I(p(x),f(x)) \neq I(f(x),p(x))$,

and $I(p(x),f(x))$ may be negative. We know from Guiasu or Kullback (Ref Chapter IV) that

$$I(p(x),f(x)) \geq \int_E p(x)\,dx \ln\left[\frac{\int_E p(x)\,dx}{\int_E f(x)\,dx}\right] \qquad (7.1)$$

where E is the interval of comparison. If E is a subset of the interval of definition for $p(x)$ then $\int_E p(x)\,dx$ may be less than one. Thus, the right-hand side of equation (7.1) may be negative, and $I(p(x),f(x))$ may be negative. Our interval of comparison will be dictated by the random sample, i.e., $E=[x_1,x_N]$, and may force negative values for $I(p(x),f(x))$. Kullback and Jeffreys extend $I(p(x),f(x))$ to a more usable measure, divergence, which retains the conceptual strength of information variation.

Kullback defines divergence, $J(p(x),f(x))$, as a measure of the difficulty of discriminating between two densities where

$$\begin{aligned} J(p(x),f(x)) &= I(p(x),f(x)) + I(f(x),p(x)) \\ &= \int[p(x)-f(x)]\ln[p(x)/f(x)]\,dx \\ &= \int[f(x)-p(x)]\ln[f(x)/p(x)]\,dx \\ &= J(f(x),p(x)) \end{aligned}$$

As Kullback points out, $J(p(x),f(x))\geq 0$ with equality if and only if $p(x)=f(x)$ almost everywhere (a.e.). A simple application of equation (7.1) will show that $J(p(x),f(x))\geq 0$

regardless of our interval of comparison. Thus, divergence offers a viable means of comparing information loss when using an entropy approximation in place of a sample. The sample density is available at M points, x_i, i=1,2,...M, and we can approximate f(x) at point y by using linear interpolation between $f(x_j)$ and $f(x_{j+1})$ where $x_j \leq y \leq x_{j+1}$. The entropy approximation is available in algebraic form and, thus, J(p(x),f(x)) can be calculated via numerical quadrature. We seek the minimum set of information functions which defines the density p(x) with minimum divergence from f(x). This set will be the active set for the given data.

Selection Procedure

The active set selection procedure of method two is analogous to multiple nonlinear regression but uses the divergence measure. The procedure includes two phases, a function addition phase (analogous to forward regression) and a function deletion phase (akin to backward regression). The procedural steps follow:

1. Generate the sample cumulative distribution at M points as in method one.

2. Produce the sample density, f(x), at the M points by numerical differentiation as in method one.

3. Produce expected values of all information functions in the potential set via quadrature or average values

as in method one. Steps 1, 2 and 3 concern data preparation while steps 4 through 6 describe the iterative procedure.

4. Use expected values in the appropriate constraint equations to find the Lagrange multipliers, i.e., the Λ vector, for the m entropy densities

$$p_j(x) = \exp\,[-\lambda_0-\lambda_j g_j(x)],\ j=1,2,\ldots m,$$

where m is the number of functions in the potential set. Each $p_j(x)$ thus contains one information function.

5. Find the value of j such that $J(p_j(x),f(x))$ is a minimum and let this value equal h. Function $g_h(x)$ is now considered a member of the active set which defines the final entropy approximation, $p(x)$. If $J(p(x),f(x))\leq EPS$ where EPS is a predefined stopping criteria, then the "function addition" phase of method two is complete. If $J(p(x),f(x))>EPS$ then function $g_h(x)$ is retained in the active set, and we iterate to find the next best function to add to the active set. Steps 4 and 5 are reaccomplished for the m-1 entropy densities $p_j(x)$ where

$$p_j(x) = \exp\,[-\lambda_0-\lambda_h g_h(x)-\lambda_j g_j(x)],\ j=1,2,\ldots m,$$

$j\neq h$. This procedure continues until a maximum of K functions (we use K=6) are active, until $J(p(x),f(x))\leq EPS$, or

until the addition of a function increases J(p(x),f(x)). We thus select the active set with minimum J(p(x),f(x)).

6. The final procedural step, the "function deletion" phase, checks for redundant information. Let k be the number of active information functions, i.e., the functions defining p(x), and consider the possibility that we have selected too many functions. Let q(x) be p(x) with one of the active functions removed. If J(p(x),q(x)) is close to zero, where "close" is defined by the user, then we lose very little information in dropping the subject function. For small divergence, we replace p(x) with q(x) and iterate to test each function for deletion. This removal step was implemented as a separate subroutine (THROUT) because it is used in method three and may be applied to method one for better results. THROUT contributed appreciably to the excellent results obtained with method two.

Results

Method two uses an approach which is conceptually similar to method one but with a different measure of fit, i.e., divergence. Divergence is an accepted information measure which evenly weights the data points and follows the information theoretic thrust of the dissertation. One must notice, however, that the selected active set is not guaranteed to be the single best set in the divergence

sense. This is not of great concern because the selected set must produce an acceptably small divergence and thus produce an acceptable fit to the sample. Method two may reach a "small" divergence before considering the single best set of functions. Method one has a slight advantage in this respect because method one uses "leaps and bounds" linear regression which considers more possible combinations of functions. With either method, the only way to ensure that the single best set is selected is to check all possible combinations of potential functions. Such a procedure is not practical, and results with both methods indicate that such a procedure is not necessary.

Table VII.I compares the results for methods one and two when applied to two sample densities from Chapter VI (normal and beta) and a third sample from a gamma distribution. The function symbols are defined in Table VI.III of the last chapter. Method two selected the correct analytic functions for the normal and gamma samples. An excellent approximation was produced for the beta although the desired analytic functions were not selected. Method one results were statistic dependent but were generally more oriented to the sample distribution. Additional samples from the same distribution families were tested with similar, although not exact, results. In all tests, if either method chose functions other than the analytic functions, the resulting approximations were still

TABLE VII.I

TEST COMPARISONS OF METHODS ONE AND TWO

Bounds	Distribution	Analytic Functions	Method 1 Anderson-Darling	Method 1 Kolmogorov-Smirnov	Method 2 Divergence
[μ-4σ, μ+4σ]	Normal μ=10.0 σ=$\sqrt{2}$	F2	F1,F5,F6,F7,F9	F2	F2
[0,1]	Beta P=4, Q=2	F3,F5	F1,F3,F7,F8,F9	F3,F5,F6,F8	F2,F5,F9
[0,17]	Gamma G=1.5, B=2	F1,F3	-	F2,F5,F6,F8	F1,F3

NOTES:

Function symbols defined in Table VI.III.

Gamma results selected on measure of maximum error between entropy and sample cumulative and not via one of the stated statistics. G=shape factor and B=scale factor.

exceptionally close to sample and analytic distributions. Of course when the analytic functions were selected, the fit to the analytic was exact. A fourth instructive comparison of the methods is made in Chapter IX for a simulation application.

Method two requires specification of the expected values for the potential information functions and definition of the approximation bounds, i.e., the interval over which the entropy approximation will apply. For the given examples, expected values were calculated via quadrature. The bounds are specified by the analyst and may be based on knowledge of the unknown distribution, quadrature results, simulation results, or the random sample. The analyst usually acquires such bounds in generation of the expected values. The interval $[x_1, x_N]$ from the sample will suffice if the expected (or average) values are calculated on this interval.

The iterations of method two for the normal sample are shown in Table VII.II. The first three iterations of Table VII.II represent the function addition phase which selects F2,F6,F1. We stop at this point because the divergence in adding F1 changes very little from the previous iterations, and, in fact, increases slightly. The analyst may have preferred to stop at F2,F6. Notice that in the function additions phase we are comparing entropy to sample densities. The entropy densities are calculated on the

TABLE VII.II

DIVERGENCE METHOD APPLIED TO NORMAL SAMPLE

A. Function Addition on $[x_1, x_N] = [6.07, 13.35]$

Iteration	Function Set	Divergence
1	F2	$J(p_2(x), f(x)) = .02257350$
2	F2,F6	$J(p_{26}(x), f(x)) = .02252025$
3	F2,F6,F1	$J(p_{261}(x), f(x)) = .02252026$

B. Function Deletion on $[\mu-4\sigma, \mu+4\sigma] = [4.34, 15.66]$

Iteration	Divergence	Action
4	$J(p_{261}(x), p_{61}(x)) = 8.01$	Retain F2
5	$J(p_{261}(x), p_{21}(x)) = 1$ (-20)	Delete F6
6	$J(p_{261}(x), p_2(x)) = 2$ (-7)	Delete F1

C. Active Set = F2

NOTE: Function symbols defined in Table VI.III.

interval $[\mu-4\sigma,\mu+4\sigma]$ whereas the sample is defined on the subset $[x_1,x_N]$; the divergence comparison is thus made on the smaller bound. We return to $[\mu-4\sigma,\mu+4\sigma]$ for function deletion (subroutine THROUT) which compares entropy densities. The final active set is F2 as desired.

Experimentation, as exemplified in Table VII.I, indicates that the divergence approach is more likely to correctly identify the analytic functions and is less sample sensitive than method one. However, method two is still sample sensitive as seen with the beta approximation. Method two selected functions F2,F5,F9 to produce a divergence $J(f(x),p(x))=.0223359$. The correct analytic functions are F3 and F5. We calculated the divergence between the analytic density, $q(x)$, and the sample, $f(x)$ to find $J(f(x),q(x))=.0249148$. Thus method two chose functions that provides a closer fit to the data, in the divergence sense, than if the correct analytic functions had been chosen. Table VII.III provides actual values of the respective densities at 17 of the 32 quadrature points used in the beta example. Sampling error and numerical differentiation account for the discrepancy between sample and analytic values.

Methods one and two are quite similar in structure and results. The nonlinear regression approach of method two appears to be less sample sensitive than method one, i.e., method two is more likely to identify the underlying

TABLE VII.III

SAMPLE, ENTROPY, AND ANALYTIC COMPARISON FOR BETA SAMPLE

I	X(I)	Sample Density	Method 2 Entropy Density	Analytic Density
1	.17069	.26205	.10144	.08248
2	.18411	.26672	.11593	.10183
3	.21237	.27656	.15229	.15088
4	.25442	.30384	.22379	.24558
5	.30870	.36929	.35396	.40673
6	.37318	.51643	.57543	.65151
7	.44545	.97566	.91564	.98034
8	.52284	1.30094	1.36148	1.36394
9	.60244	1.80312	1.82053	1.73851
10	.68131	1.61865	2.13328	2.01572
11	.75649	2.43046	2.16621	2.10842
12	.82519	1.94359	1.91065	1.96454
13	.88484	1.62170	1.47946	1.59560
14	.93323	.75704	1.01420	1.08535
15	.96885	.44177	.60802	.57145
16	.98949	.38228	.30344	.20361
17	.99430	.36861	.20920	.11200

analytic distribution. Method two is less cumbersome to use. However, method one may be preferred when a test of hypotheses, or a confidence bound, about the accuracy of the entropy approximation is desired. Method one offers flexibility in choice of statistic and is a more traditional approach. The key point is that both methods will produce excellent approximations, given a workable potential set. The choice of method is at the analyst's discretion.

The methods produce excellent results but share a common disadvantage. Both methods require a random sample of the unknown distribution and both involve numerical differentiation. Sampling and differentiation errors are two reasons for failing to explicitly identify the underlying analytic distribution. Method three provides a viable alternative which avoids these error sources.

Chapter VIII. Active Set Selection-- Method Three (Expected Values)

Introduction

This chapter presents a third method for selecting the best active set of information functions from a predefined potential set. Methods one and two, though effective, are subject to sampling and numerical differentiation errors. Method one (linear regression) produces an approximation, $p(x)$, that compromises between sample and analytic distributions with a tendency to match the sample. Method two (divergence) approximations produce similar compromises but with a strong tendency toward the underlying analytic. Method three (expected values) concentrates on the underlying analytic distribution from which the sample is generated. Method three, like methods one and two, requires the expected values of all information functions in the potential set and definition of the interval of approximation, $[a,b]$. However, method three does not use a sample cumulative or density and consequently, is faster and less complicated than previous methods.

The expected values method is based on the premise that the expected values of the potential information functions communicate sufficient information to accurately approximate the unknown distribution. Let $f(x)$ represent

the unknown underlying, analytic density, and let $\langle G \rangle_m$ be the vector of expected values for the m potential functions where $\langle G \rangle_m = [\langle g_1 \rangle, \langle g_2 \rangle, \ldots \langle g_m \rangle]^T$ and $\langle g_j \rangle = \int_a^b g_j(x) f(x)\, dx$. Our information, the $\langle G \rangle_m$ vector, is generated by the unknown analytic distribution, i.e., via quadrature, simulation or averages since f(x) is unknown. An accurate entropy approximation to f(x) must generate an accurate approximation to $\langle G \rangle_m$. For example, if the entire potential set is included in p(x), i.e.,

$$p(x) = \exp\left[-\lambda_0 - \lambda_1 g_1(x) - \ldots - \lambda_m g_m(x)\right],$$

then p(x) will generate $\langle G \rangle_m$ exactly. Now assume that f(x) is a normal distribution. We know that

$$p(x) = \exp\left[-\lambda_0 - \lambda_1 g_1(x) - \lambda_2 g_2(x)\right] = \exp\left[-\lambda_0 - \lambda_1 x - \lambda_2 x^2\right]$$

is the unique entropy characterization of the normal, and p(x) will generate the same $\langle G \rangle_m$ vector as f(x). In this normal example, $\langle g_3 \rangle$ through $\langle g_m \rangle$ represent redundant information. Jaynes states (Ref 45) and experimentation confirms that redundant information is eliminated from the entropy density, i.e., solution of the m constraint equations in our normal example will result in $\lambda_3 = \lambda_4 = \ldots \lambda_m = 0$. Such a result is predicted by our uniqueness theorems of Chapter IV.

We thus define the active set of information functions to be the minimum set of potential functions that

acceptably reproduces $\langle G \rangle_m$. This approach again emphasizes the importance of defining a large, flexible potential set, as discussed in Chapter V, so that sufficient information about the unknown density is communicated. Due to numerical difficulties, we cannot in general solve the m nonlinear constraints, for a large m, to find the unique λ_j's, j=0,1,...m. Consequently, method three builds an active set by progressively fitting the $\langle G \rangle_m$ vector and then checking for redundant functions. The approach is similar to the regression tactics of previous chapters.

Selection Procedure

Method three is an iterative procedure which we decompose into the following steps:

1. Specify [a,b] and calculate the expected value vector, $\langle G \rangle_m$, for the m dimensional potential set. The $\langle G \rangle_m$ vector is part of the assumed or "given" data. As in previous methods, we include data collection as a procedural step.

2. Use the expected values in the appropriate constraint equations to find the m entropy densities

$$p_j(x) = \exp [-\lambda_0 - \lambda_j g_j(x)], \quad j=1,2,...m,$$

where m is the number of functions in the potential set. Thus, each $p_j(x)$ contains one information function on the first iteration.

3. Use each density in step two to produce $\langle\tilde{G}\rangle_m$, an estimate to $\langle G\rangle_m$, and measure estimation error. For each entropy density we use quadrature to generate $\langle\tilde{G}\rangle_m$ where $\langle\tilde{G}\rangle_m=[\langle\tilde{g}_1\rangle,\langle\tilde{g}_2\rangle,\ldots\langle\tilde{g}_m\rangle]^T$ and $\langle\tilde{g}_k\rangle=\int_a^b g_k(x)p_j(x)\,dx$. We then calculate the error of estimation, M_j^2, where M_j^2 is the sum of errors squared;

$$M_j^2 = \sum_{i=1}^{m} (\langle g_i\rangle-\langle\tilde{g}_i\rangle)^2,\ j=1,2,\ldots m.$$

4. Select the information function which induces the best approximation to $\langle G\rangle_m$, i.e., pick the minimum M_j^2. This information function becomes part of the active set and thus part of the final approximation, p(x).

5. Check stopping criteria. If $M^2\leq$EPS, where EPS is a predefined stopping value, then we have defined an effective active set and may proceed to step six. If $M^2>$EPS then we iterate to find the next best function to add to the active set. For the second iteration, steps 2, 3, 4, and 5 are repeated for m-1 entropy densities where $p_j(x) = \exp[-\lambda_0-\lambda_h g_h(x)-\lambda_j g_j(x)]$, $j=1,2,\ldots m$, $j\neq h$, where h is the active information function. This procedure continues for m-2 densities, m-3 densities, etc.; that is, the active set grows until $M^2\leq$EPS or until K functions are active (we use K=6).

If we exceed K functions without producing a sufficiently small M^2 then we have reached an error condition.

We must assume an insufficient potential set and consider additional potential functions. Collection of a random sample to produce a frequency histogram or a numerical approximation to the unknown density may provide insight at this point. An example of an insufficient potential set was given in Chapter V for a distribution that resembled the double exponential. Again, the potential set of Chapter V should provide sufficient information for many characterizations on [a,b] as the results section of this chapter will show. Once we obtain a sufficiently small M^2, we consider the elimination of unnecessary functions.

6. Eliminate redundant information functions, i.e., apply subroutine THROUT of Chapter VII. THROUT eliminates functions from the active set, one function at a time, and evaluates the divergence between p(x) with the active set and p(x) with one less function. If the divergence is near zero ("near" is defined by the analyst) then the subject function may be deleted from the active set. This function removal step is repeated until one complete pass through the active set is accomplished without a function deletion. Active set selection is then complete.

Results

Method three was tested by generating $\langle G \rangle_m$ vectors for known distributions, producing the entropy approximation, and comparing the two densities. Method three

consistently identified the desired analytic functions, when such functions were elements of the potential set, and exactly characterized the analytic densities. The potential set of Chapters V, VI, and VII was used for our testing and is repeated in Table VIII.I for convenience. Table VIII.II presents representative test results. The normal, beta (skewed right) and gamma (skewed left) distributions of previous chapters are shown as well as six additional distributions. A tenth example is given in Chapter IX for a simulation output distribution. Graphs are not shown for most of the Table VIII.II distributions because the approximation errors are very small, i.e., $\sup_x |p(x)-f(x)| < 10^{-4}$. We discuss the examples to demonstrate the strength of method three.

TABLE VIII.I

POTENTIAL INFORMATION FUNCTIONS

Symbol	Information Function	Symbol	Information Function	Symbol	Information Function
F1	$(x-\mu)/\sigma$	F4	$\ln(x-a)$	F7	$((x-\mu)/\sigma)^3$
F2	$((x-\mu)/\sigma)^2$	F5	$\ln(b-x)$	F8	$((x-\mu)/\sigma)^4$
F3	$\ln x$	F6	$(\ln(x-a))^2$	F9	$\ln(x^2+1)$

NOTE: μ = mean; σ = standard deviation; and [a,b] = bounds.

TABLE VIII.II

ACTIVE SET SELECTION RESULTS

No.	Interval	Distribution	Analytic Functions	Method 1 Set	Method 2 Set	Method 3 Set
1	$\mu-4\sigma$, $\mu+4\sigma$	Normal $\mu=10$, $\sigma^2=2$	F2	F2	F2	F2
2	0., 1.0	Beta P=4, Q=2	F3,F5	F3,F5,F6,F8	F2,F5,F9	F3,F5
3	0,17.0	Gamma B=2, G=1.5	F1,F3	F2,F5,F6,F8	F1,F3	F1,F3
4	$\mu-4\sigma$, $\mu+4\sigma$	Normal $\mu=20$, $\sigma^2=4$	F2	-	-	F2
5	$\mu-4\sigma$, $\mu+4\sigma$	Normal $\mu=9$, $\sigma^2=3$	F2	-	-	F2
6	0., 20.	Exponential B=.6143	F1	-	-	F1
7	$1/\alpha$, α	Hyperbolic $\alpha=6.0$	F3	-	-	F3
8	a,b	Uniform a=5., b=15.	Constant	-	-	Constant
9	0, 2	Bimodal (Fig. 8.1)	Unknown	-	-	F2,F6,F7,F8

NOTE: Method 1 results using Kolmogorov-Smirnov statistic (Chapter VI).

The beta distribution provides an interesting example. The beta density has the following two parameter form:

$$f_b(x) = [C/(b-a)^{P+Q-1}](x-a)^{P-1}(b-x)^{Q-1}, \quad a \leq x \leq b$$

where $C = \Gamma(P+Q)/(\Gamma(P)\Gamma(Q))$ and $\Gamma(\cdot)$ is the gamma function. Substitution of P=4.0, Q=2.0, and [a,b]=[0,1] in $f_b(x)$ produces $f_b(x) = 20.0\ x^3(1-x)$. Table VIII.III displays the iterations of method three in selecting the active set F3,F5, i.e., ln(x) and ln(1-x). Notice that functions F8 and F2 were introduced and ultimately eliminated. The final entropy density from Table VIII.III is

$$\begin{aligned} p(x) &= \exp[-\lambda_0 - \lambda_3 \ln(x) - \lambda_5 \ln(x-b)] \\ &= \exp[2.9957 + 3\ln(x) + \ln(x-1)] \\ &= \exp[2.9957]\ x^3(x-1) \\ &= 19.999999\ x^3(x-1) \end{aligned}$$

which we round to

$$p(x) = 20.0\ x^3(x-1) = f_b(x) \text{ exactly.}$$

The normal and gamma examples produce similar results. The gamma density with shape factor G and scale parameter B is

$$f_g(x) = C\,x^{G-1}\exp[-x/B], \quad x,G,B > 0$$

TABLE VIII.III

METHOD THREE RESULTS FOR BETA EXAMPLE

A. Function Addition

Iteration	Function Set	M^2
1	F8	.024418
2	F8,F2	.016227
3	F8,F2,F5	.005436
4	F8,F2,F5,F3	1.1 (-26)

λ_0 =-2.99573, λ_2=-5.5(-13), λ_3=-3.0, λ_5=-1.0, λ_8=1.7(-15)

B. Function Deletion (THROUT)

Iteration	Divergence	Action
5	$J(p_{8253}(x), p_{253}(x))$=2. (-28)	Delete F8
6	$J(p_{8253}(x), p_{53}(x))$=8. (-25)	Delete F2
7	$J(p_{8253}(x), p_{3}(x))$=1. (+19)	Retain F5
8	$J(p_{8253}(x), p_{5}(x))$=22.805	Retain F3

C. Active Set = F5,F3

λ_0=-2.99573, λ_3=-3.0, λ_5=-1.0

where $C = 1.0/[\Gamma(G)(B^G)]$. With G=1.5 and B=2.0, we get

$$f_g(x) = .398940\ x^{.5}\exp(-.5x)$$

which is skewed left with mean 3.0 and variance 6.0. We chose bounds, [0,17.0], based on the random sample which was used in methods one and two. Table VIII.IV shows the gamma progression. We notice that our quadrature subroutine produced μ=2.9864995 and σ^2=5.8143277 on the [0,17.0] interval. Clearly, more accurate values for μ and σ^2 would be obtained on larger intervals. However, the given values are accurate for [0,17.] and the entropy approximation is concerned with this interval. The point is that expected value estimates should be computed over the interval of interest as we have done in our examples. Method three chose F1 and F3, or $x-\mu/\sigma$ and ln(x), for the active set. From Table VIII.IV,

$$\begin{aligned} p(x) &= \exp[-\lambda_0-\lambda_1(x-\mu)/\sigma-\lambda_3 \ln(x)] \\ &= \exp[-\lambda_0+\lambda_1\mu/\sigma]\exp[(-\lambda_1/\sigma)x]\ x^{-\lambda_3} \\ &= .398942\exp[-.49999998x]x^{.5} \end{aligned}$$

which again provides a rather accurate approximation to $f_g(x)$. The key to the extreme accuracy in all the examples is the fact that the data, i.e., the $\langle G\rangle_m$ vector, is accurate. The time and money invested by the analyst to

TABLE VIII.IV

METHOD THREE RESULTS FOR GAMMA EXAMPLE

A. Function Addition

Iteration	Function Set	M^2
1	F5	1.50377
2	F5,F8	.171532
3	F5,F8,F3	.001014
4	F5,F8,F3,F1	1.7 (-25)

λ_0=2.412188, λ_1=1.205646, λ_3=-.5, λ_5=5.9(-11), λ_8=1.1(-13)

B. Function Deletion (THROUT)

Iteration	Divergence	Action
5	$J(p_{5831}(x), p_{831}(x))$=2.4 (-20)	Delete F5
6	$J(p_{5831}(x), p_{31}(x))$=2.4 (-20)	Delete F8
7	$J(p_{5831}(x), p_{1}(x))$=.3155	Retain F3
8	$J(p_{5831}(x), p_{3}(x))$=15.10	Retain F1

C. Active Set = F1,F3

λ_0=2.412188, λ_1=1.205646, λ_3=-.5

produce accurate average values is well rewarded with method three.

Distribution number seven of Table VIII.II, which we call the hyperbolic, was investigated because it is similar in appearance to an exponential distribution on a bounded interval. Our intent was to see if method three would distinguish between similar distributions. The density for our hyperbolic is

$$f_h(x) = 1/[2x \ln(a)], \quad 1/a \leq x \leq a$$

or $$f_h(x) = \exp[-\ln x + \ln(a^2)]$$

Thus ln(x), F3, is the desired analytic information function, and method three must produce $\lambda_0 = -\ln(a^2)$ and $\lambda_1 = 1.0$ for an exact fit. Calculation will show for $f_h(x)$ that $\langle x \rangle = (a^2-1)/(2a-\ln a)$. For the exponential,

$$f_e(x) = \beta \exp[-\beta x], \quad 0 \leq x < \infty$$

with $\langle x \rangle = -1/\beta$. We wished to test hyperbolic and exponential distributions with the same means. We thus selected the a parameter for the hyperbolic, calculated <x>, and used <x> to find the exponential parameter. The result was two similar distributions with the same means, though the exponential is applied over a larger interval. Method three distinguished between the two distributions based on

respective $\langle G \rangle_m$ vectors, as shown in Table VIII.II, and produced accurate representations once more.

The uniform and bimodal distributions were introduced as extreme cases. The maximum entropy density for an unknown distribution given only the interval [a,b] and no further information is the uniform distribution. Method three, when given a $\langle G \rangle_m$ vector from a uniform distribution, should thus select only the constant parameter, λ_0. The method performed perfectly.

The bimodal distribution was taken from reference 14 which discussed a discrete entropy approach to develop density histograms. The bimodal density is composed of two "tent" functions,

$$
\begin{aligned}
f(x) &= 2x && 0 \le x \le 1/2 \\
&= 2(1-x) && 1/2 \le x \le 1 \\
&= 2(x-1) && 1 \le x \le 3/2 \\
&= 4-2x && 3/2 \le x \le 2 \\
&= 0 && \text{otherwise.}
\end{aligned}
$$

Our continuous entropy approximation procedure was developed for unimodal distributions as evidenced by our potential function set (Ref Chapter V). Thus our potential set does not contain the correct information functions to provide an exact fit to f(x), but a reasonable approximation results. Figure 8.1 graphs the analytic and entropy densities. Our continuous entropy approach provides a density

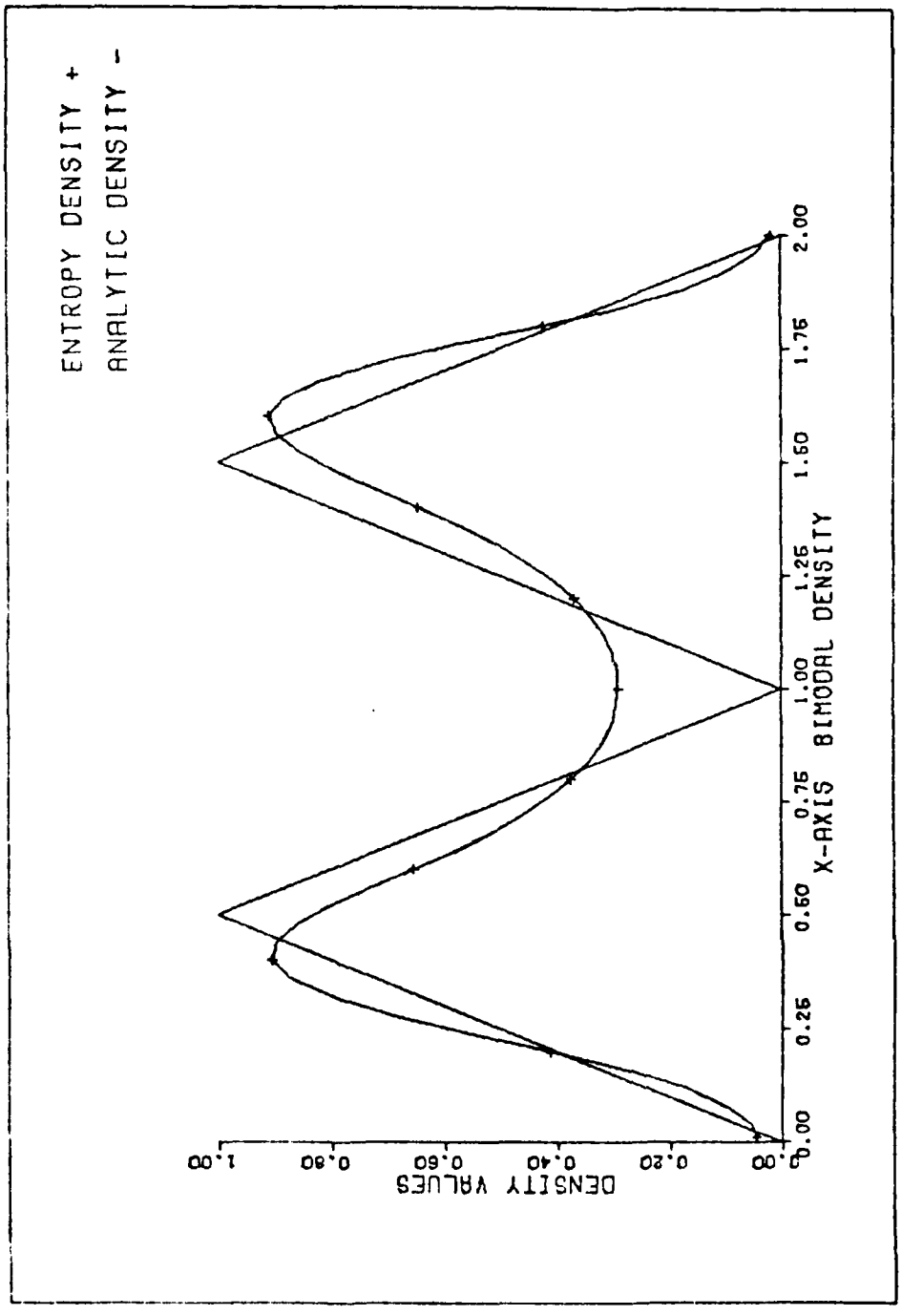


Fig. 8.1. Bimodal Approximation (Method 3)

that has the same general shape as the histogram approximation but also provides an analytic form for the approximation density, p(x). (Reference 14 contains graphs of the histogram approximation.)

Measure Sensitivity

The average values procedure of method three is based on a fit to the $\langle G \rangle_m$ vector. The method progressively selects the information functions that contribute to accomplishing this fit. Such a procedure is sensitive to the measure of fit, i.e., M^2. We selected a least squares measure, $M^2 = \sum_{i=1}^{m} (\langle g_i \rangle - \langle \tilde{g}_i \rangle)^2$, after experimentation with various forms of this measure. For successful application, M^2 must measure the "relative" contribution of an information function regardless of the size of a particular expected value. For example, if $g_8(x)=x^4$ and $g_1(x)=x$ on interval [50,100] then $\langle g_8 \rangle$ will be much larger than $\langle g_1 \rangle$. Thus the contribution of $(\langle g_8 \rangle - \langle \tilde{g}_8 \rangle)$ to M^2 will have more importance than $(\langle g_1 \rangle - \langle \tilde{g}_1 \rangle)$, and we would expect the procedure to first select the function that produces the smallest $(\langle g_8 \rangle - \langle \tilde{g}_8 \rangle)$.

Various forms of ratio tests were investigated in an effort to evenly weight the information functions, but such measures proved cumbersome. The most straightforward approach is to equalize the influence of information functions prior to testing, i.e., scale $g_i(x)$, $i=1,2,\ldots m$ such

that the $\langle g_i \rangle$ are roughly equivalent in value. Normalization of $\langle G \rangle_m$ is not sufficient because the size bias is maintained. Thus, we independently scale potential information functions if such functions promise to produce large average values on the interval of approximation. For example, we may replace moments, x^k, with normalized central moments, $(x-\mu)^k/\sigma^k$. Moments were the only information functions, in the potential set of Chapter V, that required scaling for our applications. The analyst should be aware of the possible need for function scaling for large approximation bounds, [a,b].

Summary

The goal of our entropy approximation procedure is to acceptably approximate an unknown distribution based on obtainable information. Method three has shown that we can provide an extremely accurate characterization, given the interval of approximation and the expected values of certain information functions. The accuracy of approximation depends on the accuracy of $\langle G \rangle_m$ and the flexibility of the potential set. The potential set of Chapter V is extremely productive for unimodal distributions on a bounded interval.

Method three has been implemented as a FORTRAN computer subroutine, METH3. The program uses previously discussed subroutines THROUT, which performs the function deletion step, and ENTROP, which solves the constraint

equations for the entropy density parameters. METH3 was used in all the above examples. The interval bounds [a,b] and vector $\langle G \rangle_m$ are required subroutine inputs, and the active set and p(x) are returned. The analyst can modify the potential function set by changing function cards which do not involve the principle subroutines. The program is constructed to handle 12 information functions in the potential set although the 9 functions of Chapter V (repeated in Table VIII.I) have proved sufficient in experimentation.

The excellent performance of the expected values method is complemented by its simplicity and the lack of a need for a large random sample from the unknown distribution. However, the analyst may prefer to produce a sample to make density and cumulative comparisons as in methods one and two. Such comparisons confirm the accuracy of approximation and indicate if modifications to the potential set are needed.

As a final comment, we review the purpose of the selection procedures. The purpose is to select the active set of functions for a specific approximation. If the analyst has already identified the active set, i.e., he knows the family of the unknown distribution, then he may circumvent selection procedures and simply solve the constraints to completely specify p(x). If the analyst can only generate expected values (or averages) for certain

functions, then those functions may compose the potential set, and method three may be used to select the active set. If a random sample is available then the analyst may prefer methods one or two. Finally, the analyst may prefer a combination of the above techniques or application of all three methods.

We have demonstrated that the methods work. The selection of a particular method depends on the specific problem, available data, and data accuracy. Our approach has been to first apply method three because it is the easiest to use. However, inaccurate expected values may cause unsatisfactory results with method three as demonstrated in the "interval arithmetic" application of Chapter XI. Methods one or two may then be preferred.

Chapter IX. Application to Simulation

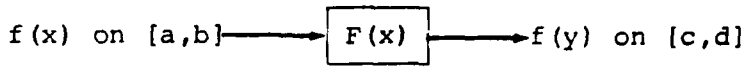
Introduction

Previous chapters have described an effective procedure for approximating an unknown distribution and providing a closed form for the approximating density, p(x). The procedure has broad application in mathematical and stochastic analysis. This chapter discusses the use of our procedure to approximate the output distribution of a computer simulation, an application which motivated the development of our method. The method was designed to support simulation studies at the Air Force Flight Dynamics Laboratory (AFFDL). Vehicle Synthesis Branch.

The strength and importance of the entropy approximation procedure as applied to simulation will be demonstrated by considering an AFFDL example problem. The general simulation problem and usual output characterization approaches are considered, followed by discussion of the entropy approach with example.

Simulation Model

A computer simulation may be viewed as a "black box" model which provides a distribution of output values based on user specified input distributions. We consider a simplified "black box" model:



where f(x) is the density function of input random variable X, F(x) is the mathematical transformation which represents the simulation, and f(y) is the output random variable density function. Notice that random variables X and Y may be vector valued although we consider univariate input and output for now. The transformation F(x) is known to the user or is available as a computer subroutine.

This model highlights two significant points about simulation. First, the input distribution must be completely specified by the simulation user. Input specification may take several forms (e.g., a set of discrete values for X, a density function for X, or a curve representing possible values, etc.), but the input is specified and thus known. Second, the usual purpose of the simulation is evaluation of the output. The user may require a sample output distribution, an average value of the output variable, or an indication of output sensitivity to input variations. The simulation, in general, becomes more useful to the user as his degree of knowledge about the output distribution increases. However, simulation output is a set of discrete values. To best serve the user, this set of values must be manipulated to describe the stochastic nature of the output in terms of a distribution or density function. Thus, the input is known, and the transformation

is known (or available as a subroutine). We seek an approximation or characterization of the output distribution.

Output Characterization

How might an analyst characterize the output distribution given the known information in the previous model? The answer to this question depends on the nature of the mathematical transformation, F(x). If F(x) is known and linear or mathematically "nice," then an analytical solution may be feasible. For example, the analyst may be able to propagate the input density, f(x), through the model using transformation of variable techniques to analytically derive f(y) without computer simulation. One might consider decomposition of F(x) into a series of less complex transformations for this purpose. For nonlinear transformations, linear approximation is a popular technique and has been successfully applied to very complex modeling problems.

Research conducted by Orr (Ref 63) provides an example of such an application to an extensive antiaircraft artillery simulation. While linear approximation was acceptably accurate for several nonlinear portions of the simulation, simplified nonlinear models were needed for a few submodels. Clearly, the analytical approach is the preferred method when possible. However, detailed modeling of real world processes seldom results in transformations which are analytically manageable. The most frequent

approach when difficult transformations are involved is Monte Carlo simulation.

Monte Carlo simulation, while a potentially accurate and reliable means to model real world systems, possesses two notable disadvantages. First, a large number of simulation trials are needed to provide adequate information about the output distribution. Thus, computational expense may restrict Monte Carlo analysis, particularly in terms of output sensitivity to input variations. Secondly, the output of a Monte Carlo simulation is a random sample. A suitable method of distribution approximation must be applied to this sample to derive meaningful stochastic information. The entropy procedure provides a distribution approximation and can reduce the number of trials or simulation calls required for approximation and for subsequent sensitivity analysis.

Entropy Approximation

The entropy approximation procedure, when applied to computer simulation, provides a usable and minimally prejudiced density function, effectively uses available or computable information, and provides analytical insight that is not afforded by Monte Carlo simulation. We discuss this application and potential computational benefits for sensitivity studies.

The entropy procedure was presented in Chapter III and expanded in subsequent chapters. The simulation output density, f(y), is approximated by an entropy density of the form

$$p(y) = \exp\left[-\lambda_0 - \lambda_1 g_1(y) - \ldots \lambda_k g_k(x)\right]$$

where $g_i(y)$, i=1,2,...k are information functions. We highlight the benefits of the entropy application to simulation by reviewing the procedural steps.

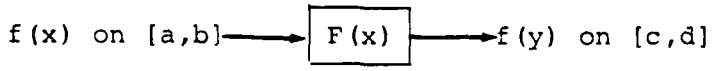
Select the Active Set. Three methods were described for selecting the active set of k functions from a predefined potential set of m functions (Chapters VI, VII, and VIII). Methods one (regression) and two (divergence) each require a Monte Carlo sample of the simulation output to select the active set. Since a Monte Carlo sample can provide a numerical approximation of the output distribution, one may question the benefit of proceeding with the entropy application. However, the entropy procedure provides at least two advantages. First, an analytical representation of the output density is provided for ease in subsequent analysis. As previously discussed, the specific information functions that are selected for this representation may provide further insight, i.e., if y and y^2 are selected then the output distribution is approximately normal. Secondly, the entropy procedure produces a

distribution that compromises between the sample and true underlying distributions. Thus, the entropy method provides a closer fit to the underlying unknown density than a single Monte Carlo run can provide. Method three (expected values) does not require a Monte Carlo sample, unless such a sample is needed to compute the expected values, and thus provides the added benefit of reduced simulation calls. (However, we recommend an initial Monte Carlo run when using method three to ensure that the potential set includes a sufficient number of functions (Chapter VIII).)

Given the initial Monte Carlo sample, subsequent Monte Carlo runs are needed only when the forms or bounds of the input distributions are changed, i.e., from one distribution to another such as from normal to beta. Variations in input distributions, such as changes in mean, variance or distribution parameters, are permissible without reselecting the active set. Thus, once the active set is defined, sensitivity analysis may be accomplished by simply generating expected values for the active set and solving for the specific p(y). A substantial change of inputs (e.g., a change of bounds or distribution families) may drastically affect the form of the output distribution, and active set selection should be reaccomplished to ensure accurate approximations. The entropy procedure thus offers a significant tool for output analysis and potential

savings in computer time via a reduced need for simulation access.

Generate Expected Values. Appendix A and Chapter III introduce numerical quadrature for generation of expected values. Consider, once more, our simplified simulation model:



From equation (3.5) we have

$$\langle g_i(y) \rangle \approx \frac{b-a}{2} \sum_{j=1}^{M} W_j \, g_i(F(x_j)) \, f(x_j), \quad i=1,2,\ldots k,$$

where W_j and x_j, $j=1,\ldots M$ represent quadrature weights and points, and $f(x_j)$ is the value of the input density at x_j. A study of this approximation for $\langle g_i(y) \rangle$ surfaces two significant benefits:

1. Only M function evaluations (or simulation calls) are needed to calculate all k expected values $\langle g_i(y) \rangle$, $i=1,2,\ldots k$ That is, the same values of $F(x_j)$ are used in all k expected values. This fact alone provides a significant improvement to Monte Carlo simulation.

2. The M simulation calls ($F(x_j)$, $j=1,2,\ldots M$) may be stored and the simulation input, f(x), modified (the interval [a,b] must remain fixed) for subsequent sensitivity analysis without further access to the simulation. This

benefit provides a notable analytical tool for sensitivity analysis.

The listed benefits are particularly significant for expensive simulations. However, one must note that the benefits of quadrature are somewhat reduced when multidimensional integrals are involved, i.e., when the input distribution is multivariate. Quadrature is an approximation to analytic integration and the accuracy of approximation depends on the number of quadrature points, M. While M may be small for one dimensional integration (16 to 32 points offer excellent accuracy in most cases), the actual number of simulation calls for n dimensional integration is M^n. Consequently, the number of simulation calls for large n can rapidly approach the number of calls for a Monte Carlo simulation. Appendix A discusses multidimensional quadrature and provides references As a rule of thumb, quadrature is effective when the number of input variables, n, is less than or equal to four.

A final point pertains to specification of the interval for output approximation, [c,d]. These bounds must be known prior to application of the entropy procedure. If the analyst knows reasonable limits for the simulation output, then such limits should be used. However, both quadrature and Monte Carlo methods of estimating the expected values supply a means to estimate [c,d]. For the Monte Carlo approach, the bounds are simply the minimum and

maximum values of the sample distribution. The same logic applies to quadrature, i.e., $[c,d] \approx [\min_i F(x_i), \max_i F(x_i)]$ where $F(x_i)$ is simulation output at quadrature point x_i.

Solve the Constraint Equations. This step is thoroughly discussed in Chapter IV to include a computer subroutine for implementation. The constraint equations are repeatedly solved in each of the active function selection procedures. Once the active set is known, then subsequent output approximations for the same input family may be accomplished by generating new expected values and solving the constraints.

Numerical Usefulness

The entropy procedure provides a minimally prejudiced, stochastic representation of the simulation output. The approach is not dependent on a specific simulation but is geared to "black box" models. The numerical usefulness is discussed in previous sections and summarized here.

The entropy procedure provides an analytic form for the simulation output distribution based on expected values of functions. Expected values may be calculated from the known input distributions with limited use of the simulation. Simulation calls may be stored for subsequent analysis of output sensitivity to input changes. When a Monte Carlo sample is needed to identify the active set, the Monte Carlo work does not have to be reaccomplished when input

parameters are modified (given fixed input bounds). Additionally, the approach offers an analytic form for the output density which is not provided with straight Monte Carlo simulation. The entropy approach thus provides potential reduction in simulation calls and stochastic output representation.

Example Application

An example application of the entropy procedure to simulation is given for the AFFDL problem of Figure 9.1. The problem is described in the first section and followed by a detailed discussion of entropy results using active set selection method one (linear regression). Methods two and three are then considered and show very similar results. All three methods provide accurate approximations to the data.

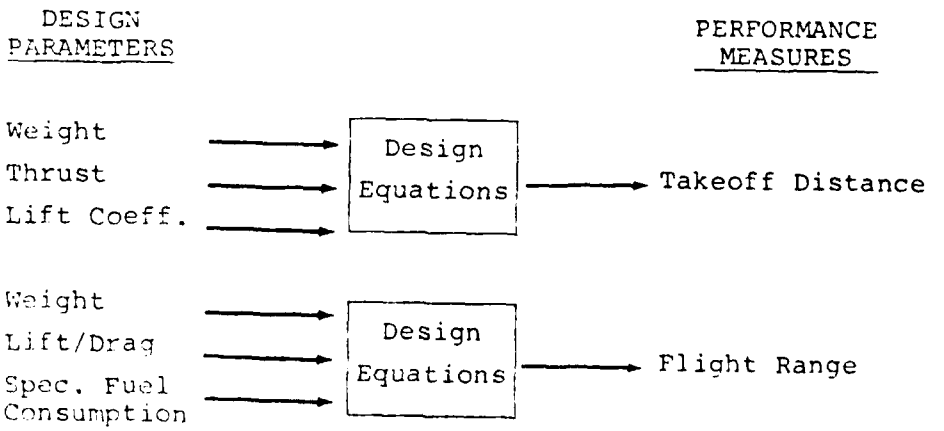


Fig. 9.1. AFFDL Simulation Model

Simulation Model. AFFDL contracted for a study of aircraft performance measures for several different aircraft with different engines (Ref 59). Managers were concerned with the stochastic nature of performance measures as related to the stochastic behavior of several design parameters. The problem of interest is graphically displayed in Figure 9.1. Our example involves the two performance measures in Figure 9.1 (takeoff distance and flight range) and the indicated design parameters for one engine (TF2--Ref 59). Variability in the aircraft production process, as reflected in the stochastic nature of design parameters, means that no two aircraft will have the same values for performance measures. The contract goal was to predict performance measure distributions based on estimated design parameter distributions. The approach was Monte Carlo simulation. Aircraft production was simulated via traditional design equations with the design parameters as simulation inputs. Various distributions were assumed for the input parameters, and a sample cumulative was produced for each performance measure. Constants in the design equations and input distributions were altered to represent different type aircraft or engines and comparative sample cumulatives were produced.

To demonstrate the entropy method, we used the contractor's design equations (Table IX.I) and input distributions, but we developed our own simulation models (one

TABLE IX.I

PERFORMANCE MODEL

Takeoff

$$\text{distance} = 8.656^{*}\left[\frac{W^{2}}{\rho \cdot S \cdot C_{L} \cdot T}\right]^{1.16} + 425.2$$

$W = W_{empty} + W_{fuel} + W_{payload}$

ρ = density ratio

S = Wing surface area

C_L = Maximum coefficient of lift

T = Thrust

Range

$$\text{range} = \left(\frac{L}{D}\right) \cdot \left(\frac{V}{C_t}\right) \cdot \ln\left(\frac{W}{W1}\right)$$

$W1 = W_{empty} + W_{payload} + W_{reserve}$

V = Cruise speed

C_t = Specific fuel consumption

L/D = Lift/drag ratio

*NOTE: Constant 8.656 later corrected by contractor to 10.81. We retained 8.656 in our examples.

for each performance measure). The models were developed to provide experimental control but also proved useful in identifying a minor contractor error in one design equation (see Table IX.I). We use the equations as indicated in Table IX.I without correction. For a specific example we consider the input distributions of Table IX.II, i.e., normal distributions with bounds $[\mu-4\sigma, \mu+4\sigma]$ where μ is the mean and σ is the standard deviation. Random samples of 500 takeoff distances and 500 range values were generated and stored. We now apply the entropy procedure to approximate the performance measure distributions.

Expected Values and Integration Bounds. The entropy approach requires estimates of expected values of information functions in the simulation output space. The estimates may be provided via average values (Monte Carlo) or numerical quadrature. Each simulation in Figure 9.1 involves three independent inputs and a univariate output, thus suggesting quadrature for expected values (Ref Appendix A and Chapter III). We use a 16 point quadrature formula for the triple integration in our example. Multidimensional quadrature is discussed in Appendix A. Table IX.III provides a comparison of quadrature and average values for simulation output means and variances with various samples and sample sizes. Average values were computed as follows: $\langle g(t) \rangle \approx \sum_{i=1}^{N} g(t_i)/N$ where N is the sample size and

TABLE IX.II

EXAMPLE INPUT DISTRIBUTIONS AND CONSTANTS

Input	Distribution*	Mean= μ	Variance= σ^2
W_{empty}	Normal	.7427(+6)	.702944(+9)
T	Normal	.3017(+6)	.149810(+7)
C	Normal	.2920(+1)	.324863(-2)
CL/CD	Normal	.2030(+2)	.102478
Ct	Normal	.63	.332522(-5)

Constant	Value
W_{fuel}	498,000 lbs.
$W_{payload}$	390,000 lbs.
ρ	.944
S	11,270 ft
V	460 kns
$W_{reserve}$	40,600 lbs.

*NOTE: All distributions bounded on $[\mu-4\sigma, \mu+4\sigma]$.

TABLE IX.III

COMPARISON OF QUADRATURE AND AVERAGE VALUES

Measure	Sample Size	Mean	Variance	Minimum Value	Maximum Value
Quadrature:					
Takeoff	16	6492.26	72050.87	5104.12	8262.58
Range	16	4880.56	14962.69	4185.34	5701.30
Averages:					
Takeoff	1000	6506.04	70559.08	5605.12	7462.57
Takeoff	500	6496.71	70648.83	5830.01	7473.79
Takeoff	500	6491.58	71019.91	5745.75	7240.77
Takeoff	500	6539.59	70677.82	5603.24	7242.84
Takeoff	100	6520.41	81384.04	5629.48	*
Takeoff	100	6543.82	83514.99	5742.07	*
Takeoff	100	6570.19	91038.30	5750.68	*
Averages:					
Range	1000	4880.91	14090.92	4492.83	5304.45
Range	500	4887.31	14965.55	4484.93	5230.29
Range	500	4901.04	14915.77	4536.23	5418.93
Range	500	4900.85	13364.17	4585.49	5347.85
Range	100	4865.78	14127.65	4563.52	*
Range	100	4947.65	15569.14	4661.69	*
Range	100	4998.05	15611.30	4623.67	*

*NOTE: Upper bounds not retained on this run.

t_i is the output sample value. As expected, a sizeable fluctuation is noted with sample size. The average value appears to approach the quadrature value as N increases. Haber (Ref 34) provides more insight to quadrature accuracy.

Table IX.III also compares sample and quadrature output bounds. These bounds must be specified for constraint solution. For consistency, the bounding method should agree with the method of estimating expected values, i.e., sample bounds if averages are used. Notice that all sample bounds are subsets of the quadrature interval. Our examples use quadrature bounds.

Method One (Linear Regression). The entropy procedure requires selection of the active set of information functions from a predefined potential set. As in previous examples, we use the Chapter V potential set (repeated in Table IX.IV for reader convenience) with quadrature values for μ, σ, and [c,d]. Selection method one (Chapter VI)

TABLE IX.IV

POTENTIAL INFORMATION FUNCTION SET

$F1 = (x-\mu)/\sigma$	$F4 = \ln(x-c)$	$F7 = ((x-\mu)/\sigma)^3$
$F2 = ((x-\mu)/\sigma)^2$	$F5 = \ln(d-x)$	$F8 = ((x-\mu)/\sigma)^4$
$F3 = \ln(x)$	$F6 = (\ln(x-c))^2$	$F9 = \ln(x^2+1)$

NOTE: μ = mean; σ = standard deviation; [c,d] = bounds.

was applied to our random samples of size 500 for takeoff distance and range simulation outputs. The regression results and statistical measures are listed in Tables IX.V and IX.VI for nine candidate sets. The function sets were chosen based on largest and second largest adjusted R^2 for a given set size. We allowed up to six active functions per set. Notice from the tables that different statistics imply different active sets. Chapter VI and Appendix B provide guidance in choice of statistic for final active set selection. We choose the Anderson-Darling, A^2, statistic because we seek accuracy in the distribution tails. Thus, E4=(F2,F4,F7,F8) is the active set for takeoff distance characterization and E2=(F2,F8) is active for flight range. The best value for each statistic is underlined in the tables.

Let us consider the approximation accuracy for the takeoff distance example of Table IX.V. The value of A^2 is .3193. Stephens (Ref 76) provides a table of critical values up to the 15% significance level. Our null hypothesis is:

> H_0: The sample distribution comes from a population with distribution function that is described by our entropy approximation, p(x).

Stephens' table gives a critical value of A^2=1.610 at the 15% level and A^2=1.933 at 10% significance. Thus we reject H_0 if our calculated value of A^2 exceeds the critical value at our chosen significance level. Our extremely small

TABLE IX.V

TAKEOFF DISTANCE REGRESSION RESULTS

Candidate Set	Information Functions	Adjusted R^2	Kolmogorov-Smirnov = D	Anderson-Darling=A^2	Cramér von Mises=W^2
E2	2,8	94.40	.0285	.4225	.0543
E3	2,4,8	94.68	.0266	.4243	.0585
E4	2,4,5,8	94.61	.0295	.3341	.0452
E5	2,3,4,5,8	97.82	.0293	.3244	.0436
E6	2,3,4,5,7,8	97.79	.0294	.3232	.0435
E2´	2,7	87.03	.0280	.4471	.0734
E3´	2,6,8	94.68	.0267	.4238	.0582
E4´	2,4,7,8	94.60	.0293	.3193	.0429
E5´	2,4,5,8,9	97.82	.0293	.3244	.0436
For E4´	λ_0=4.58, λ_2=.503, λ_4=.266, λ_7=-.0196, λ_8=.00103				

TABLE IX.VI

FLIGHT RANGE REGRESSION RESULTS

Candidate Set	Information Functions	Adjusted R^2	Kolmogorov-Smirnov=D	Anderson-Darling=A^2	Cramér von Mises=W^2
E2	2,8	94.83	.0263	.4492	.0746
E3	2,7,8	94.97	.0318	.7678	.1272
E4	2,4,7,8	95.09	.0266	.4698	.0731
E5	1,2,4,5,8	95.93	.0264	.4672	.0729
E6	1,2,4,5,7,8	97.45	.0265	.4702	.0733
E2´	2,7	86.44	.0317	.7686	.1273
E3´	2,4,8	94.73	.0285	.5380	.0902
E4´	2,6,7,8	95.09	.0266	.4698	.0731
E5´	2,4,5,7,8	95.47	.0265	.4696	.0732

values of A^2 for both takeoff distance and flight range indicate an accurate fit at significance levels much higher than 15%. Thus, we have produced accurate approximation to the output data for both simulations.

Figure 9.2 presents an entropy and sample density comparison for takeoff distance to demonstrate the accuracy. We eliminate numerical differentiation noise by comparing sample and entropy cumulative distributions in Figure 9.3. Differences between sample and entropy cumulatives are plotted versus the actual cumulatives because the entropy fit is very accurate. Figures 9.4 and 9.5 provide the same information for the flight range distribution with active set E2. The four figures show an accurate fit to the sample and thus imply an accurate representation of the unknown distributions.

Active set selection method one uses linear regression and is thus sample dependent. As demonstrated, the method provides an accurate fit to a given sample. However, application of the method to a second sample from the same distribution may result in selection of different active functions. We wish to show that the active set (and thus the entropy density) for a given sample will produce an acceptable fit to subsequent samples. If a given sample is a good approximation to the true underlying analytic distribution, then the entropy distribution which approximates this sample must also approximate the analytic distribution.

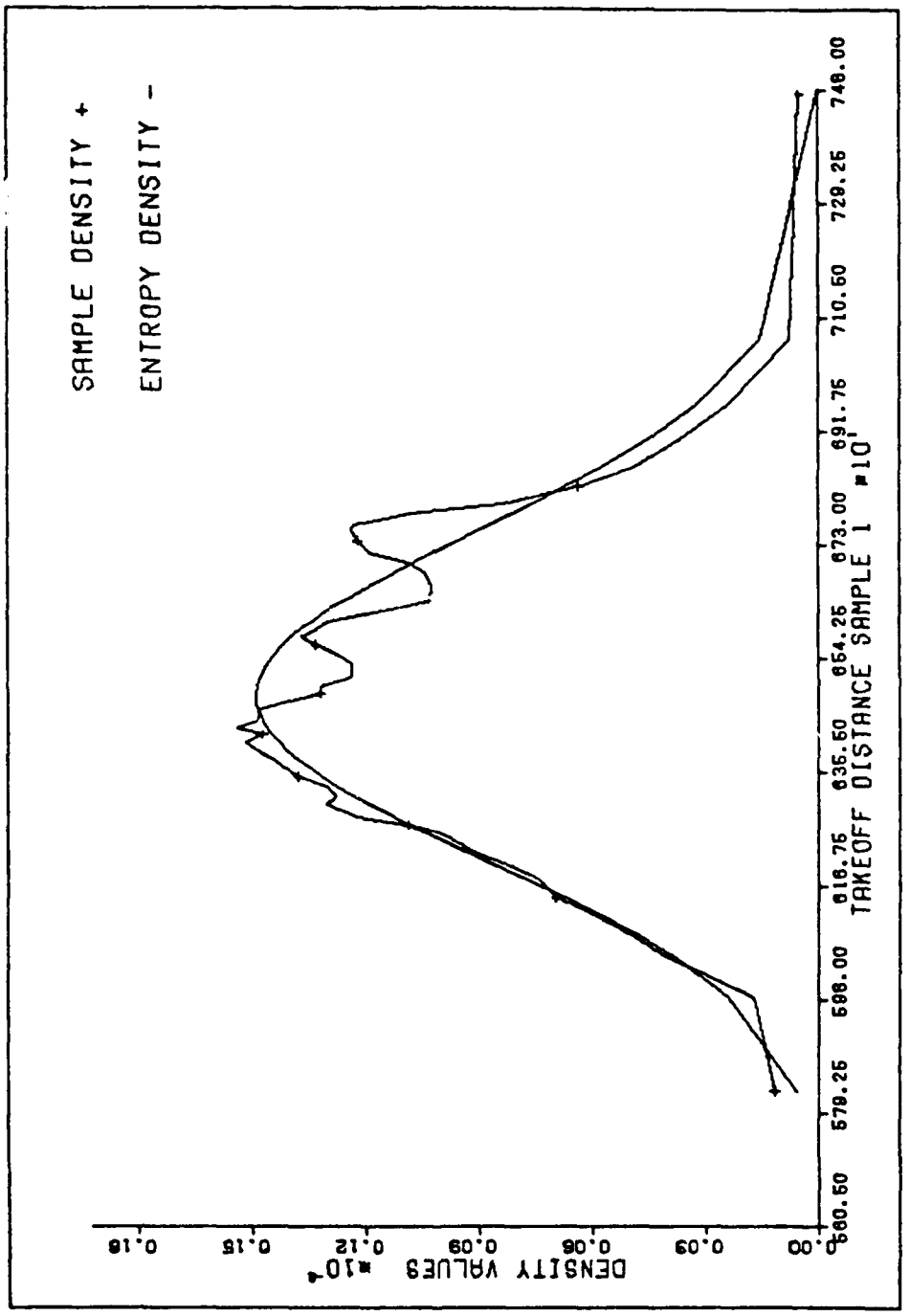


Fig. 9.2. Takeoff Densities, Sample 1 (N=500)

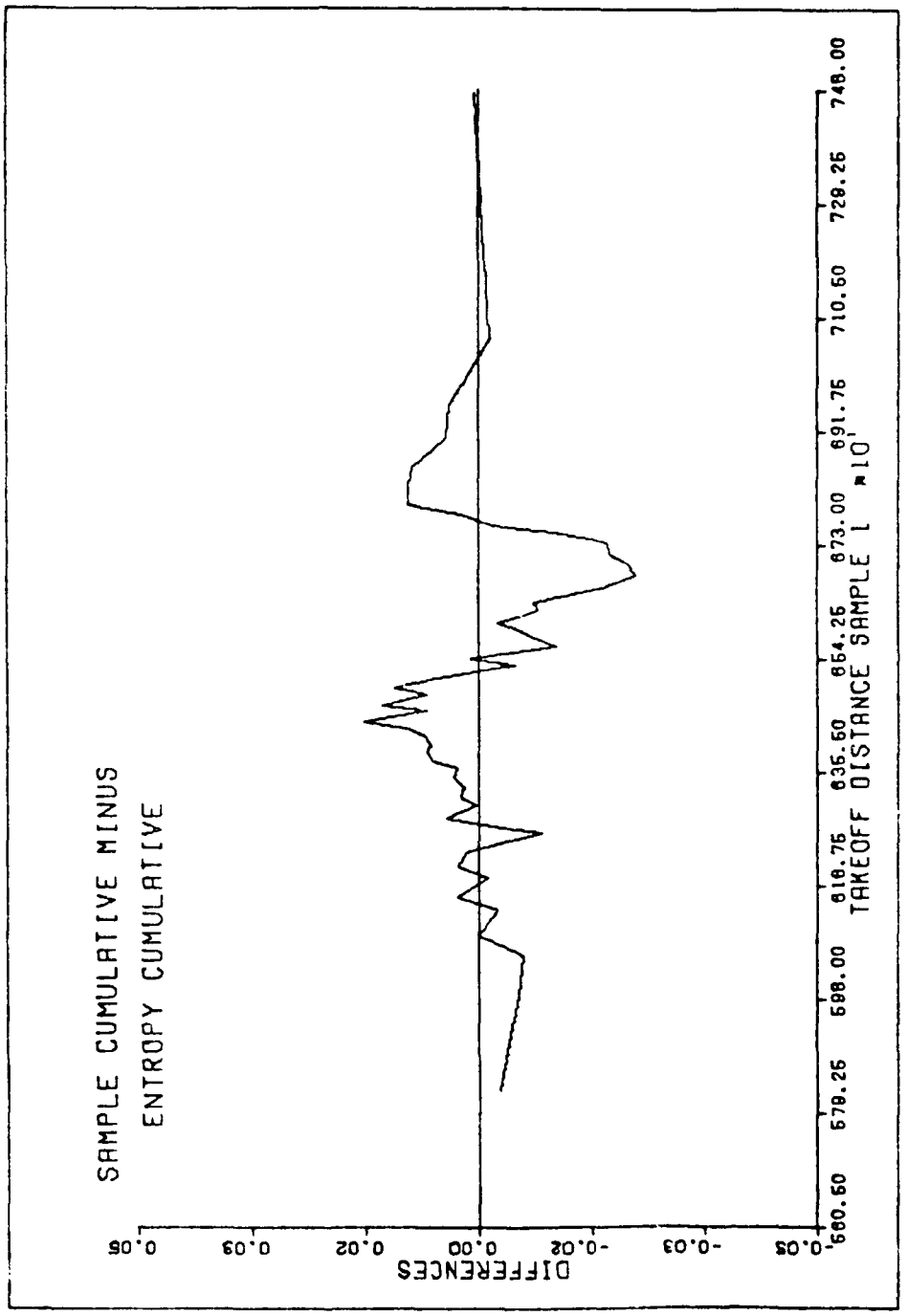


Fig. 9.3. Takeoff Cumulative Difference, Sample 1 (N=500)

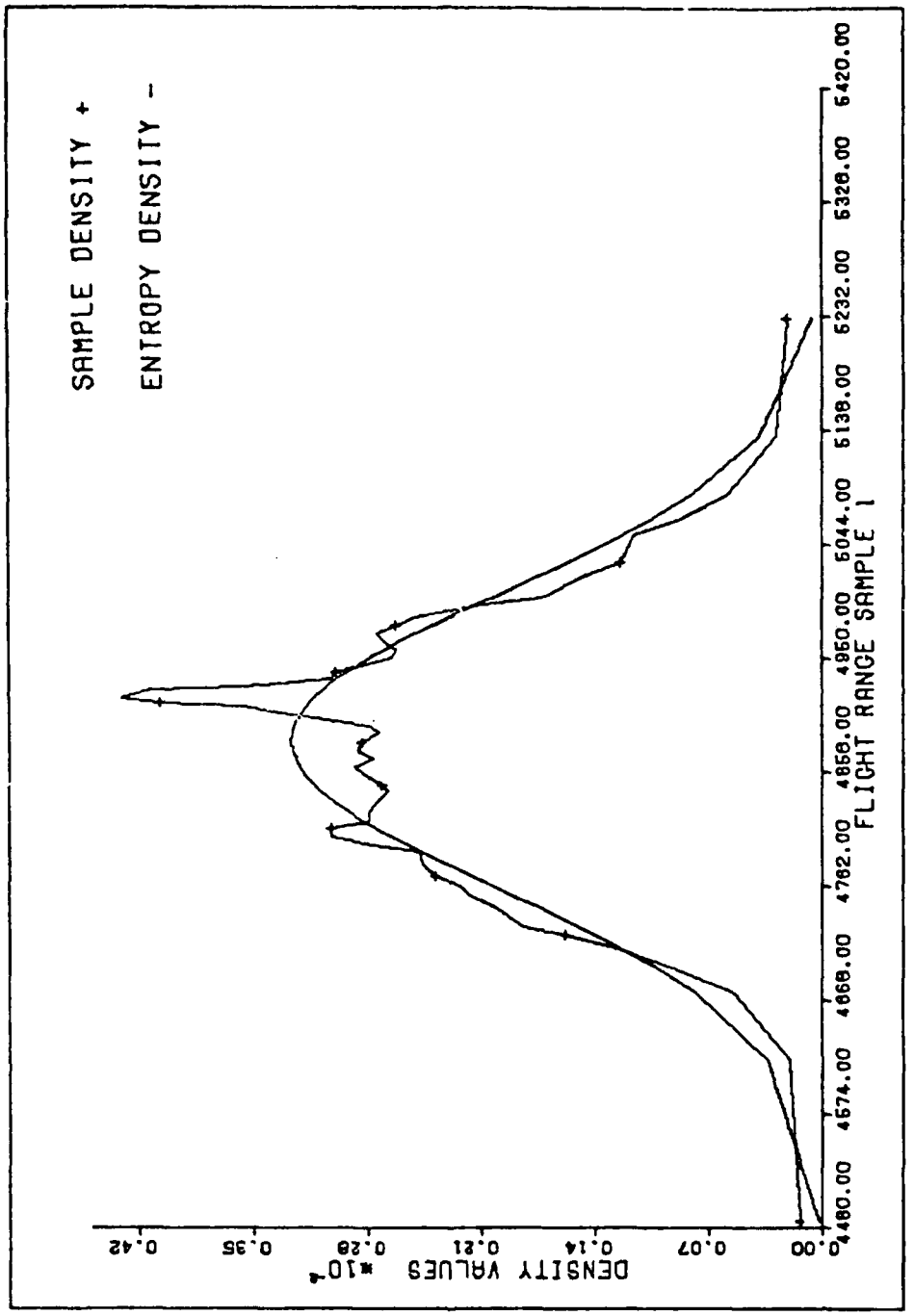


Fig. 9.4. Range Densities, Sample 1 (N=500)

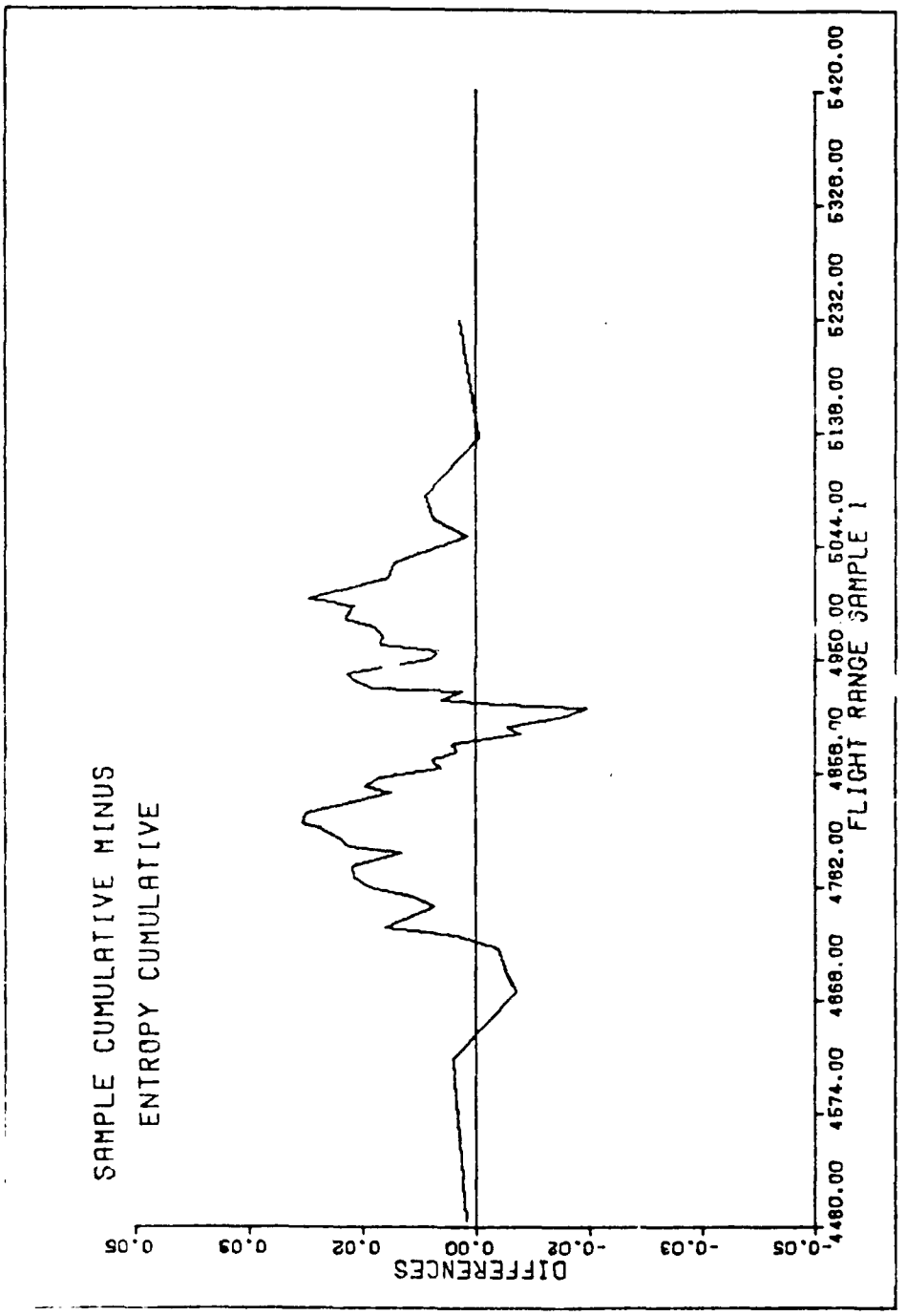


Fig. 9.5. Range Cumulative Difference, Sample 1 (N=500)

In fact, as indicated in Chapter VI, the entropy distribution is a compromise between the sample and analytic distributions. Thus, subsequent sample distributions should be well approximated by the entropy fit.

We demonstrate this prediction by generating additional samples and comparing the samples to our original entropy cumulative. Figures 9.6 and 9.7 provide cumulative comparisons for a second and third sample (again 500 deviates each) from the unknown takeoff distance distribution. Figure 9.8 and 9.9 provide equivalent comparisons for the flight range. We notice that subsequent samples, particularly for the takeoff example, fit one side or the other of the entropy approximation indicating that our original approximation is indeed a compromise. To further test our accuracy with the original entropy approximation, we generate a fourth and larger random sample (1000 deviates). Figures 9.10 through 9.13 graph results for the cumulatives. As sample size increases, the sample distribution approaches the entropy distribution (particularly in the tails) again indicating that the entropy distribution is a good approximation to the unknown, underlying analytic distribution.

Method Two (Divergence). We apply method two to the original takeoff distance sample data for comparison with method one. Method two uses the divergence measure and is described in Chapter VII. This method also produces

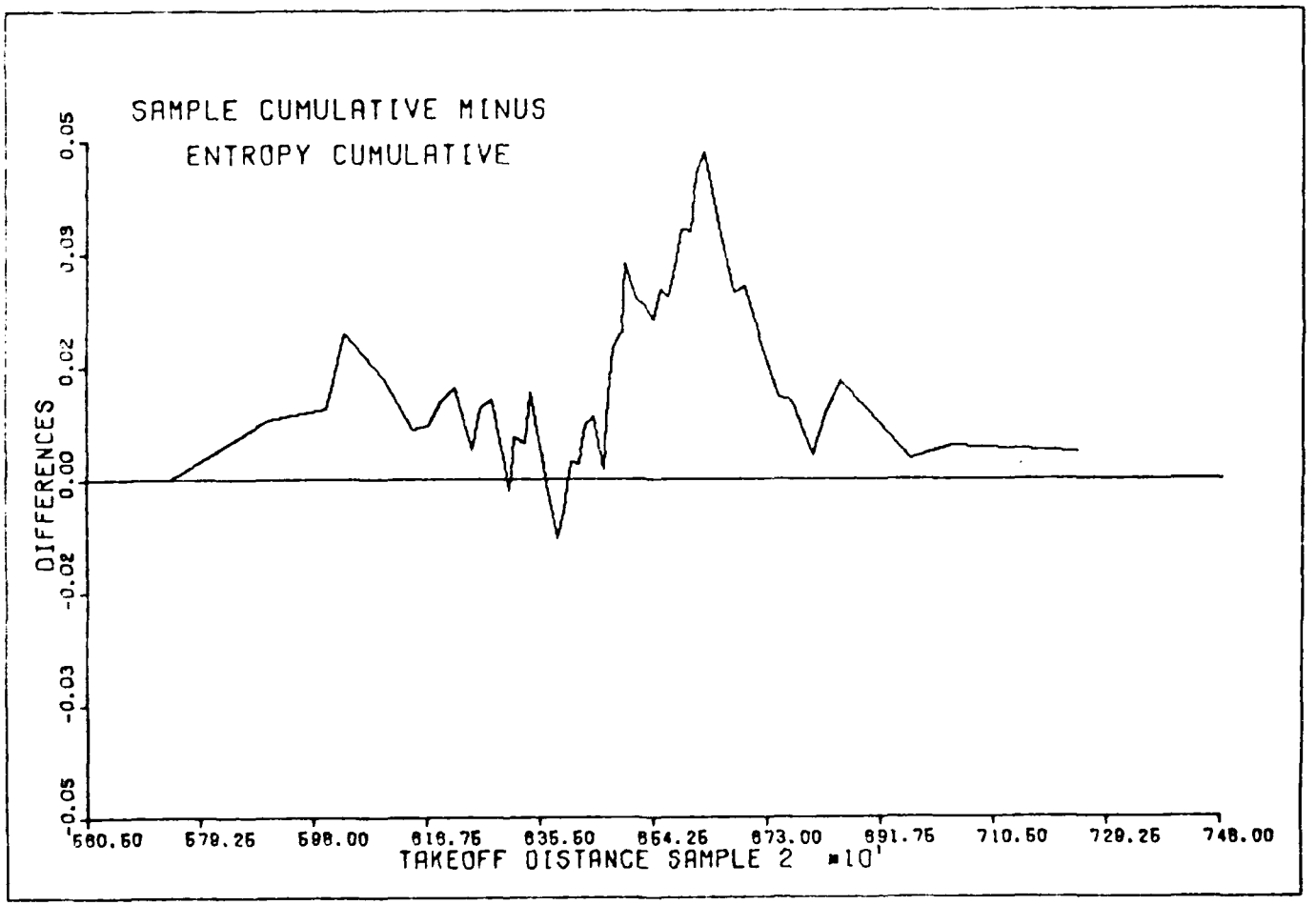


Fig. 9.6. Entropy Cumulative (Sample 1) Compared to Takeoff, Sample 2

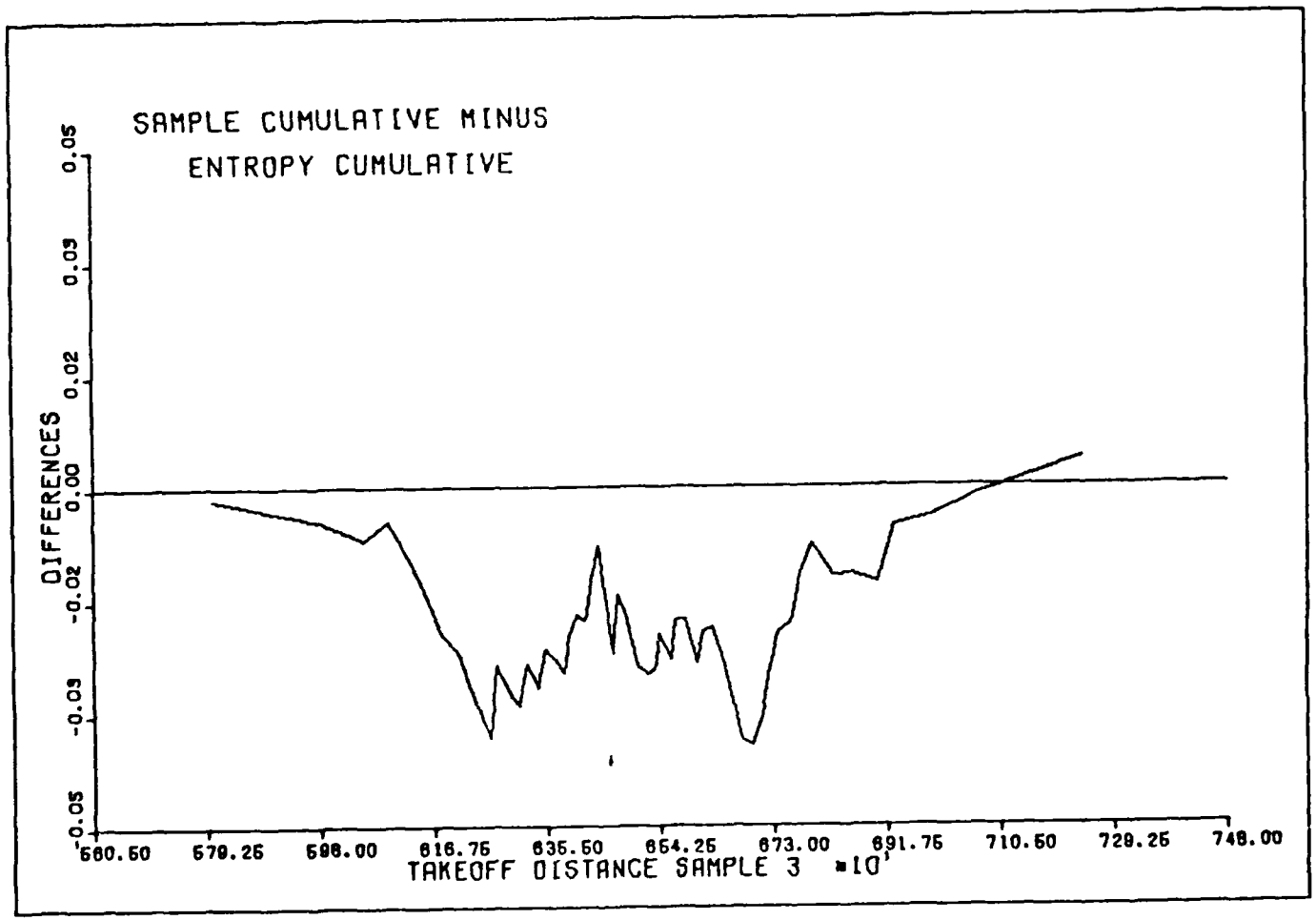


Fig. 9.7. Entropy Cumulative (Sample 1) Compared to Takeoff, Sample 3

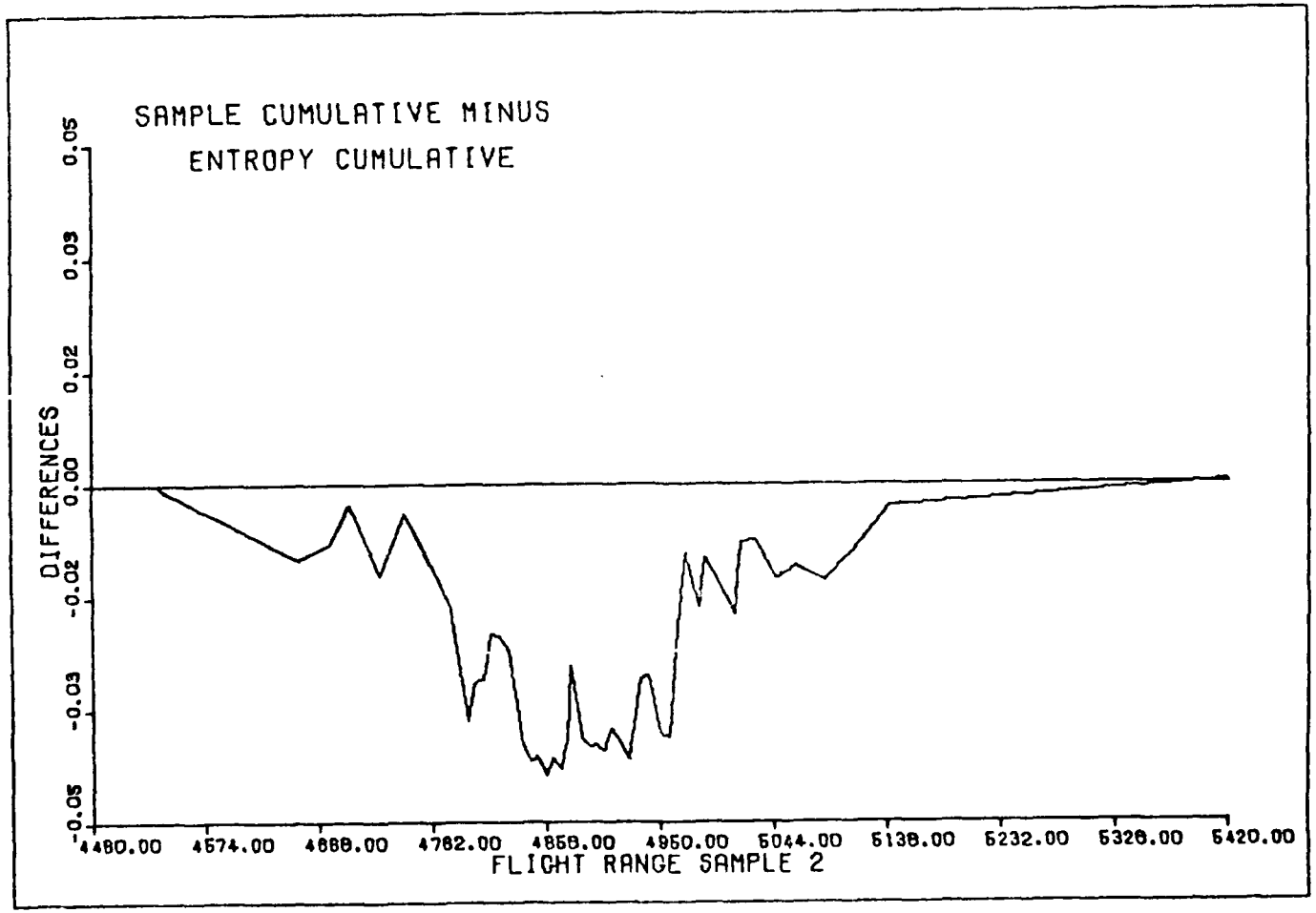


Fig. 9.8. Entropy Cumulative (Sample 1) Compared to Range, Sample 2

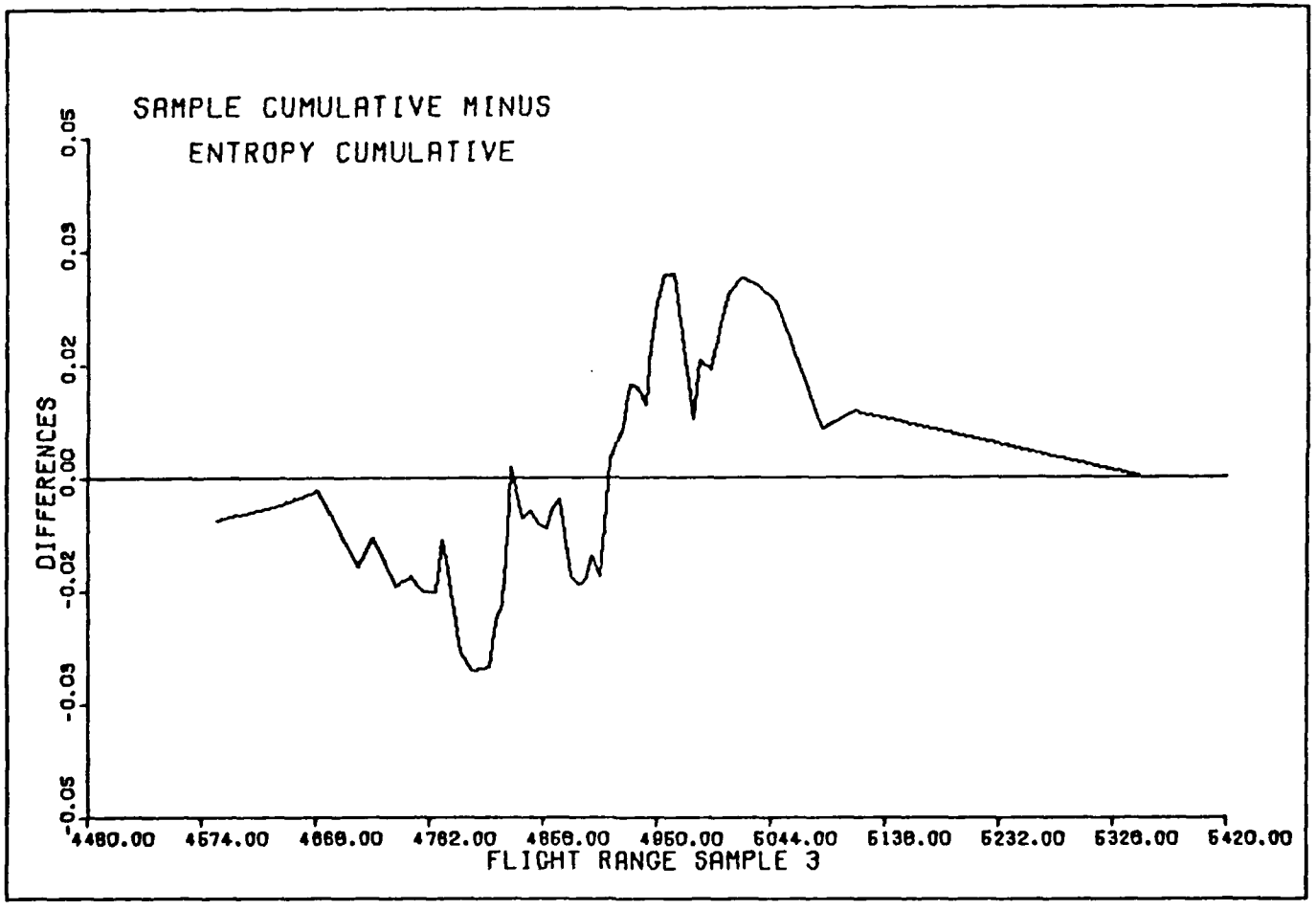


Fig. 9.9. Entropy Cumulative (Sample 1) Compared to Range, Sample 3

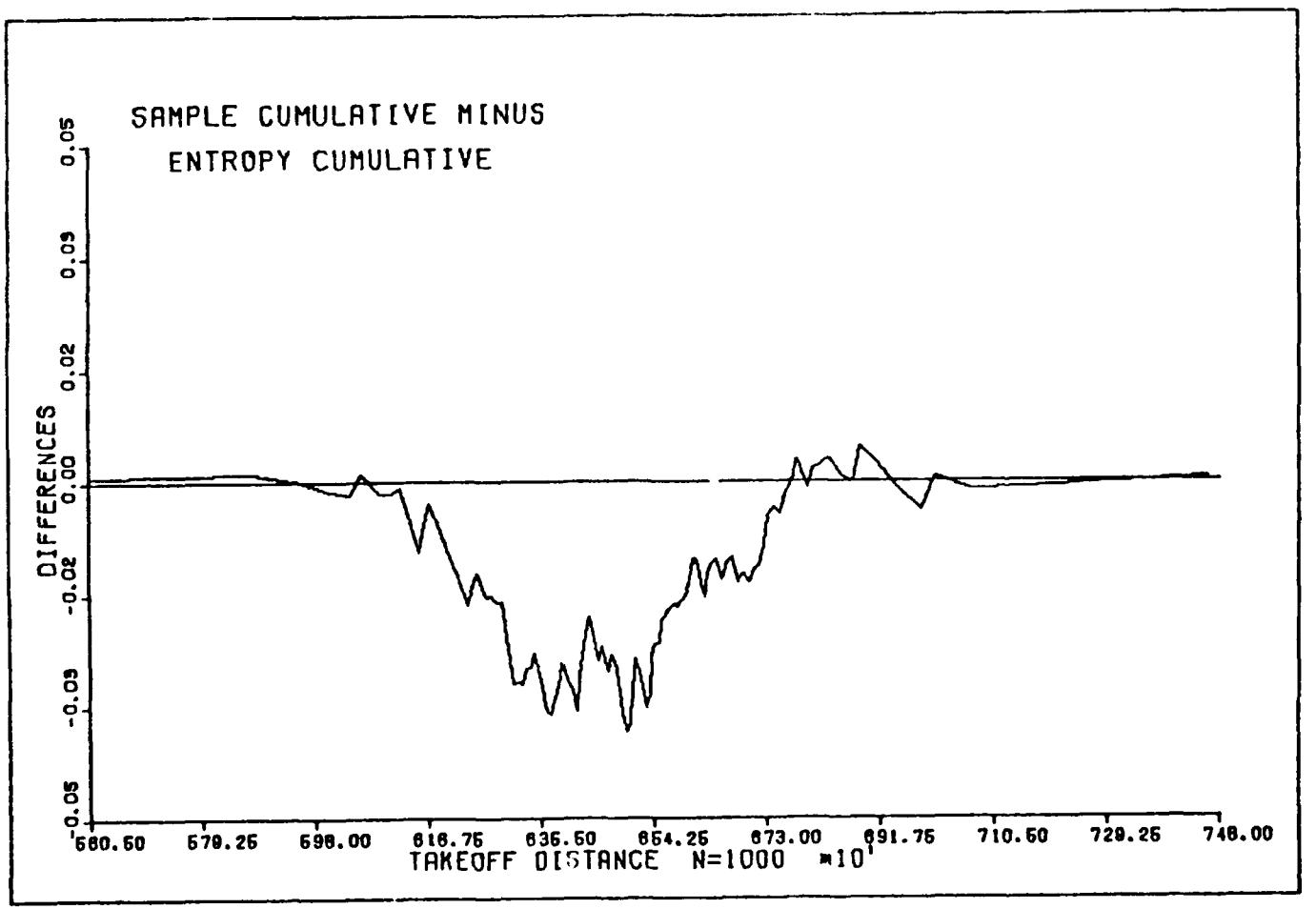


Fig. 9.10. Entropy Cumulative (Sample 1) Compared to Takeoff, Sample 4

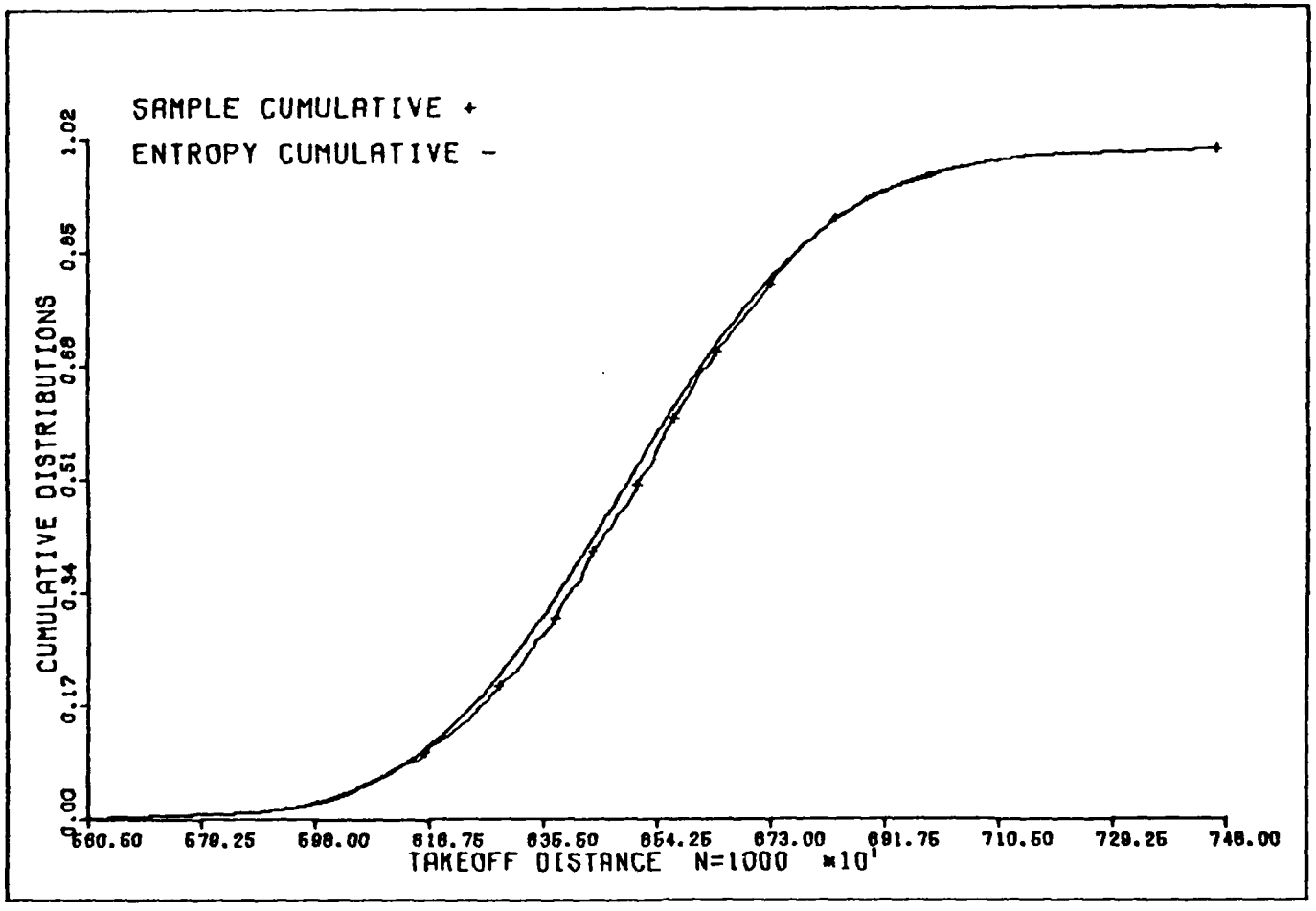


Fig. 9.11. Cumulatives for Entropy (Sample 1) and Takeoff, Sample 4

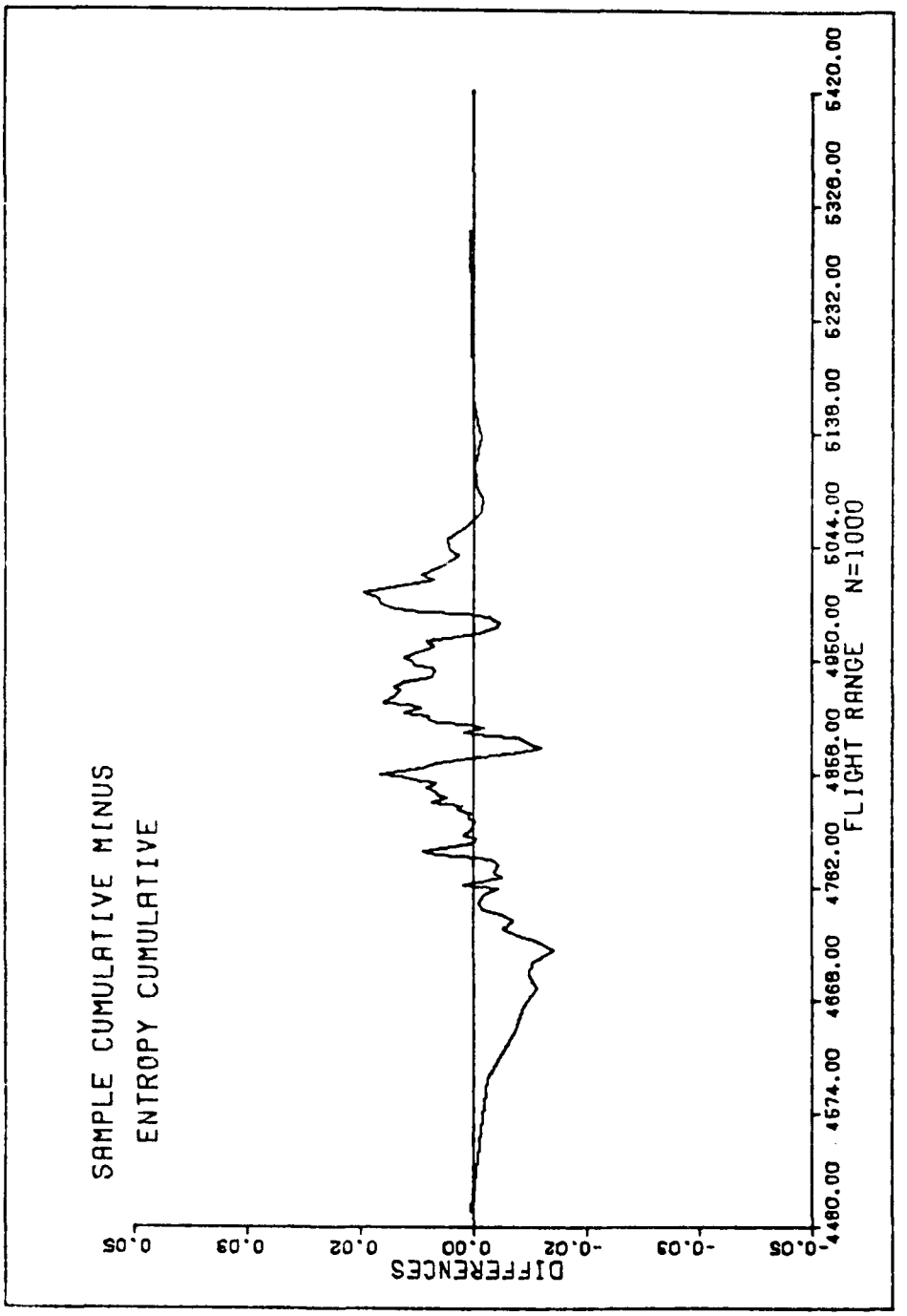


Fig. 9.12. Entropy Cumulative (Sample 1) Compared to Range, Sample 4

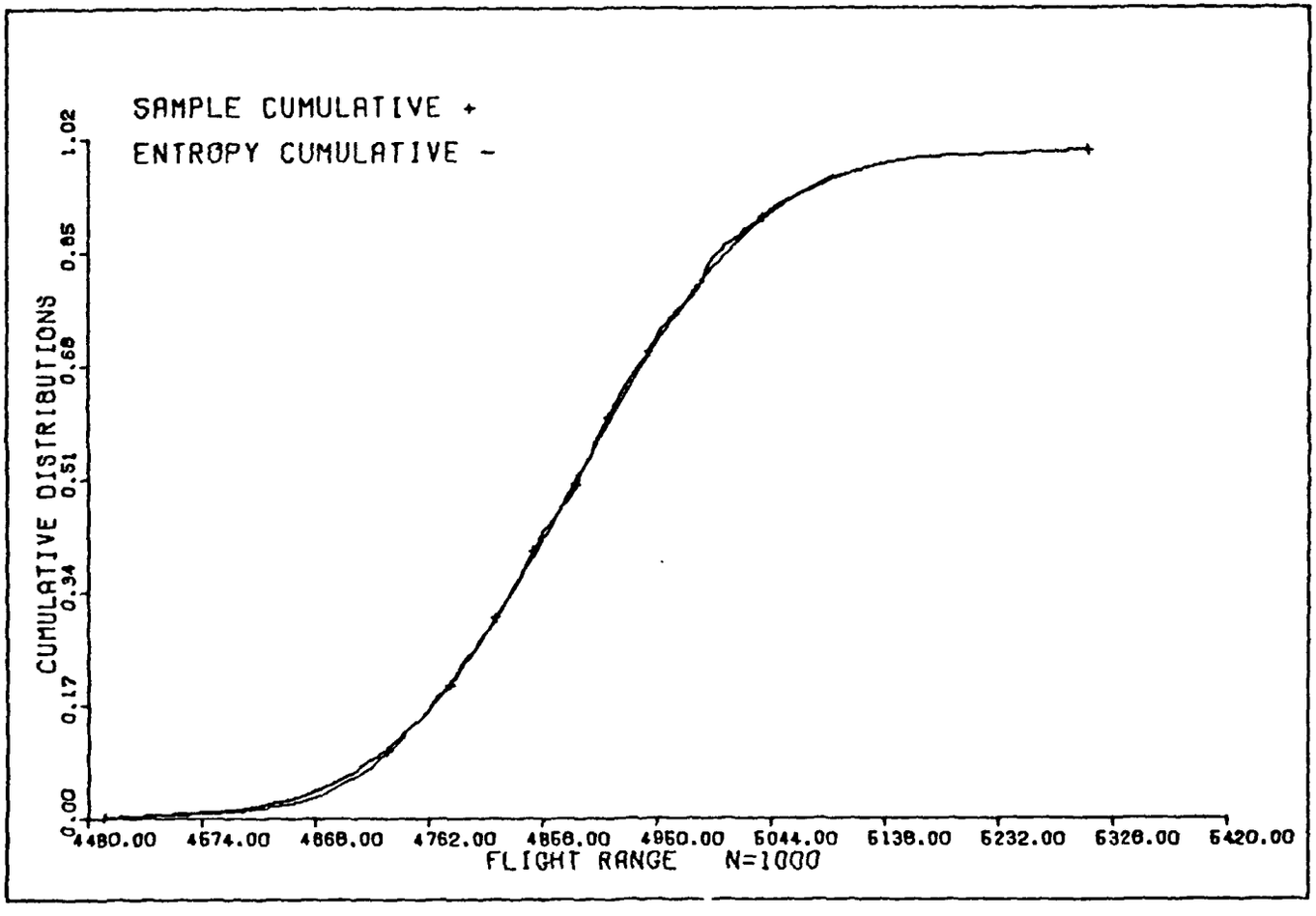


Fig. 9.13. Cumulatives for Entropy (Sample 1) and Range, Sample 4

a compromise distribution, but experimentation with known distributions has shown that the divergence measure favors a fit to the analytic distribution whereas method one favors the sample. Method two includes a function elimination step which is not currently part of method one. The elimination step (subroutine THROUT from Chapter VII) considers the entropy approximation and eliminates "redundant information" functions in an "information theoretic" sense.

The results of method two when applied to the sample takeoff distance data are shown in Table IX.VII. Notice that method two stops adding functions at set F2,F7,F4 because the addition of a fourth function increases divergence. The function deletion phase, iteration 6 of Table IX.VII, indicates that little information is communicated via function F7, and the analyst may consider eliminating F7. The small value of λ_7 also indicates a less important function. We chose to retain all three functions in our active set. Before comparing methods one and two, we consider method three.

Method Three (Expected Values). Method three (Ref Chapter VIII) does not use the sample distribution to produce the entropy approximation, but concentrates on a fit to the expected values of the potential set. Table IX.VIII presents the results of method three for the takeoff distance example. This table highlights two interesting points. Iteration four shows that adding either functions

TABLE IX.VII

DIVERGENCE METHOD APPLIED TO TAKEOFF DISTANCE SAMPLE

A. Function Addition on [x(1),x(N)] = [5830., 7473.8]

Iteration	Function Set	Divergence
1	2	$J(p_2(x),f(x)) = .044093$
2	2,7	$J(p_{27}(x),f(x)) = .042295$
3	2,7,4	$J(p_{274}(x),f(x)) = .038427$
4	2,7,4,6	$J(p_{2746}(x),f(x)) = .038841$

where f(x) is sample density

B. Function Deletion on Quadrature Bounds = [5104.1, 8262.6]

Iteration	Divergence	Action
5	$J(p(x),p_{74}(x)) = 652.8$	Retain Function 2
6	$J(p(x),p_{24}(x)) = .00532$	Retain Function 7
7	$J(p(x),p_{27}(x)) = 10.05$	Retain function 4

where $p(x) = p_{274}(x)$

C. Active Set = F2,F4,F7

$\lambda_0=4.648$, $\lambda_2=.5086$, $\lambda_4=.2572$, $\lambda_7=-.0188$

TABLE IX.VIII

METHOD THREE APPLIED TO TAKEOFF DISTANCE EXAMPLE

A. Function Addition

Iteration	Function Set	M
1	2	.015916
2	2,7	.003607
3	2,7,5	.000435
4	2,7,5,8	2.1 (-10)
	2,7,5,4	9.9 (- 7)

B. Function Deletion (THROUT)

Iteration	Divergence	Action
5	$J(p(x),p_{758}(x)) = 107.9$	Retain function 2
6	$J(p(x),p_{258}(x)) = .00540$	Retain function 7
7	$J(p(x),p_{278}(x)) = 2.326$	Retain function 5
8	$J(p(x),p_{275}(x)) = .000082$	Retain function 8

$\lambda_0=9.045$, $\lambda_2=.4940$, $\lambda_5=-.3386$, $\lambda_7=-.0192$, $\lambda_8=.00084$

C. Active Set F2,F5,F7,F8 or F2,F5,F7

F8 or F4 will produce an acceptably small M^2. Thus, more than one combination of functions can produce an acceptable approximation. We chose F8. Notice in iteration eight that deletion of F8 produces very little information loss, and the analyst may prefer to use active set F2,F5,F7, i.e., set λ_8=0.0. These points again indicate the need for analyst involvement in the selection procedure. All three selection methods are designed as analyst tools and not as stand-alone computer programs.

A summary of results for the three methods is provided in Table IX.IX. The three methods produce such close results that graphs of the distributions are inadequate for distinction. Thus, Table IX.X is included to provide a comparison of cumulative distribution values at 18 data points.

TABLE IX.IX

COMPARISON OF THREE ENTROPY METHODS FOR TAKEOFF DISTANCE DISTRIBUTION

Method	Approximation Interval	Active Set
Regression	5830.0, 7473.8	2,4,7,8
Divergence	5104.1, 8262.6	2,4,7
Expected Values	5104.1, 8262.6	2,5,7,8 (2,5,7)

AD-A083 514 AIR FORCE INST OF TECH WRIGHT-PATTERSON AFB OH SCHOO--ETC F/G 12/1
CONTINUOUS DENSITY APPROXIMATION ON A BOUNDED INTERVAL USING IN--ETC(U)
MAR 80 J E MILLER
UNCLASSIFIED AFIT/DS/MA/80-1 NL

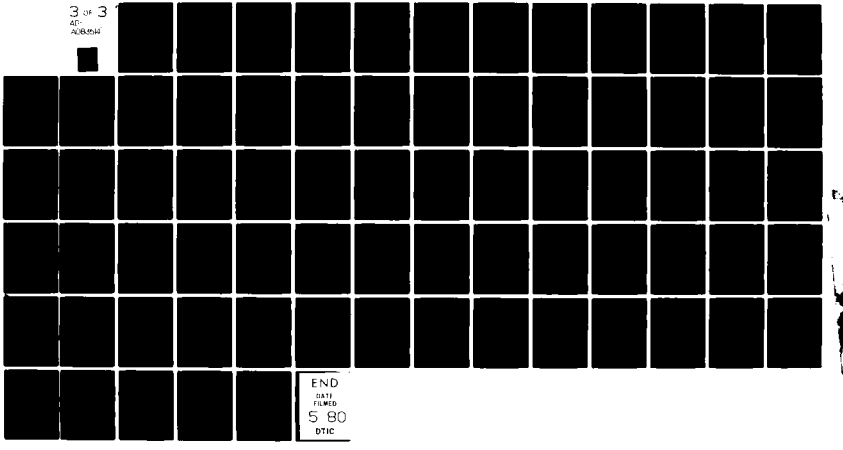


TABLE IX.X

SAMPLE AND ENTROPY CUMULATIVE DISTRIBUTIONS

I	X(I)	Cumulative Sample	Cumulative Method 1	Cumulative Method 2	Cumulative Method 3
1	5830.0	.20 (-2)	.49963(-2)	.51471(-2)	.49957(-2)
2	6083.2	.60 (-1)	.59915(-1)	.59891(-1)	.59927(-1)
3	6181.9	.12	.12141	.12108	.12142
4	6255.6	.18	.18880	.18826	.18880
5	6301.3	.24	.23981	.23920	.23981
6	6346.5	.30	.29652	.29589	.29651
7	6387.9	.36	.35282	.35225	.35281
8	6427.7	.42	.40997	.40951	.40996
9	6465.8	.48	.46619	.46589	.46618
10	6511.5	.54	.53395	.53384	.53394
11	6564.7	.60	.61109	.61122	.61109
12	6602.3	.66	.66285	.66313	.66285
13	6661.3	.72	.73752	.73797	.73753
14	6716.7	.78	.79855	.79907	.79856
15	6762.9	.84	.84206	.84257	.84208
16	6825.7	.90	.89018	.89059	.89019
17	6961.2	.96	.95595	.95604	.95595
18	7473.8	1.0	.99954	.99950	.99954

Summary

This chapter has discussed the application of our entropy procedure to computer simulation. The three active set selection methods were used on a specific example and produced consistently similar results. The choice of method is left to the analyst; previous chapters provide guidance. The entropy method provides an excellent tool for distribution characterization and a viable tool for sensitivity analysis.

The simulation application was presented in detail because such a general procedure may be applied to numerous stochastic modeling problems. The method treats the simulation like a "black box." Thus, if the analyst can formulate his problem as a special case of the "black box" model then he may apply the entropy procedure.

As a final note, the simulation input distributions were provided in our examples and must be known to implement a simulation. However, the entropy characterization procedure is also useful in defining input distributions. Chan (Ref 10) suggests such an application for a special case of our general model. Other entropy applications are discussed in Chapter XI.

Chapter X. Sensitivity

Introduction

The following sections address several aspects of sensitivity analysis. We summarize sensitivity issues from previous chapters on simulation and active set selection. The central concern of this chapter is sensitivity of the entropy approximation, $p(x)=\exp[-\lambda_0-\lambda_1 g_1(x)-\ldots\lambda_k g_k(x)]$, to errors in the expected value vector, $\langle G\rangle=(\langle g_0\rangle,\langle g_1\rangle,\ldots \langle g_k\rangle)^T$ where there are k functions in the active set. (For notational convenience only, we do not include variable x in our expected value symbols, i.e., we let $\langle g_i(x)\rangle=\langle g_i\rangle$.) We present theoretical developments from the literature as well as two numerical procedures for studying approximation sensitivity.

Simulation Sensitivity

As R. E. Shannon states,

> Sensitivity analysis is one of the most important concepts in simulation modeling. By this we mean determining the sensitivity of our final answers to the values of the parameters used [Ref 70:32].

Simulation is designed to facilitate sensitivity analysis because the analyst has complete control over the parameters (or inputs) and can vary them one at a time (or jointly) to observe the effect on simulation output. In fact, the AFFDL problem of Chapter IX was solved via

simulation to answer a sensitivity question, i.e., how do design measures vary as input design parameters change?

Our entropy procedure provides an effective tool for output comparisons by producing an accurate description of the output distribution. The inputs may then be altered and a second entropy approximation generated for density or cumulative comparisons. Frequently, graphs of the resulting output distributions will answer the analyst's sensitivity questions. While the entropy procedure provides graphs, it provides additional insight to simulation sensitivity. Notice that the entropy approximation, $p(x)$, is based on the expected value vector $\langle G \rangle$, and changes in the inputs cause subsequent changes in $\langle G \rangle$. Given a $\langle G \rangle$ vector for the output of a simulation with specified inputs, we may study the sensitivity of $p(x)$ to variations in this $\langle G \rangle$ vector, and this study is accomplished without using the simulation. Once we have established acceptable bounds for the output $\langle G \rangle$ vector, we may return to the simulation to investigate the effect on $\langle G \rangle$ of varying simulation inputs. Procedures for evaluating the sensitivity of $p(x)$ to vector $\langle G \rangle$ will be presented.

Active Set Selection

Chapters VI, VII, and VIII discuss three methods for selecting the active set of information functions from a predefined potential set. Methods one (linear regression)

and two (divergence) use random samples of the unknown distribution to develop the active set and demonstrate sample dependence. While sample sensitive, the methods produce approximations that compromise between the sample and true underlying distributions. The point of significance is that this compromising quality insures an adequate fit to subsequent samples with the active set from a previous sample. Previous chapters provide conceptual and experimental justification. Method three (expected values) selects the active set based on expected value information and is sensitive to error in this data. The three methods thus demonstrate data sensitivity. This data sensitivity is a desired property and enables accurate approximations.

The importance of specifying a broad potential set was presented in Chapter V and represents another form of sensitivity. If the potential set contains the correct functions, we can exactly recreate the unknown analytic distribution as previously demonstrated. The accuracy of approximation depends heavily on specifying enough information via the potential set. The potential set of Chapter V provides an excellent starting point and proved quite accurate in experimentation.

Approximation Sensitivity

The term "approximation sensitivity" is used here to describe the sensitivity of the approximation density,

p(x), to errors in the expected values vector <G>. We assume that the active set has been selected and the information functions are specified and fixed. Our concern centers on how changes in the expected values of the active set produce changes in p(x). We restate the problem in the Lagrange formulation, discuss theoretic implications, and then present two methods for studying this sensitivity.

The Problem. The maximum entropy procedure approximates the unknown density, f(x), by a density of the form

$$p(x) = \exp\left[-\lambda_0 - \lambda_1 g_1(x) - \ldots \lambda_k g_k(x)\right]$$

where the $g_i(x)$, i=0,1,...k, represent our active set of information functions with $g_0(x) \equiv 1$. The lambda vector, $\Lambda = (\lambda_0, \lambda_1, \ldots \lambda_k)^T$, identifies the specific p(x) and is determined by solving a system of nonlinear constraint equations:

$$F(\Lambda) = \begin{bmatrix} f_0(\Lambda) \\ \vdots \\ f_k(\Lambda) \end{bmatrix} = \begin{bmatrix} \int g_0(x) \exp[-\sum_{i=0}^{k} \lambda_i g_i(x)]dx \\ \vdots \\ \int g_k(x) \exp[-\sum_{i=0}^{k} \lambda_i g_i(x)]dx \end{bmatrix} = \begin{bmatrix} \langle g_0 \rangle \\ \\ \langle g_k \rangle \end{bmatrix} = \langle G \rangle \quad (10.1)$$

where the <G> vector is provided via quadrature, average value approximation, or other means. The only unknown is Λ. Thus our approximation problem is transformed to a problem of selecting a Λ vector, based on a given <G>, such that the resulting p(x) satisfies the constraints.

Assuming a "wise" choice of information functions, p(x) will provide an adequate representation of f(x).

Sensitivity analysis is defined as an evaluation of the change in system output effected by a systematic variation of system inputs. Our system is described by equations (10.1) with input <G> and output Λ or p(x). We thus consider sensitivity at two levels; the sensitivity of Λ to <G>, and the sensitivity of p(x) to <G>.

Theoretical Support. The Λ vector defines the explicit p(x) for a particular set of constraint values, <G>. The Lagrange multiplier formulation of our problem provides some immediate information on Λ sensitivity to <G>. As discussed by Tribus, Jaynes, Crain (Refs 82; 42; 15), and others, we may consider λ_0 as a function of λ_i, i=1,2,...k. Following Tribus' work with discrete entropy, we consider the first constraint, $f(\Lambda)=\langle g_0\rangle=1$, to produce

$$\exp[\lambda_0] = \int \exp[-\sum_{i=1}^{k} \lambda_i g_i(x)]\, dx \qquad (10.2)$$

or

$$\lambda_0(\lambda_1,\ldots\lambda_k) = \ln[\int \exp[-\sum_{i=1}^{k} \lambda_i g_i(x)]\, dx].$$

Differentiation of (10.2) produces

$$\partial\lambda_0/\partial\lambda_m = -\langle g_m\rangle \qquad (10.3)$$

Equation (10.3) states a linear relationship between λ_0 and λ_m which is weighted by $\langle g_m \rangle$. This tells us that $\lambda_0(\lambda_1,\ldots\lambda_k)$, called the "potential function" by Jaynes and Tribus, is most strongly effected by variations in the larger elements of $\langle G \rangle$. We would thus expect that $g_m(x)$ with large $|\langle g_m \rangle|$ would strongly effect $p(x)$. While providing conceptual insight, equation (10.3) provides little practical sensitivity information for our procedure. We scale the information functions in our application to reduce the values of $|\langle g_i \rangle|$, $i=1,2,\ldots k$. The scaling is a numerical convenience but also enhances the performance of the three active set selection methods. Let us pursue the theoretical relationship between Λ and $\langle G \rangle$.

Given an $\xi=(\xi_0,\xi_1,\ldots\xi_k)^T$ variation in $\langle G \rangle$ with $\xi_0=0$ ($g_0(x)\equiv 1$), we want to find $\delta=(\delta_0,\delta_1,\ldots\delta_k)^T$ where δ is the variation produced in Λ. We again extend the discrete example of Tribus (Ref 82). Consider the $p(x)$ which produces maximum entropy $S_{max}(p(x))$. Then

$$S_{max}(p(x)) = -\int p(x) \ln (p(x))\, dx$$

$$S_{max}(p(x)) = \int \sum_{i=0}^{k} \lambda_i g_i(x)\, p(x)\, dx, \text{ or}$$

$$S_{max}(\Lambda) = \sum_{i=0}^{k} \lambda_i \langle g_i \rangle = \Lambda^T \langle G \rangle \qquad (10.4)$$

As expected, the entropy is a function of the Λ and $\langle G \rangle$ vectors. If we consider the $\langle g_i \rangle$ to be the independent

variables in equation (10.4), we find $\partial S_{max}(\langle G\rangle)/\partial\langle g_j\rangle=\lambda_j$, $j=0,1,\ldots k$ or, subsequently, $\partial\lambda_j/\partial\langle g_k\rangle=\partial\lambda_k/\partial\langle g_j\rangle$. These equations highlight the interdependence of the elements of Λ and $\langle G\rangle$, i.e., a perturbation such as $\xi=(0,..0,\xi_m,0,..0)$, $\xi_m\neq 0$, may result in a δ with all nonzero elements. The theory provides insight, but fails to provide a viable means of examining the sensitivity of Λ to $\langle G\rangle$ and subsequently of $p(x)$ to $\langle G\rangle$.

Sensitivity of Λ to $\langle G\rangle$. We are given a nominal $\langle G\rangle$ vector, $\langle G\rangle_0$, and solve equations (10.1) for Λ_0 such that $F(\Lambda_0)=\langle G\rangle_0$. From Theorem 4.3, if Λ_0 exists then it is locally unique, $F(\Lambda)$ is one to one in some neighborhood of Λ_0, and F^{-1} exists with $F^{-1}(\langle G\rangle_0)=\Lambda_0$. Given perturbation ξ, we wish to find vector δ such that $F(\Lambda_0+\delta)=\langle G\rangle_0+\xi$.

The usual approach to such a problem (Refs 14; 19; 70; 80; 63) is to vary a single element of $\langle G\rangle_0$ and calculate δ. Then reset $\langle G\rangle_0$, vary a second element of $\langle G\rangle_0$, compute the respective δ, and continue to iterate. The result of this brute force linear approximation is an approximation to the partial derivatives

$$\frac{\partial\Lambda}{\partial\langle g_i\rangle_0}=\left(\frac{\partial\lambda_0}{\partial\langle g_i\rangle_0}\ ,\ \frac{\partial\lambda_1}{\partial\langle g_i\rangle_0}\ ,\ \ldots\ \frac{\partial\lambda_k}{\partial\langle g_i\rangle_0}\right)^T,\quad i=1,2,\ldots k.$$

These partials can provide an indication of the strength of a specific information function. A large value of

$\partial\lambda_j/\partial\langle g_i\rangle_0$ for one or more values of j indicates that a small error in $\langle g_i\rangle_0$ may cause a large change in Λ and thus p(x). This suggests additional effort to insure accuracy for the subject $\langle g_i\rangle_0$.

The linear approximation provides useful information about Λ sensitivity and is easily accomplished with the computer programs of Chapter IV. However, the linear approach has a recognized weakness (Refs 19; 20; 68; 80) in that constraint coupling has not been fully considered. We must include simultaneous variations of all $\langle g_i\rangle$, i=1,2,...k, $g_0(x)\equiv 1$, and observe the effect on Λ. We extend the linear investigation.

Define a k dimensional rectangle, $R_\alpha(\langle G\rangle_0)$, about vector $\langle G\rangle_0$ (k dimensions versus k+1 because $\langle g_0\rangle\equiv 1$). This rectangle is a function of parameter α where each side of the rectangle is an interval, $[\langle g_i\rangle_0-\alpha\langle g_i\rangle_0, \langle g_i\rangle_0+\alpha\langle g_i\rangle_0]$, i=1,2,...k. Thus, α denotes a confidence in our estimation of the expected values. By sampling from $R_\alpha(\langle G\rangle_0)$ and computing $\Lambda=(\Lambda_0+\delta)$, we may investigate the shape of the k+1 dimensional rectangle that is generated about Λ_0. For example, we may record the maximum deviation for each element of Λ_0 (i.e., the maximum value of δ_i for each i) that results from the allowable perturbations of $\langle G\rangle_0$. "Large" values of δ_i may suggest a reduction in α. This approach has potential use for placing α bounds on $\langle G\rangle_0$ when the maximum allowable δ_i, i=0,1,...k, are known or hypothesized.

The scheme provides a starting point for a practical attack on our central concern; how do errors in $\langle G \rangle_0$ affect the entropy approximation, p(x)?

Sensitivity of p(x) to $\langle G \rangle$. The previous sections relate sensitivity concepts from the literature and our research. They provide conceptual and theoretic insights. This section directly attacks the sensitivity question and describes a practical procedure for investigating the variation in p(x) due to errors or changes in $\langle G \rangle_0$.

As in the previous section, we are given $\langle G \rangle_0$ from which we find Λ_0 and thus p(x). We select α to define $R_\alpha(\langle G \rangle_0)$ and sample from $R_\alpha(\langle G \rangle_0)$ to produce $\langle \tilde{G} \rangle = (1., \langle \tilde{g}_1 \rangle, \ldots \langle \tilde{g}_k \rangle)^T$ where $\langle \tilde{g}_i \rangle$ is in $[\langle g_i \rangle_0 - \alpha \langle g_i \rangle_0, \langle g_i \rangle_0 + \alpha \langle g_i \rangle_0]$. Thus, $\langle \tilde{G} \rangle$ is composed of k independent, uniformly distributed random variables, $\langle \tilde{g}_i \rangle$, i=1,2,...k, and a constant, $\langle \tilde{g}_0 \rangle = 1$. Generation of the $\langle \tilde{G} \rangle$ samples is accomplished by sampling from k uniform distributions with the stated α bounds. Each $\langle \tilde{G} \rangle$ vector results in a $\tilde{\Lambda}$ which defines a $\tilde{p}(x)$, i.e., a perturbation of p(x).

The sample space, $R_\alpha(\langle G \rangle_0)$, is specified as a function of α and the expected value vector $\langle G \rangle_0$. Our procedure is designed to place confidence bounds about p(x) based on predefinition of the sample space. We generate N samples of $\langle \tilde{G} \rangle$ and produce the corresponding N densities, $\tilde{p}(x)$. Each $\tilde{p}(x)$ is a continuous density from the same

entropy family as p(x) and on the same approximation interval. To bound p(x), we specify M points on the interval of approximation and consider the maximum deviations, above and below p(x), achieved by the sample densities. This approach specifies an upper and lower bound on the nominal density, p(x), as a function of α for a given $\langle G \rangle_0$.

Figures 10.1, 10.2, and 10.3 present the results of this procedure when applied to the beta distribution of previous chapters (M=50, N=500). As expected, the bounds on p(x) grow as α increases. Similar results are shown in Figures 10.4 through 10.8 for the normal distribution; $p(x)=\exp[-\lambda_0-\lambda_1(x-\mu)^2/\sigma^2)]$, $\mu=10.$, $\sigma^2=2$. For the normal example we allowed $\langle (x-\mu)^2/\sigma^2 \rangle$ to vary. This approach is effective for determining a reasonable α bound on the $\langle G \rangle_0$ vector, i.e., for specifying an accuracy bound on the data. For the beta example we conclude that an α greater than .05 produces unsatisfactory approximation error. Other analysts may be more or less tolerant. The normal example demonstrates less sensitivity. We must insure that sufficient emphasis is placed on data collection and calculation of $\langle G \rangle_0$ to accomplish the desired level of data accuracy.

The above sensitivity model offers a practical means of evaluating the sensitivity of p(x). The results will depend on the form of p(x), i.e., the information functions in p(x), and the values of $\langle G \rangle_0$ and α. Thus, we cannot state general sensitivity results that pertain to

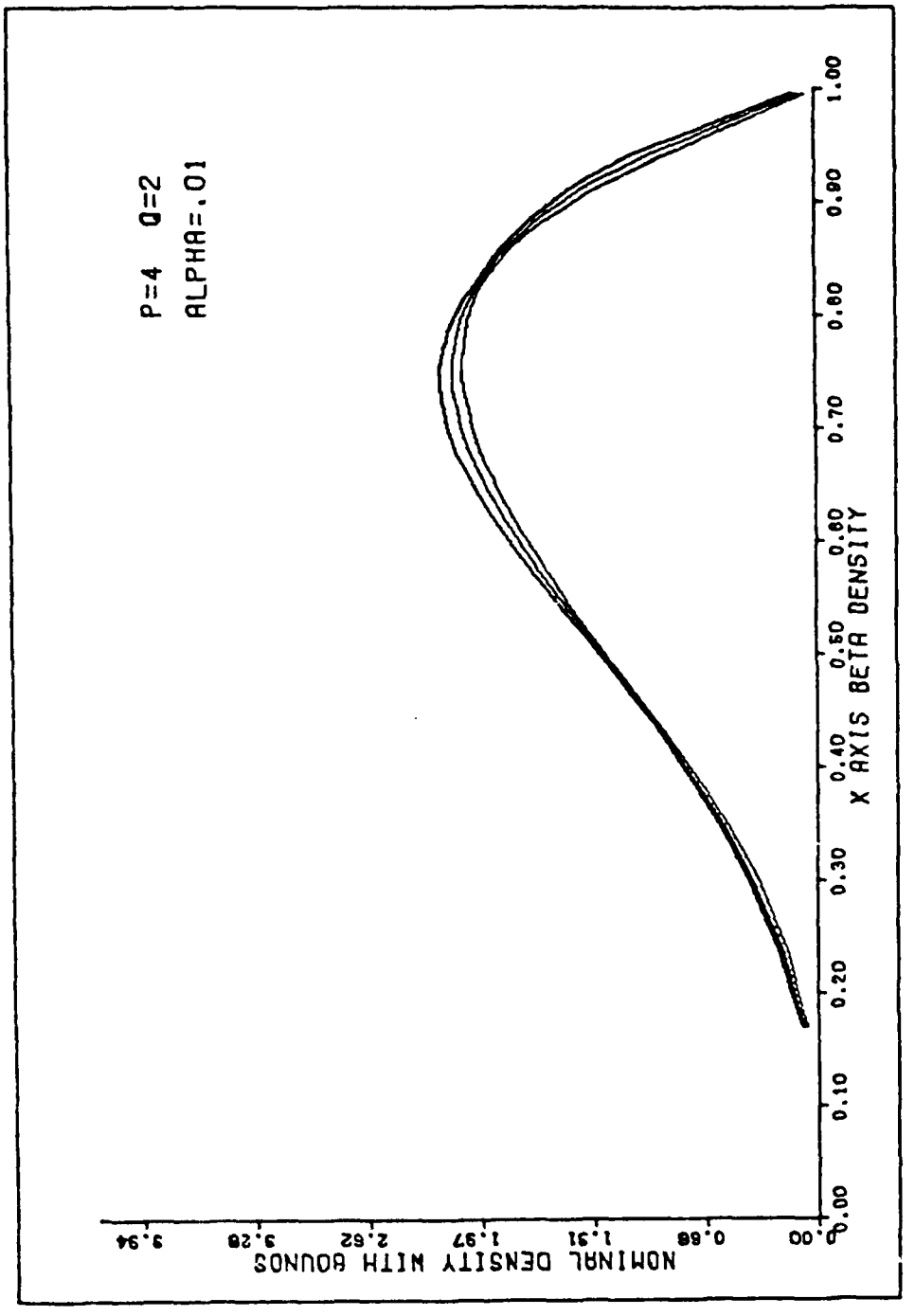


Fig. 10.1. Beta Approximation Bounds, Alpha = .01

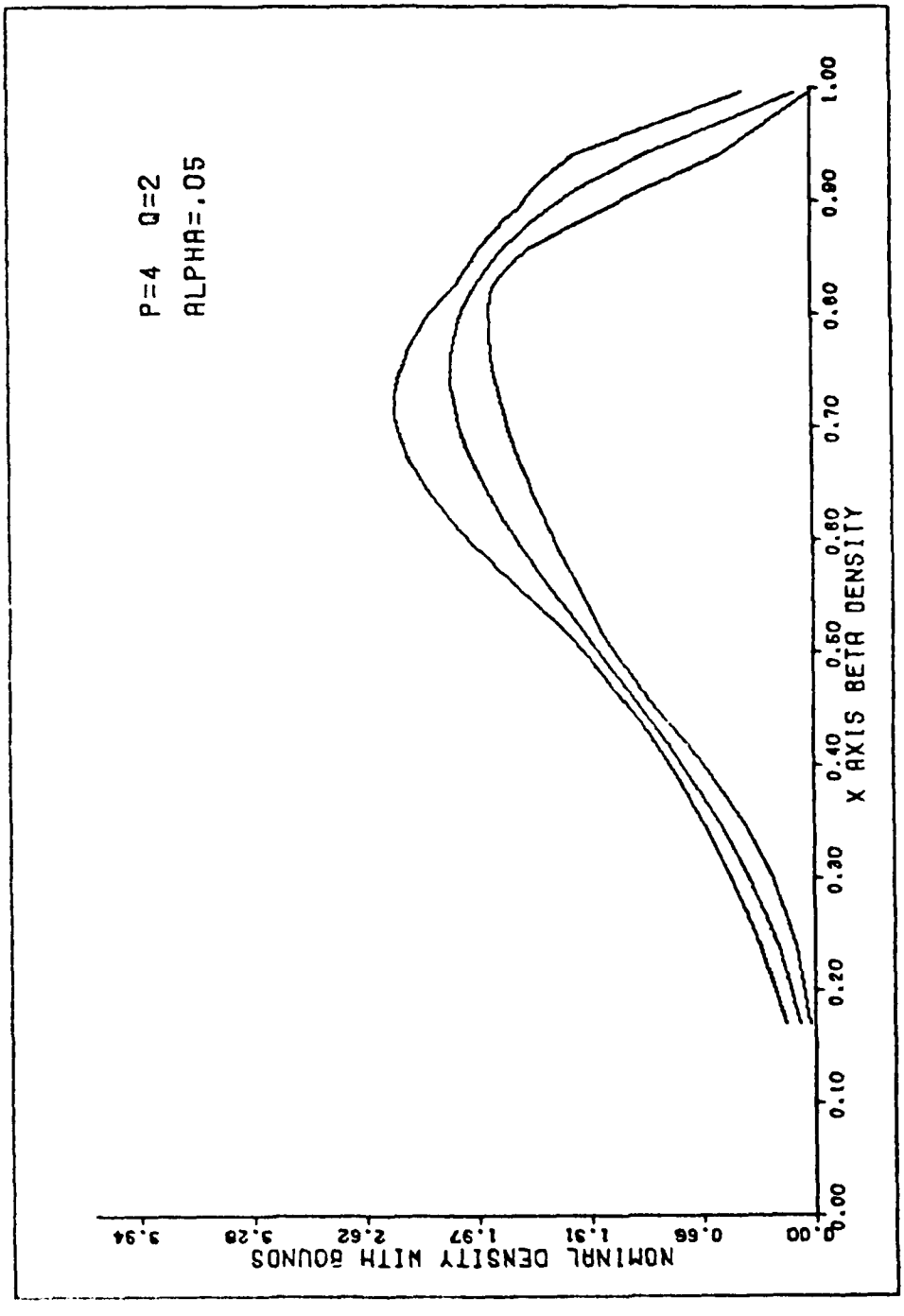


Fig. 10.2. Beta Approximation Bounds, Alpha = .05

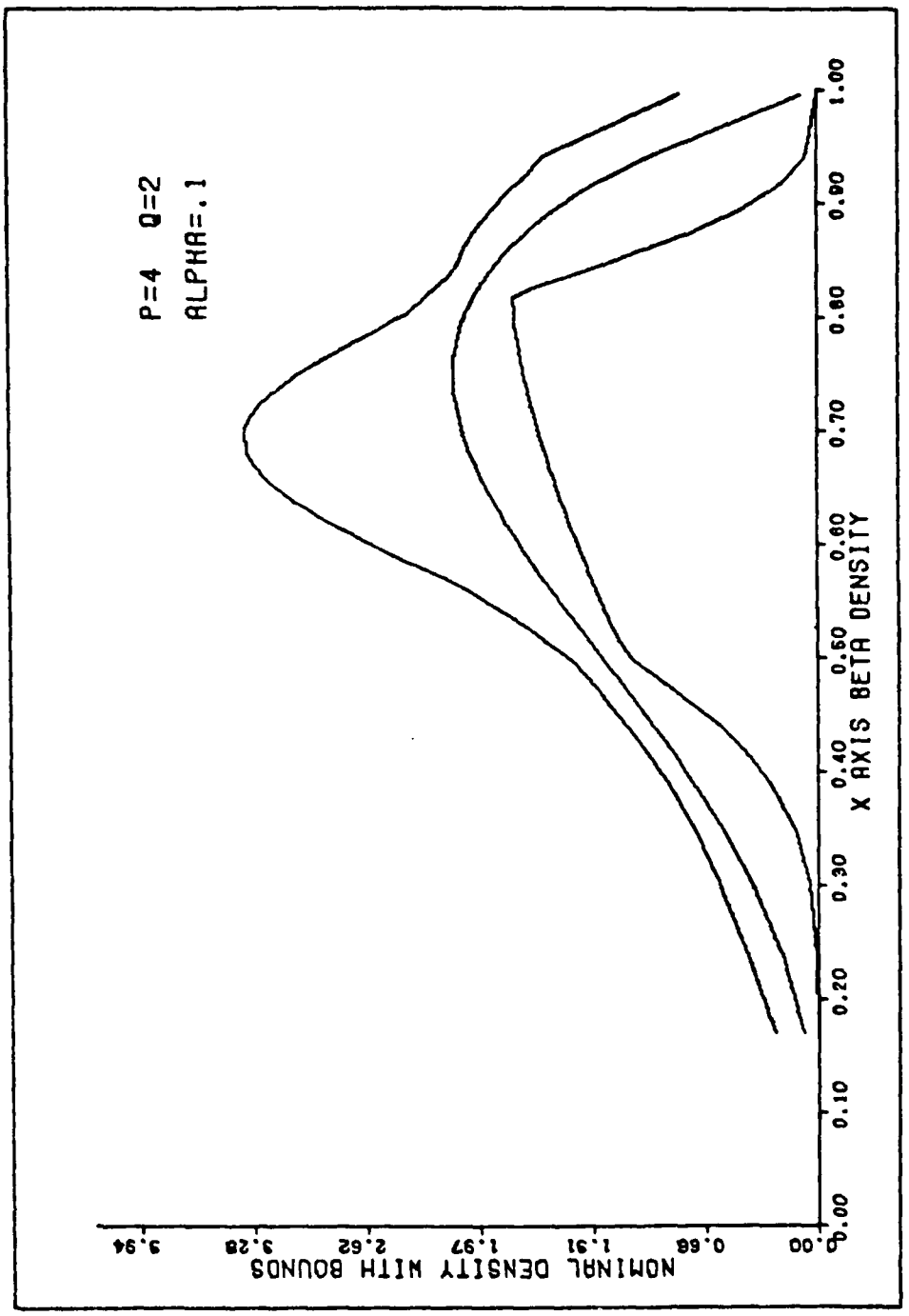


Fig. 10.3. Beta Approximation Bounds, Alpha = .1

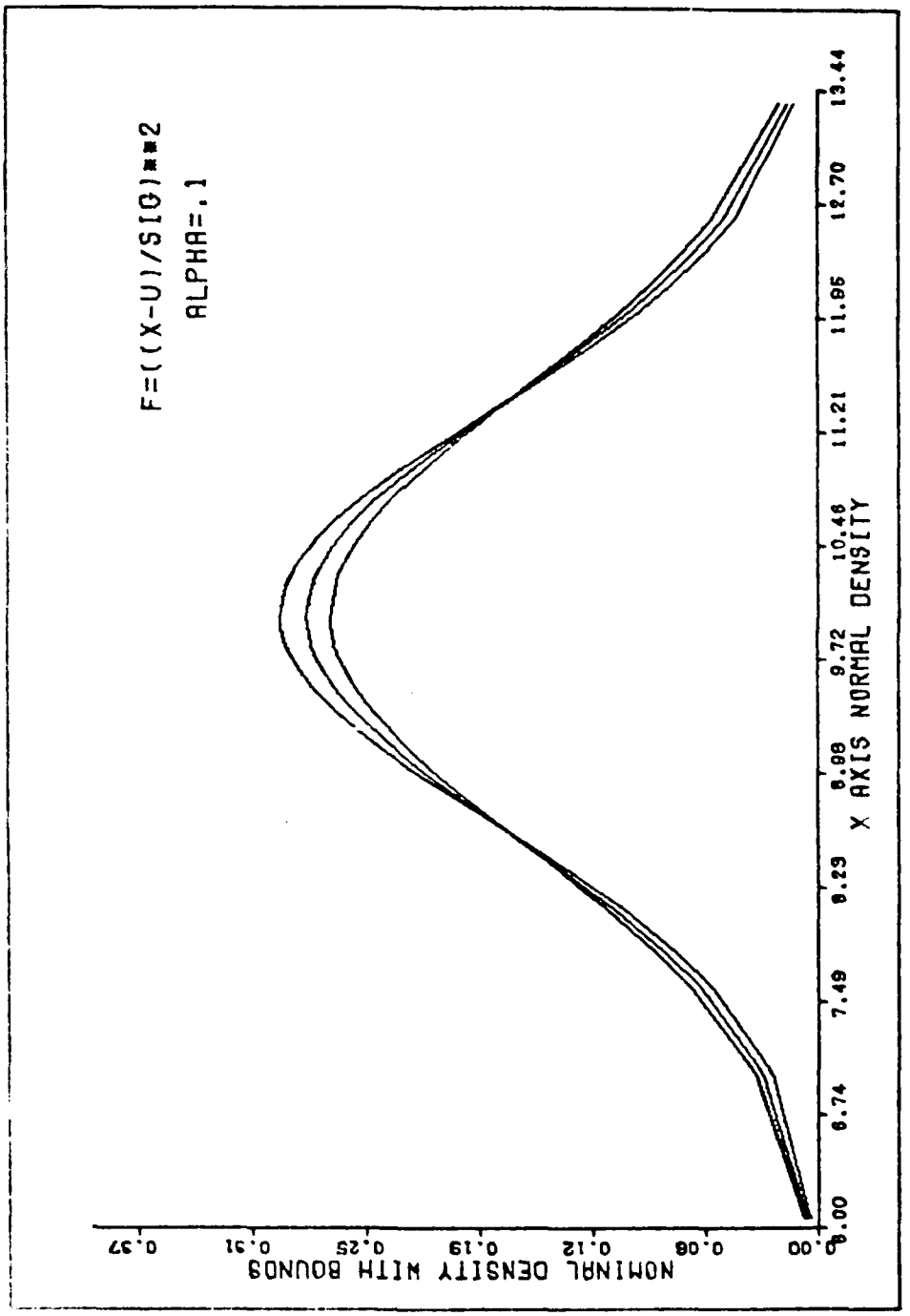


Fig. 10.4. Normal Approximation Bounds, Alpha = .1

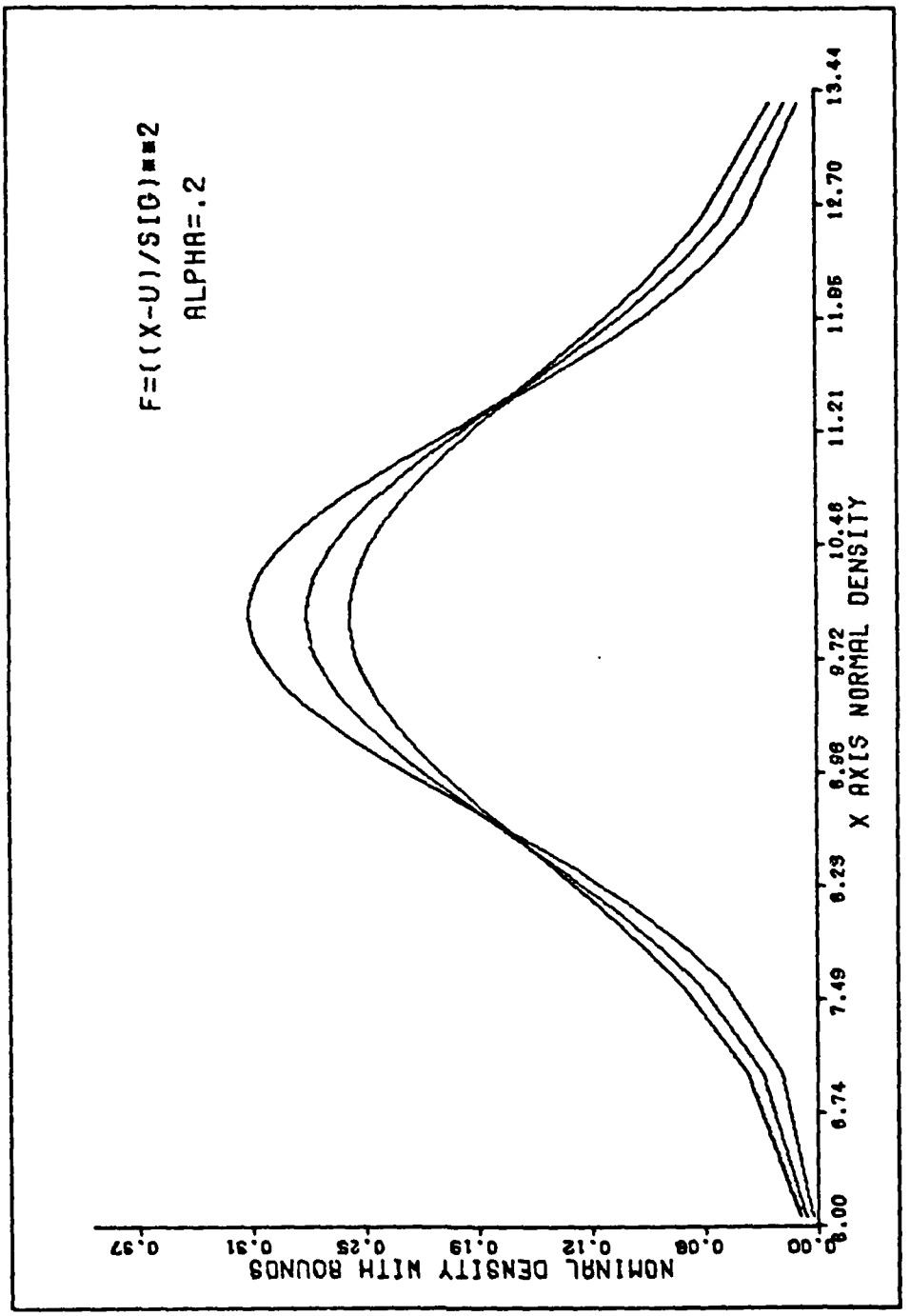


Fig. 10.5. Normal Approximation Bounds, Alpha = .2

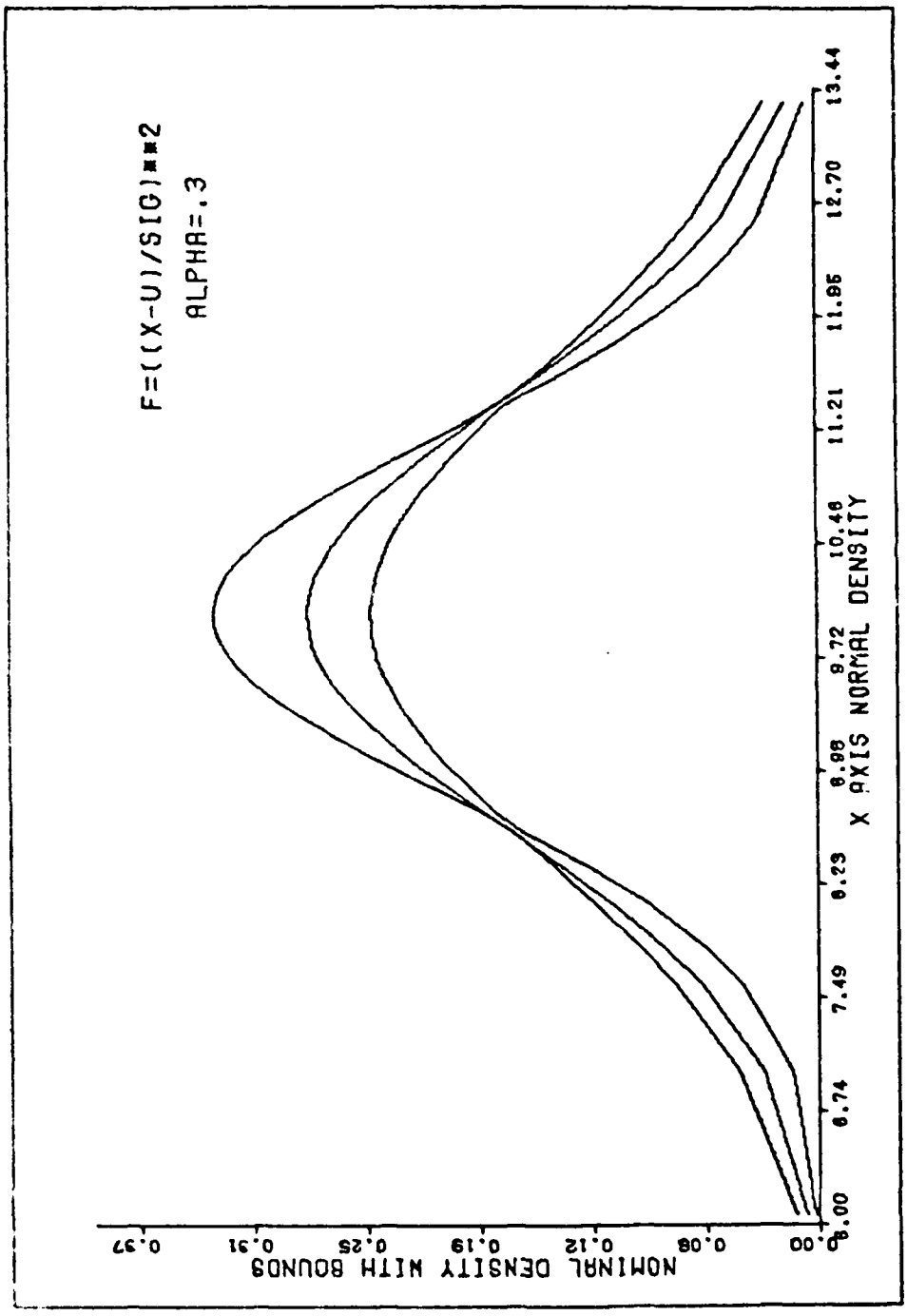


Fig. 10.6. Normal Approximation Bounds, Alpha = .3

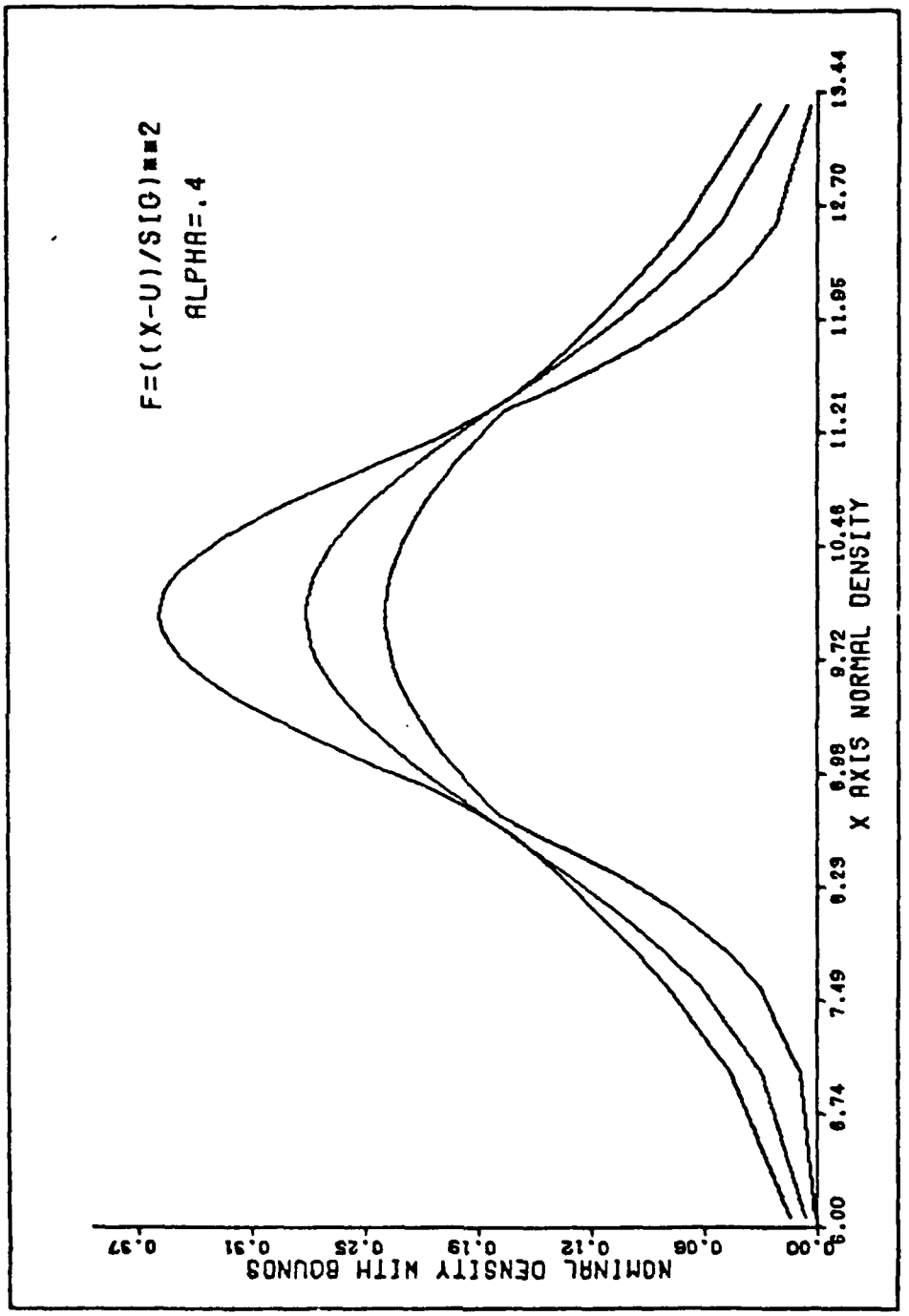


Fig. 10.7. Normal Approximation Bounds, Alpha = .4

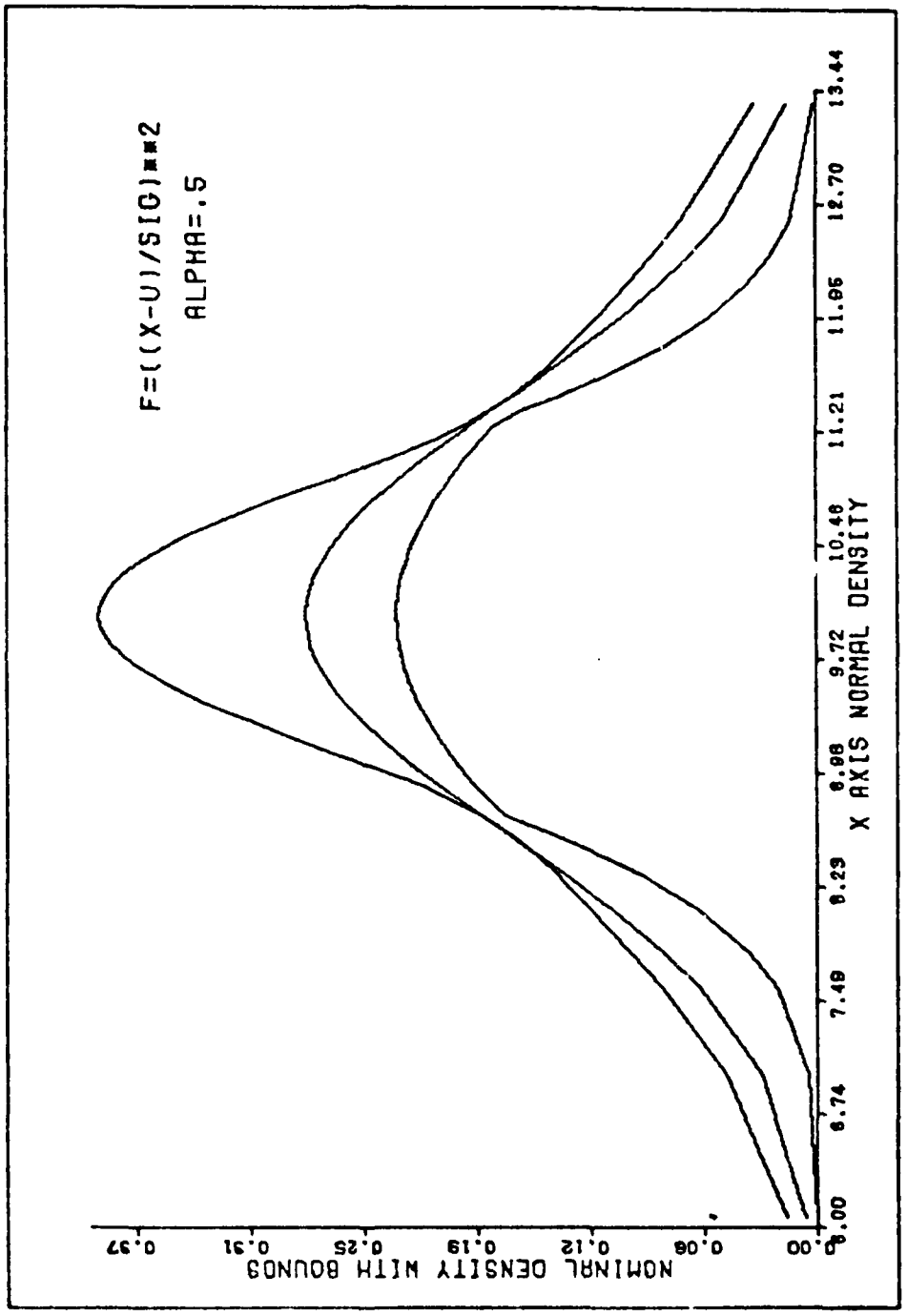


Fig. 10.8. Normal Approximation Bounds, Alpha = .5

all approximations or to all problems. Instead, we provide a procedure which the analyst may use on his specific problem. Sensitivity conclusions derived from this procedure will be subjective. The procedure is easy to implement and, once $\langle G \rangle_0$ is specified, does not require additional access to the unknown distribution. The entropy approach has thus provided sensitivity insight that other characterization procedures lack.

Various modifications of our sensitivity model are feasible. For example, the analyst may consider maximum approximation error, i.e., $\max_x |p(x)-\tilde{p}(x)|$. He may then select α bounds that insure $\max_x |p(x)-\tilde{p}(x)|$ less than some epsilon. Secondly, the effect of a single expected value $\langle g_m \rangle$, may be evaluated by using the same approach but with other elements of $\langle G \rangle_0$ fixed. A third variation results if additional information exists about the accuracy of $\langle G \rangle_0$. For example, we have assumed a uniform error for each element of $\langle G \rangle_0$. The analyst may know that the error is better approximated by another distribution. The above procedure may still be used but with $\langle \tilde{G} \rangle$ vectors produced from the known error distribution. Finally, the analyst may prefer to compare entropy cumulative distributions, $P(x)$ and $\tilde{P}(x)$, versus densities. The same sensitivity model may be used but with bounds now produced for $P(x)$. We consider a second model for sensitivity investigations in the next section.

Entropy Approximation for Sensitivity Measures

A procedure to investigate the sensitivity of p(x) to error in $\langle G \rangle_0$ has been presented. We use this sensitivity model, information theory concepts, and our entropy approximation procedure to develop a second sensitivity model. The second model demonstrates another application of our entropy approximation procedure.

In previous sections we defined the given data, $\langle G \rangle_0$, which produces entropy approximation p(x). A perturbation of $\langle G \rangle_0$ produces vector $\langle \tilde{G} \rangle$ which results in a second density, $\tilde{p}(x)$. How may we "measure" the variation in p(x) due to the perturbation in $\langle G \rangle_0$? We have

$$p(x) = \exp\left[-\sum_{i=0}^{k} \lambda_i g_i(x)\right] \text{ and}$$

$$\tilde{p}(x) = \exp\left[-\sum_{i=0}^{k} \beta_i g_i(x)\right]$$

where $\beta_i = \lambda_i + \xi_i$ and $\langle \tilde{g}_i \rangle = \langle g_i \rangle_0 + \xi_i$, $i=0,1,\ldots k$; $\xi_0 = 0$. A useful measure of variation between densities is divergence, $J(p(x),\tilde{p}(x))$ (Chapters II and VII). Thus,

$$J(p(x),\tilde{p}(x)) = \int [p(x)-\tilde{p}(x)] \ln [p(x)/\tilde{p}(x)]\, dx$$

$$= \int [p(x)-\tilde{p}(x)][-\Sigma\lambda_i g_i(x) + \Sigma\beta_i g_i(x)]\, dx$$

$$= \int \Sigma(\beta_i-\lambda_i) g_i(x) p(x)\, dx - \int \Sigma(\beta_i-\lambda_i) g_i(x) \tilde{p}(x)\, dx$$

$$= \Sigma(\beta_i-\lambda_i)\langle g_i\rangle_0 - \Sigma(\beta_i-\lambda_i)\langle \tilde{g}_i\rangle$$

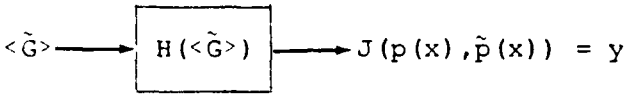
$$= \Sigma(\beta_i-\lambda_i)(\langle g_i\rangle_0 - \langle \tilde{g}_i\rangle)$$

$$= \Sigma(\lambda_i+\delta_i-\lambda_i)(\langle g_i\rangle_0-\langle g_i\rangle_0-\xi_i)$$

$$J(p(x),\tilde{p}(x)) = \Sigma(\delta_i)(-\xi_i) = -\delta^T\xi \qquad (10.5)$$

where all summations are for i=0 to k. Equation 10.5 allows rapid computation of divergence and thus a measure of information loss when p(x) is replaced by $\tilde{p}(x)$.

Combining the divergence measure with the concepts of sensitivity model one, we create a second sensitivity model in the form of a simulation:



Vector $\langle\tilde{G}\rangle$ is an independent, multivariate, uniformly distributed input random vector (that depends on α), and $J(p(x),\tilde{p}(x))$ is the univariate simulation output. Chapter IX discussed the use of our entropy approximation procedure for simulations of this form. Given α which defines our input distribution, we wish to determine the sensitivity of p(x) in terms of the measure $J(p(x),\tilde{p}(x))$. Application of our entropy procedure to the output of model two provides a complete representation of the density of $J(p(x),\tilde{p}(x))$, $p_J(y)$, for the given α. Knowledge of $p_J(y)$ enables the development of confidence bounds and statistical statements about the divergence, i.e., the probability that

divergence is less than Q for the given α is $\int_0^Q p_J(y)\,dy$. Since divergence measures variation in p(x), then divergence is a viable sensitivity measure. Thus, the simulation approach has provided a means to numerically quantify the sensitivity question.

As an example, the model was used for the beta distribution or previous sections with α=.1. We generated 500 sample input vectors, $\langle\tilde{G}\rangle$, and calculated the corresponding divergence values. Method three (Chapter VIII) was applied for the entropy characterization of $p_J(y)$. Figure 10.9 shows the entropy and sample divergence densities. The sample density was computed by numerical differentiation of the sample cumulative. With $p_J(y)$ known, we consider the impact of errors in $\langle G\rangle_0$; i.e., given that $\langle G\rangle_0$ produces p(x) and $\langle\tilde{G}\rangle$ produces $\tilde{p}(x)$, the probability that the divergence between $\tilde{p}(x)$ and p(x) is less than .05 for all $\langle\tilde{G}\rangle$ in $R_\alpha(\langle G\rangle_0)$ is $\int_0^{.05} p_J(y)\,dy$. For our example, $\text{Prob}(J\leq.05)=.608$ and $\text{Prob}(J\leq.5)=.983$ where $\text{Prob}(J<\infty)=1$. Experimentation has shown that a divergence of .05 produces an "acceptable" fit between p(x) and $\tilde{p}(x)$. For our example, however, we have a 40% chance of exceeding a .05 divergence with α of .1. As with the first sensitivity model, we conclude that α=.1 is too large.

The sensitivity model was described in terms of the divergence measure. Divergence provides an excellent relative measure, and we know that divergence "near" zero

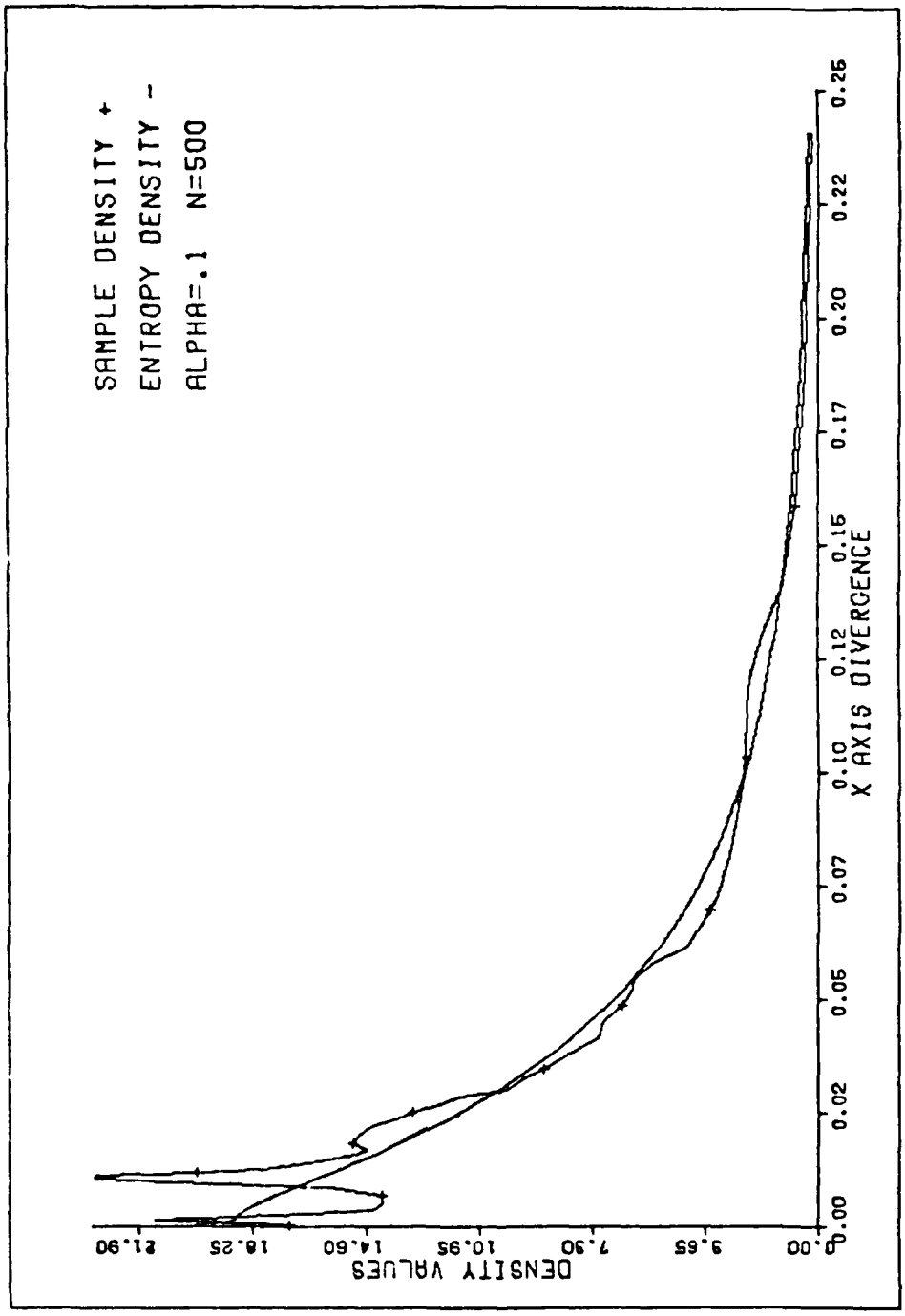
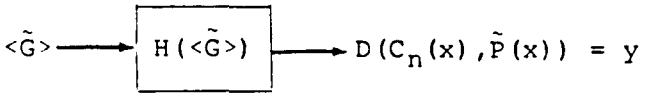


Fig. 10.9. Divergence Density Approximations

indicates "small" information exchange. We may readily determine which elements of $\langle G \rangle_0$ are most influential by systematically varying a single element and calculating divergence. The element which produces the largest divergence is the most influential. We may also compare divergence densities or probability statements for different values of α. For example, we may require $\text{Prob}(J \leq .05) > .90$ and experiment to find the appropriate α. The weakness in our model is that we cannot define a statistical meaning for a given value of divergence, i.e., is $J(p(x),\hat{p}(x))=.1$ an acceptable error? However, the sensitivity model was designed for flexibility and provides an alternative.

Divergence may be replaced with more popular measures in the sensitivity model. The goodness of fit measures of Chapter VI and Appendix B (Kolmogorov-Smirnov, etc.) are examples of measures that provide better statistical quantification of sensitivity information. We consider the Kolmogorov-Smirnov statistic as an example, $D=\sup_x |C_N(x)-P(x)|$ where $C_N(x)$ represents the sample cumulative of size N for the unknown distribution. The D statistic may be used to test the hypothesis that $C_N(x)$ was taken from distribution P(x) where P(x) is the entropy cumulative. Thus, we are given $\langle G \rangle_0$ and sample $C_N(x)$ from an unknown distribution. Vector $\langle G \rangle_0$ produces the approximation density p(x) and subsequently P(x). We define $R_\lambda(\langle G \rangle_0)$ as before and select $\langle \hat{G} \rangle$ from $R_\lambda(\langle G \rangle_0)$.

Vector $\langle\tilde{G}\rangle$ produces $\tilde{p}(x)$, and we may calculate $D(C_n(x),\tilde{P}(x))$ and $D(C_n(x),P(x))$. Our sensitivity model is as follows:



We may again apply the entropy approximation procedure to produce the density $p_D(y)$. Thus, for a given sample, $C_n(x)$, and a given α, we are able to make significant probability statements, i.e., $\text{Prob}(D \leq Q) = \int_0^Q p_D(y)\,dy$. If Q represents the critical value of the D statistic for a given significance level, then we have calculated the probability of not exceeding Q for all $\langle\tilde{G}\rangle$ which are elements of $R_\alpha(\langle G\rangle_0)$. We may thus relate the error in $\langle G\rangle_0$, determined by α, to a probability of accepting the hypothesis that a sample from the unknown distribution is a sample from our approximation distribution. While the D and other statistics are more difficult to calculate than divergence, the statistical meaning that results may be worth the effort.

Summary

We have touched on several aspects of sensitivity while concentrating on sensitivity of the approximation density, $p(x)$, to errors in the expected value vector, $\langle G\rangle_0$. Two general sensitivity models were explored with examples. One model resulted in an upper and lower bound on $p(x)$ as a function of an α error bound on $\langle G\rangle_0$. The

second model was cast in a "black box" simulation mold for use with various measures of error between distributions. Both models provide viable procedures for evaluation of system sensitivity.

Chapter XI. Other Applications

Introduction

The entropy procedure, as presented in Chapter III, is a flexible tool for characterizing or approximating unknown distributions. Application of this procedure to computer simulation has been discussed. However, the generality and flexibility of the method enables wide application. Three examples are presented in this chapter to demonstrate potential use and with the intent of stimulating thought for other applications.

Cumulative Data Versus Expected Values

The entropy procedure provides an approximation of an unknown distribution based on information about that distribution. The entropy approximation is "minimally prejudiced" in that only the available information is used and "maximum uncertainty" is maintained with respect to other information. Our development used information in the form of expected values of certain information functions. However, the entropy procedure may be applied when the available information takes other forms. As an example, the analyst may encounter distributions where the cumulative probability function is known or can be estimated at a finite number of points and expected values cannot be

estimated. We show that, using only the cumulative information, the resulting "minimally prejudiced" distribution is a piecewise uniform distribution.

Our total information consists of values of the cumulative distribution, C_i, at n points, x_i, i=1,2,...n, where $a \leq x_1 \leq x_2 \leq \ldots \leq x_n \leq b$ and [a,b] is the approximation inverval. Following the development of Chapter IV, we state the characterization problem:

$$\max S(p(x)) = \max \left[-\int_a^b p(x) \ln (p(x))\, dx\right]$$

subject to:

$$\int_a^b p(x)\, dx = 1$$

$$\int_a^{x_1} p(x)\, dx = \int_a^b \psi_1(x)\, p(x)\, dx = C_1$$

$$\vdots$$

$$\int_a^{x_n} p(x)\, dx = \int_a^b \psi_n(x)\, p(x)\, dx = C_n$$

where $\psi_i(x) = 1, \quad a \leq x \leq x_i$

$= 0$, otherwise.

The Lagrangian becomes

$$L(p(x),\Lambda) = \int_a^b -p(x) \ln p(x) - \lambda_0 p(x) - \Sigma\lambda_i \psi_i(x) p(x)\, dx + \Sigma\lambda_i C_i$$

$$= \int_a^b p(x) [\ln(1/p(x)) - \lambda_0 - \Sigma\lambda_i \psi_i(x)]\, dx + \Sigma\lambda_i C_i$$

$$= \int_a^b p(x) \ln\{(1/p(x)) \exp[-\lambda_0 - \Sigma\lambda_i \psi_i(x)]\}\, dx + \Sigma\lambda_i C_i$$

where $\Lambda = (\lambda_0, \lambda_1 \ldots \lambda_n)^T$ and all summations are from i=1 to n. We recall that $\ln(w) \leq w-1$ for all w and $\ln(w) = w-1$ if and only if w=1. Thus,

$$L(p(x),\Lambda) \leq \int_a^b p(x)\{(1/p(x)) \exp[-\lambda_0 - \Sigma\lambda_i\psi_i(x)] - 1\}\,dx + \sum_{i=1}^{n} \lambda_i C_i$$

Since we wish to maximize $L(p(x),\Lambda)$, we seek equality which occurs if and only if

$$p(x) = \exp[-\lambda_0 - \sum_{i=1}^{n} \lambda_i \psi_i(x)] \quad \text{almost everywhere.} \quad (11.1)$$

Thus, p(x) is a uniform distribution between each of the known x_i, i=1,...n; that is,

$$p(x) = \exp[-\lambda_0 - \sum_{i=1}^{k} \lambda_i \psi_i(x)], \quad a \leq x \leq x_k \qquad (11.2)$$

and p(x) is a piecewise uniform distribution.

The Lagrange multipliers are easily calculated as we show with a numerical example for n=3. Table XI.I and Figure 11.1 present the data and the interval of action for each $\psi_i(x)$ function. From the constraints we have

$$\int_a^b p(x)\,dx = \int_a^{x_1} p(x)\,dx + \int_{x_1}^{x_2} p(x)\,dx + \int_{x_2}^{x_3} p(x)\,dx + \int_{x_3}^{b} p(x)\,dx$$

Working backward from point b we find;

TABLE XI.I

SAMPLE CUMULATIVE VALUES

Symbol	Value	Cumulative
a	1.0	0
x_1	1.25	.1
x_2	1.50	.3
x_3	1.75	.6
b	2.0	1.0

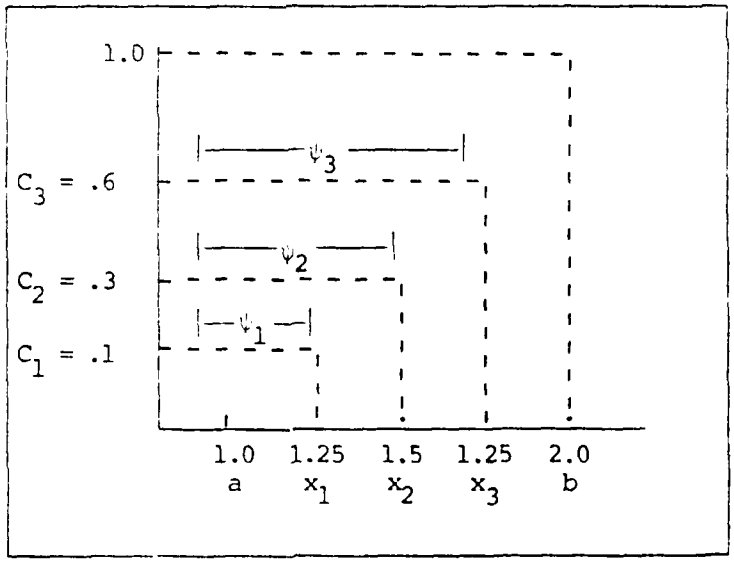


Fig. 11.1. Sample Cumulative Data

$$\int_{x_3}^{b} \exp[-\lambda_0]\, dx = 1 - C_3 \text{ or}$$

$$\lambda_0 = -\ln[(1-C_3)/(b-x_3)] = -.47004,$$

$$\int_{x_2}^{x_3} \exp[-\lambda_0-\lambda_3]\, dx = C_3 - C_2 \text{ or}$$

$$\lambda_3 = -\lambda_0 - \ln[(C_3-C_2)/(x_3-x_2)] = .287682,$$

$$\int_{x_1}^{x_2} \exp[-\lambda_0-\lambda_2-\lambda_3]\, dx = C_2 - C_1, \text{ or}$$

$$\lambda_2 = -\lambda_0-\lambda_3-\ln[(C_2-C_1)/(x_2-x_1)] = .405465,$$

and

$$\int_{a}^{x_1} \exp[-\lambda_0-\lambda_1-\lambda_2-\lambda_3]\, dx = C_1 \quad \text{to find}$$

$$\lambda_1 = -\lambda_0-\lambda_2-\lambda_3 - \ln[C_1/(x_1-a)] = .693147.$$

Thus, from equation 11.2:

$$p(x) = \begin{cases} \exp[-.916290] & 1.0 \le x \le 1.25, \\ \exp[-.223144] & 1.25 < x \le 1.5, \\ \exp[.182322] & 1.5 < x \le 1.75, \\ \exp[.470004] & 1.75 < x \le 2.0, \\ 0 & \text{otherwise.} \end{cases}$$

This simple example illustrates the concept. The entropy density reproduces the known information, i.e., the cumulative values, but does not bias our approximation in any other sense. By providing more information, such as

expected values, we increase the accuracy of approximation.

Hierarchical Models

Our entropy characterization procedure may prove quite useful in the analysis of hierarchical models. "Hierarchical model" implies a group of submodels of different degrees of detail (perhaps computer simulations) where the outputs of the more detailed submodels provide input to "higher level" submodels. As a typical example, one might envision a large scale air war game where the first modeling echelon is divided into models for the various theaters of operation. The second echelon supports the first level with models for air engagements, interdiction, or air defense. Subsequent levels provide detailed models of munitions supply, aircraft maintenance, tanker support, targeting, etc.

The entropy procedure provides a means to evaluate a hierarchical model at the submodel level. In fact, the procedure enables characterization of the output distribution of a submodel. Potential benefits include evaluation of the "degree of influence" of a specific submodel and submodel dependencies. If the models are computer simulations, then the procedure may also save substantial computer time.

Interval Arithmetic

As a third example, we consider application of the entropy procedure to interval arithmetic. Moore discusses interval arithmetic and potential uses in his article "Bounding Sets in Function Spaces with Applications to Nonlinear Operator Equations" (Ref 62). As the name implies, interval arithmetic is the application of arithmetic operations to intervals of the real line. The purpose is to specify a bound on the result of an operation. Interval operations are defined as follows:

addition, $[a,b] + [c,d] = [a+c,b+d]$;

subtraction, $[a,b] - [c,d] = [a-d,b-c]$;

multiplication, $[a,b] * [c,d] = [\min(ac,ad,bc,bd), \max(ac,ad,bc,bd)]$;

and division, $[a,b]/[c,d] = [a,b] * [(1/d),(1/c)]$.

Moore mentions several areas for application of interval arithmetic techniques: search procedures, safe starting regions and stopping criteria for iterative schemes (Ref 61), and error bounds for machine computation. The basic assumptions are that each operand is a bounded interval (as [a,b]), and that the value of the operand can fall anywhere within the stated bound. Interval arithmetic provides an absolute bound for an operation but provides no information about the distribution of the result. When working with the resulting interval, one must assume a uniform

distribution or, if an explicit value is needed, assume the middle or mean value.

Clearly, the usefulness of interval arithmetic is enhanced if the distribution of the result is known in addition to the absolute bound. In fact, knowledge of the resulting distribution may enable reduction in size of the resulting interval, with a selected degree of confidence, or selection of a more accurate mean value. Our entropy procedure can provide the desired distribution approximation. We demonstrate with the model of Figure 11.2.

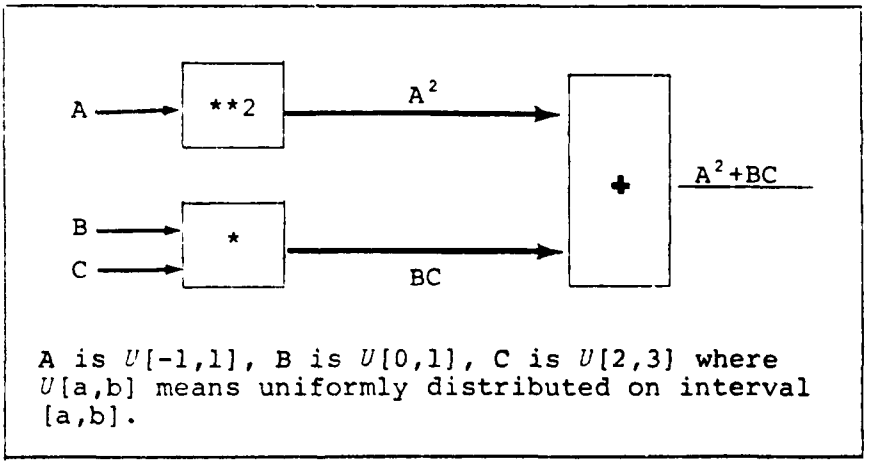


Fig. 11.2. Hierarchical Model of A^2+BC

Figure 11.2 provides a hierarchical scheme for the interval equation A^2+BC where variables A, B, and C are uniformly distributed as indicated. (This example also demonstrates the use of entropy to evaluate submodels of a simplified hierarchical model.) Applying Moore's

operations on the given intervals we find the following: A*A is $U[-1,1]$, B*C is $U[0,3]$, and AA+BC is $U[-1,4]$. We can improve these bounds by refining rules for the squaring operation, i.e., $A*A=A^2$ must be nonnegative. Our improved bounds follow: A^2 is $U[0,1]$, BC is $U[0,3]$, and A^2+BC is $U[0,4]$. In reality, the three variables (A^2, BC, and A^2+BC) are not uniformly distributed. Using transformation of variables techniques (Ref 37), the analytic distributions of these variables may be derived. For example, the density of A^2 given that A is $U[-1,1]$ is $f(A^2)=\ln(1.5)/\sqrt{A^2}$, $0<A^2\leq 1$. However, analytic derivation becomes increasingly difficult as operand distributions become more complex. The entropy approach offers a viable alternative.

Application of the entropy procedure to the problem of Figure 11.2 produces an interesting demonstration of procedure flexibility. We first generate 500 samples each for A, B, and C from independent uniform distributions and generate subsequent sample distributions for A^2, BC, and (A^2+BC). For each output distribution, we then calculate average value estimates, i.e., $\langle g_i(x)\rangle=\sum_j g_i(x_j)/500$, for the expected values of our potential information functions. Application of method three (expected value method of Chapter VIII) produces the results listed in Table XI.II. Method three provides an excellent approximation for A^2, but notice the large errors in Table XI.II for BC and (A^2+BC). Method three does not produce an acceptable fit

TABLE XI.II

RESULTS OF METHOD 3 FOR INTERVAL ARITHMETIC

Variable	Active Set	Squared Error
A^2	F2,F3,F7,F8	.001077
BC	F3	3.25683
(A^2+BC)	F5	4.53455

where Squared Error = $\sum_{i=1}^{9} (g_i(x)-\tilde{g}_i(x))^2$ and information functions are defined in Table V.III.

to the average value vectors for these two variables and, thus, does not provide acceptable distribution approximations. The large errors indicate two possible problems; either the potential set of information functions is inadequate (Ref Chapter V), or our data, i.e., the average value vector, is too inaccurate. Examination of the sample distributions (graphs are presented in Figures 11.3 through 11.8) does not indicate extreme behavior, i.e., bimodal or peaked distributions, thus our potential set seems appropriate. We investigate the expected value approximations by analytically computing a few expected values for the A^2 and BC distributions. Table XI.III presents a comparison which highlights the error between average and true expected values. As the table shows, we have tried to approximate

TABLE XI.III

AVERAGE VALUES VERSUS ANALYTIC EXPECTED VALUES

Variable	Function	Average Value	Analytic Value
$A^2=x$	$\langle x \rangle$	.339319	.333333
$A^2=x$	$\langle (x-\mu)^2 \rangle$	.090752	.088889
$A^2=x$	$\langle (x-\mu)/\sigma \rangle$	.7 (-12)	.0
$A^2=x$	$\langle \ln(x) \rangle$	-1.99329	-2.0
$A^2=x$	$\langle (x-\mu)^3/\sigma^3 \rangle$	.613545	.638877
$A^2=x$	$\langle (x-\mu)^4/\sigma^4 \rangle$	2.10257	2.14286
$A^2=x$	$\langle (\ln(x-a))^2 \rangle$	7.9370	8.0
BC=x	$\langle x \rangle$	1.23544	1.25
BC=x	$\langle (x-\mu)^2 \rangle$	.556546	.548611
BC=x	$\langle (x-\mu)/\sigma \rangle$	.1 (-11)	.0
BC=x	$\langle \ln(x) \rangle$	-.129399	-.090458
BC=x	$\langle (x-\mu)^3/\sigma^3 \rangle$	.168989	.128175
BC=x	$\langle (x-\mu)^4/\sigma^4 \rangle$	2.02580	1.97674

where μ = mean and σ = standard deviation

a distribution based on rather inaccurate information. Method three performed properly for the supplied data.

The flexibility of our procedure is significant at this point in that we have two options for increasing the accuracy of approximation. First, we can produce more accurate expected value estimates via quadrature (or averages with a larger sample) and reapply method three. Secondly, we can use the available average values but decrease their significance by concentrating on a fit to the sample distributions. We choose the second option and apply active set selection method two (divergence approach of Chapter VII) which takes advantage of sample distribution availability.

The results of method two are displayed in Table XI.IV. The accuracy of our approximations is shown in Figures 11.3 through 11.8 which provide sample and entropy comparisons for densities and cumulative distributions. The cumulative graphs include the uniform cumulative to highlight the error of assuming a uniform output distribution. Notice that the entropy and sample cumulatives are plotted over the sample points, x_i, i=1,...500, while the uniform is plotted for the entire active interval, [a,b]. Though not plotted, the entropy approximation, p(x), applies over the entire interval, and $\int_a^b p(x)\,dx = 1$. The entropy procedure clearly provides acceptable approximations to the output distributions.

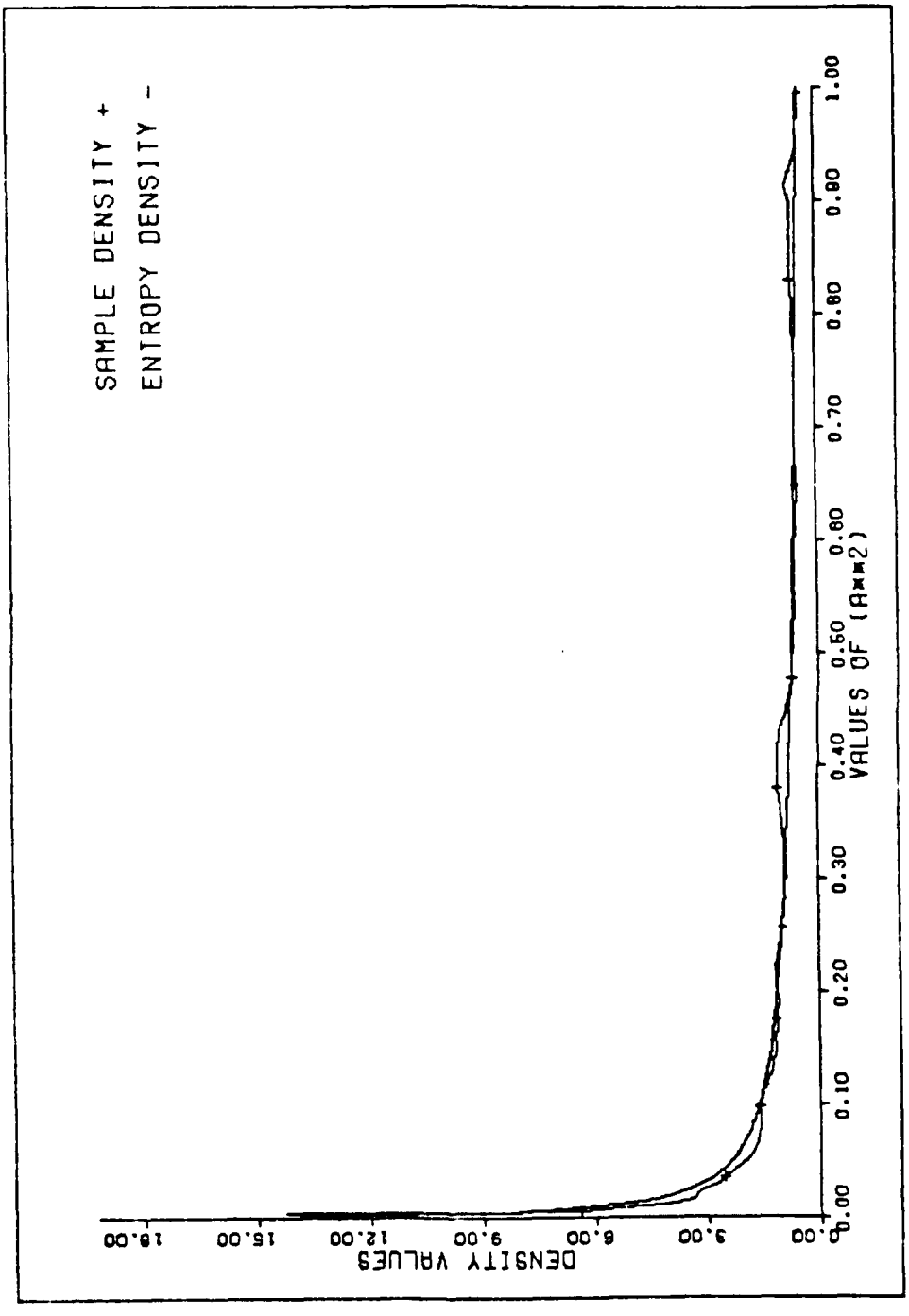


Fig. 11.3. Sample and Entropy Densities (A**2)

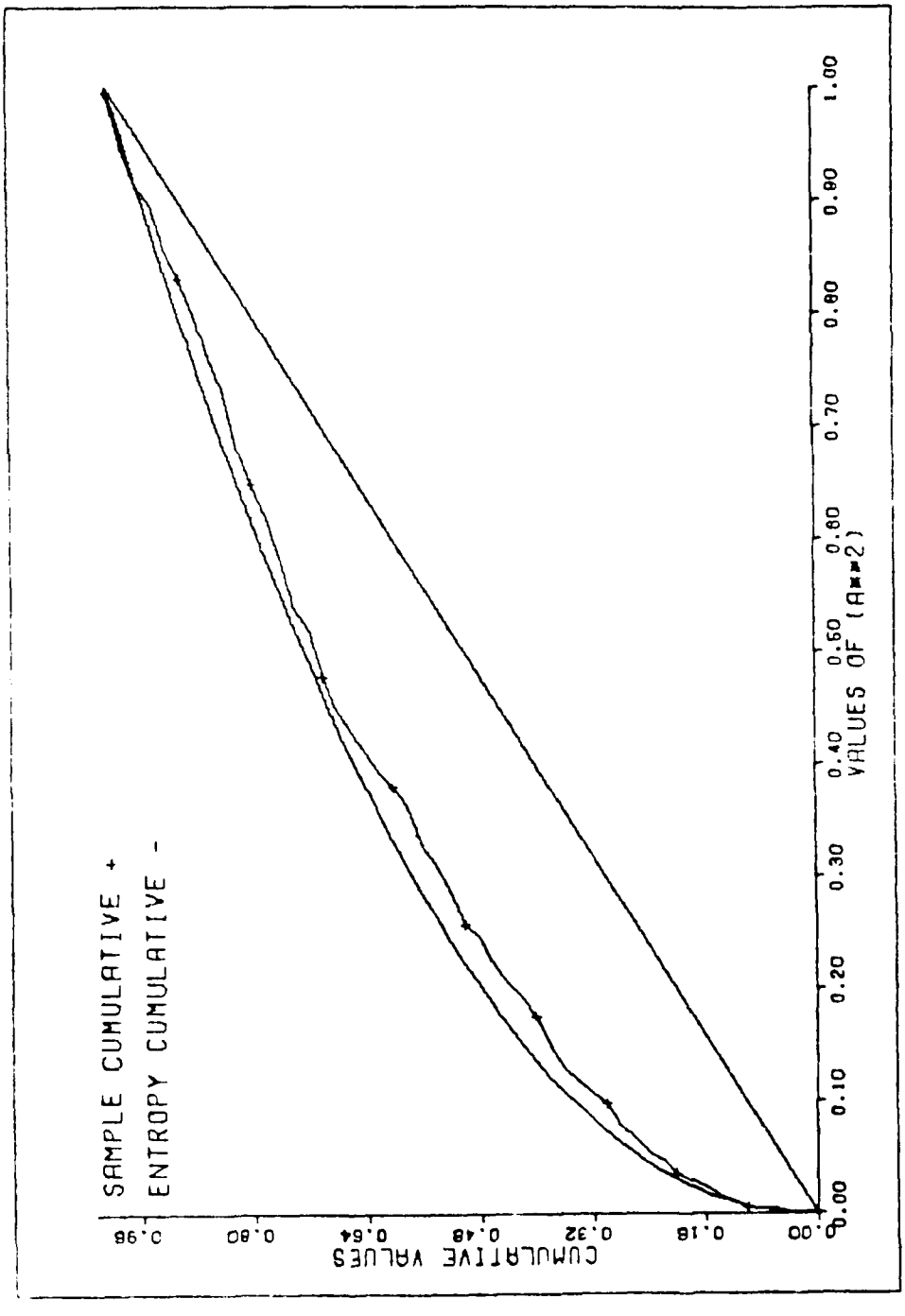


Fig. 11.4. Sample, Entropy, and Uniform Cumulatives (A**2)

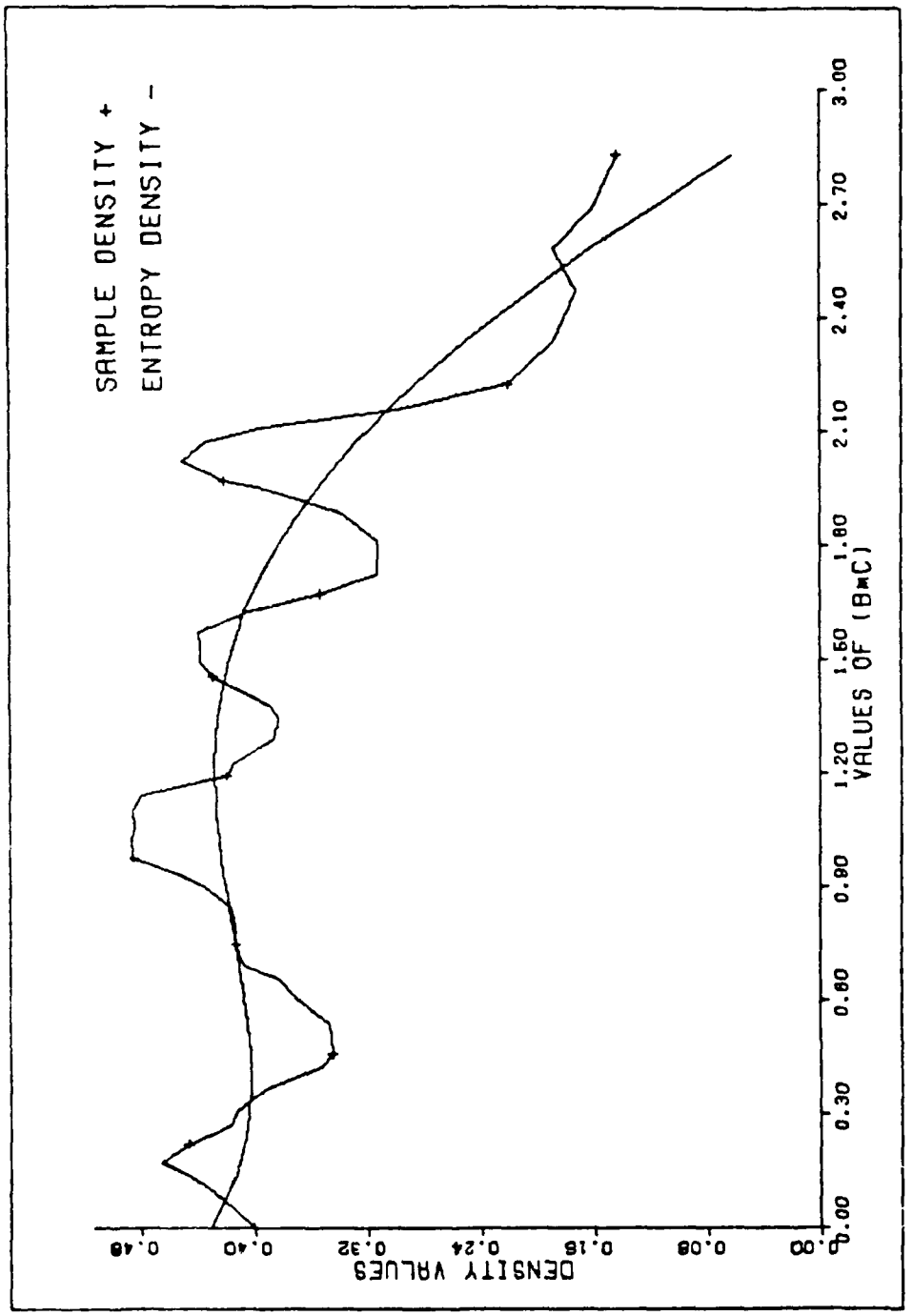


Fig. 11.5. Sample and Entropy Densities (B*C)

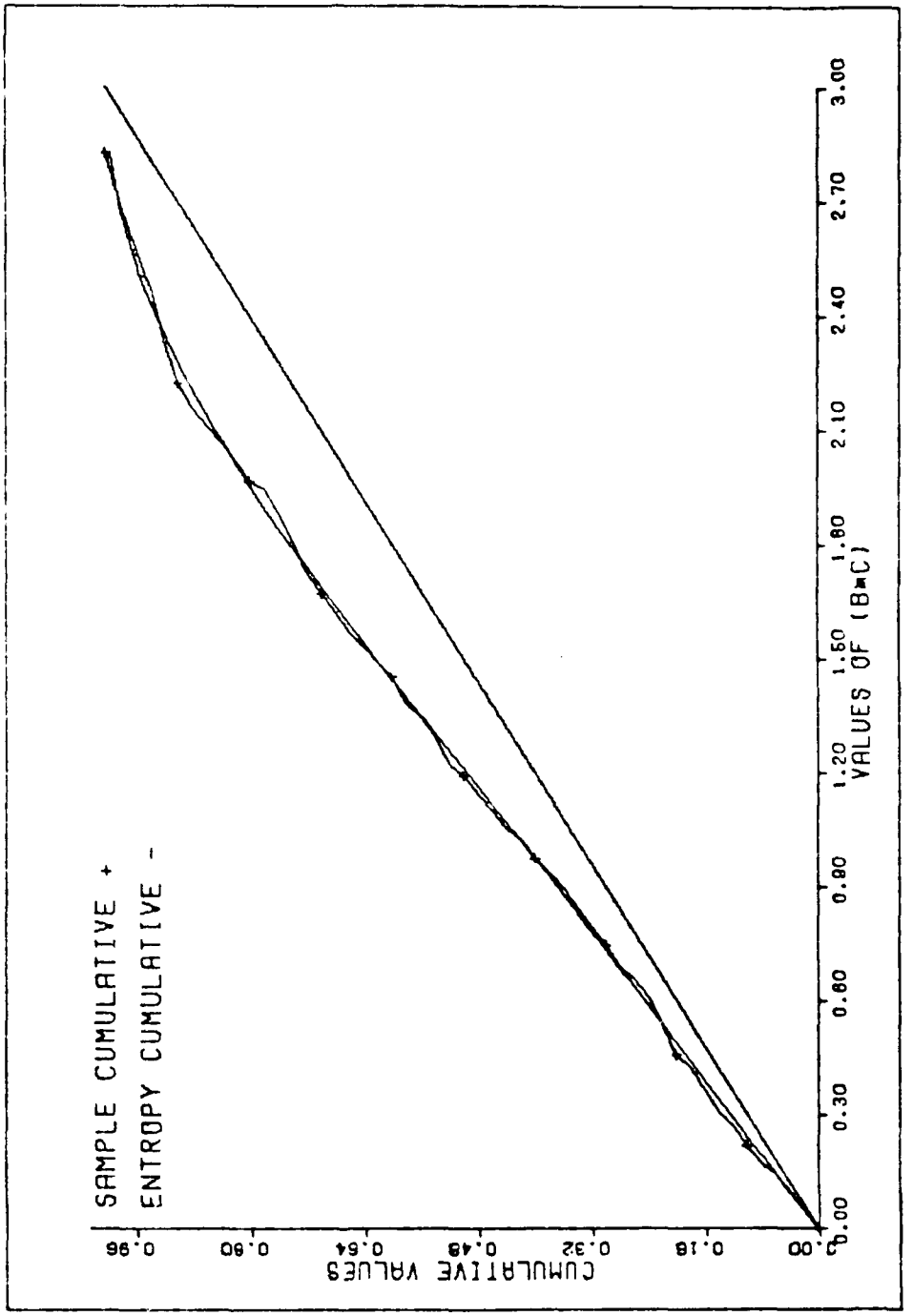


Fig. 11.6. Sample, Entropy, and Uniform Cumulatives (B*C)

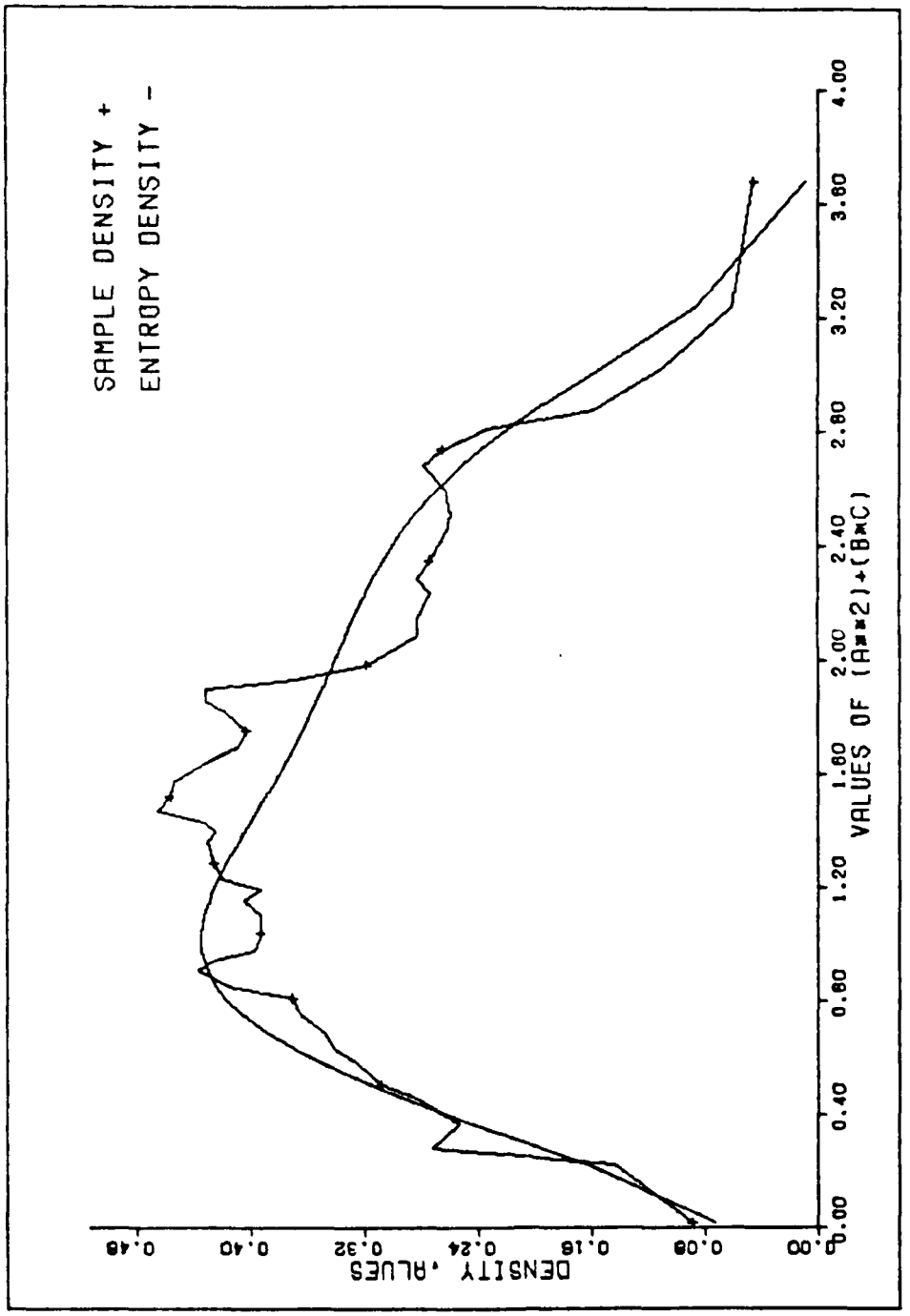


Fig. 11.7. Sample and Entropy Densities (A**2) + (B*C)

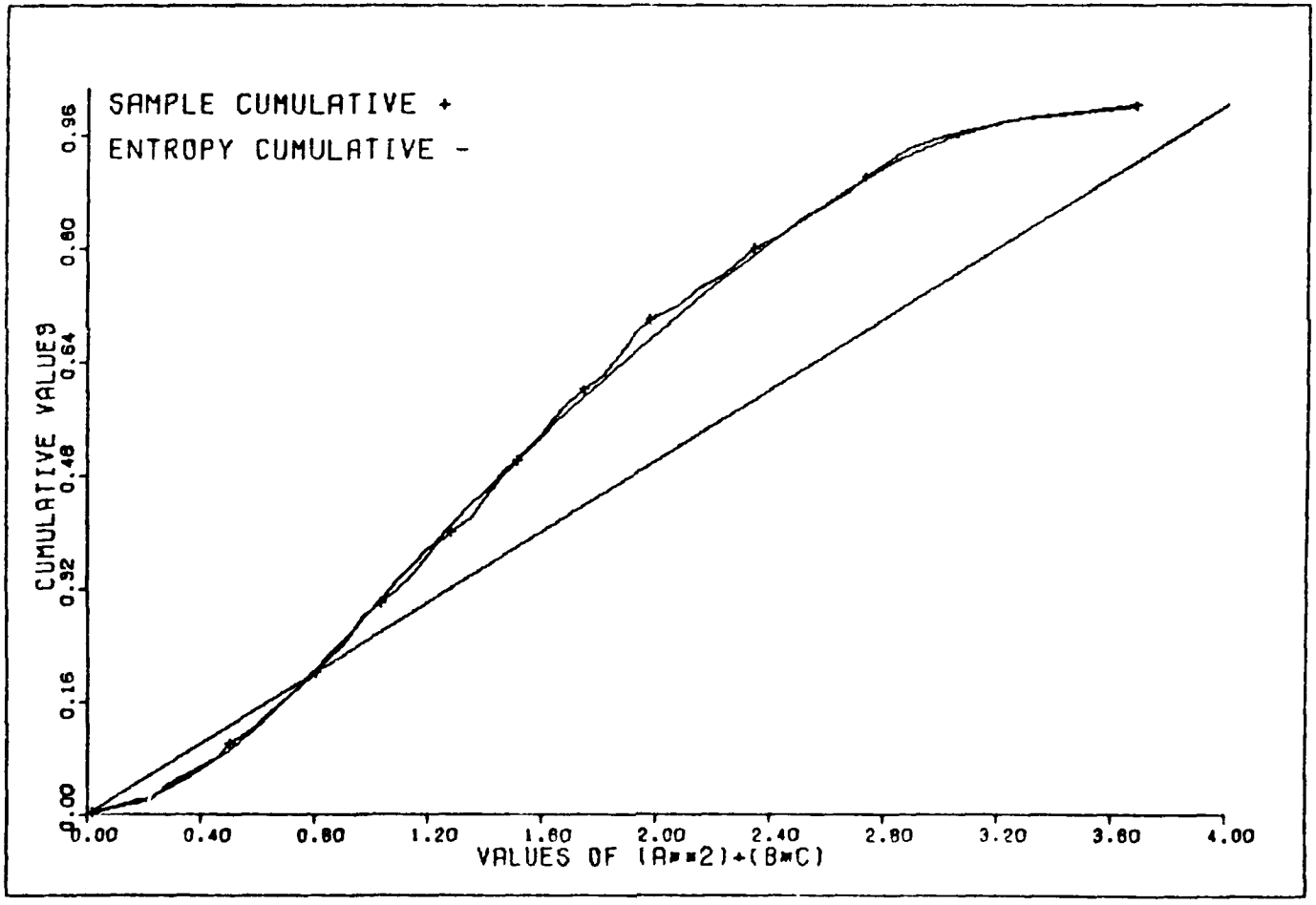


Fig. 11.8. Sample, Entropy, and Uniform Cumulatives (A**2) + (B*C)

TABLE XI.IV

RESULTS OF METHOD 2 FOR INTERVAL ARITHMETIC

Variable	Active Set	Divergence
A	F5	$J(p_5(x), f(x)) = .023823$
BC	F7,F9	$J(p_{79}(x), f(x)) = .000116$
(A^2+BC)	F3,F4,F9	$J(p_{349}(x), f(x)) = .001512$

where f(x) is the sample density and information functions are defined in Table V.III.

Summary

We have touched on three potential applications of the entropy procedure in addition to computer simulation. The procedure uses available or computable information, and the accuracy of approximation, of course, depends on the amount and accuracy of the information. Thus, the analyst must weigh information collection costs against the benefit of accurate density approximations. The possible applications for such a procedure are numerous. Our examples only represent a starting point.

Chapter XII. Summary and Future Research

Summary

We have used the concept of "maximum entropy" to develop a procedure for characterizing (or approximating) an unknown distribution based on information about that distribution. The procedure uses available information but maintains "maximum uncertainty" with respect to unspecified information and provides a "minimally prejudiced" representation of the unknown density. Our development requires information in the form of expected values of "information functions," but the procedure can be applied to other forms of information. The work is based on a constrained optimization problem and includes three procedural steps: specification of a potential set of information functions, selection of the active set for a particular approximation, and solution of the constraint equations to completely define the approximation density.

We have shown that if a solution exists to the optimization problem, then it is unique. Further, the solution density will take the following form:

$$p(x) = \exp[-\lambda_0 - \lambda_1 g_1(x) - \ldots - \lambda_k g_k(x)]$$

where the $g_i(x)$ are information functions and the λ_i are associated Lagrange multipliers. A numerical scheme for

solution of the constraints was presented. The numerical scheme may converge to a local optimum solution versus the global solution. However, under extensive testing against known distributions, the scheme has always produced the correct result.

Three separate methods were presented for selection of the active set of information functions. Methods one and two require expected value estimates and a random sample of the unknown distribution. Method three requires only expected values. Selection of a specific method is problem and data dependent. In experimentation with known analytic distributions, the methods either characterized the analytic or provided a compromise between sample and analytic distributions. Accuracy of approximation with all methods is a function of potential set specification and data accuracy.

Sensitivity aspects were addressed to include two approaches to a study of system sensitivity. Finally, several examples and example applications of the entropy approximation procedure were presented. The entire approximation procedure, to include the three information function selection methods and sensitivity studies, has been programmed for computer use.

Future Research

Our research surfaced several areas for continued investigation. These areas are highlighted in the following paragraphs.

The entropy procedure was applied to interval [a,b], assuming that the unknown density was "relatively" well behaved. In at least two examples, the bimodal distribution of Figure 8.1 and the interval arithmetic examples for BC and (A^2+BC) of Figures 11.5 through 11.8, we approximated distributions that were not entirely well behaved. For these examples we briefly investigated a piecewise application of the entropy procedure, i.e., a division of interval [a,b] into subintervals $[a,b]=[a,c_1]\ [c_1,c_2]\ldots [c_n,b]$ with application of the entropy procedure to each subinterval. This concept holds potential for more difficult distributions.

Expansion of our work to distributions on the semi-infinite and infinite intervals is feasible. Such an expansion centers on an investigation of numerical quadrature schemes and numerical procedures for solving the constraint equations. Orr (Ref 63) has accomplished some preliminary work in this area.

The research centered on characterizing univariate distributions, yet the theoretical development supports multivariate characterization. Such development may follow from the work presented in this paper.

We have discussed three methods for selection of the active set of information functions. Two of the methods involve a fit to the sample density where the sample density is produced by numerical differentiation. The methods are successful because they partially compensate for the numerical differentiation error. The development of a scheme which uses the sample cumulative, thus avoiding one level of numerical error, may prove beneficial. Such a scheme could follow the structure of method two, given a means to "efficiently" compute errors between cumulatives.

Use of the entropy procedure for hypothesis testing is a viable research area. Consider method three which produces the entropy density by forcing an approximation to the expected values of the potential information functions (Chapter VIII). As shown in Table VIII.II, when the expected values are accurate and the potential set includes the correct analytic functions, method three will accurately characterize the unknown density. For example, if the unknown density $f(x)$ is normal and the potential set includes functions x and x^2, then these functions are selected for the entropy approximation, $p(x)$, such that $p(x)=f(x)$. Such results suggest the use of this procedure to test if the unknown distribution is normal, or beta, etc. Again, the key factor in success of such an approach is accurate estimation of expected values. The suggested

research ties to recent work by Dudewicz and van der Meulen (Ref 24).

Finally, our procedure provides an effective means of approximating unknown distributions, and we have suggested several applications. Potential applications are numerous and a viable research area. Applications to risk analysis, game theory, and pattern recognition, akin to the discrete entropy applications, are examples for continued investigation.

Bibliography

1. Abramowitz, M. and I. A. Stegun (Editors). Handbook of Mathematical Functions With Formulas and Graphs and Mathematical Tables. Washington: U.S. Gov. Printing office, 1965.

2. Acton, F. S. Numerical Methods that Work. New York: Harper and Row, 1970.

3. Agmon, N., Y. Alhassid and R. D. Levine. "An Algorithm for Determining the Lagrange Parameters in the Maximal Entropy Formalism," The Maximum Entropy Formalism, edited by R. D. Levine and M. Tribus. Cambridge: MIT Press, 1979.

4. Ahiezer, N. I. and M. G. Krein. Some Questions in the Theory of Moments. R.I.: American Math. Society, 1962.

5. Ahlberg, J., E. Nilson and J. Walsh. The Theory of Splines and Their Applications. New York: Academic Press, 1967.

6. Anderson, T. W. and D. A. Darling. "A Test of Goodness of Fit," Journal of American Statistical Association, 49: 765-769 (December 1954).

7. Barnard, T. E. The Maximum Entropy Spectrum and the Burg Technique. Technical Report TR-75-01. Arlington, Virginia: Office of Naval Research, June 1975. (AD A026 626).

8. Box, G. E. P. and M. Muller. "A Note on the Generation of Random Normal Deviates," Annals of Mathematical Statistics, 29: 610-611 (June 1958).

9. Campbell, L. L. "Characterization of Entropy of Probability Distributions on the Real Line," Information and Control, 21: 329-338 (November 1972).

10. Chan, M. M. W. "System Simulation and Maximum Entropy," Operations Research, 19: 1751-1753 (November 1971).

11. Chanda, K. C. and R. W. Kulp. "On Some Nonparametric Estimators for the Linear Markov Scheme," Communications in Statistics--Theor. Meth., A7: 427-439 (1978).

12. Chen, C. H. "On Information and Distance Measures, Error Bounds, and Feature Selection," Information Sciences, 10: 159-173 (September 1976).

13. Collatz, L. Functional Analysis and Numerical Mathematics. Translated by H. Oser. New York: Academic Press, 1966.

14. Collins, R. and A. Wragg. "Maximum Entropy Histograms," Journal of Physics A Mathematical and General, 10: 1441-1464 (September 1977).

15. Crain, B. R. "Estimation of Distributions Using Orthogonal Expansions," Annals of Statistics, 2: 454-463 (May 1974).

16. -----. "More on Estimation of Distributions Using Orthogonal Expansions," Journal of American Statistical Association, 71: 741-745 (September 1976).

17. -----. "An Information Theoretic Approach to Approximating a Probability Distribution," SIAM Journal of Applied Mathematics, 32: 339-346 (March 1977).

18. Csiszár, I. "I-Divergence Geometry of Probability Distributions and Minimization Problems," Annals of Probability, 3: 146-158 (February 1975).

19. Cukier, R. I., et al. "Study of the Sensitivity of Coupled Reaction Systems to Uncertainties in Rate Coefficients. I. Theory," Journal of Chemical Physics, 59: 3873-3878 (October 1973).

20. -----. "Study of the Sensitivity of Coupled Reaction Systems to Uncertainties in Rate Coefficients. III. Analysis of the Approximations," Journal of Chemical Physics, 63: 1140-1149 (August 1975).

21. Darling, D. A. "The Kolmogorov-Smirnov, Cramér-Von Mises Tests," Annals of Mathematical Statistics, 28: 823-838 (December 1957).

22. Dhar, D. "Entropy and Phase Transitions in Partially Ordered Sets," Journal of Mathematical Physics, 19: 1711-1713 (August 1978).

23. Draper, N. R. and H. Smith, Applied Regression Analysis. New York: John Wiley and Sons, 1966.

24. Dudewicz, E. J. and E. C. van der Meulen. Entropy-Based Statistical Inference, I: Testing Hypotheses on Continuous Probability Densities, with Special Reference to Uniformity. Report No. 120. Leuven, Belgium: Department of Mathematics, Katholieke Universiteit Leuven, June 1979.

25. Dyer, A. R. "Hypothesis Testing Procedures for Separate Families of Hypothesis," Journal of American Statistical Association, 69: 140-145 (March 1974).

26. Fano, R. M. Transmission of Information. New York: Wiley, 1961.

27. Fox, R. L. Optimization Methods for Engineering Design. Massachusetts: Addison-Wesley Pub. Co., 1971.

28. Furnival, G. M. and R. W. Wilson, Jr. "Regression by Leaps and Bounds," Technometrics, 16: 499-511 (November 1974).

29. Goguen, J. A. and L. A. Carlson. "Axioms for Discrimination Information," IEEE Transactions on Information Theory, IT-21: 572-574 (September 1975).

30. Gokhale, D. V. "Approximating Discrete Distributions, with Applications," Journal of American Statistical Association, 68: 1009-1011 (December 1973).

31. -----. "Maximum Entropy Characterizations of Some Distributions," A Modern Course on Statistical Distributions in Scientific Work: Proceedings of the NATO Advanced Study Institute Held at the University of Calgary, Alberta, Canada, July 29--Aug 10, 1974. Volume 1, edited by C. P. Patil, S. Kotz, and J. K. Ord. Boston: Reidel Pub. Co., 1975.

32. Golomb, S. W. "The Information Generating Function of a Probability Distribution," IEEE Transactions on Information Theory, IT-11: 75-77 (January 1966).

33. Guiasu, S. Information Theory with Applications. New York: McGraw-Hill, Inc., 1977.

34. Haber, S. "Numerical Evaluation of Multiple Integrals," SIAM Review, 12: 481-526 (October 1970).

35. Halmos, P. R. and L. J. Savage. "Application of the Radon-Nikodym Theorem to the Theory of Sufficient Statistics," Annals of Mathematical Statistics, 20: 225-241 (June 1949).

36. Hildebrand, F. B. Introduction to Numerical Analysis (second edition). New York: McGraw-Hill, 1974.

37. Hines, W. W. and D. C. Montgomery. Probability and Statistics in Engineering and Management Science. New York: Ronald Press Co., 1972.

38. Hocking, R. R. "Criteria for Selection of a Subset Regression: Which One Should be Used?," Technometrics, 14: 967-970 (November 1972).

39. Hogg, R. V. and A. T. Craig. Introduction to Mathematical Statistics (third edition). New York: Macmillan Co., 1970.

40. Hornbeck, R. W. Numerical Methods. New York: Quantum Pub., 1975.

41. Ingarden, R. S. and A. Kossakowski. "Poisson Probability Distribution and Information Thermodynamics," Bull. Acad. Polon. Sci. Ser. Sci. Math. Astronom. Phys., 19: 83-86 (1971).

42. Jaynes, E. T. "Information Theory and Statistical Mechanics," Physical Review, 106: 620-630 (May 1957).

43. -----. "Information Theory and Statistical Mechanics. II," Physical Review, 108: 171-191 (October 1957).

44. -----. "Prior Probabilities," IEEE Transactions on Systems Science and Cybernetics, SSC-4: 227 (September 1968).

45. -----. "Where Do We Stand on Maximum Entropy?", The Maximum Entropy Formalism. Edited by R. D. Levine and M. Tribus. Cambridge Mass.: MIT Press, 1979.

46. Jeffreys, H. Theory of Probability. Oxford: Clarendon Press, 1948.

47. Johnson, N. and S. Kotz. Continuous Univariate Distributions. Two volumes. Boston: Houghton Mifflin Co., 1970.

48. Kampé de Fériet, J. Théorie de l'Information. Principe du Maximum de l'Entropie et ses Applications á la Statistique et á la Mecanique. Lille: Publications du Laboratoire de Calcul de la Faculté des Sciences de l'université de Lille, 1963.

49. Kaskey, G., et al. Statistical Techniques in Transistor Evaluation--Transformations to Normality. Technical Report No. 1, Contract Nobs-72660. Washington: Department of the Navy, Bureau of Ships, January 1961.

50. Kullback, S. and R. A. Leibler. "On Information and Sufficiency," Annals of Mathematical Statistics, 22: 79-86 (March 1951).

51. Kullback, S. Information Theory and Statistics. New York: John Wiley and Sons, Inc., 1959.

52. -----. "Approximating Discrete Probability Distributions," IEEE Transactions on Information Theory, IT-15: 444-447 (July 1969).

53. Lee, D. A., et al. Some Practical Aspects of the Treatment of Ill-posed Problems by Regularization. Technical Report No. ARL 75-0022, Wright-Patterson AFB, Ohio: Aerospace Research Laboratories, February 1975.

54. Lewis, P. A. W. "Distribution of the Anderson-Darling Statistic," Annals of Mathematical Statistics, 32: 1118-1124 (December 1961).

55. Lindgren, B. W. Statistical Theory. New York: Macmillan Co., 1962.

56. Luenberger, D. G. Optimization by Vector Space Methods. New York: Wiley, 1969.

57. -----. Introduction to Linear and Nonlinear Programming. Reading, Mass.: Addison-Wesley Pub. Co., 1973.

58. Marschak, J. "Entropy, Economics, Physics." Los Angeles: Western Management Science Institute, University of California, 1975. (ADA 001 069).

59. McDonald-Douglas Corporation. New Strategic Airlift Concepts. Vol. III, Risk Analysis. Technical Report No. AFFDL-TR-79-3062. Long Beach, California: McDonald-Douglas, June 1979.

60. Melick, H. C. Analysis of Inlet Flow Distortion and Turbulence Effects on Compressor Stability. Technical Report No. NASA-CR-114577, Moffett Field, California: Ames Research Center, NASA, March 1973.

61. Moore, R. E. and S. T. Jones. "Safe Starting Regions for Iterative Methods," SIAM Journal on Numerical Analysis, 14: 1051-1064 (December 1977).

62. -----. "Bounding Sets in Function Spaces with Applications to Nonlinear Operator Equations," SIAM Review, 20: 492-498 (July 1978).

63. Orr, G. E. Modeling Techniques for Weapon System Simulation. Unpublished report, Wright-Patterson AFB, Ohio: Department of Mathematics, Air Force Institute of Technology, 1979.

64. Parr, W. C. and W. R. Schucany. Minimum Distance and Robust Estimation. Paper presented at 41st Annual Meeting of Institute of Mathematical Statistics in San Diego, August 1978.

65. Rietz, H. L. Mathematical Statistics. Mathematical Association of America, Ill.: Open Court Pub. Co., 1927.

66. Royden, H. L. Real Analysis (second edition). New York: Macmillan Pub. Co., 1968.

67. Saaty, T. L. and J. Bram. Nonlinear Mathematics. New York: McGraw-Hill, Inc., 1964.

68. Schaibly, J. H. and K. E. Shuler. "Study of the Sensitivity of Coupled Reaction Systems to Uncertainties in Rate Coefficients. II. Applications," Journal of Chemical Physics, 59: 3879-3888 (October 1973).

69. Sen, P. K. "Estimates of the Regression Coefficient Based on Kendall's Tau," Journal of American Statistical Association, 63: 1379-1389 (December 1968).

70. Shannon, R. E. Systems Simulation. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1975.

71. Shannon, C. E. and W. Weaver. Mathematical Theory of Communications. Urbana: University of Illinois Press, 1949.

72. -----. "A Mathematical Theory of Communication," Bell System Technical Journal, 27: 379-423 (1948).

73. Sharma, B. D. and I. J. Taneja. "On Axiomatic Characterization of Information-Theoretic Measures," Journal of Statistical Physics, 10: 337-346 (1974).

74. Smith, S. A. "A Derivation of Entropy and the Maximum Entropy Criterion in Context of Decision Problems," IEEE Transactions on Systems, Man and Cybernetics, SMC-4: 152-163 (March 1974).

75. Stephens, M. A. "EDF Statistics for Goodness of Fit and Some Comparisons," Journal of American Statistical Association, 69: 730-737 (September 1974).

76. -----. "Use of Kolmogorov-Smirnov, Cramér-Von Mises and Related Statistics Without Extensive Tables," Royal Statistical Society Journal B Methodological, 32: 155-122 (May 1970).

77. Taneja, I. J. "A Joint Characterization of Directed Divergence, Inaccuracy, and Their Generalizations," Journal of Statistical Physics, 11: 169-176 (April 1974).

78. Taylor, A. E. and W. R. Mann. Advanced Calculus (second edition). Mass.: Xerox College Pub., 1972.

79. Theil, H. "A Rank-Invariant Method of Linear and Polynomial Regression Analysis," I, II, and III, Nederl. Akad. Wetensch. Proc., 53: 386-392, 521-525, 1397-1412 (1950).

80. Tomovic, R. Sensitivity Analysis of Dynamic Systems. New York: McGraw-Hill, Inc., 1963.

81. Toussaint, G. T. "Sharper Lower Bounds for Discrimination Information in Terms of Variation," IEEE Transactions on Information Theory, IT-21: 99-100 (January 1975).

82. Tribus, M. Rational Descriptions, Decisions, and Designs. New York: Pergamon Press, 1969.

83. -----. "The Widget Problem Revisited," IEEE Transactions on Systems Science and Cybernetics, SCC-4: 241-245 (1968).

84. Ulych, T. J. and T. N. Bishop. "Maximum Entropy Spectral Analysis and Autoregressive Decomposition," Reviews of Geophysics and Space Physics, 13: 183-200 (February 1975).

85. Watson, G. S. "On Chi-Square Goodness-of-Fit Tests for Continuous Distributions," Royal Statistical Society Journal, B20: 44-61 (January 1958).

86. Weidemann, H. L. and E. B. Stear. "Entropy Analysis of Estimating Systems," IEEE Transactions on Information Theory, IT-16: 264-270 (May 1970).

87. White, D. J. "Entropy and Decisions," Operations Research Quarterly, 26: 15-23 (March 1975).

88. Widder, D. V. The Laplace Transform. Princeton: Princeton University Press, 1941.

89. Wilson, G. A. and A. Wragg. "Numerical Methods for Approximating Continuous Probability Density Functions, Over [0,∞), Using Moments," Institute of Mathematics and Its Applications Journal, 12: 165-173 (October 1973).

90. Wragg, A. and D. C. Dowson. "Fitting Continuous Density Functions Over [0,∞) Using Information Theory Ideas," IEEE Transactions on Information Theory, IT-16: 226-230 (March 1970).

91. Young, T. Y. and G. Coraluppi. "Stochastic Estimation of a Mixture of Normal Density Functions Using an Information Criterion," IEEE Transactions on Information Theory, IT-16: 258-263 (May 1970).

Appendix A. Numerical Quadrature

> The primary purpose of numerical integration (also called quadrature) is evaluation of integrals which are either impossible or else very difficult to evaluate analytically [Ref 40:144].

Quadrature also offers an effective means of machine integration, and a variety of numerical integration methods are available (Ref 1). One such method which is particularly adaptable to machine computation is Gauss quadrature. We consider the general quadrature approach and then Gauss quadrature specifically.

Given the function $f(x)$ and the values of $f(x)$ at N points, x_i, $i=1,2,\ldots N$, we wish to calculate the integral $\int_a^b f(x)\,dx$. The general quadrature rule to approximate this integral follows:

$$\int_a^b f(x)\,dx \approx \sum_{i=1}^{N} W_i f(x_i) \qquad (A.1)$$

where the weights, W_i, are determined by requiring that equation A.1 be exactly true when $f(x)$ is replaced by $1, x, x^2, \ldots x^{N-1}$. Thus we have N equations of the form of equation A.1 with the N unknown W_i, $i=1,2,\ldots N$. Selecting the weights by solving these N equations will guarantee that the guadrature rule will exactly integrate any polynomial of degree N-1 or less.

Gauss quadrature improves the accuracy of the integral by using orthogonal polynomials and selecting the points x_i, i=1,2,...N, to be zeroes of the orthogonal polynomials. Gauss quadrature thus assumes that one can obtain the values of f(x) at the unevenly spaced quadrature points x_i, i=1,2,...N. Abramowitz and Stegun (Ref 1) and Hornbeck (Ref 40) provide detailed explanation and examples of Gauss forms for various sets of orthogonal polynomials. Abramowitz and Stegun provide tables for weights and quadrature points. From this reference we find the following formula for Legendre polynomials:

$$\int_{-1}^{1} f(x)\,dx = \sum_{i=1}^{N} W_i f(x_i) + R_N$$

or

$$\int_{a}^{b} f(x)\,dx = \frac{b-a}{2} \sum_{i=1}^{N} W_i f(y_i) + R_N$$

where

$$y_i = \left(\frac{b-a}{2}\right) x_i + \left(\frac{b+a}{2}\right)$$

and

$$R_N = \frac{(b-a)^{2N+1}(N!)^4}{(2N+1)((2-N)!^3} 2^{2N+1} f^{2N}(\xi)$$

The Gauss-Legendre formula will produce exact results for a polynomial of degree 2N-1 or less. Thus if f(x) is closely approximated by a polynomial of degree 2N-1, then the error of approximation, R_N, will be small. The integrals that

are involved in the characterization method of this paper generally concern continuous, well-behaved functions. Quadrature is thus an effective and accurate tool for our application. Hornbeck (Ref 40) discusses practical methods of testing quadrature accuracy and potential quadrature pitfalls.

The quadrature formulae discussed above are easily extended to multiple integrals:

$$\int_{a_1}^{b_1}\int_{a_2}^{b_2} f(x,y)\,dx\,dy \approx \left(\frac{b_2-a_2}{2}\right)\int_{a_1}^{b_1}\sum_{i=1}^{N} W_i\, f(x,y_i)\,dx$$

$$\approx \left(\frac{b_1-a_1}{2}\right)\left(\frac{b_2-a_2}{2}\right)\sum_{j=1}^{N} W_j \sum_{i=1}^{N} W_i\, f(x_j,y_i)$$

The accuracy of approximation is now reduced and more function evaluations are necessary. Specifically, we required N function evaluations for one-dimensional quadrature. Two-dimensional quadrature requires N^2 function evaluations or quadrature points. In a similar fashion, K-dimensional quadrature requires N^K quadrature points or functional evaluations. The decreased accuracy and increased number of functional evaluations are discussed in detail by Haber (Ref 34). Haber suggests that multidimensional quadrature is effective up to dimension three or four. He advises the use of other methods, such as Monte Carlo quadrature, for

integrals of dimension five or greater. Multidimensional quadrature is of particular use in application of the entropy characterization procedure to computer simulation; see Chapter IX.

Appendix B. Goodness of Fit Statistics

Statistical Tests

Selection of the active set of information functions in the regression method, Chapter VI, involves the use of "goodness of fit" statistics. That is, a random sample of the unknown distribution is available, and we wish to test the hypothesis that the given sample is from one of the entropy distributions. We choose, as active, the entropy distribution that provides the highest level of confidence in the truth of our hypothesis, i.e., the distribution with the smallest value for the selected statistic. This appendix discusses a few popular statistics for hypothesis testing to include Chi-squared (χ^2) and Empirical Distribution Function (EDF) statistics.

Several references (Refs 6; 21; 25; 54; 55; 64) define goodness of fit statistics which are appropriate for testing the hypothesis, H_0, that a given sample is from a specified distribution. M. A. Stephens (Refs 75; 76) presents an excellent summary of the more well known statistics to include advantages and disadvantages. As Stephens mentions, the classical test for goodness of fit is the χ^2 test.

The χ^2 test can be applied in our case, given that the entropy densities are completely defined prior to the

test and defined without recourse to the sample data, i.e., solution for the Λ vector in each p(x) may not depend on the test sample x(I), I=1,...N. This restriction is necessary because the χ^2 test, when estimated parameters are involved, requires a maximum likelihood parameter estimation to insure a χ^2 distribution for the test statistic (Ref 85). The requirement for complete specification of the entropy densities prior to goodness of fit testing will also apply to the EDF statistics. Solution for the entropy density parameters, the Λ vector, depends on the expected value vector <G> (Chapters IV and VI). We produce <G> by numerical quadrature in our tests of method one and do not rely on the random sample. Thus the entropy densities are specified before goodness of fit testing and without using the sample.

The analyst should notice that the above restriction on parameter estimation does not preclude the use of average value estimates of <G>, i.e., $\langle g_j(x) \rangle = \sum_{i=1}^{N} g_j(x_i)/N$. Method one may still be applied with average values and will, as demonstrated in Chapter VI, provide an excellent approximation to the unknown density. However, if the sample is used to generate <G> and thus to find p(x), then the analyst can not place the usual statistical significance on the values of calculated statistics. The statistics will still offer a measure for selecting the best entropy distribution, but we may not use the statistic in conjunction with

existing statistical tables to state a confidence in our test of hypothesis H_0. A true statistical test of hypothesis, under these circumstances, will require a second independent sample of the unknown distribution. The references provide more detail on this restriction and effective use of statistics.

Stephens (Ref 75) states that when the hypothesized distribution (the entropy distribution) is completely specified and continuous then, ". . . in general, EDF statistics give more powerful tests of H_0 than χ^2." The EDF statistics of interest are summarized below. Statistic selection is at the user's discretion. Method one can be used with χ^2 and any of the EDF statistics, or other suitable statistical tests. The user must determine which aspect of the approximation is of greatest importance to him. For example, the Anderson-Darling statistic emphasizes a fit to the tails of the unknown distribution while the Kolmogorov-Smirnov statistic measures maximum error in the approximation. Several statistics and their benefits are now considered.

EDF Statistics

Let $EN(I) = \int_a^{x(I)} p(x)\,dx$ where $p(x)$ is the entropy density for a given information function set, and $x(I)$, $I=1,\ldots N$ is the sorted random sample from the unknown

distribution. The sample cumulative at each of the I points is CUM(I) = I/N.

Kolmogorov Statistics.

$$D^+ = \max_I [(I/N) - EN(I)];$$

$$D^- = \max_I [EN(I) - (I-1)/N]; \text{ and}$$

$$D = \max [D^+, D^-].$$

The statistic of interest is D (usually called Kolmogorov-Smirnov statistic) which tests the maximum deviation of the sample cumulative from the entropy cumulative.

Cramér-von Mises Statistic, W^2.

$W_n^2 = n \int_a^b [F_n(x)-F(x)]^2 \, G[F(x)] \, dF(x)$, where F(x) is the hypothesis distribution, $F_n(x)$ is the sample, and G[F(x)] is a weight function. We use the Smirnov weight function, G=1, and integrate to obtain a computational form of the statistic:

$$W^2 = \sum_{I=1}^{N} [EN(I) - (2I-1)/2N]^2 + [1/12N]$$

This form of the Cramér-von Mises statistic is akin to the sum of errors squared and weights each data point evenly.

Anderson-Darling Statistic, A^2. If we use the Cramér-von Mises W_n^2 statistic and define G[F(x)] to be $1/\{F(x)[1-F(x)]\}$, then the result is the Anderson-Darling statistic. References 6, 54 and 75 reduce A^2 to a computational form:

$$A^2 = -\{\sum_{I=1}^{N} (2I-1)[\ln(EN(I)) + \ln(1-EN(N+1-I))]\}/N - N$$

This statistic emphasizes a fit to the tails of the distribution.

Kuiper Statistic, V.

$$V = D^+ + D^-$$

Watson Statistic, U^2.

$$U^2 = W^2 - N(<EN> - 1/2)^2$$

where $<EN> = \sum_{I=1}^{N} EN(I)/N$. U^2 adjusts for the hypothesized mean. Stephens states that both V and U^2 are useful in identifying a change in scale (variance) of the sample while D, W^2, and A^2 are more effective for a change in location (mean). The references discuss the above statistics and other variations of the above. Stephens presents a comparison that, for his purposes, favors the A^2, W^2 and U^2 statistics.

Statistics for Method One

To select the "best" information function set from the candidate regression sets of Chapter VI, we prefer the A^2 or W^2 EDF statistics. Again, the "best" set will be the set of functions that results in the smallest value of A^2 (or W^2). EDF statistics are preferred to χ^2 primarily because the EDF statistics are distribution-free. Additionally, the χ^2 test requires an "unbiased" grouping of the data which detracts from a generalized approach. Finally, our entropy functions satisfy the "continuity" and "completely defined" requirements of the EDF tests. Under these conditions, Stephens (Ref 75) states that the EDF tests should prove more powerful than χ^2.

Vita

James E. Miller, Jr. was born on 28 December 1946 in Johnson City, New York. He graduated from high school in St. Petersburg, Florida in 1965 and attended the U.S. Air Force Academy from which he received the degree of Bachelor of Science in Engineering Sciences in June 1969. Commissioned a regular officer in 1969, he received six months of formal technical instruction for subsequent duty as a Signals Intelligence Officer at intelligence units in Japan, Thailand, and Vietnam. He then attended the Georgia Institute of Technology from which he received the degree of Master of Science in Information and Computer Science in August 1974. This education was applied to duties at the Defense Intelligence Agency as a Staff Officer, Systems Analyst, and Project Manager for development of a large-scale computer support system. He entered the doctoral program in the School of Engineering, Air Force Institute of Technology, in September 1977.

Permanent Address: 3216 S. Ferncreek Ave.
Orlando, Florida 32806

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: AFIT/DS/MA/80-1	2. GOVT ACCESSION NO.: AD-A083 514
3. RECIPIENT'S CATALOG NUMBER:	
4. TITLE (and Subtitle): CONTINUOUS DENSITY APPROXIMATION ON A BOUNDED INTERVAL USING INFORMATION THEORETIC CONCEPTS	5. TYPE OF REPORT & PERIOD COVERED: Ph.D. Dissertation
	6. PERFORMING ORG. REPORT NUMBER:
7. AUTHOR(s): James E. Miller, Jr.	8. CONTRACT OR GRANT NUMBER(s):
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Air Force Institute of Technology (AFIT-EN) Wright-Patterson AFB, Ohio 45433	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: Project 2404-01-55
11. CONTROLLING OFFICE NAME AND ADDRESS: AFFDL/FIMB Air Force Flight Dynamics Laboratory Vehicle Synthesis Branch Wright-Patterson AFB, Ohio 45433	12. REPORT DATE: March 1980
	13. NUMBER OF PAGES: 250
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office):	15. SECURITY CLASS. (of this report): Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

Approved for public release; IAW AFR 190-17

Joseph P. Hipps, Major, USAF
Director of Information

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Probability density estimation
Maximum entropy
Information theory
Bayesian statistics

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This report presents the theoretical development and numerical implementation of a procedure for approximating continuous probability density functions on a bounded interval. The work is applicable to Bayesian decision models in that available information is used to update or obtain the prior distribution. The procedure is based on the solution of a constrained entropy maximization problem and requires information in the form of expected values of

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

information functions. The approach involves three steps: estimation of expected (or average) values of "potential" information functions, selection of the "active" subset of functions to define the approximation family, and simultaneous solution of the constraints to select the specific approximating density for a given set of data.

A useful set of potential information functions is developed, and three numerical methods for active set selection are demonstrated. Numerical techniques for expected value computation are discussed, and a scheme for solution of the constraints is developed and implemented. Theoretical development includes theorems on form and uniqueness. Approximation accuracy is related to potential set definition and data accuracy. The procedure is applied to several known distributions to demonstrate applicability. Applications to computer simulation and interval arithmetic models are demonstrated with specific examples.