Foundations and Trends® in
Econometrics
2:1–2 (2006)

Information and Entropy Econometrics — A Review and Synthesis

Amos Golan

now
the essence of knowledge

Information and Entropy Econometrics — A Review and Synthesis

Information and Entropy Econometrics — A Review and Synthesis

Amos Golan

Department of Economics
American University
4400 Massachusetts Avenue
NW Washington
DC 20016-8029
USA
agolan@american.edu

the essence of knowledge

Boston – Delft

Foundations and Trends® in Econometrics

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
USA
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is A. Golan, Information and Entropy Econometrics — A Review and Synthesis, Foundations and Trends® in Econometrics, vol 2, no 1–2, pp 1–145, 2006

ISBN: 978-1-60198-104-2

Foundations and Trends® in Econometrics

Volume 2 Issue 1–2, 2006

Editorial Board

Editorial Scope

Foundations and Trends® in Econometrics will publish survey and tutorial articles in the following topics:

- Identification
- Model Choice and Specification Analysis
- Non-linear Regression Models
- Simultaneous Equation Models
- Estimation Frameworks
- Biased Estimation
- Computational Problems
- Microeconometrics
- Treatment Modeling
- Discrete Choice Modeling
- Models for Count Data
- Duration Models
- Limited Dependent Variables
- Panel Data
- Dynamic Specification
- Inference and Causality
- Continuous Time Stochastic Models
- Modeling Non-linear Time Series
- Unit Roots
- Cointegration
- Latent Variable Models
- Qualitative Response Models
- Hypothesis Testing
- Interactions-based Models
- Duration Models
- Financial Econometrics
- Measurement Error in Survey Data
- Productivity Measurement and Analysis
- Semiparametric and Nonparametric Estimation
- Bootstrap Methods
- Nonstationary Time Series
- Robust Estimation

Information for Librarians
Foundations and Trends® in Econometrics, 2006, Volume 2, 2 issues. ISSN paper version 1551-3076. ISSN online version 1551-3084. Also available as a combined paper and online subscription.

Foundations and Trends® in
Econometrics
Vol. 2, Nos. 1–2 (2006) 1–145

DOI: 10.1561/0800000004

now
the essence of knowledge

Information and Entropy Econometrics — A Review and Synthesis*

Amos Golan

Department of Economics, American University, 4400 Massachusetts Avenue, NW Washington, DC 20016-8029, USA, agolan@american.edu

This work is dedicated to Marge and George Judge

Abstract

The overall objectives of this review and synthesis are to study the basics of information-theoretic methods in econometrics, to examine the connecting theme among these methods, and to provide a more detailed summary and synthesis of the sub-class of methods that treat the observed sample moments as stochastic. Within the above objectives, this review focuses on studying the inter-connection between information theory, estimation, and inference. To achieve these objectives, it provides a detailed survey of information-theoretic concepts and quantities used within econometrics. It also illustrates

* Part of this study was done during the senior fellowship at the Institute of Advanced Study and the faculty of statistics of the University of Bologna. The author thanks the Institute and the University for their Support. Detailed Comments from Douglas Miller and Mirko Degli Esposti and an anonymous reviewer on an earlier version of this manuscript, and comments from Essie Maasoumi on earlier sections of this survey, are greatly appreciated. The author also thanks Doug Miller for pointing some unknown references. Finally, the author thanks the Editor — Bill Greene — for all his helpful comments and suggestions throughout the process of composing this review.

the use of these concepts and quantities within the subfield of information and entropy econometrics while paying special attention to the interpretation of these quantities. The relationships between information-theoretic estimators and traditional estimators are discussed throughout the survey. This synthesis shows that in many cases information-theoretic concepts can be incorporated within the traditional likelihood approach and provide additional insights into the data processing and the resulting inference.

Keywords: Empirical likelihood; entropy, generalized entropy; information; information theoretic estimation methods; likelihood; maximum entropy; stochastic moments.

JEL codes: C13, C14, C49, C51

Preface

This review and synthesis is concerned with information and entropy econometrics (IEE). The overall objective is to summarize the basics of information-theoretic methods in econometrics and the connecting theme among these methods. The sub-class of methods that treat the observed sample moments as stochastic is discussed in greater detail.
Within the above objective, we restrict our attention to study the interconnection between information theory, estimation, and inference. We provide a detailed survey of information-theoretic concepts and quantities used within econometrics and then show how these quantities are used within IEE. We pay special attention for the interpretation of these quantities and for describing the relationships between information-theoretic estimators and traditional estimators.
In Section 1, an introductory statement and detailed objectives are provided. Section 2 provides a historical background of IEE. Section 3 surveys some of the basic quantities and concepts of information theory. This survey is restricted to those concepts that are employed within econometrics and that are used within that survey. As many of these concepts may not be familiar to many econometricians and economists, a large number of examples are provided. The concepts discussed

include entropy, divergence measures, generalized entropy (known also as Cressie Read function), errors and entropy, asymptotic theory, and stochastic processes. However, it is emphasized that this is not a survey of information theory. A less formal discussion providing interpretation of information, uncertainty, entropy and ignorance, as viewed by scientists across disciplines, is provided at the beginning of that section. In Section 4, we discuss the classical maximum entropy (ME) principle (both the primal constrained model and the dual concentrated and unconstrained model) that is used for estimating under-determined, zero-moment problems. The basic quantities discussed in Section 3, are revisited again in connection with the ME principle. In Section 5, we discuss the motivation for information-theoretic (IT) estimators and then formulate the generic IT estimator as a constrained optimization problem. This generic estimator encompasses all the estimators within the class of IT estimators. The rest of this section describes the basics of specific members of the IT class of estimators. These members compose the sub-class of methods that incorporate the moment restrictions within the generic IT-estimator as (pure) zero moments' conditions, and include the empirical likelihood, the generalized empirical likelihood, the generalized method of moments and the Bayesian method of moments. The connection between each one of these methods, the basics of information theory and the maximum entropy principle is discussed. In Section 6, we provide a thorough discussion of the other sub-class of IT estimators: the one that views the sample moments as stochastic. This sub-class is also known as the generalized maximum entropy. The relevant statistics and information measures are summarized and connected to quantities studied earlier in the survey. We conclude with a simple simulated example. In Section 7, we provide a synthesis of likelihood, ME and other IT estimators, via an example. We study the interconnections among these estimators and show that though coming from different philosophies they are deeply rooted in each other, and understanding that interconnection allows us to understand our data better. In Section 8, we summarize related topics within IEE that are not discussed in this survey.

Readers of this survey need basic knowledge of econometrics, but do not need prior knowledge of information theory. Those who are familiar

with the concepts of IT should skip Section 3, except Section 3.4 which is necessary for the next few sections. Those who are familiar with the ME principle can skip parts of Section 4, but may want to read the example in Section 4.7. The survey is self contained and interested readers can replicate all results and examples provided. No detailed proofs are provided, though the logic behind some less familiar arguments is provided. Whenever necessary the readers are referred to the relevant literature.

This survey may benefit researchers who wish to have a fast introduction to the basics of IEE and to acquire the basic tools necessary for using and understanding these methods. The survey will also benefit applied researchers who wish to learn improved new methods, and applications, for extracting information from noisy and limited data and for learning from these data.

Contents

1 Introductory Statement, Motivation, and Objective **1**

2 Historical Perspective **7**

3 Information and Entropy — Background, Definitions, and Examples **15**

3.1 Information and Entropy — The Basics 15
3.2 Discussion 26
3.3 Multiple Random Variables, Dependency, and Joint Entropy 28
3.4 Generalized Entropy 32
3.5 Axioms 34
3.6 Errors and Entropy 36
3.7 Asymptotics 39
3.8 Stochastic Process and Entropy Rate 45
3.9 Continuous Random Variables 46

4 The Classical Maximum Entropy Principle **49**

4.1 The Basic Problem and Solution 49
4.2 Duality — The Concentrated Model 51

4.3 The Entropy Concentration Theorem 55
4.4 Information Processing Rules and Efficiency 57
4.5 Entropy — Variance Relationship 58
4.6 Discussion 59
4.7 Example 61

5 Information-Theoretic Methods of Estimation — I 63

5.1 Background 63
5.2 The Generic IT Model 64
5.3 Empirical Likelihood 67
5.4 Generalized Method of Moments — A Brief Discussion 72
5.5 Bayesian Method of Moments — A Brief Discussion 81
5.6 Discussion 83

6 Information-Theoretic Methods of Estimation — II 85

6.1 Generalized Maximum Entropy — Basics 85
6.2 GME — Extensions: Adding Constraints 93
6.3 GME — Entropy Concentration Theorem and Large Deviations 94
6.4 GME — Inference and Diagnostics 96
6.5 GME — Further Interpretations and Motivation 99
6.6 Numerical Example 103
6.7 Summary 105

7 IT, Likelihood and Inference 107

7.1 The Basic Problem — Background 107
7.2 The IT-ME Solution 109
7.3 The Maximum Likelihood Solution 111
7.4 The Generalized Case — Stochastic Moments 112
7.5 Inference and Diagnostics 114

7.6 IT and Likelihood 117
7.7 Conditional, Time Series Markov Process 118

8 Concluding Remarks and Related Work Not Surveyed 131

References 137

1

Introductory Statement, Motivation, and Objective

All learning, information gathering and information processing, is based on limited knowledge, both *a priori* and data, from which a larger "truth" must be inferred. To learn about the true state of the world that generated the observed data, we use statistical models that represent these outcomes as functions of unobserved structural parameters, parameters of priors and other sampling distributions, as well as complete probability distributions. Since we will never know the true state of the world, we generally focus, in statistical sciences, on recovering information about the complete probability distribution, which represents the ultimate truth in our model. Therefore, all estimation and inference problems are translations of limited information about the probability density function (pdf) toward a greater knowledge of that pdf. However, if we knew all the details of the true mechanism then we would not need to resort to the use of probability distributions to capture the perceived uncertainty in outcomes that results from our ignorance of the true underlying mechanism that controls the event of interest.

Information theory quantities, concepts, and methods provide a unified set of tools for organizing this learning process. They provide a

discipline that at once exposes more clearly what the existing methods do, and how we might better accomplish the main goal of scientific learning. This review first studies the basic quantities of information theory and their relationships to data analysis and information processing, and then uses these quantities to summarize (and understand the connection among) the improved methods of estimation and data processing that compose the class of entropy and information-theoretic methods. Within that class, the review concentrates on methods that use conditional and unconditional stochastic moments.

It seems natural to start by asking what is information, and what is the relationship between information and econometric, or statistical analysis. Consider, for example, Shakespeare's "Hamlet," Dostoevsky's "The Brothers Karamazov," your favorite poem, or the US Constitution. Now think of some economic data describing the demand for education, or survey data arising from pre-election polls. Now consider a certain speech pattern or communication among individuals. Now imagine you are looking at a photo or an image. The image can be sharp or blurry. The survey data may be easy to understand or extremely noisy. The US Constitution is still being studied and analyzed daily with many interpretations for the same text, and your favorite poem, as short as it may be, may speak a whole world to you, while disliked by others.

Each of these examples can be characterized by the amount of information it contains or by the way this information is conveyed or understood by the observer — the analyst, the reader. But what is information? What is the relationship between information and econometric analysis? How can we efficiently extract information from noisy and evolving complex observed economic data? How can we guarantee that only the relevant information is extracted? How can we assess that information? The study of these questions is the subject of this survey and synthesis.

This survey discusses the concept of *information* as it relates to econometric and statistical analyses of data. The meaning of "information" will be studied and related to the basics of Information Theory (IT) as is viewed by economists and researchers who are engaged in deciphering information from the (often complex and evolving) data,

while taking into account what they know about the underlying process that generated these data, their beliefs about the (economic) system under investigation, and nothing else. In other words, the researcher wishes to extract the available information from the data, but wants to do it with minimal *a priori* assumptions. For example, consider the following problem taken from Jaynes's famous Brandeis lectures (1963). We know the empirical mean value (first moment) of, say one million tosses of a six-sided die. With that information the researcher wishes to predict the probability that in the next throw of the die we will observe the value 1, 2, 3, 4, 5 or 6. The researcher also knows that the probability is proper (sum of the probabilities is one). Thus, in that case, there are six values to predict (six unknown values) and two observed (known) values: the mean and the sum of the probabilities. As such, there are more unknown quantities than known quantities, meaning there are infinitely many probability distributions that sum up to one and satisfy the observed mean. In somewhat more general terms, consider the problem of estimating an unknown discrete probability distribution from a finite and possibly noisy set of observed (sample) moments. These moments (and the fact that the distribution is proper — summing up to one) are the only available information. Regardless of the level of noise in these observed moments, if the dimension of the unknown distribution is larger than the number of observed moments, there are infinitely many proper probability distributions satisfying this information (the moments). Such a problem is called an under-determined problem. Which one of the infinitely many solutions should one use? In all the IEE methods, the one solution chosen is based on an information criterion that is related to Shannon's information measure — entropy.

When analyzing a linear regression, a jointly determined system of equations, a first-order Markov model, a speech pattern, a blurry image, or even a certain text, if the researcher wants to understand the data but without imposing a certain structure that may be inconsistent with the (unknown) truth, the problem may become inherently under-determined. The criterion used to select the desired solution is an information criterion which connects statistical estimation and inference with the foundations of IT. This connection provides us with an

IT perspective of econometric analyses and reveals the deep connection among these "seemingly distinct" disciplines. This connection gives us the additional tools for a better understanding of our limited data, and for linking our theories with real observed data. In fact, information theory and data analyses are the major thread connecting most of the scientific studies trying to understand the true state of the world with the available, yet limited and often noisy, information.

Within the econometrics and statistical literature the family of IT estimators composes the heart of IEE. It includes the Empirical (and Generalized Empirical) Likelihood, the Generalized Method of Moments, the Bayesian Method of Moments and the Generalized Maximum Entropy among others. In all of these cases the objective is to extract the available information from the data with minimal assumptions on the data generating process and on the likelihood structure. The logic for using minimal assumptions in the IEE class of estimators is that the commonly observed data sets in the social sciences are often small, the data may be non-experimental noisy data, the data may be arising from a badly designed experiment, and the need to work with nonlinear (macro) economic models where the maximum likelihood estimator is unattractive as it is not robust to the underlying (unknown) distribution. Therefore, (i) such data may be ill-behaved leading to an ill-posed and/or ill-conditioned (not full rank) problem, or (ii) the underlying economic model does not specify the complete distribution of the data, but the economic model allows us to place restrictions on this (unknown) distribution in the form of population moment conditions that provide information on the parameters of the model. For these estimation problems and/or small and non-experimental data it seems logical to estimate the unknown parameters with minimum *a priori* assumptions on the data generation process, or with minimal assumptions on the likelihood function. Without a pre-specified likelihood, other non maximum likelihood methods must be used in order to extract the available information from the data. Many of these methods are members of the class of Information-Theoretic (IT) methods.

This survey concentrates on the relationship between econometric analyses, data and information with an emphasis on the philosophy leading to these methods. Though, a detailed exposition is provided

here, the focus of this survey is on the sub-class of IT estimators that view the observed moments as stochastic. Therefore, the detailed formulations and properties of the other sub-class of estimators that view the observed moments as (pure) zero-moment conditions will be discussed here briefly as it falls outside the scope of that review and because there are numerous recent reviews and texts of these methods (e.g., Smith, 2000, 2005, 2007; Owen, 2001; Hall, 2005; Kitamura, 2006). However, the connection to IT and the ME principle, and the inter-relationships among the estimators, is discussed here as well.

2

Historical Perspective

This section provides a look back at the history of statistical and econometrics thoughts that led to the current state of Information and Entropy Econometrics. Though IEE builds directly on the foundations of Information Theory and the Maximum Entropy Formalism, it is also greatly affected by developments within statistics and econometrics. For a nice historical perspective and synthesis of IEE during the last century with an emphasis on the more traditional methods see Bera and Bilias (2002). Figures 2.1 and 2.2 present a long-term and a short-term (brief) history of IEE.

Generally speaking, facing a sample of data, the researcher's objective is to extract the available information from the data and then use these estimated values to decide if the data supports her/his original belief or hypothesis. In a more common language, in analyzing data one is trying to fit a certain distribution or a certain (linear or nonlinear) function to the data. Translating that objective to practical (optimization) models means that the estimation problem boils down to minimizing a certain distance measure between two distributions.

This philosophy and approach, within statistics and econometrics, goes a long way back, but it is mostly emphasized in the early work of

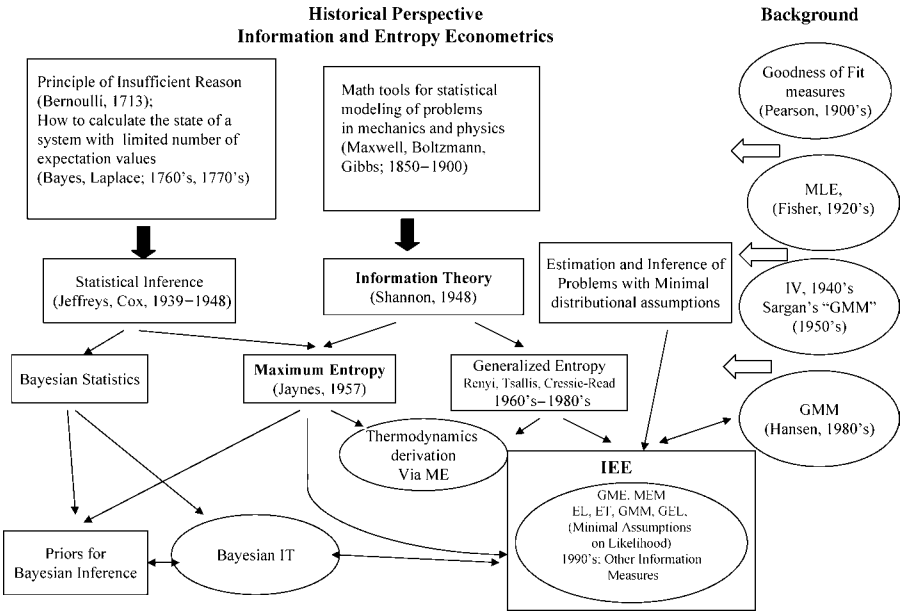


Fig. 2.1 Historical perspective: Information and entropy econometrics.

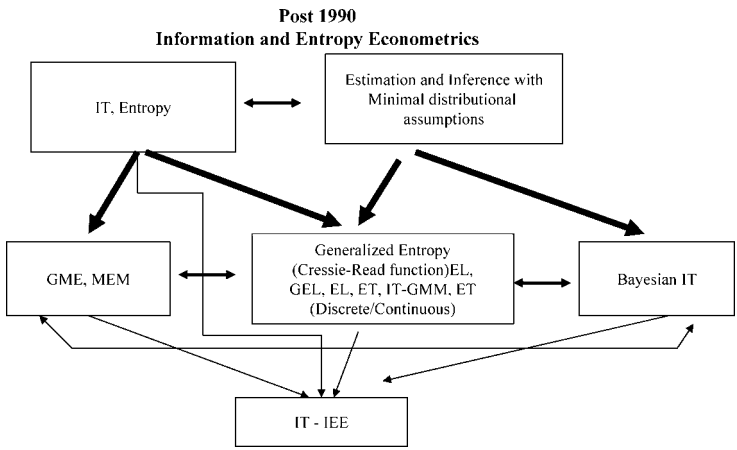


Fig. 2.2 Post 1990: Information and entropy econometrics.

Pearson on the Goodness-of-Fit measure and the Method of Moments (Pearson, 1894, 1900). That philosophy was continued by Fisher (1912, 1922) in his work on the Maximum Likelihood method and continued later with a (seemingly) competing approach known as the Minimum Chi-Square (Neyman and Pearson, 1928a,b), that led to Ferguson's Method of Moments (Ferguson, 1958). The latter approach extended the original Method of Moments to cases where there are more observed moments than unknown parameters. At the same period, Sargan (1958, 1959) developed his work on instrumental variables within a linear framework. In a parallel line of research, but via a somewhat logically simplified way, Godambe (1960) and Durbin (1960) developed the Estimating Functions approach to estimation for a similar class of problems. For a nice survey, see Bera and Bilias (2002).

In a most ingenious way Hansen (1982) developed the general theory of the Generalized Method of Moments (GMM) which builds on all of the previous work. His paper became one of the most influential papers in econometrics. Though the GMM was developed for different reasons than some of the other IT methods, it uses the same basic philosophy. The background to GMM was the research agenda in macro-econometrics in the late 1970's and early 1980's. For the models that were becoming of interest then, it was recognized that Maximum Likelihood (ML) was an unattractive method of estimation. The problem is that the implementation of ML requires the specification of the probability distribution of the data and that the ML estimator is only guaranteed to have desirable statistical properties if the method is implemented using the true distribution. Prior to the late 1970's, most empirical models (in economics) were linear and in these models it had been shown that ML under normality — the most common distributional assumption — is still consistent and asymptotically normal even if the true distribution is not normal. However, ML does not exhibit a similar robustness in the types of nonlinear models that became of interest at that period. This was a major problem for researchers because the underlying economic models (theory) did not specify the distribution of the data. Thus, in order to implement ML, it was necessary to make an arbitrary assumption on the distribution which, if incorrect, would likely undermine all subsequent inferences. GMM came about

because it was recognized that, while the underlying economic model does not specify the complete distribution of the data, the economic model does place restrictions on this distribution in the form of population moment conditions that provide information on the parameters of the model. GMM provides a convenient method for estimation of the parameters based on population moment conditions.

In approximately the same time (mid 1980's) the foundations of the Empirical Likelihood (EL) were established (Owen, 1988, 1990; Qin and Lawless, 1994). The EL is a nonparametric method for statistical inference. The EL allows the researcher to use likelihood methods but without assuming knowledge of the exact likelihood or knowledge of the exact family of distributions that generated the observed data. Using the wrong family of distributions can cause ML estimates to be inefficient and the corresponding confidence intervals and tests may be incorrect. In that respect the EL grew out of the same logic that brought about the classical ME — estimating the unknown parameters with minimal distributional assumptions.

Later on, in the 1990's the direct connection of the GMM to IEE and IT was shown (Kitamura and Stutzer, 1997; Imbens et al., 1998). This connection will be revisited shortly. At the same time, a generalized version of the EL (GEL) was introduced and it encompasses the ME and EL as well as other estimators within IEE and outside IEE (e.g., Imbens et al., 1998; Imbens, 2002; Smith, 2005, 2007; Kitamura, 2006).

In a parallel and independent research, in the late 1980's and early 1990's, the ME method was generalized (e.g., Golan and Judge, 1992; Golan et al., 1996b). The objectives of this line of research were to develop an estimation method that (i) can be used efficiently for estimating common economic data (that may be ill-behaved or ill-posed), (ii) uses minimal assumptions on the likelihood (or the data generating process), (iii) can incorporate the optimal conditions resulting from economic (or behavioral) theory, and (iv) can incorporate prior information. This method is known as the Generalized ME (GME) and it is another member of the IT family of estimators. The connection between all of these estimators is discussed in this review and synthesis.

With the above in mind, it is interesting to note the paths, outside of the social sciences, leading to the classical Maximum Entropy (ME)

via Information Theory. It seems logical to break these paths into two somewhat parallel paths.

The first path may be identified with the 18th century work of Jakob Bernoulli,[1] on the *principle of insufficient reason* and published eight years after his death in 1713, the work of Bayes (1763) and that of Laplace (1774). The fundamental question investigated by the above is to do with the basic problem of calculating the state of a system based on a limited number of expectation values (moments) represented by the data. This early work was later generalized by Jeffreys (1939) and Cox (1946) and is now known as Statistical Inference.

The second path leading to the classical Maximum Entropy, via the basics of IT, can be identified with the 19th century work of Maxwell (1859, 1876) and Boltzmann (1871), continued by Gibbs (1902) and Shannon (1948). This work is geared toward developing the mathematical tools for statistical modeling of problems in mechanics, physics, communication and information.

These two independent lines of research are similar. The objective of the first line of research is to formulate a theory/methodology that allows understanding of the general characteristics (distribution) of a system from partial and incomplete information. In the second line of research, this same objective is expressed as determining how to assign (initial) numerical values of probabilities when only some (theoretical) limited global quantities of the investigated system are known. Recognizing the common basic objectives of these two lines of research aided Jaynes (1957a,b) in the development of his classical work, the Maximum Entropy (ME) formalism. The ME formalism is based on the philosophy of the first line of research (Bernoulli, Bayes, Laplace, Jeffreys, Cox) and the mathematics of the second line of research (Maxwell, Boltzmann, Gibbs, Shannon).

The interrelationship between Information Theory, statistics and inference, and the ME principle started to become clear in the early work (1950's) of Kullback, Leibler and Lindley. Building on the basic concepts and properties of IT, Kullback and Leibler connected some of the fundamental statistics, such as sufficiency and efficiency (developed

[1] Known also as Jacque and James Bernoulli.

earlier within the context of ML by R. A. Fisher) to IT as well as provided a generalization of the Cramer–Rao inequality, and thus were able to unify heterogeneous statistical procedures via the concepts of IT (Kullback and Leibler, 1951; Kullback, 1954, 1959). Lindley (1956), on the other hand, provided the interpretation that a statistical sample could be viewed as a noisy channel (Shannon's terminology) that conveys a message about a parameter (or a set of parameters) with a certain prior distribution. In that way, he was able to apply Shannon's ideas to statistical theory by referring to the information in an experiment rather than in a message.

The interrelationship between Information Theory (IT), statistics and inference, and the ME principle may seem at first as coincidental and of interest only in a small number of specialized applications. But, by now it is clear that when these methods are used in conjunction, they are useful for analyzing a wide variety of problems in most disciplines of science. Examples include (i) work on image reconstruction and spectral analysis in medicine (such as brain scans and ECG signal analysis, magnetic resonance imaging known as MRI, X-ray computed tomography known as CT, positron emission tomography known as PET, as well as forecasting the potential spread of HIV), physics (such as tomography and deconvolution, molecular imaging and nuclear medicine as well as facial and voice recognition), chemistry and biology (such as sequences of proteins or DNA as well as modeling of species), topography (such as satellite images and constructions of maps), engineering, communication and information (such as search engines, information transmission and data updating), operations research (such as general matrix balancing and constrained optimization), political science (such as analyzing political surveys) and economics (input–output and social accounting analyses, linear and nonlinear regression analysis, Markov models and analysis of economic data while incorporating economic-theoretic information), (ii) research in statistical inference and estimation (linear and nonlinear models with minimal distributional assumptions), and (iii) ongoing innovations in information processing and IT.

The basic research objective in all of the above is how to formulate a theory/methodology that allows understanding of the general characteristics (distribution) of a system from partial and incomplete

information. That objective may be couched in the terminology of statistical decision theory and inference in which one has to decide on the "best" way of reconstructing an image (or a "message" in Shannon's work), making use of partial information about that image. Similarly, that objective may be couched within the more traditional terminology, where the basic question is how to recover the most conservative[2] estimates of some unknown function from limited data. The classical ME is designed to handle such questions and is commonly used as a method of estimating a probability distribution from an insufficient number of moments representing the only available information.

IEE is a natural continuation of IT and ME. All of the studies in IEE (developed mostly during the 1990s) build on both IT and/or ME to better understand the data while abstracting away from distributional assumptions or assumptions on the likelihood function. The outcome of these independent lines of study was a class of information-based estimation rules that differ but are related to each other. All of these types of methods perform well and are quite applicable to large classes of problems in the natural sciences and social sciences in general, and in economics in particular.

[2] By "conservative" we mean here the most uniform estimates that are based on minimal assumptions on the underlying likelihood.

3

Information and Entropy — Background, Definitions, and Examples

In this section a summary and discussion of the basics of IT is provided. Only concepts related to IEE are discussed. Readers who are familiar with these concepts should move to Section 4. The context of this section is taken from numerous sources. However, detailed discussions of most quantities discussed here can be found in the nice text of Cover and Thomas (1991). Detailed discussion of the basics and recent innovations in IT can be found in the recent survey (2004) in Foundations and Trends in Communications and Information Theory by Csiszar and Shields.

3.1 Information and Entropy — The Basics

3.1.1 Informal Definitions

The word "information" is very broadly defined. For example, the Oxford dictionary defines information as "The action of informing; formation or molding of the mind or character, training, instruction, teaching; communication of instructive knowledge." In the Webster's dictionary, "information" is defined as:

> "**1:** the communication or reception of knowledge or intelligence

> **2a**(1)**:** knowledge obtained from investigation, study, or instruction (2)**:** intelligence, news (3): facts, data **b:** the attribute inherent in and communicated by one of two or more alternative sequences or arrangements of something (as nucleotides in DNA or binary digits in a computer program) that produce specific effects **c**(1)**:** a signal or character (as in a communication system or computer) representing data (2): something (as a message, experimental data, or a picture) which justifies change in a construct (as a plan or theory) that represents physical or mental experience or another construct **d:** a quantitative measure of the content of information; *specifically*: a numerical quantity that measures the uncertainty in the outcome of an experiment to be performed."

Though by studying the above definitions it is obvious that "information" is a word packed with seemingly many meanings and somewhat vague in the more ordinary usage, it is still possible to specify a core concept of information (and related concepts, such as entropy) and use this precise mathematical definition as the basic measure that allows us to understand information, information processing, and data analysis. That precise definition is provided below. But before these concepts are precisely defined, it is helpful to briefly study the meaning of four basic concepts (information, uncertainty, entropy, and ignorance) as interpreted by scientists who work in these areas.[1] These interpretations are less vague than the dictionary definition and are more directly related to the basic questions asked by scientists in different disciplines. Though at first sight the interpretation of these concepts by different researchers seem quite different, this brief discussion below reveals that they are practically similar.

[1] The definitions of these four concepts are based on ideas and definitions sent to me by Essie Maasoumi, Ehsan Soofi, Arnold Zellner, Steve Kuhn, Jeff Perloff, Adom Giffin and Ariel Caticha, as well as definitions included in Cover and Thomas (1991), MacKay (2003), Rényi (1961, 1970), Shannon (1948), von Baeyer (2004), and comments provided by Alan Isaac, Essie Maasoumi, and Elise Golan.

Going back to the Greek philosophers, *information* may be viewed as a transfer of "form" from one medium to another. "Form" expresses relationship. Within the human context "transfer" is represented by "communication." Thus, *information* can be viewed as the "communication of relationships." In more recent history, *information* has two interconnected meanings. The informal one refers to the meaning of a message (data). The more scientific meaning emphasizes the symbols used to transmit a message or data (letters, numbers, zero–one digits, etc). In philosophical logic the *information* conveyed by a message, be it a sentence or any data, is sometimes identified with the number of possible words that it rules out. For example "A and B," conveys more information (i.e., is more informative) than A, and A conveys more information than "A or B".

Though expressed differently, the above definitions tell us that *information* reflects the decrease in ambiguity regarding a phenomenon. A common view of researchers in statistics, econometrics and other social sciences is that when we do not know a phenomenon with certainty, whatever reduces the bounds of possibility, or concentrates probability of possible outcomes, informs us. It is an addition to one's stock of knowledge; however measured and of whatever quality. For the applied researcher this means that *information* is anything, such as a fact or an opinion that affects one's estimates or decisions. It is "meaningful content." For example, the *information* in the random variable X about (the random variable) Y is the extent to which X changes the uncertainty about Y. When X is another random prospect, a particular outcome of it may increase or decrease uncertainty about Y. On average, however, if X and Y are related, knowledge of outcomes of X should decrease uncertainty about prediction of outcomes of Y. More technically, the *information* content of an outcome (of a random variable) is an inverse function of its probability. Thus, anything "surprising" (an outcome with a low probability) has a high level of *information*.

In natural sciences *information* is sometimes viewed as that quantity that represents distinguishability. *Information* is our connection to the world but it has no meaning by itself. If two pieces of *information* are indistinguishable then they are the same. Thus, *informa-*

tion only exists in the context of distinguishability. In more economic terms, and as viewed by some scientists across disciplines, like a force that induces a change in motion, *information* is whatever induces a rational agent to change her/his beliefs in a way that is prompted, but constrained, by the new *information*.

To summarize, the above different ways of reducing the vagueness of the ordinary meaning of the word *information* by researchers across the sciences, reveals that each discipline and scientist have their own interpretation and definition within the context of their research and understanding. However, a simple translation of these definitions reveals that we all (the Greek philosophers and the current logicians, the statisticians and the economists, the applied researchers in behavioral and in natural sciences) talk about data, the context of these data, data interpretation, and how to transfer data from one entity to another.

Uncertainty is a knowledge state in which it is impossible to reason logically to explain situations, events, experimental outcomes, etc. *Uncertainty* is sometimes called a "Knightian" state of limited knowledge or "doubt." Something is *uncertain* if it is possible but not known. (Here we mean "possibility" in what philosophers call an epistemic sense.) Stated differently, a proposition is uncertain if it is consistent with knowledge but not implied by knowledge. Connecting this notion of *uncertainty* with that one of information means that *uncertainty* is the amount of expected information that observations or experiments could reveal. The more they can reveal, the more *uncertainty* there is to be reduced. *Uncertainty* captures the unpredictability of an unknown prospect Y reflected in a probability distribution $f(y)$. The more concentrated is the probability distribution, the outcomes are more predictable, and hence the uncertainty about Y is lower. Therefore, the absolute bench mark of *uncertainty* is a uniform distribution (within the potential range of possibilities). To summarize, though coming from different disciplines and schools of thoughts (as with the word *information*) we all view *uncertainty* as arising from a proposition or a set of possible outcomes where none of the choices or outcomes is known with certainty. Therefore, these outcomes are represented by a certain probability distribution. The more uniform the distribution (given the

bounds) the higher the *uncertainty* that is associated with this set of propositions, or outcomes.

Entropy is expected information. It reflects what we expect to learn from observations, on average and it depends on how we measure information. In more technical words, *entropy* is a measure of uncertainty of a single random variable. Therefore, *entropy* can be viewed as a measure of uniformity. Similarly, but within a different context, *entropy* is also a measure of disorder of a system. The second law of thermodynamics states that the entropy of a (closed) system (like the universe) increases with time. It represents the progression of the system toward equilibrium that is reached at the highest level of entropy.

Entropy difference (or ratio) is a measure for comparison of uniformity. Relative *entropy* is a measure of "deviation" of uncertainty between two distributions. Within the notion of distinguishability and information, *entropy* is the tool that we use to determine the degree to which things are distinguished. *Entropy* is a tool for updating our prior probabilities (beliefs) to posterior (post-data) probabilities when new information becomes available. As with the other concepts discussed earlier, all of these different definitions converge to one coherent definition of *entropy* as an expected measure of information (or gain of information) measured relative to "uniformity," degree of distinguishablility and disorder.

Ignorance (or *absolute ignorance*) is not knowing (or not acknowledging) that there is uncertainty, and that (at a more abstract level) there is a distribution of possible outcomes or states of nature. It is a complete lack of knowledge, or information that would assist one in making a decision. Once we admit that all outcomes and states have a "law" that governs their realizations, we can debate about what that law is. Disagreements on the latter are not "ignorance." Not incorporating all that one might know to choose the "law" is being "uninformed," which is a relative concept, compared with ignorance. To summarize, unlike the previous three concepts, the concept of *ignorance* is defined similarly across disciplines. In practice, many researchers do not distinguish between the technical meanings of *ignorance* and *uncertainty*

as it is not relevant to a quantitative discussion of inference by rational agents.[2]

Finally, it is important to note that the views of *information* and its relationship with *uncertainty* expressed above are common to all researchers analyzing random processes or random data. But, for information that is absolutely certain, what is the link between *information* and probability (or a degree of rational belief) or uncertainty? If, for example I know a person's particulars, I have *information* about that person that I did not have before, but there is no real necessity for pondering *uncertainty* here. One could do that, but it is not absolutely necessary. This case is not discussed here as we are concerned here with estimation of random data representing life and behavior.

The above brief discussion on the meaning of information, entropy, uncertainty and ignorance, as commonly understood by scientists who work in these area, tells us that perhaps the meanings of the terms are approximately the same in different fields, but the objects to which they apply are different. In econometrics these objects are the data. The mathematical concepts of information and entropy are now defined.

3.1.2 Formal Definitions

Let $\boldsymbol{A} = \{a_1, a_2, \ldots, a_M\}$ be a finite set and $\boldsymbol{p}$ be a *proper* probability mass function on $\boldsymbol{A}$. "Proper" probability distribution means that all elements are nonnegative and the sum over all M elements equals exactly one. The amount of information needed to fully characterize all of the elements of this set consisting of M discrete elements is defined by $I(\boldsymbol{A}_M) = \log_2 M$ and is known as Hartley's formula (Hartley, 1928). This formula is developed within the context of communication theory. It is the logarithm of the number of possibilities (M) and, as such, it is a logarithm measure of information. Shannon (1948) built on Hartley's formula, within the context of communication process, to develop his information criterion. The Shannon's information content of an outcome a_i is $h(a_i) = h(p_i) \equiv \log_2 \frac{1}{p_i}$. As noted by Hartley (1928) the most natural choice for information measure is a logarithm function (as it

[2] As noted to me by Ariel Caticha, M. Tribus used the word "confusion" to refer to a situation where we do not even know what it is that we are ignorant or uncertain about.

captures additivity of information, discussed below). The choice of the logarithmic base corresponds to the choice one wishes to have for measuring information (Shannon, 1948). If the base 2 is used, the resulting units are "bits." A "bit" is a binary digit — a one or a zero and it is a basic unit of information. All information (data) can be specified in terms of bits. For example, the number 11000100 consists of 8 bits (or a "byte" — a unit of measurement of information storage in computer science). A random variable with two possible outcomes stores one bit of information. N such random variables store N bits of information because the total number of possible observed states/outcomes is 2^N and $\log_2(2^N) = N$ (Shannon, 1948). The choice of base 2 seems to be a natural choice (see examples below and in following sections) as it provides the most efficient (cheapest) way of coding and decoding the data.

Shannon's criterion, called *entropy*,[3] reflects the expected informational content of an outcome and is defined as

$$H(\boldsymbol{p}) \equiv \sum_{i=1}^{M} p_i \log_2 \frac{1}{p_i} = -\sum_{i=1}^{M} p_i \log_2 p_i = E[\log_2(1/p(X))] \qquad (3.1)$$

for the random variable X and with $x \log_2(x)$ tending to zero as x tends to zero. This information criterion, expressed in bits, measures the uncertainty or informational content of X that is implied by $\boldsymbol{p}$.

Example. Suppose the English language had only five letters: A, B, E, Y, and Z. Table 3.1 summarizes the informational content of each letter and the entropy under two scenarios. Scenario 1 reflects the idea that all letters are used with the same frequency, while in scenario 2 the frequency used of each letter is more realistic and is based on the relative frequency observed in the English language. In the second scenario, the entropy is lower because there is less uncertainty (more information) about the next outcome of the random variable X because, in that case,

[3] In completing his work, Shannon noted that "information" is already an overused term. The "legend" is that he approached his colleague John von Neumann, who responded: "You should call it entropy for two reasons: first, the function is already in use in thermodynamics under the same name; second, and more importantly, most people do not know what entropy really is, and if you use the word *entropy* in an argument you will win every time."

Table 3.1 Shannon information and entropy of a five letter english language.

	Scenario 1 Uniform		Scenario 2 Non-Uniform	
a_i	p_i	$h(p_i)$	p_i	$h(p_i)$
a	0.2	2.322	0.322	1.635
b	0.2	2.322	0.072	3.796
e	0.2	2.322	0.511	0.969
y	0.2	2.322	0.091	3.458
z	0.2	2.322	0.004	7.966
Sum	1		1	
Entropy		2.322		1.641

the letters "A" and "E" occur with much higher probabilities than the other three letters.

These two scenarios can be viewed as two different discrete random variables with similar outcomes but with different probabilities associated with the outcomes. Consider, a second example of a random variable X that takes the value of one with probability p, and the value of zero with probability $1-p$. In that case $H(X) = H(\boldsymbol{p}) = -p\log_2 p - (1-p)\log_2(1-p)$. $H(\boldsymbol{p})$ reaches a maximum level of one bit when $p = 1/2$ and is equal to zero when $p = 0$ or $p = 1$. From this example one learns that $H(\boldsymbol{p})$ is concave in $\boldsymbol{p}$, reaches a maximum for uniform probabilities (complete ignorance) and is equal to zero (perfect certainty) when one of the probabilities is exactly one ($p = 0$ or $p = 1$). More generally speaking, the entropy measure $H(\boldsymbol{p})$ reaches a maximum when $p_1 = p_2 = \cdots = p_M = 1/M$ (and is equal to Hartley's formula) and a minimum with a point mass function. The entropy $H(\boldsymbol{p})$ is a function of the probability distribution $\boldsymbol{p}$ and not a function of the actual values taken by the random variable. If X is a random variable with possible distinct realizations $x_1, x_2, \ldots, x_M$ with corresponding probabilities $p_1, p_2, \ldots, p_M$, the entropy $H(\boldsymbol{p})$ does not depend on the values $x_1, x_2, \ldots, x_M$ of X, but rather depends on $p_1, p_2, \ldots, p_M$ (e.g., Table 3.1).

Example (*Probability, Information, and Entropy*). For the binary probability distribution (with p and $1-p$), Table 3.2 and Figures 3.1 and 3.2 present Shannon's information, $h(a_i) = h(p_i)$, and the entropy $H(\boldsymbol{p})$ as functions of p for $p \in (0,1)$.

Table 3.2 Information and entropy of an outcome of a random variable with probabilities p and $1 - p$.

p	Information	Entropy
1.00E-05	16.61	1.81E-04
1.00E-04	13.288	0.001
0.001	9.966	0.011
0.01	6.644	0.081
0.1	3.322	0.469
0.15	2.737	0.61
0.2	2.322	0.722
0.25	2	0.811
0.3	1.737	0.881
0.35	1.515	0.934
0.4	1.322	0.971
0.45	1.152	0.993
0.5	1	1

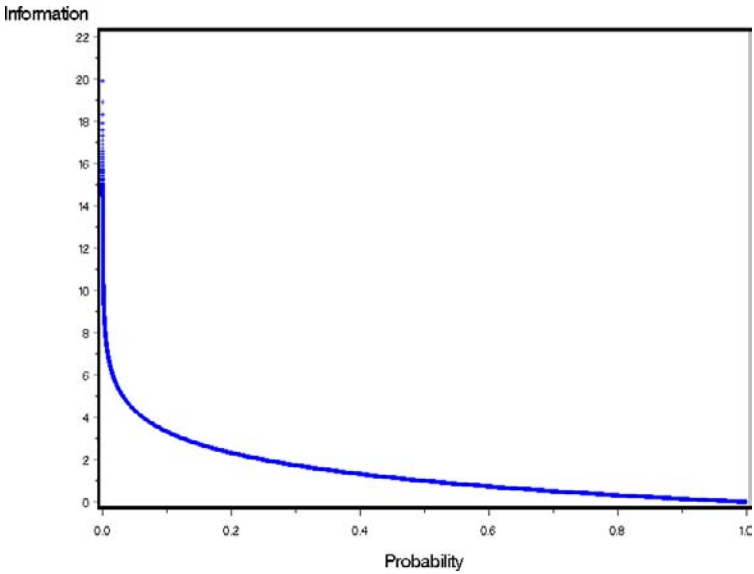


Fig. 3.1 The relationship between information and probability.

To get a better idea of the meaning of information and entropy, consider the following common example (e.g., Clarke, 2007), that analyzes the minimal number of binary questions (yes–no; 0–1) needed to determine the value of the random variable X. Let X be a random variable with three possible realizations: Blue (B), Red (R), and Yellow (Y). The

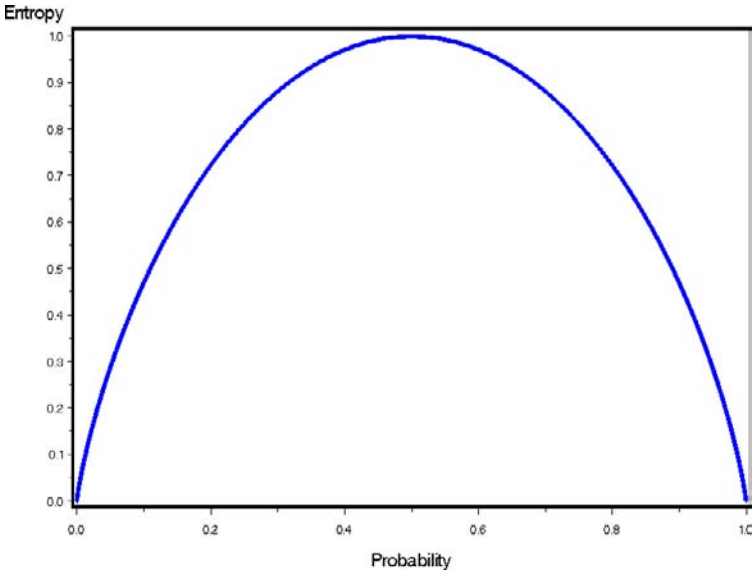


Fig. 3.2 The relationship between probability and entropy.

probability associated with each realization is $1/2$, $1/4$, and $1/4$, respectively. The entropy $H(\boldsymbol{p}) = -1/2\log_2(1/2) - 1/4\log_2(1/4) - 1/4\log_2(1/4) = 1/2\log_2 2 + 1/4\log_2 4 + 1/4\log_2 4 = 3/2$. Now, for those that are unfamiliar with the concept of entropy, but want to get an idea of the information associated with X, it is possible to proceed in a different way. In that case, they want to determine the value of X with a minimal number of binary questions. (We relate this to minimal error probability in Section 3.6.) To reach that minimum, it is possible to follow a tree diagram describing the possible events from highest to lowest probability, and start with a question on the event with the highest probability: "Is $X = \mathrm{B}$?" If the answer is "no" (half the times), our second question is "Is $X = \mathrm{R}$?" or similarly, "is $X = Y$?" In that case the answer is yes half the times and no the other half. But this answer is contingent on moving on to the second question. The expected number of binary questions is $3/2 (= 1 \times 1/2 + 2 \times 1/2)$. This example shows that the entropy of this random variable equals the *minimal* number of binary questions necessary to determine the value of X. Therefore, the entropy reflects the maximal possible amount of information about that

random variable. If for example, one starts with the "wrong" question (for an event with probability smaller than $1/2$ in that example) say, "Is $X = Y$?" the expected number of binary questions increases ($1\times 1/4 + 2\times 3/4 > 3/4$). Shannon showed that the minimum expected number of binary questions required to fully determine the value of any random variable X is between $H(X)$ and $H(X) + 1$.

Before concluding our discussion of Shannon's entropy, a different derivation of that quantity, via a combinatorial argument, is provided. This is a large sample derivation. Suppose N outcomes resulting from an experiment with K possible outcomes are observed. Let $N_1, \ldots, N_K$ be the number of observed outcomes for each state K in the N trials. Thus, $\sum_k N_k = N$ and $N_k \geq 0$. Naturally, there are K^N possible sequences resulting from the N trials. Out of the N trials, one can use the multiplicity factor, $W = \frac{N!}{\prod_k N_k!}$, to find the number of ways a particular set of frequencies (or N_k) can be realized. But first, it is helpful to define the frequency $\pi_k \equiv \frac{N_k}{N}$ or $N_k = \pi_k N$. Using the transformation $\log W = \log N! - \sum_k \log N_k!$, and Stirling's approximation ($\log x! \approx x \log x - x$ as $0 < x \to \infty$ where "$\approx$" stands for approximation) one gets

$$\begin{aligned}
\log W &\approx N \log N - N - \sum_k N_k \log N_k + \sum_k N_k \\
&= N \log N - \sum_k N_k \log N_k \\
&\overset{N\to\infty}{=} N \log N - \sum_k N\pi_k \log N\pi_k \\
&= N \log N - \sum_k N_k \log N - N \sum_k \pi_k \log \pi_k \\
&= -N \sum_k \pi_k \log \pi_k
\end{aligned}$$

and finally,

$$N^{-1} \log W \approx -\sum_k \pi_k \log \pi_k = H(\boldsymbol{\pi}).$$

To summarize, the entropy is the number of bits needed, on average, to describe a random variable and consequently it captures the average uncertainty in that random variable.[4]

3.2 Discussion

After Shannon introduced this measure, a fundamental question arose: whose information does this measure capture? Is it the information of the "sender," the "receiver" or the communication channel?[5] Similarly, is this the information in the data or the information of the observer? To try and answer this question, let us first suppose that H measures the increase in the knowledge level of the receiver after receiving the message. But this seemingly natural interpretation contradicts Shannon's idea. He used H to measure the overall capacity required in a channel to transmit a certain message at a given rate. Therefore, H is free of the receiver's level of ignorance. So what does it measure?

Going back to our earlier discussion, one answer to this question is that H is a measure of the expected amount of information in a message or in some data. To measure information, one must abstract away from any form or content of the message itself. For example, in the old-time telegraph office, where only the number of words was counted in calculating the price of a telegram, one's objective was to minimize the number of words in a message while conveying all necessary information. Likewise, the information in a message can be expressed as the number of signs (or distinct symbols) necessary to express that message in the most concise and efficient way. Any system of signs can be used, but the most reasonable (efficient) one is to express the amount of information by the number of signs necessary to express it by zeros and ones. In that way, messages and data can be compared by their informational content. Each digit takes on the values zero or one, and the information specifying which of these two possibilities occurred is called a unit of information (see earlier definition of "bit"). The answer to a question that can only be answered by "yes" and no" contains

[4] In other words, entropy is the minimal descriptive complexity of a random variable.

[5] Within the context of IT, "channel" means any process capable of transmitting information.

exactly one unit of information regardless of the meaning of that question. This unit of information is called a "bit" or a binary digit.[6]

Rényi (1961, 1970) showed that, for a (sufficiently often) repeated experiment, one needs, on average, a total of $H(\boldsymbol{p}) + \varepsilon$ binary symbols (for any positive ε) in order to fully characterize an outcome of that experiment. Thus, it seems logical to argue that the outcome of an experiment contains the amount of information $H(\boldsymbol{p})$.

The information discussed here is not "subjective" information of a particular researcher. The information contained in a single observation, or a data set, is a certain quantity that is independent of whether the observer (e.g., an economist or a computer) recognizes it or not. Consequently, $H(\boldsymbol{p})$ measures the average amount of information provided by an outcome of a random drawing governed by $\boldsymbol{p}$. In the same way, $H(\boldsymbol{p})$ is a measure of uncertainty about a specific possible outcome before observing it, which is naturally related to the amount of randomness represented by $\boldsymbol{p}$.[7]

In a more common econometric and statistical terminology, $H(\boldsymbol{p})$ can be viewed in the following way. The researcher never knows the true underlying values characterizing an economic system. Therefore, one may incorporate her/his understanding and knowledge of the system in reconstructing (estimating) the image (unknown parameters) where this knowledge appears in terms of some global macro-level quantities, such as moments. As is shown and discussed in Section 4 below, out of all possible such images, where these moment conditions are retained, one should choose the image having the maximum level of entropy, H. The entropy of the analyzed economic system measures the uncertainty (or relative ignorance) of the researcher who knows only some moments' values representing the underlying population. For a more detailed discussion of the statistical meaning of information see the nice texts by Cover and Thomas (1991) and MacKay (2003), as well as the original

[6] Shannon's realization that the binary digits could be used to represent words, sounds, images and ideas, is based on the work of Boole (1854), the 19th-century British mathematician, who invented the two-symbol logic in his work "The Laws of Thought."

[7] According to both Shannon and Jaynes, within the theory of communication and information, H measures the degree of ignorance of a communication engineer who designs the technical equipment of a communication channel because it takes into account the set of all possible messages to be transmitted over this channel during its life time.

work of Shannon (1948) and Rényi (1970) and more recent articles, within the econometric literature, by Maasoumi (1993), Soofi and Retzer (2002), Clarke (2007) and Zellner (1988, 2002).

3.3 Multiple Random Variables, Dependency, and Joint Entropy

Consider now extending the notion of entropy to more than a single random variable. For ease of notation, I use "log" and "$\log_2$" interchangeably.[8] Let X and Y be two discrete random variables with possible realizations $x_1, x_2, \ldots, x_K$ and $y_1, y_2, \ldots, y_J$, respectively. Let $p(X,Y)$ be a joint probability distribution. Now, define $P(X = x_k) = p_k$, $P(Y = y_j) = q_j$, $P(X = x_k, X = y_j) = w_{kj}$, $P(X|Y) = P(X = x_k|Y = y_j) = p_{k|j}$, and $P(Y|X) = P(Y = y_j|X = x_k) = q_{j|k}$ where $p_k = \sum_{j=1}^{J} w_{kj}$, $q_j = \sum_{k=1}^{K} w_{kj}$ and the conditional probabilities satisfy $w_{kj} = q_j p_{k|j} = p_k q_{j|k}$.

The *joint entropy* of X and Y is

$$H(X,Y) \equiv \sum_{k,j} w_{kj} \log \frac{1}{w_{kj}} = -\sum_{k,j} w_{kj} \log w_{kj}.$$

The *conditional entropy* $H(X|Y)$ is the total information in X with the condition that Y has a certain value:

$$\begin{aligned} H(X|Y) &= \sum_j q_j \left[-\sum_k p_{k|j} \log p_{k|j} \right] \\ &= \sum_j q_j \left[-\sum_k \left(\frac{w_{kj}}{q_j} \right) \log \left(\frac{w_{kj}}{q_j} \right) \right] \\ &= \sum_{k,j} w_{kj} \log \left(\frac{q_j}{w_{kj}} \right). \end{aligned}$$

The relationship among all of the above entropies is easily seen in the following quantity representing the entropy of a composite event which equals the sum of the marginal and conditional entropies (chain

[8] There is a one to one transformation of the entropy value from a log base b to a log base a : $H_b(X) = \log_b(a)[H_a(X)]$.

rule for entropies):

$$H(X,Y) = H(Y) + H(X|Y) = H(X) + H(Y|X). \tag{3.2}$$

If X and Y are independent ($w_{kj} = p_k q_j$), the above equation reduces to

$$H(X,Y) = H(X) + H(Y). \tag{3.3}$$

The *relative entropy* that provides an informational distance between two proper distributions is now defined. The relative entropy, also known as the Kullback–Liebler distance function (or cross-entropy), between the two probability mass functions $\boldsymbol{p}$ and $\boldsymbol{q}$ for the random variables X and Y is

$$D(\boldsymbol{p}||\boldsymbol{q}) \equiv \sum_{k=1}^{K} p_k \log(p_k/q_k). \tag{3.4}$$

The relative entropy $D(\boldsymbol{p}||\boldsymbol{q})$, or sometimes called $D(X||Y)$, reflects the gain in information resulting from the additional knowledge in $\boldsymbol{p}$ relative to $\boldsymbol{q}$. It is an information-theoretic distance of $\boldsymbol{p}$ from $\boldsymbol{q}$ that measures the inefficiency of assuming *a priori* that the distribution is $\boldsymbol{q}$ when the correct distribution is $\boldsymbol{p}$ (see Gokhale and Kullback, 1978). If Popeye believes the random drawing is governed by $\boldsymbol{q}$ (for example, $q_k = 1/K$ for all $k = 1, 2, \ldots, K$) while Olive Oyl knows the true probability $\boldsymbol{p}$ (which is different than uniform), then $D(\boldsymbol{p}||\boldsymbol{q})$ measures how much less informed Popeye is relative to Olive Oyl about the possible outcome. Similarly, $D(\boldsymbol{p}||\boldsymbol{q})$ measures the gain in information when Popeye learns that Olive Oyl is correct — the true distribution is $\boldsymbol{p}$, rather than $\boldsymbol{q}$. Equivalently, $D(\boldsymbol{p}||\boldsymbol{q})$ represents Popeye's loss of information when he uses $\boldsymbol{q}$. In a more information-theoretic language, if Popeye knew the true distribution of the random variable, he could construct a code with an average description length of $H(\boldsymbol{p})$ to describe the random variable. But if he uses his code for the incorrect distribution $\boldsymbol{q}$, he will need $H(\boldsymbol{p}) + D(\boldsymbol{p}||\boldsymbol{q})$ bits on the average to describe the random variable. In more econometric terms, using the incorrect likelihood, or model, when analyzing data is costly not only in terms of efficiency and precision but also may lead to an inconsistent estimator. For further discussion on this measure see Cover and Thomas

(1991). Note that $D(\boldsymbol{p}||\boldsymbol{q})$ is not a true distance and is not symmetric $[D(\boldsymbol{p}||\boldsymbol{q}) \neq D(\boldsymbol{q}||\boldsymbol{p})]$.

Finally, it is worthwhile noting the relationship between $D(\boldsymbol{p}||\boldsymbol{q})$ and the L_1 distance function. Recall that $L_1 \equiv \|\boldsymbol{p} - \boldsymbol{q}\|_1 = \sum_x |\boldsymbol{p}(x) - \boldsymbol{q}(x)|$, then

$$D(\boldsymbol{p}||\boldsymbol{q}) \geq \frac{1}{2\ln 2} L_1^2 = \frac{1}{2\ln 2} \|\boldsymbol{p} - \boldsymbol{q}\|_1^2.$$

For applications of this within the IEE literature see for example Antoine et al. (2007). The above divergence measures are also known as a class of f-divergence measures of a distribution $\boldsymbol{p}$ from $\boldsymbol{q}$. See the nice discussion in Csiszar and Shields (2004).

To find out the amount of information contained in a random variable about another random variable, called the *mutual information* between the random variables X and Y, the following quantity is defined:

$$I(X;Y) \equiv \sum_{k,j} w_{kj} \log \frac{w_{kj}}{p_k q_j} = D(w_{kj}||p_k q_j) = H(X) - H(X|Y).$$

The mutual information $I(X;Y)$ captures the reduction in uncertainty of X due to our knowledge of Y. It is the *marginal additional information* the econometrician, analyzing X, gains from knowing Y.

Following the logic of Jensen's inequality, it is possible to show that

1. $I(X;Y) \geq 0$ with equality if and only if X and Y are independent.
2. $D(\boldsymbol{p}||\boldsymbol{q}) \geq 0$ with equality if and only if $p_k = q_k$ for all k.

Generally speaking, the above quantities show that the additional information coming from another random variable, or data, results in reducing the uncertainty (or ignorance) one has about the original random variable (or data) X. Conditioning reduces entropy for non-independent random variables. Related to the above, but stated slightly different, it can be shown that given n random variables $X_i, i = 1, \ldots, n$, $H(X_1, X_2, X_3, \ldots, X_n) \leq \sum_{i=1}^n H(X_i)$ where the equality holds if and only if all the X_i's are independent.

The example below presents the different entropy quantities, defined above for a joint distribution of two discrete random variables. The

Table 3.3 Joint probabilities.

$\boldsymbol{P(x,y)}$		x			$\boldsymbol{P(y)}$
		1	2	3	
	1	0	0	1/3	1/3
y	2	1/9	1/9	1/9	1/3
	3	1/18	1/9	1/6	1/3
$\boldsymbol{P(x)}$		1/6	2/9	11/18	

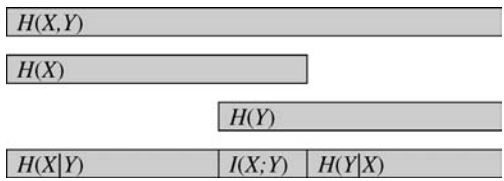
relationships among these quantities are shown as well. This example also demonstrates the relationships between information, uncertainty, uniformity, and entropy. These are represented in the entropy values of the two random variables: X and Y (where the uncertainty associated with Y, relative to X, is higher).

Example. Different entropies resulting from the joint distribution presented in Table 3.3.

Table 3.4 Entropies of Table 3.3.

$$
\begin{aligned}
H(X) &= 1.347 \\
H(Y) &= 1.585 = \log_2(3) = \mathrm{Max}(H) \\
H(X,Y) &= 2.600 \\
H(X|Y) &= 1.015 \\
H(Y|X) &= 1.252 \\
D(\boldsymbol{p}||\boldsymbol{q}) &= 0.238 \\
D(\boldsymbol{q}||\boldsymbol{p}) &= 0.237 \text{ Note: } D(\boldsymbol{p}||\boldsymbol{q}) \neq D(\boldsymbol{q}||\boldsymbol{p}) \\
I(X;Y) &= 0.333 \\
H(X,Y) &= H(Y) + H(X|Y) = H(X) + H(Y|X) \\
2.6 &= 1.585 + 1.015 = 1.347 + 1.252 \\
I(X;Y) &= H(X) - H(X|Y) = H(Y) - H(Y|X) \\
0.3326 &= 1.3472 - 1.0146 = 1.585 - 1.2524
\end{aligned}
$$

The above values can also be presented graphically:



Building on the above example, it is easy to see (Table 3.5) that, on average, more data is better: $H(X|Y) < H(X)$. Note that for a specific

Table 3.5 Conditional entropies.

$\boldsymbol{P(x,y)}$		x			$\boldsymbol{H}(X\|Y)$
		1	2	3	
	1	0	0	1	0
y	2	1/3	1/3	1/3	1.585
	3	1/6	1/3	1/2	1.459
					$\boldsymbol{H}(X\|Y) = 1.015$

outcome, say $y = 1$, $H(X|y = 1)$ is smaller than $H(X)$ while for $y = 2$ or $y = 3$ $H(X|y) > H(X)$, but on average, $H(X|Y)$, there is a gain in information.

Finally, two quantities that resemble the correlation and covariance are shown. Let X and Y be identically distributed random variables so $H(X) = H(Y)$. Then, the following measure reflects the (linear and nonlinear[9]) correlation between these two random variables:

$$r(X,Y) = 1 - \frac{H(X|Y)}{H(Y)} = \frac{I(X;Y)}{H(Y)},$$

where $0 \leq r(X,Y) \leq 1$, $r(X,Y) = 0$ iff X and Y are independent ($I(X;Y) = 0$), and $r(X,Y) = 1$ if X is a function of Y ($H(X|Y) = 0$). The following quantity connecting the different entropies of these two variables, resemble a covariance:

$$\begin{aligned} c(X,Y) &= H(X|Y) + H(Y|X) \\ &= H(X) + H(Y) - 2I(X;Y) \\ &= H(X,Y) - I(X;Y) \\ &= 2H(X,Y) - H(X) - H(Y). \end{aligned}$$

3.4 Generalized Entropy

Building on Shannon's work, a number of generalized information measures were developed. Though none of these measures exhibits the exact properties of the Shannon's entropy, these measures are used often in econometrics and provide a basis for defining IT estimators. These generalized information measures are all indexed by a single parameter α.

[9] See also discussion in the next section.

Starting with the idea of describing the gain of information, Rényi (1961) developed the entropy of order α for incomplete random variables.[10] The relevant generalized entropy measure of a *proper* probability distribution (Rényi, 1970) is

$$H_\alpha^R(\boldsymbol{p}) = \frac{1}{1-\alpha}\log\sum_k p_k^\alpha. \tag{3.5}$$

The Shannon measure is a special case of this measure where $\alpha \to 1$. Similarly, the (Rényi) cross-entropy (between two distributions: $\boldsymbol{p}$ and $\boldsymbol{q}$) of order α is

$$D_\alpha^R(\boldsymbol{x}|\boldsymbol{y}) = D_\alpha^R(\boldsymbol{p}\|\boldsymbol{q}) = \frac{1}{1-\alpha}\log\sum_k \frac{p_k^\alpha}{q_k^{\alpha-1}}, \tag{3.6}$$

which is equal to the traditional cross-entropy measure (3.4) as $\alpha \to 1$.

Building on Rényi's work, and independent of his work, a number of other generalizations were developed. These generalizations include the less known *Bergman* distance and the *f-entropy* measures. However, the more commonly used generalized measures in IEE are those that were developed during the 1980's by Cressie and Read (1984) and Tsallis (1988). The cross-entropy version of the Tsallis measure is

$$D_\alpha^T(\boldsymbol{x}|\boldsymbol{y}) = D_\alpha^T(\boldsymbol{p}\|\boldsymbol{q}) = \frac{1}{1-\alpha}\left(\sum_k \frac{p_k^\alpha}{q_k^{\alpha-1}} - 1\right). \tag{3.7}$$

It is interesting to note that the functional form of (3.5) — Rényi — resembles the CES production function, while Tsallis's Eq. (3.7) is similar to the Box-Cox function.

The commonly used Cressie–Read (1984) measure is

$$D_\alpha^{CR}(\boldsymbol{x}|\boldsymbol{y}) = D_\alpha^{CR}(\boldsymbol{p}\|\boldsymbol{q}) = \frac{1}{\alpha(1+\alpha)}\sum_k p_k\left[\left(\frac{p_k}{q_k}\right)^\alpha - 1\right]. \tag{3.8}$$

The Rényi and Tsallis entropies have been compared in Tsallis (1988) and Holste et al. (1998) to show that:

$$H_\alpha^R(\boldsymbol{x}) = [1/(1-\alpha)]\log[1 + (1-\alpha)\log H_\alpha^T].$$

[10] If $\boldsymbol{\eta}$ is an incomplete, discrete random variable with M distinct realizations, then $\sum_i p_i \leq 1$ (rather than $\sum_i p_i = 1$) where $p_i > 0$; $i = 1, \ldots, M$.

It has been further shown (Golan, 2002) that all of these measures (including Cressie–Read) are connected:

$$\begin{aligned} D^R_{\alpha+1}(\boldsymbol{p}\|\boldsymbol{q}) &= -\frac{1}{\alpha}\log[1-\alpha D^T_{\alpha+1}(\boldsymbol{p}\|\boldsymbol{q})] \\ &= -\frac{1}{\alpha}\log[1+\alpha(\alpha+1)D^{CR}_{\alpha}(\boldsymbol{p}\|\boldsymbol{q})] \end{aligned} \tag{3.9}$$

where the Tsallis and Rényi measures of order $(\alpha+1)$ are compared with that of Cressie–Read measure of order α. To make this comparison more general, Eq. (3.9) is in terms of the "cross-entropy" between the two distributions $\boldsymbol{p}$ and $\boldsymbol{q}$, where the traditional cross-entropy measure is a special case of the above for $\alpha \to 0$.

All of the above measures are commonly known as α-entropies. For completeness, it is noted that the α-entropy is also known as "Chernoff entropy." Chernoff (1952) introduced this measure in his classical work on asymptotic efficiency of hypothesis tests. Chernoff entropy is found by starting with (3.9), and letting $\alpha = 1-\beta$ with $0 < \beta < 1$.

With Shannon's entropy measure, events with high or low probability do not contribute much to the measure's value. With the generalized entropy measures for $\alpha > 1$, higher probability events contribute more to the value than do lower probability events. Unlike the Shannon's measure (3.1), the average logarithm is replaced by an average of probabilities raised to the α power. Thus, a change in α changes the relative contribution of event k to the total sum. The larger the α, the more weight the "larger" probabilities receive in the sum.[11]

3.5 Axioms

Shannon's entropy can also be derived from a set of primitive axioms. For completeness one such set of axioms is provided here. A parallel

[11] Versions of Eq. (3.9) are commonly used to investigate the linear and nonlinear dependence among random variables. For example, take the mutual information (defined as the expected information in an outcome of a random draw from Y about an outcome of a random draw from X) version of (3.9) for two discrete random variables X and Y of dimension N, and for $\alpha = 1$ yields $D^R_2(X||Y) \equiv H^R_2(Y) - [H^R_2(X,Y) - H^R_2(X)]$. This measure equals zero if and only if X and Y are statistically independent, and it equals $\log(N)$ if and only if $Y = f(X)$, where f can be any linear or nonlinear function. In general, this type of measure is used for any value of α. For more, see for example, Soofi (1994), Holste et al. (1998), Maasoumi and Racine (2002) and Racine and Maasoumi (2007).

set of axioms is discussed in Csiszar and Korner (1981). However, as noted by Shannon, "... these axioms and the following theorem are in no way necessary for the present theory [information theory] ... the real justification of these [information and entropy] definitions, however, will reside in their implication" (Shannon, 1948, p. 11).

For some discrete random variable with a set of possible realizations characterized by a known proper probability distribution $\boldsymbol{p}$, we are searching for a measure, H, describing the amount of our uncertainty of the outcome. Let $H_K(p_1, p_2, p_3, \ldots, p_K)$ be a sequence of symmetric functions satisfying the following properties:

1. *Normalization.* $H_2\left(\frac{1}{2}, \frac{1}{2}\right) = 1$. (The measure H should be unchanged if the outcomes are reordered, should reach a maximum level for uniform probabilities and should increase with the number of possible outcomes.)
2. *Continuity.* $H_2(p, 1-p)$ is a continuous function of $\boldsymbol{p}$. (Changing the probabilities by a very small amount will change the measure H by a very small amount.)
3. *Shannon's Additivity.* $H_K(p_1, p_2, p_3, \ldots, p_K) = H_{K-1}(p_1 + p_2, p_3, \ldots, p_K) + (p_1 + p_2) H_2\left(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}\right)$. (Breaking the data into two mutually exclusive sub-sets should not change the total amount of H. More formally, if a choice is broken down into two mutually exclusive choices, the original measure H should be the weighted sum of the individual broken $H's$.)

Then, for any $K = 2, 3, \ldots, H_K(p_1, p_2, p_3, \ldots, p_K) = \sum_{k=1}^{K} p_k \log_2 p_k$.

With that in mind it is beneficial to look at the axiomatic differences between the Shannon's entropy and the other entropies. These measures share three properties. First, all these entropy measures are nonnegative for any arbitrary $\boldsymbol{p}$. These measures are strictly positive except when all probabilities but one equals zero (perfect certainty). Second, these measures reach a maximum value when all probabilities are equal. Third, each measure is concave for arbitrary $\boldsymbol{p}$. In addition, the generalized entropy measures share the property that they all are monotonically decreasing functions of α for any $\boldsymbol{p}$.

The entropy measures differ in terms of their additivity properties. Following (3.2) and (3.3), Shannon entropy has the property that a composite event equals the sum of the marginal and conditional entropies:

$$H(X,Y) = H(Y) + H(X|Y) = H(X) + H(Y|X). \tag{3.10}$$

This property does not hold for the generalized measures. But, if X and Y are independent, Eq. (3.10) reduces to

$$H(X,Y) = H(X) + H(Y), \tag{3.11}$$

which is the property of *standard additivity*, that holds for both the Shannon and Rényi entropy measures, but not for the Tsallis measure which is pseudo-additive. For two *independent* subsets A and B, H_α^T is "pseudo-additive" and satisfies

$$H_\alpha^T(A,B) = H_\alpha^T(A) + H_\alpha^T(B) + (1-\alpha)H_\alpha^T(A)H_\alpha^T(B) \quad \text{for all } \alpha$$

where

$$H_\alpha^T(A,B) \equiv H_\alpha^T(X,Y) = \left(\sum_{k,j} w_{kj}^\alpha - 1\right) \Big/ (1-\alpha).$$

Finally, only the Shannon and Tsallis measures have the property of *Shannon additivity* defined above. For completeness, that property is restated now in a more general way. The total amount of information in the entire sample is a weighted average of the information in two mutually exclusive subsamples, A and B. Let the probabilities for subsample A be $\{p_1,\ldots,p_L\}$ and those for B be $\{p_{L+1},\ldots,p_K\}$, and define $p_A = \sum_{k=1}^{L} p_k$ and $p_B = \sum_{k=L+1}^{K} p_k$. Then, for all α, (*including* $\alpha = 1$),

$$\begin{aligned} H_\alpha^T(p_1,\ldots,p_K) = {} & H_\alpha^T(p_A,p_B) + p_A^\alpha H_\alpha^T(p_1/p_A,\ldots,p_L/p_A) \\ & + p_B^\alpha H_\alpha^T(p_{L+1}/p_B,\ldots,p_K/p_B). \end{aligned}$$

3.6 Errors and Entropy

In this section, it is shown that entropy measures can be used to bound the probability of error when analyzing conditional or unconditional

data. Consider a traditional estimation problem of say, the next outcome of a discrete random variable X with an underlying probability distribution $p(x)$. But rather than observing X, the random variable Y that is conditioned on X is observed. For simplicity it is assumed here that X and Y are of the same dimension. Let $\hat{X} = f(Y)$ be an estimate of X. With that estimate, the next natural question is what is the probability that $\hat{X} = X$? Let $P_e \equiv \text{Prob}\{\hat{X} \neq X\}$ be the probability of error. Then, using Fano's inequality (Fano, 1961) representing the relationship between the entropy (or the conditional entropy) and the probability of error is shown to be

$$H(P_e) + P_e \log(K - 1) \geq H(X|Y),$$

where K is the dimension of X or Y.[12] A weaker version of this inequality is

$$P_e \geq [H(X|Y) - 1]/\log(K),$$

which is just a function of the entropy. We will revisit this relationship in later sections.

Example. In this example the mathematical relationships of the above (Fano) inequalities are calculated for the unconditional and conditional cases. Consider the simple, unconditional $P(x) = (1/6,\ 2/9,\ 11/18)$ in Table 3.3. Since $X = 3$ has the highest probability, our best estimate is $\hat{X} = 3$ so the probability of error here is $P_e = 1 - p(X = 3) = 0.389$. Now, $H(P_e) = -P_e \log P_e - (1 - P_e)\log(1 - P_e) = 0.964$ and $K = 3$, so $0.964 + 0.389\log(2) = 1.353 \geq H(X) = 1.347$ and the above inequality is satisfied. Similarly, using the weaker inequality, one gets $P_e \geq [H(X) - 1]/\log(K) = 0.219$. Again, the above weak inequality is satisfied. Now consider the conditional distribution (Table 3.5). Substituting $H(X|Y) = 1.015$ into the weak inequality yields $P_e \geq [H(X|Y) - 1]/\log(K) = 0.009$, so as expected the additional information decreases the probability of error. Finally, note that if all probabilities are uniform, $H(P_e) + P_e \log(K - 1) = H(X|Y)$.

[12] In the more mathematical and information theoretic literature "K" is often called the number of elements in the alphabet of X, where "alphabet" is the set of all outcomes of X. For example, let X be a discrete random variable with alphabet $\aleph$, then $|\aleph| = K$ and $|\cdot|$ represents the number of elements in a set.

The above example shows that entropy can be used to calculate the bound on the error of the estimates, and that this error decreases as additional information becomes available.

Example (*A Simple Derivation of Fano's Inequality*). Consider a "general" die-like (unconditional) problem. For a discrete random variable Y, $k = 1,2,3,\ldots,K$, let $\text{Prob}(Y = k) = p_k$ where $p_1 \geq p_2 \geq p_3 \geq \cdots \geq p_K$. Since $Y = 1$ has the highest probability, then the minimal error probability (P_e) must be associated with the estimator $\hat{Y} = 1$, so $P_e = 1 - p_1$. We want to calculate the bound on P_e (Fano's inequality). One way to proceed is to maximize $H(\boldsymbol{p})$ subject to the P_e constraint. This is actually the ME principle discussed in Section 4. This maximization will provide a relationship between P_e and the entropy H. Specifically,

$$\begin{aligned} H(\boldsymbol{p}) &= -\sum_{k=1}^{K} p_k \log p_k = -p_1 \log p_1 - \sum_{k=2}^{K} p_k \log p_k \\ &= -p_1 \log p_1 - \sum_{k=2}^{K} P_e \frac{p_k}{P_e} \log \frac{p_k}{P_e} - P_e \log P_e \\ &= P_e H\left(\frac{p_2}{P_e}, \frac{p_3}{P_e}, \frac{p_4}{P_e}, \ldots\right) + H(P_e) \\ &\leq H(P_e) + P_e \log(K-1), \end{aligned}$$

where the last inequality holds because the maximal value of $H\left(\frac{p_2}{P_e}, \frac{p_3}{P_e}, \frac{p_4}{P_e}, \ldots\right)$ is achieved for the uniform distribution. Thus, any data Y that can be estimated with a probability of error P_e must satisfy the condition $H(Y) \leq H(P_e) + P_e \log(K-1)$ and a lower bound (weaker condition) of

$$P_e \geq \frac{H(Y) - 1}{\log(K-1)}.$$

Numerically, for an unbalanced die Y ($K = 6$ and $Y = 1,2,3,4,5,6$) let $p(Y) = 0.323, 0.258, 0.194, 0.129, 0.065, 0.032$. Then, $P_e = 1 - p_1 = 0.677$, $H(\boldsymbol{p}) = 2.286$, and $H(P_e) = 0.907$. Substituting these quantities

into the above equations yields

$$\begin{aligned} H(\boldsymbol{p}) &= 2.286 = 0.527 + 1.378 + 0.381 \\ &= 1.379 + 0.907 \\ &\leq 0.907 + 1.573 = 2.480 \end{aligned}$$

and the lower bound for the error probability is

$$P_e = 0.677 \geq \frac{H(Y) - 1}{\log(K-1)} = \frac{2.286 - 1}{2.3219} = 0.554.$$

Finally, it is noted that this example can be easily extended to the conditional case.

3.7 Asymptotics

For iid random variables $X_i,\ i = 1, 2, \ldots, n$ with $p(X_1, X_2, \ldots, X_n)$ representing the probability of observing the sequence $x_1, x_2, \ldots, x_n$, the Asymptotic Equipartition Property, AEP (which is the IT variation of the Law of Large Numbers) states that the probability $p(X_1, X_2, \ldots, X_n)$ of the observed sequence (data) is approximately 2^{-nH} where H is the entropy. More technically, for iid $X_1, X_2, \ldots, X_n$,

$$-\frac{1}{n}\log_2 p(X_1, X_2, \ldots, X_n) \overset{\Pr}{\to} H(X),$$

where $\overset{\Pr}{\to}$ stands for convergence in probability. Note that

$$\begin{aligned} -\frac{1}{n}\log_2 p(\cdot) \to H \Rightarrow -\frac{1}{n}\log_2 p(\cdot) &\cong H \\ \Rightarrow -\log_2 p(\cdot) = nH &\Rightarrow p(\cdot) = 2^{-nH}. \end{aligned}$$

Example. Consider flipping n identical coins each with a probability θ of landing "Heads." Let X be one if "Head" is up, and zero otherwise. After n flips the values $X_1, X_2, \ldots, X_n$ are observed. The probability of this specific sequence is just $p(X_1, X_2, \ldots, X_n) = p(x_1)p(x_2)\cdots p(x_n)$. If for example, $\theta = 0.5$, the coin is flipped 10 times, and 6 heads are observed, then, using the above, $p(X_1, X_2, \ldots, X_n) = 2^{-nH} = 0.00098$ (recall that $H(1/2, 1/2) = 1$). If, on the other hand, $\theta = 0.7$, and 8

"Heads" out of $n = 10$ flips are observed, the calculations show that $p(X_1, X_2, \ldots, X_n) = 0.005 \neq 2^{-nH} = 0.002$ where $H = 0.881$.

More generally, assume the sequence $x_1, x_2, \ldots, x_n$, generated from a discrete random variable $X \in \{0, 1\}$, say a coin, with the probability mass function $p(1) = \theta$ and $p(0) = 1 - \theta$, is observed. The probability of the observed sequence is $\prod_{i=1}^{n} p(x_i) = \theta^{\sum x_i (1-\theta)(n-\sum x_i)}$. For example, if the observed sample is (1, 1, 1, 0, 1), then its probability is $\theta^4(1 - \theta)$. But there are 2^n sequences of length n and, as clearly seen in the example, not all of these sequences have the same probability. However, based on the above it can be shown that the number of 1's in that sequence is close to $n\theta$ where all such sequences have (approximately) the same probability of $2^{-nH(p)} = 2^{-5H(p)} = 2^{-5(0.722)} = 0.082$.

It is possible to relate the above to the notion of a *typical set*. Let X_i, $i = 1, 2, \ldots, n$ be iid random variable drawn from some distribution $p(x)$ with alphabet $\aleph$ (with $|\aleph|$ possible realizations). All of the possible sequences $(X_1, X_2, \ldots, X_n)$ can be divided into two distinct sets: the *typical set* and the *non-typical set*. The typical set has 2^{nH} elements with a probability of *almost* one. The elements in that set are (practically) equally probable. Specifically, a typical set is the set of sequences $(x_1, x_2, \ldots, x_n) \in \aleph^n$ that satisfies $2^{-n(H+\varepsilon)} \leq p(x_1, x_2, \ldots, x_n) \leq 2^{-n(H-\varepsilon)}$ for any $\varepsilon > 0$. If a sequence is in the typical set, its probability is equal to $2^{-n(H \pm \varepsilon)}$. The number of elements in that set is no more than $2^{n(H+\varepsilon)}$. For our purposes what important is the following. Analyzing a sample of data, one can divide the set of all possible sequences into two basic sets: the *typical* one and the *non-typical* one. In the typical set, the sample's entropy is close (with high probability) to the true (but unknown) entropy. This is not the same for the non-typical set. The significance of this is that any property that holds for the typical set (or typical sequence) is true with a probability close to one. *Therefore, any property (value) of the typical set is fully consistent with the mean property (value) of a large sample.* Loosely speaking, the moments of the typical set in the observed sample are practically the same as the (unobserved) moments of the underlying population. For a more basic discussion, within the basics of information theory, see for example Cover and Thomas (1991). For discussion of typical sets and large deviations in econometrics see Kitamura and

Stutzer (2002), (Stutzer, 2003a,c), Kitamura (2006) and Kitamura and Otsu (2005).

Example (*AEP and the Relative Entropy* $D(p||q)$). Let X_i, $i = 1,2,\ldots,n$ be iid, discrete random variable drawn from some distribution $p(x)$ and $x \in \{1,2,3,\ldots,m\}$. (Recall that $p(X_1,\ldots,X_K) = \prod_{k=1}^{K} p(X_k)$ and $\frac{1}{K}p(X_1,\ldots,X_K) \to H(X)$.) Let $q(x)$ be another probability (say, prior probability) mass function on the same support $\{1,2,3,\ldots,m\}$. Then,

$$\begin{aligned}\lim -\frac{1}{n}\log_2 q(X_1,X_2,\ldots,X_n) &= \lim -\frac{1}{n}\sum_i \log_2 q(X_i) \to -E[\log_2 q(X)] \\ &= -\sum_i p(x)\log_2 q(x) = D(\boldsymbol{p}||\boldsymbol{q}) + H(\boldsymbol{p}),\end{aligned}$$

and the limit of the log-likelihood ratio $\frac{1}{n}\log\frac{p(X_1,\ldots,X_n)}{q(X_1,\ldots,X_n)}$ is

$$\begin{aligned}\lim -\frac{1}{n}\log_2 \frac{q(X_1,X_2,\ldots,X_n)}{p(X_1,X_2,\ldots,X_n)} &= \lim -\frac{1}{n}\sum \log_2 \frac{q(X_i)}{p(X_i)} \\ &\to -E\left[\log_2 \frac{q(X)}{p(X)}\right] \\ &= -\sum_i p(x)\log_2 \frac{q(x)}{p(x)} = D(\boldsymbol{p}||\boldsymbol{q}).\end{aligned}$$

The AEP is quite a powerful law for analyzing subsets of typical sequences for discrete random variables. A more powerful method (especially for econometricians) that allows us to consider sequences with the same *empirical distribution* is the *method of types*. With this method, the bounds on the number of such sequences (or a particular empirical distribution) and the probability of each sequence in that set can be calculated. The basic idea, its relationship to the theory of Large Deviations and econometrics are briefly discussed below.[13] For

[13] Study of the theory of typical sets allows us also to relate the Fisher's information and entropy even though Fisher's information matrix is defined for a family of parametric distributions, while the entropy is defined for all distributions. This comparison is done after reparameterizing any distribution, say $f(x)$ by some location parameter β and then redefining Fisher's information for the family of densities $f(x-\beta)$. Cover and Thomas show that the entropy relates to the volume of the typical set, while Fisher information is related to its surface.

detailed derivations within IT see Cover and Thomas (1991), the original work of Csiszar (1984, 1991), Sanov (1961) and the nice text of Dembo and Zeitouni (1998). For early applications in econometrics see Stutzer (2000, 2003a,b,c), Kitamura and Stutzer (2002) and Kitamura (2006).

Let $X_1, X_2, \ldots, X_n$ be a sequence of n symbols (or n observed values) from the alphabet $\aleph$ with $|\aleph|$ elements. The *empirical probability* distribution, or type, P_x of the sequence $x_1, x_2, \ldots, x_n$ is the relative proportion of occurrences of each one of the symbols of $\aleph$. Next, define P_n to be the set of types with the denominator n.

Example. Let $\aleph = \{0, 1\}$, then the set of all possible types with denominator $n = 3$ is

$$P_{n=3} = \left\{ (P(0), P(1)) : \left(\frac{0}{3}, \frac{3}{3}\right), \left(\frac{1}{3}, \frac{2}{3}\right), \left(\frac{2}{3}, \frac{1}{3}\right), \left(\frac{3}{3}, \frac{0}{3}\right) \right\}.$$

The type class of P is $T(P) = \{(x_1, x_2, \ldots, x_n) \in \aleph^n | P_x = P\}$. $T(P)$ is the set of all sequences with length n and type P_n.

Example. Continuing with above example, but let $n = 4$ and the observed sequence (sample) be $x = 0010$. The type P_x is $P_x(0) = 3/4$ and $P_x(1) = 1/4$. The type class of P_x is the set of all sequences of length $n = 4$, with three zero's and a single one, so $T(P_x) = \{0001, \ldots, 1000\}$, and the number of elements in $T(P)$ is $|T(P)| = \binom{4}{3,1} = \frac{4!}{3!1!} = 4$.

The method of types' basic theorem shows that the number of types is at most polynomial in n: $|P_n| \leq (n+1)^{|\aleph|}$. Recalling that the number of sequences is exponential in n, means that at least one type has exponentially many sequences in its type class. Further, as noted before for the AEP, the method of types is also related to the entropy and the relative entropy measures. To see that relationship, assume $X_1, X_2, \ldots, X_n$ are iid according to some true distribution $Q(x)$. The probability of the sequence $x_1, x_2, \ldots, x_n$ is $Q^n(x_1, x_2, \ldots, x_n) = 2^{-n(H(P_x)+D(P_x\|Q))}$ which is a function of its type. Naturally, if the sequence $x_1, x_2, \ldots, x_n$ is in the type class of Q, the above reduces to $Q^n(x_1, x_2, \ldots, x_n) = 2^{-nH(Q)}$.

Example. Let X be a discrete random variables with possible realizations 1, 2, 3, and 4, with the corresponding probabilities 4/8, 2/8,

1/8, and 1/8, respectively. The probability of observing a sequence with these frequencies is $2^{-nH\left(\frac{1}{2},\frac{1}{4},\frac{1}{8},\frac{1}{8}\right)}$ for n being any multiple of 8.

The lower and upper bounds on the size of the type class of P, $T(P)$ is

$$\frac{1}{(n+1)^{|\aleph|}}2^{nH(P)} \leq |T(P)| \leq 2^{nH(P)}.$$

Example. Consider the binary case ($\aleph \in \{0,1\}$). In this simple case, the type is fully defined by the number of one's (or zero's) in the sequence. The size of the type class is just $\binom{n}{k}$ and the bounds are $\frac{1}{(n+1)}2^{nH\left(\frac{k}{n}\right)} \leq \binom{n}{k} \leq 2^{nH\left(\frac{k}{n}\right)}$.

The basic theorems on types show (i) that there is only a polynomial number of types, while there are an exponential number of sequences of each type, (ii) the exact formula for the probability of any sequence of type P under some distribution Q, and (iii) the approximation for the probability of a type class. These theorems allow us to evaluate the behavior of long sequences (large data) based on the properties of the type of the sequence. For example, from (i) above it is possible to conclude that since the probability of each type depends (exponentially) on the relative entropy between the type P and the distribution Q, type classes that are far away from the true (unknown) distribution have exponentially smaller probability (the probability of a typical set goes to one as n goes to infinity). The last statement can be formalized as follows. Let $X_1, X_2, \ldots, X_n$ be iid and distributed according to $P(x)$ with a finite alphabet $\aleph$. It can be shown that

$$\text{Prob}\{D(P_{x_1,x_2,\ldots,x_n}\|P) > \varepsilon\} \leq 2^{-n\left(\varepsilon - |\aleph|\frac{\log_2(n+1)}{n}\right)}.$$

Thus, $D(P_{x_1,x_2,\ldots,x_n}\|P) \to 0$ with probability one. A stronger version of typicality exists but is not discussed here. The above takes us to the theory of Large Deviations. This theory and its implications, though at the heart of IT, are outside the scope of that review. But for completeness, I provide a brief discussion here and then point toward its econometric implications.

The theory of Large Deviations (LD) deals with estimating the probability of an infrequent event — an event with an *a priori* small probability. Consider the iid, discrete random variable X that has a Bernoulli distribution: $f(x;\theta) = \theta^x(1-\theta)^{(1-x)}$ for $x = 0, 1$. Let θ be 0.4. The theory of LD deals with questions like: "What is the probability that the expected value of the observed random variable (sequence) is greater than 0.9?" This question deals with a "large deviation" between the true mean of the distribution and the observed (sample) mean. In terms of the typical sequences discussed above, an expected value of 0.9 is just $P_{x_1,x_2} = (0.1, 0.9)$. Thus, the probability that the sample's mean is close to 0.9 is exactly the probability of that type. The probability of that large deviation, p_{LD}, is a function of the relative entropy and the sample size. In our example, p_{LD}, which cannot be approximated well enough using the central limit theorem, is approximately $2^{-nD(0.1,0.9\|0.6,0.4)}$. For $n = 5$, 10, and 100, $p_{\mathrm{LD}} = 0.064$, 0.004, and $1.2E$-24, respectively. The above is expressed in the famous Sanov's theorem (Sanov, 1961) which is sketched below.

Let E be a certain subset of the set of proper probability distributions ($E \subseteq P$). Let $X_1, X_2, \ldots, X_n$ be iid generated by the distribution $Q(x)$ with the set of possible outcomes (alphabet) $\aleph$. Then, $Q_n(E) = Q_n(E \cap P_n) \leq (n+1)^{|\aleph|} 2^{-nD(P^*\|Q)}$ for $P^* = \arg\min_{P \in E} D(P\|Q)$. Note that P^* is the closest distribution (in entropy terms) to Q in E. If the set E is the closure of its interior, then $\frac{1}{n}\log_2 Q_n(E) \to -D(P^*\|Q)$.

The relationship between this theorem and the principle of Maximum Entropy is discussed in the next section. But, for now to see the depth of that theorem imagine that you observe a sample of data, and you are asked to verify whether a certain set of conditions (say, the observed sample's moments) is consistent with your prior beliefs (the distribution $Q(x)$). What this theorem tells us is that this probability is found by minimizing the relative entropy $D(\cdot\|\cdot)$ subject to these constraints (e.g., observed sample moments). Using the Lagrangean approach yields the desired (exponential) probability distribution P^*. Specific examples are presented in Section 4. Applications to hypothesis testing and inference as well as to the likelihood ratio statistic will be discussed in Sections 6 and 7 (see also Kitamura, 2006).

3.8 Stochastic Process and Entropy Rate

Consider now a more realistic process involving dependent random variables that form a stationary process. To capture the main quantities and to relate it to familiar econometric problems, consider a first-order Markov Chain that is stationary and for simplicity is also time-invariant.

Let X be a discrete random variable with alphabet $\aleph$ and a probability mass function $p(x) = \Pr\{X = x\}$, $x \in \aleph$. Let the stochastic process $X_1, X_2, \ldots, X_T$ ($t = 1, 2, \ldots, T$ is a discrete time index) be such that $\Pr(X_{t+1} = x_{t+1} | X_t = x_t)$ for all $x_1, x_2, \ldots, x_T \in \aleph$. This process can be specified as

$$\alpha_{x_{t+1}} = \sum_{x_t} \alpha_{x_t} P_{x_t x_{t+1}} \quad \text{with} \sum_{x_t} \alpha_{x_t} = 1, \tag{3.12a}$$

where P is the stationary first-order Markov probability matrix. More compactly, this process can be specified as

$$\alpha_k(t+1) = \sum_j \alpha_j(t) P_{jk} \quad \text{with} \sum_j \alpha_j = 1 \quad \text{and} \quad \sum_k P_{jk} = 1. \tag{3.12b}$$

We will go back to this process in Section 7.

We already know that the entropy reflects the average amount of information in a random variable, or data. With a stochastic process, one wants to measure the change, or incremental increase, of (average) information with the process, or how does the entropy of the sequence increases with t. Example include measuring the increase in information due to additional period of data collected from the same process as the current data, or trying to better understand the process behind an evolving economic (time series) data. The *Entropy Rate* captures that increase.

The *Entropy Rate* of some stochastic process $\{X_t\}$ is

$$H(\aleph) \equiv \lim_{T \to \infty} \frac{1}{T} H(X_1, X_2, \ldots, X_T)$$

when the limit exists. $H(\aleph)$ reflects the entropy per symbol of the T random variables. A related (conditional) quantity is

$$H^*(\aleph) \equiv \lim_{T \to \infty} H(X_T | X_{T-1}, X_{T-2}, \ldots, X_1)$$

when the limit exists. $H^*(\aleph)$ is the *conditional entropy* — the entropy of the last random variable conditional on the past $T-1$ values. If the stochastic process is stationary, then $H(\aleph) = H^*(\aleph)$. The entropy rate is the expected number of bits per symbol necessary to describe the process. Going back to the first-order, *stationary* Markov process, these two measures are related through

$$H(\aleph) = H^*(\aleph) = H(X_2|X_1) = -\sum_{jk} \alpha_j P_{jk} \log P_{jk}$$
$$= \sum_k \alpha_k \Big[-\sum_j P_{jk} \log P_{jk} \Big].$$

Example. Consider a three-state, stationary, Markov process. Let the stationary (and known) distribution α be $\alpha = (0.5, 0.3, 0.2)'$. Using (3.12a) or (3.12b), this stationary process is characterized by

$$P_{jk} = \begin{bmatrix} 0.553 & 0.284 & 0.163 \\ 0.465 & 0.312 & 0.223 \\ 0.420 & 0.322 & 0.258 \end{bmatrix} = \begin{bmatrix} \boldsymbol{p}_1 \\ \boldsymbol{p}_2 \\ \boldsymbol{p}_3 \end{bmatrix} \quad \text{where} \quad \begin{bmatrix} H(\boldsymbol{p}_1) = 1.415 \\ H(\boldsymbol{p}_2) = 1.521 \\ H(\mathbf{p}_3) = 1.556 \end{bmatrix}.$$

The entropy of state X at period t is $H(X_t) = H(\alpha) = 1.485$, while $H(\aleph) = 1.475$. Note that $H(P) = 4.492$, while $H(\boldsymbol{p}_j)$ equals 1.415, 1.521, 1.556, respectively for $j = 1, 2, 3$. If, on the other hand, all probabilities are uniform ($\alpha = (1/3, 1/3, 1/3)'$ so $p_{jk} = 1/3$ for all j and k) then $H(\aleph) = 1.585$, $H(\boldsymbol{p}_j) = 1.585$ for each $j = 1, 2, 3$ and $H(P) = 4.755$. As expected, all entropies of this (more uncertain) case are higher. It is harder to describe that process that can be thought of as a more "noisy" process.

3.9 Continuous Random Variables

A brief discussion of the differential entropy, defined for a continuous random variable, is now provided. Let X be a continuous random variable with probability distribution P and density $f(x)$ with respect to some dominating measure μ. The entropy is

$$H(P) \equiv \int f(x) \log \frac{1}{f(x)} d\mu(x), \tag{3.13}$$

where this differential entropy does not have all of the properties of the discrete entropy (3.1) and there are some basic philosophical questions regarding that quantity. However, this measure can be generalized to the "cross-entropy" equivalent of the discrete case. Let Q be another probability distribution with density $q(x)$ with respect to the same measure μ. Then,

$$D(P||Q) \equiv \int f(x) \log \frac{f(x)}{q(x)} d\mu(x)$$

and it measures the information discrepancy between the two distributions P and Q.

Example. The differential entropy for X_i $(i = 1, \ldots, n)$ that are multivariate normal with mean θ and covariance matrix Σ is

$$\begin{aligned} & H(X_1, \ldots, X_T) \\ & \quad = H(\textit{multivariate normal}) \\ & \quad = -\int f(\boldsymbol{x}) \left[-\frac{1}{2} (\boldsymbol{x} - \theta)' \Sigma^{-1} (\boldsymbol{x} - \theta) - \log(\sqrt{2\pi})^T |\Sigma|^{1/2} \right] d\boldsymbol{x} \\ & \quad = \frac{1}{2} \log(2\pi e)^T |\Sigma|, \end{aligned}$$

where $f(\boldsymbol{x})$ is the probability density function of $X_1, \ldots, X_T$, "$|\cdot|$" stands for the determinant and the natural log is used here.

Similarly the Shannon *mutual information* between two random variables can be specified as the relative entropy between their joint distribution and the product of their marginal distributions. For the random variables X and Y with joint distribution P_{XY} and marginal distributions P_X and P_Y with densities p_{XY}, p_{X}, and p_{Y}, respectively, the Shannon mutual information is

$$I(X;Y) = D(P_{XY}||P_X \times P_Y).$$

This measure can naturally be extended as a conditional measure with respect to another variable. In Sections 4 and 5 the differential entropy is used for specific estimation models. However, since most models within IEE use the discrete entropy, not much background is provided here but the reader is referred to some nice and detailed discussions of

the continuous (differential) entropy and related measures (Soofi, 1994, 2000; Maasoumi and Racine, 2005; and Clarke, 2007). For use of the differential entropy for stochastic process and the extension of quantities like the differential entropy rate and spectral estimation (with the normal process), see Burg (1975) and Cover and Thomas (1991, Chap. 11).[14]

[14] A number of tests using entropy exist and are not fully covered in this review. These tests include a test for normality based on the sample's entropy (Vasicek, 1976), and testing of fit for exponentiality, based on the cross entropy (divergence) measure $D(\cdot||\cdot)$ is introduced by Ebrahimi et al. (1992). See these studies for details.

4

The Classical Maximum Entropy Principle

4.1 The Basic Problem and Solution

Facing the fundamental question of drawing inferences from limited and insufficient data, Jaynes proposed the ME principle, which he viewed as a generalization of Bernoulli and Laplace's Principle of Insufficient Reason. The ME formalism was introduced by Jaynes (1957a,b, 1963) and is based on Shannon's entropy measure. This formalism was extended and applied by a large number of researchers (e.g. Levine, 1980; Levine and Tribus, 1979; Tikochinsky et al., 1984; Skilling, 1989a; Hanson and Silver, 1996; Golan et al., 1996b). Axiomatic foundations for this approach were developed by Shore and Johnson (1980); Skilling (1989b) and Csiszar (1991). See also Jaynes (1984) and his nice text (Jaynes, 2003) for additional discussion.

In more recent work within econometrics and statistics, and facing similar problems of how to estimate the unknown distribution of a random variable with minimal assumptions on the underlying likelihood function, methods similar in philosophy to the ME principle are developed. These methods include the Empirical Likelihood, the Generalized EL, versions of the Generalized Method of Moments as well as the Bayesian method of Moments. In these methods the Shannon's

entropy measure is substituted for other measures within the class of α-entropies or for the differential entropy. These methods are discussed in the next section. Using the tools of the calculus of variations the classical ME is summarized below.[1]

Suppose the first T (zero or "pure") moment conditions of an unknown K-dimensional ($K > T + 1$), proper probability distribution $\boldsymbol{p}$ corresponding to the K-dimensional, discrete random variable, are observed. In the linear case, these T pure moments are $y_t = \sum_k x_{tk} p_k$, where X is a $T \times K$ design matrix and $\boldsymbol{y}$ are the observed (zero) sample moments. Similarly, these moments can be expressed as

$$\boldsymbol{y} = X\boldsymbol{p}, \quad \text{or} \quad \boldsymbol{y} - X\boldsymbol{p} = 0.$$

A number of examples help clarify these notations. Consider the case where the researcher only knows the sample's mean ($\bar{X}$) and variance ($\hat{\sigma}^2$) of a single discrete random variable X with alphabet $\aleph$ and $|\aleph| = K$ (i.e., a K-*dimensional* discrete random variable). Though these moments are based on a sample of N observations, the researcher does not have that sample. In that case $T = 2$, and let $x_{1k} = x_k$ and $x_{2k} = (x_k - \bar{X})^2$, so $y_1 = \bar{X} = \sum_k x_{1k} p_k = \sum_k x_k p_k$ and $y_2 = \hat{\sigma}^2 = \sum_k x_{2k} p_k = \sum_k (x_k - \bar{X})^2 p_k$. Economic examples and applications of such cases include the estimation of income distribution or size distribution of firms based on observed first moments, optimal portfolio and asset choices based on observed assets' means as well as estimating Input–Output coefficients, Social Accounting Matrices or Markov transition probabilities from observed aggregated data. Consider another example where the researcher observes all of the N observations in the above sample and, rather than form a likelihood function, wishes to estimate the natural weight of each observation p_i ($i = 1, 2, \ldots, N$) in that sample. In that case $y_1 = \bar{X} = \sum_{i=1}^{N} x_i p_i$ and $y_2 = \hat{\sigma}^2 = \sum_{i=1}^{N} (x_i - \bar{X})^2 p_i$. This case is discussed further in Section 5. The ME principle is now presented.

Given the T (pure) moments $\boldsymbol{y}$, our objective is to estimate the K-dimensional, unknown distribution $\boldsymbol{p}$ for the case where $K > T + 1$ (an under-determined problem). Using here the natural log (rather

[1] ME is a standard variational problem.

than $\log_2$) the ME formulation is

$$\text{ME} = \begin{cases} \hat{\boldsymbol{p}} = \arg\max\{H(\boldsymbol{p}) \equiv -\sum_k p_k \log p_k\} \\ \text{s.t. } \boldsymbol{y} - X\boldsymbol{p} = 0; \quad \sum_k p_k = 1 \end{cases}. \tag{4.1}$$

Similarly, the cross-entropy (CE) formulation when prior information $\boldsymbol{q}$ is available is

$$\text{CE} = \begin{cases} \tilde{\boldsymbol{p}} = \arg\min\{D(\boldsymbol{p}||\boldsymbol{q}) \equiv \sum_k p_k \log(p_k/q_k)\} \\ \text{s.t. } \boldsymbol{y} - X\boldsymbol{p} = \boldsymbol{0}; \quad \sum_k p_k = 1 \end{cases}. \tag{4.2}$$

Note that under the ME formulation one maximizes the entropy while under the CE formulation the entropy difference between $\boldsymbol{p}$ and $\boldsymbol{q}$ is minimized. The two procedures are similar in the sense that the ME (4.1) can be viewed as the CE (4.2) with uniform priors ($\boldsymbol{q}$'s).

The CE solution is

$$\tilde{\boldsymbol{p}}_k = \frac{q_k \exp\left(\sum_{t=1}^T \tilde{\lambda}_t x_{tk}\right)}{\sum_k q_k \exp\left(\sum_{t=1}^T \tilde{\lambda}_t x_{tk}\right)} \equiv \frac{q_k \exp\left(\sum_{t=1}^T \tilde{\lambda}_t x_{tk}\right)}{\Omega}, \tag{4.3}$$

where $\Omega(\tilde{\boldsymbol{\lambda}}) \equiv \sum_k q_k \exp\left(\sum_{t=1}^T \tilde{\lambda}_t x_{tk}\right)$ is a normalization factor known also as the partition function and $\tilde{\boldsymbol{\lambda}}$ is the T-dimensional vector of estimated Lagrange multipliers. If $\tilde{\boldsymbol{p}}$ is the solution to such an optimization problem, it can be shown that $I(\boldsymbol{p};\boldsymbol{q}) = I(\boldsymbol{p};\tilde{\boldsymbol{p}}) + I(\tilde{\boldsymbol{p}};\boldsymbol{p})$ for any $\boldsymbol{p}$ satisfying the set of constraints in (4.1) or (4.2). This is the analogous to the Pythagorean Theorem in Euclidean geometry, where $I(\boldsymbol{p};\boldsymbol{q})$ can be regarded as the analog for the squared Euclidean distance.

4.2 Duality — The Concentrated Model

The ME and CE formulations above are constructed in terms of a constrained optimization (call it the primal model) where the optimization is carried with respect to the $\boldsymbol{p}$'s. However, it is possible to construct both the ME and CE as an unconstrained dual model, which is equivalent to a concentrated likelihood function. The advantages of the dual formulation is that first, an unconstrained model is simpler

(and computationally superior), second by moving from the probability space to the Lagrange multipliers' space the dimension of the model decreases significantly (recall K is much greater than $T+1$), and third, that formulation allows a direct comparison with the more traditional likelihood methods.

To derive the dual formulation (concentrated model), one starts by constructing the Lagrangean for model (4.2) but without the proper probability requirement $\left(\sum_k p_k = 1\right)$. The CE solution (4.3), that already satisfies the proper probability requirement, is then inserted into the first right-hand side term of $\ell(\boldsymbol{\lambda})$ that yields the concentrated CE model

$$\begin{aligned}\ell(\boldsymbol{\lambda}) &= \sum_k p_k \log(p_k/q_k) + \sum_t \lambda_t \left[y_t - \sum_k x_{tk} p_k \right] \\ &= \sum_k p_k(\boldsymbol{\lambda}) \left[\sum_t \lambda_t x_{tk} - \log(\Omega(\boldsymbol{\lambda})) \right] + \sum_t \lambda_t \left[y_t - \sum_k x_{tk} p_k \right] \\ &= \sum_t \lambda_t y_t - \log(\Omega(\boldsymbol{\lambda})). \end{aligned} \tag{4.4}$$

In a more condensed notations, the primal–dual relationship is

$$\begin{aligned}\underset{\boldsymbol{p}\in P}{\operatorname{Min}} I(\boldsymbol{p},\boldsymbol{q}) &= \underset{\boldsymbol{\lambda}\in D}{\operatorname{Max}} \{\boldsymbol{\lambda}'\boldsymbol{y} - \log\Omega(\boldsymbol{\lambda})\} \\ &= \underset{\boldsymbol{\lambda}\in D}{\operatorname{Max}} \left\{ \sum_t y_t \lambda_t - \log\left[\sum_k q_k \exp\left(\sum_{t=1}^{T} \lambda_t x_{tk} \right) \right] \right\}. \end{aligned} \tag{4.5}$$

Looking at (4.5) it is quite clear that it has the same form of a likelihood function. In fact, under *uniform priors* $(\boldsymbol{q})$, it is equivalent to the ML Logit for a discrete choice model. Detailed comparison is done in Section 6.

In a slightly more general notation let the moments be represented by $\sum_k p_k g_t(X) = E[g_t]$ where X is a $T \times K$ matrix. Then, the relationship between the Lagrange multipliers and the data is easily seen from

$$-\frac{\partial \log\Omega}{\partial \lambda_t} = E[g_t] \tag{4.6}$$

while the higher moments are captured via

$$\frac{\partial^2 \log \Omega}{\partial \lambda_t^2} = \mathrm{Var}(g_t) \quad \text{and} \quad \frac{\partial^2 \log \Omega}{\partial \lambda_t \partial \lambda_s} = \mathrm{Cov}(g_t g_s). \tag{4.7}$$

The formulations above are for the discrete case. A detailed example of that case is provided at the end of this section (Section 4.6). To show that above principle can be used for continuous random variables, a few examples are provided below.

Example (*Single Constraint: Normalization*). Maximizing the differential entropy subject to normalization $\left(\int f(x)dx = 1\right)$ yields

$$\hat{f}(x) = \exp(-1 - \hat{\mu}),$$

where $\hat{\mu}$ is the Lagrange multiplier associated with the only constraint. If X is confined to some range (a,b) then $f(x) = 1/(b-a)$ is uniformly distributed with mean $(a+b)/2$ and variance $(b-a)^2/12$, $\mu = \log(b-a) - 1$ and $H(\boldsymbol{x}) = -\log(b-a)$. This example shows that the uniform distribution is the ME density function resulting from maximizing the differential entropy subject to one condition — the requirement that the density function is proper (normalization). This is consistent with the basic philosophy behind the ME principle (and the principle of insufficient reason) where the maximal level of entropy is associated with the uniform distribution — a state of maximal uncertainty.

Example (*Two Constraints: Normalization and First Moment*). For a nonnegative random variable X, if one knows that the expectation is $\int xf(x)dx = m$, then the ME method yields $\hat{f}(x) = (1/m)\exp(-x/m)$ which is the exponential distribution with mean m, variance m^2, and $H = 1 + \log(m)$.

Example (*Three Constraints: Normalization and Two First Moments*). Let a random variable X be distributed over the whole real line with the two first moments $\int xf(x)dx = m$ and $\int (x-m)^2 f(x)dx = \sigma^2$. For simplicity, let $m = 0$. Maximizing the differential entropy subject to these constraints (and as usual the proper density requirement) yields (for the Lagrange multipliers $\mu, \lambda_1, \lambda_2$),

$$\hat{f}(x) = \exp(-1 - \hat{\mu} - \hat{\lambda}_1 x - \hat{\lambda}_2 x^2) = \sqrt{1/2\pi\sigma^2}\exp(-x^2/2\sigma^2),$$

where $\hat{\mu} = (1/2)\log(2\pi\sigma^2) - 1$, $\hat{\lambda}_1 = 0$, $\hat{\lambda}_2 = 1/2\sigma^2$ and $H = (1/2)\ln(2\pi e\sigma^2)$. Note that this is just the normal distribution. Thus, the normal density function, is the ME density function resulting from maximizing the entropy subject to the first two moments. This provides an additional motivation for using the normal density function in applied work. This example is revisited in Section 5.5 dealing with Zellner's Bayesian method of moments.

Finally, a short discussion on hypothesis tests and entropy is provided. A more detailed discussion and some applications are summarized in the next sections. Based on Eq. (4.1) and/or (4.2) hypothesis tests can be constructed. The basic idea is briefly summarized below via a simple example. Let $X_1, X_2, \ldots, X_N$ be a sample of size N generated iid from the probability distribution P_0. Within the context of comparing two probability distributions, consider two competing hypotheses: $H_0^1{:}P_0 = P_1$ and $H_0^2{:}P_0 = P_2$. Let α and β be the two probabilities of error: $\alpha = \text{Prob}(P_0 = P_2|H_0^1 \text{ is true})$ and $\beta = \text{Prob}(P_0 = P_1|H_0^2 \text{ is true})$. The likelihood ratio test provides an optimal test (Neyman–Pearson Lemma) for testing these two hypotheses:

$$L \equiv \frac{P_1(X_1, \ldots, X_N)}{P_2(X_1, \ldots, X_N)} > K \quad \text{for } K \geq 0.$$

To relate that test to the relative entropy, note that

$$l = \log L = \log \frac{P_1(X_1, \ldots, X_N)}{P_2(X_1, \ldots, X_N)} = N[D(\tilde{P}_{X(N)}||P_2) - D(\tilde{P}_{X(N)}||P_1)],$$

and

$$\begin{aligned} L \;&\equiv\; \frac{P_1(X_1, \ldots, X_N)}{P_2(X_1, \ldots, X_N)} > K \\ &\Leftrightarrow [D(\tilde{P}_{X(N)}||P_2) - D(\tilde{P}_{X(N)}||P_1)] > \frac{1}{N}\log K. \end{aligned}$$

Building on the above, the relationship between the likelihood ratio test and the relative entropy leads to the entropy ratio test formulated and discussed later. Further, loosely speaking, the above relationship, is related to Stein's Lemma (see for example Cover and Thomas, 1991) stating that, for the above hypotheses and probability of errors $\alpha(N)$

and $\beta(N)$, both as functions of the sample size N, the following statement holds:

$$\lim_{N\to\infty} \frac{1}{N} \log \beta(N) = -D(P_1||P_2).$$

For example, the relative entropy corresponding to the test whether the sample came from an underlying normal distribution with mean zero[2] but different variances is

$$D(P_1||P_2) = \frac{1}{2}\left[\log\left(\frac{\sigma_2^2}{\sigma_1^2}\right) + \frac{\sigma_1^2}{\sigma_2^2} - 1\right]$$
$$\text{for } P_1 \sim N(0,\sigma_1^2) \quad \text{and } P_2 \sim N(0,\sigma_2^2).$$

As a second (extreme) example consider the hypothesis contrasting P's resulting from a certain Bernoulli distribution, say $P_1 \sim \text{Bernoulli}\left(\theta = \frac{1}{3}\right)$ and $P_2 \sim \text{Bernoulli}(\theta = 1)$. In that case, one gets

$$D(P_1||P_2) = \frac{1}{3}\log\frac{1/3}{1} + \frac{2}{3}\log\frac{2/3}{0} \to \infty.$$

In that case, $D(\cdot) \to \infty$, it can be said that, for large samples, it is possible to distinguish the two probability distributions P_1 and P_2 with probability one.

4.3 The Entropy Concentration Theorem

The Entropy Concentration Theorem, ECT, (Jaynes, 1978a,b) provides another convincing rationale for the ME principle. This theorem states that out of all distributions that satisfy the observed data (moments), a significantly large portion of these distributions are concentrated sufficiently close to the one of maximum entropy. Similarly, the subset of

[2] Recall the simpler case of deriving the entropy of a random variable X that is distributed as $f(x) = (1/\sqrt{2\pi\sigma^2}) \times \exp(-x^2/2\sigma^2)$ and has the entropy

$$H(f) = -\int_{-\infty}^{\infty} f(x)\ln f(x)dx = -\int f(x)\left(-\frac{x^2}{2\sigma^2} - \ln\sqrt{2\pi\sigma^2}\right)dx$$
$$= \frac{E[X^2]}{2\sigma^2} + \frac{1}{2}\ln 2\pi\sigma^2 = \frac{1}{2} + \frac{1}{2}\ln 2\pi\sigma^2.$$

distributions (satisfying the data) that have significantly lower entropy than the maximum are the subset of *atypical* distributions.

More formally, consider a random experiment with K possible states/realizations at each trial, so in N trials there are K^N possible outcomes. The word "state" used here means a realization of a *single* trial, while "outcome" refers to the experiment as a whole. Each outcome yields a set of observations $\{N_k\}$, or frequencies $\{f_k = N_k/N; 1 \leq k \leq K\}$, with an entropy of $H(\boldsymbol{f}) = -\sum_k f_k \log f_k$. Let $\mathscr{P}$ be the subclass of all possible outcomes that could be observed in the N trials and satisfy the $T < K$ linearly independent constraints in (4.1) or (4.2). The ECT states that a high percentage of outcomes in the class $\mathscr{P}$ will have entropy in the range

$$\begin{aligned} H^* - \Delta H \leq H(\boldsymbol{f}) &= H(f_1, \ldots, f_K) \leq H^* \\ &\equiv \underset{p \in \mathscr{P}}{\text{Max}} H\left[\boldsymbol{p} | \boldsymbol{y} - X\boldsymbol{p} = \boldsymbol{0}, \sum p_k = 1\right] \\ &= H(\hat{\boldsymbol{p}}), \end{aligned} \tag{4.8}$$

where $\Delta H \equiv \chi^2_{(K-T-1;\alpha)}/2N$ and α is the upper α percentile of the χ^2 distribution with $(K - T - 1)$ degrees of freedom. Other distributions $\{f_k^o\}$ that are consistent with the constraints (sample data) will have entropy levels smaller than H^*. Their concentration near this upper bound is given by the above ECT.

This theorem tells us that asymptotically, $2N\Delta H$ is distributed over the class $\mathscr{P}$ as a $\chi^2_{(K-T-1)}$ independently of the structure of the T constraints. Hence, approximately $(1 - \alpha)100\%$ of the frequencies satisfying the observed data/constraints, have entropy within the range specified by (4.8).

Example. Suppose a die is tossed $N = 100$ times. Though one would expect the observed mean to be approximately 3.5 (for a fair die), the observed mean is 5. In this case there are two constraints: normalization and the first moment. Within the above notation (Eqs. (4.1) and (4.2), $T = 1$ and $K = 6$ and, $y = 5$. The ME solution $\hat{\boldsymbol{p}}$ is 0.021, 0.038, 0.072, 0.136, 0. 255, and 0.478, respectively for $k = 1, \ldots, 6, H(\hat{\boldsymbol{p}}) = 1.973$ and $\hat{\lambda} = 0.908$. Applying the CTE (for $\alpha = 0.05$), $\chi^2_{(6-1-1;0.05)} = \chi^2_{(4;0.05)} = 9.488$, so $\Delta H = 0.0474$, and $1.9256 \leq H \leq 1.973$. If, on the other hand,

we have a larger sample of $N = 1000$, then, the CTE yields the more concentrated interval: $1.9683 \leq H \leq 1.973$.

As noted by Jaynes, an important feature of the ECT is that, for example the above 95% concentration range $H^* - \frac{(9.488/2)}{N} \leq H \leq H^*$ is valid asymptotically for *any* random experiment with four degrees of freedom (though naturally H^* will vary based on the experiment/problem analyzed). Consider a higher significance level of $\alpha = 0.005$. For $N = 100$, 99.5% of all outcomes (allowed by the sample's constraints) have entropy in the range of width $\Delta H = (2N)^{-1}\chi^2_{(4;0.005)} = 14.86(2N)^{-1} = 0.0743$, so $1.899 \leq H \leq 1.973$, while for $N = 1000, 1.966 \leq H \leq 1.973$. The ECT provides a compelling argument in favor of using the ME principle. It shows that for large N, the overwhelming majority of all distributions consistent with our limited information (the T constraints) have entropy value very close to the maximum. The width (ΔH) of the concentration region decreases at a fast rate of N^{-1}. The ECT also provides an indirect motivation for using the normal distribution (see example on Section 4.2 showing that the normal density function is the MaxEnt density function resulting from maximizing the differential entropy subject to the first two moments).

4.4 Information Processing Rules and Efficiency

Zellner (1988, 2002) showed that Bayes theorem is an efficient Information Processing Rule (IPR) and it satisfies the information conservation principle (ICP) stating that the total input information equals the total output information. Any IPR that satisfies that principle is defined as efficient in the sense that the ratio of output to input information is one. Any IPR that does not satisfy that rule is inefficient (ratio < 1) or "produces" information (ratio > 1) within the estimation process. Following Zellner's work, it is possible to show that the ME principle is an efficient IPR. To show this, I provide some background below.

Information in the inputs and outputs can be measured in terms of their entropy level. The two inputs are the data density or likelihood function, $F(\boldsymbol{y}, \boldsymbol{\theta}) \equiv L(\boldsymbol{y}, \boldsymbol{\theta})$, and the prior distribution on $\boldsymbol{\theta}$, $q(\boldsymbol{\theta})$. The two outputs are the post-data (or posterior in the Bayesian context)

distribution of $\boldsymbol{\theta}$, $p(\boldsymbol{\theta}|\boldsymbol{y})$ and the marginal distribution $m(\boldsymbol{y})$ which is the partition function Ω in the ME/CE case. The respective information contained in the inputs and outputs is represented as the expectations, with respect to $p(\boldsymbol{\theta}|\boldsymbol{y})$, of the logarithm of the inputs and outputs. For example, the negative entropy in the data density or likelihood function is given by $\sum_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\boldsymbol{y}) \log F(\boldsymbol{y}, \boldsymbol{\theta})$. With these definitions and background, it is possible to show that the CE is an efficient (or 100% efficient) IPR for any prior.

Let $I_1(\tilde{\boldsymbol{p}}) \equiv -\sum_k \tilde{p}_k \log q_k$ and $I_2(\tilde{\boldsymbol{p}}) \equiv -\sum_k \tilde{p}_k \log[\exp(-\sum_i \tilde{\lambda}_i x_{ik})] = \sum_{k,i} \tilde{p}_k \tilde{\lambda}_i x_{ik}$ be the input information. Similarly, let $O_1(\tilde{\boldsymbol{p}}) \equiv -\sum_k \tilde{p}_k \log \tilde{p}_k$ and $O_2(\tilde{\boldsymbol{p}}) \equiv -\sum_k \tilde{p}_k \log \Omega(\tilde{\boldsymbol{\lambda}}) = -\log \Omega(\tilde{\boldsymbol{\lambda}})$ be the two output information. We want to show that

$$O_1(\tilde{\boldsymbol{p}}) + O_2(\tilde{\boldsymbol{p}}) - [I_1(\tilde{\boldsymbol{p}}) + I_2(\tilde{\boldsymbol{p}})] = 0. \tag{4.9}$$

Substitute the input and output information into Eq. (4.9) yields

$$-\sum_k \tilde{p}_k \log \tilde{p}_k - \log \Omega(\tilde{\boldsymbol{\lambda}}) + \sum_k \tilde{p}_k \log q_k - \sum_{k,i} \tilde{p}_k \tilde{\lambda}_i x_{ik}. \tag{4.9a}$$

The sum of the first and third elements in Eq. (4.9a) is identically equal to the negative of the CE (see Eq. (4.2)). The optimal value of the concentrated CE model (4.4) is

$$\sum_i \tilde{\lambda}_i y_i - \log \Omega(\tilde{\boldsymbol{\lambda}}) \equiv \sum_k \tilde{p}_k \log \tilde{p}_k - \sum_k \tilde{p}_k \log q_k. \tag{4.10}$$

Finally, substituting Eq. (4.10) for the first and third terms in Eq. (4.9a) and canceling the $\log \Omega$ terms, yields

$$\sum_i \tilde{\lambda}_i y_i - \sum_{k,i} \tilde{p}_k \tilde{\lambda}_i x_{ik} = \sum_i \tilde{\lambda}_i \left(y_i - \sum_k \tilde{p}_k x_{ik} \right), \tag{4.11}$$

which must be zero under the first-order conditions for an optimal solution $\tilde{\boldsymbol{p}}$ and $\tilde{\boldsymbol{\lambda}}$.

4.5 Entropy — Variance Relationship

It is possible to show that for any under-determined problem, out of all possible solutions (that are consistent with the observed

data/moments), the ME (or CE) solution $\tilde{\boldsymbol{p}}$ yields the estimated $\boldsymbol{\lambda}$ with the smallest possible variances. Intuitively speaking, the higher the entropy of $\boldsymbol{p}$ (or the uncertainty of the corresponding distribution), the higher the value of the Fisher information matrix of $\boldsymbol{\lambda}, I(\boldsymbol{\lambda})$. Since $\mathrm{Var}(\boldsymbol{\lambda})$ is the inverse of $I(\boldsymbol{\lambda})$, the above statement is confirmed. Similarly, a higher entropy of $\boldsymbol{p}$ means that the T moment constraints do not have much information. Since these observed data do not add much information, one expects that from sample to sample, the variability of the unknown/unobserved parameters $\boldsymbol{\lambda}$, represented by $\mathrm{Var}(\boldsymbol{\lambda})$, will be small. In a more precise way, Cover and Thomas show a direct relationship between Fisher's information matrix and the relative entropy. They show that for the parametric family $\{p_\theta(x)\}$

$$\lim_{\tilde{\theta}\to\theta} \frac{1}{(\theta-\tilde{\theta})^2} D(p_\theta || p_{\tilde{\theta}}) = \frac{1}{\ln 4} I(\theta),$$

where $I(\theta)$ is Fisher's information matrix and "ln" stands for the natural logarithm.

4.6 Discussion

With his basic formulation of the ME principle, Jaynes was able to provide some new insight into the ongoing debate on "probabilities vs. frequencies" by *defining* the notion of probabilities via Shannon's entropy measure. His principle states that in any inference problem, the probabilities should be assigned by the ME principle, which maximizes the entropy subject to the requirement of proper probabilities and any other available information.[3]

[3] In the fields of economics and econometrics, it was probably Davis (1941) who conducted the first work within the spirit of ME. He conducted this work before the work of Shannon and Jaynes, and therefore he did not use the terminology of IT/ME. In his work, he estimated the income distribution by (implicitly) maximizing the Stirling's approximation of the multiplicity factor subject to some basic requirements/rules. For a discussion of Davis's work and of the earlier applications and empirical work within IEE in economics/econometrics see Zellner (1991) and Maasoumi (1993). Other related studies that preceded Jaynes's classical papers on ME include the work of Esscher on the Esscher transform (1932), and discussed nicely by Gerber (1980). The transform looks like an exponential tilting approach similar in spirit to the ME procedure proposed by Efron and Tibshirani (1993). In a different context, Pearson's work provides a way to fit density functions from a finite set of moments. Though, this is not an IT method, it has many of the same features.

Two basic questions keep coming up in the literature: Is the ME principle "too simple?" and does the ME principle "produces something from nothing?" The answer to the above questions is contained in the simple explanation that, under the ME principle, only the relevant information is used, while all irrelevant details are eliminated from the calculations by an averaging process that averages over them. Therefore, it does not produce "something" from "nothing" but rather it only makes use of the available observed information where that information enters as constraints in the optimization. Maximizing the entropy subject to no constraints but the proper probability distribution requirement yields the uniform distribution that represents a state of complete uncertainty (or ignorance). Introducing the observed moments into the optimization takes the distribution away from uniformity. The more information there is in the data, the further away the resulting distribution is from uniformity or from a state of complete ignorance. In that way, one can view the ME method as a method that yields the most uninformed distribution that is consistent with the observed sample moments. Going back to our earlier discussion of information in Section 3.1.1, if we also view information as what constrained our beliefs (or the assignment of our probabilities), then information is represented as constraints on probabilities. Thus, the ME method is a natural way for handling information.

This basic principle and its philosophy motivate much of the Bayesian literature on "How to decide on prior probabilities." The ME principle provides a coherent way of doing so. For example, see the series of proceedings based on the annual International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering started in 1980.[4]

See Solomon and Stephens (1978) for a recent article that nicely describes that procedure, and Bera and Bilias (2002) for discussion of Pearson's work, and others, within the IEE framework.

[4] Stating it differently (A. Caticha, private communication) entropy (or relative entropy) can be thought of as a tool for updating prior probabilities to posterior probabilities when new information (constraints) becomes available. The usual Gibbs–Shannon–Jaynes entropy is the special case that applies to discrete probabilities when the prior happens to be uniform. The information can be in the form of data (and this reproduces Bayes' rule) or in the form of any other kind of constraint such as expected values. In this approach Bayesian and ME methods are unified into one coherent whole. Further, under that view,

4.7 Example

Consider the die example discussed in Section 1. Assuming the observed mean value after n tosses is $y = 5$. The ME formulation is

$$\begin{aligned} &\underset{\{p\}}{\text{Max}}\ H(\boldsymbol{p}) = -\sum_{k=1}^{6} p_k \log_2 p_k \\ &\text{s.t} \sum_k p_k x_k = y \quad \text{and} \quad \sum_k p_k = 1, \end{aligned}$$

where $k = 1, 2, \ldots, 6$, $x_k = 1, 2, \ldots, 6$ and $\log_2$ is used. The solution is

$$\hat{p}_k = \frac{2^{-\hat{\lambda} x_k}}{\sum_{k=1}^{6} 2^{-\hat{\lambda} x_k}} \equiv \frac{2^{-\hat{\lambda} x_k}}{\Omega}.$$

For $y = 5$, the solution $\hat{\boldsymbol{p}}$ is 0.021, 0.038, 0.072, 0.136, 0.255, and 0.478, respectively for $k = 1, \ldots, 6$, $H(\hat{\boldsymbol{p}}) = 1.973$ and $\hat{\lambda} = 0.908$.

Next, we ask the question: "What is the probability that the mean observed value, after n tosses, is no less than 5 if the die is a fair one?" Following the Large Deviations (LD) derivations (Section 3.7) $Q_n(E) \triangleq 2^{-nD(\hat{P}\|Q)}$ for $\hat{P} = \arg\min_{P \in E} D(P\|Q)$ such that the constraint $\sum_k p_k x_k \geq 5$ is satisfied, and where $\triangleq$ means the two terms are equal to the first order in the exponent (e.g., $d_n \triangleq f_n \equiv \lim_{n\to\infty} \frac{1}{n} \log \frac{d_n}{f_n} = 0$). For uniform Q (a fair die), $D(\hat{P}\|Q) = 0.612$. The probability that the observed mean of n tosses of a fair die is no less than 5 is $2^{-nD(\hat{P}\|Q)} = 2^{-n(0.612)}$. For $n = 10$, 100, and 1000 that probability is 0.014, $3.7E$-19, and $5.4E$-185, respectively. More examples are presented in the section on Inference below.

Similarly, using the relative entropy to test the competing hypotheses of a fair die vs. a die with mean of 5 ($\hat{\boldsymbol{p}}$ above), one gets

$$\begin{aligned} [D(\hat{P}||P_2(y=5)) - D(\hat{P}||P_1(\text{uniform}:\ y=3.5))] &= 0 - 0.612 \\ &= -0.612, \end{aligned}$$

entropy needs no interpretation in terms of heat, or as a measure of multiplicity, disorder or amount of information. Entropy is merely a tool for inference.

while the hypothesis of a fair die vs. a die with mean 6 yields

$$[D(\hat{P}||P_2(y=6)) - D(\hat{P}||P_1(\text{uniform}: \ y=3.5))] \to \infty.$$

In the second case, it is possible to distinguish the two probability distributions with probability one (as $n \to \infty$).

The ECT in this example is already investigated and discussed in Section 4.3 above.

5

Information-Theoretic Methods of Estimation — I: Basics and Zero Moments

5.1 Background

The ME principle opened the way for a whole new class of estimation methods. All of these IT methods have the objective of extracting all of the available information from the data, but with minimal assumptions on the underlying distribution generating the data. All of these methods can be constructed as maximization (minimization) of a certain information criterion subject to the observed sample information. In this section I discuss these methods under a unified optimization framework.

Among econometricians who want to analyze data with minimal assumptions or avoid specifying a likelihood function, the IT class of estimators is increasingly popular. This class includes the Empirical Likelihood (EL), the Generalized EL (GEL), the Generalized Maximum Entropy (GME) and the Generalized Method of Moments (GMM), as well as the Bayesian Method of Moments (BMOM). Though most of these moment based estimators are not originally designed to uncover the complete distribution but rather only the (structural) parameters that appear in the moment condition, these methods can also be viewed

as solutions to under-determined problems (in the probability space). If one tries to avoid distributional assumptions, or assumptions on the likelihood function, all problems become under-determined (ill-posed) in probability space. There are always more unknowns than knowns regardless of the sample size. To solve such problems, either the number of unknowns is reduced via the moment conditions that are introduced, or there is a need for a certain criterion to choose among the many solutions (sets of estimates) that are consistent with the observed data, or usually a combination of both of the above is used. The common feature of the methods within the class of IT estimators is their entropy-type criterion that is used to choose one of the many solutions that are consistent with the observed information that is represented in terms of moments or data. This criterion, the entropy of order α (or higher order entropy), is more familiar to econometricians as the Cressie–Read criterion (Eq. (3.9)). Though all of these IT methods are related via that criterion, they differ in two basic ways: the pre-specified α-level of the generalized entropy objective function, and the way the observed data enter into the optimization.

The relationship of this criterion to entropy and other information-theoretic measures is discussed in Kitamura and Stutzer (1997), Imbens et al. (1998), Golan et al. (1996b), (Mittelhammer et al., 2000, Chapters 12–13), (Golan, 2002), Kitamura (2006), and Smith (2000, 2004, 2005).

5.2 The Generic IT Model

Let $Y = \{Y_1, Y_2, \ldots, Y_T\}$ be a sample of iid observations from an unknown distribution F_0. There is an N-dimensional parameter vector $\boldsymbol{\theta}_0$ that is related to F_0 in the sense that the information about $\boldsymbol{\theta}_0$ and F_0 is available in the form of $M \geq N$ moments (or functionally independent unbiased estimating functions). An IT "likelihood" of this parameter vector is defined by considering the distributions supported on the sample, where Y_i is assigned a probability p_i where $\boldsymbol{p}' = \{p_1, p_2, \ldots, p_T\}$. For a specified value of the parameter vector, say $\boldsymbol{\theta}_1$, the IT (empirical) likelihood is defined as the maximal value of some function $f(\cdot)$, defined below, over all such probability distribu-

tions satisfying the relationship between $\boldsymbol{y}$, $\boldsymbol{p}$, and $\boldsymbol{\theta}_1$ that are specified via the M-dimensional vector equation $\boldsymbol{g}(\boldsymbol{y}, \boldsymbol{p}, \boldsymbol{\theta}_1) = [\mathbf{0}]$. Often the elements of the function $g(\cdot)$ are called parametric information functions (PIF's). Under that approach, one starts by defining the feasible set of proper probability distributions supported on the sample observations. The feasible set is characterized by a set of M restrictions on the unknown probabilities $\boldsymbol{p}$. These restrictions are based on the PIFs specified as parametric functions of the data. Usually these functions are the moments. Given the T observations and M PIF restrictions (moment conditions), the objective is to obtain estimates of the probability $\boldsymbol{p}$. These estimates represent the (unobserved empirical) weight of each observed data point. In most common cases $M \ll T$ (where "$\ll$" stands for "much smaller"), implying the problem of estimating $\boldsymbol{p}$ based on the observed information in the PIFs is under-determined.

With the above in mind, in the class of Information-Theoretic estimators rather than starting with a pre-specified likelihood, the observed data are used to estimate the empirical distribution (or natural weights) $\boldsymbol{p}$, that is most consistent with the M observed sample moments. Regardless of the sample size, there are infinitely many sets of "weights" that are consistent with the observed sample, making the problem an under-determined problem. Like the ME formulation, a simple way to solve such an under-determined (ill-posed) problem is to transform it into a well-posed, constrained optimization problem. This is done by minimizing a certain criterion subject to the observed sample moments or any other function of the data. But which solution should one choose? Or stating it differently, what criterion should be minimized? The objective function used in all the IT estimators is the generalized entropy measure (entropy of order α), also known as the Cressie–Read function (see Eq. (3.9)) for different values of the α, where Shannon's entropy is just a special case of that function. All of these measures are entropy divergence measures reflecting a certain entropy distance between two distributions; say a prior and an empirical one. Using these divergence measures form a basis for optimal decision making and for statistical inference.

Let the vector $\boldsymbol{q}$ be a proper distribution on the same support as $\boldsymbol{p}$, say $\boldsymbol{q}' = \{q_1, q_2, \ldots, q_T\}$. These $\boldsymbol{q}$'s can be viewed as some priors. Let

$f(\boldsymbol{p}||\boldsymbol{q})$ be any function defined on $\boldsymbol{p}$ and $\boldsymbol{q}$ denoting some well defined distance between these two proper probability distributions. A class of possible estimators for solving the above under-determined problem is

$$\text{Generic Estimator} = \begin{cases} \boldsymbol{p}^* = \arg\min\{f(\boldsymbol{p}\|\boldsymbol{q})\} \\ \text{s.t.} \\ g_m(\boldsymbol{y}, \boldsymbol{p}, \boldsymbol{\theta}_1) = [\mathbf{0}];\ m = 1, 2, \ldots, M \\ \sum_{i=1}^{T} p_{\mathrm{i}} = 1;\ i = 1, 2, \ldots, T \text{ and } M < T - 1 \end{cases}.$$

The IT class of estimators is a subset of the above class of estimators. Substitute the generalized entropy function (3.6), $D_{\alpha}^{R}(\boldsymbol{p}\|\boldsymbol{q}) = \frac{1}{1-\alpha}\log\sum_i \frac{p_i^{\alpha}}{q_i^{\alpha-1}}$, (or Eq. (3.9)), representing an information-divergence measure between $\boldsymbol{p}$ and $\boldsymbol{q}$, for $f(\cdot)$ yields

$$\text{IT Estimators} = \begin{cases} \widehat{\boldsymbol{p}} = \arg\min\{f(\boldsymbol{p}\|\boldsymbol{q}) = D_{\alpha+1}^{R}(\boldsymbol{p}\|\boldsymbol{q})\} \\ \text{s.t.} \\ g_m(\boldsymbol{y}, \boldsymbol{p}, \theta_1) = [\mathbf{0}];\ m = 1, 2, \ldots, M \\ \sum_{i=1}^{T} p_{\mathrm{i}} = 1;\ i = 1, 2, \ldots, T \text{ and } M < T - 1 \end{cases}.$$

The class of IT estimators $\{\widehat{\boldsymbol{p}}\}$ is a subset of the estimator class $\{\boldsymbol{p}^*\}$ where each specific estimator in that class (to be discussed below) depends on the pre-specified α and on the exact specification of $g_m(\cdot)$. If, for example, $\alpha \to 0$, $f(\boldsymbol{p};\boldsymbol{q}) = D_{\alpha+1}^{R}(\boldsymbol{p},\boldsymbol{q})$ becomes the Kullback–Liebler divergence measure $D(\boldsymbol{p}||\boldsymbol{q}) \equiv \sum_{i=1}^{T} p_i \log(p_i/q_i)$, and the CE solution $\tilde{\boldsymbol{p}}$ results. If in addition, the $\boldsymbol{q}$'s are uniform ($q_i = 1/T$ for all $i = 1, 2, \ldots, T$) $\tilde{\boldsymbol{p}} = \hat{\boldsymbol{p}}$ which is equivalent to using the negative of the Shannon's entropy $H(p) \equiv -\sum_{i=1}^{T} p_i \log_2 p_i$ as the objective function, resulting in the ME solution $\hat{\boldsymbol{p}}$.

The formulation above is presented as a constrained optimization problem. Like the ME, the solution is derived via the Lagrangean. Once the optimal solution $\widehat{\boldsymbol{p}}$ is found, the dual, unconstrained formulation (or the concentrated model) can be formulated. This is done via the same route discussed in Section 4.2, Eq. (4.4). The transformation from the primal (constrained) model to the dual (unconstrained) model is one of the great advantages of that framework. By doing so, it is

possible to move from the higher dimensional probability space to the much lower dimension of the parameter space. Further, the Lagrange multipliers, which are directly related to the estimated parameters, also reflect the contribution of each constraint (moment) to the optimal value of the objective function. In the IT class, the objective is an informational criterion, meaning the estimated Lagrange multipliers reflect the marginal information of each constraint (data point, moment, etc.). It is the same Lagrange multipliers that enter as the parameters in the estimated probability distribution. This issue and its relationship to hypothesis tests are discussed in Sections 6 and 7. A brief discussion of the members of the IT family of estimators that have one common property: they all use zero-moment conditions, is now provided. Another member of that family, that uses stochastic moment conditions, is discussed in Section 6.

5.3 Empirical Likelihood

Letting $\alpha \to -1$ subject to the same set of $M+1$ restrictions yields the Empirical Likelihood method. Given the M moments, our objective is to estimate the T-dimensional, unknown distribution $\boldsymbol{p}$. Let $\boldsymbol{y}$ be a T dimensional random vector characterized by an unknown T-dimensional distribution $\boldsymbol{p}$ with a vector of unknown parameters $\boldsymbol{\theta}$ and $g_m(y_t;\theta)$ represents the M moments of the distribution $\boldsymbol{p}$. For example, if $M=2$, $g_m(y_t;\theta)$ may be $\sum_t p_t y_t = \theta_1$ and $\sum_t p_t y_t^2 = \theta_2$. Similarly if $\boldsymbol{y}$ is a function of a set of covariates X, $\boldsymbol{y} = f(X)$, these two moments can be expressed accordingly. The M (pure) moments can be expressed as

$$\sum_t p_t g_m(y_t;\theta) = 0,$$

where θ is an unknown parameter (or a vector of parameters). Using the framework discussed above and letting $\alpha \to -1$ the EL criterion is simply the probability of the observed sample or its natural logarithm (i.e., the log empirical likelihood):

$$\prod_{t=1}^{T} p_t \quad \text{or} \quad \sum_{t=1}^{T} \log(p_t) \quad \text{or} \quad \frac{1}{T}\sum_{t=1}^{T} \log(p_t).$$

Following Owen (1990, 1991, 2001), DiCiccio et al. (1991) and Qin and Lawless (1994), the EL approach for choosing the probability distribution $\boldsymbol{p}$ is

$$\operatorname*{Max}_{\boldsymbol{p}} \frac{1}{T}\sum_{t=1}^{T}\log p_t \tag{5.1}$$

subject to the structural and general constraints

$$\sum_t p_t g_m(y_t;\theta) = 0, \tag{5.2}$$

$$\sum_t p_t = 1, \tag{5.3}$$

$$p_t \geq 0 \qquad . \tag{5.4}$$

The corresponding Lagrangean and first-order conditions with respect to $\boldsymbol{p}$ are

$$L_{\mathrm{EL}} = \frac{1}{T}\sum_t \log p_t - \sum_m \lambda_m \left[\sum_t p_t g_m[y_t,\theta]\right] + \eta\left(1 - \sum_t p_t\right) \tag{5.5}$$

$$\frac{\partial L}{\partial p_t} = \frac{1}{T}\frac{1}{p_t} - \sum_m \lambda_m g_m(y_t,\theta) - \eta = 0, \tag{5.6}$$

from which it follows that

$$\sum_t p_t \frac{\partial L}{\partial p_t} = \frac{1}{T}T - \eta = 0, \tag{5.7}$$

so $\eta = 1$. The resulting optimal estimated weights (probabilities) are

$$\hat{p}_t = T^{-1}\left[\sum_m \hat{\lambda}_m g_m(y_t;\theta) + 1\right]^{-1}. \tag{5.8}$$

For completion, consider the simple linear model discussed in Section 4.[1] Let X be an $M \times T$ design matrix X, the M observed

[1] Note that now, in order to be consistent with the notation in this section, the indices for $\boldsymbol{p}$ and X are different than those in Section 4, but the meanings of $\boldsymbol{p}$ and X and their relationship to $\boldsymbol{y}$ are unchanged.

moments are $\boldsymbol{y} = X\boldsymbol{p}$ (or $\boldsymbol{y} - X\boldsymbol{p} = \mathbf{0}$), $\boldsymbol{p}$ is a T-dimensional proper probability distribution and $T > M$. Then,

$$\text{EL} = \begin{cases} \hat{\boldsymbol{p}} = \arg\max \left\{ \dfrac{1}{T} \sum_t \log p_t \right\} \\ \text{s.t.} \;\; \boldsymbol{y} - X\boldsymbol{p} = \mathbf{0};\; \sum_t p_t = 1;\; p_t \geq 0 \end{cases}$$

and

$$\hat{p}_t = T^{-1} \left[\sum_m \hat{\lambda}_m x_{mt} + 1 \right]^{-1}. \tag{5.9}$$

Example (*The Linear Regression Model*). Consider the linear regression model $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\boldsymbol{y}$ is a T-dimensional vector of observed data, X is a $T \times K$ matrix and $\boldsymbol{\varepsilon}$ is a T-dimensional random vector with mean zero. The EL model is

$$\begin{aligned} &\ell(\boldsymbol{\beta}, \theta; \boldsymbol{y}) \equiv \underset{\boldsymbol{p}, \beta}{\text{Max}} \left\{ \sum_{i=1}^{T} \log p_i \right\} \\ &\text{s.t.} \\ &\sum_{i=1}^{T} p_i \boldsymbol{x}_i \left(y_i - \sum_k x_{ik} \beta_k \right) = \mathbf{0} \\ &\sum_{i=1}^{T} p_i = 1; \quad p_i \geq 0. \end{aligned} \tag{5.10}$$

For $\hat{p}_i(\text{EL}) = 1/T$ for all $i = 1, \ldots, T$ (uniform), the EL solution is equivalent to the least squares solution.[2] If, for example and loosely speaking, the X's are correlated with the $\boldsymbol{\varepsilon}$, a set of instruments S that are correlated with X but not with $\boldsymbol{\varepsilon}$ can be used. In that case, rather than using the moments $E[X'\boldsymbol{\varepsilon}]$ in the linear equation, as is done in (5.10) above, we are using the instruments to

[2] Note that in the OLS case, the parameter vector is just identified by the moment conditions so the EL estimator is just the method of moments' estimator which is the OLS in that case.

form the moments $E[S'\boldsymbol{\varepsilon}]$:

$$\begin{aligned}
&\ell(\boldsymbol{\beta},\theta;\boldsymbol{y}) \equiv \underset{\boldsymbol{p},\boldsymbol{\beta}}{\text{Max}} \left\{ \sum_{i=1}^{T} \log p_i \right\} \\
&\text{s.t.} \\
&\sum_{i=1}^{T} p_i \boldsymbol{s}_i \left(y_i - \sum_k x_{ik}\beta_k \right) = \mathbf{0} \\
&\sum_{i=1}^{T} p_i = 1; \quad p_i \geq 0.
\end{aligned} \tag{5.11}$$

If all probabilities are uniform ($p_i = 1/T$ for all $i = 1,\ldots,T$), the EL solution is equivalent to the traditional instrumental variable (IV) solution[3]: $\hat{\boldsymbol{\beta}}_{IV} = (S'X)^{-1}S'\boldsymbol{y}$.

For more details, including examples and test-statistics, see Owen (2001), Qin and Lawless (1994) and Mittelhammer, Judge, and Miller (2001, Chap. 12). For recent advancements and a survey of the EL (and GEL) see Smith (2000, 2005), Ramalho and Smith (2002), Kitamura (2006) and Schennach (2004).

Two notes in conclusion. First, substituting the objective function within the EL framework for the entropy of order α (or the Cressie Read function — Eq. (3.9)) takes us back to the generic IT estimator discussed in the previous section. This idea goes back to the work of Imbens et al. (1998) that discuss three special cases of that objective function. As discussed earlier, in that class of estimators, the researcher has to specify the α in Eq. (3.9). However, only the case where $\alpha \to 0$ is fully consistent with Shannon's entropy. Second, Smith (1997) considered a more general class of estimators which he called Generalized EL (GEL). Under the GEL, the constrained optimization model is transformed into a concentrated model. Rather than working with the larger probability space the problem is respecified as a concentrated (likelihood-like) model in the parameters space (or the Lagrange multipliers space). This idea is similar to the original work of Agmon et al. (1979) who were the first to construct the concentrated ME model (Eq. (4.4)), and then applied and extended to the Generalized ME by Golan and Judge (1993), Golan et al. (1994), Miller (1994)and Golan

[3] See previous footnote.

et al. (1996b). See for example Golan et al. (1996b) for detailed examples and derivations. The GME is discussed in Section 6.

5.3.1 Efficiency and Information

Building on Section 4.4, it is easy to show that the EL is an efficient (or 100% efficient) IPR. A 100% efficient IPR is defined as an IPR where the output and input information are equal. To simplify exposition, the linear case is presented here. Let

$$I_1(\hat{\boldsymbol{p}}) \equiv -\sum_t \hat{p}_t \log q_t = -\sum_t \hat{p}_t \log \frac{1}{T} = -\log \frac{1}{T} = \log(T)$$

and

$$I_2(\hat{\boldsymbol{p}}) \equiv -\sum_t \hat{p}_t \log \left[\sum_m \hat{\lambda}_m x_{mt} + 1\right]^{-1} = \sum_t \hat{p}_t \log \left[\sum_m \hat{\lambda}_m x_{mt} + 1\right]$$

be the input information. Similarly, let $O(\hat{\boldsymbol{p}}) \equiv -\sum_t \hat{p}_t \log \hat{p}_t$ be the output information. Like the ME (CE) we want to show that the output information equals the input information:

$$-\sum_t \hat{p}_t \log \hat{p}_t - \log(T) - \sum_t \hat{p}_t \log \left[\sum_m \hat{\lambda}_m x_{mt} + 1\right] = 0. \tag{5.12}$$

Substituting $\hat{p}_t$ Eqs. (5.8) and (5.9) into Eq. (5.12) yields

$$\begin{aligned}
&-\sum_t \hat{p}_t \log \left[(T^{-1}) \left(\sum_m \hat{\lambda}_m x_{mt} + 1\right)^{-1}\right] - \log(T) \\
&\quad -\sum_t \hat{p}_t \log \left(\sum_m \hat{\lambda}_m x_{mt} + 1\right) \\
&= \log(T) + \sum_t \hat{p}_t \log \left(\sum_m \hat{\lambda}_m x_{mt} + 1\right) - \log(T) \\
&\quad -\sum_t \hat{p}_t \log \left(\sum_m \hat{\lambda}_m x_{mt} + 1\right) = 0.
\end{aligned} \tag{5.13}$$

As one would expect, the EL is a 100% efficient IPR. This result is now used to compare the input information of the EL and ME. Holding

the prior information under both models to be the same, $I_1(\hat{\boldsymbol{p}}) = I_1(\tilde{\boldsymbol{p}})$, it is sufficient to compare $I_2(\hat{\boldsymbol{p}})$ and $I_2(\tilde{\boldsymbol{p}})$ — the "likelihood" functions of both methods. It was shown earlier that for all $\boldsymbol{\lambda} \neq \mathbf{0}$ $O(\hat{\boldsymbol{p}}) < O_1(\tilde{\boldsymbol{p}})$. Consequently, it follows that $I_2(\hat{\boldsymbol{p}}) < I_2(\tilde{\boldsymbol{p}})$ which completes this argument. This argument shows that the EL model uses more input information than the ME model. This additional information enters through the EL's likelihood function.

Example (*EL and the ECT*). Consider the die problem of Section 4, but now one uses the EL rather than the ME to estimate the probabilities:

$$\begin{aligned} &\underset{\{p\}}{\text{Max}} \ \tfrac{1}{K}\sum_{k=1}^{6}\log p_k \\ &\text{s.t} \\ &\sum_k p_k x_k = y \quad \text{and} \quad \sum_k p_k = 1 \end{aligned}$$

for $k = 1, 2, \ldots, 6$, $x_k = 1, 2, \ldots, 6$ and $y = 5$. The EL solution is $\hat{\boldsymbol{p}}_{\text{EL}} = 0.044,\ 0.053,\ 0.069,\ 0.098,\ 0.167,$ and $0.570,$ and $H(\hat{\boldsymbol{p}}_{\text{EL}}) = 1.909$. Recalling the ME solution of $\hat{\boldsymbol{p}}$ is 0.021, 0.038, 0.072, 0.136, 0.255, and 0.478, respectively for $k = 1, \ldots, 6$ and $H(\hat{\boldsymbol{p}}) = 1.973$, one can calculate ΔH via the ECT. For $N = 100$ and $\alpha = 0.05$, $\chi^2_{(4;0.05)} = 9.488$. Thus, $\Delta H = 0.0474$ and $1.9256 \leq H \leq 1.973$. Consequently, the EL solution is outside the 95% concentration range. In the ECT terms, the $\hat{\boldsymbol{p}}_{\text{EL}}$ is an atypical set. Figure 5.1 shows these two estimated distributions.

5.4 Generalized Method of Moments — A Brief Discussion

5.4.1 Background

The GMM framework and its many applications within econometrics are beyond the scope of this review as it demands its own review and books. However, the literature on GMM and IT is briefly discussed here. The focus here is (i) to show the relationship between the GMM and IT methods and (ii) to show that though the GMM and other related, alternative methods, are developed for handling over-identified problems, if one looks at it in a different space — the unknown discrete

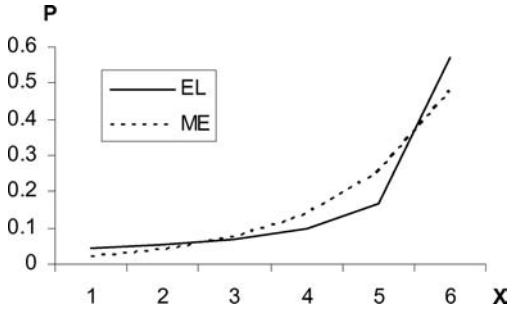


Fig. 5.1 The EL and ME estimated distributions for the die problem with mean of 5.

probability space — then it is exactly the under-determined problem solved via the ME principle.

Under its original specification (Hansen, 1982), the GMM estimator makes use of functions of the random variables and the unknown parameters (zero-moment functions) that have expectation of zero when evaluated at the true (unknown) values of the parameters. One advantage of the GMM is that it is easily connected to economic theory where the zero-moment conditions are expressed as resulting from agents' optimization. Under the GMM, the unknown parameters are estimated by setting the sample means of the moment conditions as close as possible to zero. The other main advantage of the GMM is that it can handle the exactly identified case (number of moment conditions equals the number of unknown parameters) and the over identified case where there is more information (observed moments) than unknown parameters to be estimated. Hansen's innovation here was to set up a linear combination of all the moment conditions where the dimension of the linear combinations equals the dimension of the unknown parameters.[4] In this way, all the available information in the data (moments) can be captured. Originally, Hansen proposed a two-step GMM estimator where the first step is used for estimating that linear combination. But, in practice, that first step is inefficient and has an impact on the small sample properties of the GMM estimator. Since Hansen's (1982)

[4] The GMM method generalizes the method of moments. See Bera and Bilias (2002) for a historical perspective and synthesis.

paper, much work has been done on the choice of the "optimal" linear combination. A large portion of that work connects the information-theoretic estimators with the GMM estimator (for over-identified estimation problems) in order to increase the efficiency.

Since Hansen's seminal paper, the GMM became a unifying framework for estimation and inference in econometrics. The success of the GMM is not only due to its properties and applicability to economic problems, but also because it encompasses many of the commonly used estimation methods (including maximum likelihood, ordinary least squares, generalized least squares, instrumental variables, and two-stage least squares). Generally speaking, though GMM has desirable large sample properties, it is well known that its small sample performance is quite poor for certain applications (Hansen et al., 1996). Examples include possible bias of the two-step GMM estimator and lack of accuracy of the confidence intervals. In view of this, an IT alternative model was proposed originally by Kitamura and Stutzer (1997). Within the linear regression model framework, their model is an innovation on the ME formulation. They use the relative entropy measure $D(\boldsymbol{p}\|\boldsymbol{q})$ subject to the observed sample's moments to estimate the (implicit) empirical weights associated with each observation. They then derive its dual formulation, as discussed in Section 4, which relates the Lagrange multipliers to the estimated $\boldsymbol{\beta}$'s. At approximately the same time, Imbens et al. (1998) developed their IT approach for inference in moment condition problems. Using the same logic as above, they use the relative entropy measure to construct a class of unconstrained optimization problems that provide alternatives to the GMM. Loosely speaking, the main advantage of the IT estimators is to provide alternatives to the original GMM that utilize the available, observed information more efficiently and therefore produce more efficient estimators. By constructing the estimator in its concentrated (dual unconstrained) form (Eq. (4.4)), it becomes computationally more efficient. For recent reviews and advancements of GMM see Hall (2005). For a more detailed discussion of GMM and its relationships to EL and GEL see Imbens (2002), Kitamura (2006) and Otsu (2006). For a nice synthesis of the method of moments, GMM and information-theoretic methods see Bera and Bilias (2002). A short discussion of GMM and IT is now provided.

5.4.2 GMM and IT

A more concise discussion of GMM and the different alternative IT estimators is now discussed. These alternative estimators provide the information-theoretic interpretation to the over-identified GMM estimator. The discussion here is restricted to the main logic behind the class of IT estimators' alternatives to GMM. A detailed discussion of each one of these models is outside the scope of this review.

Using the framework in Imbens (2002), let θ^* be a K-dimensional vector ($\theta^* \in \Theta \subset \mathbb{R}^K$). Let $\boldsymbol{y}$ be a random vector of dimension P with its supports in $y \subset \mathbb{R}^P$. Define the moment function as $g : y \times \Theta \rightarrow \mathbb{R}^M$ which is a known vector valued function such that $E[g(y,\, \theta^*)] = 0$ and $E[g(y,\, \theta)] \neq 0$ for all $\theta \in \Theta$ and $\theta \neq \theta^*$. Given our random sample $y_1, \ldots, y_T$ we wish to find an estimator for θ^* and to evaluate the large sample properties of that estimator.

In the just identified case where $K = M$ (dimension of θ equals the dimension of g) we can estimate θ^* by solving

$$\frac{1}{T}\sum_i g(y_i,\, \hat{\theta}_{\mathrm{GMM}}) = 0. \tag{5.14}$$

Replacing the sample mean by the expectation yields the unique solution that is equal to θ^*. This estimator is unique and consistent (Hansen, 1982; Newey and McFadden, 1994).

If we have more moment conditions than unknown parameters ($M > K$) then there is no solution to (5.14). Hansen (1982) idea was to generalize the above optimization problem to a minimization of the quadratic problem $Q_{W,T}(\theta)$:

$$Q_{W,T}(\theta) = \frac{1}{T}\left[\sum_i g(y_i,\, \theta)\right]' W \left[\sum_i g(y_i,\, \theta)\right] \tag{5.15}$$

for some $M \times M$ positive definite and symmetric (weight) matrix W. Hansen (1982) and Newey and McFadden (1994) develop the large sample properties of the minimand of Eq. (5.15), $\hat{\theta}_{\mathrm{GMM}}$. In the just identified case, the choice of the weight matrix W is irrelevant because, for large samples, $\hat{\theta}_{\mathrm{GMM}}$ will be equal to the value of θ that sets the average moments exactly to zero and the solutions for model (5.14) and (5.15) coincide.

In the overidentified case, on the other hand, the choice for the weight matrix is crucial. Loosely speaking, the optimal choice for W (in terms of minimizing the asymptotic variance) is just the inverse of the covariance of the moments. Unfortunately, this choice is unknown to the empirical researcher. Therefore, Hansen proposed a non-optimal solution, known as the two-step GMM. In the first step, one can obtain an estimate of θ^* by minimizing $Q_{W,T}(\theta)$ using some arbitrary (positive definite and symmetric) weight matrix W. With this initial estimate $\breve{\theta}$, the optimal W can be estimated via

$$\hat{W}^{-1} = \left[\frac{1}{T} \sum_i g(y_i, \breve{\theta}) g(y_i, \breve{\theta})' \right]^{-1}.$$

Substituting $\hat{W}$ for W in (5.15), in the second stage, we estimate θ^* by minimizing $Q_{\hat{W},T}(\theta)$. The resulting estimator $\hat{\theta}_{\mathrm{GMM}}$ has the same first-order asymptotic distribution as the minimand of the quadratic form (Eq. (5.15)) with the true W. For more details, specific tests and further interpretations for this basic framework see for example Hansen (1982) and Imbens (2002). Some tests are also discussed below.

As discussed earlier, though the two-stage GMM has appealing large sample properties and it provides a new framework for analyzing problems characterized by over identified set of functions (moments), it is not efficient for small samples, is biased in some cases and there may be lack of precision in evaluating confidence intervals. A number of alternative estimators were developed. Here, just the class of IT estimators is discussed briefly. The basic idea is to find a way to estimate the weight matrix W. But rather than to estimate it directly (which is inefficient with the information the researcher usually possesses) it is estimated indirectly. If one assumes that the observed sample data are discrete with some support and unknown probabilities, then, the basic parameters of interest (appearing in the moment conditions) can be expressed as functions of these support points and probabilities. In other words, one can estimate the unknown weights (natural probabilities) associated with each observation. These probabilities are functions of the Lagrange multipliers associated with each moment condition which are one-to-one functions of the unknown parameters θ. In that way, rather

than estimating W directly, the parameters of interest are estimated directly while using all of the available information efficiently. Further more, one can view the estimated Lagrange multipliers as the "shadow values" of the informational content of each one of the moment conditions. This method will work whether the problem is exactly identified or over-identified in the parameter space. In fact, under that view, unless one knows (or assumes) the likelihood function, for all common estimation problems where $T > M$, and regardless whether $M = K$ or $M > K$, the problem in the probability space is always under-determined. To fix ideas, consider the following example.

Consider a sample of iid $y_1, \ldots, y_T$ univariate observations from an unknown distribution F with mean θ and $E[y^2] = m(\theta)$ where $m(\cdot)$ is a known function. Examples include the single parameter problem (Qin and Lawless, 1994) $E[y] = \theta$ and $E[y^2] = m(\theta) = 2\theta^2 + 1$. The objective is to estimate the unknown parameter θ. The information about F can be expressed via the two estimating functions $\sum_i p_i y_i - \theta = 0$ and $\sum_i p_i y_i^2 - 2\theta^2 - 1 = 0$. Thinking of this problem in probability space, means that this problem is exactly the ME problem discussed earlier. Given $M = 2$ observed moments (and the third requirement that $\sum_i p_i - 1 = 0$), the full distribution of dimension $T > 3$ is estimated via the ME principle. The ME solution is the MaxEnt (exponential) distribution with the Lagrange multipliers being the "real" parameters of interest and are one-to-one related to θ. Explicitly,

$$\begin{aligned}
&\underset{p,\theta}{\text{Min}}\ D(\boldsymbol{p}||\boldsymbol{q}) \equiv \sum_i p_i \log(p_i/q_i) \\
&\text{s.t} \\
&\sum_i p_i y_i - \theta = 0 \\
&\sum_i p_i y_i^2 - 2\theta^2 - 1 = 0 \\
&\sum_i p_i - 1 = 0
\end{aligned}$$

and for generality, the cross-entropy $D(\boldsymbol{p}||\boldsymbol{q})$ is used here, where $\boldsymbol{q}$ is a vector of prior probabilities (or empirical distribution) that is taken to

be uniform. Solving the Lagrangean yields the MaxEnt distribution

$$\tilde{p}_i = \frac{q_i \exp(\tilde{\lambda}_1 y_i + \tilde{\lambda}_2 y_i^2)}{\sum_i q_i \exp(\tilde{\lambda}_1 y_i + \tilde{\lambda}_2 y_i^2)} = \frac{q_i \exp(\tilde{\lambda}_1 y_i + \tilde{\lambda}_2 y_i^2)}{\Omega(\tilde{\boldsymbol{\lambda}})},$$

where $\tilde{\theta} = -\frac{\tilde{\lambda}_1}{4\tilde{\lambda}_2}$.

As was shown earlier, that problem can be written as an unconstrained (concentrated) model with respect to the Lagrange multipliers $\boldsymbol{\lambda}$ (see Eq. (4.4)), which in turn yields $\tilde{\theta}$. Then, estimating the Covariance of these Lagrange multipliers, or any other function of them, is straight forward. See for example (Golan et al., 1996b, Appendix 3C), for derivation of the covariance matrix for both the primal (constrained) and concentrated (unconstrained) entropy models and the transformation between the covariance in parameter space to the probability space.

The above problem can be solved with any one of the IT methods under the generic IT estimators of Section 5.2. Consider for example the EL method. Substitute the entropy (or entropy divergence) objective for the EL objective $\sum_i \log p_i$ and maximize with respect to $\boldsymbol{p}$ and θ, subject to the same three constraints, yields the EL solution

$$\hat{p}_{i(\mathrm{EL})} = [\hat{\lambda}_1 y_i + \hat{\lambda}_2 y_i^2 + 1]^{-1} \quad \text{with } \hat{\theta} = -\frac{\hat{\lambda}_1}{4\hat{\lambda}_2}.$$

In both cases, the solutions $\tilde{F}_{\mathrm{ME}}$ and $\hat{F}_{\mathrm{EL}}$ satisfy the observed moment functions.[5] The same can be done with any one of the IT estimators (different values of α in the entropy of order α, or the Cressie–Read function, Eq. (3.9)). To relate it back to the GMM estimator, if rather than specifying W, one wishes to estimate the common distribution of the T observations, a natural choice is the (uniform) empirical distribution of $1/T$. But within the over-identified GMM case the choice of equal/uniform weights does not satisfy the requirement that the expected value of the random variable is θ. Instead, with the above approach, one searches for the distribution F that is the closest as possible to the empirical distribution $1/T$. In information-theoretic terms,

[5] Qin and Lawless (1994) present sampling experiments based on that example. Golan and Judge (1996) use sampling experiments to contrast the EL with the ME for that example.

one searches for the least informed distribution out of the infinitely many distributions satisfying the observed moments representing the only available information. That distribution is related to the unknown parameters of interest via the Lagrange multipliers associated with each observed moment (or data point). These moments represent the only available information. They capture the amount of information contained in each one of the moment equations and are the basic parameters that determine the estimated $\boldsymbol{p}$'s. Similarly, through their one-to-one relation to θ they provide the estimated distribution $\hat{F}$.

There are a number of advantages for that approach. First, one does not have to start by estimating W which results in increase efficiency. Though the IT estimators, such as the EL and GEL, use the same set of moments (information), they remove some of the imprecision resulting from estimating the weight matrix. Second, the estimated Lagrange multipliers have information-theoretic interpretation. They capture the informational content of each observed moment. They capture the marginal informational contribution of each moment function to the optimal value of the objective function. (Note that the objective function is an information measure). Third, the structural parameters of interest are functions of these multipliers. Fourth, as is shown in the literature, under that approach, confidence intervals and other likelihood-like tests can be easily performed. Fifth, the IT estimators can be constructed as concentrated (unconstrained) models making them computationally more efficient. This result goes back to Agmon et al. (1979) within the ME.

To summarize, the two-step GMM and all other zero-moment IT estimators discussed here are first-order asymptotically efficient. Qin and Lawless (1994) and Imbens (1997) show that for properly specified moment conditions, the EL and the two-step GMM estimators are equivalent to the order of $O_P(N^{-1/2})$. Imbens (2002) discusses the differences among these estimators by calculating approximations to their finite-sample distributions. These results are based on Newey and Smith (2004) and Imbens and Spady (2001). These approximations show that the two-step GMM has a number of terms that do not appear in the GEL class of estimators. Some of these terms increase in magnitude as the number of moments increases. In terms of bias,

the bias of the GMM estimator increases (linearly) with the number of moment conditions. The bias under the GMM estimator is linear in the number of over-identifying restrictions, with the coefficient equals to the correlation between the moments and their derivatives. In contrast, the bias (up to the order analyzed) in the exponential tilting (entropy) estimator is not affected by the number irrelevant moments. It seems that the main reason for the differences between the IT estimators and the GMM for estimating over-identified problems is that under the IT framework that goes back to Jaynes's (1957a, b) work on the ME principle, the information is contained in the constraints and adding an irrelevant information means the value of the objective function is unchanged. Within the over-identified GMM structure adding irrelevant information may have a direct impact on W which impacts the values of the estimated parameters.

See Imbens (2002) for a detailed discussion on GMM and EL, the earlier work of Chamberlain (1987) on efficiency (within GMM), Kitamura and Stutzer (1997) for discussion of the dependent case for the GMM and the Kullback–Liebler discrepancy function, and Mittelhammer et al. (2000) for a general discussion of GMM and EL.

5.4.3 GMM and Tests

A basic question within the GMM framework is how can one confirm the validity, and test the informational content, of the over-identifying set of restrictions. These tests are easily formulated within the IT class of estimators. A simpler version of these tests was discussed in Section 3. A generalized version of these tests is discussed below (Section 6) within the context of the Generalized ME. These tests are the empirical likelihood ratio test, the Wald test, and the Lagrange multipliers tests. Imbens (2002) shows that the leading terms in these tests are identical to the leading term of the test developed in Hansen's original paper. The EL ratio test is based on the difference between the restricted and unrestricted values of the EL (or entropy of order α) objective functions. Under a certain set of regularity conditions, two times the difference is distributed as a χ^2 with degrees of freedom equals to the number of over-identifying conditions for the test statistic under the null hypothesis.

A test analogous to the traditional Wald test is based on the difference between the mean moments and their probability limit under the null hypothesis of zero. Like Hansen's (1982) test the average moments are weighted by their covariance matrix. Unlike the GMM framework where the covariance is estimated based on an initial estimates of the unknown parameters, under the IT estimators, the estimated probabilities themselves are used.

The third test is analogous to the Lagrange multiplier test. Imbens (2002) discusses two versions of this test. The first uses a generalized inverse for the covariance in order to compare the Lagrange multipliers to zero. The second, proposed by Imbens et al. (1998) is to use the inverse of the covariance matrix directly. The second version is simpler and is found to perform better in sampling experiments. Like the other two test statistics, both of these Lagrange multipliers tests (in large samples) have a χ^2 distribution with degrees of freedom equals the number of over-identifying restrictions.

For recent developments and specific applications in IT-GMM see recent work by Smith (2007), Kitamura and Stutzer (2002), Antoine et al. (2007), Hall et al. (2007a), Wu and Perloff (2007), Judge and Mittelhammer (2007), Journal of Econometrics Special issue on IEE (2002, 2007), Econometric Reviews (2008) and the recent text by Hall (2005).

5.5 Bayesian Method of Moments — A Brief Discussion

The above IT methods were discussed within the philosophy of "sampling theory." However, with the same objective of estimating with minimal *a priori* assumptions on the likelihood, much work has been done within the Bayesian philosophy. This trend started with the seminal work of Zellner on the Bayesian Method of Moments, BMOM, (Zellner, 1994, 1996a, 1997; Tobias and Zellner, 2001). As with the other IT estimators discussed earlier, the idea behind the BMOM method is to estimate the unknown parameters with minimum assumptions on the likelihood function. As stated by (Zellner, 1997, p. 86), "The BMOM approach is particularly useful when there is difficulty in formulating an appropriate likelihood function. Without a likelihood function, it is

not possible to pursue traditional likelihood and Bayesian approaches to estimation and testing. Using a few simple assumptions, the BMOM approach permits calculation of post-data means, variances and other moments of parameters and future observations."

To avoid a likelihood function Zellner proposed to maximize the differential (Shannon) entropy (3.13) subject to the empirical moments of the data. This yields the most conservative (closest to uniform) post data density. In that way the BMOM uses only assumptions on the realized error terms which are used to derive the post data density. Stating it differently, under the BMOM one gets around the need to specify priors by deriving the posterior directly from an ME argument. To do so, the BMOM equates the posterior expectation of a function of the parameter to its sample value and chooses the posterior to be the (differential) maximum entropy distribution subject to that constraint.

Consider the linear model $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ and define $\text{Data} \equiv (\boldsymbol{y}, X)$. The first assumption out of the two basic assumptions is $X'\text{E}[\boldsymbol{\varepsilon}|\text{Data}] = \mathbf{0}$. Thus, $\hat{\boldsymbol{\beta}}_{\text{LS}} = E(\boldsymbol{\beta}|\text{Data}) = (X'X)^{-1}X'\boldsymbol{y}$ so the posterior mean of $\boldsymbol{\beta}$ is the least squares estimate and $E(\boldsymbol{\varepsilon}|\text{Data}) = \hat{\boldsymbol{\varepsilon}}$. The second assumption is $\text{Var}(\boldsymbol{\varepsilon}|\sigma^2, \text{Data}) = \sigma^2 X(X'X)^{-1}X'$ where σ^2 is the error variance. Thus, $\text{Var}(\boldsymbol{\beta}|\sigma^2, \text{Data}) = \sigma^2(X'X)^{-1}$ and finally, $\text{Var}(\boldsymbol{\beta}|\text{Data}) = s^2(X'X)^{-1}$, where s^2 is an estimator of σ^2. With these assumptions, and under that framework, the posterior mean and variance of $\boldsymbol{\beta}$ are the traditional estimates in large sample Bayesian approaches.

Going back to the ME principle, if one maximizes the (differential) entropy of the posterior density subject to the above two moment constraints (and the requirement of a proper density function) on $\boldsymbol{\beta}$, then $E(\boldsymbol{\beta}|\text{Data}) = (X'X)^{-1}X'\boldsymbol{y}$ and $\text{Var}(\boldsymbol{\beta}|\text{Data}) = s^2(X'X)^{-1}$. Then, recalling our earlier derivation that the normal distribution is the maximum entropy under the first two moments (Section 4.2), we get $\boldsymbol{\beta} \sim N(\hat{\boldsymbol{\beta}}, s^2(X'X)^{-1})$. Interestingly enough, this is the exact same density for $(\boldsymbol{\beta}|\text{Data})$ one gets from the usual analysis. Thus, Zellner was able to show that the usual analysis is optimal in a maximum entropy sense. Note that this is called the BMOM because the moment constraints enter under the posterior distribution. The BMOM coincides with the Bayesian posterior under a normal likelihood and diffuse

(non-informative) prior. For recent discussion of Bayesian methods and information processing see Zellner (2007) and Clarke (2007). They also provide a discussion on the choice of moments and other BMOM examples where more than just two moments, or other transformations on the parameters, are used.

5.6 Discussion

So far the generic IT estimator and some specific methods within the IT family of estimation rules were discussed. The connection to information theory was developed and discussed. All the methods discussed so far have one thing in common: the moment conditions used are zero-moment conditions, $g_m(\boldsymbol{y}, \boldsymbol{p}, \boldsymbol{\theta}_1) = [\mathbf{0}]$. However, different values of α (in the objective function) are used within both the discrete and differential entropies. In the next section I discuss another member of that family that is different than the above methods by treating all moment conditions as stochastic.

6

Information-Theoretic Methods of Estimation — II: Stochastic Moments

6.1 Generalized Maximum Entropy — Basics

A basic property of the classical ME (or CE) approach, as well as the EL, GEL, GMM, and BMOM, is that the moment conditions have to be exactly fulfilled (zero-moment conditions). This property is satisfactory for (relatively) large samples or for well behaved samples. Unfortunately, in both the social and natural sciences we are often trying to understand small and/or ill-behaved (and often non-experimental) data where the zero-moments' restrictions may be too costly. Another basic concern for researchers is how to incorporate in the estimation procedure information resulting from economic-theoretic behavior such as agents' optimization. As was discussed in Section 5.4 above, a main advantage of the GMM is that it is easily connected to economic theory where the zero-moment conditions may be formulated as functions of agents' optimization. But even within the GMM framework, as well as within the other IT estimators discussed in Section 5, the task of incorporating that information may become very complicated, or even impossible. A main reason for that is the need to incorporate all of that information in terms of (zero) moment conditions or in terms of restrictions on these moments.

In this section, another member of the IT class of estimators is discussed. This estimator accommodates for the basic two problems raised above. It uses a more flexible set of moment conditions in the optimization. This provides a greater flexibility resulting in more stable estimates for finite and/or ill-behaved data and provides the researcher with a general framework for incorporating economic theoretic and other behavioral information in a simple way that is consistent with information theory. This information can be in terms of linear, nonlinear or inequality functions and does not have to be formulated in terms of zero-moment functions. Generally speaking, the classical ME is reformulated with *stochastic* moment conditions. We define the term "stochastic moments" as moment conditions, or functions of the random variables and the unknown parameters, with additive terms that have expectation of zero. These moments, or functions, can be conditional or unconditional.

The stochastic moments can be introduced in two ways. First, by allowing for some additive noise (with mean zero) for each one of the moment conditions. Second, by viewing each observation as a noisy moment resulting from the same data generating process. In that view, each observed data point can be treated as a composite of two components: signal and noise. The Generalized Maximum Entropy (GME) model that was developed in the early 1990's has the above view in mind and treats the moments (or each observation) as stochastic.[1] For detailed background, development and applications of the GME see the text of Golan et al. (1996b). For more examples and properties see also Golan et al. (1996c, 1997) as well as earlier work by Golan and Judge (1992) and Miller (1994). To introduce the idea here, rather than repeat previous formulations and examples, I describe the GME for the traditional linear regression model and for the more general framework — the second case. Section 7 provides another example of the GME, connects it to stochastic moments within discrete choice modeling, and connects it with the ML method.

[1] For axiomatic derivation of the GME, which is an extension of the ME axiomatic literature, see Golan and Perloff (2002).

Consider the linear regression model with T observations and K explanatory variables,

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{6.1}$$

where $\boldsymbol{y}$ is a T-dimensional vector of observed random variable, X is a $T \times K$ matrix of exogenous variables, $\boldsymbol{\beta}$ is a K-dimensional vector of the unknown parameters that we want to recover from the data, and $\boldsymbol{\varepsilon}$ is a T-dimensional vector of the unobserved and unobservable random errors. In line with tradition, it is assumed that the set of true (unknown) parameters is bounded: $\boldsymbol{\beta} \in B$ where B is a convex set. The LS solution for that model: $\boldsymbol{\beta}_{\text{LS}} = (X'X)^{-1}X'\boldsymbol{y}$.

Rather than search for the point estimates $\boldsymbol{\beta}$, each β_k is viewed as the mean value of some well defined random variable $\boldsymbol{z}$. The unobserved error vector $\boldsymbol{\varepsilon}$ is also viewed as another set of unknowns, and similar to the signal vector $\boldsymbol{\beta}$, each ε_t is constructed as the mean value of some random variable $\boldsymbol{v}$. Similar to other IT estimators discussed earlier, our main objective here is to estimate the unknown $\boldsymbol{\beta}$ with minimal distributional assumptions. Under the GME framework, however, we estimate simultaneously the full distribution of each β_k and each ε_t (within their support spaces) with minimal distributional assumptions.

Without loss of generality, let each (k) element of $\boldsymbol{\beta}$ be bounded below by $\underline{z}_k$ and above by $\bar{z}_k$:

$$B = \{\boldsymbol{\beta} \in \Re^K | \beta_k \in (\underline{z}_k, \bar{z}_k),\ k = 1, 2, \ldots, K\}. \tag{6.2}$$

Let $\boldsymbol{z}_k$ be an M-dimensional vector $\boldsymbol{z}_k \equiv (\underline{\boldsymbol{z}}_k, \ldots, \bar{\boldsymbol{z}}_k)' = (z_{k1}, \ldots, z_{kM})'$ for all $k = 1, 2, 3, \ldots, K$, and Z is a $K \times M$ matrix consisting of the individual M-dimensional vectors $\boldsymbol{z}_k$ and the elements z_{km}. Let $\boldsymbol{p}_k$ be an M-dimensional proper probability distribution defined on the set $\boldsymbol{z}_k$ such that

$$\beta_k = \sum_m p_{km} z_{km} \equiv E_{\boldsymbol{p}_k}[\boldsymbol{z}_k] \quad \text{or} \quad \boldsymbol{\beta} = E_P[Z]. \tag{6.3}$$

In this formulation, the observed data, $\boldsymbol{y}$, are viewed as the mean process Z with a probability distribution P that is defined on the supports $\boldsymbol{z}_k$'s and is conditional on X. Thus, the econometrician chooses the support space $\boldsymbol{z}_k$, and then uses the data to estimate the P's which in turn

yield the point estimates ($\boldsymbol{\beta}$). Note that $\boldsymbol{\beta}$ can be constructed as the median or any other quantity of interest, within this framework.

Similarly, assume that $\boldsymbol{\varepsilon} \in V$ where V is a convex set that is *symmetric around zero*. As done with the $\boldsymbol{\beta}$'s above, each error term is redefined as

$$\varepsilon_t = \sum_j w_{tj} v_j \equiv E_{w_t}[\boldsymbol{v}] \quad \text{or} \quad \boldsymbol{\varepsilon} = E_W[V]. \tag{6.4}$$

That is, the observed errors are viewed as elements taken as random draws from a certain distribution with probability weights $\{w_{tj}\}$. The dimension of the matrices V and W is $T \times J$. For simplicity, the above support spaces are constructed as discrete and bounded. It is possible, within the same framework, to construct unbounded and continuous supports (Golan and Gzyl, 2002, 2003, 2006).

Having reformulated $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$, the linear model is specified as

$$y_t = \sum_{k=1}^{K}\sum_{m=1}^{M} z_{km} p_{km} x_{tk} + \sum_j v_j w_{tj}, \quad \text{or} \quad y = XE_P[Z] + E_W[V].$$

The GME estimation method is

$$\text{GME} = \begin{cases} \hat{\boldsymbol{p}} = \arg\max_{\boldsymbol{p},\boldsymbol{w}} \Big\{ \{H(\boldsymbol{p}) + H(\boldsymbol{w})\} \\ \qquad \equiv -\sum_k \sum_m p_{km} \log p_{km} - \sum_t \sum_j w_{tj} \log w_{tj} \Big\} \\ \text{s.t.} \\ y_t = \sum_{k=1}^{K} \sum_{m=1}^{M} z_{km} p_{km} x_{tk} + \sum_j v_j w_{tj} \\ \qquad\qquad\qquad (\text{or } \boldsymbol{y} = XE_P[Z] + E_W[V]), \\ \sum_m p_{km} = 1, \sum_j w_{tj} = 1. \end{cases} \tag{6.5}$$

The estimated probabilities for the signal vector $\boldsymbol{\beta}$ are

$$\hat{p}_{km} = \frac{\exp(-z_{km}\sum_t \hat{\lambda}_t x_{tk})}{\sum_m \exp(-z_{km}\sum_t \hat{\lambda}_t x_{tk})} \equiv \frac{\exp(-z_{km}\sum_t \hat{\lambda}_t x_{tk})}{\Omega_k(\hat{\boldsymbol{\lambda}})}, \tag{6.6}$$

and the estimated probabilities for the noise vector $\boldsymbol{\varepsilon}$ are

$$\hat{w}_{tj} = \frac{\exp(-\hat{\lambda}_t v_j)}{\sum_j \exp(-\hat{\lambda}_t v_j)} \equiv \frac{\exp(-\hat{\lambda}_t v_j)}{\Psi_t(\hat{\boldsymbol{\lambda}})}. \tag{6.7}$$

The estimated values of $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$ are

$$\hat{\beta}_k \equiv \sum_m z_{km}\hat{p}_{km} \tag{6.8}$$

and

$$\hat{\varepsilon}_t \equiv \sum_j v_j \hat{w}_{tj}. \tag{6.9}$$

As with the ME, it is possible to transform the primal optimization GME model to a dual, concentrated model which is a function of the Lagrange multipliers $\boldsymbol{\lambda}$ (Golan and Judge, 1992; Golan et al., 1996b):

$$\begin{aligned}\underset{p\in P, w\in W}{\text{Max}} H(P,W) &= \underset{\boldsymbol{\lambda}\in D}{\text{Min}} \left\{ \sum_t y_t\lambda_t + \sum_k \log\Omega_k(\boldsymbol{\lambda}) + \sum_t \log\Psi_t(\boldsymbol{\lambda}) \right\} \\ &= \underset{\boldsymbol{\lambda}\in D}{\text{Min}} \left\{ \sum_t y_t\lambda_t + \sum_k \log\left[\sum_m \exp(-z_{km}\sum_t \lambda_t x_{tk})\right] \right. \\ &\qquad \left. + \sum_t \log\left[\sum_j \exp(-\lambda_t v_j)\right] \right\}.\end{aligned}$$

The concentrated model is solved by minimizing with respect to $\boldsymbol{\lambda}$, to obtain the optimal $\boldsymbol{\lambda}$. The optimal $\boldsymbol{\lambda}$ is then used to obtain the P's via Eq. (6.6), with which the set of $\boldsymbol{\beta}$'s is recovered via Eq. (6.8). The Hessian matrix of the GME problem is negative definite for $P, W \gg 0$ and thus satisfies the sufficient condition for a unique global minimum (Golan et al., 1996b).[2]

The GME minimizes the joint entropy distance between the data and the state of complete uncertainty (the uniform distribution). It is a dual-loss function that assigns equal weights to prediction and precision.[3] Equivalently, it can be viewed as a shrinkage estimator that shrinks the data to the priors (uniform distributions) and toward the center of their supports.

[2] Here, the GME is expressed in term of discrete and bounded support spaces. For specification of continuous and unbounded supports see Golan and Gzyl (2002, 2006).

[3] The objective function puts equal weights on the $\boldsymbol{p}$ and $\boldsymbol{w}$ entropies. If the researcher prefers to use unequal weights, it is possible. See Golan et al. (1996c) for a discussion of why unequal weights might be used and the implications for estimation.

The estimated probabilities provide the full distribution (within its pre-specified support) of each one of the parameters of interest ($\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$). The $\boldsymbol{\beta}$'s are direct functions of the Lagrange multipliers ($\boldsymbol{\lambda}$). These multipliers reflect the marginal information of each observation. Like the EL (and GEL) they capture that natural weight of each observation and convey that information in the estimated exponential distributions $\hat{p}_{km}$ and $\hat{w}_{tj}$. This is shown in Section 6.5.

Like all IT methods, the available information is represented here as constraints in the basic (primal) constrained optimization model. There is no additional (hidden) information or implicit assumptions. Once this optimization is solved, the concentrated (unconstrained) model is constructed. As discussed previously, the transformation from the primal to the dual, concentrated model means that the problem is transformed from the probability space to the Lagrange multipliers' space. This means that the complexity of our GME model is not changed as the number of support points increases. This is because the real parameters here are the Lagrange multipliers. Finally, if additional information, such as information on the underlying distribution, the covariance structure or economic-theoretic information is available, it can be easily incorporated within this framework. This is shown in Section 6.2. Two applications of the GME are now presented.

Example (*Stochastic Moments*). Consider now the case where rather than optimizing with respect to each observed data points (method 6.5 or its concentrated version), one optimizes with respect to the observed moments, but in line with the GME method, these moments are viewed as stochastic (zero moments with additive noise):

$$\sum_{t} x_{tk} y_t = \sum_{t,k,m} x_{tk} x_{tk} z_{km} p_{km} + \sum_{t,j} x_{tk} v_j w_{tj},$$

where the errors' support space V shrinks to zero as the sample size increases (Golan et al., 1996b, Chap. 6.6). In line with the EL linear model of Section 5.3, the above can be written as

$$\sum_{t=1}^{T} \boldsymbol{x}_t \left(y_i - \sum_{k,m} x_{tk} z_{km} p_{km} - \sum_{j} v_j w_{tj} \right) = \mathbf{0}.$$

The (Relaxed) Stochastic Moments GME is

$$\text{GME} = \begin{cases} \hat{\boldsymbol{p}} = \arg\max_{\boldsymbol{p},\boldsymbol{w}} \Big\{ \{H(\boldsymbol{p}) + H(\boldsymbol{w})\} \\ \quad \equiv -\sum_k \sum_m p_{km} \log p_{km} - \sum_t \sum_j w_{tj} \log w_{tj} \Big\} \\ \text{s.t.} \\ \sum_{t=1}^{T} \boldsymbol{x}_t \left(y_i - \sum_{k,m} x_{tk} z_{km} p_{km} - \sum_j v_j w_{tj} \right) = \mathbf{0} \\ \sum_m p_{km} = 1, \ \sum_j w_{tj} = 1. \end{cases} \tag{6.5a}$$

The estimated probabilities for the signal vector $\boldsymbol{\beta}$ are

$$\hat{p}_{km} = \frac{\exp\left(-z_{km}\hat{\lambda}_k \sum_t x_{tk} x_{tk}\right)}{\sum_m \exp\left(-z_{km}\hat{\lambda}_k \sum_t x_{tk} x_{tk}\right)} \equiv \frac{\exp\left(-z_{km}\hat{\lambda}_k \sum_t x_{tk} x_{tk}\right)}{\Omega_k(\hat{\boldsymbol{\lambda}})},$$

where now the dimension of the Lagrange multipliers $\boldsymbol{\lambda}$ is K. The estimated probabilities for the noise vector $\boldsymbol{\varepsilon}$ are

$$\hat{w}_{tj} = \frac{\exp\left(-v_j \sum_k \hat{\lambda}_k x_{tk}\right)}{\sum_j \exp\left(-v_j \sum_k \hat{\lambda}_k x_{tk}\right)} \equiv \frac{\exp\left(-v_j \sum_k \hat{\lambda}_k x_{tk}\right)}{\Psi_t(\hat{\boldsymbol{\lambda}})},$$

and finally, $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\varepsilon}}$ are found via Eqs. (6.8) and (6.9). Following previous discussion, the concentrated model (as a function of the $K\boldsymbol{\lambda}$'s) is:

$$\begin{aligned} \underset{p\in P, w\in W}{\text{Max}} H(P,W) &= \underset{\boldsymbol{\lambda}\in D}{\text{Min}} \left\{ \sum_{t,k} y_t x_{tk} \lambda_k + \sum_{\text{k}} \log\Omega_{\text{k}}(\boldsymbol{\lambda}) + \sum_t \log \Psi_t(\boldsymbol{\lambda}) \right\} \\ &= \underset{\boldsymbol{\lambda}\in D}{\text{Min}} \left\{ \sum_t y_t x_{tk} \lambda_k + \sum_{\text{k}} \log \left[\sum_m \exp\left(-z_{km}\lambda_k \sum_t x_{tk} x_{tk} \right) \right] \right. \\ &\qquad \left. + \sum_t \log \left[\sum_j \exp\left(-\sum_k \lambda_k v_j x_{tk} \right) \right] \right\}. \end{aligned} \tag{6.10}$$

(Golan et al., 1996b, Chap. 6) developed this class of estimation methods. They discuss different ways to incorporate the sample data

in the GME optimization. Examples include the individual noisy observation shown in model 6.5, the stochastic moments' case shown in above example, and weighted stochastic moments. They also develop the large sample results. It is just noted here, that it is immediate to see that for (6.10 or 6.5a), as $T \to \infty$, $\hat{\boldsymbol{\beta}}_{\text{GME}} = \hat{\boldsymbol{\beta}}_{\text{LS}}$ as long as the support space Z spans the true values of $\boldsymbol{\beta}$. See Golan et al. (1996b) for small and large sample properties of the GME. For additional results and applications see also Golan and Gzyl (2006). For completion, the above model is formulated below in terms of our six-sided die example.

Example (*GME and the Dice Problem*). Within the GME, the (stochastic first moment) die problem is

$$\underset{\{\boldsymbol{p},\boldsymbol{W}\}}{\text{Max}} \; H(\boldsymbol{P},\boldsymbol{W}) = -\sum_{k=1}^{6} p_k \log p_k - \sum_{j=1}^{J} w_j \log w_j$$

s.t

$$\sum_k p_k x_k + \sum_j w_j v_j = y, \quad \sum_k p_k = 1, \text{ and } \sum_j w_j = 1,$$

where $x_k = 1, 2, \ldots, 6$, $y \in (1, 6)$, $\boldsymbol{v}$ is a J-dimensional support space and the natural log is used. The concentrated model is

$$\begin{aligned} \ell(\lambda) &= \lambda y + \log\left[\sum_k \exp(-\lambda x_k)\right] + \log\left[\sum_j \exp(-\lambda v_j)\right] \\ &= \lambda y + \log\Omega + \log\Psi. \end{aligned}$$

Though this example may seem very simple, it does reflect a wide range of problems. For example, consider two separate samples with two different means. In that case, the ME cannot be used (with a very high probability there is no unique probability distribution that satisfies the two sets of moments). However, under the GME framework a solution always exists. In a related context, consider the work of Hellerstein and Imbens (1999) on combining data sets. They are using a GMM framework to solve that estimation problem. The above example shows that the GME approach can also be used for estimating the same types of problems.

Finally, we refer the reader to a less known method called the maximum entropy on the mean (MEM) method (Bercher et al., 1996; Besnerais et al., 1999; Gamboa and Gassiat, 1997; Gzyl, 1993). This method is an extension of the classical ME method and closely related to the GME. In fact, the MEM can be constructed as a special case of the GME with non-stochastic moments. This method is not discussed further in that review.

6.2 GME — Extensions: Adding Constraints

Equality, inequality or nonlinear restrictions can be imposed within the GME model. To do so, one needs to incorporate these restrictions within the IT–GME optimization model. These restrictions reflect additional information the researcher possess and then test. This additional information can be incorporated within the IT–GME framework only due to the fact that all moments/data enter as stochastic. This stochastic nature allows for the additional freedom that does not exist within the zero-moment IT methods discussed previously. There are many applications where constraints are incorporated into the model where these constraints reflect theoretical information resulting from economic behavior and/or statistical information related to the nature of the randomness. See for example Golan et al. (1996b), or a collection of applications in industrial organization (Perloff et al., 2007).

Consider the following example to correct for the statistical nature of the data. Assume our data exhibit first-order autocorrelation: $\varepsilon_t = \rho\varepsilon_{t-1} + \varpi_t$, where ρ is the autocorrelation coefficient with $\rho \in (-1, 1)$ and ϖ is a vector of independently and identically distributed errors with mean zero. That additional set of restrictions can be incorporated in model (6.5) so that if there is first-order autocorrelation in the data, our model will capture it; however, the "model" does not force that correlation to exist and bias the estimates in the absence of autocorrelation. The GME with first-order autocorrelation version of (6.5) is

$$\underset{\{\boldsymbol{p},\boldsymbol{w},\rho\}}{\text{Max}} \left\{ -\sum_{k,m} p_{km} \log p_{km} - \sum_{tj} w_{tj} \log w_{tj} \right\}$$

s.t.

$$y_t = \sum_{k=1}^{K}\sum_{m=1}^{M} z_{km} p_{km} x_{tk} + \varepsilon_t \quad (\text{or } \boldsymbol{y} = X E_P[Z] + \boldsymbol{\varepsilon})$$

$$\varepsilon_t = \varpi_t = \sum_j w_{tj} v_j \quad \text{for } t = 1$$

$$\varepsilon_t = \rho\varepsilon_{t-1} + \varpi_t = \rho\varepsilon_{t-1} + \sum_j w_{tj} v_j \quad \text{for } t = 2, 3, \ldots, T$$

$$\sum_m p_{km} = 1; \quad \sum_j w_{tj} = 1.$$

It is worth noting here, that under the above framework, autocorrelation is not forced on the model, but rather it is picked up if it exists in the data. In a similar way, all other covariance structures, such as higher order autocorrelations and heteroskedasticity can be captured. Soft data coming from economic theory, such as game theoretic restrictions, can be incorporated in the same way. In that case, it is first needed to construct the optimal (economic) conditions and then incorporate them as stochastic version of these conditions within the optimization model (e.g., Golan et al., 2000; Golan, 2001). In both cases (statistical and theoretic conditions), the constraints provide additional information that reduces the feasible solution space. Tests to verify the relationship between these additional restrictions and the data are discussed below. For examples of GME with nonlinear constraints see Golan et al. (1996a). For examples of GME with inequality constraints see Golan et al. (1997) and Golan et al. (2001).

6.3 GME — Entropy Concentration Theorem and Large Deviations

Since the GME is a natural extension of the classical ME, it is possible to build on the ME results and provide more motivations and further results for the GME. I concentrate here on two basic ideas: The Entropy Concentration Theorem, ECT, and Large Deviations.

The ECT, discussed in Section 4.3 for the classical ME is easily extended to the GME and provides another powerful motivation for

that method. Let $H^*(\hat{\boldsymbol{p}}, \hat{\boldsymbol{w}}) \equiv H(\hat{\boldsymbol{p}}) + H(\hat{\boldsymbol{w}})$ be the entropy value for model (6.5) or its concentrated counterpart model. Let $\mathscr{P}$ be the subclass of all possible outcomes that could be observed from our sample data that satisfy (i) the constraints (stochastic moments) and/within (ii) the support spaces Z and V (in Eq. (6.5)). The ECT, applied to the GME, states that a significant percentage of outcomes in the class $\mathscr{P}$ (and within the pre-specified support spaces Z and V) will have an entropy in the range

$$\begin{aligned} H^* - \Delta H \leq H(\boldsymbol{p}, \boldsymbol{w}) \leq H^* &\equiv \underset{p \in \mathscr{P}}{\text{Max}} H(\text{Eq. (6.5)}) \\ &= H^*(\hat{\boldsymbol{p}}, \hat{\boldsymbol{w}}) \equiv H^*(\hat{\boldsymbol{p}}) + H^*(\hat{\boldsymbol{w}}), \end{aligned}$$

where $\Delta H \equiv \chi^2_{(C;\alpha)} \Big/ 2T$, T is the number of observations, α is the upper percentile of the χ^2 distribution with C degrees of freedoms (which changes based on the structure of the model). Other, atypical distributions $\{p_{km}, w_{tj}\}$, that are conditional on Z and V and are consistent with the constraints (sample data), will have entropy levels smaller than H^* and their concentration near this upper bound is given by the above ECT.

Example (*Calculating* ΔH). For the stochastic moment problem (6.5a) $0 \leq H(\boldsymbol{p}, \boldsymbol{w}) \leq K \log M + T \log J$. There are K stochastic moments (K basic unknown parameters, or in terms of probabilities there are $KM + TJ$ unknowns and $2K + T$ constraints), so $C = KM + TJ - (2K + T) = K(M - 2) + T(J - 1)$ and $2T\Delta H = \chi^2_{(C;\alpha)}$. For the stochastic data problem (6.5) the bounds on the entropy are the same: $0 \leq H(\boldsymbol{p}, \boldsymbol{w}) \leq K \log M + T \log J$. However, there are T stochastic constraints in this case, so $C = K(M - 1) + T(J - 2)$.

A brief discussion of Large Deviations interpretation for the GME model is now provided. (Note however, that a complete formal model is not developed here.) For $\boldsymbol{\beta} = \boldsymbol{f}(\boldsymbol{y}, X|Z, V)$ we want to find the probability that the estimated parameters (conditioned on the observed data) have values similar to the center of their supports, meaning all prior probabilities on these supports are uniform. Following Section 3 (and examples in Section 4), define the set E as a set of all proper probability distributions satisfying the observed moment conditions represented

explicitly in (6.5), or (6.5a):

$$E \equiv \left\{ P,W|Z,V \text{ and } \sum f_s(\boldsymbol{y},X) \geq c_s,\ s=1,\ldots,S \right\},$$

where $S=T$ for the stochastic data model (6.5) and $S=K$ for the stochastic moments model (6.5a). To find the closest distribution to the prior distribution $\{P^0,W^0\}$, minimize $D(P,W||P^0,W^0)$ subject to the observed stochastic sample's data (or moments) as specified by (6.5) and (6.5a). This yields the estimated probabilities $\hat{p}_{km}$ and $\hat{w}_{tj}$ (together with $\hat{\boldsymbol{\lambda}}$). These are the estimated probabilities, conditionals on Z and V, that satisfy the observed data and that are closest (in entropy) to the priors. For example, for the uniform priors ($p^0_{km}=\frac{1}{M}$ for $m=1,\ldots,M$ and $w^0_{tj}=\frac{1}{J}, j=1,\ldots,J$), $2^{-(T-1)D(\hat{P},\hat{W}||P^0,W^0)}$ is the probability that conditional on the observed data (and Z and V), all estimated probabilities are uniform, and therefore the estimated $\boldsymbol{\beta}$ are all at the center of their Z support spaces and all the $\boldsymbol{\varepsilon}$'s are zeros. For example, if all signal supports Z are symmetric about zero, the above describes the probability that all $\boldsymbol{\beta}$'s are zero.

6.4 GME — Inference and Diagnostics

There are a number of basic statistics and diagnostics available for the GME. Most of these measures are just applications and extensions of methods discussed in earlier sections. These statistics are part of the output provided in SAS, LIMDEP, SHAZAM, and other software that include the GME procedure.

Because the GME is an IT method, it makes sense to start with the information measures known as the normalized entropy measures. These measures quantify the relative informational content in the data. For each variable k, the entropy measure is a continuous function from zero to $\log(M)$. For convenience in making comparisons, these entropies are normalized to the zero–one interval. The normalized entropy measures $S(\cdot)$ for the GME model is

$$S(\hat{\boldsymbol{p}}) = \frac{-\sum_{k,m} \hat{p}_{km} \log \hat{p}_{km}}{K \log M},$$

where $S(\hat{\boldsymbol{p}}) \in [0,1]$. Note that this measure is conditional on the choice of Z. This measure is further discussed in the context of discrete choice models in the next section.

Similarly, the same measures can be used to evaluate the information in each one of the variables $k = 1, 2, \ldots, K$:

$$S(\hat{p}_k) = \frac{-\sum_m \hat{p}_{km} \log \hat{p}_{km}}{\log M}.$$

These variable-specific information measures reflect the relative contribution (of explaining the dependent variable) of each one of the independent variables. The above normalized entropy measure can be connected to Fano's inequality and the probability of errors (Section 3.6). An example and further discussion is provided in Section 7.5.

The asymptotic variance of the GME model is

$$\text{Var}(\hat{\beta}) = \frac{\sigma^2(\hat{\beta})}{\varpi^2(\hat{\beta})}(X'X)^{-1},$$

where

$$\sigma^2(\hat{\beta}) \equiv \frac{1}{T}\sum_i \hat{\lambda}_i^2$$

and

$$\varpi^2(\hat{\beta}) \equiv \left[\frac{1}{T}\sum_i \left(\sum_j v_{ij}^2 \hat{w}_{ij} - \left(\sum_j v_{ij}\hat{w}_{ij}\right)^2\right)^{-1}\right]^2.$$

The Entropy Ratio (ER) test — which corresponds to the likelihood ratio, or empirical ratio, test[4] — measures the entropy discrepancy between the constrained (say, $\boldsymbol{\beta} = \boldsymbol{\beta}_0$) and the unconstrained ($\hat{\boldsymbol{\beta}}$) models:

$$\text{ER} = \frac{2\varpi^2(\hat{\beta})}{\sigma^2(\hat{\beta})}|H_U(\hat{\beta}) - H_R(\beta = \beta_0)| \cong 2|H_U(\hat{\beta}) - H_R(\beta = \beta_0)|,$$

[4] See also discussion in Section 5. Many of these tests are similar to other tests within the IT family of estimators reflecting entropy differences between the unconstrained and constrained cases as reflected by the null hypotheses.

where H_R is the restricted hypothesis and H_U is the unrestricted hypothesis. For example, the entropy-ratio statistic for testing the null hypothesis (H_0) that all parameters are zero is

$$\mathrm{ER}(\boldsymbol{\beta} = \mathbf{0}) = 2H_R(\boldsymbol{\beta} = 0) - 2H_U(\hat{\boldsymbol{\beta}}).$$

Under certain regularity assumptions (see for example, Owen, 1990; and Qin and Lawless, 1994; Golan and Gzyl, 2006; Mittelhammer and Cardell, 1996), $\mathrm{ER}(\boldsymbol{\beta} = \mathbf{0}) \rightarrow \chi^2_K$ as $T \rightarrow \infty$, when the restriction is true and K is the number of restrictions. The approximate α -level confidence interval for the estimates is obtained by setting $\mathrm{ER}(\bullet) \leq C_\alpha$, where C_α is chosen so that $\Pr(\chi^2_K < C_\alpha) = \alpha$ where C_α is the critical value of the χ^2_K statistic (with K degrees of freedom) at a significance level of α. Similarly, it is possible to test any other hypothesis of the form $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ for all, or any subset, of the parameters.

A goodness of fit measure for the GME estimator is

$$R^* = 1 - \frac{H_U(\hat{\boldsymbol{\beta}})}{H_R(\boldsymbol{\beta} = \mathbf{0})},$$

where $R^* = 0$ implies no informational value of the dataset, and $R^* = 1$ implies perfect certainty or perfect in-sample prediction. This measure, R^*, is the same as the information index of Soofi (1996) and is directly related to the normalized entropy measure (Golan et al., 1996b) and to Fano's inequality (probability of error).

The Wald Test (WT) to examine hypotheses about possible convex combinations of $\boldsymbol{\beta}$ can also be used. For the null hypothesis H_0: $L\boldsymbol{\beta} = c$, where L is a set of linearly independent combinations of some, or all, of the $\boldsymbol{\beta}$'s and c is a specific value (such as zero), the Wald Statistic is

$$WT = (L\boldsymbol{\beta} - c)'(L(\mathrm{Var}(\hat{\boldsymbol{\beta}}))L')^{-1}(L\boldsymbol{\beta} - c),$$

which, under H_0 has a central χ^2 with degrees of freedom equal to $\mathrm{Rank}(L)$. All of these tests can be performed directly on the Lagrange multipliers $\boldsymbol{\lambda}$, which determine $\boldsymbol{\beta}$ in the GME framework.

Additional tests, like the tests for over-identifying restrictions (Section 5.4.3), can be formulated here as well within the same framework.

6.5 GME — Further Interpretations and Motivation

6.5.1 Independence of Signal and Noise

Using the objective functional $H(\boldsymbol{p},\boldsymbol{w}) = H(\boldsymbol{p}) + H(\boldsymbol{w})$ in our basic IT–GME model (6.5) means one uses the traditional assumption of independence between the signal and noise. It is important to confirm that the post-data estimates obey the same property. To do so, rather than starting with the above separable function $H(\cdot)$, it is possible to define the entropy objective on a joint (support) set B^*. Let

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} = [X, I]\begin{pmatrix}\boldsymbol{\beta}\\ \boldsymbol{\varepsilon}\end{pmatrix} \equiv X^*\boldsymbol{\delta}. \tag{6.11}$$

As before, we view $\boldsymbol{\beta} = E_P[\boldsymbol{z}]$ and $\boldsymbol{\varepsilon} = E_P[\boldsymbol{v}]$ where $\boldsymbol{z} \in B$, $\boldsymbol{v} \in V$ so $B^* = B \times V$.

Let Q be the prior information, within the relevant subspaces, so $dQ(\delta) = dQ_\beta(z) \otimes dQ_\varepsilon(v)$ where the priors are taken to be continuous uniform distributions defined on the lower and upper bounds.[5] The continuous/differential version of the IT–GME solution is

$$dP_{\text{GME}}(\boldsymbol{\delta}, \tilde{\boldsymbol{\lambda}}) = \frac{e^{\langle -\tilde{\boldsymbol{\lambda}}, X^*\boldsymbol{\delta}\rangle}}{\Omega(\tilde{\boldsymbol{\lambda}})} dQ(\boldsymbol{\delta}), \tag{6.12}$$

where $\langle \boldsymbol{a}, \boldsymbol{b}\rangle$ denotes the Euclidean scalar (inner) product of vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ and $\Omega(\cdot)$ is the partition function. Letting $\boldsymbol{\alpha} = X'\boldsymbol{\beta}$, and omitting "$\sim$" for simplicity, the general partition function (over B^*) is

$$\begin{aligned}\Omega(\lambda) &= \int_{B^*} e^{-\langle X'\lambda, \delta\rangle} dQ(\delta) = \int_B \int_V e^{-\langle \alpha, \delta\rangle} dQ_\beta(z) dQ_\varepsilon(v)\\ &= \int_B e^{-\langle \alpha, \boldsymbol{z}\rangle} dQ_\beta(z) \int_V e^{-\langle \lambda, \boldsymbol{v}\rangle} dQ_\varepsilon(v) = \Omega_\beta(\lambda)\Psi_\varepsilon(\lambda).\end{aligned} \tag{6.13}$$

Finally,

$$dP_{\text{GME}}(\boldsymbol{\delta}, \boldsymbol{\lambda}) = \frac{e^{-\langle \lambda, XZ\rangle}}{\Omega_\beta(\lambda)} dQ_\beta(z) \frac{e^{-\langle \boldsymbol{\lambda}, \boldsymbol{v}\rangle}}{\Psi_\varepsilon(\lambda)} dQ_\varepsilon(\boldsymbol{v}) = dP_\beta(z) dP_\varepsilon(v). \tag{6.14}$$

[5] Note that this formulation is more general and allows us to couch all the different data representations under a single model (see Golan et al., 1996b, Chap. 6, and Golan and Gzyl, 2002, 2003, 2006).

Starting with the general case of P_{GME}, we ended up with the two distinct elements $dP_\beta(z)$ and $dP_\varepsilon(v)$, implying the signal and noise are independent and thus the GME rule does not violate the basic *a priori* independence assumption used in the linear model (6.1).

6.5.2 The Continuous Limit

The continuous limit of the GME is now investigated. To do so, the behavior of the GME at the limit of $M \to \infty$ and $J \to \infty$ is studied, while holding everything else unchanged. This is related to the notion of *super-resolution* (e.g., Bercher et al., 1996; Besnerais et al., 1999; Gamboa and Gassiat, 1997).[6] Within our previous definition, let $\underline{z}_k \leq \beta_k \leq \bar{z}_k$ and let $B = \times_{k=1}^{K}[\underline{z}_k, \bar{z}_k]$.

Assuming uniform *a priori* information (within B) for each covariate, implies

$$dQ(\zeta) = \overset{K}{\underset{k=1}{\otimes}} \frac{d\zeta_j}{(\bar{z}_k - \underline{z}_k)}, \tag{6.15}$$

where, as above, Q reflects our prior knowledge. The post-data P is $dP(\zeta) = \rho(\zeta)dQ(\zeta)$. Maximizing the differential relative entropy (or similarly minimizing the cross-entropy)

$$D(P||Q) = \int_B \rho(\zeta)\log\rho(\zeta)dQ(\zeta)$$

subject to the data and the normalization requirements yields

$$dP_\lambda(\zeta) = \frac{e^{-\langle \boldsymbol{\lambda}, X^*\zeta\rangle}}{\Omega(\boldsymbol{\lambda})} dQ(\zeta). \tag{6.16}$$

Finally, using the same notations as in Eq. (6.13), the partition function is

$$\begin{aligned}\Omega(\boldsymbol{\lambda}) &= \int_B e^{-\langle \boldsymbol{\lambda}, X\zeta\rangle} dQ(\zeta) = \int_B e^{-\langle X'\boldsymbol{\lambda}, \zeta\rangle} dQ(\zeta) \\ &= \prod_{k=1}^{K} \int_{\underline{z}_k}^{\bar{z}_k} e^{-\alpha_k\zeta_k} \frac{d\zeta_k}{\bar{z}_k - \underline{z}_k} = \prod_{k=1}^{K} \frac{e^{-\alpha_k\underline{z}_k} - e^{-\alpha_k\bar{z}_k}}{\alpha_k(\bar{z}_k - \underline{z})}.\end{aligned} \tag{6.17}$$

[6] Viewing this problem as inherently ill-posed this limit can be thought of as a certain type of "consistency" which is different than the traditional consistency of investigating the properties as the number of observations increases.

The continuous partition function for each $k = 1, 2, \ldots, K$ is

$$\Omega_k(\hat{\lambda}) = \frac{e^{-\hat{\alpha}_k \underline{z}_k} - e^{-\hat{\alpha}_k \bar{z}_k}}{\hat{\alpha}_k(\bar{z}_k - \underline{z}_k)}, \tag{6.18}$$

and the point estimate, $\hat{\beta}_k$, is

$$\begin{aligned}\hat{\beta}_k &= \frac{1}{(\bar{z}_k - \underline{z}_k)} \left(-\frac{\partial}{\partial \hat{\alpha}_k}\right) \int_{\underline{z}_k}^{\bar{z}_k} e^{-\hat{\alpha}_k \zeta_k} d\zeta_k \\ &= \frac{1}{(\bar{z}_k - \underline{z}_k)} \left\{ \frac{\underline{z}_k e^{-\hat{\alpha}_k \underline{z}_k} - \bar{z}_k e^{-\hat{\alpha}_k \bar{z}_k}}{\hat{\alpha}_k} + \frac{e^{-\hat{\alpha}_k \underline{z}_k} - e^{-\hat{\alpha}_k \bar{z}_k}}{\hat{\alpha}_k^2} \right\}. \end{aligned} \tag{6.19}$$

Similarly, the continuous version of the partition function for the noise terms is

$$\Psi_i(\hat{\lambda}) = \frac{e^{-\hat{\lambda}_i \underline{v}} - e^{-\hat{\lambda}_i \bar{v}}}{\hat{\lambda}_i(\bar{v} - \underline{v})} \tag{6.20}$$

and

$$\hat{e}_i = \frac{1}{(\bar{v} - \underline{v})} \left\{ \frac{\underline{v} e^{-\hat{\lambda}_i \underline{v}} - \bar{v} e^{-\hat{\lambda}_i \bar{v}}}{\hat{\lambda}_i} + \frac{e^{-\hat{\lambda}_i \underline{v}} - e^{\hat{\lambda}_i \bar{v}}}{\hat{\lambda}_i^2} \right\}. \tag{6.21}$$

Within the GME model (6.5), for a given sample and a given prior support space B^*, as the number of elements in this support goes to infinity ($M \to \infty, J \to \infty$) the amount of information approaches its upper bound. This implies, that for a given data set and a given $B^* = B \times V$, the continuous version of the GME yields (on average) the best possible estimates for this type of estimation rule. Nevertheless, it is important to note that it is possible (in some special cases) for the GME with a finite number of support points to be superior (in terms of MSE) to the continuous GME version as long as the number of support points is greater than two. For related work on the continuous approximation see LaFrance (1999) for the uniform case and Golan and Gzyl (2002, 2006) for the other distributions.[7]

[7] In a different setup, and within the MEM approach, Gamboa and Gassiat (1997) show that for the case of ill-posed noisy moment problems, as the number of discrete support points goes to infinity the MEM and Bayes approaches converge to the same rule. However, their result is quite different than the one presented here for the GME. Specifically, with

6.5.3 The Natural Weights

Within the GME model (6.5), the "natural weight" of each observation (or the "natural empirical distribution") is just a function of the T Lagrange multipliers

$$\pi_i(\hat{\lambda}) = \frac{\exp(-\hat{\lambda}_i)}{\sum_i \exp(-\hat{\lambda}_i)}. \tag{6.22}$$

As $T \to \infty$, $\boldsymbol{\pi} \to$ uniform distribution. Unlike the other IT estimators discussed earlier, under the GME method, the weights are direct function of the information in the data. Specifically, the estimated Lagrange parameters reflect the contribution (in terms of information) to the optimal level of the entropy (objective) and as such they capture the contribution of each observation to the "explanation" of the signal. The resulting empirical distribution, $\boldsymbol{\pi}$, is a function of these $\boldsymbol{\lambda}$'s.

6.5.4 Efficiency

Following our earlier discussion on efficient IPR's, it is possible to demonstrate that the GME is also a 100% efficient IPR. Recall that a 100% efficient IPR is one that satisfies the "information conservation principle" where the input information equals the output information. Thus, there is no loss of information in the inversion process.

For the GME (uniform priors), the two input information components are

$$\begin{aligned} \text{Input}_1 &\equiv -\sum_k \sum_m \tilde{p}_{km} \log q_{km} - \sum_i \sum_j \tilde{w}_{ij} \log w_{ij}^0 \\ &= K \log M + T \log J \end{aligned} \tag{6.23}$$

$$\begin{aligned} \text{Input}_2 &\equiv -\sum_{k,m} \tilde{p}_{km} \log[\Omega_k(\tilde{\boldsymbol{\alpha}})] - \sum_{t,j} \tilde{w}_{tj} \log[\Psi_t(\tilde{\boldsymbol{\lambda}})] \\ &= \sum_{k,m} \tilde{p}_{km} \tilde{\alpha}_k z_{km} + \sum_{t,j} \tilde{w}_{tj} \tilde{\lambda}_t v_j, \end{aligned} \tag{6.24}$$

---

the objective of solving an ill-posed noisy moments problem, they start with a continuous model that can be solved only by a special "discretization." Each one of these ill-posed "discretized" cases is solved by ME. Then, they investigate the limit of this set of solutions. See also the general discussion in O'Sullivan (1986).

where the RHS of (6.23) holds only for the uniform priors case. The output information components are

$$\text{Output}_1 \equiv -\sum_k \sum_m \tilde{p}_{km} \log \tilde{p}_{km} - \sum_i \sum_j \tilde{w}_{ij} \log \tilde{w}_{ij} \tag{6.25}$$

$$\text{Output}_2 \equiv -\sum_{k,m} \tilde{p}_{km} \Omega_k - \sum_{t,j} \tilde{w}_{tj} \log \Psi_t = -\sum_k \log \Omega_k - \sum_t \log \Psi_t. \tag{6.26}$$

To show 100% efficiency one needs to prove that the input to output ratio is one. Following our earlier formulation and taking into account the first-order conditions and noting that all the partition functions cancel each other, we have

$$\sum_i \tilde{\lambda}_i y_i - \sum_k \sum_m \tilde{p}_{km} z_{km} x_{ik} - \sum_i \sum_j \tilde{w}_{ij} v_j, \tag{6.27}$$

which must be zero for an optimal solution. This proves that the GME inversion procedure is a 100% efficient information processing rule. Though it is beyond the scope of this review, it is noted that if one's objective is to compare estimation rules in term of their *a priori* information (or set of basic assumptions), as was done earlier for the EL and ME, a possible way is to compare all 100% efficient IPR's in terms of their input information.

6.6 Numerical Example

Using simulated data for the linear model $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, this example presents a small number of numerical applications of estimation and inference for the GME method. It is not aimed at comparing methods of estimations, but rather as a brief tutorial. The data for that experiment were generated as follows: $\boldsymbol{x}_1 = \mathbf{1}$, $\boldsymbol{x}_2$ and $\boldsymbol{x}_3$ are taken from a Uniform(0, 20), and $\boldsymbol{x}_4 = \boldsymbol{x}_3 + \text{Normal}(0,\ 0.2)$, $\boldsymbol{\beta} = (1, -2, 3, 0)'$, and $\boldsymbol{\varepsilon}$ is iid Normal(0, 2). The support space $\boldsymbol{z}$ for each $k = 1, \ldots, 4$, with $M = 5$, is symmetric around zero $[\boldsymbol{z}_k = (-100, -50, 0, 50, 100)]$, and $J = 3$ for $\boldsymbol{v}$ with the traditional three empirical sigma rule (e.g., Pukelsheim, 1994).

Using uniform priors for both signal and noise the estimated $\boldsymbol{\beta}$'s for $T = 10$ are 1.898, −2.070, 1.364, and 1.609 for the GME and 3.116,

-2.130, 0.901, and 2.021 for the OLS. In both models the only coefficient that is statistically significant at the 5% level is the intercept. For $T = 1000$ the estimated $\boldsymbol{\beta}$'s are 1.278, -2.007, 3.160, and -0.16966 for the GME and 1.285, -2.008, 3.330, -0.3401 for the OLS. In this case all coefficients are significant.

Hypothesis A. Let $H_0 : \beta_4 = 0$ vs. $H_1 : \beta_4 \neq 0$. Using the ER test for $T = 10$:

$$\begin{aligned} \text{ER} &= 2[H^*(\text{unrestricted}) - H^*(\beta_4 = 0)] \\ &= 2(17.417 - 17.416) = 0.002 < \chi^2_{(1,0.05)} \end{aligned}$$

so the null hypothesis is not rejected.

Hypothesis B. Let $H_0 : \beta_2 \geq 0$ vs. $H_1 : \beta_2 < 0$. Using the ER test, for $T = 10$:

$$\text{ER} = 2(17.417 - 17.225) = 0.384 < \chi^2_{(1;0.05)}$$

so the null hypothesis is not rejected. Repeating hypothesis B for $T = 1000$, one gets

$$\text{ER} = 2(1,104.3 - 1,078.1) = 52.4 > \chi^2_{(1;0.055)}$$

and the null hypothesis is rejected.

In terms of large deviation (LD) analysis for the same hypothesis, one gets for $T = 10$

$$\text{Prob}(\beta_2 \geq 0) = 2^{-TD(\hat{P},\hat{W}||P^0,W^0)} = 2^{-10\times 0.199} \cong 0.252,$$

and for $T = 1000$

$$\text{Prob}(\beta_2 \geq 0) = 2^{-TD(\hat{P},\hat{W}||P^0,W^0)} = 2^{-1000\times 26.9} \cong 0.0000.$$

Both results are consistent with the ER (χ^2) tests above.

In terms of the Entropy Concentration Theorem, for $T = 10$, $H^* = 17.417$ and $(2 \times 10)\Delta H = \chi^2_{(C;\alpha)}$ with $C = K(M-1) + T(J-1) - T = 26$, so $\chi^2_{(C;\alpha)} = \chi^2_{(26;.05)} = 38.885$ so $\Delta H = 1.944$. Thus, 95% of possible distributions, consistent with the sample data and the support spaces Z and $\boldsymbol{v}$, have entropy in the interval [15.473, 17.417].

Similar calculations for $T = 1000$ yield $\chi^2_{(C;\alpha)} = \chi^2_{(1016;.05)} \cong 43.8$ and $\Delta H = 43.8/2000 = 0.022$, so 95% of sample's consistent distributions have entropy $H \in [1,104.28, 1,104.30]$

Finally, for $T = 1000$, and GCE with uniform priors the hypothesis $H_0 : \boldsymbol{\beta} = \mathbf{0}$ vs. $H_1 : \boldsymbol{\beta} \neq \mathbf{0}$ is tested. ER $= 2208.6 > \chi^2_{(C,\alpha)}$ so cannot accept the null hypothesis of $\boldsymbol{\beta} = \mathbf{0}$. In terms of LD, the probability that all probabilities are uniform within their supports (so $\boldsymbol{\beta} = \mathbf{0}$), given the data, is $2^{-TD(\hat{P},\hat{W}||P^0,W^0)} = 4.02E - 226$ which is consistent with the ER test above. A similar test, applied just for the signal probabilities, P, yields a probability of 0.018.

6.7 Summary

So far the generic IT estimator and two basic members of that class (those with zero-moment conditions — Section 5 — and those with stochastic moment conditions — Section 6) were discussed. In the next section, a simple example is used to further connect the two and to relate it to some of the more traditional estimation methods.

A note on computation is in place. For the ME and GME models, one can use the concentrated (dual), unconstrained model. Computationally, these are efficient methods that are easy to compute. A number of leading software packages have integrated the ME and GME procedures in them. This includes, SAS, LIMDEP (for discrete choice) and SHAZAM. Since the concentrated model is unconstrained, it can easily be used within any other software that allows for optimization, using matrix operations. In addition, using optimization software, such as GAMS, makes it easy to construct these methods in their primal (constrained optimization) form, which makes it easier for estimating time-series data and nonlinear problems. In contrast to the ME and GME models, the EL, GEL, and GMM models are often computationally more complex, even when the concentrated (unconstrained) model is used. See Imbens (2002) and Kitamura (2006) for brief discussions on this issue.

7

IT, Likelihood and Inference — Synthesis via a Discrete Choice, Matrix Balancing Example

7.1 The Basic Problem — Background

To study the interconnection between IT, likelihood and inference, the problem of Matrix Balancing (or contingency tables) is investigated and studied. To introduce the problem, consider the familiar matrix-balancing problem, where the goal is to fill the cells of a matrix from aggregated data. Given the row and column sums we wish to recover the cells (each pixel) of the matrix. This basic problem is a simplified version of a large class of common estimation problems discussed below.

To make our example more visual, consider the following matrix:

	x_1		x_j		x_K
y_1	P_{11}				P_{1K}
y_i			P_{ij}		
y_K	P_{K1}				P_{KK}

The bold symbols $(\boldsymbol{x}, \boldsymbol{y})$ reflect the data we have, and the P's reflect the unknown quantities the researcher wishes to estimate from the data. Each y_i is the sum of all the cells in its respective row (e.g., $\boldsymbol{y}_1 = P_{11} + \cdots + P_{1K}$). Similarly, each x_j is the sum of all the elements in its respective column. In many problems, data on $\boldsymbol{y}$ and $\boldsymbol{x}$ are in terms of real values (say dollars). In these cases, $\boldsymbol{y}$ and $\boldsymbol{x}$ can be normalized to satisfy the above requirement. Once the P's are estimated, the quantities of interest (such as dollars or flows in each cell) can be recovered by a renormalization process.

Before solving this problem, a short discussion of a number of common applications of this type of model, is given. First, let $\boldsymbol{y}$ be a K-dimensional vector of proportions for each one of the kth states in period (t), and let $\boldsymbol{x}$ be a K-dimensional vector of proportions for each state k in period $t+1$, then P is a $(K \times K)$ matrix of first-order Markov transition probabilities. If more than two periods of data exist, the same framework holds with T K-dimensional vectors $\boldsymbol{y}_t$ and $\boldsymbol{x}_t$. Other examples within economics include social accounting (and input–output) balancing from cross section and/or time-series data (Golan and Vogel, 2000), unordered multinomial choice and other discrete choice problems, as well as management problems like the "traveling sales-person," airline routing, distribution of goods among retailers, and work on spectral analysis. This basic problem is also quite common in other disciplines such as medicine, physics, chemistry, biology, topography, engineering, communication, information, operations research, and political science. Within these disciplines, practical examples include work on image reconstruction, analysis of political surveys, tomography[1] and much more. In all of these problems the observed data can be normalized to the [0, 1] interval and the P's could therefore be viewed as probabilities.

This example is chosen because it encompasses a large class of linear models, it allows a direct derivation of both the IT model and the ML, and it allows us to further develop, study, and utilize all the quantities defined and discussed in Sections 3–6 for estimation and inference.

[1] See for example Golan and Volker (2001). For a nice discussion of hypothesis testing, within the classical ME, for these types of problems, see Good (1963).

7.2 The IT-ME Solution

What do we know and observe?

1. The hard data: $\boldsymbol{x}$ and $\boldsymbol{y}$. These data are just the row and column sums.
2. We also know (assume) that the linear relationship between the data ($\boldsymbol{x}$ and $\boldsymbol{y}$) and the unobserved P is:

$$y_i = \Sigma_j p_{ij} x_j \tag{7.1}$$

and

$$\Sigma_i p_{ij} = 1. \tag{7.2}$$

These two sets of equations reflect all we know about these data. It reflects the interrelationship between the (*unknown*) elements in the matrix and the (known) sums of each row and column, and it reflects the fact that, in this example, the elements of each column can be viewed as a proper probability distribution.

Our objective is to estimate the matrix P with minimal assumptions.

Is it possible to achieve this goal with the data we have? To answer this, it is helpful to first look at what we know and what we want to know. Looking at the table above it is clear that there are all together K^2 unknowns. Also it is clear that there are K known quantities (or data points) in the first equation and an additional set of K known quantities from the second equation. All together there are $2K$ knowns. Therefore (for all $K > 2$), the problem is under-determined; there are infinitely many P's that satisfy these two equations. To solve such a problem we resort to the ME (or CE) principle.

Let P^0 be the set of prior probabilities. We then solve this problem by minimizing the relative entropy $D(\boldsymbol{p}||\boldsymbol{p}^0)$ subject to the set of observed moments (7.1) and the set of proper probabilities' requirements (7.2).[2]

[2] Note, that rather than using the relative entropy in (7.3), we can use the EL objective to get:

$$L = \frac{1}{K}\sum_{k=1}^{K} \log(p_k) + \sum_i \lambda_i (y_i - \sum_j p_{ij} x_j) + \sum_j \mu_j (1 - \sum_i p_{ij}).$$

However, our interest here is to connect the traditional ME and ML.

The Lagrangean is

$$L = D(\boldsymbol{p}||\boldsymbol{p}^0) + \sum_i \lambda_i \left(y_i - \sum_j p_{ij} x_j \right) + \sum_j \mu_j \left(1 - \sum_i p_{ij} \right). \quad (7.3)$$

Using the natural log, the estimated coefficients are

$$\tilde{p}_{ij} = \frac{p_{ij}^0 \exp(\tilde{\lambda}_i x_j)}{\sum_i p_{ij}^0 \exp(\tilde{\lambda}_i x_j)} \equiv \frac{p_{ij}^0 \exp(\tilde{\lambda}_i x_j)}{\Omega_j(\tilde{\lambda}_i)}. \quad (7.4)$$

If the priors are all uniform, $p_{ij}^0 = \frac{1}{K}$ for all i and j, than $\tilde{p}_{ij} = \hat{p}_{ij}$ (ME = CE).

The concentrated, dual, formulation is

$$\begin{aligned}
\ell(\lambda) &= \sum_i \sum_j p_{ij} \log(p_{ij}/p_{ij}^0) + \sum_i \lambda_i \Big(y_i - \sum_j p_{ij} x_j \Big) \\
&= \sum_i \sum_j p_{ij} \log \left[\frac{p_{ij}^0 \exp(\lambda_i x_j)}{\Omega_j} \right] - \sum_i \sum_j p_{ij} \log p_{ij}^0 \\
&\quad + \sum_i \lambda_i y_i - \sum_i \sum_j \lambda_i p_{ij} x_j \\
&= \sum_i \lambda_i y_i - \sum_i \sum_j p_{ij} \log \Omega_j + \sum_i \sum_j p_{ij} \log p_{ij}^0 \\
&\quad - \sum_i \sum_j p_{ij} \log p_{ij}^0 + \sum_i \sum_j p_{ij} \lambda_i x_j - \sum_i \sum_j p_{ij} \lambda_i x_j \\
&= \sum_i \lambda_i y_i - \sum_i \sum_j p_{ij} \log \Omega_j(\lambda) \\
&= \sum_i \lambda_i y_i - \sum_j \log \Omega_j(\lambda). \quad (7.5)
\end{aligned}$$

Maximizing the dual, unconstrained (concentrated) problem $\ell(\lambda)$, with respect to $\boldsymbol{\lambda}$ and equating to zero yields $\tilde{\boldsymbol{\lambda}}$ which, in turn, yields the estimates $\tilde{p}_{ij}$ via Eq. (7.4). As was discussed previously, the concentrated model (7.5) is computationally much more efficient. It is worth noting here that, like similar problems, there are exponent functions in (7.5) and therefore it is useful to normalize the data. Like other discrete choice models, if the normalization is done correctly, the P's are

not affected (though the Lagrange multipliers are affected). A normalization used often is dividing each element of $\boldsymbol{x}$ and $\boldsymbol{y}$ by $\text{Max}\{x_j, y_i\}$.

7.3 The Maximum Likelihood Solution

To contrast the concentrated (dual) CE/ME estimators with the ML consider the following approach. Given the above re-normalized data where each $y_i \in [0,1]$ and ignoring the priors (or similarly, assuming all priors are uniform), the likelihood function can be expressed as

$$L = \prod_{j=1}^{K} p_{1j}^{y_1} p_{2j}^{y_2} \dots p_{Kj}^{y_K}, \tag{7.6}$$

and the log-Likelihood function is

$$\log(L) \equiv \ell = \sum_i \sum_j y_i \log p_{ij}. \tag{7.7}$$

Letting P be the logistic (exponential) distribution:

$$p_{ij} = \frac{\exp(\beta_i x_j)}{\sum_i \exp(\beta_i x_j)} \equiv \frac{\exp(\beta_i x_j)}{1 + \sum_{i=2}^{K} \exp(\beta_i x_j)} \equiv \frac{\exp(\beta_i x_j)}{\Omega_j(\boldsymbol{\beta})} \quad \text{for } i = 2, \dots, K \tag{7.8}$$

and

$$p_{1j} = \frac{1}{1 + \sum_{i=2}^{K} \exp(\beta_i x_j)} \equiv \frac{1}{\Omega_{\text{j}}(\boldsymbol{\beta})} \quad \text{for } i = 1 \tag{7.9}$$

where $\Omega_{\text{j}}(\boldsymbol{\beta}) \equiv 1 + \sum_{i=2}^{K} \exp(\beta_i x_j)$.

Substituting (7.8) and (7.9) into (7.7) yields

$$\begin{aligned} \ell &= \sum_i \sum_j y_i \log \left[\frac{\exp(\beta_i x_j)}{1 + \sum_{i=2}^{K} \exp(\beta_i x_j)} \right] \\ &= \sum_i \sum_j y_i \beta_i x_j - \sum_i \sum_j y_i \log \Omega_j(\boldsymbol{\beta}) \\ &= \sum_i y_i \beta_i - \sum_j \log \Omega_j(\boldsymbol{\beta}), \end{aligned} \tag{7.10}$$

which is just a simple version of the ML multinomial logit.

In this problem, for uniform priors ($p_{ij}^0 = \frac{1}{K}$ for all i and j), the ML-logit is *equivalent* to the dual unconstrained model (7.5) with $\boldsymbol{\beta} = \boldsymbol{\lambda}$.

Finally, the first-order conditions (FOC) of (7.10), and similarly of (7.5), are just

$$\frac{\partial \ell}{\partial \lambda_i} = y_i - \sum_j p_{ij} x_j = 0 \tag{7.11}$$

and the diagonal elements of the Hessian are

$$\frac{\partial^2 \ell}{\partial \lambda_i^2} = -\sum_j \frac{\partial p_{ij}}{\partial \lambda_i} x_i = -\sum_j [-p_{ij}(1-p_{ij})x_j]x_j = \sum_j p_{ij}(1-p_{ij})x_j^2, \tag{7.12}$$

where

$$\frac{\partial p_{ij}}{\partial \lambda_i} = -p_{ij}(1-p_{ij})x_j \tag{7.13}$$

and the non-diagonal elements of the Hessian are

$$\frac{\partial^2 \ell}{\partial \lambda_i \partial \lambda_l} = -\sum_j p_{ij} p_{lj} x_j^2. \tag{7.14}$$

The information matrix is

$$I(\boldsymbol{\lambda}) = E\left(-\frac{\partial^2 \ell}{\partial \lambda_i \partial \lambda_l}\right) = \begin{cases} -\sum_j p_{ij}(1-p_{ij})x_j^2 & \text{for } i = l \\ \sum_j p_{ij} p_{lj} x_j^2 & \text{for } i \neq l. \end{cases} \tag{7.15}$$

The covariance matrix and the standard errors are easily calculated from (7.15).

So far, we have seen that for that discrete problem, the ME and ML (logit) are equivalent. Next, the above formulation is generalized.

7.4 The Generalized Case — Stochastic Moments

Consider the more realistic case where the observed moments are noisy. In that case one can handle it, within the IT–ME philosophy, as stochastic constraints. (This is similar in spirit to the model described in Chapter 6.) Following Golan et al. (1996b) and Golan et al. (1996c) Eq. (7.1)

can be written as

$$y_i = \sum_j p_{ij} x_j + e_i. \tag{7.16}$$

As discussed earlier, both $\boldsymbol{x}$ and $\boldsymbol{y}$ are normalized to the [0, 1] interval, so each error $e_i \in [-1, 1]$. Following the GME literature (7.16) can be specified as

$$y_i = \sum_j p_{ij} x_j + \sum_h v_h w_{ih}, \tag{7.17}$$

where $\boldsymbol{w}_i$ is an H-dimensional $v$ector of weights satisfying

$$\sum_h w_{ih} = 1 \quad \text{and} \quad \sum_h v_h w_{ih} \equiv e_i \tag{7.18}$$

and $\boldsymbol{v}$ is an H-dimensional support with $H \geq 2$ and symmetric around zero. Building on the IT–CE model of Section 7.2 and on Section 6, the new generalized model is

$$\begin{aligned}
&\underset{\boldsymbol{p},\boldsymbol{w}}{\text{Min}} \{D(\boldsymbol{p}, \boldsymbol{w} || \boldsymbol{p}^0, \boldsymbol{w}^0)\} = \underset{\boldsymbol{p},\boldsymbol{w}}{\text{Min}} \{D(\boldsymbol{p} || \boldsymbol{p}^0) + D(\boldsymbol{w} || \boldsymbol{w}^0)\} \\
&\text{s.t.} \\
&y_i = \sum_j p_{ij} x_j + \sum_h v_h w_{ih} \\
&\sum_i p_{ij} = 1; \quad \sum_h w_{ih} = 1
\end{aligned} \tag{7.19}$$

The optimization yields

$$\tilde{p}_{ij} = \frac{p^0_{ij} \exp(\tilde{\lambda}_i x_j)}{\sum_i p^0_{ij} \exp(\tilde{\lambda}_i x_j)} \equiv \frac{p^0_{ij} \exp(\tilde{\lambda}_i x_j)}{\Omega_j(\tilde{\lambda}_i)} \tag{7.20}$$

and

$$\tilde{w}_{ih} = \frac{w^0_{ih} \exp(\tilde{\lambda}_i v_h)}{\sum_h w^0_{ih} \exp(\tilde{\lambda}_i v_h)} \equiv \frac{w^0_{ih} \exp(\tilde{\lambda}_i v_h)}{\Psi_i(\tilde{\boldsymbol{\lambda}})}, \tag{7.21}$$

where w^0_{ih} are the prior probabilities defined over the support space $\boldsymbol{v}$ and are taken to be uniform.

The concentrated, dual, IT (GCE) estimation rule is

$$\ell(\boldsymbol{\lambda}) = \sum_i \lambda_i y_i - \sum_j \log \Omega_j(\lambda) - \sum_i \log \Psi_i(\boldsymbol{\lambda}). \tag{7.22}$$

This is an IT GCE/GME which is a generalized ML-logit. If $e_i = 0$ for all i, then the ML = ME = GME/GCE. By allowing stochastic moments, and using the relative entropy $D(\cdot||\cdot)$ as the objective, all the errors are "pushed" toward zero but are not forced to be exactly zero. In that way, the sample's moments are allowed (but not forced) to be different than the underlying population moments, a flexibility that seems natural for finite data sets. This flexibility means that the resulting estimates are more stable (lower variances) relative to estimates resulting from zero-moment restrictions.

With that basic framework, the IT model summarized here can be easily extended to include time-series data as well as any other information that may be captured via its cross moments with $\boldsymbol{x}$ and $\boldsymbol{y}$. See for example Golan and Vogel (2000) and Golan et al. (2007).

7.5 Inference and Diagnostics

As before, one can define the information measures (Golan, 1988).

$$S(\tilde{P}) \equiv \frac{-\sum_i \sum_j \tilde{p}_{ij} \log \tilde{p}_{ij}}{K \log(K)} \quad \text{for uniform priors, or}$$

$$S(\tilde{P}) \equiv \frac{-\sum_i \sum_j \tilde{p}_{ij} \log \tilde{p}_{ij}}{-\sum_i \sum_j p^0_{ij} \log p^0_{ij}}$$

and

$$S(\tilde{\boldsymbol{p}}_j) \equiv \frac{-\sum_i \tilde{p}_{ij} \log \tilde{p}_{ij}}{\log K} \quad \text{or} \quad S(\tilde{\boldsymbol{p}}_j) \equiv \frac{-\sum_i \tilde{p}_{ij} \log \tilde{p}_{ij}}{\sum_i p^0_{ij} \log p^0_{ij}},$$

where both sets of measures are between zero and one with one reflecting uniformity (complete ignorance: $\boldsymbol{\lambda} = \boldsymbol{\beta} = \mathbf{0}$) of the estimates, and zero reflecting perfect knowledge. The first measure reflects the information in the whole system, while the second one reflects the information in each moment/column j. Note that similar information measures can be constructed for each desired sub-matrix of P. Similar information measures like $I(\tilde{P}) = 1 - S(\tilde{P})$ are also used (e.g., Soofi, 1996).

The normalized entropy can easily be connected to Fano's inequality, formulating a bound on the probability of error, discussed in earlier sections. In the context of this section, consider the follow-

ing example. Assume the researcher knows that the highest probability for each vector in the matrix is on the diagonal (e.g., within a Markov process it means the highest probability is associated with "no change" in states between two consecutive periods). For each probability vector j $(j = 1, 2, \ldots, K)$ let the estimator $\hat{p}_j = p_{jj}$. The probability of error in each vector (proper probability distribution) j is $p_{ej} = 1 - \hat{p}_j = 1 - p_{jj}$. Recall that $H(p_{ej})$ is just the entropy of the two elements p_{ej} and $1 - p_{ej}$. With these definitions, Fano's inequality is

$$H(p_{ej}) + p_{ej} \log(K - 1) \geq H(\boldsymbol{p}_j),$$

and the weaker inequality (providing a bound on the error) is

$$p_{ej} \geq \frac{[H(\boldsymbol{p}_j) - 1]}{\log K} = S(\boldsymbol{p}_j) - \frac{1}{\log K}.$$

Once an error probability (as a function of the entropy or the normalized entropy) is constructed for each one of the K probability distributions, it can be easily extended for the full matrix.

Following the likelihood literature (the traditional likelihood ratio test), the empirical likelihood literature (Owen, 1988, 1990; Hall, 1990; Qin and Lawless, 1994), the derivations of Section 6.3 and building on the quantities presented in Sections 3 and 5, an entropy ratio test can be used. Let ℓ_Ω be the unconstrained likelihood, and ℓ_ω be the constrained one where, say $\boldsymbol{\beta} = \boldsymbol{\lambda} = \mathbf{0}$. Then, the log-likelihood ratio statistic is $-2 \log \frac{\ell_\Omega}{\ell_\omega}$. Given the equivalence between the ML and the ME, the *log-likelihood* value of the constrained problem is just the value of $\text{Max}(H)$ while $\log \ell_\omega = K \log K$. Thus, the log-likelihood, or entropy-ratio statistic is just

$$W(\text{IT} - \text{ME}) = 2 \log \ell_\omega - 2 \log \ell_\Omega = 2K \log(K)[1 - S(\tilde{P})]$$

or

$$W(\text{CE}) = 2H(P^0)[1 - S(\tilde{P})].$$

For the non-uniform priors and $S(\tilde{P}) \equiv H(\tilde{P})/H(P^0)$.

Under the null hypothesis, W(IT–ME) converges in distribution to $\chi^2_{(K-1)}$. Finally, McFadden (1974) Pseudo-R^2 gives the proportion of

variation in the data that is explained by the model (a measure of model fit):

$$\text{Pseudo} - R^2 \equiv 1 - \frac{\log \ell_\Omega}{\log \ell_\omega} = 1 - S(\tilde{P}).$$

Finally, to see the similarity between the $D(\boldsymbol{p}||\boldsymbol{p}^0)$, the CE objective, and the χ^2 statistic, in a different way, consider the following. Let $\{p_{ij}\}$ be a set of K observed distributions where each distribution is over K observations. Let the null hypothesis be $H_0 : P = P^0$. Then,

$$\chi^2_{(K-1)} = \sum_i \sum_j \frac{1}{p^0_{ij}}(p_{ij} - p^0_{ij})^2.$$

But, a second-order approximation of $D(\boldsymbol{p}||\boldsymbol{p}^0)$ yields

$$D(\boldsymbol{p}||\boldsymbol{p}^0) \equiv \sum_i \sum_j p_{ij} \log(p_{ij}/p^0_{ij}) \cong \frac{1}{2} \sum_i \sum_j \frac{1}{p^0_{ij}}(p_{ij} - p^0_{ij})^2,$$

which is just the log-likelihood ratio statistic of this estimator. Since two times the log-likelihood ratio statistic corresponds approximately to χ^2, the relationship is clear. All statistics discussed here apply for both the ME and GME methods.

This derivation emphasizes the basic difference between the ME and CE (or GME and GCE) approaches. Under the ME (or GME) approach, one investigates how "far" the data pull the estimates away from a state of complete ignorance (uniform distribution). Thus, a high value of χ^2 implies the data tell us something about the estimates, or similarly, there is valuable information in the data. Under the CE (or GCE) approach, the question becomes how far the data take us from our initial (*a priori*) beliefs — the priors. A high value of χ^2 implies that our prior beliefs are rejected by the data. For some discussion and background on goodness of fit statistics for multinomial type problems see, for example, Koehler and Larntz (1980) and Greene (2008). Further discussion of diagnostics and testing for ME–ML model (under zero-moment conditions) appears in Soofi (1994). He provides measures related to the normalized entropy measures discussed above and provides a detailed formulation of decomposition of these information concepts.

7.6 IT and Likelihood

Though coming from different philosophies and approaching the estimating problem differently, under the formulation used here, the ME is equivalent to the ML for the problem discussed here. More generally, for discrete choice problems with zero-moment conditions, the ME and the ML-logit are the same. This equivalency allows us to understand some of the inherent relationship between IT and ML. In investigating both primal (constrained) and dual (unconstrained) models, it is easy to see that the basic unknowns of interest are the Lagrange multipliers: $\boldsymbol{\lambda} \equiv -\boldsymbol{\beta}$. Using the normalization $\lambda_1 = \beta_1 = 0$, the objective is to recover these $K-1$ unknowns which in turn yield the $K \times K$ matrix P.[3]

In both the likelihood and the IT–ME approaches the basic parameters are the parameters that enter the exponential distribution. In the ML-logit model, the likelihood (or underlying distribution) is specified *a priori*. In the IT approach, no likelihood is specified. However, using Shannon's entropy as the objective (or the Kullback–Liebler entropy divergence measure) the solution to the optimization problem is the exponential (logistic) distribution. With this in mind, it is possible to have better interpretation to the estimated parameters (known also as the MaxEnt distribution). These parameters are not only the parameters of the estimated distribution, they also represent the additional information in each data point (moment). It is the marginal information of each one of the observed moments.

So far the simpler — zero moments — case was discussed. Extending the synthesis to the stochastic moments example reveals that the IT–GCE (or GME) model is a generalization of the ME–ML model (e.g., Golan et al., 1996b,c). This is a semi parametric, IT model that is more efficient than the ML for all finite data, yet exhibits the

[3] Note that since we seek to evaluate the impact of the $x_j's$ on each p_{ij}, just like with the traditional discrete choice models, the more interesting parameters are the marginal effects. The marginal quantities for the classical cross-entropy are

$$\frac{\partial \tilde{p}_{ij}}{\partial x_j} = \tilde{\lambda}_i \tilde{p}_{ij} - \tilde{p}_{ij} \frac{\sum_i p_{ij}^0 \exp(\tilde{\lambda}_i x_j)\tilde{\lambda}_i}{\Omega_j} = \tilde{p}_{ij}\left(\tilde{\lambda}_i - \sum_i \tilde{p}_{ij}\tilde{\lambda}_i\right)$$

.

same level of complexity. Looking at the concentrated model (7.22), one sees that the parameters of interest in the generalized model are the same as in the zero-moment model. There is no added complexity. By treating the observed data as stochastic, greater flexibility is achieved.

Finally, and of most importance is to note that by redeveloping that model from an IT perspective, all results and quantities developed and defined in Sections 3–6 apply here as well. Of utmost importance are (i) the ECT that provides a new rationale for using the ML for large samples and the GME for smaller samples, and (ii) the large deviations interpretation that under that IT framework can be immediately applied, and used, for the ML.

7.7 A Specific Example: Conditional, Time Series Markov Process

Consider a first-order (conditional) Markov Chain that is stationary and for simplicity is also time-invariant, discussed in Section 3.8. Let Y be a discrete random variable with alphabet $\aleph$ and a probability mass function $p(y) = \text{Prob}\{Y = y\},\ y \in \aleph$. Let the stochastic process $Y_1, Y_2, \ldots, Y_T$ be such that $\text{Prob}(Y_{t+1} = y_{t+1} | Y_t = y_t)$ for all $y_1, y_2, \ldots, y_T \in \aleph$ and the time subscript t. To simplify notations (similar to Section 3.8), let the indices j and k represent the states at period $t+1$ and t, respectively. Then, for each individual (observation) $i = 1, 2, \ldots, N$, this process can be specified as

$$y_j(t+1) = \sum_k y_k(t) P_{kj} \quad \text{with} \quad \sum_j y_j = 1 \quad \text{and} \quad \sum_k P_{jk} = 1, \quad (7.23)$$

where P is the stationary first-order Markov probability matrix. Let y_{itj} be the observed state of individual i at period $t = 1, 2, \ldots, T$. Then Eq. (7.23) becomes

$$y_{itj} = \sum_{k=1}^{K} p_{kj} y_{i,t-1,k}, \quad (7.24)$$

where y_{itj} is a K-dimensional vector of binary variables for each individual that takes the value $y_{itj} = 1$ if state j is observed at time t and

$y_{itj} = 0$ for all other $K-1$ states.[4] This Markov process is conditional on individual characteristics and other data Z. Within the spirit of the IT class of estimators, we would like to introduce the Data Z with minimal assumptions. If we do not know the exact functional form connecting Y and Z, a possible way to incorporate this relationship is via the cross moments.

Let Z be a matrix of individual covariates with the elements z_{its}, so $P_{t+1} = \boldsymbol{f}(y_t(\boldsymbol{z}_{ts}), \boldsymbol{z}_{ts}) = \boldsymbol{f}(\boldsymbol{z}_{ts})$, or more explicitly, $p_{kj}(t+1) = f_s(y_t(\boldsymbol{z}_{ts}), \boldsymbol{z}_{ts}) = f_s(\mathbf{z}_{ts})$. Since, the researcher does not know $\boldsymbol{f}$, in this example (the relationship between the observed data, the unknown probabilities and the covariates Z) it can be captured via the cross moments:

$$\sum_{t=2}^{T}\sum_{i=1}^{N} y_{itj} z_{its} = \sum_{t=1}^{T-1}\sum_{i=1}^{N}\sum_{k=1}^{K} p_{kj} y_{itk} z_{its}. \tag{7.25}$$

Given the above framework, our objective here is to estimate the K^2 unknown P's using different IT estimators discussed above. By using the same framework for all estimators, we are able to see explicitly the different assumptions and information used for each method and the relationship among these different estimators.

7.7.1 Different IT Estimators

Case A. The Zero (pure) Moments

The Classical ME (Uniform Priors) is

$$\begin{aligned} &\underset{\boldsymbol{p}}{\text{Max}}\ H(\boldsymbol{p}) \\ &\text{s.t.} \\ &\sum_{t=2}^{T}\sum_{i=1}^{N} y_{itj} z_{its} = \sum_{t=1}^{T-1}\sum_{i=1}^{N}\sum_{k=1}^{K} p_{kj} y_{itk} z_{its}; \quad \sum_{k} p_{kj} = 1 \end{aligned}$$

[4] For example, if $K = 6$ and the ith individual is in state $j = 2$ in time t, then $y_{it2} = (0\ 1\ 0\ 0\ 0\ 0)$.

with the solution

$$\hat{p}_{kj} = \frac{\exp\left(-\sum_{t=1}^{T-1}\sum_{i,s} y_{itk} z_{its} \hat{\lambda}_{sj}\right)}{\sum_j \exp\left(-\sum_{t=1}^{T-1}\sum_{i,s} y_{itk} z_{its} \hat{\lambda}_{sj}\right)} \equiv \frac{\exp\left(-\sum_{t=1}^{T-1}\sum_{i,s} y_{itk} z_{its} \hat{\lambda}_{sj}\right)}{\Omega_k}. \tag{7.26}$$

The Concentrated (Dual) ME is

$$\ell(\lambda) = \sum_{t=2}^{T}\sum_{j=1}^{K}\sum_{i,s} y_{itj} z_{its} \lambda_{sj} + \sum_k \log \Omega_k(\lambda).$$

Solving with respect to $\boldsymbol{\lambda}$, and substituting in Eq. (7.26) yields $\hat{P}$'s.

The Cross Entropy (Non-Uniform Priors) is

$$\underset{\boldsymbol{p}}{\text{Min}} \{D(\boldsymbol{p}||\boldsymbol{p}^0)\}$$

s.t.

$$\sum_{t=2}^{T}\sum_{i=1}^{N} y_{itj} z_{its} = \sum_{t=1}^{T-1}\sum_{i=1}^{N}\sum_{k=1}^{K} p_{kj} y_{itk} z_{its}; \quad \sum_k p_{kj} = 1$$

with the solution

$$\tilde{p}_{kj} = \frac{p^0_{kj} \exp\left(\sum_{t=1}^{T-1}\sum_{i,s} y_{itk} z_{its} \tilde{\lambda}_{sj}\right)}{\sum_j p^0_{kj} \exp\left(\sum_{t=1}^{T-1}\sum_{i,s} y_{itk} z_{its} \tilde{\lambda}_{sj}\right)} \equiv \frac{p^0_{kj} \exp\left(\sum_{t=1}^{T-1}\sum_{i,s} y_{itk} z_{its} \tilde{\lambda}_{sj}\right)}{\Omega_k}. \tag{7.27}$$

Note that the CE is an extension of the ME for non-uniform priors. If all priors are uniform (i.e., $p^0_{kj} = 1/K$ for all k and j), then the CE solution is the same as the ME solution ($\tilde{P} = \hat{P}$).

The Concentrated (Dual) CE is

$$\ell(\lambda) = \sum_{t=2}^{T}\sum_{j=1}^{K}\sum_{i,s} y_{itj} z_{its} \lambda_{sj} - \sum_k \log \Omega_k(\lambda),$$

and, similar to the concentrated ME, we solve with respect to $\boldsymbol{\lambda}$, and then substitute the estimated $\boldsymbol{\lambda}$'s into Eq. (7.27) which yields $\tilde{P}$'s.

The Maximum Likelihood — Logit To construct the zero-moment ML-Logit model, one starts by specifying the traditional log-likelihood function and substituting in the desired functional form connecting the P's and the data. Choosing the exponential (or logistic) distribution yields

$$\begin{aligned}
\log(L) \equiv \ell &= \sum_{i,t,j} y_{itj} \log p_{kj}\left(z_{its}\right) \\
&= \sum_{i,t,j} y_{itj} \log \left[\frac{\exp\left(\sum_{t=1}^{T-1} \sum_{i,s} y_{itk} z_{its} \beta_{sj}\right)}{\sum_j \exp\left(\sum_{t=1}^{T-1} \sum_{i,s} y_{itk} z_{its} \beta_{sj}\right)} \right] \\
&= \sum_{i,t,s,j} y_{itj} z_{its} \beta_{sj} - \sum_{i,t,k} y_{itk} \log \sum_j \exp\left(\sum_{t=1}^{T-1} \sum_{i,s} y_{itk} z_{its} \beta_{sj} \right) \\
&= \sum_{i,t,s,j} y_{itj} z_{its} \beta_{sj} - \sum_k \log \Omega_k(\boldsymbol{\beta}).
\end{aligned}$$

For completion, the information matrix is

$$I(\boldsymbol{\lambda}) = E\left(-\frac{\partial^2 \ell}{\partial \boldsymbol{\lambda} \partial \boldsymbol{\lambda}'} \right).$$

Solving with respect to the unknown parameters $\boldsymbol{\lambda}$, and substituting them into the logistic distribution yields the optimal solution $\hat{P}_{\mathrm{ML}}$. This solution is equivalent to the ME solution $\hat{P}$. Thus, for that problem the ML Logit is equivalent to the ME method (and to the CE with uniform priors).

The Generic-IT It is possible to construct this problem within the generic IT framework of Section 5.2. It is best to start with the primal (constrained) model.

$$\begin{aligned}
&\underset{\boldsymbol{p}}{\mathrm{Min}} \left\{ f(\boldsymbol{p}||\boldsymbol{p}^0) = D^R_{\alpha+1}(\boldsymbol{p}||\boldsymbol{p}^0) \right\} \\
&\text{s.t.} \\
&\sum_{t=2}^{T} \sum_{i=1}^{N} y_{itj} z_{its} = \sum_{t=1}^{T-1} \sum_{i=1}^{N} \sum_{k=1}^{K} p_{kj} y_{itk} z_{its}; \quad \sum_k p_{kj} = 1.
\end{aligned}$$

Note, that in this case, one needs to choose α prior to the estimation. Then, going through the same steps discussed above (constructing the Lagrangean and solving) yields $\tilde{P}_{\text{Generic}-\text{IT}}$. For example, for $\alpha \to 0$, $\tilde{P}_{\text{Generic}-\text{IT}} = \tilde{P}_{\text{CE}}$ and if all priors are uniform $\tilde{P}_{\text{Generic}-\text{IT}} = \tilde{P}_{\text{CE}} = \hat{P} = \hat{P}_{\text{ML}}$. If, $\alpha \to -1$, then the EL solution (developed below) results. Finally, similar to the above methods, the generic IT model can be specified as an unconstrained, concentrated model (see Sections 5 and 6).

The Empirical Likelihood is

$$\underset{\boldsymbol{p}}{\text{Max}} \left\{ \frac{1}{KJ} \sum_{k,j} \log p_{kj} \right\}$$

s.t.

$$\sum_{t=2}^{T} \sum_{i=1}^{N} y_{itj} z_{its} = \sum_{t=1}^{T-1} \sum_{i=1}^{N} \sum_{k=1}^{K} p_{kj} y_{itk} z_{its}; \quad \sum_{k} p_{kj} = 1.$$

The solution is

$$\hat{p}_{kj} = (KJ)^{-1} \left[\sum_{t=1}^{T-1} \sum_{i,s} y_{itk} z_{its} \hat{\lambda}_{sj} + 1 \right]^{-1}.$$

Note that the $\hat{P}_{\text{EL}} \neq \hat{P}_{\text{ME}} = \hat{P}_{\text{ML}}$. Like the previous models, the EL can be specified as a constrained optimization model (above) or as a concentrated, unconstrained model (see Section 5).

Finally, consider the case where the researcher has additional information in terms of parametric information functions (PIF's) $\boldsymbol{g}(\boldsymbol{y}, \boldsymbol{p}, \boldsymbol{\theta}) = [\mathbf{0}]$ where $\boldsymbol{y}$ is conditional on Z. The above EL (or similarly the GEL or the Generic-IT estimator) can be formulated in terms of that information. Just substitute these PIF's for the cross-moment conditions above and optimize with respect to the unknown parameters. This problem is of a lower dimension and utilizes more information (information that enters via the PIF's).

Case B. The Stochastic Moments Models

We now discuss the class of estimators where the moment conditions are specified as stochastic conditions. Rewriting the zero-moment

conditions (Eq. (7.25)) as stochastic yields

$$\sum_{t=2}^{T}\sum_{i=1}^{N} y_{itj}z_{its} = \sum_{t=1}^{T-1}\sum_{i=1}^{N}\sum_{k=1}^{K} p_{kj}y_{itk}z_{its} + \sum_{t=1}^{T-1}\sum_{i=1}^{N} z_{its}\varepsilon_{itj}$$
$$= \sum_{t=1}^{T-1}\sum_{i=1}^{N}\sum_{k=1}^{K} p_{kj}y_{itk}z_{its} + \sum_{t=1}^{T-1}\sum_{i=1}^{N}\sum_{m=1}^{M} z_{its}w_{itjm}v_m, \tag{7.25a}$$

where the additive noise $\boldsymbol{\varepsilon} \in [-1, 1]$ has zero mean, $\boldsymbol{v}$ is the errors' support space of dimension $M \geq 2$ (and if desired for a certain problem it can be specified to be inversely related to the sample size) and W is the set of probability distributions defined on the same support such that their expected value is ε_{itj} ($\varepsilon_{itj} = \sum_m w_{itjm}v_m$ and $1 = \sum_m w_{itjm}$).

The GME is

$$\max_{\boldsymbol{p},\boldsymbol{w}} \{H(\boldsymbol{p}, \boldsymbol{w})\} = \max_{\boldsymbol{p},\boldsymbol{w}} \{H(\boldsymbol{p}) + H(\boldsymbol{w})\}$$

s.t.

$$\sum_{t=2}^{T}\sum_{i=1}^{N} y_{itj}z_{its} = \sum_{t=1}^{T-1}\sum_{i=1}^{N}\sum_{k=1}^{K} p_{kj}y_{itk}z_{its} + \sum_{t=1}^{T-1}\sum_{i=1}^{N}\sum_{m=1}^{M} z_{its}w_{itjm}v_m$$
$$\sum_i p_{ij} = 1; \quad \sum_m w_{itjm} = 1$$

with the solutions

$$\hat{p}_{kj} = \frac{\exp\left(-\sum_{t=1}\sum_{i,s} y_{itk}z_{its}\hat{\lambda}_{sj}\right)}{\sum_j \exp\left(-\sum_{t=1}\sum_{i,s} y_{itk}z_{its}\hat{\lambda}_{sj}\right)} \equiv \frac{\exp\left(-\sum_{t=1}\sum_{i,s} y_{itk}z_{its}\hat{\lambda}_{sj}\right)}{\Omega_k(\hat{\boldsymbol{\lambda}})}$$

and

$$\hat{w}_{itjm} = \frac{\exp\left(-\sum_s z_{its}v_m\lambda_{sj}\right)}{\sum_m \exp\left(-\sum_s z_{its}v_m\lambda_{sj}\right)} \equiv \frac{\exp\left(-\sum_s z_{its}v_m\lambda_{sj}\right)}{\Psi_{itj}(\hat{\boldsymbol{\lambda}})}$$

GME (Concentrated Model) is
Following the same logic as before, the concentrated model is

$$\begin{aligned}\ell(\lambda) &= \sum_{t=2}^{T}\sum_{j=1}^{K}\sum_{i,s} y_{itj} z_{its}\lambda_{sj} + \sum_{k}\log\left[\sum_{j}\exp\left(-\sum_{t=1}\sum_{i,s} y_{itk} z_{its}\lambda_{sj}\right)\right] \\ &\quad + \sum_{i,t,j}\log\left[\sum_{m}\exp\left(-\sum_{s} z_{its} v_m \lambda_{sj}\right)\right] \\ &= \sum_{t=2}^{T}\sum_{j=1}^{K}\sum_{i,s} y_{itj} z_{its}\lambda_{sj} + \sum_{k}\log\Omega_k(\boldsymbol{\lambda}) + \sum_{i,t,j}\log\Psi_{itj}(\boldsymbol{\lambda}). \qquad (7.28)\end{aligned}$$

Solving with respect to $\boldsymbol{\lambda}$, and substituting in the above solutions $(\hat{p}_{kj})$ yields $\hat{P}_{\text{GME}} \neq \hat{P}_{\text{ME}} = \hat{P}_{\text{ML}}$. As discussed earlier, the level of complexity of the GME is exactly the same as the ME and all other IT estimators discussed. This is because the real parameters of interest are the $\boldsymbol{\lambda}$'s and their dimension is not affected by M. Further, in this example, the GME is just a generalization of the ME (or ML Logit).

GCE (Concentrated Model) is
For brevity of exposition, only the concentrated model is presented below. For non-uniform priors P^0, the GCE concentrated model is

$$\begin{aligned}\ell(\lambda) &= \sum_{t=2}^{T}\sum_{j=1}^{K}\sum_{i,s} y_{itj} z_{its}\lambda_{sj} - \sum_{k}\log\left[\sum_{j} p^0_{kj}\exp\left(\sum_{t=1}\sum_{i,s} y_{itj} z_{tts}\lambda_{sj}\right)\right] \\ &\quad - \sum_{i,t,j}\log\left[\sum_{m} w^0_{itjm}\exp\left(\sum_{s} z_{its} v_m \lambda_{sj}\right)\right] \\ &= \sum_{t=2}^{T}\sum_{j=1}^{K}\sum_{i,s} y_{itj} z_{its}\lambda_{sj} - \sum_{k}\log\Omega_k(\boldsymbol{\lambda}) - \sum_{i,t,j}\log\Psi_{itj}(\boldsymbol{\lambda}),\end{aligned}$$

where $\Omega_k(\boldsymbol{\lambda}) = \sum_j p^0_{kj}\exp(\sum_{t=1}\sum_{i,s} y_{itk} z_{its}\lambda_{sj})$, $\Psi_{itj}(\boldsymbol{\lambda}) = \sum_m w^0_{itjm}$ $\exp(\sum_s z_{its} v_m \lambda_{sj})$ and where all errors' priors, W^0, are taken to be uniform within their supports. If, in addition, all priors P^0's are uniform then $\tilde{P}_{\text{GCE}} = \hat{P}_{\text{GME}} \neq \hat{P}_{\text{ML}}$

7.7.2 Entropy, Information, and Inference

So far we looked at different IT estimators for this problem. We now, look at some basic information measures, rates of information, and possible tests for that problem. These tests can be used for each one of the methods discussed, be it a zero moment or a stochastic moment model and for each value of α — the parameter in the generalized entropy (or Cressie Read) function.

Entropy Rate
The *Entropy Rate* of the stochastic process $\{Y_t\}$ is

$$H(\aleph) \equiv \lim_{T\to\infty} \frac{1}{T} H(Y_1, Y_2, \ldots, Y_T)$$

when the limit exists. In that case $H(\aleph)$ reflects the incremental increase of the (average) information with the process. It captures the amount of entropy change with t. The conditional entropy of the last random variable given the past is

$$H^*(\aleph) \equiv \lim_{T\to\infty} H(Y_T | Y_{T-1}, Y_{T-2}, \ldots, Y_1)$$

when the limit exists. In this example (stationary process)

$$\begin{aligned} H(\aleph) &= H^*(\aleph) = H(Y_2(Z)|Y_1(Z)) \\ &= -\sum_{kj} y_k P_{kj}(Z) \log P_{kj}(Z) = \sum_j y_j \left[-\sum_k P_{kj}(Z) \log P_{kj}(Z) \right]. \end{aligned}$$

For the practical user, this measure can be used to measure the incremental information (or entropy) contribution of each additional period of observed data.

Joint Entropy and Marginal Entropy

$$H(\hat{P}) = -\sum_{kj} \hat{P}_{kj} \log \hat{P}_{kj}, \quad H(\hat{\boldsymbol{p}}_k) = -\sum_j \hat{P}_{kj} \log \hat{P}_{kj}.$$

These measures reflect the total entropy in the signal P or in each vector $\boldsymbol{p}_k$.

Joint Normalized Entropy (Information Measure) and Marginal Normalized Entropy

$$S(\hat{P}) = \frac{\left[-\sum_{kj} \hat{P}_{kj} \log \hat{P}_{kj}\right]}{K \log K}, \quad S(\hat{\boldsymbol{p}}_k) = \frac{\left[-\sum_{j} \hat{P}_{kj} \log \hat{P}_{kj}\right]}{\log K}$$

These measures are just a normalized (to the zero–one interval) version of the previous measures. They are related to the entropy ratio statistic, the pseudo R^2 (Sections 5 and 6 and below) and to the probability of errors (Fano's inequality) discussed below.

The Relative Entropy

$$D(\tilde{\boldsymbol{p}}||\boldsymbol{p}^0) = \sum_{kj} \tilde{P}_{kj} \log(\tilde{P}_{kj}/P^0_{kj}).$$

This measure reflects the entropy distance between the post data (posterior) P's and the priors.

For conditional entropies and the relationship among all of these measures see Section 3.3 and Tables 3.3–3.5.

Probability of Errors

Roughly speaking, if Y is a function of Z, then let $\hat{Y} = f(Z)$ be an estimate of Y. What is the probability that $\hat{Y} = Y$? Let $P_e = \text{Prob}\{\hat{Y} \neq Y\}$ be the probability of error. With a slight abuse of notations (for simplicity) we get

$$H(P_e) + P_e \log(K-1) \geq H(Y|Z).$$

The weaker version of this inequality is

$$P_e \geq \frac{[H(Y|Z) - 1]}{\log(K)} = S(Y|Z) - \frac{1}{\log K}.$$

Both of the above, provides us with a direct relationship between the estimated entropy and the probability of error.

Large Deviations Interpretation

For $p_{kj}(t+1) = f_s(y_t(\boldsymbol{z}_{ts}), \boldsymbol{z}_{ts}) = f_s(\boldsymbol{z}_{ts})$ we want to find the probability that the estimated transitions (conditioned on the observed data)

are equal to the prior probabilities $\boldsymbol{p}^0$:

$$\text{Prob}\left\{\frac{1}{T-1}\sum_t f_s(\boldsymbol{z}_{ts}) \geq c_s,\ s=1,\ldots,S\right\}.$$

Following Sections 3 and 6, define the set E as a set of all proper probability distributions satisfying the observed moment conditions $\boldsymbol{f}(\cdot)$ represented explicitly in (7.25):

$$\begin{aligned} E &\equiv \left\{P\middle|\sum_t f_s(\boldsymbol{z}_{ts}) \geq c_s,\ s=1,\ldots,S\right\} \\ &= \left\{P\middle|\sum_{t=1}^{T-1}\sum_{i=1}^{N}\sum_{k=1}^{K} p_{kj}y_{itk}z_{its} \right. \\ &\qquad \left. \geq \sum_{t=2}^{T}\sum_{i=1}^{N} y_{itj}z_{its},\ s=1,\ldots,S;\ j=1,\ldots K\right\}, \end{aligned}$$

where the second equality reflects our explicit example. To find the closest distribution to the prior distribution P^0, minimize $D(P||P^0)$ subject to the observed moments (see (7.26)). This yields the estimated probabilities $\tilde{p}_{kj}$ (together with $\tilde{\boldsymbol{\lambda}}$). See Eq. (7.27). These are the estimated probabilities that satisfy the observed data and that are closest (in entropy) to the priors. For example, for the uniform priors ($p^0_{kj} = \frac{1}{K}$ for $j = 1,\ldots,J$), $2^{-(T-1)D(\tilde{P}||P^0)}$ is the probability that conditional on the observed data (moments), all transition probabilities are uniform. The same idea can be extended for the stochastic moment (GCE) case.

The Entropy Concentration Theorem

This case is more consistent with the original ECT for the classical ME (Jaynes, 1978a,b). Since the simple zero moment case was already discussed in Section 4, a more general ECT is discussed here. The main extension is that now the possible states/realizations are conditioned on the covariates and that the sample moments are viewed as stochastic. Let $H^*(\hat{\boldsymbol{p}}, \hat{\boldsymbol{w}}) \equiv H(\hat{\boldsymbol{p}}) + H(\hat{\boldsymbol{w}})$ be the entropy value for the conditional stochastic moment problem. Let $\mathscr{P}$ be the subclass of all possible outcomes that could be observed from our sample data that satisfy the constraints (stochastic moments). The ECT, applied

to that GME model, states that a significant percentage of outcomes in the class $\mathscr{P}$ will have an entropy in the range

$$H^* - \Delta H \leq H(\boldsymbol{p}, \boldsymbol{w}) \leq H^* \equiv \underset{p \in P}{\text{Max}} H \text{ (or the value of Eq. (7.28))}$$
$$= H^*(\hat{\boldsymbol{p}}, \hat{\boldsymbol{w}}) \equiv H^*(\hat{\boldsymbol{p}}) + H^*(\hat{\boldsymbol{w}}),$$

where $2(TN)\Delta H \equiv \chi^2_{(C;\alpha)}$, N is the number of individuals and T is the number of time periods, α is the upper a percentile of the χ^2 distribution with C degrees of freedoms. The degree of freedom, C, depends whether one uses zero-moment conditions or stochastic moment conditions. Other distributions $\{p_{km}, w_{tj}\}$, that are conditional on V and are consistent with the constraints (sample data), will have entropy levels smaller than H^* and their concentration near this upper bound is given by the above ECT. For the zero-moment conditions, the above becomes

$$H^* - \Delta H \leq H(\boldsymbol{p}) \leq H^* \equiv \underset{p \in \mathscr{P}}{\text{Max}} H \text{ (of the ME)} = H^*(\hat{\boldsymbol{p}}).$$

Hypothesis Tests
Consider the hypothesis $H_0 : \boldsymbol{\lambda}_s = \mathbf{0}$ *for some* $s \in S$ vs. the alternative $H_1 : \boldsymbol{\lambda}_s \neq \mathbf{0}$ *for some* $s \in S$. Let ℓ_Ω be the value of the maximal entropy (subject to all the data). Let ℓ_ω be the optimal entropy value subject to all the data and the restriction $\boldsymbol{\lambda}_s = \mathbf{0}$. Using the ER test, we have

$$W = 2|\ell_\omega(\boldsymbol{\lambda}_s = \mathbf{0}) - \ell_\Omega(\boldsymbol{\lambda}' s \text{ are not constrained})| \rightarrow \chi^2_{(K-1)}$$

under the null.

Hypotheses and Models with Stochastic Moments
All statistics and tests described above can be used for all IT estimators (zero-moment models, stochastic moments, and for all levels of α in the α-entropy). However, it is important to note that in all stochastic moments' models (GME, GCE) these tests are "more conservative." By "more conservative" I mean that given a sample of data and a certain null hypothesis, W is smaller and the probability of a large deviation is larger, for the stochastic moments' estimators. To understand that point, recall that when one uses stochastic moments,

as defined in Section 6, the magnitude of the estimated Lagrange multipliers must be, on average, smaller. Therefore, the optimal relative entropy (in the objective) is smaller; the distance between the priors (uniform or non-uniform) and the estimated probabilities (post-data) is smaller (on average) relative to the zero-moment conditions. Looking at the above tests and probability of large deviations, this statement is clear.

8

Concluding Remarks and Related Work Not Surveyed

The basic information quantities and concepts, used within econometrics, for estimation and inference were reviewed. It was also shown that all methods within IEE are special cases of the generic IT estimator of Section 5.2:

$$\text{IT Estimators} = \begin{cases} \widehat{\boldsymbol{p}} = \arg\min\{f(\boldsymbol{p}\|\boldsymbol{q}) = D^{R}_{\alpha+1}(\boldsymbol{p}\|\boldsymbol{q})\} \\ \text{s.t.} \\ g_m(\boldsymbol{y}, \boldsymbol{p}, \boldsymbol{\theta}_1) = [\mathbf{0}];\ m = 1, 2, \ldots, M \\ \sum_{i=1}^{T} p_{\text{i}} = 1;\ i = 1, 2, \ldots, T \text{ and } M \ll T - 1. \end{cases}$$

In Section 3, a number of basic IT quantities were reviewed. These quantities and ideas were employed in following sections. However, within the objectives of the current review, I just discussed estimation and related inference issues. Therefore, a large number of IT related topics that fall within IEE were not discussed. A very short list of these topics is provided below.

First is the literature on IT–GMM. Though, I have discussed it here, this class of methods within IEE, deserves much deeper review. However, the connection of GMM to IT and ME was discussed in detail. To summarize, the original framework of the IT–GMM, uses

$\alpha \to 0$, which is the relative entropy D, together with zero-moment conditions (see Kitamura and Stutzer, 1997 and Imbens et al., 1998). Some other immediate applications include specific cases of the GEL (e.g., Smith, 2004, 2005) and the Euclidean Empirical Likelihood, where the L_2 norm is substituted for the objective in the generic IT-estimator. Though much work is being done on IT-estimators (all with zero-moments conditions) that are directly related to the EL, GEL, and GMM, I did not cover most of it in this review. But I did show exactly where they fit within IEE and some of the philosophy behind it. One of the main motivations for moving toward IT methods is that it is known by now that the GMM estimators may be substantially biased in finite samples, especially so when there are large numbers of unconditional moment conditions. Therefore, a number of alternative IT estimators that are first-order equivalent to the GMM have been proposed. See, for example the nice discussion in Smith (2007) on efficient IT inference in the case of conditional moment restrictions and the work of Kitamura et al. (2004) and Tripathi and Kitamura (2003). They develop an efficient "local" version of the objective in the IT (minimum discrepancy) estimator (EL and GEL) methods that avoid the necessity of explicit estimation of the conditional Jacobian and variance matrices of the conditional moment restrictions, and provide empirical conditional probabilities for the observations.

Second, is the analysis of complex, nonlinear systems where the generalized entropy (Eq. (3.9)) with values of $\alpha > 0$ is used to capture linear and nonlinear dependence among random variables. Quantities such as the Lyapunov exponents (measuring the nonlinearity of a system and whether the system is chaotic or not), fractal and multifractal dimensions and correlation dimensions are just a few examples. All of these quantities describe the amount of information, or information decay (related to entropy rate discussed earlier), in a system and are used to investigate nonlinear (dynamic) systems within parametric and non-parametric frameworks. For example, take the mutual information (defined as the expected information in an outcome of a random draw from Y about an outcome of a random draw from X) version of (3.9) for two discrete random variables X and Y of

dimension N, and for $\alpha = 1$:

$$I_2^R(X;Y) \equiv H_2^R(Y) - [H_2^R(X,Y) - H_2^R(X)].$$

This measure equals zero if and only if X and Y are statistically independent, and it equals $\log(N)$ if and only if $Y = f(X)$, where f can be any linear or nonlinear function. In general, this type of measure is used for any value of α where α is directly related to the system's (embedding) dimension, or where α is related to (multi) fractal dimension in a nonlinear-chaotic system. For details see the large literature on nonlinear systems and chaos. Within econometrics, see for example, Soofi (1994), Maasoumi and Racine (2002), Racine and Maasoumi (2007) and Ullah (2002). For related work on the basic notions of complexity within IT see Cover and Thomas (1991) and Csiszar and Shields (2004).

The third concept is called minimum description length (MDL) and is based on the idea that the best way to extract information from the observed data is when one uses the shortest possible description code. Thus, the best statistical model to fit the data is the one that leads to the shortest description, while taking into account that the model itself must be described (in that code) as well. For a tutorial on MDL see, for example Bryant and Cordero-Brana (2000). For deeper formulation with statistical implications, the relationship between MDL and ML and model selection criteria, such as the BIC, see (Csiszar and Shields, 2004, Section 8) and Clarke (2007).

Fourth, for a deeper and comprehensive discussion of IT see the texts of Cover and Thomas (1991) and MacKay (2003), as well as the original work of Shannon (1948) and Rényi (1970) and the tutorial of Csiszar and Shields (2004). For more recent articles, within the econometric literature, see for example Maasoumi (1993), Soofi and Retzer (2002), Clarke (2007) and Zellner (1999, 2002).

The fifth concept dealing with recent, and evolving, issue of much concern to econometricians is how to identify the informational content of instruments in the case of over identified problems. Stated differently, a common problem facing researchers, within the GMM framework, is what possible set of moments, from a possible set of candidates, should be used in the estimation. This issue was discussed in Section 5.4.3, but

more is needed. For example, one can use the inherent logic of IT to construct such tests. After all, the sample moments enter as constraints within an optimization and the objective is an informational one. Thus, it seems natural to base that choice on the informational content of the moment conditions relative to the desired inferences. This idea is starting to evolve in recent research. Recent work shows that the entropy measure can be used toward this goal. Examples include the work of Hall et al. (2007a) and Hall et al. (2007b). See also the text of Hall (2005) for more detailed discussions.

Sixth, I did not provide here a survey of Bayesian IT models (except for a summary of Zellner's BMOM). There is much research in that area and tight connections between Bayes rule and Bayesian econometrics. This body of research demands its own review. It was just noted here that the ME principle is used often in Bayesian methods as the proper way to assign the most uninformed prior probabilities. For basic reviews and analysis of recent work see Zellner (1996b, 1999, 2007), Clarke (2007) and the annual on Maximum Entropy and Bayesian Methods.

Seventh, the class of regularization methods often used for ill-conditioned data was not discussed here. Though it fits within the generic IT estimator, this class deserves its own review and study. The objective of these models is to extract the signal from the (very noisy and often ill-conditioned, highly collinear) data. To do so, a penalty function is introduced in the dual objective function yielding regularization methods that exhibit relatively good risk characteristics. Within the more traditional maximum-entropy approach, (e.g., Csiszar, 1991, 1996; Donoho et al., 1992; Bercher et al., 1996; Ertl et al., 1996; Gzyl, 1998) regularization methods are used for analyzing such noisy problems. One approach, for example, for recovering the unknown probability distribution/s is to subtract from the entropy objective, or from the negative of the cross-entropy objective, a regularization term penalizing large values of the errors. Typically, the regularization term is chosen to be $\delta \sum_i \varepsilon_i^2$, where ε_i is the error term and $\delta > 0$ is the regularization parameter to be determined prior to the optimization (e.g., Donoho et al., 1992), or is iteratively determined during repeated optimization cycles (e.g., Ertl et al., 1996). However, this approach, or variations of it, works fine when one knows the underlying distribution of the ran-

dom variables (and the true variances). For example, the assumption of normal noise is frequently used. This assumption leads to the χ^2 distribution with a number of (random) constraints minus one degree of freedom. (Note here that the GME, or GCE, can be viewed as "regularization" methods where instead of a direct regularization parameter, the support spaces are specified.) Finally, the solution to this class of noisy models was also approached via the maximum entropy in the mean formulation (e.g., Navaza, 1986; Bercher et al., 1996). In that formulation the underlying philosophy is to specify an additional convex constraint set that bounds the possible solutions.

For a detailed comparison and discussion of other entropy and non-entropy regularization methods, as well as the maximum entropy on the mean, see for example Donoho et al. (1992), (Golan et al., 1996b, Chap. 8), Bercher et al. (1996) and Golan (2001).

Finally, another topic that was covered here but needs much more (both in terms of review and development) is to do with hypothesis tests and IT. Details of many tests and examples of different hypotheses, as well as a brief review of large deviations, were provided. But more survey and more research is needed. The basics of that subjects can be found in Cover and Thomas (1991) and Csiszar and Shields (2004). Application of large deviations in econometrics, within IEE can be found in the work of Stutzer (2000, 2003a,c), Kitamura and Stutzer (2002), and Kitamura (2006).

The above list of topics reflects many of the issues studied within IEE, but is not comprehensive. Interested readers can find good complements for this survey as well as current advances in IEE in recent volumes of the Journal of Econometrics (2002, 2007) and Econometric Reviews (2007). Readers who are interested in IT and Bayesian econometrics/statistics should look at the series of annual proceedings of the Maximum Entropy and Bayesian methods conferences.

References

Agmon, N., Y. Alhassid, and R. D. Levine (1979), 'An algorithm for finding the distribution of maximal entropy'. *Journal of Computational Physics* **30**, 250–259.

Antoine, B., H. Bonnal, and E. Renault (2007), 'On the efficient use of the informational content of estimating equations: Implied probabilities and Euclidean likelihood'. *Journal of Econometrics* **138**, 461–487.

Bayes, R. T. (1763), 'An essay toward solving a problem in the doctrine of chances'. *Philosophical Transactions of the Royal Soceity of London* **53**, 370–418.

Bera, A. K. and Y. Bilias (2002), 'The MM, ME, ML, EL, EF and GMM approaches to estimation: A synthesis'. *Journal of Econometrics* **107**, 51–86.

Bercher, J. F., G. L. Besnerais, and G. Demoment (1996), 'The maximum entropy on the mean method, noise and sensitivity'. In: J. Skiling and S. Sibisi (eds.): *Maximum Entropy and Bayesian Studies*. Kluwer.

Besnerais, L. G., J.-F. Bercher, and G. Demoment (1999), 'A new look at entropy for solving linear inverse problems'. *IEEE Transactions on Information Theory* **45**, 1565–1578.

Boltzmann, L. W. (1871). *Weiner Berichte* **63**, 397–418; 679–711; 712–732. Translated later.

Boole, G. (1854), *An Investigation of the Laws of Thought*. London: MacMillan. Reprinted by Dover Pub., Inc., New York, 1958.

Bryant, P. G. and O. I. Cordero-Brana (2000), 'Model selection using the minimum description length pronciple'. *The American Statistician* **54**, 257.

Burg, J. P. (1975), 'Maximum Entropy and Spectral Analysis'. Ph.D. thesis, Stanford University Press.

Chamberlain, G. (1987), 'Asymptotic efficiency in estimation with conditional moment restrictions'. *Journal of Econometrics* **34**, 305–334.

Chernoff, H. (1952), 'A measure of asymptotic efficiency for tests of hypothesis based on the sum of observations'. *Annals of Mathematical Statistics* **23**, 493–507.

Clarke, B. (2007), 'Information optimality and Bayesian modeling'. *Journal of Econometrics* **138**, 405–429.

Cover, T. M. and J. A. Thomas (1991), *Elements of Information Theory*. New York: John Wiley & Sons.

Cox, R. T. (1946), 'Probability, frequency and reasonable expectation'. *American Journal of Physics* **14**, 1–13.

Cressie, N. and T. R. C. Read (1984), 'Multinomial goodness-of-fit tests'. *Journal of Royal Statistical Soceity B* **46**, 440–464.

Csiszar, I. (1984), 'Sanov property, generalized I-projection and a conditional limit theorem'. *The Annals of Probability* **12**, 768–793.

Csiszar, I. (1991), 'Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems'. *The Annals of Statistics* **19**, 2032–2066.

Csiszar, I. (1996). In: K. M. Hanson and R. N. Silver (eds.): *Maximum Entropy and Bayesian Studies*. Dordrecht: Kluwer.

Csiszar, I. and J. Korner (1981), *Information Theory: Coding Theorems for Discrete*. Budapest: Akademiai Kiado and New York: Academic Press.

Csiszar, I. and P. C. Shields (2004), 'Information theory and statistics: A tutorial'. *Foundations and Trends in Communications and Information Theory* **1**(4), 417–528.

Davis, H. T. (1941), *The Theory of Econometrics*. Indiana: The Principia Press.

Dembo, A. and O. Zeitouni (1998), *Large Deviations Techniques and Applications*. Second edition. Springer Verlag.

DiCiccio, T. J., P. Hall, and J. Romano (1991), 'Empirical likelihood is Bartlett-correctable'. *The Annals of Statistics* **19**, 1053–1991.

Donoho, D. L., I. M. Johnstone, J. C. Hoch, and A. S. Stern (1992), 'Maximum entropy and the nearly black object'. *Journal of the Royal Statistical Society Series B* **54**, 41–81.

Durbin, J. (1960), 'Estimation of parameters in time-series regression models'. *Journal of the Royal Statistical Society Series B* **22**, 139–153.

Ebrahimi, N., M. Habibullah, and E. S. Soofi (1992), 'Testing exponentiality based on Kullback-Leibler information'. *Journal of the Royal Statistical Society Series B* **54**(3), 739–748.

Econometric Reviews (2007). Special issue on IEE (Forthcoming).

Efron, B. and R. J. Tibshirani (1993), *An Introduction to the Bootstrap*. Chapman & Hall.

Ertl, K., W. von der Linden, V. Dose, and A. Weller (1996), 'Maximum entropy based reconstruction of soft X-ray emissivity profiles in W7-AS'. *Nuclear Fusion* **36**(11), 1477–1488.

Esscher, F. (1932), 'On the pProbability function in the collective theory of risk'. *Skandinavisk Aktuariedskrift* **15**, 175–195.

Fano, R. M. (1961), *Transmission of Information: A Statistical Theory of Communication*. New York: John Wiley.

Ferguson, T. S. (1958), 'A method of generating best asymptotically normal estimates with application to estimation of bacterial densities'. *Annals of Mathematical Statistics* **29**, 1046–1062.

Fisher, R. A. (1912), 'On an absolute criterion for fitting frequency curves'. *Messenger of Mathematics* **41**, 155–160.

Fisher, R. A. (1922), 'On the mathematical foundations of theoretical statistics'. *Philosophical Transactions of the Royal Society of London Series A* **222**, 309–368.

Gamboa, F. and E. Gassiat (1997), 'Bayesian methods and maximum entropy for ill-posed problems'. *The Annals of Statistics* **25**, 328–350.

Gerber, H. U. (1980), 'A characterization of certain families of distributions via essche transforms and independence'. *Journal of the American Statistical Association* **75**(372), 1015–1018.

Gibbs, J. W. (1902), *Elementary Principles in Statistical Mechanics*. New Haven, CT: Yale University Press.

Gokhale, D. and S. Kullback (1978), *The Information in Contingency Tables*. New York: Marcel Dekker.

Golan and D. Volker (2001), 'A generalized information theoretical approach to tomographic reconstruction'. *Journal of Physics A: Mathematical and General,* pp. 1271–1283.

Golan, A. (1988), *A discrete-stochastic model of economic production and a model of production fluctuations – theory and empirical evidence*. PhD Dissertation, UC Berkeley.

Golan, A. (2001), 'A simultaneous estimation and variable selection rule'. *Journal of Econometrics* **101**(1), 165–193.

Golan, A. (2002), 'Information and entropy econometrics – Editor's view'. *Journal of Econometrics* **107**(1–2), 1–15.

Golan, A. and H. Gzyl (2002), 'A generalized maxentropic inversion procedure for noisy data'. *Applied Mathematics and Computation* **127**, 249–260.

Golan, A. and H. Gzyl (2003), 'Priors and information-theoretic estimation'. *American Statistical Association Proceedings*.

Golan, A. and H. Gzyl (2006), 'Ill conditioned linear estimation problems: An information theoretic approach with priors'. Working Paper.

Golan, A. and G. Judge (1992), 'Recovering and processing information in the case of underdetermined economic inverse models'. Unpublished Manuscript, University of California, Berkeley.

Golan, A. and G. Judge (1993), *Recovering the parameters and forecasting in the case of ill-posed non-stationary inverse problems*. Unpublished manuscript, University of California, Berkeley.

Golan, A. and G. Judge (1996), 'A maximum entropy approach to empirical likelihood estimationand inference'. Unpublished Paper, University of California, Berkeley. (Presented at the 1997 Summer Econometric Society meetings).

Golan, A., G. Judge, and L. Karp (1996a), 'A maximum entropy approach to estimation and inference in dynamic models or counting fish in the sea with maximum entropy'. *Journal of Economic Dynamics and Control* **20**, 559–582.

Golan, A., G. Judge, and D. Miller (1996b), *Maximum Entropy Econometrics: Robust Estimation with Limited Data*. New York: John Wiley & Sons.

Golan, A., G. Judge, and J. Perloff (1996c), 'A generalized maximum entropy approach to recovering information from multinomial response data'. *Journal of the American Statistical Association* **91**, 841–853.

Golan, A., G. Judge, and J. Perloff (1997), 'Estimation and inference with censored and ordered multinomial response data'. *Journal of Econometrics* **79**, 23–51.

Golan, A., G. G. Judge, and S. Robinson (1994), 'Recovering information from incomplete or partial multisectoral economic data'. *Review of Economics and Statistics* **76**(3), 541–549.

Golan, A., L. Karp, and J. M. Perloff (2000), 'Estimating firms' mixed price and advertising strategies: Coke and Pepsi'. *Journal of Business Economic Statistics* **18**(4), 398–409.

Golan, A., J. Lane, and E. McEntarfe (2007), 'The dynamics of worker reallocation within and across industries'. *Economica* **74**, 1–20.

Golan, A. and J. Perloff (2002), 'Comparison of maximum entropy and higher-order entropy estimators'. *Journal of Econometrics* **107**(1–2), 195–211.

Golan, A., J. Perloff, and Z. Shen (2001), 'Estimating a demand system with non-negativity constraints: Mexican meat demand'. *Review of Economics and Statistics* **83**(3), 541–550.

Golan, A. and S. Vogel (2000), 'Estimation of non-stationary social accounting matrix coefficients with supply-side information'. *Economic Systems Research* **12**(4), 447–471.

Good, I. J. (1963), 'Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables'. *Annals of Mathematical Statistics* **34**, 911–934.

Greene, W. H. (2008), *Econometric Analysis*. Sixth edition. Prentice Hall.

Gzyl, H. (1993), *The Maximum Entropy Method*. Singapore: Wirld Scientific Publishing.

Gzyl, H. (1998). In: G. Erickson (ed.): *Maximum Entropy and Bayesian Studies*. Dordrecht: Kluwer.

Hall, A. R. (2005), *Generalized Method of Moments*, Advanced Texts in Econometrics Series. Oxford University Press.

Hall, A. R., A. Inoue, K. Jana, and C. Shin (2007a), 'Information in eneralized method of moments estimation and entropy based moment selection'. *Journal of Econometrics* **138**, 488–512.

Hall, A. R., A. Inoue, and C. Shin (2007b), 'Entropy based moment selection in the presence of weak identification'. *Econometric Reviews*. (Forthcoming).

Hall, P. (1990), 'Pseudo-likelihood theory for empirical likelihood'. *The Annals of Statistics* **18**, 121–140.

Hansen, L. P. (1982), 'Large sample properties of generalized methods of moments estimators'. *Econometrica* **50**, 1029–1054.

Hansen, L. P., J. Heaton, and A. Yaron (1996), 'Finite-sample properties of some alternative GMM estimators'. *Journal of Business and Economic Statistics* **14**, 262–280.

Hanson, K. M. and R. N. Silver (1996), *Maximum Entropy and Bayesian Methods*. Dordrecht: Kluwer.

Hartley, R. V. L. (1928), 'Transmission of information'. *Bell System Technical Journal,* pp. 535–563.

Hellerstein, J. K. and G. W. Imbens (1999), 'Imposing moment restrictions from auxiliary data by weighting'. *The Review of Economics and Statistics* **81**, 1–14.

Holste, D., I. Grobe, and H. Herzel (1998), 'Bayes' estimators of generalized entropies'. *Journal of Physics A: Mathematical and General* **31**, 2551–2566.

Imbens, G. and R. Spady (2001), 'The performance of empirical likelihood and its generalizations'. Unpublished working paper, Department of Economics, University of California, Berkeley.

Imbens, G. W. (1997), 'One-step estimators for over-identified generalized method of moments models'. *Review of Economic Studies* **64**, 359–383.

Imbens, G. W. (2002), 'Generalized method of moments and empirical likelihood'. *Journal of Business & Economic Statistics* **20**, 493–507.

Imbens, G. W., P. Johnson, and R. H. Spady (1998), 'Information-theoretic approaches to inference in moment condition models'. *Econometrica* **66**, 333–357.

Jaynes, E. T. (1957a), 'Information theory and statistical mechanics'. *Physics Review* **106**, 620–630.

Jaynes, E. T. (1957b), 'Information theory and statistical mechanics II'. *Physics Review* **108**, 171–190.

Jaynes, E. T. (1963), 'Information theory and statistical mechanics'. In: K. W. Ford (ed.): *Statistical Physics*. New York: W. A. Benjamin, Inc., MIT Press, pp. 181–218.

Jaynes, E. T. (1978a), 'Information theory and statistical mechanics II'. In: K. W. Ford (ed.): *Statistical Physics*. New York: W.A. Benamin, Inc., pp. 181–218.

Jaynes, E. T. (1978b), 'Where do we stand on maximum entropy'. In: R. D. Levine and M. Tribus (eds.): *The Maximum Entropy Formalism*. MIT Press, pp. 15–118.

Jaynes, E. T. (1984), 'Prior information and ambiguity in inverse problems'. In: D. W. McLaughlin (ed.): *Inverse Problems, SIAM Proceedings*. Providence, RI: American Mathematical Society, pp. 151–166.

Jaynes, E. T. (2003), *Probability Theory: The Logic of Sciences*. Cambridge University Press.

Jeffreys, H. (1939), *Theory of Probability*. Oxford: Clarendon Press (later editions: 1948, 1961, 1967, 1988).

Journal of Econometrics (2002), 'Special issue on IEE'. **107**, 1–374.

Journal of Econometrics (2007), 'Special issue on IEE'. **138**, 379–585.

Judge, G. and R. C. Mittelhammer (2007), 'Estimation and inference in the case of competing sets of estimating equations'. *Journal of Econometrics* **138**, 513–531.

Kitamura, Y. (2006), 'Empirical likelihood methods in econometrics: Theory and practice'. Cowles Foundation Discussion Paper No. 1569.

Kitamura, Y. and T. Otsu (2005), 'Minimax estimation and testing for moment condition models via large deviations'. Manuscript, Department of Economics, Yale University.

Kitamura, Y. and M. Stutzer (1997), 'An information theoretic alternative to generalized method of moments estimation'. *Econometrica* **65**(4), 861–874.

Kitamura, Y. and M. Stutzer (2002), 'Corrections between entropic and linear projections in asset pricing estimation'. *Journal of Econometrics* **107**, 159–174.

Kitamura, Y., G. Tripathi, and H. Ahn (2004), 'Empirical likelihood based inference in conditional moment restriction models'. *Econometrica* **72**, 1667–1714.

Koehler, K. J. and K. Larntz (1980), 'An empirical investigation of goodness-of-fit statistics for sparse multinomial'. *Journal of the American Statistical Association* **75**, 336–344.

Kullback, S. (1954), 'Certain inequalities in information theory and the Cramer-Rao inequality'. *The Annals of Mathematical Statistics* **25**, 745–751.

Kullback, S. (1959), *Information Theory and Statistics*. New York: John Wiley & Sons.

Kullback, S. and R. A. Leibler (1951), 'On information and sufficiency'. *Annals of Mathematical Statistics* **22**, 79–86.

LaFrance, J. T. (1999), 'Inferring the nutrients with prior information'. *American Journal of Agricultural Economics* **81**, 728–734.

Laplace, P. S. (1774), 'Memoire sur la probabilite des causes par les evenemens'. *Mémoires de l'Académie Royale des Sciences* **6**, 621–656. Reprinted in Laplace (1878-1912), **8**, 27–65.

Levine, R. D. (1980), 'An information theoretical approach to inversion problems'. *Journal of Physics A* **13**, 91–108.

Levine, R. D. and M. Tribus (1979), *The Maximum Entropy Formalism*. Cambridge, MA: MIT Press.

Lindley, D. V. (1956), 'On a measure of the information provided by an experiment'. *Annals of Mathematics* **27**, 986–1005.

Maasoumi, E. (1993), 'A compendium to information theory in economics and econometrics'. *Econometric Reviews* **12**, 137–181.

Maasoumi, E. and J. Racine (2002), 'Entropy and predictability in the stock market returns'. *Journal of Econometrics* **107**, 291–312.

MacKay, D. J. C. (2003), *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press.

Maxwell, J. C. (1859), 'Illustrations of the dynamical theory of gases'. In: W. D. Niven (ed.): *Collected Works*. London (1890), Vol. I: pp. 377–409.

Maxwell, J. C. (1876), 'On Boltzmann's theorem on the average distribution of energy in a system of material points'. In: W. D. Niven (ed.): *Collected Works*. London (1890), Vol. II: pp. 713–741.

McFadden, D. (1974), 'Conditional logit analysis of qualitative choice behavior'. In: P. Zarembka (ed.): *Frontiers of Econometrics*. New York: Academic Press, pp. 105–142.

Miller, D. J. (1994), Entropy and information recovery in linear economic models. PhD Dissertation, UC Berkeley.

Mittelhammer, R. and S. Cardell (1996), *On the Consistency and Asymptotic Normality of the Data Constrained GME Estimator of the GML* (mimeo). Pullman, WA: Washington State University.

Mittelhammer, R. C., G. G. Judge, and D. J. Miller (2000), *Econometric Foundations*. Cambridge: Cambridge University Press.

Navaza, J. (1986), 'The use of non-local constraints in maximum-entropy electron density reconstruction'. *Acta Crystallographica* **A42**, 212–223.

Newey, W. and D. McFadden (1994), 'Large sample estimation and hypothesis testing'. In: Engle and McFadden (eds.): *The Handbook of Econometrics* Vol. 4. New York: North-Holland.

Newey, W. K. and R. J. Smith (2004), 'Higher order properties of GMM and generalized empirical likelihood estimators'. *Econometrica* **72**, 219–255.

Neyman, J. and E. S. Pearson (1928a), 'On the use and interpretation of certain test criteria for purposes of statistical inference: Part I'. *Biomentrica* **20A**, 175–240.

Neyman, J. and E. S. Pearson (1928b), 'On the use and interpretation of certain test criteria for purposes of statistical inference: Part II'. *Biomentrica* **20A**, 263–294.

O'Sullivan, F. (1986), 'A statistical perspective on ill-posed inverse problems'. *Statistical Science* **1**, 502–527.

Otsu, T. (2006), 'Generalized empirical likelihood under weak identification'. *Econometric Theory* **22**, 513–527.

Owen, A. (1988), 'Empirical likelihood ratio confidence intervals for a single functional'. *Biometrika* **75**(2), 237–249.

Owen, A. (1990), 'Empirical likelihood ratio confidence regions'. *The Annals of Statistics* **18**(1), 90–120.

Owen, A. (1991), 'Empirical likelihood for linear models'. *The Annals of Statistics* **19**(4), 1725–1747.

Owen, A. (2001), *Empirical Likelihood*. Chapman & Hall/CRC.

Pearson, K. (1894), 'Contribution to the mathematical theory of evolution'. *Philosophical Transactions of the Royal Society of London Series A* **185**, 71–110.

Pearson, K. (1900), 'On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling'. *Philosophical Magazine Series* **50**, 157–175.

Perloff, J., L. Karp, and A. Golan (2007), *Estimating Market Power and Strategies*. Cambridge University Press.

Pukelsheim, F. (1994), 'The three sigma rule'. *American Statisticia* **48**, 88–91.

Qin, J. and J. Lawless (1994), 'Empirical likelihood and general estimating equations'. *The Annals of Statistics* **22**, 300–325.

Racine, J. and E. Maasoumi (2007), 'A versatile and robust metric entropy test of time-reversibility, and other hypotheses'. *Journal of Econometrics* **138**, 547–567.

Ramalho, J. J. S. and R. J. Smith (2002), 'Generalized empirical likelihood non-nested tests'. *Journal of Econometrics* **107**, 99–125.

Rényi, A. (1961), 'On measures of entropy and information'. In: *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability,* Vol. I. Berkeley: University of California Press, pp. 547–561.

Rényi, A. (1970), *Probability Theory*. Amsterdam: North-Holland.

Sanov, I. N. (1961), 'On the probability of large deviations of random variables'. *Selected Translations in Mathematical Statistics and Probability* **I**, 213–244.

Sargan, J. D. (1958), 'The estimation of economic relationships using instrumental variables'. *Econometrica* **26**, 393–415.

Sargan, J. D. (1959), 'The estimation of relationships with autocorrelated residuals by the use of instrumental variables'. *Journal of the Royal Statistical Society Series B* **21**, 91–105.

Schennach, S. M. (2004), 'Exponentially tilted empirical likelihood'. Discussion Paper, University of Chicago.

Shannon, C. E. (1948), 'A mathematical theory of communication'. *Bell System Technical Journal* **27**, 379–423; 623–656.

Shore, J. E. and R. Johnson (1980), 'Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy'. *IEEE Transactions on Information Theory* **IT-26**(1), 26–37.

Skilling, J. (1989a), 'The axioms of maximum entropy'. In: J. Skilling (ed.): *Maximum Entropy and Bayesian Methods in Science and Engineering*. Dordrecht: Kluwer Academic, pp. 173–187.

Skilling, J. (1989b), 'Classic maximum entropy'. In: J. Skilling (ed.): *Maximum Entropy and Bayesian Methods*. Dordrecht: Kluwer Academic Publishing, pp. 45–52.

Smith, R. J. (1997), 'Alternative semi-parametric likelihood approaches to generalized method of moments estimation'. *Economic Journal* **107**, 503–519.

Smith, R. J. (2000), 'Empirical likelihood estimation and inference'. In: M. Salmon and P. Marriott (eds.): *Applications of Differential Geometry to Econometrics*. Cambridge: Cambridge University Press, pp. 119–150.

Smith, R. J. (2004), *GEL criteria for moment condition models*. University of Warwick.

Smith, R. J. (2005), Local GEL methods for conditional moment restrictions. Working Paper, University of Cambridge.

Smith, R. J. (2007), 'Efficient information theoretic inference for conditional moment restrictions'. *Journal of Econometrics* **138**, 430–460.

Solomon, H. and M. A. Stephens (1978), 'Approximations to density functions using pearson curves'. *Journal of the American Statistical Association* **73**(361), 153–160.

Soofi, E. S. (1994), 'Capturing the intangible concept of information'. *Journal of the American Statistical Association* **89**(428), 1243–1254.

Soofi, E. S. (1996), 'Information theory and bayesian statistics'. In: D. A. Berry, K. M. Chaloner, and J. K. Geweke (eds.): *Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellnge*. New York: Wiley, pp. 179–189.

Soofi, E. S. (2000), 'Principal information theoretic approaches'. *Journal of the American Statistical Association* **95**(452), 1349–1353.

Soofi, E. S. and J. J. Retzer (2002), 'Information indices: Unifications and applications'. *Journal of Econometrics* **107**, 17–40.

Stutzer, M. J. (2000), 'Simple entropic derivation of a generalized black-scholes model'. *Entropy* **2**, 70–77.

Stutzer, M. J. (2003a), 'Fund managers may cause their benchmarks to be priced 'Risks''. *Journal of Investment Management*. (Reprinted in Fong (ed.): *The World of Risk Management*, World Scientific Publishing, 2006).

Stutzer, M. J. (2003b), 'Optimal asset allocation for endowments: A large deviations approach'. In: Satchell and Scowcraft (eds.): *Advances in Portfolio Construction and Implementation*. Butterworth-Heineman.

Stutzer, M. J. (2003c), 'Portfolio choice with endogenous utility: A large deviations approach'. *Journal of Econometrics* **116**, 365–386.

Tikochinsky, Y., N. Z. Tishby, and R. D. Levine (1984), 'Consistent inference of probabilities for reproducible experiments'. *Physics Review Letters* **52**, 1357.

Tobias, J. and A. Zellner (2001), 'Further results on Bayesian method of moments analysis of the multiple regression model'. *International Economic Review* **42**(1), 121–139.

Tripathi, G. and Y. Kitamura (2003), 'Testing conditional moment restrictions'. *Annals of Statisics* **31**, 2059–2095.

Tsallis, C. (1988), 'Possible generalization of Boltzmann-Gibbs statistics'. *Journal of Statistical Physics* **52**, 479.

Ullah, A. (2002), 'Uses of entropy and divergence measures for evaluating econometric approximations and inference'. *Journal of Econometrics* **107**, 313–326.

Vasicek, O. (1976), 'A test for normality based on sample entropy'. *Journal of the Royal Statistical Society Series B* **38**(1), 54–59.

von Baeyer, H. C. (2004), *Information: The New Language of Science*. Harvard University Press.

Wu, X. and J. M. Perloff (2007), 'GMM estimation of a maximum entropy distribution with interval data'. *Journal of Econometrics* **138**, 532–546.

Zellner, A. (1988), 'Optimal information processing and bayes theorem'. *American Statistician* **42**, 278–284.

Zellner, A. (1991), 'Bayesian methods and entropy in economics and econometrics'. In: W. T. Grandy Jr. and L. H. Schick (eds.): *Maximum Entropy and Bayesian Methods*. Amsterdam: Kluwer, pp. 17–31.

Zellner, A. (1994), 'Bayesian method of moments/instrumental variable (BMOM/IV) analysis of mean and regression models'. *1994 Proceedings of the Section on Bayesian Statistical Science of the American Statistical Association*. 1995 (Paper presented at the August 1994 ASA meetings).

Zellner, A. (1996a), 'Bayesian method of moments/instrumental variable (BMOM/IV) analysis of mean and regression models'. In: J. C. Lee, W. C. Johnson, and A. Zellner (eds.): *Modeling and Prediction: Honoring Seymour Geisser*. Springer-Verlag, pp. 61–75.

Zellner, A. (1996b), 'Models, prior information, and Bayesian analysis'. *Journal of Econometrics* **75**, 51–68.

Zellner, A. (1997), 'Bayesian method of moments (BMOM): Theory and applications'. In: T. B. Fomby and R. C. Hill (eds.): *Advances in Econometrics: Applying Maximum Entropy to Econometric Problems*. Greenwich: JAI Press, pp. 85–105.

Zellner, A. (1999), 'New information-based econometric methods in agricultural economics: Discussion'. *American Journal of Agricultural Economics* **81**, 742–746.

Zellner, A. (2002), 'Information processing and Bayesian analysis'. *Journal of Econometrics* **107**, 41–50.

Zellner, A. (2007), 'Some aspects of the history of Bayesian information processing'. *Journal of Econometrics* **138**, 388–404.