APPROXIMATING MULTIVARIATE DISTRIBUTIONS WITH CUMULATIVE RESIDUAL ENTROPY: A STUDY ON DYNAMIC INTEGRATED CLIMATE-ECONOMY MODEL

BY

MUHAMMED SUTCU

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Industrial Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2015

Urbana, Illinois

Doctoral Committee:

Associate Professor Ramavaparu S. Sreenivas, Chair
Associate Professor Ali E. Abbas, Director of Research
Associate Professor Carolyn L. Beck
Professor Rizwan Uddin

ABSTRACT

The complexity of real world decision problems is exacerbated by the need to make decisions with only partial information. How to model and make decisions in situations where only partial preference information is available is a significant challenge in decision analysis practice. In most of the studies, the probability distributions are approximated by using the mass function or density function of the decision maker. In this dissertation, our aim is to approximate representative probability and utility functions by using cumulative distribution functions instead of density/mass functions. This dissertation consists of four main sections. The first two sections introduce the proposed methods based on cumulative residual entropy, the third section compares the proposed approximation methods with the methods in information theory literature, and the final section of the dissertation discusses the cumulative impact of integrating uncertainty into the DICE model.

In the first section of the dissertation, we approximate discrete joint probability distributions using first-order dependence trees as well as the recent concept of cumulative residual entropy. We formulate the cumulative residual Kullback-Leibler (KL)-divergence and the cumulative residual mutual information measures in terms of the survival function. We then show that the optimal first-order dependence tree approximation of the joint distribution using the cumulative Kullback-Leibler divergence is the one with the largest sum of cumulative residual mutual information pairs.

In the second part of the dissertation, we approximate multivariate probability distributions with cumulative probability distributions rather than density functions in maximum entropy formulation. We use the discrete form of maximum cumulative residual entropy to approximate

joint probability distributions to elicit multivariate probability distributions using their lower order assessments.

In the third part of the dissertation, we compare several approximation methods to test the accuracy of different approximations of joint distributions with respect to the true distribution from the set of all possible distributions that match the available information. A number of methods have beeb presented in the literature for joint probability distribution approximations and we specifically compare those approximation methods that use information theory to approximate multivariate probability distributions.

Finally, we study whether uncertainty significantly affects decision making especially in global warming policy decisions and integrate climatic and economic uncertainties into the DICE model to ascertain the cumulative impact of integrating uncertainty on climate change by applying cumulative residual entropy into the DICE model.

To my lovely wife NUR

ACKNOWLEDGEMENTS

I would like to express my special appreciation and thanks to my advisor Professor Dr. Ali Abbas, you have been a tremendous mentor for me. I would like to thank you for encouraging my research and for allowing me to grow as a research scientist. Your advice on both research as well as on my career have been priceless.

I would also like to thank my committee members, Professor Ramavarapu "RS" Sreenivas, Professor Carolyn Beck, Professor Rizwan Uddin for serving as my committee members even at hardship. I also want to thank you for their encouragement, and insightful comments.

During the period of five years, many friends are helpful to color my life. I have to acknowledge all my colleagues in research room 205 for their assistances in many aspects. They are Andrea Hupman, Sara Behdad, Yerkin Abdildin, Zhengwei Sun, James Schreiner, Xi Chen, Yuan Zhao, Aida Williams, and Aaron Joseph. I am also grateful the faculty and staff of the Industrial and Enterprise Systems Engineering department. Special thanks go to Holly Kizer, Amy Summers and Kathleen Ricker.

To my parents in Turkey who have waited so long for me to finish and prayed all the time for me, I could not find a word in both languages English or Turkish which could describe how much I am grateful to you.

Finally, for the special ones who have been with me no matter where or when, for their unconditional love, patience and support: my beloved wife NUR and my two children Sare Ayse and "…".Nothing in a simple paragraph can express my feelings, the love I have for them all.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION **1**

1.1. Overview 1

1.2. Decision Making and Climate Change 3

1.3. DICE Model 5

1.4. Climate Change and Uncertainty 7

1.5. Problem Definition 8

1.6. Approximation with Partial Information 10

1.7. Contributions of the Dissertation 17

CHAPTER 2: REVIEW OF BASIC CONCEPTS, TERMINOLOGY, AND NOTATIONS OF INFORMATION THEORY **19**

2.1. Introduction 19

2.2. Basic Concepts, Terminology, and Notations of Information Theory in Probability 19

2.3. Entropy of Discrete Random Variables 20

2.4. Entropy of Continuous Random Variables 23

2.5. The differences between Shannon Entropy and Cumulative Residual Entropy 26

2.6. Relative Entropy and Mutual Information 29

2.7. Maximum Entropy Formulations 32

2.8. Review of Previous Work on Related Research 36

CHAPTER 3: FIRST ORDER DEPENDENCE TREES WITH CUMULATIVE RESIDUAL ENTROPY **39**

3.1. Introduction 39

3.2. Cumulative Residual KL-Divergence and Mutual Information 40

3.3. First Order Dependence Trees using Cumulative Residual Entropy 41
3.4. Monte Carlo Simulation for CRE First Order Dependence Tree 47
3.5. Comparison of CRE Based First Order Dependence Trees with Chow and Liu's First Order Dependence Tree 49
3.6. Decision Problems where CRE Based First Order Dependence Trees are More Suitable or Preferable 64
3.7. Summary of the Results 64
3.8. Conclusion 65
CHAPTER 4: CONSTRUCTION OF JOINT PROBABILITY DISTRIBUTIONS BASED ON MAXIMUM CUMULATIVE RESIDUAL ENTROPY 67
4.1. Introduction 67
4.2. Maximum Cumulative Residual Entropy.....67
4.3. Application of Maximum CRE using Monte Carlo Simulation 73
4.4. Comparison of Maximum Entropy and Maximum Cumulative Residual Entropy 77
4.5. Summary of the Results 90
4.6. Conclusion 90
CHAPTER 5: COMPARISON OF INFORMATION THEORY BASED APPROXIMATION METHODS 92
5.1. Introduction 92
5.2. Information Theory based Approximation Methods 93
5.3. Approximating the Four Binary Variable Probability Distribution 97
5.4. Approximating the Four Three-outcome Variable Joint Distributions 105
5.5. Summary of the Results 112
CHAPTER 6: APPLICATION TO DICE MODEL: APPLYING CUMULATIVE RESIDUAL ENTROPY INTO THE DICE MODEL 114
6.1. Introduction 114

6.2. Dynamic Integrated Climate Economy Model 114
6.3. Global Warming Policies used in DICE Model 117
6.4. Analysis of the Model 118
6.5. Results of the Analysis 121
CHAPTER 7: CONCLUSIONS 133
7.1. Summary of the Results 133
7.2. Future Work 134
REFERENCES 136

CHAPTER 1: INTRODUCTION

1.1. Overview

The huge amount of carbon based fuel consumption is one of the most urgent challenges facing the earth. Over the past century, there is increasing evidence that the atmospheric concentrations of greenhouse gases, mainly carbon dioxide (CO2), methane (CH4) and nitrous oxide (N2O), have risen significantly. Measurements of CO_2 show that CO_2 concentration in the atmosphere has increased from about 330 ppm in 1960 to about 389 ppm in 2010 and reached the 400ppm milestone on May 9, 2013, up from around 280ppm ongoing rise in the CO_2 concentration before the Industrial Revolution (EPA, 2013). That is a huge increase (around 43%) and may lead to catastrophic global warming and climate change.

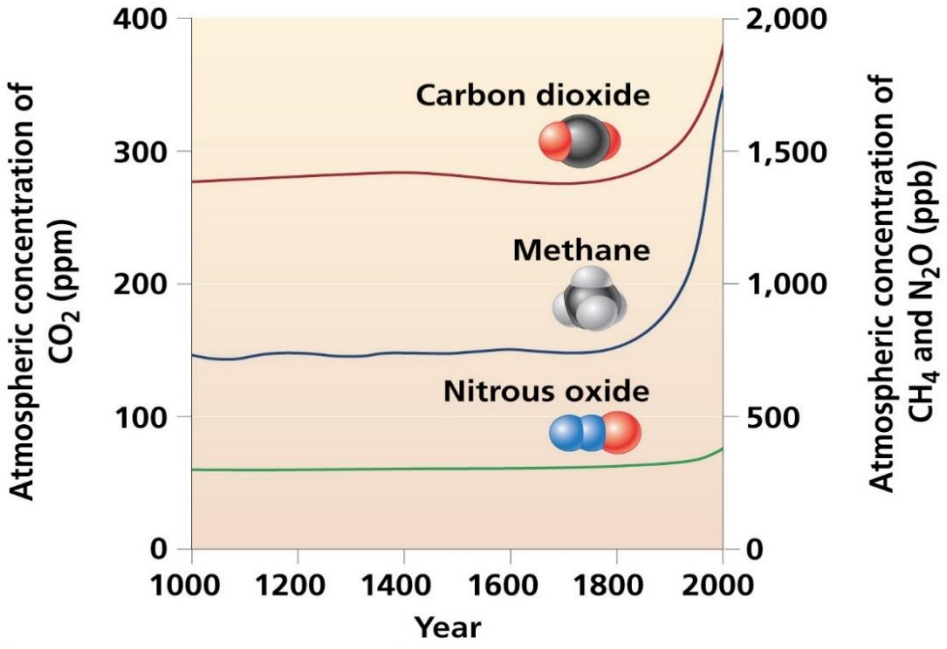


Figure 1.1: Atmospheric Concentration of Greenhouse Gases from 1000 to 2000

Human activities, particularly the burning of fossil fuels to produce energy, tropical deforestation, industrial processes, and some agricultural practices, have emerged as one of the defining factors for increasing concentration of greenhouse gases in the atmosphere. An overwhelming amount of research (Fisher and Narain, 2003; Kunreuther et.al., 2013; Peterson, 2006; Parmeson and Yohe, 2003; Walther et.al., 2003; Tol, 2002a, 2002b; Stern et.al., 2006; Stern, 2008, academic books (Nordhaus, 1994; Nordhaus and Boyer, 2000; Nordhaus, 2007,2008, and 2013), and reports of national and international environmental organizations and agencies such as the Intergovernmental Panel on Climate Change (IPCC), the United States Environmental Protection Agency (EPA), the World Nature Organization (WNO), and the European Environment Agency (EEA) point out the potential impacts of climate change and global warming on water resources, food production, health, economy and the environment such as falling crop yields, rising sea levels, increasing air pollution, decreasing in water availability, increasing frequency of heavy precipitation events, and rising number of species facing extinction.

Increasing concentrations of CO2 and other anthropogenic greenhouse gases, which act as an insulator or blanket above the earth, intensify the naturally-occurring greenhouse effect and accelerate the increase in the earth's global temperature in direct proportion to rising carbon dioxide levels. The global average temperature shows an increase of approximately 0.778°C since the early 20th Century (EPA, 2014). The red line in Figure 1.2 indicates the rise in surface air temperatures from 1955 to 2010 in degrees Celsius, whereas the black line shows a corresponding increase in atmospheric carbon dioxide concentrations in parts per million (ppm).

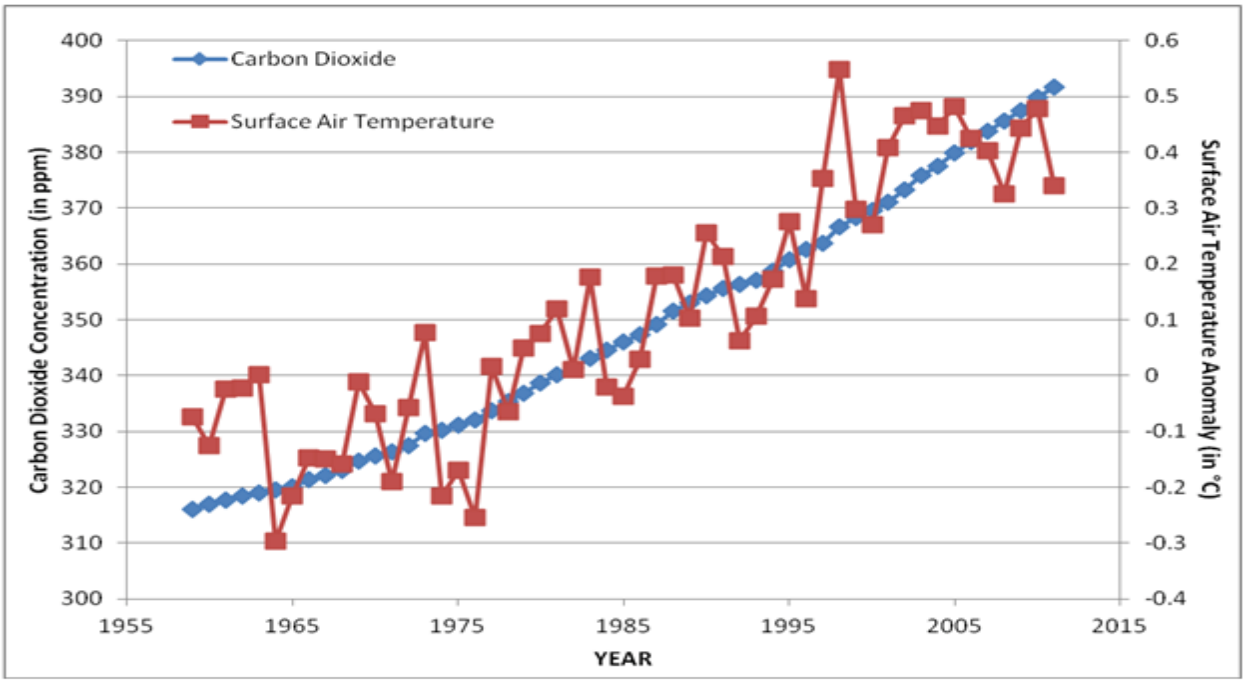


Figure 1.2: Global annual average Surface Air Temperature and Carbon Dioxide Concentration

1.2. Decision Making and Climate Change

The issue of climate change has proven one of the most controversial, widely-discussed, and difficult decision problems involving significant scientific and socio-economic uncertainties (Nordhaus, 1991; Tol, 1997; Stern, 2006; Baker, 2009). The changing climate impacts society and ecosystems in a broad variety of ways, causing unpredictable and often catastrophic precipitation events and/or drought, influencing agricultural crop yields, creating new health concerns for humans and animals, transforming and damaging forests and other ecosystems in alarming ways, threatening our access to critical natural resources such as water, and even creating potential implications for national and global security.

Both advances in climate change research and the adverse effects it has demonstrated have revealed a great need for urgent action—specifically, creating more ambitious targets for limitation and reduction of greenhouse gas emissions.

Scientists actively work to understand past and future climate trends by using observations and theoretical models to match past climate data, make future projections, and link causes and effects in climate change. Most climate change literature focuses on integrated assessment models (IAM) to solve the economic impacts of global warming and slow climate change through the reduction of emissions (Stern, 1977a, 1977b; Anthoff and Tol, 2009; Hope et.al., 2006; Manne et.al., 1995 & 2006). IAM models are advantageous in that that they can address important issues such as the efficient allocation of abatement problems and accepted damages by specifying the costs and benefits of various abatement policies, using a detailed description of both economic and environmental improvements.

There are three main integrated assessment models used by the EPA (United States Environmental Protection Agency); DICE (Dynamic Integrated Climate-Economy), FUND (Climate Framework for Uncertainty, Negotiation and Distribution), and PAGE (Policy Analysis of the Greenhouse Effect). The DICE model is an integrated economic and geophysical model of the economics of climate change developed by Nordhaus and colleagues (Norhaus, 1977a, 1977b). The FUND model, originally developed by Richard Tol (1997) and improved by David Anthoff and Richard Tol (2009), performs cost-benefit and cost-effectiveness analyses of greenhouse gas emission reduction policies and examines equity of climate change and climate policy. The PAGE model, developed by Chris Hope with John Anderson, Paul Wenman, and Erica Plambeck (2006), projects future increases in global mean temperature (GMT), the economic costs of damages caused by climate change, implementation of mitigation policies, and adaptation measures.

Other renowned models and analyses include the Stern Review, a comprehensive analysis of the economic aspects of global warming released for the British government on 2006 by economist Nicholas Stern; and the MERGE model, developed by Manne et.al. (1995 and 2004),

which explicitly accounts for the economy wide impacts of rising energy costs by considering alternative sources of energy supply such as hydroelectricity, nuclear energy, and fossil fuels such as oil, natural gas, and coal.

Most of these models are purely deterministic approaches. The DICE model, which provides an economic analysis of the problem of global warming, is the focus of this dissertation because DICE is an open access model that solves on an EXCEL spread sheet or GAMS program code. It therefore has a large audience familiar with the basics of the model. Also, the US government especially EPA uses the DICE model, in combination with FUND and PAGE, to determine the social cost of carbon. Due to the reasons discussed above, we choose DICE model in our dissertation.

1.3. DICE Model

The Dynamic Integrated Climate-Economy model (DICE) is an integrated assessment model of climate change developed by William Nordhaus and colleagues that integrates both the economic costs and benefits of greenhouse gas controls with an aggregate model linking economic growth with climate change to reduce emission and slow greenhouse warming (Norhaus, 1977a, 1977b). A simplified analytical and empirical model that represents the economics, policy, and scientific aspects of climate change, the DICE model attempts to quantify how the atmospheric concentration of CO_2 negatively affects economic output through its impact on global average surface temperature. First proposed in a discussion paper for the Cowles Foundation (Nordhaus, 1991) and iterated over the past two decades (Nordhaus, 1994; Nordhaus and Boyer, 2000; Nordhaus, 2007,2008), the most up-to-date version of the DICE model was released with current discussion in 2013 (Nordhaus, 2013).

DICE is mainly a policy optimization model with an economic objective function measuring the economic welfare of consumers or producers. The DICE model includes estimates of both the costs of reducing carbon dioxide emissions and the long term future climate impacts from climate change, which enables benefits and costs of carbon dioxide emissions to be weighed in order to optimize controls in the near term, thus maximizing the welfare function to evaluate alternative policies (Figure 6-2).

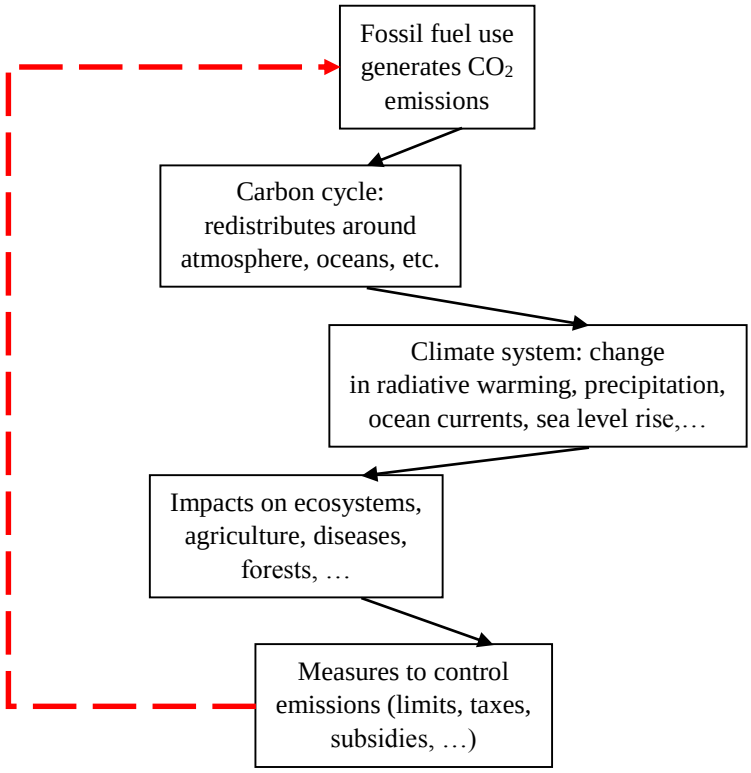


Figure 1.3: Schematic flow chart of a full integrated assessment model for climate change science, economics, and policy (Nordhaus, 2013)

The social cost of carbon is estimated using Integrated Assessment Models (IAMs), of which Nordhaus' DICE is the oldest and one of the best respected models in the literature. Our goal in this dissertation is to incorporate uncertainty into the DICE climate change model.

1.4. Climate Change and Uncertainty

The impact of carbon emissions on the environment is directly observable, and the equivalent monetary values determined for them are very costly: approximately $18 per ton of CO2 increasing by about 2.1% per year to $50 in 2055 (Newbold et.al., 2010). Therefore, it is crucial to take actions for slowing global warming carefully because carbon emission is very expensive and the total cost of carbon emission would amount to hundreds of billions of dollars. However, parameters such as population, CO2 emission, and discount rate create uncertainties in determining both the costs of emissions reductions and the damages from climate change that may not be resolved in this century. Therefore, it is important to analyze the environmental policies by integrating uncertain parameters into the DICE model because in DICE model there are eight uncertain parameters (Nordhaus, 1998; Pizer, 1999; Nordhaus, 2013) which are assumed as deterministic and their mean values are used in the DICE model.

The importance of addressing the element of uncertainty in scientific calculations is such that ignoring or avoiding it could very well lead to under- or overestimating results. In some decision or situations, it is preferable to avoid/ignore uncertainty, while in some cases it is completely eliminated. However, in most economic models, but particularly in those addressing climate change impact, one or more parameters may be highly sensitive in the sense that a slight change of its value results in a significant change in output. Hence proper modeling of uncertainty is an important aspect in science. Several studies have attempted to incorporate uncertainty into the climate change models to deal with global environmental change (Peck and Teisberg, 1993; Heal and Kristrom, 2002; McInerney and Keller, 2008; Baker and Shittu, 2008; Baker, 2009; Schmidt et.al, 2011; Funke and Paetz, 2011;Babonneau et.al., 2011; Haurie et.al., 2012; Keppo and van der Zwann, 2012) .

Like most integrated assessment models, the original DICE model is also deterministic and cannot determine the optimal policy unless the decision maker knows (or is assumed to know) the climate's response to emissions with certainty. However, accurate evaluation of investments in climate change mitigation must take climatic and economic uncertainties into consideration because climate change, long-term economic development, and their interactions are highly uncertain. Therefore, it is important to incorporate uncertainty into the usual cost-benefit climate change models.

A few studies have tried to incorporate uncertainty into the DICE model to evaluate climate change policies. Pizer (1999) uses a non-recursive stochastic programming approach for determining optimal climate change policy under uncertainty. Newbold and Daigneault (2009) run several simulations to explore the impact of uncertainty on future climate and economic trends by constructing two probability density functions about climate uncertainty. Sokolov et al. (2009) and Webster et al. (2012) used Monte Carlo simulation methods and estimated climate uncertainty conditional on different policy scenarios. Hu et.al. (2012) apply a robust simulation approach, which finds the worst-case performances to evaluate environmental policies by assuming the uncertain parameters of the DICE model follow a multivariate normal distribution.

1.5. Problem Definition

Nordhaus (2013) used the DICE model to evaluate and compare a number of different environmental policies, including (i) baseline scenario, (ii) optimal tax scenario, (iii) limit temperature increase to 2 Celsius degree, (iv) Stern scenario, and (v) Copenhagen Accord scenario. Nordhaus pointed out eight critical uncertain parameters in the DICE model and we have selected these eight major parameters for further study: uncertainties about the growth rate of total factor productivity (ga0), the rate of de-carbonization (dsig), the asymptotic population growth

(popasym), the cost of the backstop technology (pback), the damage-output coefficient (a2), the transfer coefficient of carbon dioxide (b12), the equilibrium temperature-sensitivity coefficient (t2xCO2), and the total availability of fossil fuels (fosslim). The following table (Table 1.1) shows the marginal distributions of eight variables.

Table 1.1: Marginal distributions of uncertain variables of DICE model

Variable	Definition of the Variable	Mean	St.Dev	Unit
ga0	Rate of Growth of Total Factor Productivity	0.079	0.004	per year
dsig	Rate of De-carbonization	-0.001	0.002	per year
t2xCO2	Equilibrium Temperature-Sensitivity Coefficient	2.900	1.110	Celcius per CO2 doubling
a2	Damage Parameter	0.003	0.001	Fraction of global output
pback	Price of backstop technology	344	138	$ per ton of carbon replaced
popasym	Asymptotic global population	10500	1892	millions
b12	Transfer coefficient in Carbon Cycle	0.088	0.017	per decade
fosslim	Total Resources of Fossil Fuels	6000	1200	billions of tons of carbon

However Nordhaus assumes that there is no uncertainty in the DICE model and fixes all the uncertain parameters and use these parameters at their mean values. We use the values given by Nordhaus in his books “A Question of Balance” (2008) and “DICE 2013R Manual” (2013) and assume that all the marginal distributions are from known families. All the variables in this analysis are normally distributed.

Our purpose is to integrate climatic and economic uncertainties into the DICE model to understand whether uncertainty has any significant effect on the performances of the policies. In the DICE model, there are eight different variables and each is discretized to three different values (see Figure-1.4).

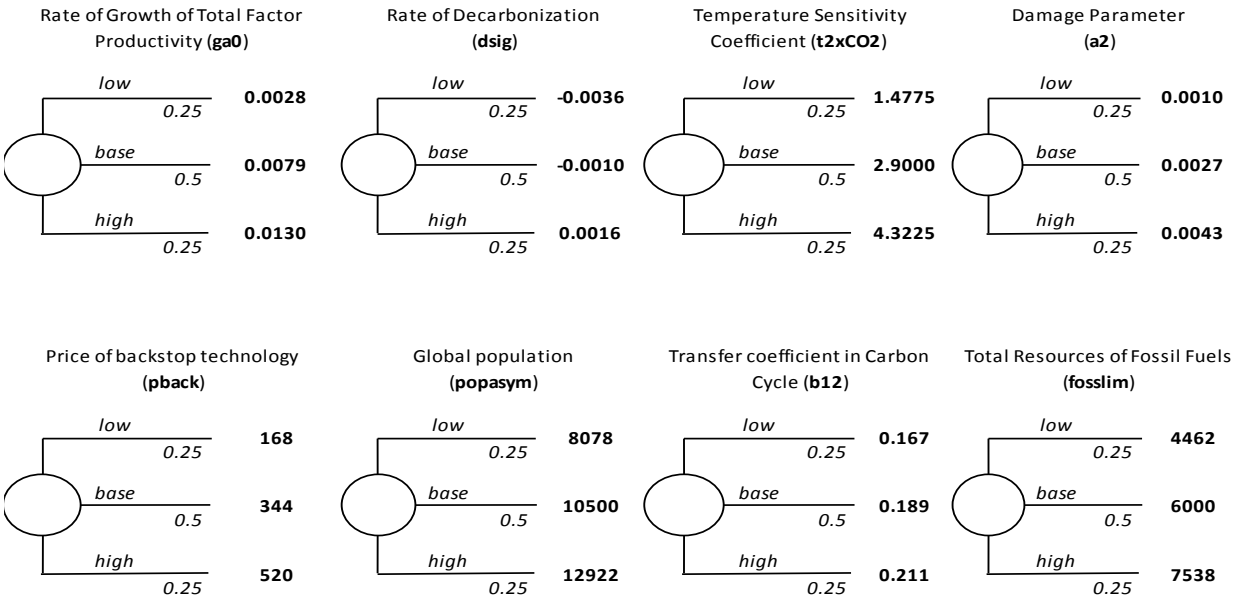


Figure 1.4: Decision Trees of Each Uncertain Variables

We then draw the decision tree with the uncertain parameters and the alternative climate change policies. Figure 1.5 shows the decision tree of DICE model with 8 uncertain parameters.

Our aim is to study the effect of uncertainty in global warming policy decisions and integrate climatic and economic uncertainties into the DICE model to find out what will be the cumulative impact of integrating uncertainty into the DICE model. We use the uncertain parameters in our analysis to generate joint probability distributions of eight uncertain parameters to evaluate and compare the expected performances of different policies.

1.6. Approximation with Partial Information

Modeling real-world decision situations such as those involving climate change often requires multivariate distributions, into which information about random events and their interactions is encoded. For instance, in the DICE model, there are eight different uncertain parameters and each is discretized to three different values, requiring construction of a $3\times3\times3\times3\times3\times3\times3\times3$ multi-

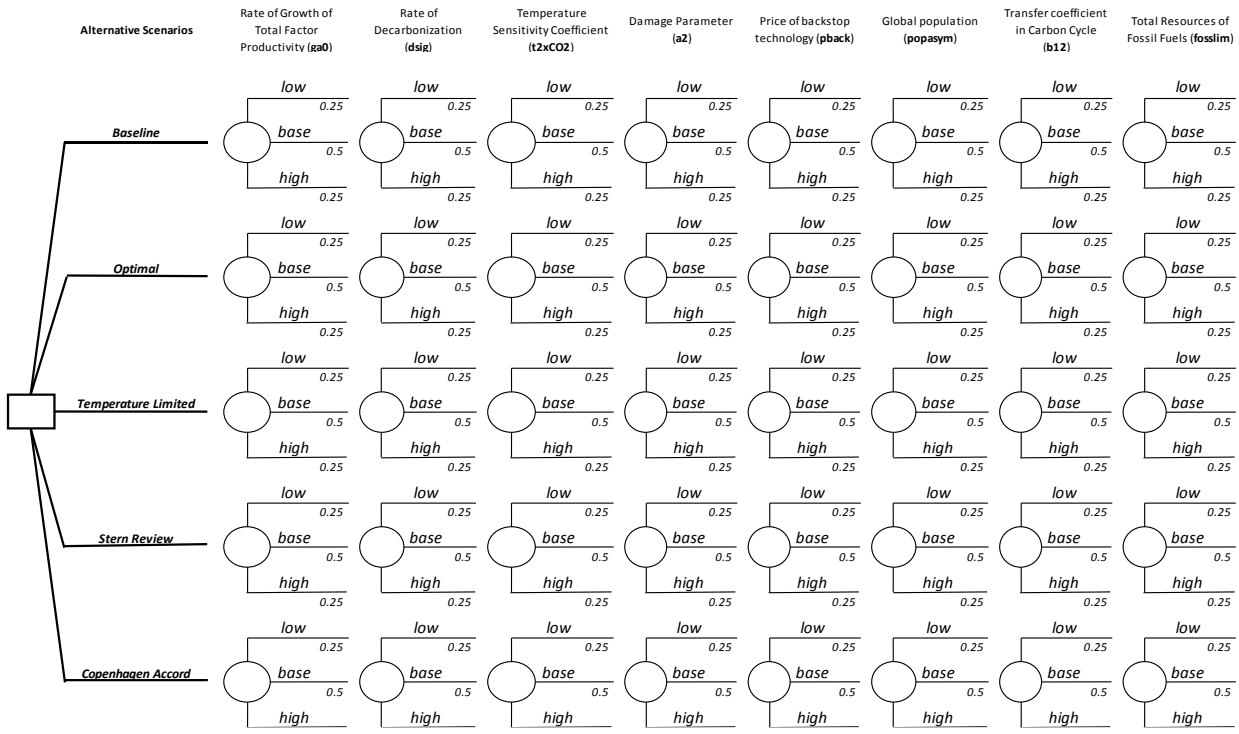


Figure 1.5: Decision Tree of DICE Model with 8 Uncertain Parameters

variate probability distribution. In order to construct multivariate distributions concerning the specific circumstances and preferences of a given decision situation, information needs to be elicited from the decision maker (DM). In real life decisions, however, the amount of information we can collect from the decision maker is limited because eliciting more information from the decision maker is difficult as the number of variables and lower order assessments increases (Howard, 1968; Ku and Kullback, 1969; Abbas, 2006), time consuming (Baker, 2009; Pearl, 1988; Keeney, 1973), expensive (Clemen and Reilly, 1999; Abbas, 2010), and because the decision maker could be unable or unwilling to make decisions (Matheson and Howard, 1969; Abbas, 2002, 2006, 2009, 2010).

1.6.1. Approximation with Independence Assumption

Many approximation methods have been discussed to approximate higher order utility and probability functions using lower order assessments in the decision analysis literature. In most of these studies, the probability distributions or utility functions are approximated by some simplifying assumptions such as probability independence (Keeney, 1973; Fishburn and Keeney, 1974; Brodley, 1982; Howard and Matheson, 1984; Pearl, 1988) or utility independence (Richard, 1975; Keeney, 1971, 1972, 1973, and 1974; Fishburn and Keeney, 1975).

A simple way to explain the concept of independence in probability theory is that the occurrence of one does not affect the probability of the other. Independence in probability theory is formulized as the product of the probabilities of the two individual events or product of their marginal distributions. In correspondence with multivariate probability distributions, the construction of multiattribute utility functions is also simplified as utility independence assumptions are satisfied. Independence concepts in utility and conditions under which preferences for some attributes are invariant with respect to others are discussed by Keeney and Raiffa (1976). If every attribute is utility-independent of its complement, then the functional form of the utility function reduces to a multilinear form. A stronger independence condition also holds if every subset of the attributes is utility-independent of its complement and the utility function has either the multiplicative or the additive form.

Utility or probability independence simplifies the construction of the multiattribute probability or utility function significantly. However, very strong conditions must hold in order to use independence forms. There are several situations in utility theory in which at least one attribute is not utility-independent of its complemen or, in other words, not all the attributes are utility dependent on their complement. Thus, the multilinear form propounded by Keeney and Raiffa

(1976) is not applicable. This situation is called as partial utility independence in the decision analysis literature.

Partial utility independence among attributes is used in the literature to approximate utility functions in different functional forms. Bell (1988) first characterized the functional forms of utility function based on the number of switches and called as m-switch utility functions. Abbas and Bell (2011) extended this idea to independent multiple attribute utility functions that lead to a maximum of one-switch change and discussed the independence conditions of one-switch for multiattribute utility functions (Abbas and Bell, 2012). Then Abbas and Chudziak (2013) applied one-switch utility functions in annuity payment decision to derive the functional forms of multiple attribute utility functions that lead to a maximum of one-switch change in preferences.

1.6.2. Approximation in the case of Dependence among Variables

In most decision problems, preferences of the decision maker may change with the different values of the variables or influences the likelihood of the other variables. Probability dependence refers to any situation in which random variables do not satisfy a mathematical condition of probabilistic independence Probability independence asserts that that the occurrence of one does not affect the probability of the other. Similarly, two random variables are independent if the realization of one does not affect the probability distribution of the other. This property is not always appropriate in most of the decision situations. Thus, in the case of dependency among variables, it is important to incorporate dependence relationships among variables into the decision situation when constructing probability distributions and utility functions of decision maker.

Several methods are proposed to construct joint probability distributions by incorporating dependence relationship among variables. Barron and Barrett (1996) compared three

approximation methods and evaluate these according to the quality of the decisions resulting from the approximated weights. Hazen (1986) explored the use of partial preference information in multiattribute decision making. Kirkwood and Sarin (1985) derived conditions to determine whether a pair of alternatives can be ranked, given the partial information about weighting constants, and presents an algorithm that partially rank-orders the complete set of alternatives based on the pairwise ranking information. Weber (1987) presented a ranking method that allows the DM to provide preference information in the form of pairwise comparisons of the alternatives. Clemen et. al. (2000), Reilly (2000), Lowell (1994), and Smith et.al (1992) performed several analyses to compare the accuracy of methods that assess probability dependence and showed that the optimal decision alternative may change when dependence between the variables is incorporated. Montiel and Bickel (2012) generated the set of all possible discrete distributions that expressed given information and approximated joint probability distributions on the Hit-and-Run Sampler algorithm. Clemen and Reilly (1999) used copulas to construct joint probability distributions based on lower-order assessments.

Several graphical methods are also proposed to construct joint probability distributions by incorporating dependence relationship among variables. Howard (1989) explored the use of graphs to capture probability dependence and called as knowledge maps. Bedford and Cooke (2001) showed multivariate probability distributions graphically for dependent random variables which they called vines and derived a general formula for the density of a vine dependent distribution. Boutilier et al. (2004) proposed a qualitative graphical representation of preferences that reflects conditional dependence and independence of preference statements under a *ceteris paribus* interpretation. Abbas et.al. (2010) presented a new method for constructing joint probability

distributions of continuous random variables using isoprobability contours without assessing directly the dependence, or association, between the variables.

Incorporating preference dependence between the attributes while constructing multiattribute utility functions are also discussed by several authors in the field of decision analysis. One approach introduced by Matheson and Howard (1968) that constructs a deterministic value function over the attributes and then assigns a utility function over the value function to represent the decision maker's preferences. This method is also discussed by Abbas and Howard by an example of a value function of a "peanut butter and jelly sandwich" (Abbas and Howard, 2005). Abbas (2010) also discussed a variety of methods for constructing multiattribute utility functions. Chajewska et.al. (2000) showed how density estimation techniques can be applied to approximate a density function from a database of partially elicited utility functions. Abbas (2009) introduced the multiattribute utility copula, which is a new functional forms that can be used to model preferences over utility dependent attributes. Abbas (2013) extended multiattribute utility copulas and proposes a method to construct utility copula functions using univariate utility assessments at the boundary values. Wang and Dyer (2012) estimated multivariate distributions through the use of a decision tree based on copulas.

1.6.3. Approximation by Using Elements of Information Theory

Several approaches to approximate joint probability distributions use elements of information theory. Dependence trees constitute one such well-known method. Chow and Liu (1968) first described how to construct a second-order product approximation of a joint probability distribution using first order dependence trees with a product of conditional, and pairwise distributions, where each child has only one parent. A more extended tree construction algorithm for dependence trees was outlined in Meila (1999). Ku and Kullback (1969) generalized Chow and Liu's algorithm,

allowing any lower-order marginal distributions to be used in the approximation. Keefer (2004) presented a model for approximating probability dependence among binary events. Sutcu and Abbas (2014) determined the best first order dependence tree approximation using the concept of cumulative residual entropy, which is an alternative measure of entropy that uses cumulative probability distributions.

Maximum entropy is also used widely to approximate joint probability distributions and multiattribute utility functions. The principle of maximum entropy was first expounded by E. T. Jaynes in two papers in 1957 where he emphasized a natural correspondence between statistical mechanics and information theory (Jaynes 1957a, 1957b). The principle of maximum entropy is often used to obtain prior probability distributions for Bayesian inference and used to approximate multiattribute utility functions. Chan (1971) discussed what probability distribution should employ if only range of a system parameter is known in discrete event simulation. Thomas (1979) described a generalized maximum entropy principle for dealing with decision problems involving uncertainty if only partial information is available. Smith (1993) computed the moments of joint distributions or value lotteries and then use these moments in maximum entropy formulation to compute approximate value lotteries or certain equivalents. Mackenzie (1994) used maximum entropy formulation to approximate multivariate distribution with given marginal and pairwise correlations. Abbas (2006) explored the use of the maximum entropy principle to approximate joint distributions using any number of lower order assessments. Abbas (2002) presented a graphical method to determine the maximum entropy distribution between upper and lower probability bounds and provide an interpretation for the shape of the maximum entropy distribution subject to fractile constraints. He also showed the formulation of maximum entropy problems given upper and lower bounds on moments and probabilities (Abbas, 2005). Dai et.al (2007)

studied and quantified the uncertainties in the software reliability modeling with correlated parameter(s) by combining the maximum entropy approach principle into the Bayesian approach.

Also, by using the analog between probability and utility, maximum entropy formulation is applied to utility. Abbas (2002, 2006), who showed how to apply maximum entropy to single attribute utility functions when only partial information is available can be credited with the seminal work in assigning utility values using the maximum entropy method. When a decision situation has more than one attribute, Abbas (2006) demonstrated one solution method that assigns a single attribute utility function over a value function. Abbas (2004) presented a maximum entropy method to find an optimal question-algorithm to elicit von Neumann and Morgenstern utility values and select the minimum number of questions needed for utility elicitation. In another paper, he provided a new method to measure of utility dependence, presented moments and cross-moments of utility functions, and derived the functional form of a utility function that satisfies some given moment assessment (Abbas, 2007). Hadfi and Ito (2012) extended the maximum entropy utility principle into an asymptotic maximum entropy utility principle for preference elicitation in a situation subject to a large predictive uncertainty with a small learning sample. Herfert and La Mura (2004) used the utility functions directly in entropy definition and approximated the utility functions of the decision maker by using the maximum entropy utility formulation.

1.7. Contributions of the Dissertation

In this dissertation, our aim is to help DMs to make better decision using both their preferences and the information available. Our current work addresses the problems where partial information about the decision situations is known. We approximate representative probability if only partial information is elicited from the decision maker by using cumulative distribution

functions instead of density/mass functions. Using joint cumulative probability distributions has many properties; (i) always non-negative, (ii) valid for both continuous and discrete cases, (iii) easy to implement because of using cumulative distribution functions instead of density functions.

Our first contribution is to approximate joint probability distributions of a set of discrete random variables using a product of second order conditional and marginal distributions based on cumulative residual entropy. We construct optimum first-order tree approximation of the joint distribution with respect to the Cumulative Residual Kullback Leibler divergence if its dependence tree has the maximum sum of cumulative residual mutual information pairs.

Our second contribution is to propose an approximation method similar to maximum entropy principle to construct representative joint probability distributions from its lower order assessments by using maximum cumulative residual entropy approach.

Finally, our third contribution is to take into consideration all the uncertain variables and their interactions together and characterize the uncertainty in climate change models to understand whether uncertainty significantly affects climate change decisions with regard to climate change policies. To do this, we integrate uncertainty into the well-known DICE model to understand whether uncertainty has any significant effect on the performances of the policies.

CHAPTER 2: REVIEW OF BASIC CONCEPTS, TERMINOLOGY, AND NOTATIONS OF INFORMATION THEORY

2.1. Introduction

This chapter reviews the fundamental concepts, terminology, and notations of information theory in probability that will be used in the remaining sections of the dissertation. We first define the marginal, joint, and conditional probability distributions and we also define them for survival functions. Then, we explain the entropy definitions of discrete random variables and continuous random variables in Section 2.3 and 2.4, respectively. Section 2.5 discusses the differences between traditional entropy and cumulative residual entropy. We explain the KL-divergence and mutual information in Section 2.6. Section 2.7 defines the maximum entropy approaches in the literature. Finally, we review the previous work on related research in Section 2.8.

2.2 Basic Concepts, Terminology, and Notations of Information Theory in Probability

This section presents the basic notation and definitions that will be used in the remaining sections of the dissertation. Let

$$F_x(x) = P(X \leq x) \tag{2.1}$$

be the marginal cumulative distribution function of the random variable X , and let

$$F(x, y) = P(X \leq x, Y \leq y) \tag{2.2}$$

be the bivariate cumulative distribution function of random variables X and Y .

Define a marginal survival function for variable X as

$$S_x(x) = 1 - F_x(x) = P(X > x) \tag{2.3}$$

Note that $S_x(x)$ is the probability that X is alive at the value x and a bivariate survival function for random variables X and Y as

$$S(x, y) = P(X > x, Y > y) = 1 - F_x(x) - F_y(y) + F(x, y) \tag{2.4}$$

The conditional survival function between two variables (X given Y) is

$$S_{x|y}(x \mid y) \triangleq \frac{S(x, y)}{S_y(y)} = S(X \geq x \mid Y \geq y) \tag{2.5}$$

where the variable X is alive at the time x given that the Y survived to just before time y .

2.3. Entropy of Discrete Random Variables

In this section, we first define the discrete form of Shannon's entropy and its interpretations; then we define the discrete form of cumulative residual entropy which use cumulative functions instead of probability mass functions.

2.3.1 Interpretation of the Discrete Shannon's Entropy

In information theory, entropy is a measure of average uncertainty associated with random variable. The concept was proposed by Claude E. Shannon in his 1948 paper "A Mathematical Theory of Communication"[ ]. Entropy of a discrete random variable is formulized as

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log\big(p(x_i)\big) \tag{2.6}$$

where $p(x)$ is the probability mass function of a discrete random variable. This measurement defines how much information needed to explain the outcome of a variable. In general, the base of

the logarithm is "base=2". The entropy will then be measured in bits. The entropy is a measure of the average uncertainty in the random variable. It is the number of bits on average required to describe the random variable. We will provide the following example of entropy expression to gain better understanding.

Example 2.1: Consider a discrete variable X with four possible outcomes and corresponding probabilities as shown below. What is the entropy of this random variable?

$$p(x)=\begin{cases} ½ & if\ x=0 \\ ¼ & if\ x=1 \\ ⅛ & if\ x=2 \\ ⅛ & if\ x=3 \end{cases} \tag{2.7}$$

Let us calculate the entropy of X and we use binary logarithm (base 2) in the entropy expression which is commonly used.

$$\begin{aligned} H(X) &= -\frac{1}{2}\log_2\left(\frac{1}{2}\right)-\frac{1}{4}\log_2\left(\frac{1}{4}\right)-\frac{1}{8}\log_2\left(\frac{1}{8}\right)-\frac{1}{8}\log_2\left(\frac{1}{8}\right) \\ &= \frac{7}{4} \end{aligned} \tag{2.8}$$

Suppose that we wish to determine the value of X with the minimum number of binary questions. The resulting expected number of binary questions required is 1.75. This turns out to be the minimum expected number of binary questions required to determine the value of X. Moreover, an interval is found to the minimum expected number of binary questions required to determine X as

$$H(X) \le Expected\ number\ of\ questions \le H(X)+1 \tag{2.9}$$

The entropy of random variable X is eventually the lower bound of expected number of questions to determine variable X. This formulation is very important to our decision analysis process, and simplifies it with binary questions.

2.3.2 Interpretation of Cumulative Residual Entropy

After several decades, Rao et.al.(2004) developed an alternative entropy measurement formulation by extending Shannon's entropy from density functions to cumulative distribution functions of random variables. Let X be a random vector in R^N, then cumulative residual entropy of X is

$$\mathcal{E}(X) = -\int_{R_+^N} P\left(|X| > \lambda\right) \log\left(P\left(|X| > \lambda\right)\right) d\lambda \tag{2.10}$$

where $X = \left(X_1, X_2, \cdots, X_N\right)$, $\lambda = (\lambda_1, \lambda_2, \cdots, \lambda_N)$ and $|X| > \lambda$ means $|X_i| > \lambda_i$ and $R_+^N = \left(x_i \in R^N; x_i \geq 0\right)$.

Shannon's entropy $H(X)$ of X is computed solely using the probabilities $P(X = t)$ and interprets the entropy as a measure of the "uncertainty" in X. If, X denotes the life span of a machine, the appropriate probabilities to consider are $P(X > t)$ in which the life span exceeds a given time t and not $P(X = t)$ in which the life span equals t. Cumulative Residual Entropy possesses more general mathematical properties than the Shannon entropy. This measure is always non-negative and its definition is valid for both continuous and discrete cases. Let's look at the following example to understand better how cumulative residual entropy measures the uncertainty.

Example 2.2: Consider the same discrete variable X with four possible outcomes and corresponding probabilities in Example 1. In this example, however, we will calculate the entropy with new entropy expression. First, we define the cumulative distribution function of the discrete

probability distribution of X and the complementary distribution function (survival function) of X:

$$P(x)=\begin{cases}1/2 & if\ x=0\\ 3/4 & if\ x=1\\ 7/8 & if\ x=2\\ 1 & if\ x=3\end{cases} \qquad and \qquad \bar{P}(x)=\begin{cases}1/2 & if\ x=0\\ 1/4 & if\ x=1\\ 1/8 & if\ x=2\\ 0 & if\ x=3\end{cases} \tag{2.11}$$

Thus, the entropy is calculated by using the complementary cumulative distribution functions as

$$\begin{aligned} H(X) &= -\frac{1}{2}\log_2\left(\frac{1}{2}\right)-\frac{1}{4}\log_2\left(\frac{1}{4}\right)-\frac{1}{8}\log_2\left(\frac{1}{8}\right)-0\log_2(0) \\ &= \frac{11}{8} \end{aligned} \tag{2.12}$$

By using the formulation at (4), an interval is found to

$$1.375 \le expected\ number\ of\ questions \le 2.375 \tag{2.13}$$

Shannon entropy and Rao entropy are consistent in the discrete cases, in that both measurement expressions are always non-negative and its definition is valid for both discrete cases. Unfortunately, extending Shannon's discrete entropy to a continuous (differential) entropy case poses some challenges. We will introduce and compare both measurement expressions for continuous cases in the following section.

2.4 Entropy of Continuous Random Variables

In this section, we first define the continuous form of the traditional entropy measure (also called differential entropy in the literature) and then define the continuous form of the cumulative residual entropy.

2.4.1 Interpretation of the Differential Entropy

As in the discrete case, it is not difficult to handle the analogous definitions and results of the continuous case. Entropy of a continuous random variable X with cumulative function $F(X) = P(X \le x)$ and density function $F'(x) = f(x)$ is formulized as

$$H(X) = -\int f(x)\log\big(f(x)\big)dx \tag{2.14}$$

This measurement also defines how much information needed to explain the outcome of a variable; however, differential entropy does not share all properties of discrete entropy. We will provide the following example of differential entropy expression to show the difference.

Example 2.3: Consider a random variable distributed uniformly from 0 to a, and the density of uniform distribution is

$$f(x) = \begin{cases} 1/a & 0 \le x < a \\ 0 & \text{otherwise} \end{cases} \tag{2.15}$$

Then its differential entropy is calculated as

$$h(X) = -\int_0^a \frac{1}{a}\log\left(\frac{1}{a}\right)dx = \log a \tag{2.16}$$

Unlike discrete entropy, differential entropy can be negative. If $a < 1$, then $h(X) < 0$, if $a = 1$, then $h(X) = 0$, and entropy takes positive values if a is bigger than 1. The differential entropy is inconsistent in the sense that differential entropy may take any value on the real line. The possibility of negative entropy shows that expected number of questions cannot be interpreted and

found in the continuous case. Moreover, estimating empirical distributions by using maximum entropy is impossible in continuous cases due to the possibility of negative entropy.

2.4.2. Interpretation of the Cumulative Residual Entropy

As we discussed in discrete case, Rao et.al. (2004) develop an alternative entropy measurement formulation by extending differential entropy from density functions to cumulative distribution functions of random variables. Let X be a random vector in R^N, then cumulative residual entropy of X is

$$\mathcal{E}(X) = -\int_{R_+^N} P\left(|X| > \lambda\right) \log\left(P\left(|X| > \lambda\right)\right) d\lambda \tag{2.17}$$

Cumulative residual entropy is always non-negative and its definition is valid for both continuous and discrete cases. Let's look at the following example to understand better how cumulative residual entropy measures the uncertainty.

Example 2.4: Consider the same continuous uniform variable X which is distributed from 0 to a in Example-3. Let's calculate the entropy with new entropy expression. First we define the cumulative distribution function of the uniform distribution of X and complementary cumulative distribution function (tail distribution) of X.

$$F(x) = \begin{cases} \frac{x}{a} & 0 \le x < a \\ 0 & \text{otherwise} \end{cases} \qquad \textit{and} \qquad \bar{F}(x) = \begin{cases} \frac{(a-x)}{a} & 0 \le x < a \\ 0 & \text{otherwise} \end{cases} \tag{2.18}$$

So, the entropy is calculated by using the complementary cumulative distribution functions as

$$\begin{aligned}\mathcal{E}(X) &= -\int_0^a P(|X| > x)\log\big(P(|X| > x)\big)\,dx \\ &= -\int_0^a \left(\frac{a-x}{a}\right)\log\left(\frac{a-x}{a}\right)dx \\ &= \frac{1}{4}a\end{aligned} \tag{2.19}$$

As we see from the previous example, cumulative residual entropy is nonnegative unlike the Shannon's entropy. Differential entropy is inconsistent in the sense that the entropy of uniform distribution in an interval of length a is $\log(a)$, which is zero if $a = 1$ and negative if $a < 1$.

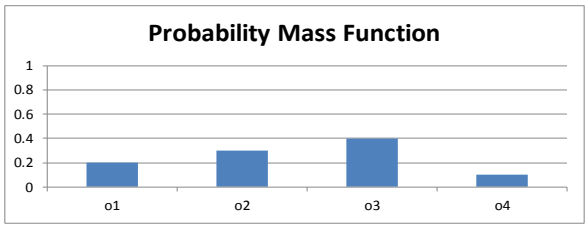
2.5. The differences between Shannon Entropy and Cumulative Residual Entropy

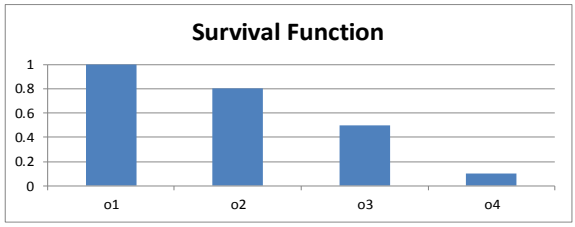
The traditional entropy definition uses probability mass functions or density functions in entropy formulation. On the other hand, cumulative residual entropy uses cumulative functions in both discrete and continuous cases. For any discrete random variable X, Shannon's entropy $H(X)$ of X is computed solely using the probabilities $P(X = t)$ and interprets the entropy as a measure of the "uncertainty" in X. If, X denotes the life span of a machine, the price of a stock, the number of properly functioning components required in a complex system, the appropriate probabilities to consider are $P(X > t)$ which represents life span exceeds a given time t and not $P(X = t)$ which is life span equals t.

For instance, the discrete variable X has a probability mass function

$$p(x) = \begin{cases} 0.2 & if\ \ outcome - 1 \\ 0.3 & if\ \ outcome - 2 \\ 0.4 & if\ \ outcome - 3 \\ 0.1 & if\ \ outcome - 4 \end{cases}$$

So, the entropy of random variable X is calculated by Shannon entropy and CRE entropy as





$$H(X) = 0.2\log 0.2 + 0.3\log 0.3 + 0.4\log 0.4 + 0.1\log 0.1$$

$$\mathcal{E}(X) = 1\log 1 + 0.8\log 0.8 + 0.5\log 0.5 + 0.1\log 0.1$$

Figure 2.1: The Entropy of Random Four Outcome Variable *X* Calculated by Shannon Entropy and CRE Entropy

CRE entropy is calculated by considering the uncertainty that exceeds a given time, whereas Shannon entropy calculates the entropy which life span equals a certain time.

Cumulative residual entropy is more advantageous for the following reasons:

- It is always non-negative in both discrete and continuous cases
- It provides consistent definitions in both the continuous and discrete domains
- It uses cumulative functions which are more regular than the density functions, because the density is computed as the derivative of the cumulative functions; and
- It can be easily computed from the sample data but eliciting density functions is a difficult task.

On the other hand, Cumulative Residual Entropy can only be applied to numeric variables. It cannot be applied to decision situation where the variables are non-numeric. For instance, you are planning to go outside but there is 40% chance of rain after 2pm and are not sure whether to carry an umbrella. Because this decision situation includes a non-numeric variable, cumulative residual entropy is inapplicable.

In general, we can say that Shannon's entropy is defined for distributions with densities. The entropy of a discrete distribution is always positive, while the differential entropy of a continuous variable may take any value on the extended real line. It is "inconsistent" in the sense that the differential entropy of a uniform distribution in an interval of length a is $\log(a)$, which is zero if $a=1$, negative if $a<1$, and positive if $a>1$.

Cumulative Residual Entropy overcomes the problems mentioned above, retaining many of the important properties of Shannon entropy while preserving the well-established principle that the logarithm of the probability of an event should represent the information content in the event. For a variable X, maximum entropy and maximum cumulative residual entropy with given constraints are defined as shown in Table 2.1.

Table 2.1: Maximum Entropy and Maximum Cumulative Residual Entropy Formulations

Maximum Entropy	**Maximum CRE**
$p^*(x) = \arg\max - \int p(x_i)\log(p(x_i))$ *s.t.* $\int h_i(x)\,p(x_i) = \mu_i$ $\int p(x_i) = 1$ $p(x_i) \geq 0$	$S^*(x) = \arg\max - \int S(x)\log(S(x))$ *s.t.* $\int r_i(x)S(x) = \alpha_i$ $\int S(x) \leq E(X)$
Solution: $p^*(x) = e^{-\alpha_0 - \alpha_1 h_1(x) - \alpha_2 h_2(x) - \ldots - \alpha_n h_n(x) - 1}$	***Solution:*** $S^*(x) = \mathrm{e}^{-1-\lambda_1 r_1(x) - \lambda_2 r_2(x) - \cdots \lambda_n r_n(x)}$

Shannon entropy can be negative or positive depending on the value of the variable but Rao entropy is always non-negative. Also, the idea in Rao's entropy definition is smooth and consistent. The distribution function is more regular because it is defined in an integral form unlike the density function, which is defined as the derivative of the distribution. The definition also preserves the well-established principle that the logarithm of the probability of an event should represent the information in the event.

The main advantages of our proposed methods are that the entropy measure is always non-negative and consistent in both discrete and continuous domains, and use cumulative functions instead of density functions. The limitation and drawback of proposed methods is that it cannot be applied to decision situation where the variables are non-numeric.

2.6. Relative Entropy and Mutual Information

In this section, we define the Kullback-Leibler divergence or relative entropy, and then the mutual information which is the special case of the relative entropy. We also define the cumulative residual entropy based Kullback-Leibler divergence measure and mutual information.

2.6.1 Relative Entropy or Kullback-Leibler Divergence

After Shannon's entropy formulation, Kullback and Leibler (1951) extended the entropy definition and introduced a new measure which is a non-symmetric measure of the difference or distance between two probability distributions. For discrete probability distributions P and Q, the KL-divergence of Q from P is defined to be

$$D_{KL}\left(P\|Q\right)=\sum_{i=1}^{n}p(x_i)\log\left(\frac{p(x_i)}{q(x_i)}\right) \tag{2.20}$$

Relative entropy is a measure of the information lost when approximate distribution Q, is used to approximate true distribution, P.The Kullback-Leibler distance is also known as the cross entropy, or relative entropy. The measure is always non-negative and zero if and only if the two probability distributions are identical.

Following on the entropy definition provided by Rao et al., Baratpour and Rad (2012) defined a measure of the difference or distance between two survival functions and called it cumulative Kullback-Leibler divergence. The Cumulative Residual KL-Divergence measure is defined as

$$CKL(S_F : S_G) = \int_0^\infty S_F(x) \ln \frac{S_F(x)}{S_G(x)} dx - [E(F) - E(G)] \tag{2.21}$$

where S_F and S_G are the survival functions of variables F and G respectively. Also, $E(F)$ and $E(G)$ are the expected values of variables F and G respectively.

2.6.2. Mutual Information

In probability, mutual information is a special case of a more general quantity called relative entropy, which is a measure of the distance between two probability distributions. Mutual information between two variables is the Kullback-Leibler distance between their joint distribution and product of their marginal when they are mutually independent (Cover and Thomas, 1991). Mutual information is a quantity that measures the mutual dependence of the two random variables. The interpretation for the mutual information is how much tells us about the random variable, X, when we know the outcome of another random variable, Y. Mutual information quantity is always non-negative and zero if and only if two random variables are mutually independent.

For two discrete variables X and Y whose joint probability distribution is $p(x, y)$, the mutual information between them, denoted $I(X;Y)$, is given by

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \ln \left(\frac{p(x,y)}{p_x(x) p_y(y)} \right) \tag{2.22}$$

where $p_x(x)$ and $p_y(y)$ are the marginal distributions of variables X and Y respectively.

In the case of continuous random variables, the double summation is replaced with double integration, and the mutual information is given by

$$I(X;Y) = \iint_{y \in Y, x \in X} p(x,y) \ln \left(\frac{p(x,y)}{p_x(x) p_y(y)} \right) dxdy \tag{2.23}$$

where $p(x,y)$ is joint probability density distribution, $p_x(x)$ and $p_y(y)$ are the marginal probability density distribution functions of X and Y respectively. It is symmetric, always non-negative and equal to zero if and only if variables are mutually independent. The relationship between entropy, conditional entropy and mutual information is expressed in Figure-2.2.

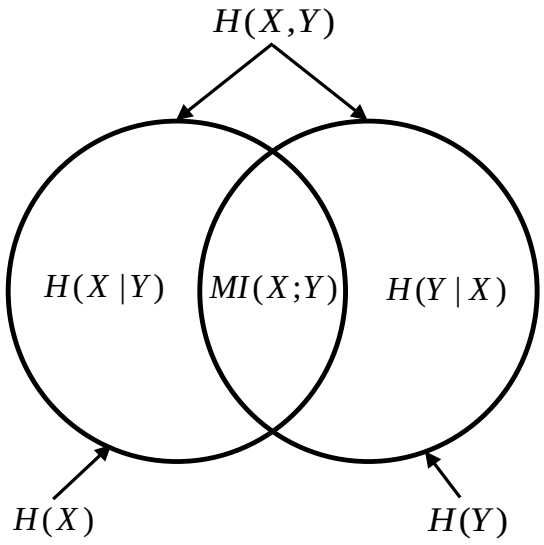


Figure 2.2: The Relationship between Entropy, Conditional Entropy and Mutual Information

After Kullback-Leibler's divergence measure, Wang et.al. (2003) defined a quantity similar to mutual information using cumulative residual entropy measure which is called Cross Cumulative Residual Entropy (CCRE) as

$$CCRE(X:Y) = \mathcal{E}(X) - E[\mathcal{E}(Y \mid X)] \tag{2.24}$$

The cross cumulative residual entropy is not symmetric as the traditional mutual information formulation. However, cross cumulative residual entropy can easily be symmetrized; the symmetric cross cumulative residual entropy (SCCRE) is given by:

$$CCRE_{symmetric}(X:Y) = \frac{1}{2}\left(CCRE(X,Y) + CCRE(Y,X)\right) \tag{2.25}$$

Cross cumulative residual mutual information quantity is also always non-negative and zero if and only if two random variables are mutually independent.

2.7. Maximum Entropy Formulations

In this section, we define the maximum entropy formulations in information theory literature. We first explain the traditional maximum entropy formulation and then the maximum cumulative residual entropy for single variable cases.

2.7.1 Interpretation of the Jaynes Maximum Entropy

Laplace might be considered the father of maximum entropy, having proposed the underlying theme 200 years ago in his "Principle of Insufficient Reason": when one has no information to distinguish between the probabilities of two events, the best strategy is to consider them equally likely (Laplace, 1774). Jaynes, a more recent pioneer of maximum entropy, extends the idea to maximize entropy subject to certain constraints representing the incomplete information (Jaynes,

1957a, 1957b). The idea of Max-Ent is to estimate a target probability distribution by finding the probability distribution of maximum entropy subject to a set of constraints that represent the incomplete information. The Maximum Entropy principle states that out of all distributions, consistent with a given set of constraints, choose one that maximizes entropy. Intuitively, the principle of maximum entropy is simple: models all which is known and assume nothing about that which is unknown. The Maximum Entropy principle states that out of all distributions consistent with a given set of constraints choose one that maximizes entropy. The maximum entropy probability mass function for a discrete variable, X , having n outcomes, when no further information is available is

$$\begin{aligned} & p(x)_{maxent} = \arg\max - \sum_{i=1}^{n} p(x_i) \log(p(x_i)) \\ & s.t. \\ & \sum_{i=1}^{n} p(x_i) = 1 \\ & p(x_i) \geq 0, \quad where\ i = 1, 2, \cdots, n \end{aligned} \tag{2.26}$$

This formulation yields a probability mass function with equal probability for each outcome,

$$p(x_i) = \frac{1}{n} \quad i = 1, 2, \cdots, n \tag{2.27}$$

This probability mass function has an entropy of $\log(n)$, which is the maximum value that can be measured when equal probabilities assigned for each outcome. Consider additional information is elicited from decision maker. In that partial information case, additional constraints, and indicator function or moment constraints are taken into account in maximum entropy formulation. Then, the convex optimization problem turns to a problem as

$$f(x)_{maxent} = \arg\max -\int_a^b f(x)\log(f(x))dx$$

$$s.t.$$

$$\int_a^b h_i(x) f(x)dx = \mu_i \qquad i = 1,2,\cdots,n \tag{2.28}$$

$$\int_a^b f(x)dx = 1$$

$$f(x) \geq 0$$

where $[a,b]$ is the interval of the variable, $h_i(x)$ is an indicator function or a moment constraint, and μ_i is given constant. Using the method of Lagrange multipliers, one can obtain the maximum entropy density estimation. Taking the partial derivative with respect to $f(x)$, equating it to zero, and rearranging the equations gives

$$f^*(x) = e^{-\alpha_0 - \alpha_1 h_1(x) - \alpha_2 h_2(x) - \ldots - \alpha_n h_n(x) - 1} \tag{2.29}$$

Example 2.5: If only the first moment is available, then the maximum entropy solution is $f^*(x) = e^{-\alpha_0 - \alpha_1 x - 1}$, which is the density of exponential distribution. If the first and second moments are available, then the maximum entropy solution is $f^*(x) = e^{-\alpha_0 - \alpha_1 x - \alpha_2 x^2 - 1}$, nothing but the Gaussian density distribution, where α_i is the Lagrange multiplier coefficient for each constraint.

2.7.2 Interpretation of Maximum Cumulative Residual Entropy

The maximum cumulative residual entropy is also calculated similar to Maximum Entropy. Let *X* be a non-negative random variable, and $r_1,\cdots,r_n$ are indicator function or moment constraints. Put

$$R_i(t) = \int_0^t r_i(s)ds \tag{2.30}$$

Suppose $E(R_i(X)) = \alpha_i$, $1 \leq i \leq n$ is given. In terms of survival function, $S(x)$ we can write

$$\int_0^{\infty} r_i(s)S(x)ds = \alpha_i \tag{2.31}$$

and define maximum cumulative residual entropy as

$$\begin{aligned} &\mathcal{E}^* = \arg\max - \int S(x)\log(S(x))dx \\ &s.t. \\ &\quad \int r_i(x)S(x)dx = \alpha_i \qquad i = 1,2,\cdots,n \\ &\quad x, F(x) \geq 0 \end{aligned} \tag{2.32}$$

and in all cases, let us write the maximum cumulative residual entropy distribution as

$$\mathcal{E}^*(x) = \exp\left(\sum_{i=1}^{n} \lambda_i r_i(x)\right) \tag{2.33}$$

Using cumulative probability distribution in maximum entropy formulation finally results in complementary cumulative distribution form. The answer is not exactly identical to Shannon's maximum entropy because the density of a function is found in Shannon's entropy case. The maximum cumulative residual entropy formulation calculates a Weibull distribution instead of an exponential distribution.

Example 2.6: Consider $r_1(x) = 1$ and $r_2(x) = t$. Then

$$\int_0^{\infty} r_1(x)S(x)dx = \int_0^{\infty} P(X > x)dx = E[X] \tag{2.34}$$

$$\int_0^{\infty} F(x)r_2(x)dx = \int_0^{\infty} P(X > x)xdx = \frac{1}{2}E[X^2] \tag{2.35}$$

With a positive support constraint and under first and second constraints, the maximum cumulative residual entropy is

$$\mathcal{E}^* = e^{-\lambda_1 - 2\lambda_2 x} \tag{2.36}$$

So, the maximum CRE is the cumulative exponential distribution with mean $\lambda = \frac{E[X^2]}{2E[X]}$. Here the maximum cumulative residual entropy formulation finds an exponential distribution which is a special case of Weibull distribution. So, both the traditional maximum entropy approach and the maximum cumulative residual entropy approach find an exponential distribution and almost approximate similar distributions,

2.8. Review of Previous Work on Related Research

We now review the previous work in this section. We first discuss Chow-Liu's first order dependence tree and then the Abbas' maximum entropy approximation method.

2.8.1. Chow-Liu's First Order Dependence Trees (Chow and Liu, 1968)

In probability theory and statistics, the Chow–Liu tree (Chow & Liu, 1968) is an efficient method for constructing a second-order product approximation of a joint probability distribution. Chow and Liu show that a probability distribution of first order dependence tree structure is the best approximation to the true distribution if its dependence tree has the maximum sum of mutual information pairs from all such first order dependence trees. Chow and Liu provide a simple algorithm for constructing the optimal tree and determine which conditional probabilities are to be used in the product approximation. The method is based on evaluating the mutual information pairs of variables at each stage. Then, the algorithm simply adds the maximum mutual information pairs to the tree. In first order dependence trees, each variable is conditioned on at most one variable, and there cannot be a cycle between the variables. A four-variate joint probability distribution $P(X_1, X_2, X_3, X_4)$ can be approximated as in Figure 2.3.

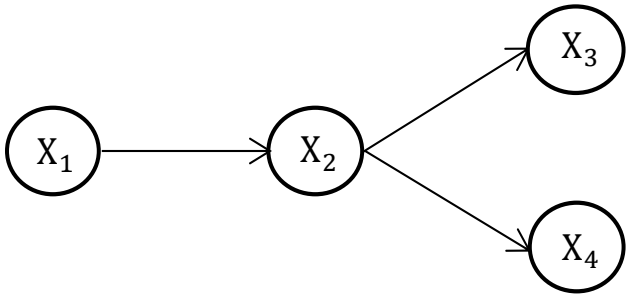


Figure 2.3: Example of a four-dimensional dependence tree

$$P^t(X_1, X_2, X_3, X_4) = P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_2)P(X_4 \mid X_2) \tag{2.37}$$

Chow and Liu show that a probability distribution of the first order dependence tree structure is the best approximation to the true distribution if its dependence tree has the maximum sum of mutual information pairs from all such first-order dependence trees.

2.8.2. Maximum Entropy with Lower Order Assessments (Abbas, 2006)

The maximum entropy principle was first defined by Jaynes (1957). The principle of maximum entropy states that the probability distribution which best represents the current state of knowledge is the one with largest entropy. Abbas (2006) considered an extension to the maximum entropy formulation. He used the discrete form of maximum entropy to approximate joint probability distributions given lower order assessments. For example, the maximum entropy formulation of a joint distribution of four variables given the knowledge of the pairwise joint assessments is

$$\begin{aligned} & p^*_{i,j,k,l} = \arg\max - \sum_{i,j,k,l} p_{i,j,k,l} \log\left(p_{i,j,k,l}\right) \\ & s.t. \\ & \sum_{k,l} p_{i,j,k,l} = p_{ij..} \quad \sum_{j,l} p_{i,j,k,l} = p_{i.k.} \quad \sum_{j,k} p_{i,j,k,l} = p_{i..l} \\ & \sum_{i,l} p_{i,j,k,l} = p_{.jk.} \quad \sum_{i,k} p_{i,j,k,l} = p_{.j.l} \quad \sum_{i,j} p_{i,j,k,l} = p_{..kl} \\ & \sum_{i,j,k,l} p_{i,j,k,l} = 1, \quad p_{i,j,k,l} \geq 0 \quad \forall i, j, k, l \end{aligned} \tag{2.38}$$

where the subscripts refer to the variables in the order of assessment provided. For example, p_{ijkl} refers to the joint probability of the i^{th} outcome (branch of the tree) of the first variable, the j^{th} outcome of the second variable, the k^{th} outcome of the third variable, and the l^{th} outcome of the fourth variable. A dot " " means that variable has been summed over. For example, $p_{23..}$ refers to the pairwise joint probability of the second branch of the first variable and the third branch of the second variable. This formulation produces a joint distribution

$$p^*_{i,j,k,l} = e^{-1-\lambda_0-\lambda_{ij..}-\lambda_{i.k.}-\lambda_{i..l}-\lambda_{.jk.}-\lambda_{.j.l}-\lambda_{..kl}} \tag{2.39}$$

where λ's are the Lagrange multipliers for the corresponding constraints and the same subscript notation as the probability notation is used like $\lambda_{ij..}$ corresponds to the constraint which has the pairwise assessment in equation 2.38, and λ_0 is a normalizing constant.

CHAPTER 3

FIRST-ORDER DEPENDENCE TREES WITH CUMULATIVE RESIDUAL ENTROPY

3.1. Introduction

In this chapter we determine the best first-order dependence tree approximation using the concept of cumulative residual entropy (CRE), an alternative measure of entropy that uses cumulative probability distributions. In contrast to discrete entropy where probabilities can be assigned to non-numeric variables, CRE requires numeric variables for the construction of a cumulative distribution.

In this chapter, we first formulate the concepts of Kullback-Leibler (KL)-divergence and mutual information in terms of cumulative residual entropy. Then we derive the optimal first-order dependence tree approximation of the joint distribution in terms of the cumulative residual KL-divergence. We show that the optimal tree approximation is the one with the highest sum of cumulative residual mutual information pairs. This result parallels the Chow-Liu dependence tree formulation based on Shannon's entropy. We then use a Monte Carlo simulation to show that our method is an alternative approximation method to Chow-Liu's first-order dependence tree method.

The remainder of this chapter is structured as follows: Section 3.2 presents cumulative residual entropy to KL-divergence and mutual information definitions. Section 3.3 discusses the optimal cumulative residual entropy-based dependence tree. Section 3.4 presents a Monte Carlo simulation to quantify and compare the accuracy of our CRE approximation with the Chow-Liu approximation.

3.2. Cumulative Residual KL-Divergence and Mutual Information

We apply cumulative residual entropy to KL-divergence and mutual information. Our definitions differ from Baratpour and Rad's (2012) cumulative Kullback-Leibler divergence, and Wang et.al's (2003) cross cumulative residual entropy definitions in that we simplify the cumulative KL divergence by removing the expected values of random variables and taking the absolute value of the expression.

Definition 3.1: *Cumulative Residual Kullback-Leibler Divergence*

The cumulative residual KL divergence between probability distributions S_T and S_A is

$$KL_{CRE}(S_T \parallel S_A) = \left| \sum S_T(x) \log \frac{S_T(x)}{S_A(x)} \right| \tag{3.1}$$

where $KL_{CRE}(S_T \parallel S_A) \geq 0$ and equality holds if and only if $S_T = S_A$.

We also define another quantity similar to mutual information and called it as Cumulative Residual Mutual Information (MI_{CRE}).

Definition 3.2: *Cumulative Residual Mutual Information*

The cumulative residual mutual information $MI_{CRE}(X,Y)$, between variables X and Y is

$$MI_{CRE} = \left| \sum_{x \in X} \sum_{y \in Y} S(x,y) \left[\log \left(\frac{S(x,y)}{S_x(x) S_y(y)} \right) \right] \right| \tag{3.2}$$

The cross-cumulative residual entropy is not symmetric and is defined based on conditional cumulative residual entropy. Our definition, however, is symmetric and expressed as a cumulative residual Kullback-Leibler divergence of the product of the marginal survival functions of two random variables from the joint survival function of random variables.

3.3. First-order Dependence Trees using Cumulative Residual Entropy

In first-order dependence trees, each variable is conditioned on at most one variable, and there cannot be a cycle between the variables. Figure-3.1 shows an example of a first-order dependence tree of four variables.

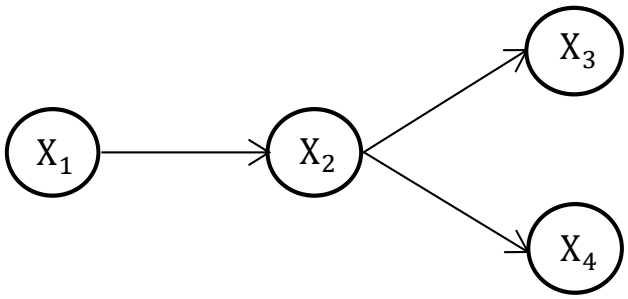


Figure 3.1: Example of a four-dimensional dependence tree

A four-variate joint probability distribution $P(X_1, X_2, X_3, X_4)$ can be approximated as in Figure-3.1 using a first-order dependence tree as

$$P^t(X_1, X_2, X_3, X_4) = P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_2)P(X_4 \mid X_2) \tag{3.3}$$

We define the optimum first-order dependence tree formulation with respect to the Cumulative Residual KL-divergence measure.

Theorem 1:

The first-order dependence tree approximation is an optimum first-order tree approximation of the joint distribution with respect to the Cumulative Residual KL-divergence if its dependence tree has the maximum sum of cumulative residual mutual information pairs.

Proof of Theorem 1:

In this proof, we follow the proof of Chow-Liu's first-order dependence tree theorem but apply it to the survival functions and the two proposed measures defined in Section 3.2: cumulative

residual KL-divergence (KL_{CRE}) and cumulative residual mutual information (MI_{CRE}). Let S_A be a second order product approximation (first-order dependence tree). The optimal first-order dependence tree is determined by minimizing the cumulative residual KL-divergence between true distribution S_T and approximate distribution S_A as $S_{A^*} = \arg\min KL_{CRE}[S_T \,||\, S_A]$. We first have the equation

$$\begin{aligned} KL_{CRE}(S_T \,||\, S_A) &= -\left| \sum S_T(x) \log \frac{S_T(x)}{S_A(x)} \right| \\ &= -\left| \sum S_T(x) \log S_T(x) + \sum S_T(x) \log \prod_{i=1}^{n} S_A(x_i \mid x_{j(i)}) \right| \end{aligned} \tag{3.4}$$

The first term of the right hand side of equation (3.4) is cumulative residual entropy of true distribution (S_T), $\mathcal{E}(S_T) = -\sum S_T(x) \log S_T(x)$. So, re-arranging equation (3.4) gives

$$KL_{CRE}(S_T \,||\, S_A) = -\left| -\mathcal{E}(S_T) + \sum S_T(x) \log \prod_{i=1}^{n} S_A(x_i \mid x_{j(i)}) \right| \tag{3.5}$$

We can write conditional survival function $S_A(x_i \mid x_{j(i)})$ as $\frac{S_A(x_i, x_{j(i)})}{S_A(x_{j(i)})}$, then equation (3.5) can be written as

$$KL_{CRE}(S_T \,||\, S_A) = -\left| -\mathcal{E}(S_T) + \sum_{i=1}^{n} \sum_{x_i, x_{j(i)}} S_T(x_i, x_{j(i)}) \log \frac{S_A(x_i, x_{j(i)})}{S_A(x_{j(i)})} \right| \tag{3.6}$$

Multiplying the numerator and the denominator of last term of equation (3.6) by marginal survival function, $S_A(x_i)$

$$KL_{CRE}(S_T \,||\, S_A) = -\left| -\mathcal{E}(S_T) + \sum_{i=1}^{n} \sum_{x_i, x_{j(i)}} S_T(x_i, x_{j(i)}) \log \frac{S_A(x_i, x_{j(i)})}{S_A(x_{j(i)})} \times \frac{S_A(x_i)}{S_A(x_i)} \right| \tag{3.7}$$

Our aim to multiply by is to re-arrange the equation (3.6) and obtain cumulative residual mutual information. By using the logarithm of a product is the sum of the logarithms of the factors rule, we can rewrite equation (3.7) as

$$KL_{CRE}(S_T \,||\, S_A) = \\ -\left|-\mathcal{E}(S_T) + \sum_{i=1}^{n}\sum_{x_i} S_T(x_i, x_{j(i)}) \log S_A(x_i) + \sum_{i=1}^{n}\sum_{x_i, x_{j(i)}} S_T(x_i, x_{j(i)}) \log \frac{S_A(x_i, x_{j(i)})}{S_A(x_i) S_A(x_{j(i)})}\right| \tag{3.8}$$

In order to minimize the cumulative residual KL information, we expect that the true and approximate distribution satisfy the equality condition that achieves the maximal value with $S_A(x_i, x_{j(i)}) = S_T(x_i, x_{j(i)})$. We rewrite the equation (3.8) by substituting $S_A(x_i, x_{j(i)})$ with $S_T(x_i, x_{j(i)})$, and have

$$KL_{CRE}(S_T \,||\, S_A) = \\ -\left|-\mathcal{E}(S_T) + \sum_{i=1}^{n}\sum_{x_i} S_T(x_i) \log S_T(x_i) + \sum_{i=1}^{n}\sum_{x_i, x_{j(i)}} S_T(x_i, x_{j(i)}) \log \frac{S_T(x_i, x_{j(i)})}{S_T(x_i) S_T(x_{j(i)})}\right| \tag{3.9}$$

Using the rule of subadditivity rule of absolute values, we can rewrite the equation (3.9) as

$$KL_{CRE}(S_T \,||\, S_A) \le -\left|-\mathcal{E}(S_T)\right| + \left|\sum_{i=1}^{n}\sum_{x_i} S_T(x_i) \log S_T(x_i)\right| \\ + \left|\sum_{i=1}^{n}\sum_{x_i, x_{j(i)}} S_T(x_i, x_{j(i)}) \log \frac{S_T(x_i, x_{j(i)})}{S_T(x_i) S_T(x_{j(i)})}\right| \tag{3.10}$$

So, minimizing the cumulative residual KL divergence is same as minimizing the right hand side of the equation (3.10). First and second terms of right hand side of equation (3.10) are independent to the dependence tree, therefore minimizing the cumulative residual KL divergence is equivalent to maximizing the sum of cumulative residual mutual information in each branch.

To illustrate the implications of Theorem-1, we now apply the cumulative residual entropy approach to the same probability distribution used in Chow-Liu (1968) to compare the two approaches.

Example 3.1:

Consider four binary variables where each variable takes on values "0" and "1". Table-3.1 shows the outcomes and corresponding probabilities of joint distribution.

Table 3.1: $2\times2\times2\times2$ Joint Probability Distribution

X1	**X2**	**X3**	**X4**	**P(x_1,x_2,x_3,x_4)**
0	0	0	0	**0.10**
0	0	0	1	**0.10**
0	0	1	0	**0.05**
0	0	1	1	**0.05**
0	1	0	0	**0.00**
0	1	0	1	**0.00**
0	1	1	0	**0.10**
0	1	1	1	**0.05**
1	0	0	0	**0.05**
1	0	0	1	**0.10**
1	0	1	0	**0.00**
1	0	1	1	**0.00**
1	1	0	0	**0.05**
1	1	0	1	**0.05**
1	1	1	0	**0.15**
1	1	1	1	**0.15**
				$\sum p_i =$ **1.000**

To compare both methods we calculate the mutual information and cumulative residual mutual information between pairs of variables. All combination of pairs of variables and mutual information and cumulative residual mutual information quantities are given at Table-3.2.

Table 3.2: Mutual Information Quantities between pairs of variables

Pair of variables	MI	CRE_{MI}
X1-X2	0.07900	0.11175
X1-X3	0.00005	0.00249
X1-X4	0.00510	0.02610
X2-X3	0.18900	0.17872
X2-X4	0.00510	0.02383
X3-X4	0.00510	0.02383

We consider two error measures for the deviation of the approximation distribution from true distribution: the absolute deviation, and least squares error. Table 3.3 provides the formulae for error measures used in this paper.

Table 3.3: Error Measures

Error Measure	Formula
Absolute Deviation	$\sum_{i=1}^{n} \left\| p_i - p_i^* \right\|$
Least Squares Error	$\sum_{i=1}^{n} \left(p_i - p_i^* \right)^2$

We construct the optimal first-order dependence trees using mutual information and cumulative residual mutual information pairs. Figure-3.2 shows the optimal dependence tree approximations. The first three diagrams in Figure-3.2 are identical to those in Chow-Liu's paper. The fourth one is the dependence tree obtained using cumulative residual entropy.

For comparison purposes of two approximation methods, we calculate the absolute deviation and least square error between the true distribution and approximate distributions. From Table 3.4, we see that the absolute deviation between the first dependence tree approximation of Chow-Liu and true distribution is 0.260055 and the least square errors is 0.005984. The absolute deviation and least squares error between dependence trees found by CRE and true distribution are identical to the

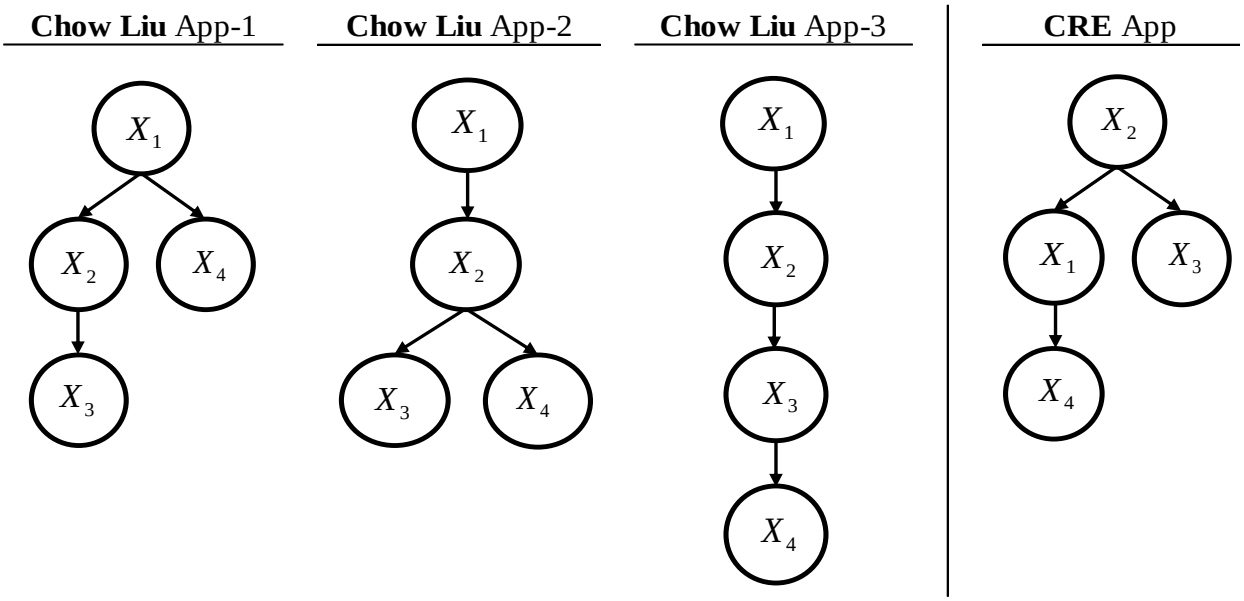


Figure 3.2: Optimal Tree Approximations calculated by Chow-Liu and CRE method

absolute deviation and least squares error between the first dependence tree approximation of Chow-Liu and true distribution. On the other hand, the absolute deviation and least squares error between second and third approximations of Chow-Liu and true distribution is higher than the absolute deviation and least squares error between dependence tree found by CRE method and true distribution; so these two dependence tree are not the best approximation.

Table 3.4: Optimal Tree Approximations calculated by Chow-Liu and CRE method

	Absolute Deviation	Least Squares Error
Chow-Liu 1	**0.2601**	**0.005984**
Chow-Liu 2	**0.2752**	**0.005989**
Chow Liu 3	**0.2752**	**0.005989**
CRE	**0.2601**	**0.005984**

For this specific example, CRE method approximates a first-order dependence tree same as Chow-Liu's first approximation and a better first-order dependence tree than second and third

approximations of Chow-Liu method. Therefore, we can say that our proposed method is at least as good as Chow-Liu's method. We now measure the performance of CRE and Chow-Liu methods in the long run by a Monte Carlo simulation in section 3.4.

3.4. Monte Carlo Simulation for CRE First-order Dependence Tree

We conducted a simulation to compare Chow-Liu method with CRE method. Simulation steps are shown in Figure-3.3. For numeric illustration, we discuss the simulation steps in terms of a several multi-variate distributions each variable has different values.

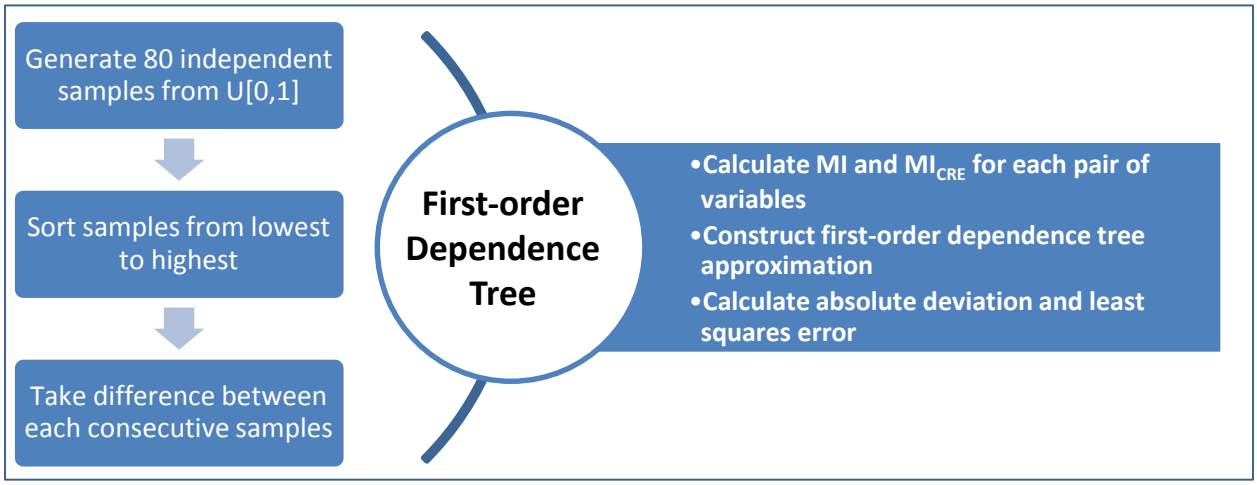


Figure 3.3: Monte Carlo Simulation Steps for Chow-Liu and CRE Approximations

We generated 10 million discrete joint probability distribution samples to check performance and accuracy of Chow-Liu's method and cumulative residual entropy method. Table 3.5 displays a summary of mean and variance of errors of second order joint probability distributions calculated by Chow-Liu and cumulative residual entropy methods.

Table 3.5: Comparison of Four-variate Approximations of Chow-Liu and CRE method

	Absolute Deviation		Least Squares Error	
	Mean	St.Dev.	Mean	St.Dev.
Chow-Liu	**0.5888**	**0.0549**	**0.0077**	**0.0018**
CRE	**0.5879**	**0.0545**	**0.0074**	**0.0017**

From Table 3.5, for the case of $3\times3\times3\times3$ joint distributions, we have found that Chow-Liu and CRE methods' results are almost exactly same after 10 million runs. The mean of absolute deviation for Chow-Liu method is 0.5888, and for CRE method is 0.5879. The ratio of the means of absolute deviation of Cho-Liu's method to the CRE method is less than $(0.5888/0.5879)\approx0.15\%$. Also, the mean of the least squares error between Chow-Liu approximate distribution and true distribution is 0.0077, and for CRE method, the least square error is 0.0074 which are a small deviation in many problems.

For convenience, we also ran another simulation with several different combination of joint distributions, including three binary variables, three three-outcome variables, three four-outcome variables, three five-outcome variables, four binary variables, four three-outcome variables, four four-outcome variables, and four five-outcome variables. Table 3.6 displays a summary of mean and variance of errors of our second-order joint probability distributions calculated by Chow-Liu and cumulative residual entropy methods.

From the simulation results in Table 3.6, we first observe that the mean and variance of errors are very close in the long run for the two methods, which means that the two approximation methods are very close but not identical.

Second, we observe that the mean value of absolute deviation for our CRE approximation method and traditional Chow-Liu method increases with the number of outcomes of a variable, while the mean value of least squares error decreases, implying that these two methods are sensitive to the number of variables and its outcomes.

Table 3.6: Simulation results of Chow-Liu and CRE-based approximations

	Absolute Deviation				Least Squares			
	Chow-Liu		CRE		Chow-Liu		CRE	
	Mean	St.dev	Mean	St.dev	Mean	St.dev	Mean	St.dev
Three binary variables	0.1810	0.1256	0.1822	0.1264	0.0074	0.0094	0.0077	0.0099
Three variables each has three values	0.3886	0.0866	0.3934	0.0914	0.0095	0.0049	0.0098	0.0053
Three variables each has four values	0.4883	0.0596	0.4942	0.0620	0.0073	0.0017	0.0071	0.0019
Three variables each has five values	0.5505	0.0469	0.5484	0.0441	0.0045	0.009	0.0042	0.008
Four binary variables	0.3612	0.1121	0.3548	0.1118	0.0138	0.0084	0.0136	0.0083
Four variables each has three values	0.5888	0.0549	0.5879	0.0545	0.0077	0.0018	0.0074	0.0017
Four variables each has four values	0.6638	0.0319	0.6557	0.0311	0.0032	0.0004	0.0030	0.0004
Four variables each has five values	0.6980	0.0254	0.6915	0.0241	0.0015	0.0003	0.0015	0.0003

3.5. Comparison of CRE Based First Order Dependence Trees with Chow and Liu's First Order Dependence Tree

The Chow-Liu approximation and CRE-based approximation give similar accuracy results and sometimes approximate exactly the same distribution in the long run. As seen in figure 3.4, for a four binary variable case they both approximated the same first-order dependence trees for more than 89% of the samples.

In this study, we aim to identify where the CRE-based first-order dependence tree differs from Chow-Liu's first-order dependence tree and explore under what conditions or situations the former method gives more accurate results. To further analyze and compare these two methods in the long run, we generate different multivariate probability distributions. We analyze the accuracy of approximations for three- and four-variate distributions with increasing numbers of values for each variable. We start with three binary random variables and then we add a new value to each variable to observe the changes in percentage of similarities and differences between these two approximation methods.

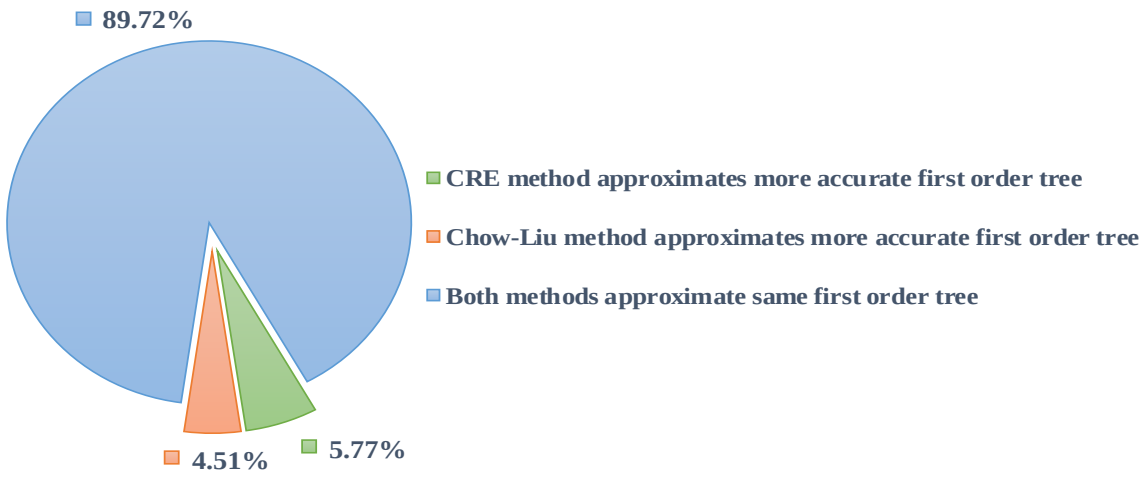


Figure 3.4: Simulation results of absolute deviation of Chow-Liu and CRE-based approximations

We do the same procedure for four random variable cases with two, three, four, five, six and seven values. We first analyze how much percent of the approximation of the sampling distributions are the same, then analyze the approximation percentage of the sampling distributions where the CRE-based first-order dependence tree method is more accurate and vice versa. We further analyze how the approximate distributions' accuracy changes with the number of random variables. In this case, we hold the number of variable values constant and increase the number of variables from three to six. Finally, we analyze how the accuracy changes with the dependence structure among variables. In our analysis, we use total correlation measure, a dependency measure in information theory that can be used to quantify the dependency among a set of random variables. We do our analysis for three- and four-variate probability distributions with two, three, and four values of each variable. In this analysis, three different accuracy measures are used: absolute deviation (AD), least squares (LS), and maximum deviation (MD).

3.5.1. Effects of Increasing Number of Values of Each Variable: Three Variable Case

We start with three binary random variables and add a new value to the variables to observe the changes in the percentage of similarities and differences between these two approximation methods. Figure-3.5 shows the results for the three-variable case using absolute deviation as the accuracy measure.

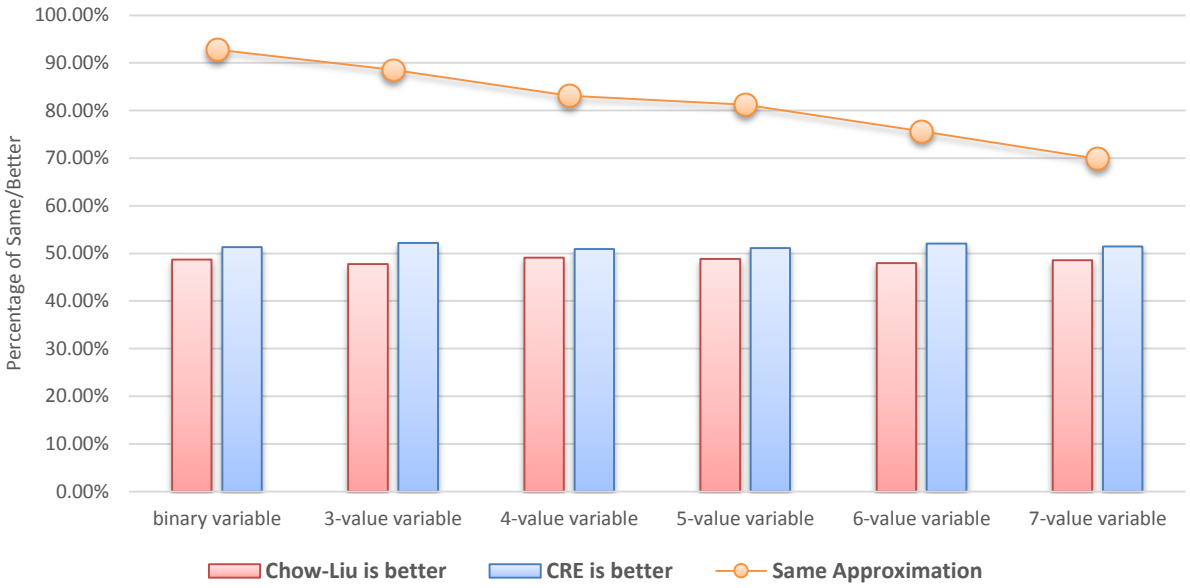
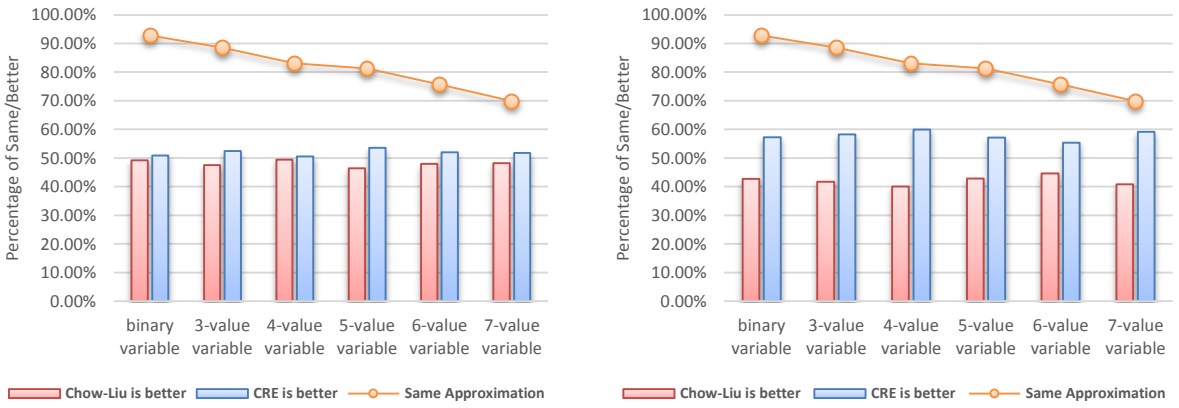


Figure 3.5: Comparison of Chow-Liu and CRE-based First-order Dependence Tree Approximations based on Absolute Deviation Measure for Three-Variable Case

From Figure 3.5, we can say that the difference between these two approximation methods increases as the number of values of each variable increases. For the binary variable case, both methods approximated the same first-order dependence trees more than 92% of the samples and this percentage decreased to 70% when each variable has seven different values. These results make sense because there is only three variables and three different pairwise combinations of variables; X1-X2, X1-X3, and X2-X3. Therefore, we can only find three possible first-order dependence trees and most likely these two methods approximate the same first-order dependence tree most of the time.

Another observation from Figure 3.5 is that the CRE-based first-order dependence tree method is slightly more accurate than the Chow-Liu method, in that the percentages for CRE method are slightly above the 50% and somewhat below 50% for Chow-Liu method.

We also run the same simulation for least squares measure and maximum deviation measure. Figure-3.6 shows the results of three variable case (Figure 3.6-a: Least Squares, Figure 3.6-b: Maximum Deviation).



(a) Least Squares **(b)** Maximum Deviation

Figure 3.6: Comparison of Chow-Liu and CRE-based First-order Dependence Tree Approximations based on Least Squares and Maximum Deviation Measures for Three Variable Case

Figure 3.6 shows similar behavior to those in Figure 3.5. However, in Figure 3.6-b, the difference between Chow-Liu and CRE methods increases when maximum absolute deviation is used as the accuracy measure. So, we may say that Chow-Liu method obtained more extreme deviations from the true probability distributions than that obtained by the CRE-based first-order dependence tree method.

3.5.2. Effects of Increasing Number of Values of Each Variable: Four Variable Case

We now do the same procedure for a four-random-variable case with two, three, four, five, six and seven values. Figure-3.7 shows the results for the four-variable case using absolute deviation as the accuracy measure.

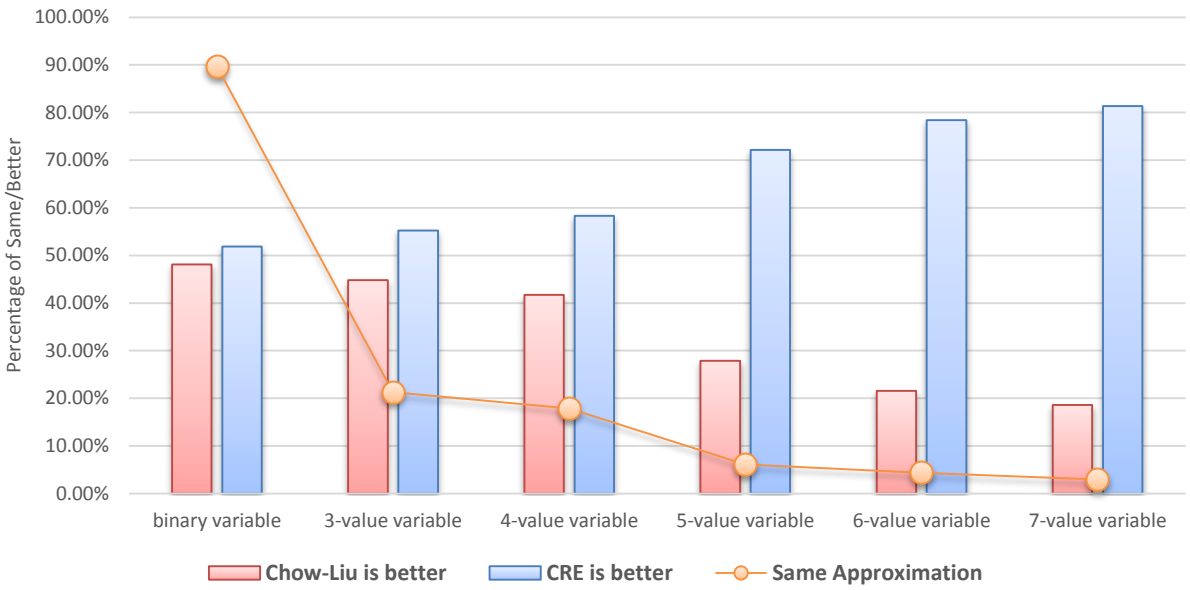
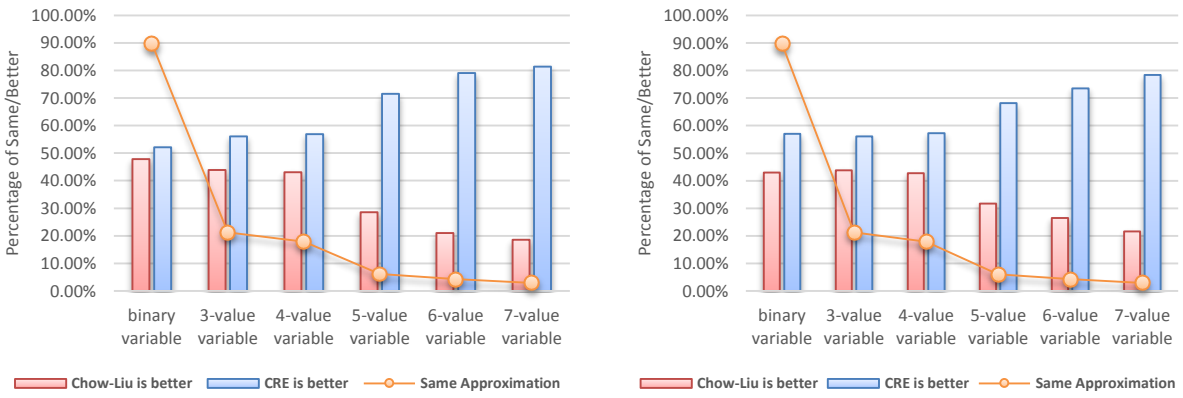


Figure 3.7: Comparison of Chow-Liu and CRE-based First-order Dependence Tree Approximations based on Absolute Deviation Measure for Four Variable Case

The results for the four-variable case are different from those for the three-variable case. In this case, four variables yield 125 possible first-order dependence trees. In the four-binary-variable case, these two methods approximated the same first-order dependence trees more than 89% of the samples. However, these two approximation methods approximate the same distribution for less than 20% of the samples for three or more values of variables. The results show that the percentage of approximations from the same first-order dependence tree decreases exponentially as the number of values of variables increase for AD, LS, and MD measures. For a low number of values of variables, the approximations yielded by both methods are nearly identical, but the CRE-based first-order dependence tree methods is more accurate than the Chow-Liu method as the number of values of variables increases. In Figure 3.8, we also plot the least squares and maximum deviation results, which are very similar to absolute deviation results.

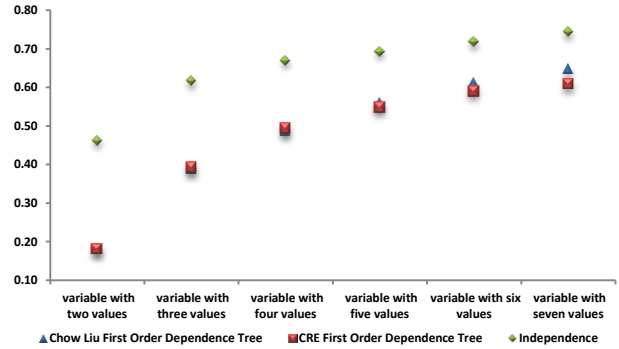


(a) Least Squares **(b)** Maximum Deviation

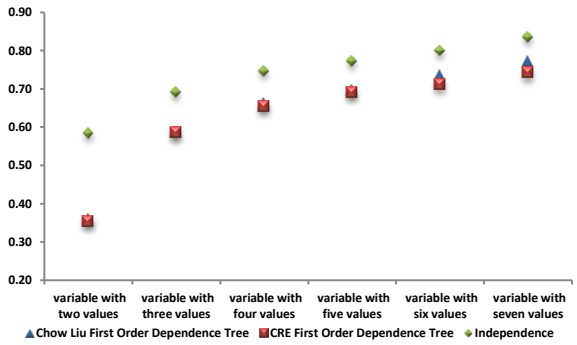
Figure 3.8: Comparison of Chow-Liu and CRE-based First-order Dependence Tree Approximations based on Least Squares and Maximum Deviation Measures for Four Variable Case

We also calculate the mean accuracy (error) values of approximations for each measure for three and four variable cases with increasing number of values for each variable and for independence cases. Figure-3.9 shows the results of mean accuracy values: although there is a difference for the percentage results of the four-variate case, the average accuracy values for each method are very close. Also, the error values increase with the number of values of variables increases for absolute deviation measure, however, decreases for least squares and maximum deviation. For higher number of values, the CRE-based first-order dependence tree is slightly better than the Chow-Liu method. These results are consistent with the previous results.

We can say that for higher values of variables, the CRE-based first-order dependence tree method outperforms the Chow-Liu first-order dependence tree method. We also observe that as the number of values of variables increases, the accuracy values of the independence case converge to the accuracy values of Chow-Liu and CRE-based first-order dependence tree approximation approaches.

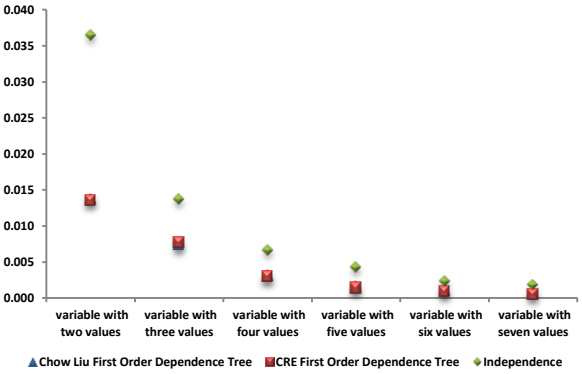


(a) Absolute Deviation of Three Variable Case

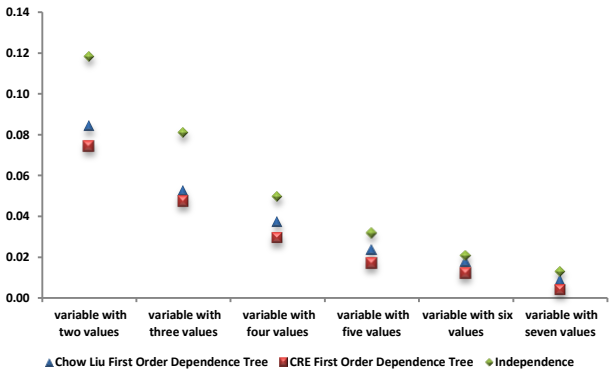


(b) Absolute Deviation of Four Variable Case

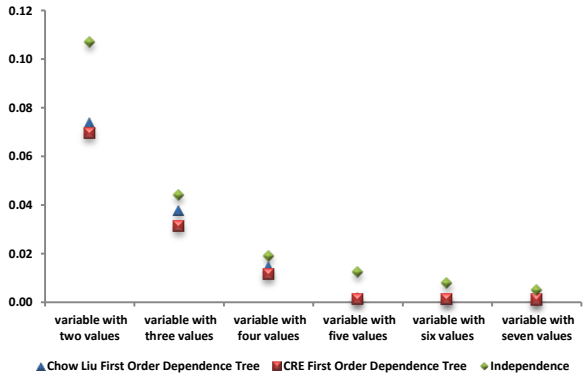
(c) Least Squares of Three Variable Case



(d) Least Squares of Four Variable Case



(e) Maximum Deviation of Three Variable Case



(f) Maximum Deviation of Four Variable Case

Figure 3.9: The Mean Accuracy (Error) Values of Each Measure for Three and Four Variable Cases with Increasing Number of Values of Each Variable

3.5.3. Effects of Increasing the Number of Variables in Multivariate Distributions

We now analyze how accuracy of the approximate distributions changes as the number of random variables increases in a multivariate distribution. In this case, we hold the number of values

of variables constant and increase the number of variables from three to six. We do our analysis in two parts. First, we analyze for binary variables, then we analyze the case each variable has three values. Table-3.7 shows the results of mean values of each measure for different multivariate distributions.

Table 3.7: Simulation Results of Mean Values of Error Measures as the Number of Variables Increases

	Absolute Deviation		**Least Squares**		**Maximum Deviation**	
	Chow-Liu	**CRE**	**Chow-Liu**	**CRE**	**Chow-Liu**	**CRE**
Three binary variables	0.1810	0.1822	0.0092	0.0098	0.0844	0.0745
Four binary variables	0.3612	0.3548	0.0138	0.0136	0.0738	0.0696
Five binary variables	0.5391	0.5181	0.0116	0.0108	0.0576	0.0545
Six binary variables	0.6267	0.6144	0.0109	0.0104	0.0431	0.0408
Three variables each has three values	0.3886	0.3934	0.0074	0.0072	0.0525	0.0473
Four variables each has three values	0.5888	0.5779	0.0074	0.0077	0.0377	0.0314
Five variables each has three values	0.6800	0.6719	0.0034	0.0032	0.0168	0.0160
Six variables each has three values	0.7174	0.7085	0.0013	0.0012	0.0075	0.0075

The mean value of absolute deviation for the Chow-Liu method is 0.5888 and for the CRE method is 0.5779. The ratio of the mean of absolute deviation of the Cho-Liu method to that of the CRE method is less than (0.5888/0.5779)=0.19%. Also, the mean of the least squares error between the Chow-Liu approximate distribution and the true distribution is 0.0074, whereas for the CRE method the least square error is 0.0077, a small deviation in many problems.

In Table-3.7 we first observe that the mean error value is very close in the long run for each of the two methods, which means that the two approximation methods are very close but not identical. We can say that these two methods are very similar and approximate almost identical joint probability distributions for lower number of variables and values of variables. So, our proposed

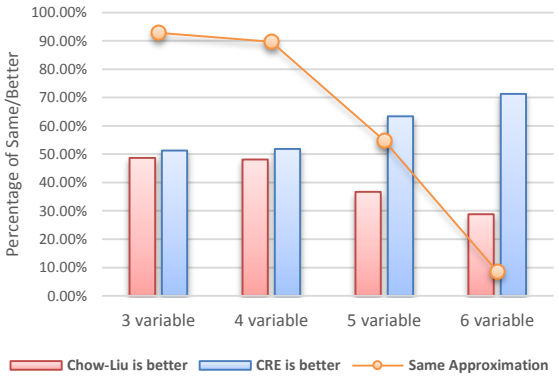
method can be used as an alternative method to the Chow-Liu method if cumulative functions are present.

In Table-3.7 we also observe that the mean values of absolute deviation for our CRE approximation method and traditional Chow-Liu method both increase with the number of variables while the mean value of least squares error and maximum deviation decrease. This implies that these two methods are sensitive to the number of variables and its outcomes.

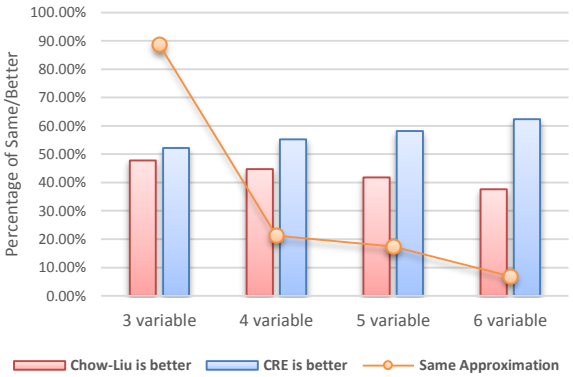
Also, we explore the changes in percentage of similarities and differences between these two approximation methods as the number of variables increases in the multivariate distribution. Figure-3.10 shows the results for the binary case (Figure 3.10-a-c-e) and the three-value case (Figure 3.10-b-d-f) for each error measure. We found that the Chow-Liu and CRE methods approximated the same first-order dependence tree for more than 90% of the samples in the three- and four-binary variable cases. However, the percentage of sample distributions that were identical for both methods decreased sharply as the number of variables increased. Also, where the two methods approximate different first-order dependence trees, we analyze the percent of the approximation of the sampling distributions where CRE-based first-order dependence tree provides more accurate approximations and vice versa.

The results of these two approaches, as shown in Figure 3.10 and Table 3.7, are very similar to those obtained with increasing number of values of variables. For lower number of variables, these two methods show very similar behavior; however,the CRE-based first-order dependence tree method provides better approximations than the Chow-Liu method as the number of variables increases. We can see that the approximations provided by the CRE-based first-order dependence tree are more than 60% more accurate than those for first-order dependence trees for a 5-binary

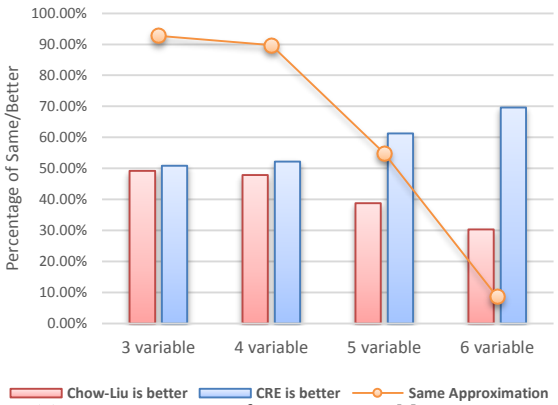
variable case, and more than 70% accurate for 6-binary variable case (Figure-a-c-e). This similarity also extends to the approximations for three-value variables . (Figure-b-d-f).



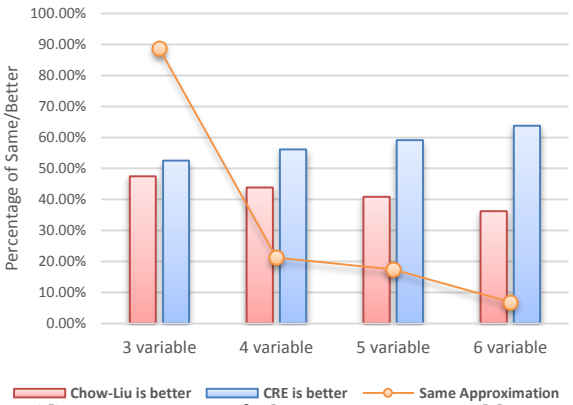
(a) Absolute Deviation of Binary Variable Case



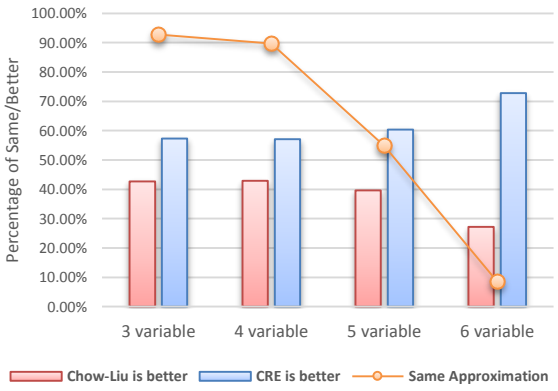
(b) Absolute Deviation of Three-outcome Variable Case



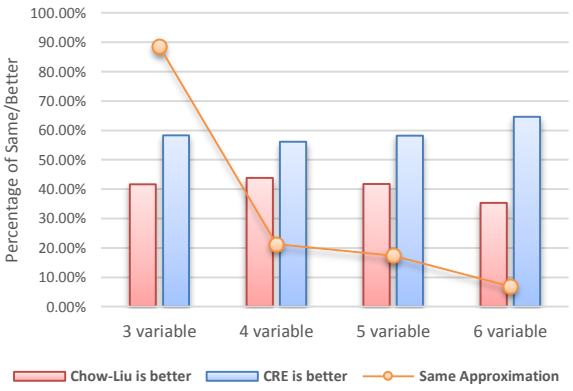
(c) Least Squares of Binary Variable Case



(d) Least Squares of Three-outcome Variable Case



(e) Maximum Deviation of Binary Variable Case



(f) Maximum Deviation of Three-outcome Variable Case

Figure 3.10: The Percent of the Approximation of the Sampling Distributions That CRE-based and Chow-Liu First-order Dependence Tree Methods are Same and Different

Thus, the CRE method outperforms the Chow-Liu method for more than four variables. Moreover, these results indicate that the effect of increasing the number of values of each variable is similar to the effect of increasing the number of random variables.

3.5.4. Effects of Changing the Dependence Structure Between Variables

We now discuss the effects of changing dependence structure among variables. We use total correlation (TC) as a dependency measure. In probability theory. and particularly in information theory, total correlation (Watanabe, 1960) is the amount of information shared among the variables. It quantifies the dependency among random variables. Total correlation is non-negative and that it is zero if and only if the random variables are independent ($TC \geq 0$).

Let $\{X_1, X_2, \cdots, X_n\}$ be discrete random variables. The total correlation is defined as

$$TC(X_1, X_2, \cdots, X_n) = \sum_{x_1 \in X_1} \sum_{x_2 \in X_2} \cdots \sum_{x_n \in X_n} p(x_1, x_2, \cdots, x_n) \log \frac{p(x_1, x_2, \cdots, x_n)}{p(x_1) \times p(x_2) \times \cdots \times p(x_n)} \quad (3.11)$$

where $p(x_1, x_2, \cdots, x_n)$ is the probability mass function of variables $X_1, X_2, \cdots, X_n$, and $p(x_1)$ is the marginal probability distribution of variable X_1.

This total correlation formulation can be reduced to the simpler difference of entropies as

$$TC(X_1, X_2, \cdots, X_n) = \left[\sum_{i=1}^{n} H(X_i) \right] - H(X_1, X_2, \cdots, X_n) \quad (3.12)$$

where $H(X_i)$ is the entropy of variable X_i, $i = 1, 2, \cdots, n$.

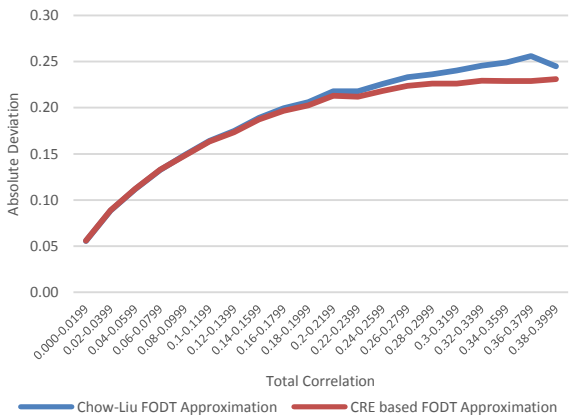
We run a simulation with several different combination of joint distributions, including three binary variables, three three-outcome variables, three four-outcome variables, four binary variables, four three-outcome variables, and four four-outcome variables. Figure 3.11, 3.12, and 3.13 display the mean values of errors of our second-order joint probability distributions calculated

by Chow-Liu and cumulative residual entropy methods for absolute deviation, least squares, and maximum deviation measures, respectively.

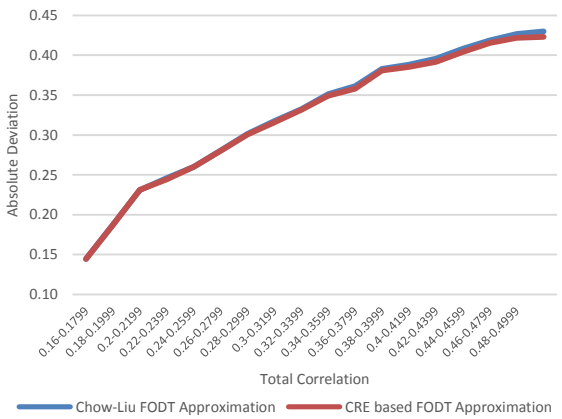
From Figure 3.11, 3.12, and 3.13, we can say that the cumulative residual approximation and the Chow-Liu approximations give almost identical accuracy results when the correlation between variables are low. As the correlation among variables increases, the error of approximations increases and CRE-based first-order dependence tree approximation is better than Chow-Liu first-order dependence tree approximation for all accuracy measures. So, we can say that CRE-based first-order dependence tree approximation shows higher accuracy as the correlation increases. For instance, for four three-outcome variable, when the average total correlation is between 0.2 and 0.2199, mean absolute deviation of Chow-Liu method is ~0.449, and mean absolute deviation of the CRE-based approximation method is ~0.442. On the other hand, when the average total correlation is between 0.5-0.519, then the absolute deviation of the CRE-based method is ~0.677, and the absolute deviation of Chow-Liu method is ~0.708.

Another observation from Figure 3.11, 3.12, and 3.13 is that the difference between the two methods is small when the variables are binary. However, the the difference between the two methods noticeably increases with the number of values for each variable. Thus, these two methods are sensitive to the number of values of variables, or we can say that the dimension of the probability distribution has an effect on the accuracy of the approximations.

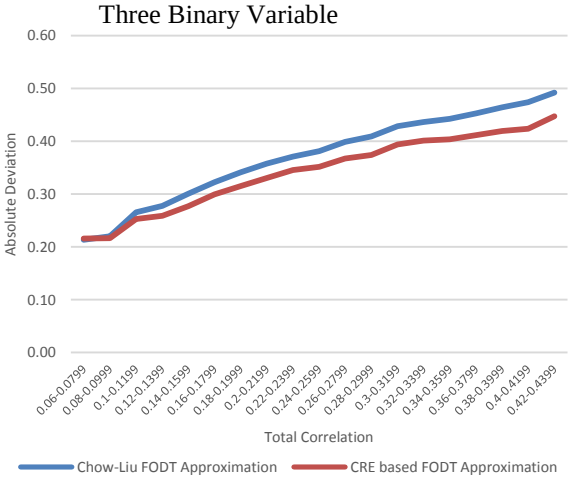
Note that the error of approximations increases as both the correlation and dimension of the multivariate probability distribution increases. Also note that the CRE-based method gives much better approximation when the total correlation is high.



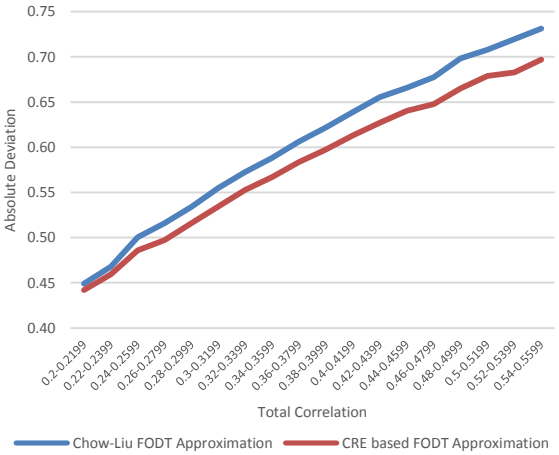
(a) Total Correlation vs. Absolute Deviation for Three Binary Variable



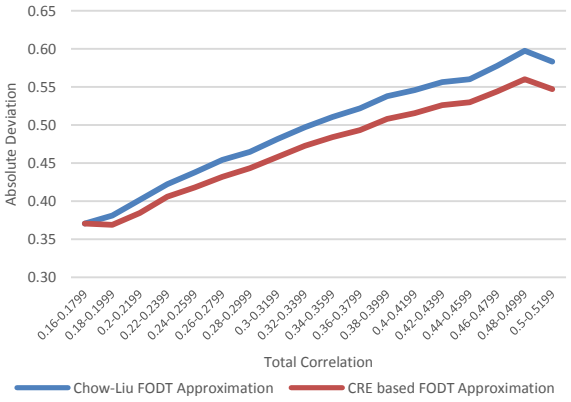
(b) Total Correlation vs. Absolute Deviation for Four Binary Variable



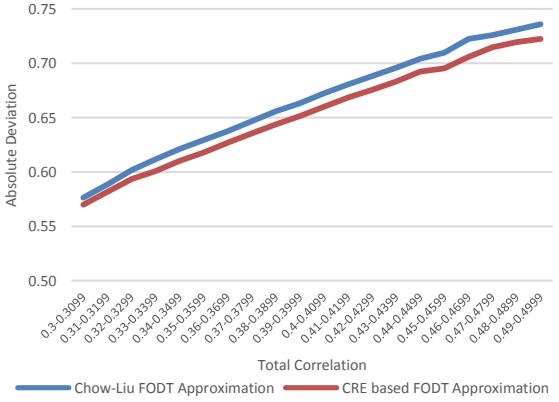
(c) Total Correlation vs. Absolute Deviation for Three Three-outcome Variable



(d) Total Correlation vs. Absolute Deviation for Four Three-outcome Variable

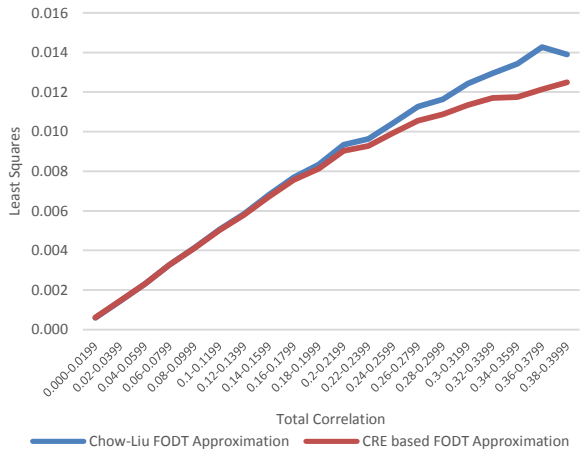


(e) Total Correlation vs. Absolute Deviation for Three Four-outcome Variable

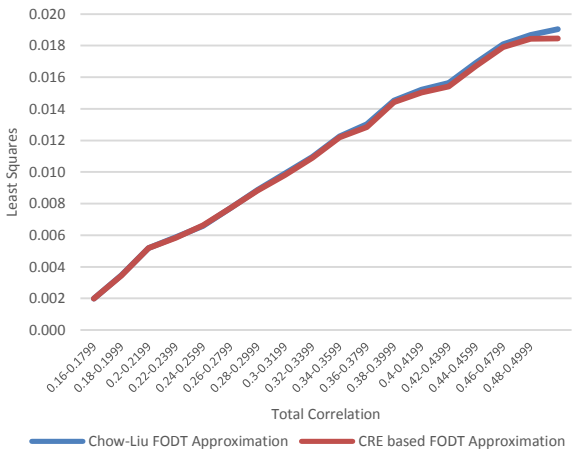


(f) Total Correlation vs. Absolute Deviation for Four Four-outcome Variable

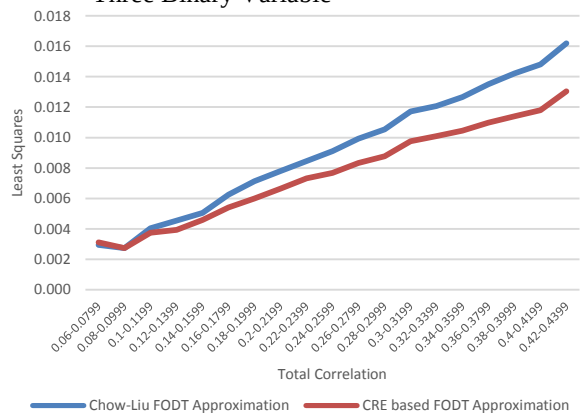
Figure 3.11: Average Values of Errors of Second-Order Joint Probability Distributions Calculated by Chow-Liu and Cumulative Residual Entropy Methods Using Absolute Deviation as Accuracy Measure



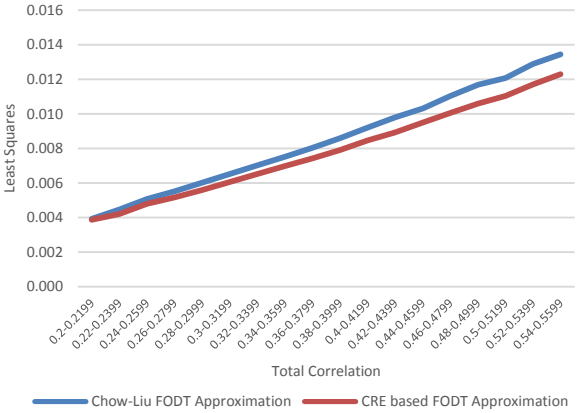
(a) Total Correlation vs. Least Squares for Three Binary Variable



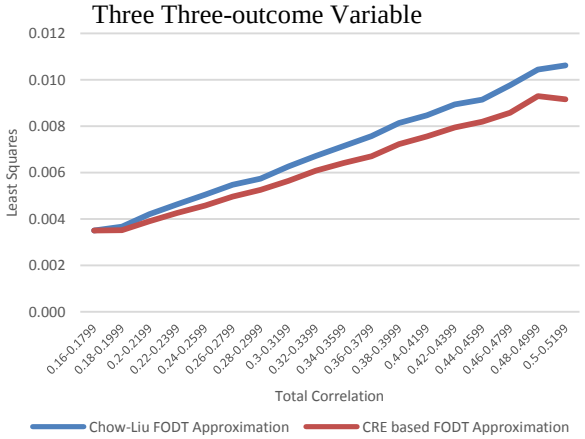
(b) Total Correlation vs. Absolute Deviation for Four Binary Variable



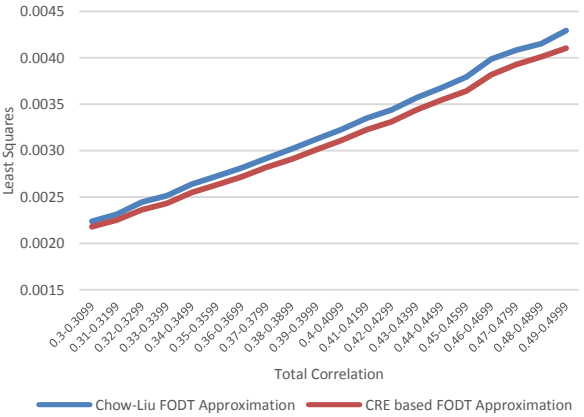
(c) Total Correlation vs. Least Squares for Three Three-outcome Variable



(d) Total Correlation vs. Least Squares for Four Three-outcome Variable

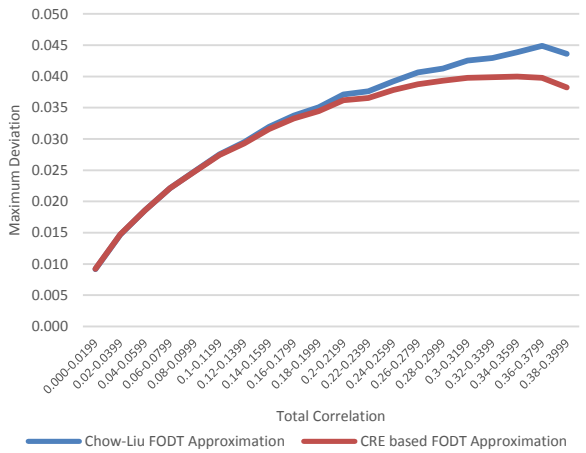


(e) Total Correlation vs. Least Squares for Three Four-outcome Variable

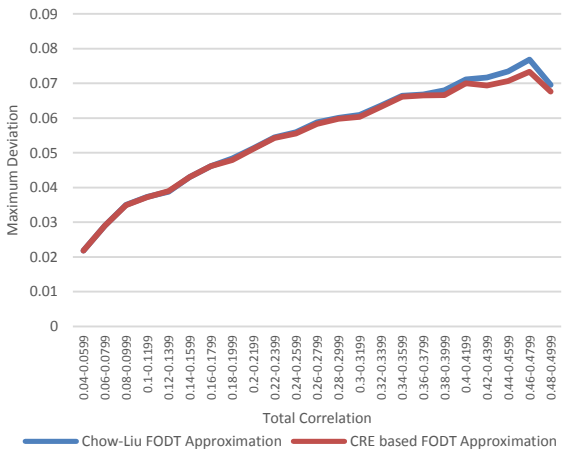


(f) Total Correlation vs. Least Squares for Four Four-outcome Variable

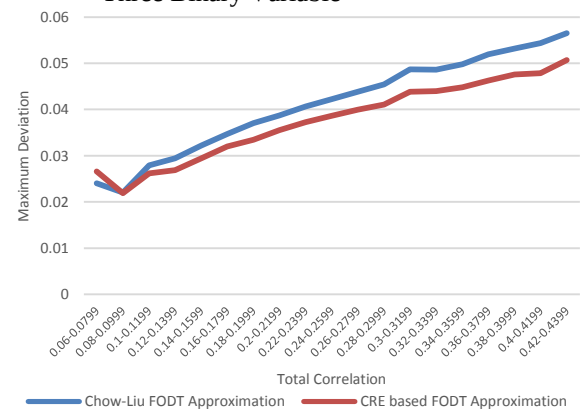
Figure 3.12: Average Values of Errors of Second-Order Joint Probability Distributions Calculated by Chow-Liu and Cumulative Residual Entropy Methods Using Least Squares as Accuracy Measure



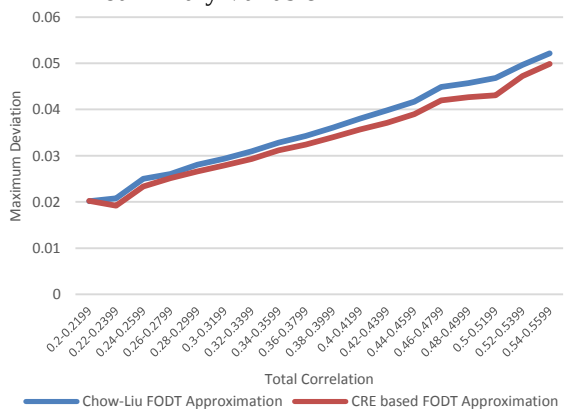
(a) Total Correlation vs. Maximum Deviation for Three Binary Variable



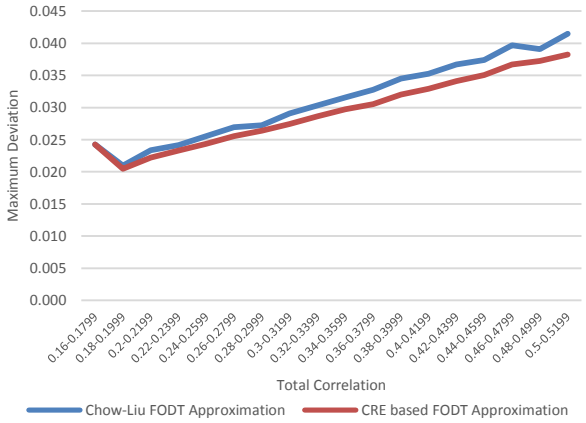
(b) Total Correlation vs. Maximum Deviation for Four Binary Variable



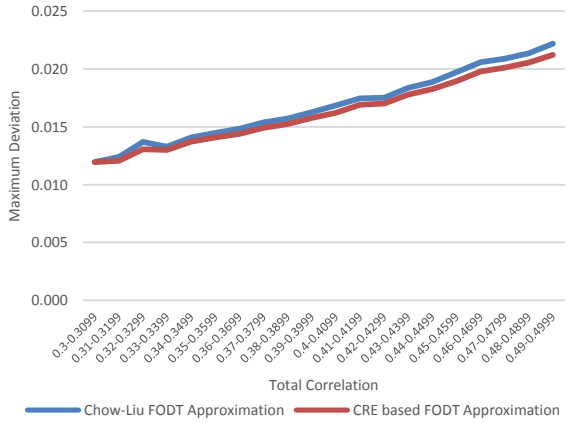
(c) Total Correlation vs. Maximum Deviation for Three Three-outcome Variable



(d) Total Correlation vs. Maximum Deviation for Four Three-outcome Variable



(e) Total Correlation vs. Maximum Deviation for Three Four-outcome Variable



(f) Total Correlation vs. Maximum Deviation for Four Four-outcome Variable

Figure 3.13: Average Values of Errors of Second-Order Joint Probability Distributions Calculated by Chow-Liu and Cumulative Residual Entropy Methods Using Maximum Deviation as Accuracy Measure

3.6. Decision Problems where CRE Based First Order Dependence Trees are More Suitable or Preferable

In the case of censored data, in which the observed value of some variable is only partially known, the CRE-based first-order dependence tree must be used. The problem of censored data is related to the problem of missing data, where the observed value of some variable is entirely unknown. Censoring is an important issue in survival analysis, representing a particular type of missing data. Thus, using survival functions instead of density functions in entropy formulation helps to take into account the censored data.

Also, survival functions generally should be used if the variable of interest is the time of an event, so CRE-based first-order dependence trees are preferable. Some examples of decision problems involving survival functions include time until onset of disease, time until stock market crash, time until equipment failure, time until earthquake, and so on. Let X be the time elapsed until a particular event occurs, such as death, infection, the appearance of a tumor, the development of some diseases, and so forth. Also, there are several other areas in which survival functions are used to make decisions, such as quality control stages in manufacturing production or preliminary medical examinations to determine whether a patient is suitable for treatment.

3.7. Summary of the Results

- The CRE-based first-order dependence tree approach can only be applied to numeric variables, but the Chow-Liu first-order dependence tree method can be applied to both numeric and non-numeric variables.
- For lower correlations, lower number of variables, and lower number of values of variables, both methods approximate almost the same first-order dependence tree.

- The CRE-based first-order dependence tree method gives more accurate approximation when the number of variables more than four.
- The accuracy of approximations decreases as the correlation and the dimension of the multivariate probability distribution increases
- The CRE-based method gives much better approximation when the total correlation and/or dimension of the multivariate distribution is high.
- For binary variable cases, both methods show almost same performance.
- The CRE-based first-order dependence tree method gives more accurate approximation as the number of values of variables increases.
- The CRE-based first-order dependence tree needs be used if some of the data are censored.
- The CRE-based first-order dependence tree is preferable if the variable consists of the time elapsed until an event occurs.

3.8. Conclusion

In this chapter, we discussed the problem of approximating multidimensional discrete probability distributions using first-order dependence trees. We showed that the optimal first-order dependence tree approximation in terms of the cumulative residual KL divergence is the one with the largest sum of cumulative residual mutual information pairs. We then ran a Monte Carlo simulation to illustrate the performance of the approximation. The results show that the cumulative residual approximation and the Chow-Liu approximations gives similar accuracy results. However, the results of these two approximation methods are not identical. So, we characterize when CRE-based first-order dependence tree yields approximations different from those of Chow-Liu's first-order dependence tree and explore under what conditions or situations that CRE-based first-order dependence tree method gives more accurate results.

We can say that CRE-based first-order dependence tree at least as good as the Chow-Liu's first-order dependence tree method. The only drawback of our proposed method is that it can only be applied to numeric variables. So, we can conclude that cumulative residual entropy method can be used as an alternative method to Chow-Liu's method if cumulative functions, especially survival functions, are present. Therefore, we don't need to calculate the density functions to approximate first-order dependence trees, and by using CRE method we can directly approximate first-order dependence trees from cumulative functions.

CHAPTER 4

CONSTRUCTION OF JOINT PROBABILITY DISTRIBUTIONS BASED ON MAXIMUM CUMULATIVE RESIDUAL ENTROPY

4.1. Introduction

This chapter proposes an approximation method based on the maximum cumulative residual entropy principle to construct representative joint probability distributions from its lower-order assessments. We formulize a maximum cumulative residual entropy that is similar to maximum entropy, but we use cumulative functions rather than density functions. Using cumulative functions in our proposed approach simplifies the maximum entropy formulation, reduces the number of assessments, captures dependence into model, and approximates less error prone joint probability distributions when only partial information is available. The maximum cumulative residual entropy principle keeps the probability distribution consistent with observed constraints. It yields a probability distribution which is "most likely" to have represented the observed data. The remainder of this chapter as follows: In section 2, we discuss the maximum cumulative residual entropy formulation. Section 3 presents a Monte Carlo simulation application of the new approach and discusses and interprets the results of simulation. Section 4 discusses the comparison of two approximation methods. Section 4 presents the conclusion remarks of the chapter.

4.2. Maximum Cumulative Residual Entropy

Laplace proposed in his publication "Principle of Insufficient Reason" that when one has no information to distinguish between the probabilities of two events, the best strategy is to consider them equally likely (Laplace, 1774). In the same analogy, if the only available information about

the probability distribution is the number of outcomes, the maximum entropy formulation assumes a uniform distribution with *n* outcomes as

$$p^*(x) = \frac{1}{n} \tag{4.1}$$

Moreover, if the marginal distributions of variables are known, then the maximum entropy formulation assumes independence between variables and produces a joint distribution which is equal to product of marginal distributions. For instance, if only marginal distributions of three variables, x, y, z are available, then the maximum entropy produces a joint distribution as

$$p^*(x, y, z) = p_x(x)\, p_y(y)\, p_y(y) \tag{4.2}$$

The maximum cumulative residual entropy formulation also produces same results with the information number of outcomes, and marginal survival function of random variables. If the only available information is the number of outcomes, then the maximum cumulative residual entropy finds a uniform survival function with n outcomes.

$$S^*(x) = \frac{n - x}{n} \tag{4.3}$$

If additionally marginal survival functions are known, the maximum cumulative residual entropy is consistent with maximum entropy and produces a joint distribution which is equal to product of marginal survival functions.

$$S^*(x, y, z) = S_x(x) S_y(y) S_y(y) \tag{4.4}$$

However, information of number of outcomes and marginal distributions is not enough to add dependence between variables into joint distribution. If we want to incorporate dependence

between variables into the maximum cumulative residual entropy formulation, we need to consider at least pairwise assessments between variables in addition to the marginal distributions of the variables. For instance, in the case of three variable decision problems, the available information should be marginal and/or pairwise assessments and the maximum cumulative residual entropy formulation of a three-variate joint distribution with the pairwise assessments is

$$
\begin{aligned}
S^*(x_{1i}, x_{2j}, x_{3k}) = \arg\max & - \sum_{x_{1i}, x_{2j}, x_{3k}} S(x_{1i}, x_{2j}, x_{3k}) \ln(S(x_{1i}, x_{2j}, x_{3k})) \\
s.t. & \\
& \sum_{x_{3k}} S(x_{1i}, x_{2j}, x_{3k}) = S(x_{1i}, x_{2j}) \\
& \sum_{x_{2j}} S(x_{1i}, x_{2j}, x_{3k}) = S(x_{1i}, x_{3k}) \\
& \sum_{x_{1i}} S(x_{1i}, x_{2j}, x_{3k}) = S(x_{2j}, x_{3k}) \\
& S(x_{1i}, x_{2j}, x_{3k}) \geq 0
\end{aligned}
\tag{4.5}
$$

where x_{1i} refers to i^{th} outcome of the first variable, and $S(x_{1i}, x_{2j}, x_{3k})$ refers to the three variable joint survival function of the i^{th} outcome of the first variable, j^{th} outcome of the second variable, and k^{th} outcome of the third variable. The solution to the maximum cumulative residual entropy for three variables using pairwise assessments is $S^*(x) = e^{-1-\lambda_0-\lambda_{ij.}-\lambda_{i.k}-\lambda_{.jk}}$.

Example 4.1:

Assume we want to find the maximum cumulative residual entropy for the joint distributions of three binary variables using its pairwise assessments. The binary variables take values “0” and “1” in this example. The probability mass function of three binary variables is given as

Table 4.1: $2\times2\times2$ Joint Probability Distribution

X1	X2	X3	p(x_1, x_2, x_3)
0	0	0	0.380
0	0	1	0.097
0	1	0	0.169
0	1	1	0.178
1	0	0	0.087
1	0	1	0.048
1	1	0	0.009
1	1	1	0.032
			$\sum p_i =$ **1.000**

There are a total of 12 pairwise cumulative residual assessments that can be found and these pairwise assessments are shown in Table-4.2.

Table 4.2: Pairwise Probabilities for given $2\times2\times2$

Pairwise Assessment	Probability
$S_{11\_}$	1.000
$S_{12\_}$	0.388
$S_{21\_}$	0.176
$S_{22\_}$	0.040
S_{1_1}	1.000
S_{1_2}	0.354
S_{2_1}	0.176
S_{2_2}	0.040
$S_{\_11}$	1.000
$S_{\_12}$	0.354
$S_{\_21}$	0.388
$S_{\_22}$	0.209

The pairwise assessment, $S_{12\_}$, define a bivariate survival function that the first variable has values bigger than one, the second variable has values bigger two, and the third variable takes any

value. This problem can be solved as an optimization problem that maximizes a concave function subject to some given constraints. In the analysis part, we don't use the full pairwise assessments because six of them are redundant and we eliminate them from our optimization analysis to simplify the approximation process. Then, the maximum cumulative residual entropy formulation for a three binary variable distribution with the independent constraints can be formulized as

$$
\begin{aligned}
S^*(x_1,x_2,x_3) &= \arg\max_{S(x_1,x_2,x_3)} -\sum_{x_1,x_2,x_3} S(x_1,x_2,x_3)\ln(S(x_1,x_2,x_3)) \\
&s.t. \\
&\sum_{x_3} S(x_1=1,x_2=1,x_3) = 1.000 \qquad \sum_{x_3} S(x_1=2,x_2=2,x_3) = 0.041 \\
&\sum_{x_3} S(x_1=1,x_2=2,x_3) = 0.388 \qquad \sum_{x_2} S(x_1=1,x_2,x_3=2) = 0.354 \\
&\sum_{x_3} S(x_1=2,x_2=1,x_3) = 0.176 \qquad \sum_{x_1} S(x_1,x_2=2,x_3=2) = 0.209 \\
&S(x_{1i},x_{2j},x_{3k}) \geq 0
\end{aligned}
\tag{4.6}
$$

We solve this optimization problem in MATLAB R2013a using the function "fmincon" to find the approximate probability distribution using given constraints. For comparison purposes, we also calculate the maximum entropy probability distribution. Table-4.3 shows the results of the analysis.

Table 4.3: Results of Max-Entropy and Max CRE for Three Binary Variables

Outcomes	True Distribution	Maximum Entropy	Maximum CRE
p_{111}	0.380	0.382	0.382
p_{112}	0.097	0.095	0.096
p_{121}	0.170	0.168	0.169
p_{122}	0.178	0.180	0.180
p_{211}	0.087	0.085	0.085
p_{212}	0.048	0.050	0.050
p_{221}	0.009	0.011	0.011
p_{222}	0.031	0.029	0.029

As we can see from the results, the approximate probability distributions are very close to each other. We compared the values of true distribution with the values obtained using maximum entropy and maximum cumulative residual entropy approximations. For this specific example, the results are almost identical. The maximum cumulative residual entropy approximation method performs similar to maximum entropy method. As a result, we can say the maximum cumulative residual entropy approach is also a good approximation method and an alternative method to maximum entropy to estimate the probability distributions using lower order assessments.

Also, we calculate the error between true distribution and maximum entropy formulation, and also the error between true distribution and maximum cumulative residual entropy formulation. We use absolute deviation and least squares error to measure the error. Table-4.4 shows the absolute deviation and the least squares error between the three binary variable joint distribution and two approximate distribution (maximum entropy and maximum CRE) using pairwise assessments.

Table 4.4: Error Results between True Distribution and Approximate Distributions

Error Measure	Maximum Entropy	Maximum CRE
Abs. Deviation	0.01809	0.01805
LSE	4.091×10^{-5}	4.091×10^{-5}

From Table-4.4, we observe that there is no a significant difference between the approximation provided by the maximum entropy approach and the maximum cumulative residual entropy approach. The absolute deviation of maximum entropy is 0.01809 and the absolute deviation of maximum cumulative residual entropy is only 0.01805. The difference between two approximation methods is less than 0.22% which is so small and the two methods are almost identical for the

specific example. Now, we discuss the performance of the maximum entropy and maximum cumulative residual entropy approximation in the long run using a Monte Carlo simulation.

4.3. Application of Maximum CRE using Monte Carlo Simulation

In this section, we discuss the performance of maximum entropy and maximum CRE approximation based on absolute deviation and least squares error using a Monte Carlo simulation. In this analysis, we conduct a simulation with four variables and each of four variables has three different outcomes. We conduct the simulation to approximate a four variate joint distribution using its lower assessments. We generate one million $3\times3\times3\times3$ outcome discrete joint probability distribution samples.

4.3.1. Monte Carlo Simulation Steps for Maximum Cumulative Residual Entropy

We conduct a simulation to compare the Chow-Liu method with the CRE method. Simulation steps are shown in Figure-4.1.

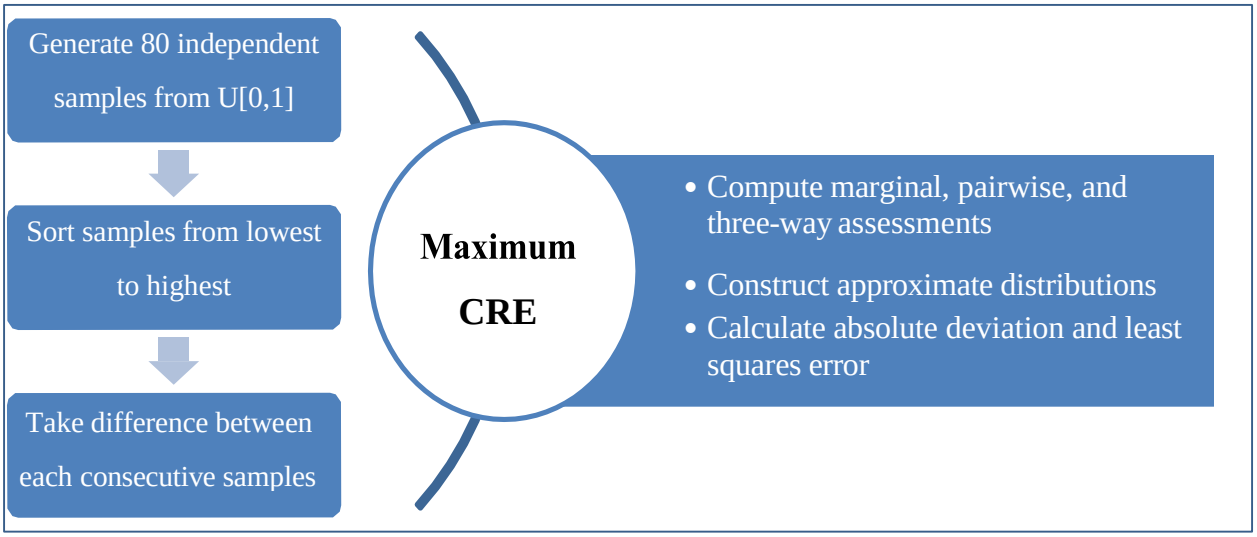


Figure 4.1: Monte Carlo Simulation Steps for Maximum Entropy and Maximum CRE Approximations

We apply a Monte Carlo simulation to maximum entropy and maximum CRE using lower assessments and then compare them to discuss their performances. We generated one million

discrete joint probability distribution samples. After that, we calculate the absolute deviation and least square errors between the true probability distribution and approximations using lower order assessments.

4.3.2. Compute the Marginal, Two-way, and Three-way Assessments of Joint Distributions

If no information is available, then maximum CRE and maximum entropy assign equal probability to each outcome. If only marginal distributions are available as a lower order assessment, then maximum entropy and maximum CRE assume probability independence among variables. Our aim is to approximate more accurate joint probability distributions, so we need to incorporate dependencies among variables into our analysis. We first of all, we compute the lower order assessments; (i) marginal distributions, (ii) pairwise distributions, and (iii) three-way distributions. Figure-4.2 shows the assessment of marginal distributions for each of the four variables. The variables take 3 different values, "1", "2", and "3". Figure-4.3 shows an example of the assessments of pairwise joint distribution for variables X_1 and X_2.

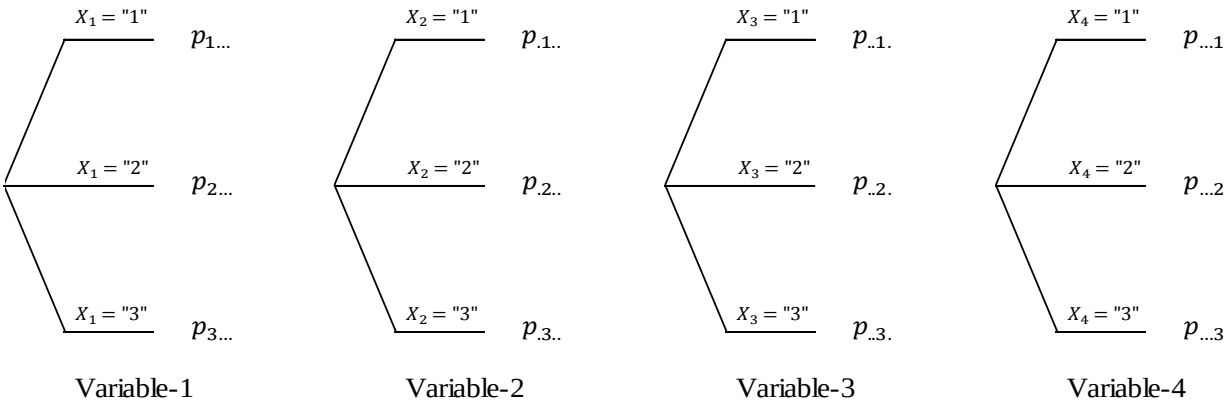


Figure 4.2: Probability Trees of Marginal Distributions using in Monte Carlo Simulation

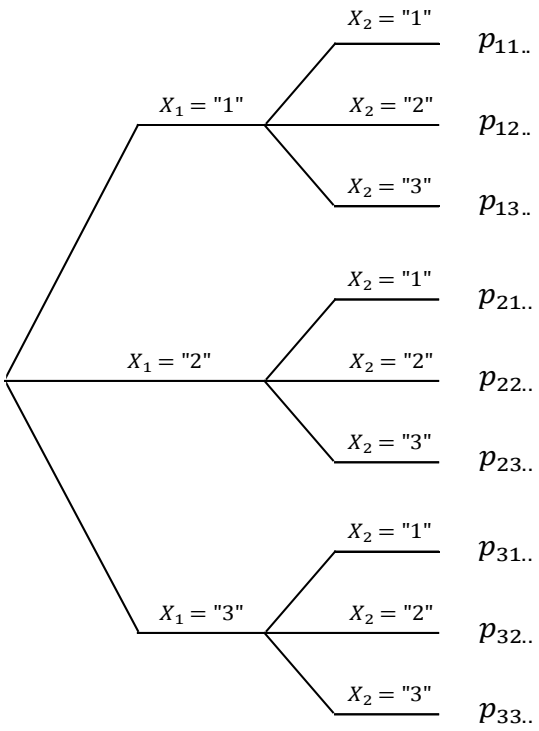


Figure 4.3: Elicitation of Pairwise Joint Distributions for variables X_1 and X_2

4.3.3. Results of Monte Carlo Simulation

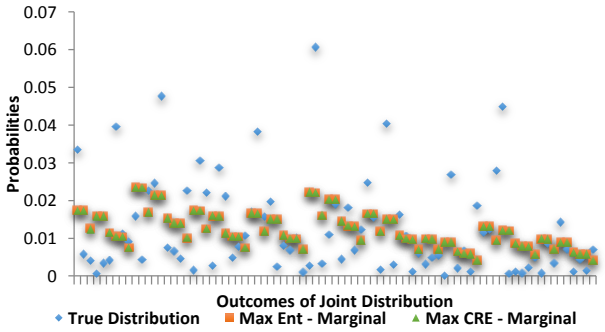
Table 4.5 displays a summary of mean and variance of errors of approximate joint probability distributions calculated by maximum entropy and maximum cumulative residual entropy methods.

The absolute deviation and least squares errors for maximum entropy and maximum CRE is almost same in the long run. But, we can say our proposed approach is a little bit better than the maximum entropy approach. For instance, we see that the mean of the absolute deviation of the pairwise assessments using maximum entropy is 0.52514, on the other hand the mean using maximum CRE is 0.52509. There is a small difference between the errors of two methods but this difference should be negligible.

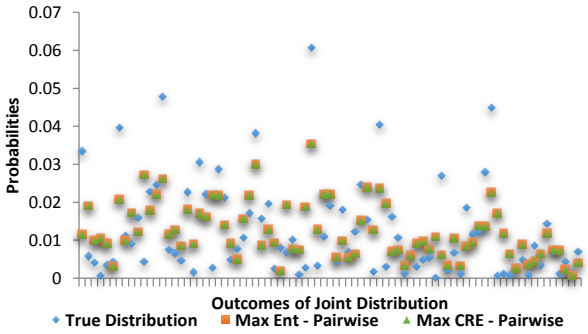
Table 4.5: Simulation Results for Max-Ent and Max CRE Approximate Distributions

Assessments	Absolute Deviation		Least Squares Error	
	Maximum Entropy	Maximum CRE	Maximum Entropy	Maximum CRE
Marginal (Independent)	0.69088	**0.69088**	0.01074	**0.01074**
Pairwise	0.52467	**0.52454**	0.0059854	**0.005985**
Three-way	0.21233	**0.21233**	0.0009298	**0.0009297**

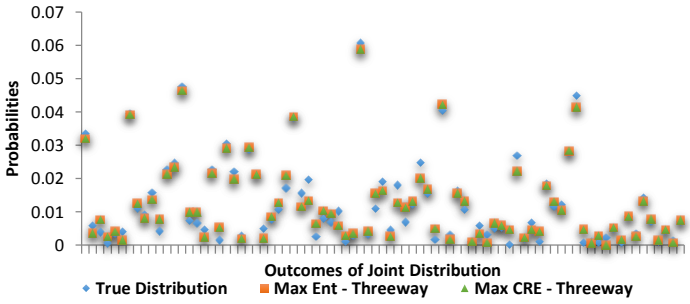
Another observation from the Table-4.5 is that when the dependences among variables are incorporated into the problem, then we can provide more representative joint probability distribution approximations. The value of absolute deviation and least squares error decrease significantly if the decision maker provides higher lower-order assessments in a decision situation.



a) Marginal Assessment



b) Pairwise Assessment



c) Three-way Assessment

Figure 4.4: Comparison of Approximate Joint Distributions Calculated by Each Lower Order Assessment

We also plot the approximate distributions of each lower assessment and true distribution to compare them visually. Figure-4.4 shows the plots of approximate joint distributions given each lower assessment and the true distribution. Each plot shows the true distribution, maximum entropy and maximum CRE approximate distributions given different assessments. Also, we can see from Figure-4.4, three-way assessment contains most of the information needed to approximate the true distribution. The outcomes of approximate distribution using three-way assessments are very close the outcomes of the true distribution. Finally, CRE-based maximum entropy provides at least as good approximation for a discrete multivariate probability distribution as that provided by the maximum entropy.

4.4. Comparison of Maximum Entropy and Maximum Cumulative Residual Entropy

In this section, we analyze when maximum CRE approximation is different from the traditional maximum entropy formulation and explore the situations where maximum CRE method gives more accurate results. To compare and discuss the performances of these two methods, we applied Monte Carlo simulation to generate different discrete multivariate probability distributions. We analyze the accuracy of approximations for three and four variate distributions with increasing number of outcomes of each variable. We start with a three-variable case and perform the same procedure for four random variable case with two, three, four, and five outcomes of each variable. We first analyze how much percent of the approximation of the sampling distributions are identical and then analyze the percent of the approximation of the sampling distributions where maximum CRE gives a better accuracy of approximation.

We further analyze how accuracy of the approximate distributions changes as the number of random variables changes. In this case, we hold the number of outcomes of variables constant and increase the number of variables from three to five. We finally analyze the relationship between

accuracy of the approximations and correlation among variables. In our analysis, we use total correlation measure as a measure of dependency among a set of random variables. We do our analysis for three and four variate probability distributions with two, three, and four outcomes of each variable. There are three different accuracy measures used in this analysis; absolute deviation (AD), least squares (LS), and maximum deviation (MD).

4.4.1. Effects of Increasing Number of Outcomes of Each Variable: Three Variable Case

Our first analysis is for three random variable case whee the number of outcomes of each variable is increased from two to five. We first analyze what percentage of the approximation of the sampling distributions are the same. Then, if these two methods approximate different joint probability distributions, we figure out for what percentage of sampling distributions that maximum CRE gives better accuracy of approximation and for what percentage of sampling distributions that maximum entropy method gives better accuracy of approximation. Figure 4.5 shows the results for three variable case using absolute deviation as the accuracy measure.

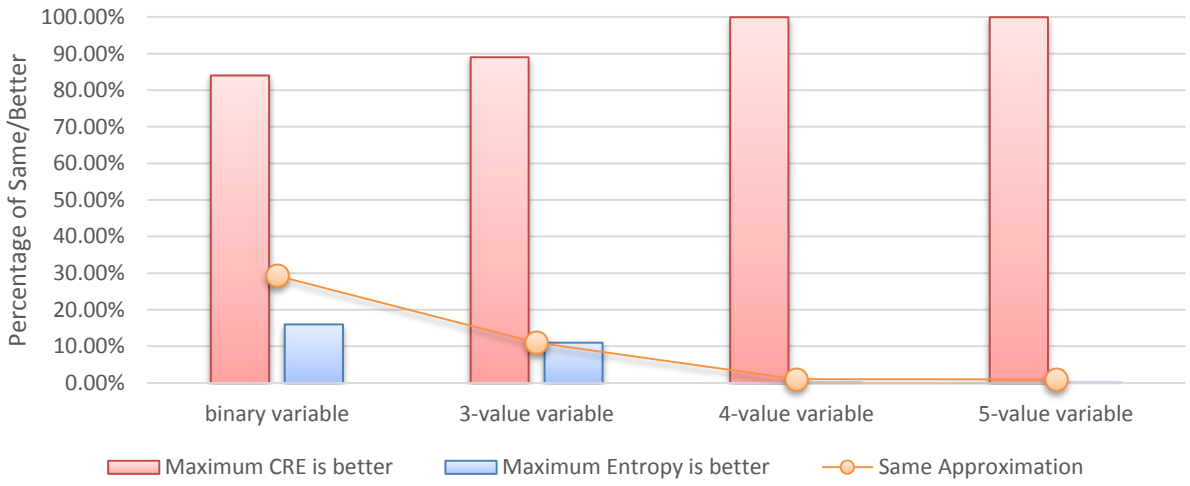
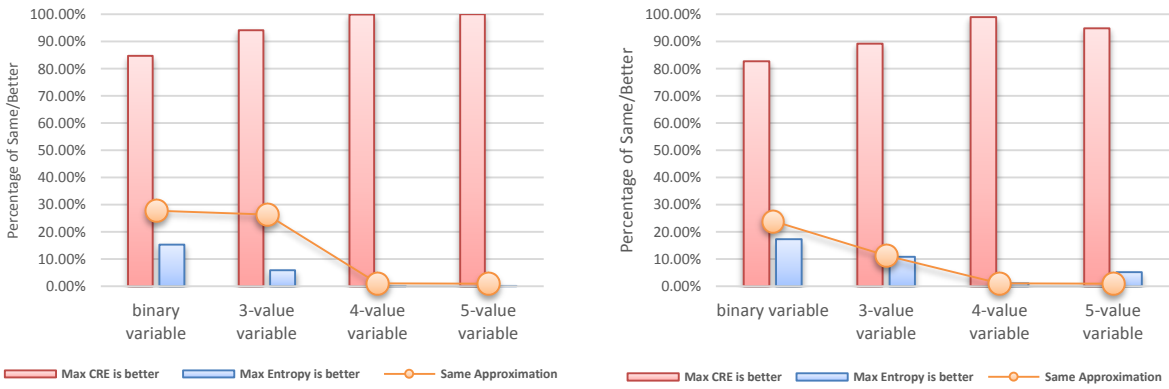


Figure 4.5: Comparison of Maximum Entropy and Maximum CRE Approximation Methods based on Absolute Deviation for Three Variable Case

From Figure 4.5, we can say that the maximum CRE approximation method gives significantly more accurate results than the traditional maximum entropy method, and the difference between

these two approximation methods increases as the number of outcomes of each variable increases. These two approximation methods sometimes approximate very similar joint probability distributions, but they are not same. To explore what percentage of both methods approximate the same joint probability distribution, we assume that if the difference between absolute deviation of maximum entropy and absolute deviation of maximum CRE method is less than a value ε, then these two approximation are same. We checked the difference as $\left|AD(\max CRE) - AD(\max entropy)\right| \le \varepsilon$. In the three-variable case, the value of ε is assumed as 0.0001. The result shows that when the variables are binary, almost 30% of the approximations are same. However, when the number of outcomes of each variable increases, the percentage of same approximation noticeably decreases to 0.01%. Moreover, as the number of outcomes of variables increases, the percentage of sampling distributions that maximum CRE gives better accuracy of approximation and the percentage of sampling distributions that maximum entropy method gives better accuracy of approximation is getting closer.



(a) Least Squares **(b)** Maximum Deviation

Figure 4.6: Comparison of Maximum Entropy and Maximum CRE Approximation Methods based on Least Squares and Maximum Deviation Measures for Three Variable Case

Another observation from Figure-4.5 is that the maximum CRE method is better than the traditional maximum entropy method and gives much higher accuracy results in the case of three-

variate joint probability distributions with any number of outcomes. However, in the case of four variables, the maximum entropy and maximum CRE methods give similar accuracy of approximation. We also run the same simulation for least squares measure and maximum deviation measure. Figure-4.6 shows the results of three variable case (Figure 4.6-a: Least Squares, Figure 4.6-b: Maximum Deviation).

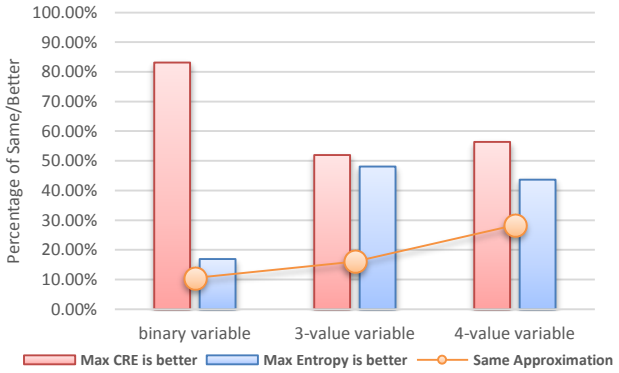
Figure 4.6 shows behavior similar to that shown in Figure 4.5. So, in general we may say that the maximum CRE approximation method obtained more accurate approximate distributions than what was observed by maximum entropy approximation method in all three different accuracy measures.

4.4.2. Effects of Increasing Number of Outcomes of Each Variable: Four Variable Case

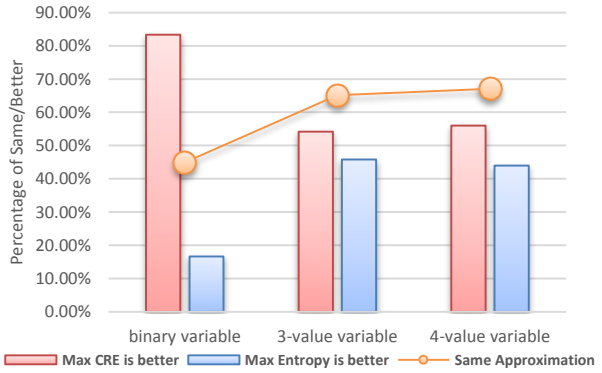
We now perform the steps involved in the three random variable case with two, three, four, and five outcomes. In a three-variable case we do only the analysis for pairwise assessments. In this case, however, we perform the analysis with two different lower order assessments: pairwise assessment and three-variate assessment. Figure 4.7 shows the results of a four variable case using pairwise assessments (Figure 4.7-a-c-e) and three-way assessments (Figure 4.7-b-d-f) when absolute deviation (Figure 4.7-a-d), least squares (Figure 4.7-b-e), and maximum deviation (Figure 4.7-c-f) are used as the accuracy measure.

The results for the four-variable case are different from those for the three variable case. In this case, the results show that the percentage of approximating the same joint probability distribution increases with the number of outcomes of variables for absolute deviation, least squares, and maximum deviation measures for both pairwise and three-way assessments. For the binary variable case, maximum CRE gives more accurate accuracy results; however, maximum CRE and maximum entropy approximate very similar joint probability distributions as the number of

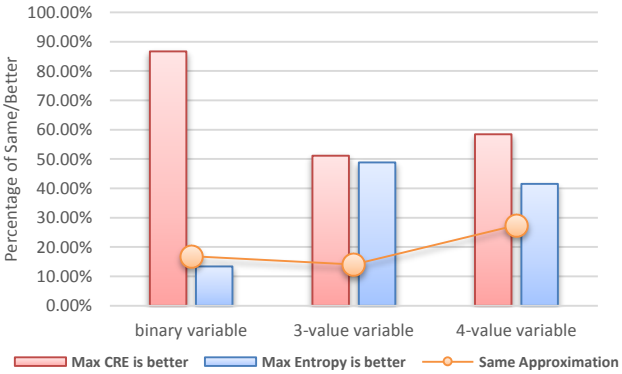
outcomes of variables increases. For a three-way assessment case, maximum CRE gives better accuracy results for binary, three-outcome, and four-outcome variable cases.



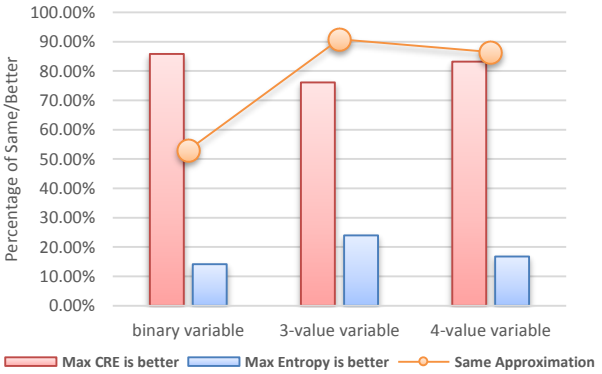
(a) Absolute Deviation of Four Variable Case using Pairwise Assessment



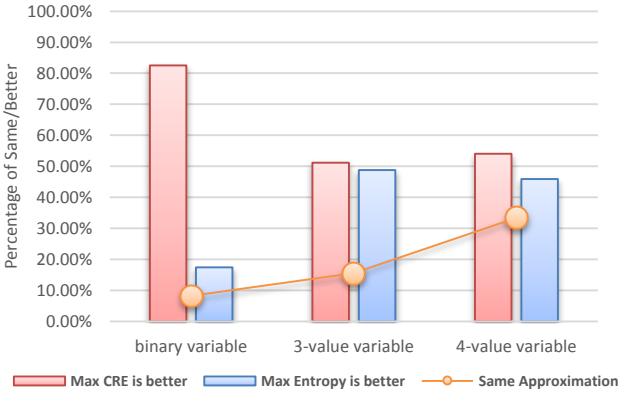
(b) Absolute Deviation of Four Variable Case using Three-way Assessment



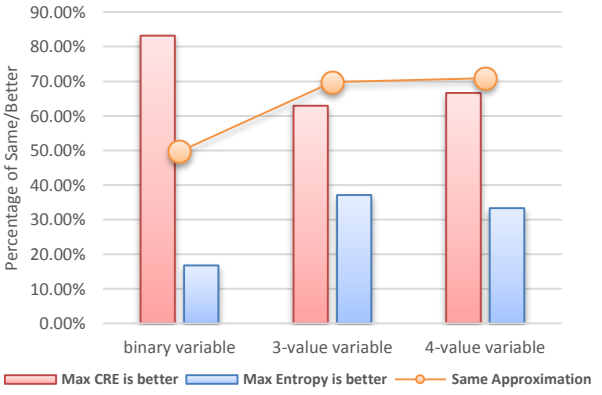
(c) Least Squares of Four Variable Case using Pairwise Assessment



(d) Least Squares of Four Variable Case using Three-way Assessment



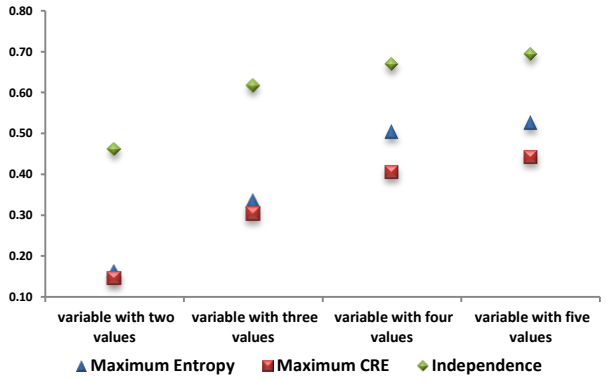
(e) Maximum Deviation of Four Variable Case using Pairwise Assessment



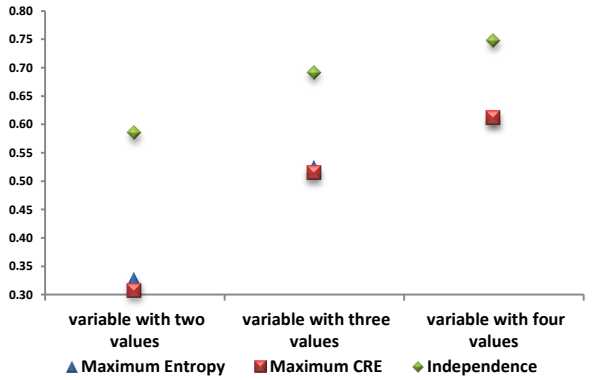
(f) Maximum Deviation of Four Variable Case using Three-way Assessment

Figure 4.7: Comparison of Maximum Entropy and Maximum CRE Approximation Methods based on Least Squares and Maximum Deviation Measures for Four Variable Case

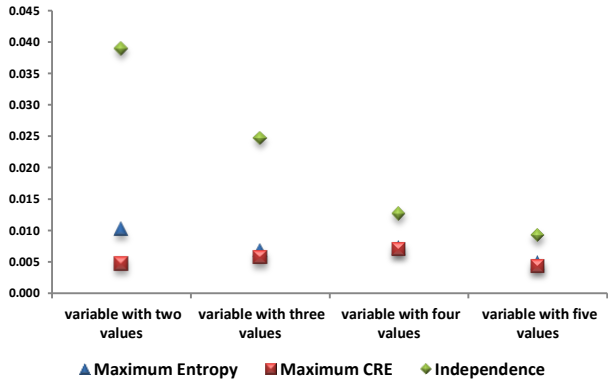
We also calculate the mean accuracy (error) outcomes of approximations for each measure for three- and four-variable cases with an increasing number of outcomes for each variable and for an independence case. Figure-4.8 shows the results of mean accuracy values.



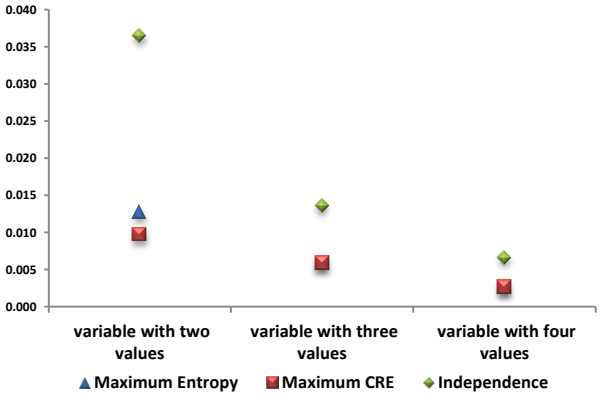
(a) Absolute Deviation of Three Variable Case



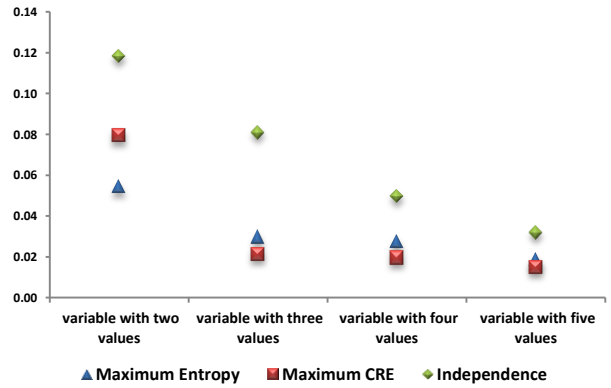
(b) Absolute Deviation of Four Variable Case



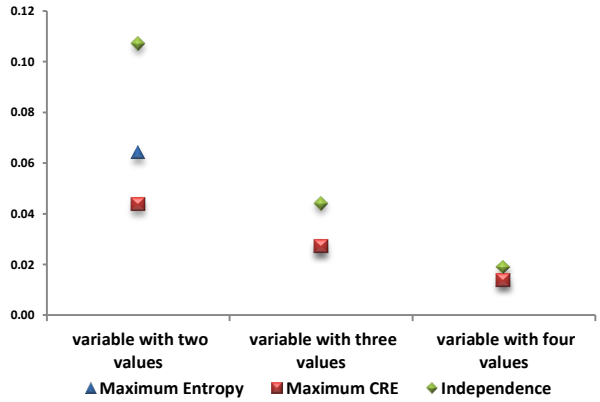
(c) Least Squares of Three Variable Case



(d) Least Squares of Four Variable Case



(e) Maximum Deviation of Three Variable Case



(f) Maximum Deviation of Four Variable Case

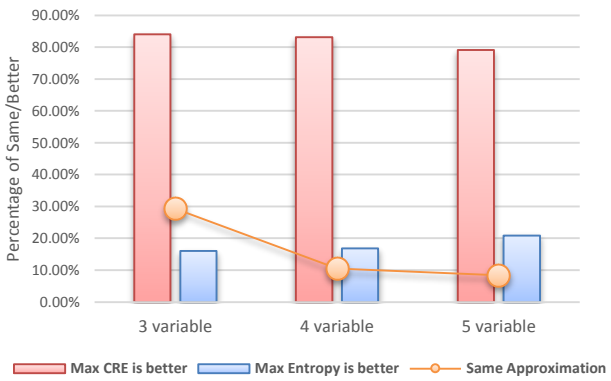
Figure 4.8: The Mean Accuracy (Error) Values of Each Measure for Three and Four Variable Cases with Increasing Number of Outcomes of Each Variable

The results show that although the percentage results of four variate case differ, the average accuracy values for each method is very close. Also, the error values increases with the number of outcomes of variables for absolute deviation measure but decreases for least squares and maximum deviation. For higher number of outcomes of variables, the maximum CRE method is slightly better than the maximum entropy method. Also, the maximum CRE approximation method outperforms the maximum entropy method (i) in three variable case for both binary and three-outcome cases and (ii) when the variables are binary. These results are consistent with the previous results.

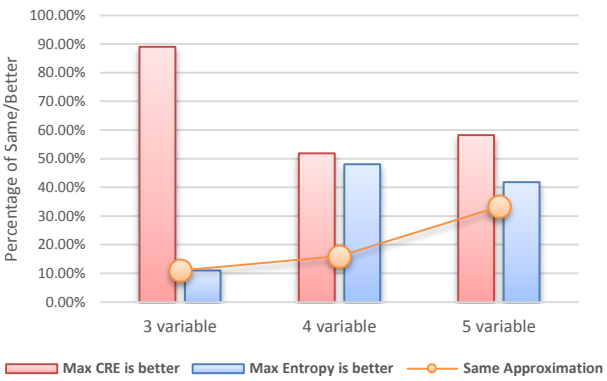
4.4.3. Effects of Increasing the Number of Variables in Multivariate Distributions

We now analyze how accuracy of the approximate distributions changes as the number of random variables increases in joint probability distributions. We first analyze what percentage of the approximation of the sampling distributions is identical. If they differ, we analyze for what percentage of sampling distributions maximum CRE approximates more accurately and for what percentage of sampling distributions maximum entropy approximates more accurately. In our analysis for binary and three-outcome variables we increase the number of variables from three to five. Figure-4.9 shows the results for binary case (Figure 4.9-a-c-e) and three-value case (Figure 4.9-b-d-f) for each error measure.

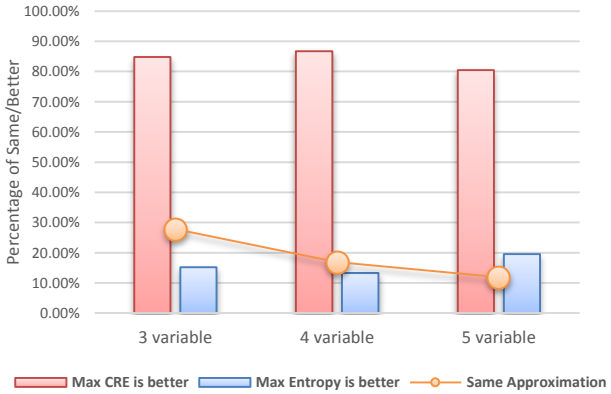
We found that the maximum entropy and maximum CRE methods approximated the same first order dependence tree for less than 40% of the samples in the three- and four-variable cases. Moreover, the percentage of sample distributions for which these two methods are same decreased for the binary variable case and increased for the three-outcome case as the number of variables increased.



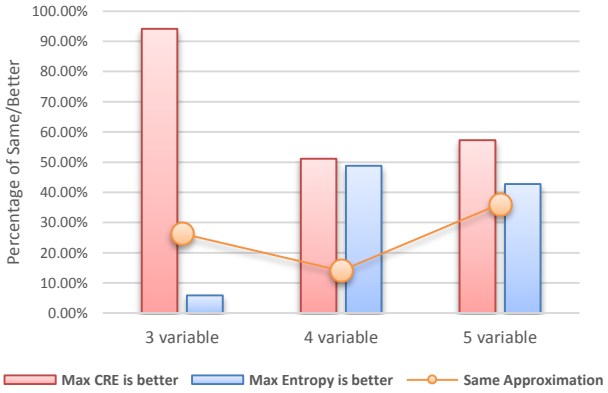
(a) Absolute Deviation of Binary Variable Case



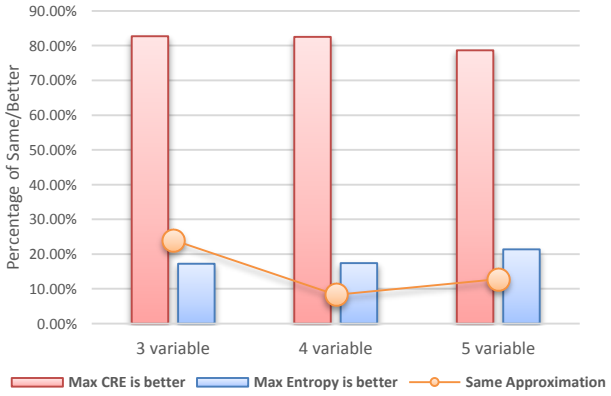
(b) Absolute Deviation of Three-outcome Variable Case



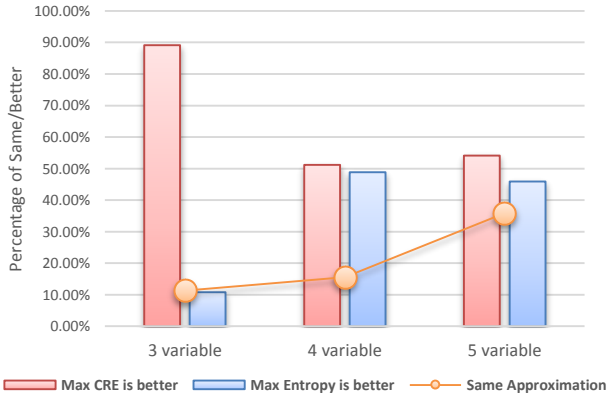
(c) Least Squares of Binary Variable Case



(d) Least Squares of Three-outcome Variable Case



(e) Maximum Deviation of Binary Variable Case



(f) Maximum Deviation of Three-outcome Variable Case

Figure 4.9: The Percent of the Approximation of the Sampling Distributions That Maximum CRE and Maximum Entropy Methods are Same and Different

If we look at the results of these two approaches in Figure 4.9, the results are very similar to those obtained with increasing number of outcomes of variables. When the variables have three

outcomes, these two methods show very similar behavior, but the maximum CRE method is better than the maximum entropy methods where the variables are binary. We can see that maximum CRE yields more than 80% more accurate approximated joint distribution for the three binary variable case (Figure 4.9-a-c-e)and more than 90% for 3 three-outcome variable case (Figure 4.9-b-d-f). However, the results for the four three-outcome variable and five three-outcome variable case are different from the results obtained from the other joint probability distributions. In the case of four and five three-outcome variables, the maximum CRE and maximum entropy approximation methods are very similar and approximate almost identical joint probability distributions.

So, we can say that the maximum CRE method outperforms the maximum entropy method when (i) the number of variables are three, or (ii) the variables are binary. Moreover, these results indicate that the effect of increasing the number of outcomes of each variable is similar to the effect of increasing the number of random variables.

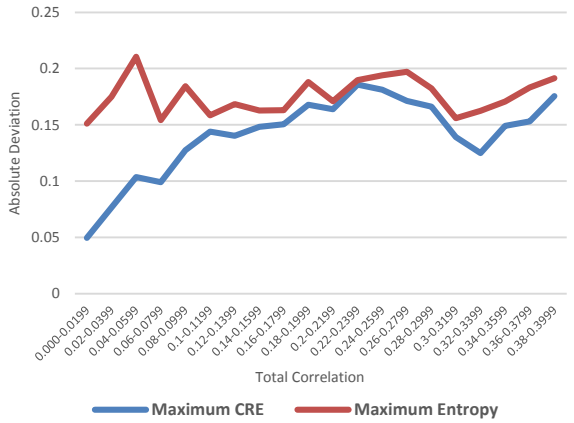
4.4.4. Effects of Changing the Dependence Structure Between Variables

We now analyze the effects of changing dependence structure among variables. We use total correlation (TC) as a dependency measure. We discussed the total correlation in detail in Chapter 3. The total correlation is defined as

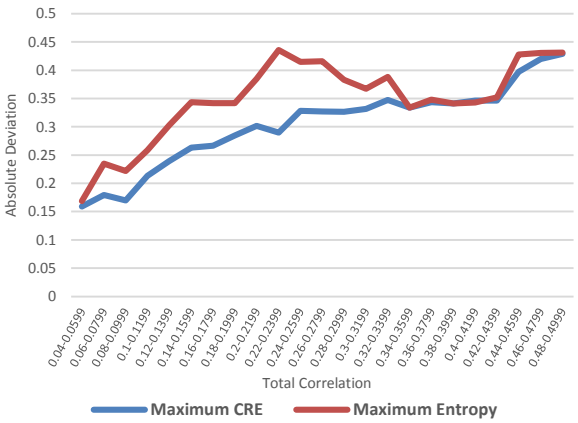
$$TC(X_1, X_2, \cdots, X_n) = \sum_{x_1 \in X_1} \sum_{x_2 \in X_2} \cdots \sum_{x_n \in X_n} p(x_1, x_2, \cdots, x_n) \log \frac{p(x_1, x_2, \cdots, x_n)}{p(x_1) \times p(x_2) \times \cdots \times p(x_n)} \qquad (4.7)$$

where $p(x_1, x_2, \cdots, x_n)$ is the probability mass function of variables $X_1, X_2, \cdots, X_n$, and $p(x_1)$ is the marginal probability distribution of variable X_1.

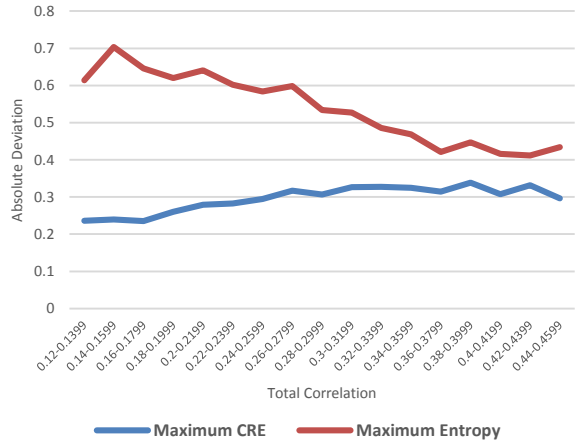
We run a simulation with several different combination of joint distributions including; three binary variables, three three-outcome variables, three four-outcome variables, four binaryvariables



(a) Total Correlation vs. Absolute Deviation for Three Binary Variable

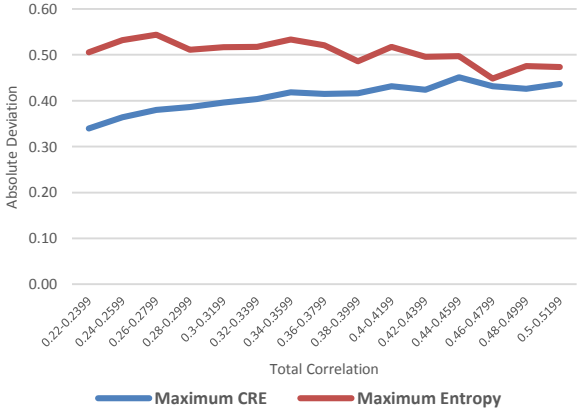


(b) Total Correlation vs. Absolute Deviation for Four Binary Variable

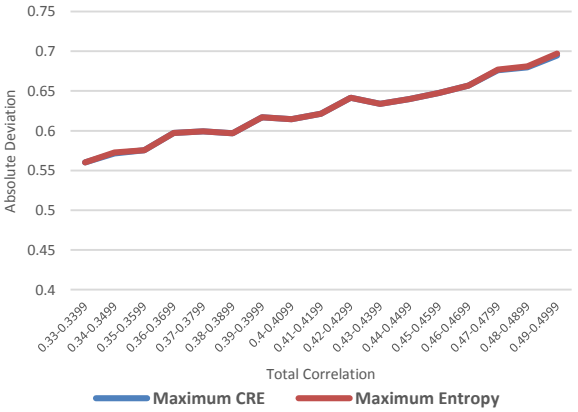


(c) Total Correlation vs. Absolute Deviation for Three Three-outcome Variable

(d) Total Correlation vs. Absolute Deviation for Four Three-outcome Variable

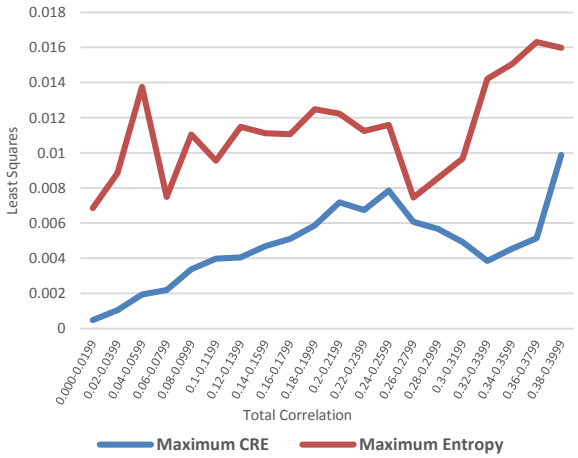


(e) Total Correlation vs. Absolute Deviation for Three Four-outcome Variable

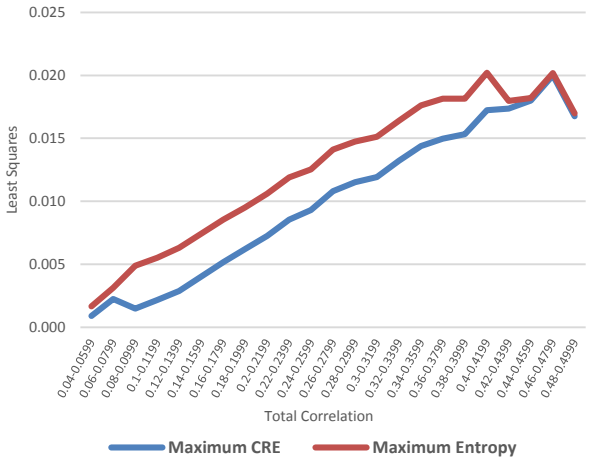


(d) Total Correlation vs. Absolute Deviation for Four Four-outcome Variable

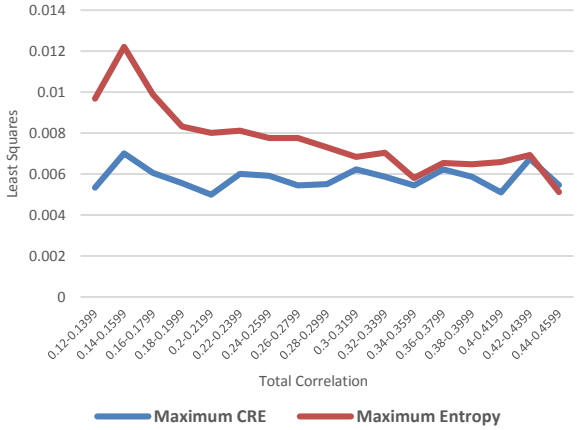
Figure 4.10: Average Values of Errors of Joint Probability Distributions Calculated by Traditional Maximum Entropy and Maximum Cumulative Residual Entropy Methods Using Absolute Deviation as Accuracy Measure



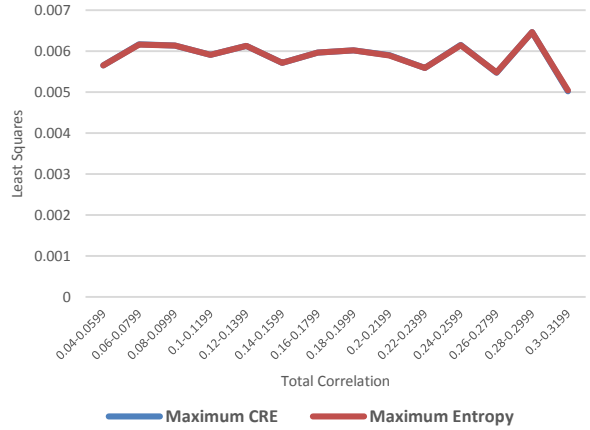
(a) Total Correlation vs. Least Squares for Three Binary Variable



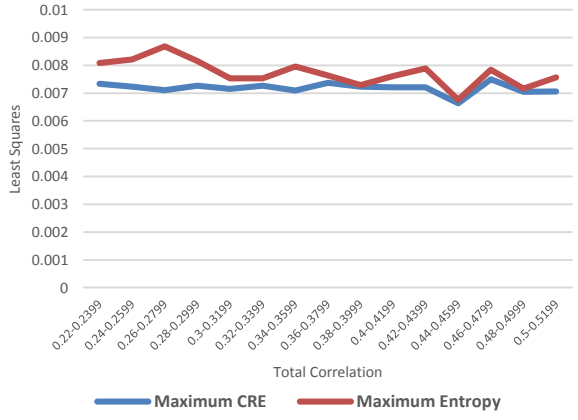
(b) Total Correlation vs. Least Squares for Four Binary Variable



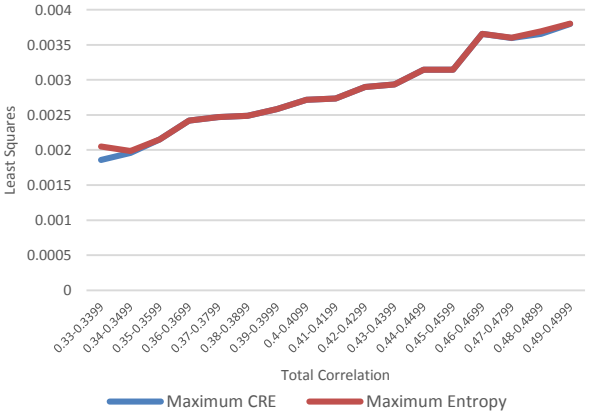
(c) Total Correlation vs. Least Squares for Three Three-outcome Variable



(d) Total Correlation vs. Least Squares for Four Three-outcome Variable

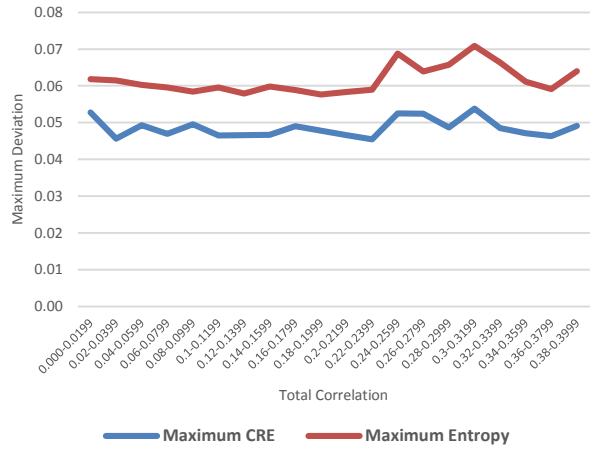


(d) Total Correlation vs. Least Squares for Three Four-outcome Variable

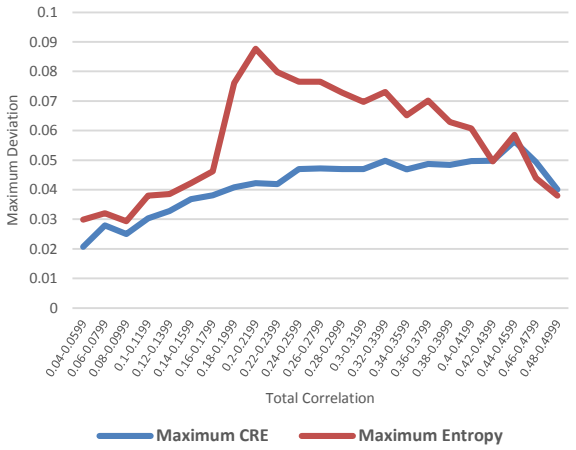


(f) Total Correlation vs. Least Squares for Four Four-outcome Variable

Figure 4.11: Average Values of Errors of Joint Probability Distributions Calculated by Traditional Maximum Entropy and Maximum Cumulative Residual Entropy Methods Using Least Squares as Accuracy Measure

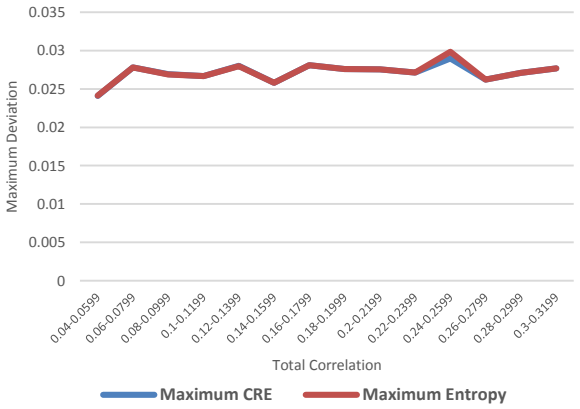


(a) Total Correlation vs. Maximum Deviation for Three Binary Variable

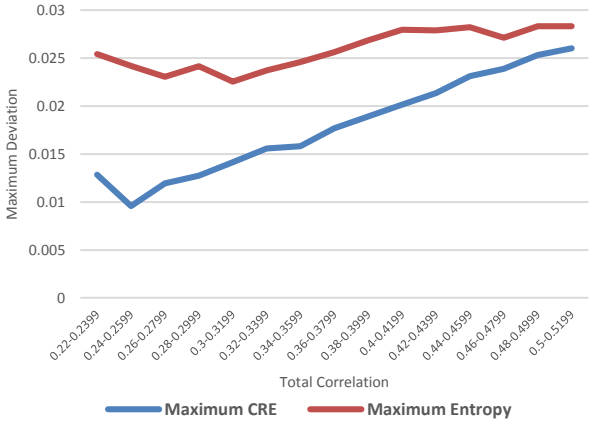


(b) Total Correlation vs. Maximum Deviation for Four Binary Variable

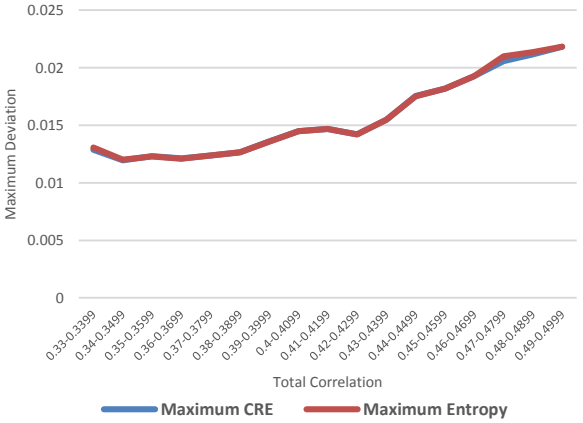
(c) Total Correlation vs. Maximum Deviation for Three Three-outcome Variable



(d) Total Correlation vs. Maximum Deviation for Four Three-outcome Variable



(e) Total Correlation vs. Maximum Deviation for Three Four-outcome Variable



(f) Total Correlation vs. Maximum Deviation for Four Four-outcome Variable

Figure 4.12: Average Values of Errors of Joint Probability Distributions Calculated by Traditional Maximum Entropy and Maximum Cumulative Residual Entropy Methods Using Maximum Deviation as Accuracy Measure

four three-outcome variables, and four four-outcome variables. Figure 4.10, 4.11, and 4.12 display the mean values of errors of approximated joint probability distributions calculated by Maximum entropy and maximum CRE methods for absolute deviation, least squares, and maximum deviation measures, respectively. For the three-variable case, maximum CRE method demonstrates better performance at lower total correlation. As the total correlation between variables increases, maximum CRE and maximum entropy methods give closer accuracy results.

Overall the results show that the maximum CRE approximation method shows performance comparable or better than maximum entropy approximation method for all accuracy measures. For instance, for three three-outcome variable, when the average total correlation is between 0.14 and 0.1599, the mean absolute deviation of maximum entropy approximation method is ~0.7035 and the mean absolute deviation of maximum CRE approximation method is ~0.2398 .On the other hand, when the average total correlation is between 0.42 and 0.4399, the absolute deviation of maximum CRE is ~0.3317, and the absolute deviation of maximum entropy is ~0.4113.

We also observe from Figure 4.10, 4.11, and 4.12 that the difference between the two methods is considerable when the total correlation is low. However, when the total correlation between variables increases, the difference between the two methods noticeably decreases. So, these two methods are sensitive to correlation among variables. We can also say that the dimension of the probability distribution has an effect on the accuracy of the approximations because maximum CRE and maximum entropy methods show almost identical accuracy results at four three-outcome and four four-outcome variable cases, but in other cases, there is a significant difference between maximum entropy and maximum CRE methods.

4.5. Summary of the Results

- Maximum CRE approach can only be applied to numeric variables, but the traditional maximum entropy method can be applied both numeric and non-numeric variables.
- There is a significant difference between maximum CRE and maximum entropy methods in the case of lower correlation, lower number of variables, and lower number of outcomes of variables.
- The accuracy of approximations decreases as the total correlation and the dimension of the multivariate probability distribution increases
- For binary variable cases and lower total correlation, maximum CRE approximation method gives much better approximation.
- Maximum CRE and traditional maximum entropy methods get closer as the number of outcomes of variables, total correlation, and number of variables increases.

4.6. Conclusion

Eliciting the decision maker's preferences with regard to a decision situation is one of the main steps in decision analysis. In this chapter, we are interested in using a method similar to the maximum entropy approach to elicit multivariate probability functions using lower-order assessments. Such a model can capture dependence into the model and approximate less error-prone probability functions. The main advantage of the proposed approach is to use cumulative functions in entropy formulation instead of density functions.

We simulate four-variate joint probability distribution using a Monte Carlo simulation to illustrate its performance and show the applicability of the proposed approach using each lower order assessment. We can say that the maximum cumulative residual entropy approach is at least as good as the maximum entropy approach.

This chapter was also intended to familiarize the reader with the maximum cumulative residual entropy approach and its ability to incorporate many types of decision situations with partial information. Maximum CRE applications would help readers to approximate smoother and better probability distributions by using cumulative distributions. In this chapter, we compare the maximum entropy and maximum cumulative residual entropy methods to explore the situations where maximum CRE method gives more accurate results and where these two methods are different.

CHAPTER 5:

COMPARISON OF INFORMATION THEORY BASED APPROXIMATION METHODS

5.1. Introduction

In this chapter, we compare several approximation methods to find the best approximate probability distribution based on the given information. The objective of our research is to test the accuracy of different approximations of joint distributions with respect to the true distribution from the set of all possible distributions that match available information. There are a number of methods presented in the literature for joint probability distribution approximations. We specifically compare approximation methods that use information theory to approximate multivariate probability distributions in the literature. The approximation methods that are used in our analysis are listed in Table 5.1.

Table 5.1: Approximation methods proposed in the literature

Approximation Method	Author(s), Year
Chow-Liu first order dependence tree approximation	Chow&Liu, 1968
CRE based first order dependence tree approximation	Sutcu&Abbas, 2014
Maximum entropy approximation	Abbas, 2006
Maximum CRE approximation	Sutcu&Abbas, 2014
Lewis product approximation	Lewis, 1959
Brown's maximum entropy approximation	Brown, 1959
Ku and Kullback's lower-order marginal distributions approximation	Ku&Kullback, 1969
Keefer's binary event approximation	Keefer,2004
Kirkwood Superposition approximation	Kirkwood, 1935
Independence approximation	

Our aim is to analyze the performance of our proposed approximation methods. The traditional approach in the multivariate probability distribution approximation concentrates on probability mass/density function, whereas the proposed methods use cumulative distributions, especially survival functions, instead of probability mass/density functions. To the best of our knowledge, however, this extension has never been examined in the literature. The primary innovation of our approach is to integrate survival functions into the maximum entropy, KL divergence measure, and mutual information formulations.

The remainder of this chapter is organized as follows. We first discuss previous approximation methods based on information theory. Then, we apply our proposed methods and the previous methods to the two different cases to test the accuracy of the approximation methods: (i) four binary variables and (ii) four-variable decision problems that take three different values each. We then analyze the effects of increasing both the number of variables and the number of alternatives or outcomes and discuss our results.

5.2. Information Theory based Approximation Methods

Much of the recent literature has used information theory for joint probability approximations. We already discussed the Chow-Liu first-order dependence tree and the maximum entropy method in Chapter 3. We also discussed in detail our proposed approaches, CRE-based first-order dependence tree and maximum CRE, in Chapter 3 and 4, respectively. We now describe briefly previous studies listed in Table 5.1.

5.2.1. Lewis' Product Approximation (1959)

Lewis (1959) considered the problem of approximating a multivariate distribution by a product of several lower order distributions. Lewis showed that the product approximation has the

minimum information based on Kullback-Liebler divergence and maximum entropy. Several product approximations to a three variate probability distribution $P(x_1, x_2, x_3)$ can be listed as

$$\begin{array}{ll} 1.\ P(x_1)P(x_2)P(x_3) & 3.\ P(x_1, x_2)P(x_3 \mid x_1) \\ 2.\ P(x_1, x_2)P(x_3) & 4.\ P(x_1, x_3)P(x_2 \mid x_3) \end{array} \tag{5.1}$$

However, Lewis's method does not compose the best approximation based on the partial information. This method selects all the product approximations under certain conditions and from those further selects the best approximation by calculating the entropies of all lower order assessments under certain conditions.

5.2.2. Brown's maximum entropy approximation (1959)

After Lewis' product approximation, Brown (1959) presented an iterative method to approximate joint probability distributions using any subset of the variables or any lower distributions. This approximation method iterates each step and converges to give an approximation that has the minimum information. Brown's approximation method is more general than the product approximation and can be applied to any set of component distributions. In this method, the iteration starts with a product approximation and stops when convergence is attained. For instance for a three-variate probability distribution $P(x_1, x_2, x_3)$, one of the iteration steps would be

$$\begin{array}{ll} p^1 = p(x_3 \mid x_2)\, p(x_1, x_2) & \text{a product approximation (first step)} \\ p^2 = p^1(x_1, x_2, x_3)\dfrac{p(x_1, x_3)}{p^1(x_1, x_3)} & \text{(second step)} \\ p^3 = p^2(x_1, x_2, x_3)\dfrac{p(x_2, x_3)}{p^2(x_2, x_3)} & \text{(third step)} \end{array} \tag{5.2}$$

This method can be applied to any set of lower order assessments. Also, the approximation give more accurate results if more and higher order distributions are used. However, the iteration procedure cannot compose the best approximation because numerous iterations are required to attain full convergence which is very difficult when the number of variables increase.

5.2.3. Ku and Kullback's lower-order marginal distributions approximation (1969)

Ku and Kullback obtained an approximation distribution by a convergent iterative algorithm in terms of lower order marginal distributions. Ku and Kullback extended Brown's approximation by using the lower order marginals. In this method, lower order assessments or marginal restraints are used as constraints and the convergent iterative procedure satisfy these marginal restraints. For instance, a three variate probability distribution $p(x_1, x_2, x_3)$ can be approximated with the subset of second order marginals $p(x_1, x_2), p(x_1, x_3), p(x_2, x_3)$ as

$$p^*(x_1, x_2, x_3) = a(x_1, x_2)b(x_1, x_3)c(x_2, x_3)\pi(x_1, x_2, x_3) \tag{5.3}$$

where $a(x_1, x_2), b(x_1, x_3), c(x_2, x_3)$ are determined to satisfy the marginal restraints.

5.2.4. Keefer's binary event approximation (2004)

Keefer presented a method to approximate probabilistic dependence between variables which is called as "underlying event (UE)". This method requires the assessment of only one conditional probability in addition to the marginal probabilities. However, Keefer's method can be applied only to the binary events in decisions. First of all, the method assess the marginal probabilities and then choose the largest and second largest values to assess the conditional probabilities. For instance, for a n project model, the method for calculating the conditional probabilities in terms of assessed probabilities as

$$p_{i|0} = \frac{p_i}{p_0} = \frac{p_i}{\left(\frac{p_j}{p_{j|k}}\right)} = p_{j|k}\left(\frac{p_i}{p_j}\right) \qquad i = 1, 2, ..., \mathrm{n} \tag{5.4}$$

where p_i is the value of one of the event's probability. Here, p_j is the largest value and defined as

$$p_j = \max\{p_i,\ i = 1, 2, ..., n\} \tag{5.5}$$

and p_k is the second largest value and defined as

$$p_k = \max\{p_i,\ i \neq j,\ i = 1, 2, ..., n\} \tag{5.6}$$

5.2.5. Kirkwood Superposition approximation (1935)

The superposition approximation was introduced by John Kirkwood in 1935 to approximate multivariate discrete probability distributions. The method generally works by using all the product of probabilities over all subset of variables. The Kirkwood approximation for a discrete probability distribution

$$p(x_1, x_2, \cdots, x_n) = \frac{\dfrac{\prod\limits_{\tau_{n-1} \subseteq V} p(\tau_{n-1})}{\prod\limits_{\tau_{n-2} \subseteq V} p(\tau_{n-2})}}{\prod\limits_{\tau_1 \subseteq V} p(\tau_1)} \tag{5.7}$$

where $\prod\limits_{\tau_i \subseteq V} p(\tau_i)$ is the product of probabilities over all subsets of variables of size i in variable set V. For the three-variable case, the approximation reduces to simply

$$p^*(x_1, x_2, x_3) = \frac{p(x_1, x_2)\, p(x_1, x_3)\, p(x_2, x_3)}{p(x_1)\, p(x_2)\, p(x_3)} \tag{5.8}$$

where the numerator contains the product of pairwise assessments of three-variate distribution and the denominator contains the product of marginal distributions. The main concern in the superposition method is that the Kirkwood approximation does not generally produce a valid probability distribution because superposition approximation method violates the normalization condition (sum of the probability is not equal to 1).

5.3. Approximating the Four Binary Variable Probability Distribution

We now provide an example of joint probability distribution of four binary variables and the approximations derived from this joint distribution using its lower order assessments. An example of the four binary distribution is given in Table 5.2.

Table 5.2: Four Binary Joint Distribution

	X1	X2	X3	X4	P(x_1,x_2,x_3,x_4)
1	1	1	1	1	0.0186
2	1	1	1	2	0.0478
3	1	1	2	1	0.0037
4	1	1	2	2	0.1949
5	1	2	1	1	0.0524
6	1	2	1	2	0.0527
7	1	2	2	1	0.0101
8	1	2	2	2	0.1062
9	2	1	1	1	0.0646
10	2	1	1	2	0.0179
11	2	1	2	1	0.0109
12	2	1	2	2	0.0266
13	2	2	1	1	0.0629
14	2	2	1	2	0.0761
15	2	2	2	1	0.2433
16	2	2	2	2	0.0114

In this example, the variables may take two different values: “1” or “2”. The first four column in Table 5.2 describe the different combinations of joint distribution. The fifth column contains the probabilities of the joint events generated by using the method described in Chapter 3. For

instance, the probability of the fourth combination of the variables where variable 1 and variable 2 take the values "1", and variable 3 and variable 4 take the value "2" is 0.1949. The true four binary distribution and the approximations generated using the information taken from the true distribution is given in Table-5.3.

Table 5.3: Approximations of Four Binary Distribution

X1	X2	X3	X4	P(x_1,x_2,x_3,x_4)	Chow-Liu	CRE FODT	MaxEnt (pairwise)	MaxEnt (three-way)	MaxEnt CRE (pairwise)	MaxEnt CRE (three-way)	Lewis*	Brown*	Ku& Kullback*	Keefer*	Kirkwood	Indepen-dence
1	1	1	1	0.0186	0.0063	0.0163	0.0110	0.0221	0.0110	0.0221	0.0101	0.0102	0.0104	0.0237	0.0238	0.0343
1	1	1	2	0.0478	0.0762	0.0771	0.0845	0.0443	0.0845	0.0443	0.0977	0.0988	0.0964	0.0697	0.0243	0.0392
1	1	2	1	0.0037	0.0115	0.0299	0.0175	0.0002	0.0175	0.0002	0.0185	0.0174	0.0168	0.0436	0.0009	0.0532
1	1	2	2	0.1949	0.1399	0.1416	0.1519	0.1984	0.1519	0.1984	0.1794	0.1695	0.1749	0.1279	0.2236	0.0605
1	2	1	1	0.0524	0.0236	0.0136	0.0205	0.0489	0.0205	0.0489	0.0198	0.0197	0.0202	0.0468	0.0454	0.0548
1	2	1	2	0.0527	0.0654	0.0645	0.0555	0.0562	0.0555	0.0562	0.0439	0.0436	0.0425	0.0313	0.0672	0.0629
1	2	2	1	0.0101	0.0434	0.0250	0.0358	0.0136	0.0358	0.0136	0.0364	0.0378	0.0364	0.0859	0.0087	0.0846
1	2	2	2	0.1062	0.1201	0.1184	0.1097	0.1027	0.1097	0.1027	0.0805	0.0837	0.0863	0.0574	0.0646	0.0968
2	1	1	1	0.0646	0.0345	0.0384	0.0316	0.0611	0.0316	0.0611	0.0188	0.0190	0.0195	0.0131	0.1120	0.0362
2	1	1	2	0.0179	0.0306	0.0133	0.0217	0.0214	0.0217	0.0214	0.0206	0.0208	0.0203	0.0386	0.0285	0.0414
2	1	2	1	0.0109	0.0455	0.0507	0.0376	0.0144	0.0376	0.0144	0.0248	0.0235	0.0226	0.0173	0.0097	0.0560
2	1	2	2	0.0266	0.0404	0.0175	0.0291	0.0231	0.0291	0.0231	0.0272	0.0257	0.0265	0.0509	0.0075	0.0640
2	2	1	1	0.0629	0.1301	0.1262	0.1354	0.0664	0.1354	0.0664	0.1458	0.1448	0.1484	0.1018	0.0539	0.0578
2	2	1	2	0.0761	0.0263	0.0436	0.0328	0.0726	0.0328	0.0726	0.0363	0.0361	0.0352	0.0680	0.0640	0.0663
2	2	2	1	0.2433	0.1716	0.1664	0.1770	0.2398	0.1770	0.2398	0.1923	0.1997	0.1922	0.1342	0.2468	0.0895
2	2	2	2	0.0114	0.0347	0.0575	0.0484	0.0149	0.0484	0.0149	0.0479	0.0497	0.0513	0.0897	0.0191	0.1024

* After 6 iterations

Column-6 of Table-5.3 shows the results of Chow-Liu first-order dependence tree, column-7 shows the results of CRE-based first-order dependence tree, column-8 shows the results of maximum entropy with pairwise assessments, column-9 shows the results of maximum entropy with three-way assessments, column-10 shows the results of maximum CRE with pairwise assessments, column-11 shows the results of maximum CRE with three-way assessments, from column-12 to column-16, the columns show the results of Lewis, Brown, Ku&Kullback, Keefer's binary, and Kirkwood's superposition approximation methods respectively. The last column (column-17) shows the results of the independence approximation. Also, Lewis, Brown, Ku&Kullback, and Keefer's binary event approximation are iterative methods, and the iterative process is stopped after six iterations.

We measure the accuracy of the each methods, especially the errors of each approximation, which makes the performance results more informative. We analyze and observe the performance of the approximations with respect to the original distribution for three different measures. We use absolute deviation (AD), least squares (LS), and maximum deviation (MD) measures in our analysis. For this specific four binary variable example, we calculate the results with respect to the three error measures. The results are shown in Table-5.4.

Table 5.4: Example of Approximations of Four Binary Variable Distribution

Approximation Method	Absolute Deviation(AD)	Least Squares (LS)	Maximum Deviation (MS)
Chow-Liu Dependence Tree	0.4955	0.0215	0.0769
CRE First Order Dependence Tree	0.4872	0.0219	0.0717
Maximum Entropy (Pairwise)	0.4502	0.0002	0.0725
Maximum Entropy (Three-way)	**0.0559**	**0.0199**	**0.0035**
Maximum CRE Entropy (Pairwise)	0.4502	0.0002	0.0725
Maximum CRE Entropy (Three-way)	**0.0559**	**0.0199**	**0.0035**
Lewis Product Approximation	0.4554	0.0202	0.0829
Brown's Approximation	0.4567	0.0199	0.0819
Ku&Kullback's Approximation	0.4553	0.0206	0.0855
Keefer Binary Approximation	0.5229	0.0282	0.0970
Kirkwood Approximation	**0.2353**	**0.0064**	**0.0474**
Independence Approximation	0.6987	0.0634	0.1538

In this example, we see that the maximum entropy approximation (MaxEnt-three way) and the maximum CRE approximation (MaxEnt CRE- three way) outperform the other methods; the third best performance is achieved by the Kirkwood's superposition approximation method. On the other hand, independence approximation is clearly the worst approximation. This shows that if dependence or relationship exists among variables, then the accuracy of the approximation decreases due to the independence approximation among variables.

To further evaluate the performances of approximation methods in the long run, we run a simulation to compare the performance of the different approximation methods and show which method(s) is the best. For each of the approximation methods, we generated 1.000.000 discrete joint probability distribution samples. Table-5.5 displays a summary of mean of errors of approximate distributions calculated by approximation methods in Table 5.1.

Overall the best performance was obtained with the maximum entropy approximation (MaxEnt-three way) and the maximum CRE approximation (MaxEnt CRE- three way) methods.

Table 5.5: Simulation Results of Approximations of Four Binary Variable Distribution

Approximation Method	Absolute Deviation(AD)	Least Squares (LS)	Maximum Deviation (MS)
Chow-Liu Dependence Tree	0.3612	0.1121	0.0506
CRE First Order Dependence Tree	0.3548	0.1118	0.0499
Maximum Entropy (Pairwise)	0.3113	0.0099	0.0443
Maximum Entropy (Three-way)	**0.0739**	**0.0006**	**0.0046**
Maximum CRE Entropy (Pairwise)	0.3112	0.0099	0.0443
Maximum CRE Entropy (Three-way)	**0.0732**	**0.0006**	**0.0046**
Lewis Product Approximation	0.4220	0.0189	0.0680
Brown's Approximation	0.3775	0.0149	0.0588
Ku&Kullback's Approximation	0.3383	0.0118	0.0519
Keefer Binary Approximation	0.4633	0.0228	0.0766
Kirkwood Approximation	**0.2126**	**0.0074**	**0.0509**
Independence Approximation	0.5848	0.0363	0.1062

These two methods consistently achieve a better performance in the long run. However, there is no general agreement regarding which method performs best. The third method that performs best is the Kirkwood's superposition approximation method. These results are acceptable and justifiable because all of these three approximation methods use three-variate lower order assessments which are the highest lower order assessments in a four variable joint distribution.

The other approximation methods are not as good as these three methods because they use the conditional assessments or pairwise assessments as the highest assessment in the analysis. The results also show that the amount of available information and the accuracy of the approximations are positively related. If more information is available, then more accurate joint distributions are approximated.

We also performed another analysis to find out how many percent of the samples performs best, second best, third best, and worst in the long run. Table 5.6 presents the percentage of time each approximation is the best, second best, third best, and worst. The results in Table-5.6 shows that, best approximation for 1% of the samples. After sampling one million four binary joint probability distribution, maximum entropy and maximum CRE are the best approximation methods for 99% of the samples, and Kirkwood is the Although the difference between maximum entropy and maximum CRE is small in the long run, the maximum CRE is slightly better than maximum entropy method. As we can see from the Table-5.6 and the Figure-5.1, the third best approximation method is the Kirkwood's superposition method but the accuracy of the Kirkwood's superposition approximation method is far away from the maximum entropy and maximum CRE approximations.

In our analysis, we didn't take into account the independence approximation. The independence approximation is the worst approximation method for more than 99% of the samples. So, we remove the accuracy results of the independence approximation method to make the results of the approximation methods more informative and understandable.

As we can see from Figure 5.1, the results show that the other approximation methods provide a reasonable approximation, but the results of errors are higher compared to the maximum entropy and maximum CRE methods. On the other hand, maximum entropy, maximum CRE, and

Table 5.6: Percentage of Distributions for Different Approximations and Measures for Four Binary Variable Example

Approximation	Measure	% of Distributions Best	% of Distributions Second Best	% of Distributions Third Best	% of Distributions Worst
Chow-Liu	AD	0.0%	0.0%	5.7%	8.4%
	LS	0.0%	0.0%	5.5%	8.5%
	MD	0.0%	0.0%	4.2%	8.2%
CRE FODT	AD	0.0%	0.0%	6.5%	6.9%
	LS	0.0%	0.0%	6.7%	7.3%
	MD	0.0%	0.0%	5.3%	7.1%
MaxEnt (pairwise)	AD	0.0%	0.0%	6.9%	3.1%
	LS	0.0%	0.0%	11.3%	2.7%
	MD	0.0%	0.0%	15.7%	2.9%
MaxEnt (three-way)	AD	48.3%	50.4%	1.3%	0.0%
	LS	48.6%	50.7%	0.5%	0.0%
	MD	49.5%	50.1%	0.1%	0.0%
MaxEnt CRE (pairwise)	AD	0.0%	0.0%	16.7%	4.4%
	LS	0.0%	0.0%	14.1%	3.9%
	MD	0.0%	0.0%	14.3%	4.1%
MaxEnt CRE (three-way)	AD	50.8%	48.7%	0.5%	0.0%
	LS	51.0%	49.0%	0.0%	0.0%
	MD	50.5%	49.5%	0.0%	0.0%
Keefer	AD	0.0%	0.0%	1.7%	27.6%
	LS	0.0%	0.0%	1.5%	27.3%
	MD	0.0%	0.0%	1.4%	26.7%
Lewis	AD	0.0%	0.0%	4.1%	19.7%
	LS	0.0%	0.0%	4.2%	20.1%
	MD	0.0%	0.0%	1.7%	19.9%
Brown	AD	0.0%	0.0%	6.0%	14.1%
	LS	0.0%	0.0%	5.1%	15.0%
	MD	0.0%	0.0%	4.3%	12.1%
Ku&Kullback	AD	0.0%	0.0%	6.7%	8.9%
	LS	0.0%	0.0%	7.1%	7.8%
	MD	0.0%	0.0%	6.3%	8.3%
Kirkwood	AD	0.9%	0.9%	43.9%	6.9%
	LS	0.4%	0.3%	44.0%	7.4%
	MD	0.0%	0.4%	46.7%	10.7%

Kirkwood superposition approximation approaches require more computational effort because these three methods use three-way assessments in the approximation procedure.

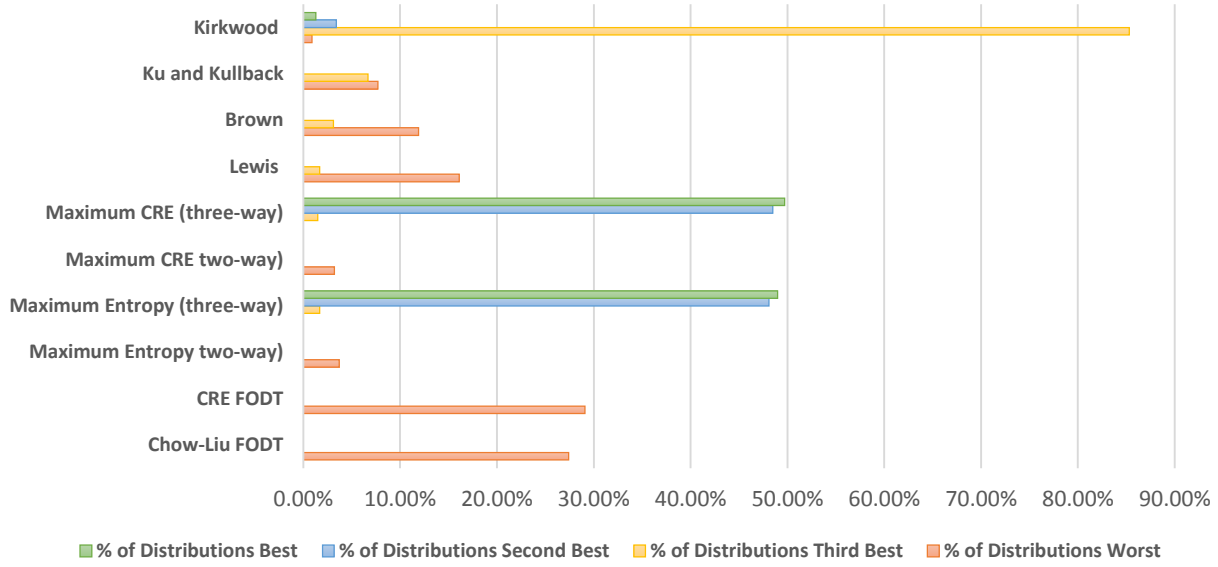
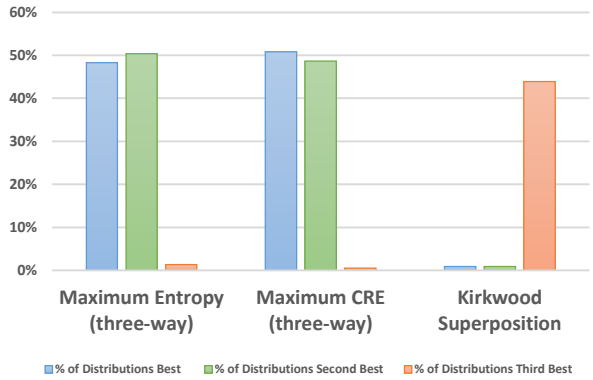


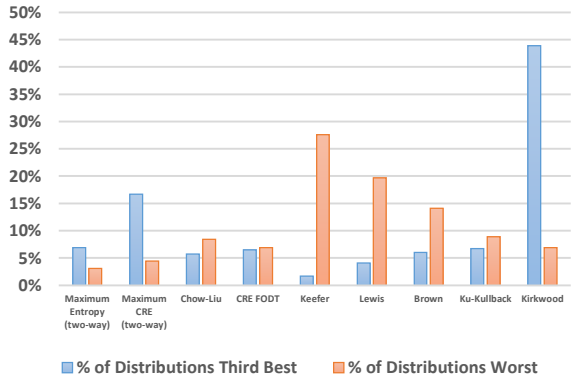
Figure 5.1: Percentage of Four Binary Distribution Approximations of Approximation Methods for Absolute Deviation Measure

For clearer presentation and interpretation, the results are divided into two groups. The first group consists of the percentage of the best three approximation methods for each error measure in Figure 5-2 (a,b,c). The second group consists the percentage of the third best and the worst approximations for each error measure in Figure 5-2 (d,e,f). The first group of the results shows that the maximum entropy and maximum CRE methods are very similar and approximate almost identical joint probability distributions. So, we can say these two methods can be used as an alternative methods to one another.

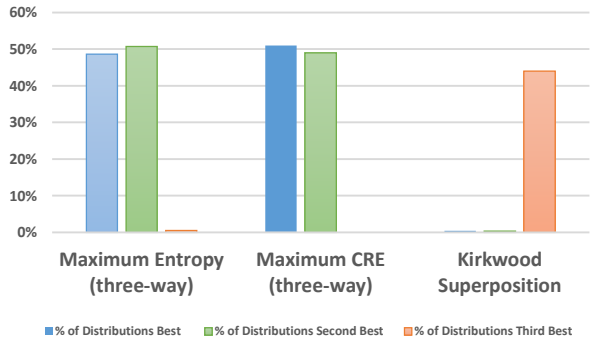
From Figure 5.2-d-e-f, we have found that Keefer's binary event approximation method is the second worst approximation method after independence approximation. The percentage of the maximum CRE (pairwise) as the third best approximation is more than 10%, and the best approximation after Kirkwood's superposition approximation. Maximum entropy (pairwise) follows the maximum CRE approximation and gives closer approximation to the maximum CRE



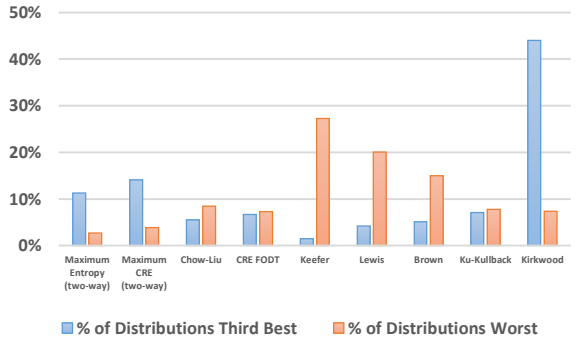
b) Percentage of Best Three Approximation Methods for Absolute Deviation Measure



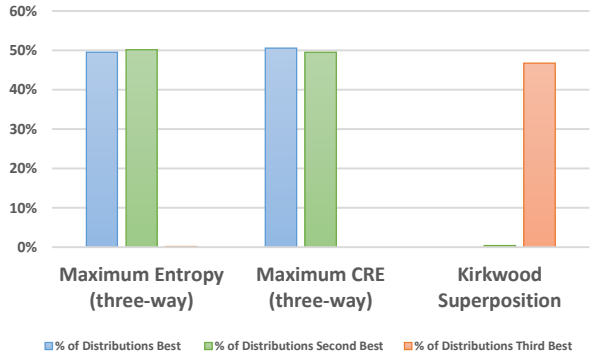
a) Percentage of Third Best and Worst Approximations for Absolute Deviation Measure



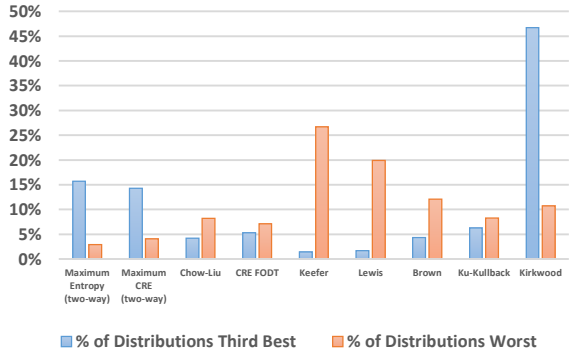
d) Percentage of Best Three Approximation Methods for Least Squares Measure



c) Percentage of Third Best and Worst Approximations for Least Squares Measure



f) Percentage of Best Three Approximation Methods for Maximum Deviation Measure



e) Percentage of Third Best and Worst Approximations for Maximum Deviation

Figure 5.2: Percentage of Approximations of Four Binary Distributions for Different Error Measures

approximation. Lewis', Brown's, and Ku-Kullback approximation methods give also similar results to the maximum CRE and maximum entropy results, but these methods requires more computational effort. The accuracy of the approximation result can be improved by increasing the

number of iterations. We can do so many steps of an iterative method as necessary to reach proper accuracy. Although a higher number of iteration steps implies a better approximation, the cost per iteration usually increases exponentially. So, in our analysis, we spend the same number of iteration steps to get to the result.

5.4. Approximating the Four Three-outcome Variable Joint Distributions

We now provide another analysis with four variables, each of which contains three different values or outcomes. In this case, we have 81 different combinations or alternatives. As in the previous case, we first show an a specific example of a four-variate distribution where each variable takes three different values, along with approximations from lower order assessments derived from use of the true distribution. We then generate one million four-variate joint distributions and compare the performances of the different approximations with respect to the original distribution. An example of the four-variate probability distribution is given in Table-5.7. In the case that each variable has three values, the variables take three different values: “1”, “2”, or “3”.The first four columns in Table 5.7 describe the different combinations of joint distribution. The fifth column contains the probabilities of the joint events generated by using the method described in Chapter 3. For instance, the probability of the combination of the variables where variable 1 and variable 2 take the values “1”, variable 3 takes the value “2” and the variable-4 takes the value “3” is “0.0023”.

We apply the same error measures used in the four binary joint distribution example to analyze and observe the performance of the approximations. Keefer’s binary event approximation method is only applied to binary variables, so we discard Keefer’s approximation method from the analysis. The results of four variable joint distribution are shown in Table-5.8.

Table 5.7: Approximations of Four Binary Distribution

X1	X2	X3	X4	$P(x_1,x_2,x_3,x_4)$	Chow-Liu	CRE FODT	MaxEnt (pairwise)	MaxEnt (three-way)	MaxEnt CRE (pairwise)	MaxEnt CRE (three-way)	Lewis*	Brown*	Ku& Kullback*	Keefer	Kirkwood	Indepen-dence
1	1	1	1	0.0188	0.0149	0.0148	0.0128	0.0172	0.0128	0.0172	0.0143	0.0131	0.0130	-	0.0173	0.0159
1	1	1	2	0.0209	0.0131	0.0130	0.0121	0.0224	0.0122	0.0224	0.0147	0.0134	0.0123	-	0.0243	0.0153
1	1	1	3	0.0029	0.0165	0.0163	0.0151	0.0031	0.0152	0.0031	0.0151	0.0138	0.0150	-	0.0057	0.0160
1	1	2	1	0.0211	0.0139	0.0151	0.0156	0.0194	0.0156	0.0194	0.0129	0.0139	0.0157	-	0.0202	0.0112
1	1	2	2	0.0146	0.0124	0.0135	0.0148	0.0154	0.0147	0.0154	0.0132	0.0143	0.0149	-	0.0150	0.0108
1	1	2	3	0.0023	0.0102	0.0111	0.0122	0.0032	0.0123	0.0031	0.0136	0.0146	0.0121	-	0.0045	0.0112
1	1	3	1	0.0064	0.0123	0.0114	0.0113	0.0098	0.0112	0.0098	0.0123	0.0128	0.0114	-	0.0082	0.0129
1	1	3	2	0.0126	0.0142	0.0131	0.0138	0.0104	0.0138	0.0104	0.0126	0.0131	0.0140	-	0.0094	0.0124
1	1	3	3	0.0221	0.0143	0.0133	0.0141	0.0210	0.0141	0.0210	0.0130	0.0135	0.0140	-	0.0199	0.0130
1	2	1	1	0.0223	0.0197	0.0164	0.0166	0.0228	0.0166	0.0227	0.0154	0.0169	0.0168	-	0.0234	0.0189
1	2	1	2	0.0036	0.0173	0.0144	0.0146	0.0054	0.0146	0.0055	0.0146	0.0159	0.0146	-	0.0066	0.0182
1	2	1	3	0.0224	0.0217	0.0180	0.0228	0.0202	0.0226	0.0202	0.0188	0.0206	0.0224	-	0.0189	0.0190
1	2	2	1	0.0221	0.0133	0.0166	0.0147	0.0188	0.0148	0.0188	0.0139	0.0130	0.0147	-	0.0200	0.0133
1	2	2	2	0.0112	0.0119	0.0149	0.0130	0.0106	0.0130	0.0106	0.0131	0.0122	0.0128	-	0.0087	0.0128
1	2	2	3	0.0185	0.0098	0.0123	0.0134	0.0224	0.0134	0.0224	0.0169	0.0158	0.0131	-	0.0241	0.0134
1	2	3	1	0.0033	0.0123	0.0126	0.0110	0.0062	0.0110	0.0062	0.0132	0.0124	0.0111	-	0.0065	0.0154
1	2	3	2	0.0097	0.0141	0.0145	0.0125	0.0086	0.0125	0.0085	0.0125	0.0117	0.0125	-	0.0086	0.0148
1	2	3	3	0.0212	0.0143	0.0147	0.0159	0.0195	0.0159	0.0195	0.0161	0.0151	0.0158	-	0.0166	0.0154
1	3	1	1	0.0183	0.0192	0.0180	0.0165	0.0195	0.0165	0.0196	0.0170	0.0166	0.0165	-	0.0175	0.0194
1	3	1	2	0.0222	0.0169	0.0158	0.0166	0.0189	0.0166	0.0189	0.0187	0.0183	0.0167	-	0.0191	0.0186
1	3	1	3	0.0152	0.0212	0.0199	0.0196	0.0172	0.0196	0.0172	0.0180	0.0177	0.0192	-	0.0169	0.0195
1	3	2	1	0.0008	0.0158	0.0183	0.0172	0.0059	0.0171	0.0058	0.0153	0.0152	0.0172	-	0.0069	0.0136
1	3	2	2	0.0196	0.0142	0.0164	0.0174	0.0194	0.0174	0.0194	0.0168	0.0167	0.0175	-	0.0203	0.0131
1	3	2	3	0.0216	0.0117	0.0135	0.0135	0.0167	0.0136	0.0168	0.0162	0.0162	0.0134	-	0.0176	0.0137
1	3	3	1	0.0157	0.0148	0.0139	0.0131	0.0094	0.0132	0.0094	0.0146	0.0149	0.0134	-	0.0097	0.0157
1	3	3	2	0.0175	0.0170	0.0160	0.0173	0.0209	0.0173	0.0210	0.0160	0.0164	0.0175	-	0.0196	0.0151
1	3	3	3	0.0172	0.0172	0.0162	0.0167	0.0200	0.0167	0.0199	0.0155	0.0159	0.0165	-	0.0184	0.0158
2	1	1	1	0.0091	0.0091	0.0094	0.0093	0.0076	0.0093	0.0076	0.0102	0.0093	0.0093	-	0.0078	0.0115
2	1	1	2	0.0151	0.0080	0.0082	0.0079	0.0104	0.0078	0.0104	0.0095	0.0087	0.0079	-	0.0091	0.0110
2	1	1	3	0.0040	0.0100	0.0103	0.0082	0.0101	0.0082	0.0101	0.0083	0.0076	0.0082	-	0.0103	0.0115
2	1	2	1	0.0163	0.0084	0.0081	0.0093	0.0129	0.0093	0.0129	0.0078	0.0084	0.0095	-	0.0125	0.0080
2	1	2	2	0.0007	0.0075	0.0072	0.0080	0.0050	0.0080	0.0050	0.0072	0.0078	0.0081	-	0.0055	0.0077
2	1	2	3	0.0064	0.0062	0.0059	0.0055	0.0055	0.0055	0.0055	0.0063	0.0068	0.0056	-	0.0046	0.0081
2	1	3	1	0.0011	0.0075	0.0075	0.0084	0.0059	0.0085	0.0059	0.0091	0.0094	0.0084	-	0.0065	0.0093
2	1	3	2	0.0022	0.0086	0.0086	0.0093	0.0027	0.0093	0.0027	0.0084	0.0088	0.0094	-	0.0036	0.0089
2	1	3	3	0.0190	0.0087	0.0087	0.0080	0.0138	0.0080	0.0138	0.0073	0.0076	0.0079	-	0.0128	0.0093
2	2	1	1	0.0161	0.0155	0.0134	0.0156	0.0145	0.0155	0.0146	0.0143	0.0156	0.0155	-	0.0147	0.0136
2	2	1	2	0.0073	0.0136	0.0118	0.0122	0.0102	0.0122	0.0102	0.0122	0.0133	0.0122	-	0.0106	0.0131
2	2	1	3	0.0220	0.0170	0.0148	0.0159	0.0206	0.0160	0.0206	0.0133	0.0146	0.0159	-	0.0197	0.0137
2	2	2	1	0.0008	0.0104	0.0115	0.0115	0.0081	0.0115	0.0081	0.0108	0.0101	0.0114	-	0.0075	0.0096
2	2	2	2	0.0101	0.0093	0.0103	0.0091	0.0061	0.0092	0.0061	0.0092	0.0086	0.0090	-	0.0059	0.0092
2	2	2	3	0.0088	0.0077	0.0085	0.0079	0.0055	0.0078	0.0055	0.0101	0.0095	0.0078	-	0.0061	0.0096
2	2	3	1	0.0177	0.0096	0.0107	0.0107	0.0119	0.0107	0.0119	0.0126	0.0118	0.0106	-	0.0111	0.0110
2	2	3	2	0.0184	0.0111	0.0122	0.0109	0.0195	0.0108	0.0195	0.0108	0.0101	0.0108	-	0.0214	0.0106
2	2	3	3	0.0043	0.0112	0.0124	0.0116	0.0090	0.0117	0.0090	0.0118	0.0111	0.0115	-	0.0094	0.0111
2	3	1	1	0.0113	0.0145	0.0141	0.0146	0.0143	0.0146	0.0142	0.0150	0.0147	0.0146	-	0.0139	0.0139
2	3	1	2	0.0103	0.0127	0.0124	0.0133	0.0122	0.0133	0.0122	0.0150	0.0146	0.0134	-	0.0111	0.0134
2	3	1	3	0.0149	0.0159	0.0156	0.0131	0.0101	0.0131	0.0102	0.0123	0.0120	0.0130	-	0.0106	0.0140
2	3	2	1	0.0164	0.0119	0.0121	0.0127	0.0124	0.0127	0.0125	0.0114	0.0114	0.0129	-	0.0117	0.0098
2	3	2	2	0.0174	0.0107	0.0109	0.0116	0.0171	0.0116	0.0172	0.0113	0.0113	0.0118	-	0.0180	0.0094
2	3	2	3	0.0064	0.0088	0.0090	0.0077	0.0106	0.0077	0.0106	0.0093	0.0093	0.0077	-	0.0096	0.0098
2	3	3	1	0.0157	0.0111	0.0113	0.0122	0.0167	0.0122	0.0167	0.0133	0.0136	0.0122	-	0.0172	0.0113
2	3	3	2	0.0151	0.0128	0.0129	0.0144	0.0136	0.0145	0.0135	0.0132	0.0136	0.0145	-	0.0147	0.0109
2	3	3	3	0.0038	0.0129	0.0131	0.0116	0.0043	0.0116	0.0043	0.0108	0.0111	0.0116	-	0.0052	0.0113
3	1	1	1	0.0027	0.0120	0.0152	0.0146	0.0059	0.0145	0.0058	0.0157	0.0143	0.0143	-	0.0068	0.0120
3	1	1	2	0.0115	0.0105	0.0134	0.0123	0.0148	0.0123	0.0147	0.0145	0.0132	0.0121	-	0.0145	0.0116
3	1	1	3	0.0222	0.0132	0.0168	0.0149	0.0158	0.0150	0.0158	0.0151	0.0138	0.0150	-	0.0149	0.0121
3	1	2	1	0.0079	0.0111	0.0078	0.0089	0.0129	0.0089	0.0129	0.0071	0.0077	0.0087	-	0.0132	0.0084
3	1	2	2	0.0135	0.0100	0.0070	0.0075	0.0084	0.0075	0.0084	0.0066	0.0071	0.0074	-	0.0084	0.0081
3	1	2	3	0.0052	0.0082	0.0058	0.0062	0.0052	0.0061	0.0052	0.0069	0.0074	0.0062	-	0.0040	0.0085
3	1	3	1	0.0174	0.0099	0.0096	0.0105	0.0092	0.0106	0.0092	0.0111	0.0115	0.0103	-	0.0091	0.0097
3	1	3	2	0.0059	0.0114	0.0111	0.0116	0.0077	0.0116	0.0078	0.0102	0.0106	0.0113	-	0.0077	0.0094
3	1	3	3	0.0117	0.0115	0.0112	0.0115	0.0180	0.0115	0.0180	0.0107	0.0111	0.0115	-	0.0165	0.0098
3	2	1	1	0.0162	0.0160	0.0170	0.0185	0.0172	0.0186	0.0173	0.0171	0.0188	0.0187	-	0.0157	0.0143
3	2	1	2	0.0206	0.0141	0.0149	0.0145	0.0160	0.0146	0.0159	0.0145	0.0159	0.0145	-	0.0153	0.0137
3	2	1	3	0.0222	0.0177	0.0187	0.0220	0.0257	0.0220	0.0257	0.0191	0.0209	0.0227	-	0.0247	0.0144
3	2	2	1	0.0126	0.0108	0.0088	0.0082	0.0086	0.0082	0.0086	0.0078	0.0073	0.0082	-	0.0083	0.0100
3	2	2	2	0.0032	0.0097	0.0078	0.0065	0.0078	0.0065	0.0078	0.0066	0.0062	0.0065	-	0.0070	0.0097
3	2	2	3	0.0034	0.0080	0.0065	0.0065	0.0028	0.0065	0.0028	0.0087	0.0081	0.0067	-	0.0040	0.0101
3	2	3	1	0.0059	0.0100	0.0108	0.0101	0.0089	0.0101	0.0089	0.0121	0.0113	0.0101	-	0.0093	0.0116
3	2	3	2	0.0194	0.0115	0.0124	0.0103	0.0194	0.0103	0.0195	0.0102	0.0096	0.0102	-	0.0204	0.0111
3	2	3	3	0.0059	0.0116	0.0125	0.0127	0.0029	0.0128	0.0029	0.0135	0.0126	0.0131	-	0.0052	0.0116
3	3	1	1	0.0188	0.0127	0.0152	0.0151	0.0146	0.0150	0.0146	0.0153	0.0149	0.0149	-	0.0160	0.0146
3	3	1	2	0.0056	0.0111	0.0133	0.0136	0.0070	0.0136	0.0071	0.0151	0.0148	0.0135	-	0.0075	0.0140
3	3	1	3	0.0215	0.0140	0.0167	0.0156	0.0243	0.0156	0.0242	0.0148	0.0145	0.0158	-	0.0239	0.0147
3	3	2	1	0.0081	0.0105	0.0078	0.0079	0.0070	0.0080	0.0070	0.0069	0.0069	0.0078	-	0.0068	0.0103
3	3	2	2	0.0045	0.0094	0.0070	0.0072	0.0051	0.0072	0.0050	0.0069	0.0069	0.0071	-	0.0048	0.0099
3	3	2	3	0.0058	0.0077	0.0058	0.0055	0.0064	0.0055	0.0064	0.0067	0.0067	0.0056	-	0.0055	0.0103
3	3	3	1	0.0142	0.0098	0.0096	0.0100	0.0195	0.0100	0.0195	0.0107	0.0110	0.0099	-	0.0201	0.0118
3	3	3	2	0.0109	0.0112	0.0110	0.0117	0.0091	0.0118	0.0090	0.0106	0.0109	0.0116	-	0.0084	0.0114
3	3	3	3	0.0081	0.0114	0.0112	0.0110	0.0047	0.0110	0.0048	0.0104	0.0107	0.0112	-	0.0067	0.0119

(*) After 6 iterations

Table 5.8: Example of Approximations of Four Three-Outcome Variable Distribution

Approximation Method	Absolute Deviation(AD)	Least Squares (LS)	Maximum Deviation (MD)
Chow-Liu Dependence Tree	0.40842	0.00296	0.01456
CRE First Order Dependence Tree	0.39752	0.00291	0.01750
Maximum Entropy (Pairwise)	0.38322	0.00278	0.01637
Maximum Entropy (Three-way)	**0.22996**	**0.00095**	**0.00819**
Maximum CRE Entropy (Pairwise)	0.38299	0.00278	0.01632
Maximum CRE Entropy (Three-way)	**0.22973**	**0.00095**	**0.00818**
Lewis Product Approximation	0.40124	0.00290	0.01443
Brown's Approximation	0.39202	0.00285	0.01441
Ku&Kullback's Approximation	0.38263	0.00277	0.01640
Keefer Binary Approximation	-	-	-
Kirkwood Approximation	**0.24433**	**0.00104**	**0.00826**
Independence Approximation	0.43624	0.00321	0.01457

From Table 5.8, for the case of $3\times3\times3\times3$ joint distributions, we have found that the results are very similar to those obtained with the four-binary joint distribution case. In this case, we see that the maximum entropy approximation (MaxEnt-three way) and the maximum CRE approximation (MaxEnt CRE- three way) are also the best approximations; the third best performance is achieved by the Kirkwood's superposition approximation method.

For convenience, we also run another simulation with $3\times3\times3\times3$ joint distributions. Table 5.9 displays a summary of mean of errors of four three-outcome joint probability distributions calculated by methods in Table 5.1.

Table 5.9: Simulation Results of Approximations of Four Three-Outcome Variable Distribution

Approximation Method	Absolute Deviation(AD)	Least Squares (LS)	Maximum Deviation (MD)
Chow-Liu Dependence Tree	0.5888	0.0030	0.0146
CRE First Order Dependence Tree	0.5719	0.0029	0.0175
Maximum Entropy (Pairwise)	0.5251	0.0060	0.0258
Maximum Entropy (Three-way)	**0.2128**	**0.0009**	**0.0084**
Maximum CRE Entropy (Pairwise)	0.5251	0.0060	0.0251
Maximum CRE Entropy (Three-way)	**0.2128**	**0.0009**	**0.0085**
Lewis Product Approximation	0.5896	0.0077	0.0329
Brown's Approximation	0.5594	0.0069	0.0299
Ku&Kullback's Approximation	0.5296	0.0061	0.0273
Keefer Binary Approximation	-	-	-
Kirkwood Approximation	**0.2856**	**0.0019**	**0.0171**
Independence Approximation	0.6904	0.0107	0.0441

From Table 5.9, we can see that the mean of absolute deviation for the Maximum CRE (three-way) is 0.212867, and for the maximum entropy (three-way) method is 0.212868. The ratio of the means of absolute deviation of maximum CRE method to the maximum entropy method is less than %0.01. Also, the mean of the least squares error between maximum CRE approximate distribution and true distribution is 0.0009, and for maximum entropy method, the least square error is 0.0009, a small deviation in many problems. The Kirkwood superposition approximation is the third best approximation with an absolute deviation error of 0.2856, and least squares error of 0.0019.

We also performed the percentage analysis for this example to find out what percentage of the samples performs best, second best, third best, and worst in the long run. Table 5.6 presents the percentage of time each approximation is the best, second best, third best, and worst. The results are shown in Table 5.10.

Table 5.10: Percentage of Distributions for Different Approximations and Measures for Four Three-Outcome Variable Example

Approximation	Measure	% of Distributions Best	% of Distributions Second Best	% of Distributions Third Best	% of Distributions Worst
Chow-Liu	AD	0.0%	0.0%	0.0%	27.4%
	LS	0.0%	0.0%	0.0%	27.1%
	MD	0.0%	0.0%	0.1%	29.7%
CRE FODT	AD	0.0%	0.0%	0.0%	29.1%
	LS	0.0%	0.0%	0.0%	26.6%
	MD	0.0%	0.0%	0.3%	22.5%
MaxEnt (pairwise)	AD	0.0%	0.0%	0.0%	3.7%
	LS	0.0%	0.0%	0.0%	3.9%
	MD	0.4%	1.2%	0.8%	5.0%
MaxEnt (three-way)	AD	49.0%	48.1%	1.7%	0.0%
	LS	49.2%	48.8%	0.7%	0.0%
	MD	39.3%	45.1%	10.1%	0.0%
MaxEnt CRE (pairwise)	AD	0.0%	0.0%	0.0%	3.2%
	LS	0.0%	0.0%	0.0%	3.5%
	MD	0.0%	1.4%	1.2%	4.7%
MaxEnt CRE (three-way)	AD	49.7%	48.5%	1.5%	0.0%
	LS	50.6%	47.5%	1.3%	0.0%
	MD	39.1%	49.2%	8.8%	0.0%
Lewis	AD	0.0%	0.0%	1.7%	16.1%
	LS	0.0%	0.0%	1.0%	17.5%
	MD	0.0%	0.0%	3.1%	15.9%
Brown	AD	0.0%	0.0%	3.1%	11.9%
	LS	0.0%	0.0%	2.6%	12.5%
	MD	0.0%	1.1%	2.7%	10.4%
Ku&Kullback	AD	0.0%	0.0%	6.7%	7.7%
	LS	0.0%	0.0%	5.6%	8.3%
	MD	0.9%	0.8%	7.2%	8.5%
Kirkwood	AD	1.3%	3.4%	85.3%	0.9%
	LS	0.2%	3.7%	88.8%	0.6%
	MD	20.3%	1.2%	65.7%	3.3%

The results are very similar to what we found for the four binary variable case. The results for absolute deviation measure show that, after sampling one million four three-outcome joint probability distribution, maximum entropy and maximum CRE are the best approximation

methods for almost 99% of the samples, and the Kirkwood superposition approximation is the best approximation for 1.3% of the samples. In our example, we again didn't take into account the independence approximation because the independence approximation is the worst approximation method for more than 99% of the samples. We also didn't consider the Keefer's approximation method because this method can only be applied to binary variables.

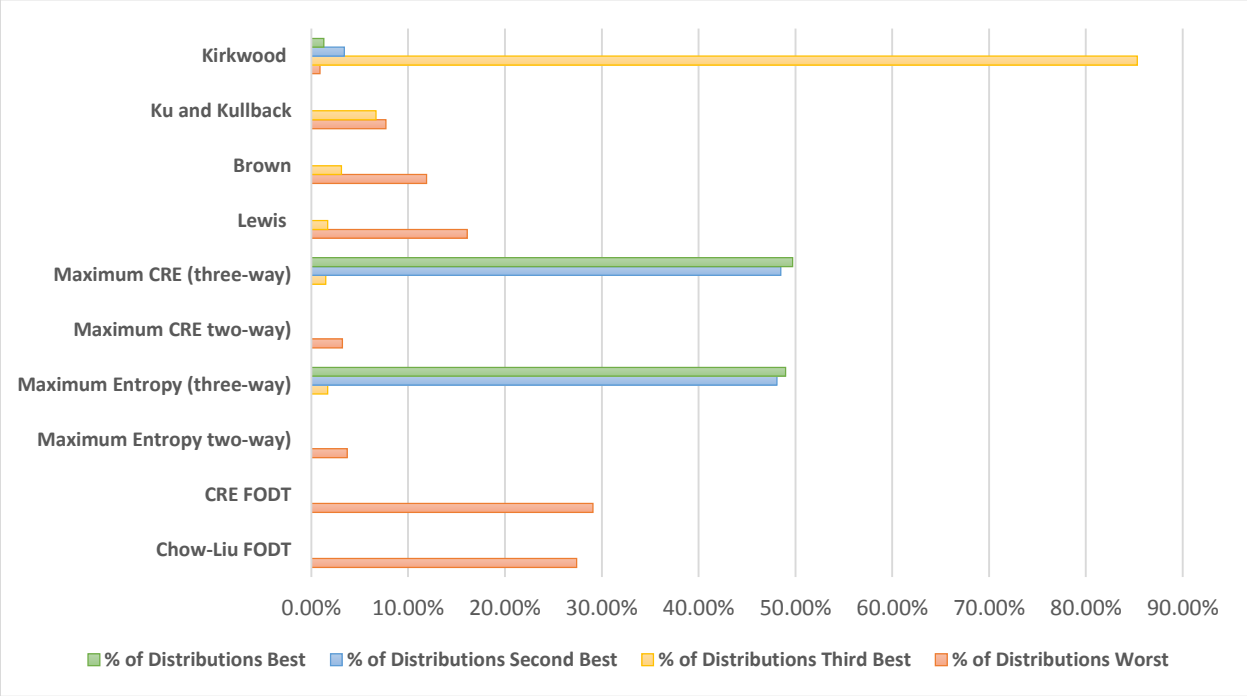
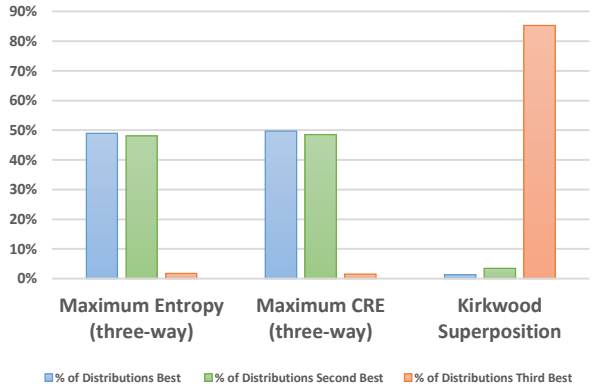
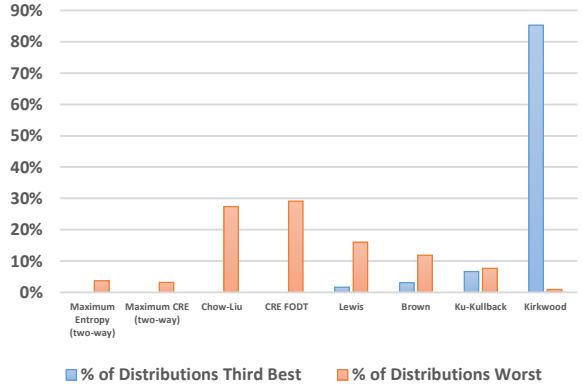


Figure 5.3: Percentage of Four Three-Outcome Distribution Approximations for Absolute Deviation Measure

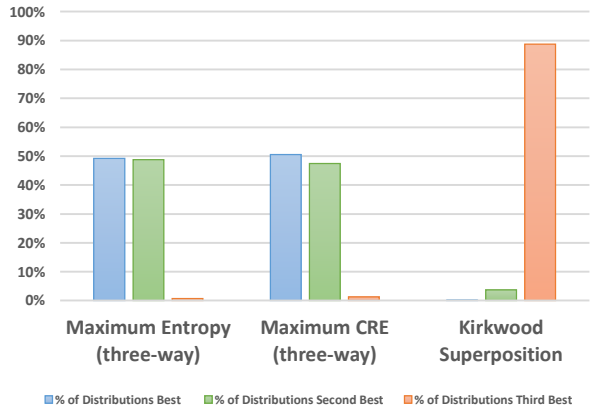
Maximum CRE and maximum entropy methods are obviously the best approximation methods so far. We now discuss the other approximation methods to understand how they perform in the long run. For our detailed analysis, we again divided the results into two groups for clearer presentation and interpretation. Our first group shows the results of each error measure of the best three approximation methods: maximum CRE, maximum entropy, and Kirkwood's superposition approximation. The second group consists of the percentage of the third best and the worst approximations of each error measure. Figure 5.4 shows the results of each group.



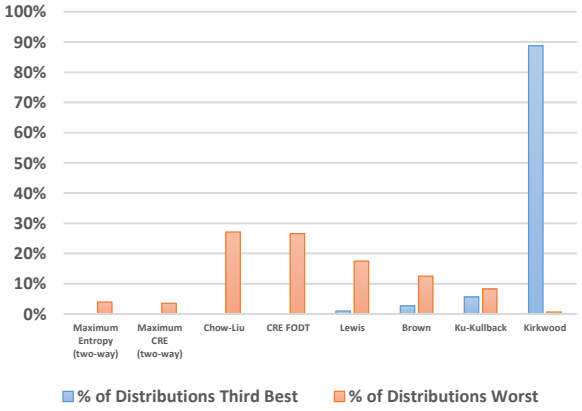
a) Percentage of Best Three Approximation Methods for Absolute Deviation Measure



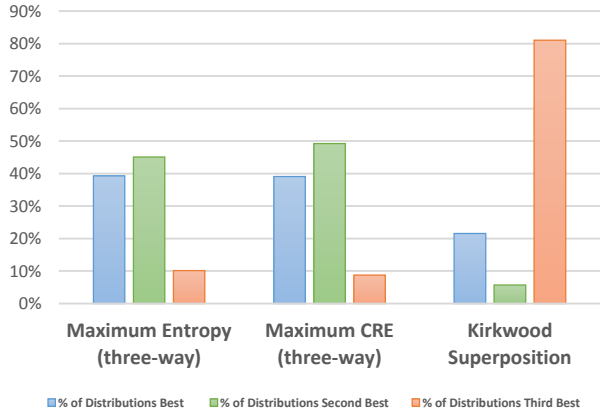
d) Percentage of Third Best and Worst Approximations for Absolute Deviation



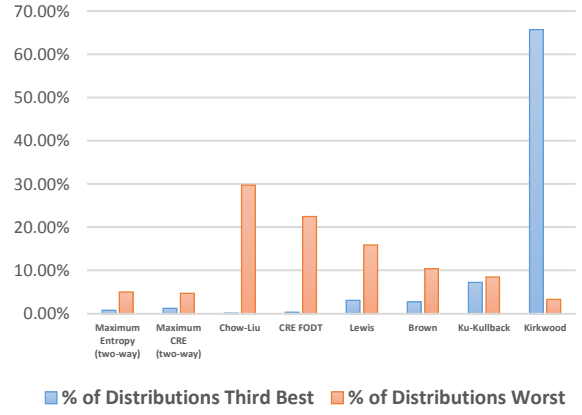
b) Percentage of Best Three Approximation Methods for Least Squares Measure



e) Percentage of Third Best and Worst Approximation Methods for Least Squares



c) Percentage of Best Three Approximation Methods for Maximum Deviation Measure



f) Percentage of Third Best and Worst Approximations for Maximum Deviation

Figure 5.4: Percentage of Approximations of Four Three-Outcome Distributions for Different Error Measures

The first group of the results shows that the maximum entropy and maximum CRE methods are very similar; thus we can say these two methods can be used as an alternative methods to one another. Another observation from the first group of the results is that the Kirkwood superposition approximation is the third best approximation. However, this approximation method is not consistently gives good results because superposition approximation method violates the normalization condition and gives the worst results more than 7% of the time in the long run.

From second group of the results, we have found that Chow-Liu first-order dependence tree approximation method is the second worst approximation method after independence approximation, and the Ku&Kullback's approximation method is the best approximation after maximum CRE (three-way), maximum entropy (three-way), and the Kirkwood's superposition approximation. We can see from Table 5.9 and Figure 5.4 that the mean of errors are very close in the long run for the methods which means that these approximation methods are very close but they are not same. We can say that these methods are very similar and approximate almost identical joint probability distributions.

Another observation from the simulation results is that Lewis's, Brown's, and Ku&Kullback's methods are much harder to apply because these methods are iterative methods. The iteration procedure cannot compose the best approximation because numerous iterations are required to attain full convergence which is very difficult when the number of variables increase.

5.5. Summary of the Results

- ✓ The maximum entropy approximation (MaxEnt-three way) and the maximum CRE approximation (MaxEnt CRE- three way) outperform the other methods.

- ✓ Kirkwood superposition approximation are dominated by maximum entropy and maximum CRE methods and dominate the rest of the approximation methods. So, the third best performance is achieved by the Kirkwood's superposition approximation method.
- ✓ Independence approximation is clearly the worst approximation. After independence approximation, Keefer's binary event approximation is the second worst approximation.
- ✓ Maximum entropy (pairwise), maximum CRE (pairwise), Chow-Liu first-order dependence tree, CRE-based first-order dependence tree, Lewis, Brown, and Ku&Kullback approximation methods give similar accuracy results, however, the computational effort grow exponentially for the iterative approximation methods and makes these methods difficult to apply especially when the number of variables increases.

CHAPTER 6

APPLICATION TO DICE MODEL: APPLYING CUMULATIVE RESIDUAL ENTROPY INTO THE DICE MODEL

6.1. Introduction

This chapter is divided into four sections. Section 1 introduces the DICE model. Section 2 explains the global warming policies used in the DICE model. In section 3, we study the uncertain parameters used in the model and discuss how to use them in our analysis. In section 4, we run simulation experiments for five different policies, and discuss the results of the analysis.

6.2. Dynamic Integrated Climate Economy Model

The Dynamic Integrated Climate-Economy model, referred to as the DICE model, is an integrated assessment model of climate change developed by William Nordhaus and colleagues that integrates both the economic costs and benefits of greenhouse gas controls with an aggregate model linking economic growth with climate change to reduce emission and slow greenhouse warming. The DICE model is a simplified analytical and empirical model that represents the economics, policy, and scientific aspects of climate change. The DICE model attempts to quantify how the atmospheric concentration of CO_2 negatively affects economic output through its impact on global average surface temperature. The model appears to have first been proposed in a discussion paper for the Cowles Foundation in 1992, and the final version of the model is published with updated discussion of the model in 2013. Figure 6-1 shows a schematic flow chart of a full integrated assessment model for climate change science, economics, and policy of the DICE model.

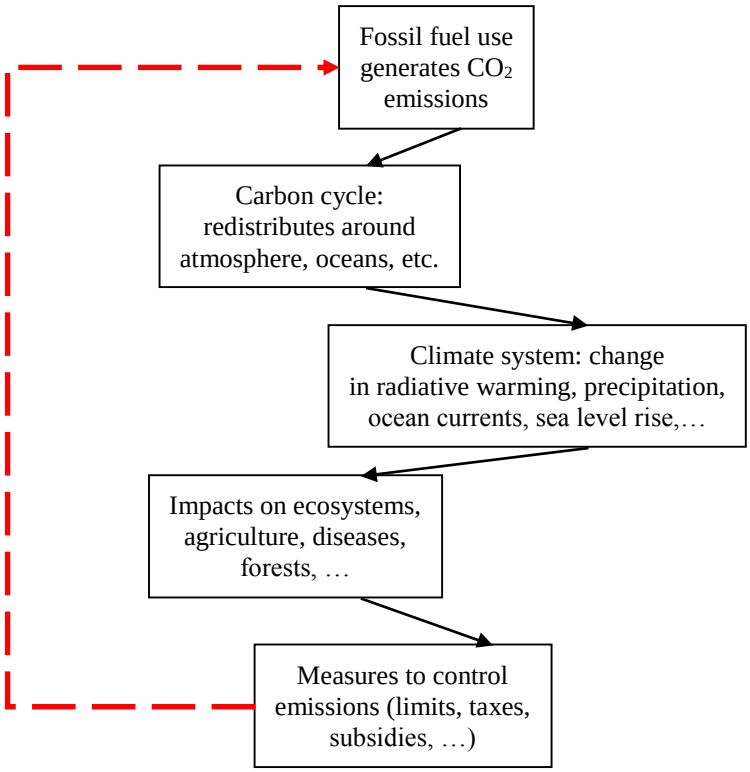


Figure 6.1: Schematic flow chart of a full integrated assessment model for climate change science, economics, and policy (Nordhaus, 2013)

The DICE model is a policy optimization model with an economic objective function that measures economic welfare. The DICE model includes the estimates of both the costs of reducing both carbon dioxide emissions and long term future impacts on climate, enabling costs and benefits of carbon dioxide emissions to be weighed in order to help determine the optimal level of near-term controls. So, the main goal of the DICE model is to maximize the welfare function to evaluate alternative policies.

The DICE Model seeks to choose a policy that maximizes the social welfare function, W, that is the discounted sum of the population-weighted utility of per capita consumption. The objective

function that DICE seeks to maximize is

$$\max \sum_{t=1}^{T_{max}} U[c(t), L(t)]R(t) \tag{6.1}$$

where $c(t)$ is per capita consumption, $L(t)$ is total population, $R(t)$ is the discount factor and $U\left[c(t), L(t)\right]$ is the total worldwide utility of consumption.

The utility function in equation 1.1 is defined as

$$U[c(t), L(t)] = L(t)\left[\frac{[c(t)]^{1-\alpha}}{1-\alpha}\right] \tag{6.2}$$

where utility is equal to population multiplied by per-capita consumption. Per capita consumption is adjusted by an elasticity parameter (α) to account for disparities in equality. If α is close to zero, then the consumptions are close substitutes, with low aversion to inequality; if α is high, then the consumptions are highly differentiated, which reflects high inequality aversion.

$R(t)$ in objective function is the discount factor and defined as

$$R(t) = (1+\rho)^{-t} \tag{6.3}$$

where the pure rate of social time preference, ρ , is the discount rate which provides the welfare weights on the utilities of different generations. Equation 6.3 takes the total utility in each period and discounts it back to the present using the social discount rate (ρ).

The objective function is non-linear in the DICE model and solved by non-linear optimization. Nordhaus divides the constraints into two groups: economic constraints and environmental

constraints. The economic constraints of the DICE model are (i) output of the economy defined with Cobb-Douglas production function and (ii) the capital stock. Environmental constraints are (i) flow of emissions of greenhouse gases, (ii) concentration of GHGs in the atmosphere, (iii) change in temperature in atmosphere and shallow oceans, (iv) damage resulting from increasing temperatures, and (v) economic cost of climate change policies.

6.3. Global Warming Policies used in DICE Model

We now briefly explain the five different environmental policies that are used in our analysis. These policies are selected from the DICE-2013R model (Nordhaus and Sztorc, 2013).

1) Baseline: The baseline policy includes existing policies as of 2010; no new policies are included to slow or reverse greenhouse warming. This policy enables individuals and companies to take precautions or steps to slow down climate change, but governments are assumed to take no steps to control or limit greenhouse-gas emissions.

2) Optimal: The opptimal policy involves weighing the present value of the costs of climate change abatement against the present value of its benefits. This policy sets emission reduction levels to maximize the value of net economic consumption and is the best possible policy for emissions reductions, given our estimated economic, technological, and geophysical constraints.

3) Temperature-Limited: The aim of this policy is to limit the increase in the global temperature to 2^{o}C from the 1900 average (pre-industrial level). The constraints are adapted to not exceed 2^{o}C.

4) Stern Review: The Stern Review-recommended policy uses very low discount rates and is implemented using a time discount rate of 0.1 percent per year and a consumption elasticity of 1, leading to low real interest rates and generally to higher carbon prices and emissions control rates.

5) Copenhagen Accord: This policy is based on the Copenhagen Accord, a continuation of the Kyoto Protocol. In this scenario, developed countries are assumed to implement deep emissions reductions over the next four decades, with developing countries following gradually.

6.4. Analysis of the Model

In this section, we discuss the uncertain variables and their parameters and analyze the model. Our purpose is to integrate climatic and economic uncertainties into the DICE model and find out the cumulative impact of integrating uncertainty on climate change by applying CRE into the DICE model. The aim of the DICE model is to maximize the present value of net welfare for the next sixty time periods, with each period representing five years. We run the simulation for five different scenarios from 2010 to 2300. In the DICE model, there are eight different variables and each is discretized to three different values ($3\times3\times3\times3\times3\times3\times3\times3$ joint distribution). We calculate our analysis for pairwise assessments of maximum entropy and maximum approximation CRE methods. We also calculate deterministic and independence cases.

Nordhaus pointed out eight critical uncertain parameters in the DICE model and we have selected these eight major parameters for further study: i) uncertainties about the growth rate of total factor productivity (ga0), ii) the rate of de-carbonization (dsig), iii) the asymptotic population growth (popasym), iv) the cost of the backstop technology (pback), v) the damage-output coefficient (a2), vi) the transfer coefficient of carbon dioxide (b12), vii) the equilibrium temperature-sensitivity coefficient (t2xCO2), and viii) the total availability of fossil fuels (fosslim). Earlier studies have shown that these parameters have the largest impact on both outcomes and policies. The following table (Table 6.1) shows the marginal distributions of eight variables. We use the values given by Nordhaus in his books “A Question of Balance” (2008) and “DICE 2013R Manual” (2013) and

assume that all the marginal distributions are from known families. All the variables in this analysis are normally distributed.

Table 6.1: Marginal distributions of uncertain variables of DICE model

Variable	Definition of the Variable	Mean	St.Dev	Unit
ga0	Rate of Growth of Total Factor Productivity	0.079	0.004	per year
dsig	Rate of De-carbonization	-0.001	0.002	per year
t2xCO2	Equilibrium Temperature-Sensitivity Coefficient	2.900	1.110	Celcius per CO2 doubling
a2	Damage Parameter	0.003	0.001	Fraction of global output
pback	Price of backstop technology	344	138	$ per ton of carbon replaced
popasym	Asymptotic global population	10500	1892	millions
b12	Transfer coefficient in Carbon Cycle	0.088	0.017	per decade
fosslim	Total Resources of Fossil Fuels	6000	1200	billions of tons of carbon

We discretized each variable by using the McNamee and Celona's (1990) discretization method. We especially use the shortcut of the McNamee-Celona which is called equal areas method. This method divides the cumulative distribution function into intervals between the P100 and the P75, the P75 and the P25, and the P25 and the P0. This produces a weighting of 0.25, 0.50, and 0.25, respectively. This method weights the 10^{th} (P10), 50^{th} (P50), and 90^{th} (P90) percentiles of probability distribution by 0.250, 0.500, and 0.250, respectively. 10% (Low), 50% (Base) and 90% (High) percentiles for each uncertainty by using equal areas method are given in Table 6.2.

Table 6.2: Low, Base and High Percentiles of uncertain variables of the DICE model

	Percentiles		
Variable	Low(10%)	Base(50%)	High(90%)
ga0	0.0739	0.0790	0.0841
dsig	-0.0036	-0.0010	0.0016
t2xCO2	1.4775	2.9000	4.3225
a2	0.0010	0.0027	0.0043
pback	168	344	520
popasym	8075	10500	12925
b12	0.0662	0.0880	0.1098
fosslim	4462	6000	7538

We also show the variables in a decision tree format with their three different outcomes using the low, base and high percentiles of each variable in Figure-6.2.

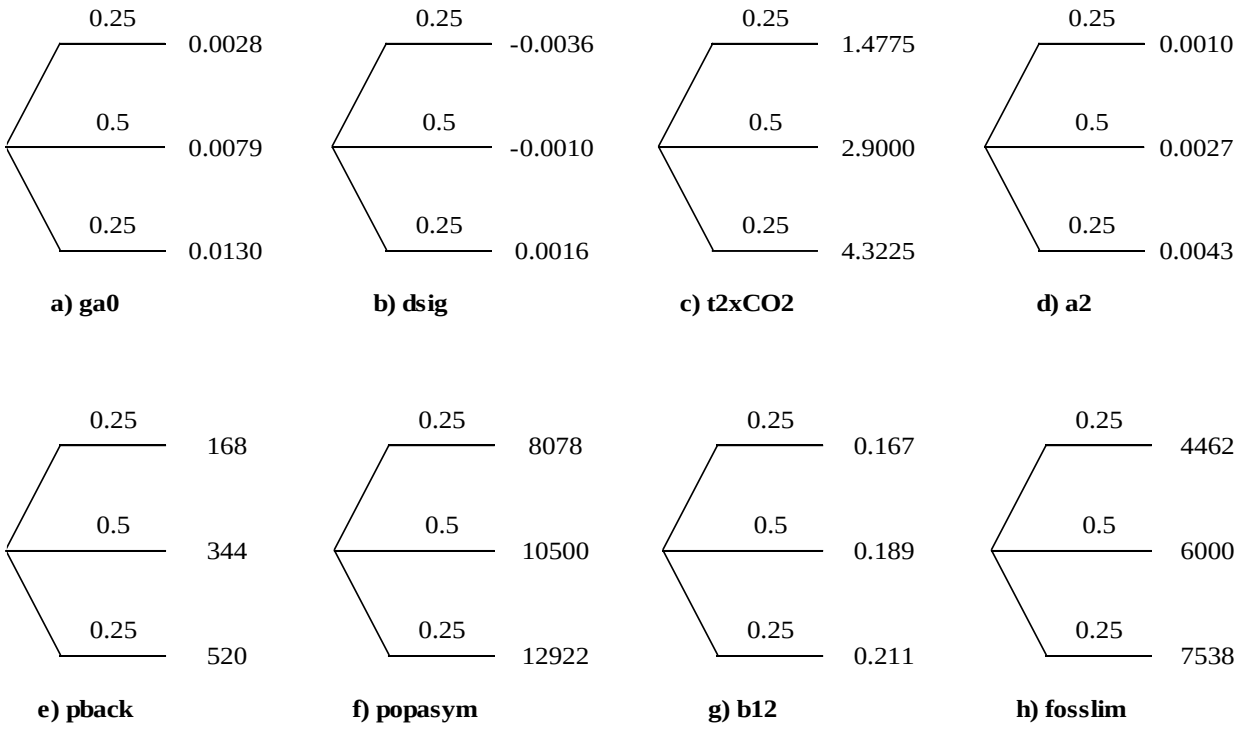


Figure 6.2: Decision trees of eight discretized variables of DICE model

6.5. Results of the Analysis

Nordhaus (2013) used the DICE model to evaluate and compare a number of different environmental policies including (i) baseline scenario, (ii) optimal tax scenario, (iii) limit temperature increase to 2 Celsius degree, (iv) Stern scenario, and (v) Copenhagen Accord scenario. There are eight uncertain parameters in the DICE model; however Nordhaus assumes that there is no uncertainty in the DICE model and fixes all the uncertain parameters to their mean values. In this study, our aim is to integrate uncertainty into the DICE model to understand whether uncertainty has any significant effect on the performances of the policies. We take an approach to handle the dependence between variables and uncertainty in the DICE model instead of fixing the uncertain parameters to their mean values. We incorporate uncertainty into the DICE model by discretizing uncertain parameters. Then we use the uncertain parameters in our analysis to generate joint probability distributions of eight uncertain parameters using our proposed methods to evaluate and compare the expected performances of different policies.

To understand whether uncertainty significantly affects decision making and welfare consequences, three different optimizations are performed. We first assume that there is no uncertainty in the model and fix the values of each uncertain parameter into their mean values same as the Nordhaus' analysis. Our second approach assumes that the parameters are uncertain but they are independent. Finally, we don't assume any independence among variables and take into consideration dependence among variables and integrate it into the DICE model.

In this study, we consider five different types of performance measures to compare and evaluate different global warming policies. We prefer to use performance measures often used in the literature. The performance measures are (i) carbon price or social cost of carbon, which is

optimized from the second period, (ii) carbon concentration in the atmosphere, (iii) average global temperature increase above preindustrial level, (iv) total carbon emission, and (v) net output of abatement cost and climate damages.

The following figures show the results of the analysis. Figure 6.3 and 6.4 show the results for the optimal carbon prices, figure 6.5 and 6.6 show the results of average carbon concentration in the atmosphere, figure 6.7 and 6.8 show the results of global temperature increase above preindustrial level, figure 6.9 and 6.10 show the results of total carbon emission, figure 6.11 and 6.12 show the results of net output of abatement cost and climate damages for five climate change policies.

The results of the simulation experiments indicate that uncertainty does in fact matter and uncertainties about the future cannot be eliminated. Ignoring uncertainty could limit our capability to take corrective actions in the future, resulting in poor policies, leading to inefficient use of resources, and decreasing our ability to avoid many of the more severe consequences of climate change.

In the short run, the difference between the deterministic, independence, and uncertain cases is very small. In the long run, however, a deterministic assumption in which all uncertainties are ignored leads to an underestimate of average carbon prices between 2100 and 2200, and after 2200, a deterministic case leads to an overestimate of carbon prices for the baseline scenario. For the other more aggressive scenarios, the carbon prices for a deterministic case is much higher than the uncertain case after 2050.

The results for the carbon concentration in the atmosphere shows behavior similar to those for the carbon prices. In the 21st century, ignoring uncertainty overestimated carbon concentration in the

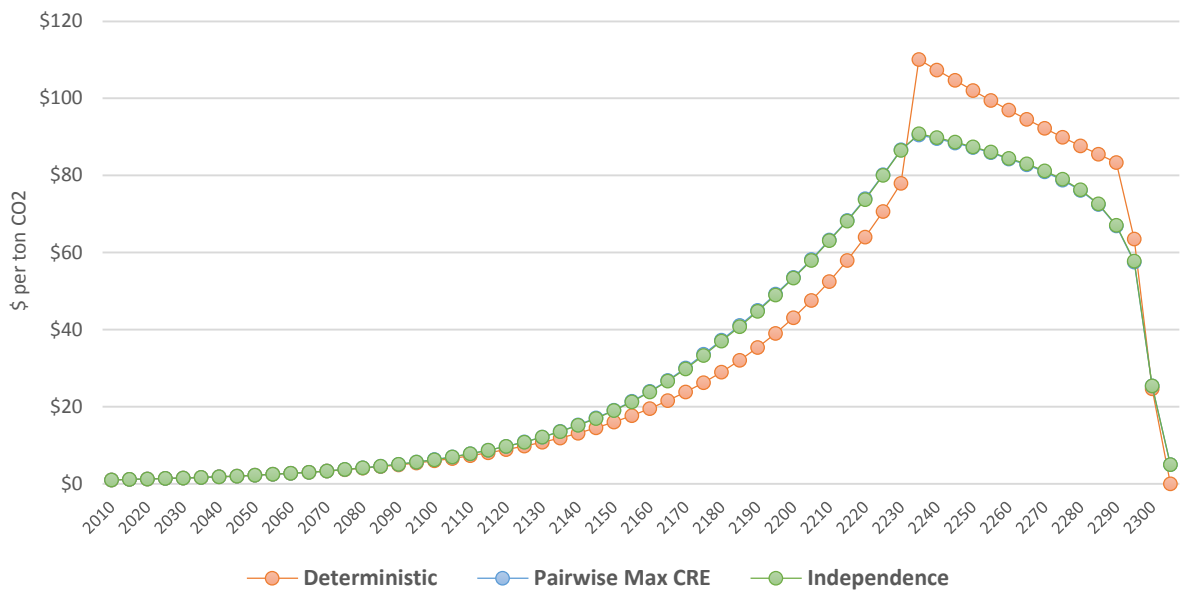
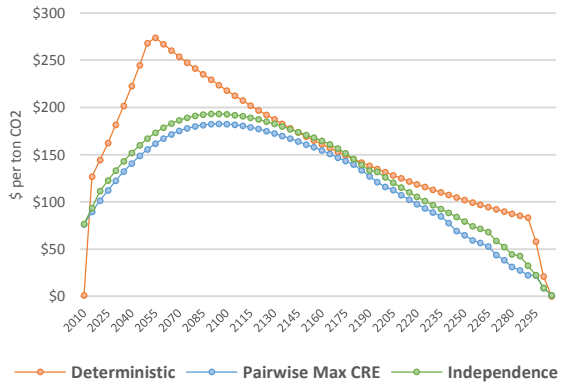


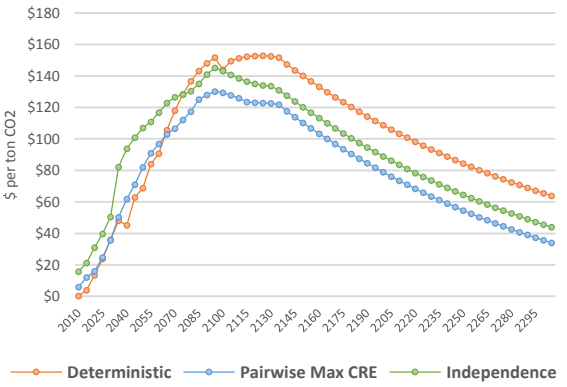
Figure 6.3: Carbon Prices under No Policy 2010-2300 for Deterministic, Independence, and Dependence Cases

(a) Optimal Tax Scenario

(b) Limit 2ºC Scenario



(c) Stern Scenario



(d) Copenhagen Accord

Figure 6.4: Carbon Prices under Alternative Policies from 2010 to 2300 for Deterministic, Independence, and Dependence Cases

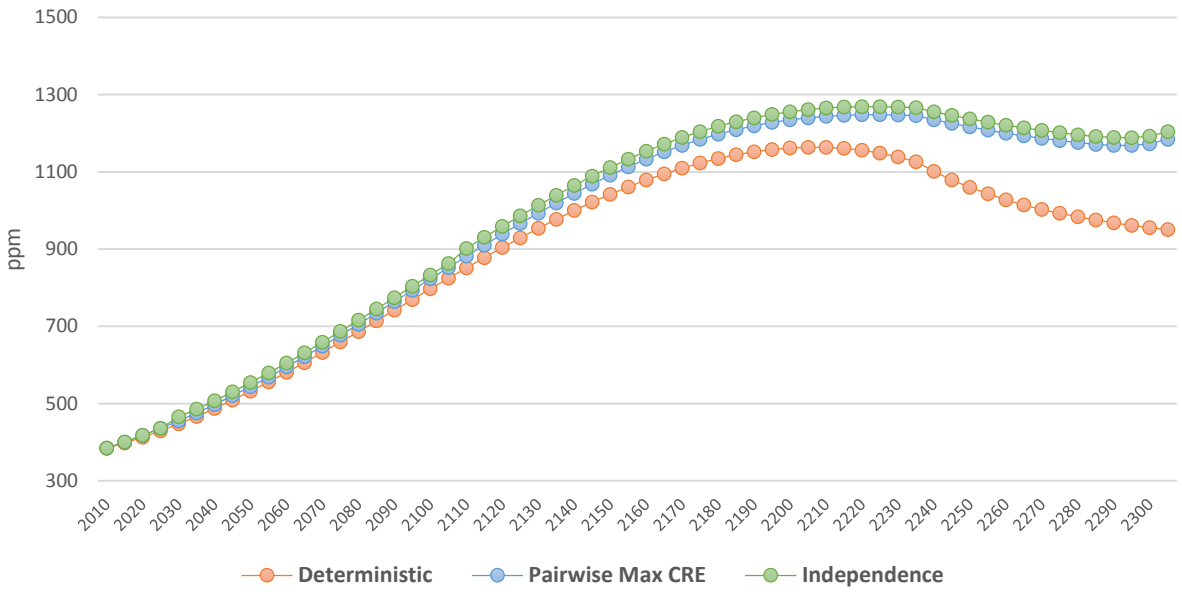
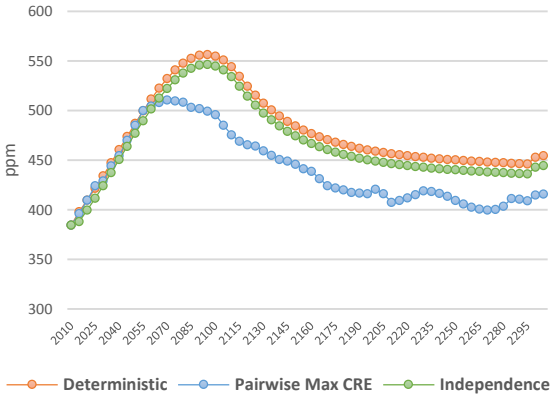
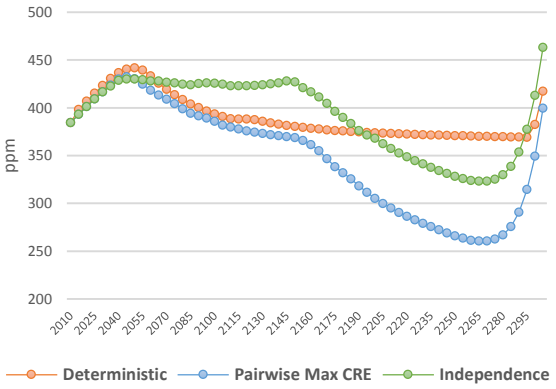


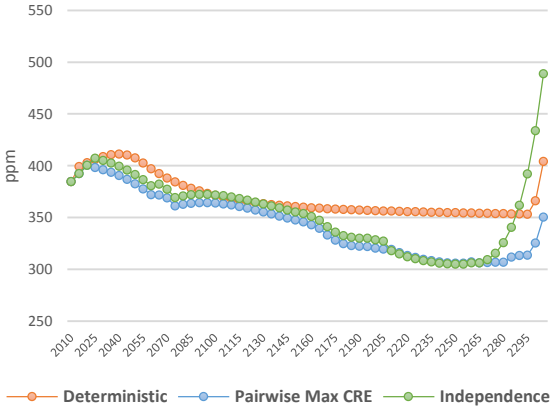
Figure 6.5: Carbon Concentration under No Policy 2010-2300 for Deterministic, Independence, Dependence Cases



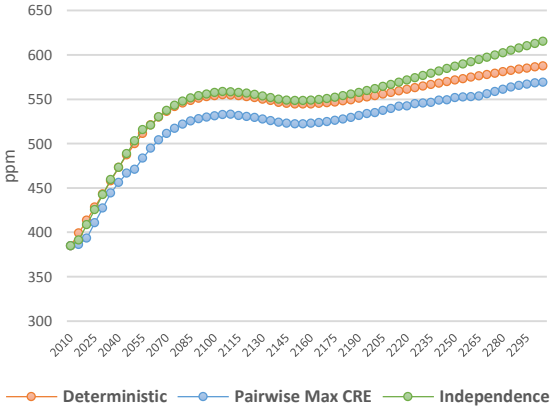
(a) Optimal Tax Scenario



(b) Limit 2ºC Scenario



(c) Stern Scenario



(d) Copenhagen Accord

Figure 6.6: Carbon Concentration under Alternative Policies from 2010 to 2300 for Deterministic, Independence, and Dependence Cases

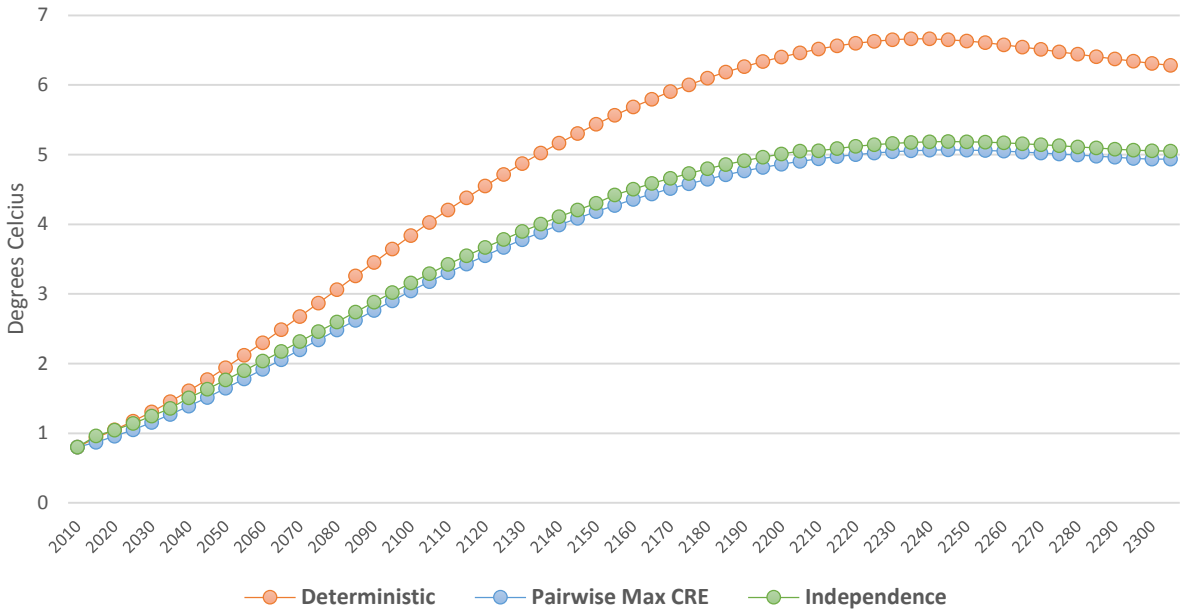
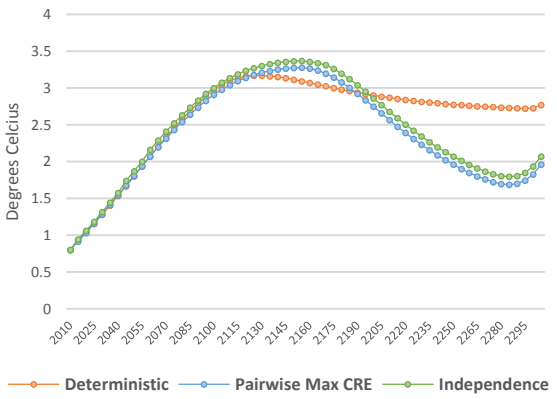
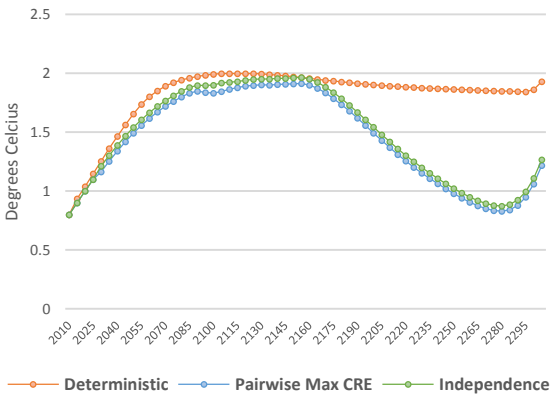


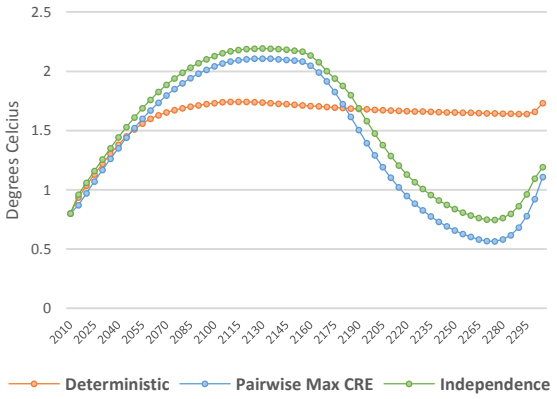
Figure 6.7: Temperature Increase under No Policy 2010-2300 for Deterministic, Independence, Dependence Cases



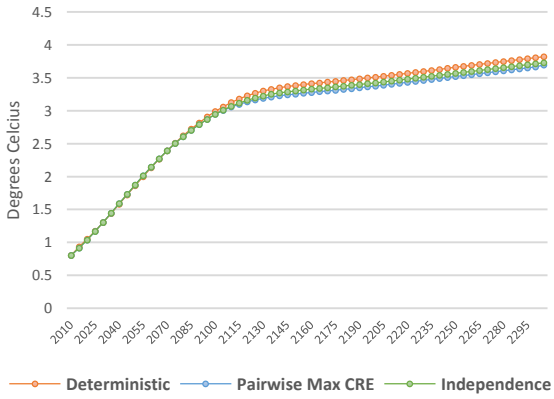
(a) Optimal Tax Scenario



(b) Limit 2ºC Scenario



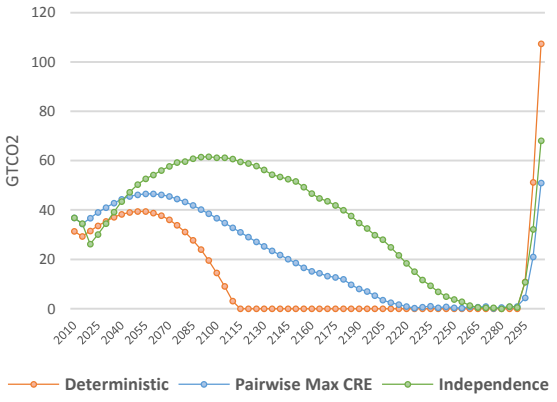
(c) Stern Scenario



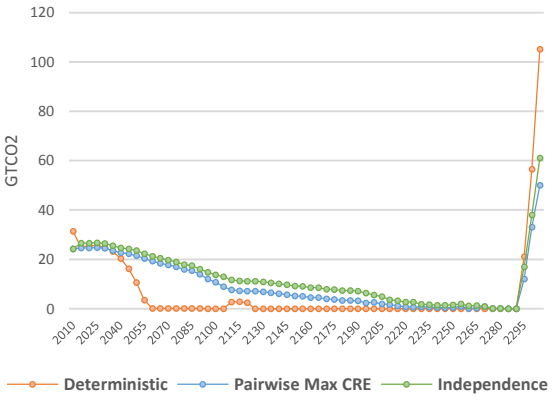
(d) Copenhagen Accord

Figure 6.8: Temperature Increase under Alternative Policies from 2010 to 2300 for Deterministic, Independence, and Dependence Cases

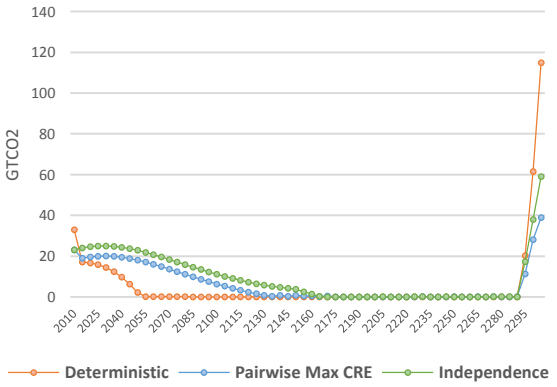
Figure 6.9: Total Carbon Emission under No Policy 2010-2300 for Deterministic, Independence, Dependence Cases



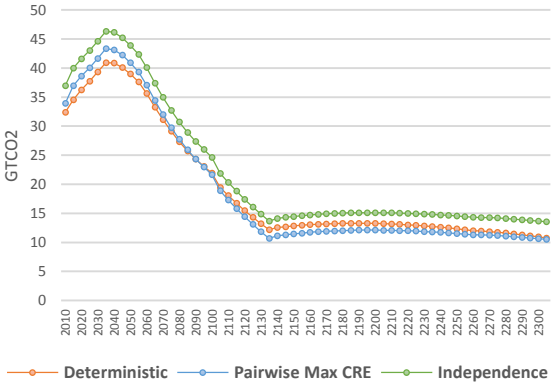
(a) Optimal Tax Scenario



(b) Limit 2ºC Scenario



(c) Stern Scenario



(d) Copenhagen Accord

Figure 6.10: Total Carbon Emission under Alternative Policies from 2010 to 2300 for Deterministic, Independence, and Dependence Cases

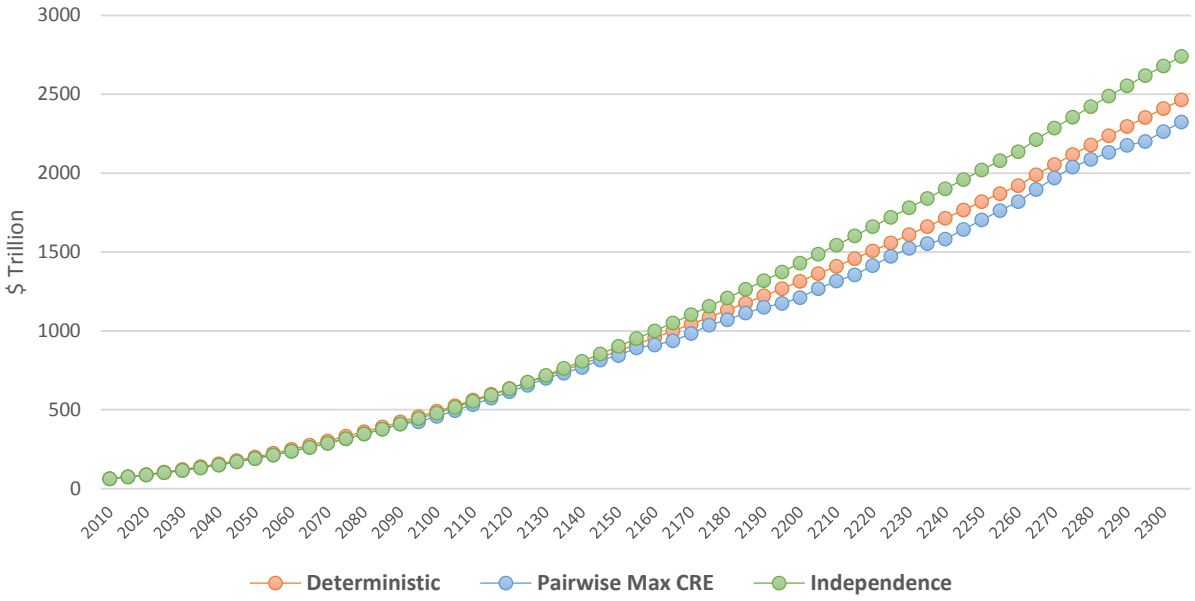
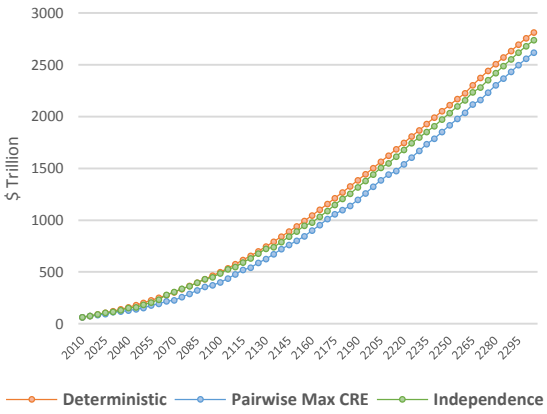
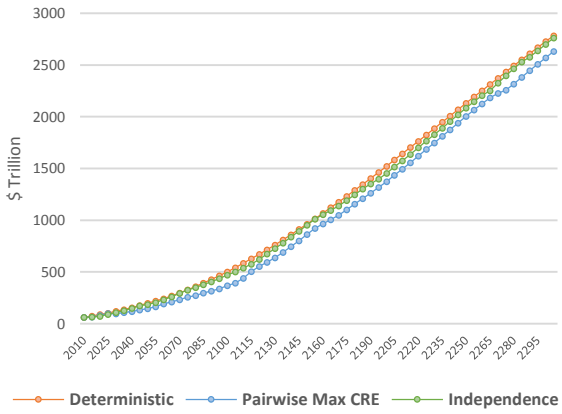


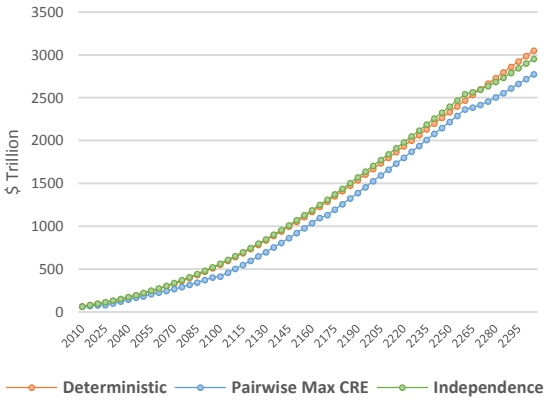
Figure 6.11: Net Output under No Policy 2010-2300 for Deterministic, Independence, and Dependence Cases



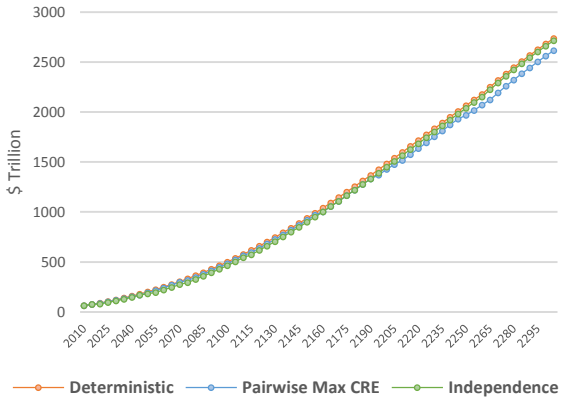
(a) Optimal Tax Scenario



(b) Limit 2ºC Scenario



(c) Stern Scenario



(d) Copenhagen Accord

Figure 6.12: Net Output under Alternative Policies from 2010 to 2300 for Deterministic, Independence, and Dependence Cases

atmosphere. By starting the 22nd century, the rate of carbon concentration decreases significantly when uncertainty is considered. Uncertainty clearly has a dramatic effect on the future path of carbon concentration and average carbon prices, but with a relatively small consequence for immediate decisions. Also, these results show that the carbon concentration in the atmosphere is negatively correlated to carbon prices. As the carbon prices increases, the carbon concentration decreases gradually.

When we analyze the results for the temperature increase performance measure (Figure 6.7 and Figure 6.8) we can say that the results are very similar for deterministic and uncertain cases up to 2100. However after 2150, the deterministic method overestimates the temperature increase. The temperature goes almost constant under deterministic case after 2150 but decreases sharply in the case of uncertainty for optimal tax, limit 2ºC, and Stern scenarios. Significantly, the expected change in mean global temperature associated with the optimal, Stern, and limit 2ºC scenarios is never greater than 3.5 ºC above the pre-industrial norm, peaking in the year 2150 for three scenarios. For the Baseline and Copenhagen Accord scenarios, the temperature is increasing gradually for all time periods and gives similar result under uncertain, independent, and deterministic conditions. Also, temperature increase is inversely proportion to carbon prices and directly proportional to carbon concentration. These results are reasonable because as the price of carbon increases, people and companies prefer to use less carbon-based fuel, which helps to decrease both carbon concentration and average temperature increase from the preindustrial level.

For total carbon emission, the results are very different compared to the other performance measures. The results are closer under uncertain and certain situations for baseline scenario; however, in the other four scenarios the total carbon emission decreases to zero around 2060 under deterministic conditions which doesn't make sense. If we assume that all the values of uncertain

parameters fix to their mean values, this results show that the optimal tax, limit-2°C, Stern, and Copenhagen scenarios are very powerful and achieve the desired goals in a short time. However, in reality, it takes a long time to reduce carbon emission, and the damages resulting from the temperature increase near zero in the long run. If we compare the results with carbon concentration then we can also say that this is impossible in a short period of time. The results under uncertainty, however, make sense because the carbon emission decreases to zero after 2200, which means that after 200 years the policies have influenced people and companies to decrease carbon emission and halt the temperature increase.

Ignoring uncertainty in the net output of abatement cost and climate damages initially appears very similar under uncertain and deterministic conditions; however the deterministic case leads to an overestimation of costs and damages for all scenarios because carbon intensity may decrease at a faster rate under uncertainty and abatement costs are assumed to be directly proportional to carbon intensity.

From Figure 6.11 and 6.12, the estimate values of net values of abatement costs and climate damages under certainty is very similar to net values under uncertainty. Also, the graphs show very similar behavior as time periods increases. It seems that the deterministic case is competitive with the uncertain case because net present values under deterministic and uncertain cases are very close. However, assuming certainty about the future while making global warming decisions is clearly not true in the majority of the policies and performance measures. Although the net present values under deterministic and uncertain cases are close, the average values of temperature increase and carbon emissions show that making decisions under certainty about the global warming may lead to catastrophic climate change.

Ignoring uncertainty in temperature increase, net output of abatement cost and climate damages, and carbon price performance measures causes overestimation of the benefits for all scenarios, but ignoring uncertainty causes slightly underestimation for total carbon emission and significant underestimation for carbon concentration. If there is not enough information to assume that the parameters are deterministic or independent, then dependence among variables and uncertainty should be considered in global warming decisions, as dependence and uncertainty may have a significant impact on climate change and global warming decisions.

We also calculate the average values and standard deviations of the each policy for all performance measures under deterministic and uncertain cases. The following table (Table 6.3) shows the mean and standard deviation values of five output variables from 2010 to 2300; the optimal carbon price, carbon concentration in the atmosphere , temperature increase from the preindustrial level, total carbon emission, and the net output of abatement cost and climate damages.

The results in Table 6.3 show that if uncertainty is taken into account, there is a significant impact on all performance measures and substantial change on the average values of all performance measures. However, the independence case results are slightly different than those of the pairwise assessment case. In this instance, dependence among variables by using pairwise assessments may have very little impact on the average values of performance measures.

Also, if we analyze the results for carbon prices, we can easily say that the deterministic case leads to an overestimate of carbon prices for all scenarios. Moreover, the average value of global temperature increase associated with the Limit 2ºC and Stern scenarios is lower than 2.0 ºC above the pre-industrial norm under deterministic and uncertain cases; however, there is a 0.4 ºC

difference, which is huge for our earth and for us, between the average temperature increase under deterministic case and the average temperature increase under uncertainty.

Table 6.3: Mean and Standard Deviation of Each Performance Measure under Each Policy 2010-2300 for Deterministic, Independence, and Dependence Cases

Scenarios	Method	Carbon Price[1]		Carbon Concentration[2]		Temperature[3]		Total Carbon Emission[4]		Net Output[5]	
		Mean	St.Dev	Mean	St.Dev	Mean	St.Dev	Mean	St.Dev	Mean	St.Dev
Base	Deterministic	34.33	36.72	893.09	246.77	4.78	1.96	52.65	35.98	1040.83	731.69
	Pairwise Max CRE	34.00	32.77	967.87	289.95	3.71	1.43	64.69	35.43	985.12	689.34
	Independence	33.98	32.82	984.26	295.90	3.83	1.44	71.73	28.51	1123.18	829.76
Optimal	Deterministic	112.22	52.99	476.55	41.23	2.60	0.62	13.32	20.45	1171.02	860.20
	Pairwise Max CRE	86.76	41.15	440.35	36.72	2.35	0.70	21.12	17.77	1039.70	799.05
	Independence	91.66	41.75	466.72	40.87	2.44	0.71	35.92	21.55	1126.36	831.78
Limit 2C	Deterministic	143.83	64.95	389.67	22.08	1.81	0.27	6.63	16.95	1177.75	859.55
	Pairwise Max CRE	128.14	50.79	347.94	56.23	1.44	0.39	9.74	10.39	1066.94	823.19
	Independence	134.26	49.48	391.67	39.74	1.48	0.40	12.36	11.17	1145.03	848.00
Stern	Deterministic	155.64	65.65	370.25	19.24	1.61	0.20	5.42	17.49	1290.23	941.65
	Pairwise Max CRE	120.40	53.65	343.70	30.02	1.40	0.56	6.59	9.13	1167.55	889.50
	Independence	130.17	54.31	355.78	37.50	1.52	0.53	9.45	11.97	1302.19	937.06
Copenhagen	Deterministic	100.28	40.29	539.68	46.12	3.02	0.85	19.42	10.16	1151.36	836.98
	Pairwise Max CRE	79.21	34.79	519.21	44.53	2.94	0.79	19.19	11.34	1103.49	795.44
	Independence	91.97	35.39	547.35	52.19	2.97	0.81	22.20	11.34	1122.52	836.18
		(1) $ per ton		(2) ppm		(3) Degrees Celcius		(4) GTCO2		(5) $ trillion	

CHAPTER 7: CONCLUSIONS

7.1. Summary of the Results

Estimating the functional form of multivariate probability distributions with partial information is important because the decision maker is often unable or unwilling to provide precise information, but the decision maker may be able to assess certain relations among subsets of attributes. So, in decision analysis literature, most of the real world problems are often solved on the basis of given certain conditions and assumptions. Most of the decision problems assume independence among variables. However, in real life decisions, the decision maker must make decisions which involve trade-offs and uncertainties among variables. Assuming independence simplifies the approximation process, but the accuracy of the decisions reduces due to the loss of information. If we know more information about decision maker's preferences, this helps to approximate or elicit more accurate joint probability distribution, but eliciting more information from the decision maker is difficult, time consuming and expensive. So, our main goal in this work is to address the problems where partial information about the decision situations is known.

Our first contribution, in Chapter 3, is to approximate joint probability distributions of a set of discrete random variables using a product of second order conditional and marginal distributions based on cumulative residual entropy. We construct optimum first order tree approximation of the joint distribution if its dependence tree has the maximum sum of cumulative residual mutual information pairs.

Our second contribution, in Chapter 4, is to propose an approximation method similar to maximum entropy principle to construct representative joint probability distributions from its lower order assessments by using maximum cumulative residual entropy approach.

Chapter 5 compares several approximation methods to find the best approximate probability distribution based on the given information. . The objectives of this chapter is to test the accuracy of different approximations of joint distributions with respect to the true distribution from the set of all possible distributions that match available information.

Chapter 6 takes into account the climatic and economic uncertainties because climate change, long-term economic development, and their interactions are highly uncertain. Climatic and economic uncertainties are integrated into the DICE model, and to find out what will be the cumulative impact of integrating uncertainty on climate change by applying CRE to the DICE model.

7.2. Future Work

The existing research can be extended in several ways:

- A new method to compare approximate probability distributions or utility functions with respect to the true distribution from the set of all possible distributions that match same available information can be developed.
- It is essential to construct utility functions for the decision maker if only the decision maker's partial preference information. Abbas used the analogy between probability and utility and introduced the utility density functions, implementing it in his maximum entropy formulation (Abbas, 2002 and 2006). In a similar way, the maximum CRE entropy formulation can be further studied to apply utility functions. One future research direction would be to construct an alternative measure of uncertainty by using the analogy between utility and probability to incorporate the multiattribute utility function into the cumulative residual entropy.

- In this study, first order dependence trees with CRE and maximum entropy approximations can only be used with CRE Kullback Leibler divergence. There are also several other measures in the literature like Bregman divergence (Bregman, 1967) and Csiszar divergence (1963 and 1967). Another future research direction would be to use other divergences such as Bregman divergences or Csiszar divergence.

- Another future direction of this research involves taking climate change into account in buying a new car and analyzing the effects of climate change on buying decisions: e.g., whether climate change has a major impact on a decision maker's decision or decision maker is focused on the price of the car and disregards altogether its carbon footprint.
- Abbas (2002 and 2006) uses the analogy between probability and utility to produce an entropy definition for utility functions. But a formulation similar to mutual information and KL-divergence in probability has not been extended for utility functions. By using the analogy between utility and probability, mutual information and KL-divergence measures can be extended to utility functions using the Cumulative Residual entropy measure.

REFERENCES

Abbas, A.E. 2002a. An entropy approach for utility assignment in decision analysis. In 22nd International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering. 659(1), 328-338. AIP Conference Proceedings.

Abbas, A.E. 2002b. Entropy Methods for Univariate Distributions in Decision Analysis. In 22nd International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering. 659(1), 339-349. AIP Conference Proceedings.

Abbas, A.E. 2004. Entropy Methods for Adaptive Utility Elicitation. IEEE Transactions on Systems, Science and Cybernetics. 34 (2), 169-178.

Abbas, A.E. 2005. Maximum Entropy Distributions between Upper and Lower Bounds. In 25nd International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering. 803, 25-42. AIP Conference Proceedings.

Abbas, A.E. 2006. Entropy methods for joint distributions in decision analysis. Engineering Management, IEEE Transactions on, 53(1), 146-159.

Abbas, A.E. 2006. Maximum entropy utility. Operations Research 54(2): 277-290.

Abbas, A.E. 2007. Moments of Utility Functions and Their Applications; European Journal of Operational Research, 180: 378-395; 2007.

Abbas, A.E. 2009. Multiattribute Utility Copulas. Operations Research, 57 (6), 1367-1383.

Abbas, A.E. 2010. Constructing Multiattribute Utility Functions for Decision Analysis. INFORMS Tutorials in Operations Research, editor Hasenbein, J. Hanover, Maryland, 62-98.

Abbas, A.E. 2010. General decompositions of multiattribute utility functions with partial utility independence. Journal of Multicriteria Decision Analysis, 17, 37-59.

Abbas, A.E. 2013. Utility Copula Functions Matching all Boundary Assessments. Operations Research, 61(2), 359-371.

Abbas, A.E. and Bell, D.E. 2011. One-Switch Independence for Multiattribute Utility Functions. Operations Research, 59 (3), 764-771.

Abbas, A.E. and Bell, D.E. 2012. One-Switch Conditions for Multiattribute Utility Functions. Operations Research, 60 (5), 1199-1212.

Abbas, A.E. and Chudziak, J. 2013. One-Switch Utility Functions with Annuity Payments. Applied Mathematics and Computation, 219 (14), 7699-7710.

Abbas, A.E. and Howard, R.A. 2005. Attribute Dominance Utility. Decision Analysis, 2(4), 185-206.

Abbas, A.E., Budescu, D.V., and Gu, R. 2010. Assessing Joint Distributions with Isoprobability Contours; Management Science, 56(6): 997-1011.

Anthoff, D., and Tol, R.S. 2009. The impact of climate change on the balanced growth equivalent: an application of FUND. Environmental and Resource Economics, 43(3), 351-367.

Babonneau, F., Haurie, A., Loulou, R., and Vielle, M. 2012. Combining stochastic optimization and Monte Carlo simulation to deal with uncertainties in climate policy assessment. Environmental Modeling and Assessment, 17, 51–76.

Baker, E. 2009. Optimal policy under uncertainty and learning about climate change: a stochastic dominance approach. Journal of Public Economic Theory, 11(5), 721–747.

Baker, E., and Shittu, E. 2008. Uncertainty and endogenous technical change in climate policy models. Energy Economics, 30(6), 2817–2828.

Baratpour, S., and Rad, A.H. 2012. Testing goodness-of-fit for exponential distribution based on cumulative residual entropy. Communications in Statistics-Theory and Methods. 41(8), 1387-1396.

Barron, F.H., and Barrett, B.E. 1996. Decision quality using ranked attribute weights. Management Science, 42(11), 1515-1523.

Bedford, T., Cooke, R.M. 2001. Probability density decomposition for conditionally dependent random variables modeled by vines. Annals of Mathematics and Artificial intelligence, 32(1-4), 245-268.

Bell, D.E. 1988. One-switch utility functions and a measure of risk. Management Science. 34(12) 1416-1424.

Bordley, R.F. 1982. A Multiplicative Formula for Aggregating Probability Assessments. Management Science, 28(10), 1137-1148.

Boutilier, C., Brafman, R., Domshlak, C., Hoos, H., and Poole, D. 2004. CP-nets: A tool for representing and reasoning about conditional ceteris paribus preference statements. Journal of Artificial Intelligence Research (JAIR), 21, 135–191.

Bregman, L.M. 1967. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. USSR computational mathematics and mathematical physics, 7(3), 200-217.

Brook, D., and Evans, D.A. 1972. An approach to the probability distribution of CUSUM run length. Biometrika, 59(3), 539-549.

Brown, D.T. 1959. A note on approximations to discrete probability distributions. Information and Control, 2(4), 386-392.

Chajewska, U., Koller, D. and Parr, R. 2000. Making Rational Decisions using Adaptive Utility Elicitation. Proceedings of the 17th National Conference on Artificial Intelligence. 363-369.

Chan, M.M. 1971. Technical Note—System Simulation and Maximum Entropy. Operations Research, 19(7), 1751-1753.

Chow, C.K., and Liu, C.N. 1968. Approximating discrete probability distributions with dependence trees. IEEE Transactions on Information Theory. vol IT 14, no 3, 462-467.

Clemen, R.T. and Reilly, T. 1999. Correlations and copulas for decision and risk analysis. Management Science., vol 45(2), 208–224.

Clemen, R.T., Fischer, G.W., and Winkler, R.L. 2000. Assessing dependence: Some experimental results. Management Science 46(8), 1100-1115.

Clemen, R.T., Fischer, G.W., and Winkler, R.L. 2000. Assessing dependence: Some experimental results, Management Science., vol. 46, no. 8, pp. 1100–1115

Cover, T. M., and Thomas, J. 1991. Elements of information theory. Wiley-interscience.

Csiszar, I. 1963. Eine Informations Theoretische Ungleichung und ihre Anwendung auf den Beweisder Ergodizitat von marko_schen ketten. Magyar. Tud. Akad. Mat. Kutato Int. Kozl, Vol. 8, 85-108.

Csiszar, I. 1967. On Topological Properties of f-Divergences. Studia Math. Hungarica, Vol. 2, 329-339.

Dai, Y.S., Xie, M., Long, Q., and Ng, S.H. 2007. Uncertainty analysis in software reliability modeling by Bayesian analysis with maximum-entropy principle. Software Engineering, IEEE Transactions on, 33(11), 781-795.

European Environment Agency (EEA) http://www.eea.europa.eu/

Fishburn, P.C., and Keeney, R.L. 1974. Seven Independence Concepts and Continuous Multiattribute Utility Functions. Journal of Mathematical Psychology, 11(3), 294-327.

Fishburn, P.C., and Keeney, R.L. 1975. Generalized utility independence and some implications. Operations Research, 23(5), 928-940.

Fisher, A., and Narain, U. 2003. Global warming, endogenous risk, and irreversibility. Environmental and Resource Economics, 25(4), 395–416.

Funke, M., and Paetz, M. 2011. Environmental policy under model uncertainty: a robust optimal control approach. Climatic Change, 107, 225–239.

Gull, S.F., and Skilling, J. 1984. Maximum entropy method in image processing. Communications, Radar and Signal Processing, IEE Proceedings F, 131(6), 646-659.

Hadfi, R., and Ito, T. 2012. Asymptotic Maximum Entropy Principle for Utility Elicitation under High Uncertainty and Partial Information. Proceedings of the Twenty-Fifth International Florida Artificial Intelligence Research Society Conference

Haurie, A., Tavoni,M., and van der Zwaan, B.C.C. 2012. Modeling uncertainty and the economics of climate change: recommendations for robust energy policy. Environmental Modeling and Assessment, 17, 1–5.

Hazen, G.B. 1986. Partial information, dominance, and potential optimality in multiattribute utility theory. Operations research, 34(2), 296-310.

Heal, G., and Kristrom, B. 2002. Uncertainty and climate change. Environment and Resource Economics, 22, 3–39.

Herfert, M., and La Mura, P. 2004. Estimation of Consumer Preferences via Ordinal Decision-Theoretic Entropy. Working Paper 64, Leipzig Graduate School of Management (HHL).

Hope, C. 2006. The marginal impact of CO2 from PAGE2002: An integrated assessment model incorporating the IPCC's five reasons for concern. Integrated assessment, 6(1).

Howard R.A., and Matheson J.E. 1984. Influence diagrams, In the Principles and Applications of Decision Analysis, Vol. II. Strategic Decisions Group, Menlo Park: California, USA.

Howard, R.A. 1964. Decision Analysis: Applied Decision Theory. Proceedings of the Fourth International Conference on Operational Research, 55-71, Wiley-Interscience.

Howard, R.A. 1989. Knowledge maps. Management Science, 35(8), 903-922.

Høyland, K., and Wallace, S. W. 2001. Generating scenario trees for multistage decision problems. Management Science, 47(2), 295-307.

Hu, Z., Cao, J., and Hong, L.J. 2012. Robust simulation of global warming policies using the DICE model. Management Science, 58(12), 2190-2206.

Intergovernmental Panel on Climate Change (IPCC) http://www.ipcc.ch/

Jaynes, E.T. 1957a. Information theory and statistical mechanics. Physical review, 106(4), 620.

Jaynes, E.T. 1957b. Information theory and statistical mechanics. II. Physical review, 108(2), 171.

Keefer, D. 2004. The underlying event model for approximating probabilistic dependence among binary events. IEEE Transactions on Engineering Management, 51(2), 173–182.

Keeney, R.L. 1971. Utility independence and preferences for multiattributed consequences. Operations Research, 19(4), 875-893.

Keeney, R.L. 1972. Concepts of Independence in Multiattribute Utility Theory (No. TR-78). Massachusetts Institute of Tech Cambridge Operations Research Center.

Keeney, R.L. 1973. A decision analysis with multiple objectives: The Mexico City airport. The Bell Journal of Economics and Management Science, 101-117.

Keeney, R.L. 1973. Risk independence and multiattributed utility functions. Econometrica: Journal of the Econometric Society, 27-34.

Keeney, R.L. 1974. Multiplicative utility functions. Operations Research, 22(1), 22-34.

Keeney, R.L. and Raiffa, H. 1976. Decisions with Multiple Objectives. Wiley: New York.

Kelly, D.L. 2007. Using copulas to model dependence in simulation risk assessment. ASME International Mechanical Engineering Congress and Exposition, IMECE2007-41284, pp. 81-89.

Keppo, I., and van der Zwaan, B. 2012. The impact of uncertainty in climate targets and CO2 storage availability on long-term emissions abatement. Environmental Modeling and Assessment, 17, 177–191.

Kirkwood, C.W., and Sarin, R.K. 1985. Ranking with partial information: A method and an application. Operations Research, 33(1), 38-48.

Kirkwood, J.G. 1935. Statistical mechanics of fluid mixtures. The Journal of Chemical Physics, 3, 300.

Ku, H.H., and Kullback, S. 1969. Approximating discrete probability distributions. IEEE Transactions on Information Theory, 15(4), 444–447.

Kullback, S. and Leibler, R.A. 1951. On information and sufficiency. Ann. Math. Statist., vol. 2, pp. 79–86

Kunreuther, H., Heal, G., Allen, M., Edenhofer, O., Field, C.B., and Yohe, G. 2013. Risk management and climate change. Nature Climate Change, 3, 447–450.

Laplace P. 1774. Memoire sur la Probabilite des Causes par les Evenements. l'Academie Royale des Sciences, 6, 621-656. English translation by S.M. Stigler in 1986 as "Memoir on the Probability of the Causes of Events" in Statistical Science, 1(3), 359-378.

Lewis, P.M. 1959. Approximating probability distributions to reduce storage requirements. Information and control, 2(3), 214-225.

Lowell, D.G. 1994. Sensitivity to relevance in decision analysis. Ph. D. thesis, Engineering and Economic Systems, Stanford University, Stanford,CA.

Mackenzie, G.R. 1994. Approximately maximum entropy multivariate distributions with specified marginals and pairwise correlations. PhD Dissertation, Univ. of Oregon, Oklahoma

Manne, A.S., Mendelsohn, R., and Richels, R. 1995. MERGE: A Model for Evaluating Regional and Global Effects of GHG Reduction Policies. Energy Policy, 23(1):17-34.

Manne, A.S., Richels, R.G., 2004. MERGE: An integrated assessment model for global climate change. Stanford University, Stanford, CA, USA.

Matheson, J.E., and Howard, R.A. 1968. An introduction to decision analysis. In The Principles and Applications of Decision Analysis. vol. I, Howard RA, Matheson JE (eds). Strategic Decisions Group: Menlo Park, CA,USA.

McInerney, D., and Keller, K. 2008. Economically optimal risk reduction strategies in the face of uncertain climate thresholds. Climatic Change, 91, 29–41.

McNamee P, and Celona, J. 1990. Decision Analysis with Supertree, 2nd ed. (The Scientific Press, South San Francisco).

Meila, M. 1999. An Accelerated Chow and Liu Algorithm: Fitting Tree Distributions to High-Dimensional Sparse Data. Proceedings of the Sixteenth International Conference on Machine Learning, pp. 249–257.

Miller, A.C., and Rice, T.R. 1983. Discrete approximations of probability distributions. Management Science, 29(3), 352-362.

Montiel, L.V., and Bickel, J.E. 2012. A simulation-based approach to decision making with partial information. Decision Analysis, 9(4), 329-347.

Newbold, S.C., and Daigneault, A. 2009. Climate response uncertainty and the benefits of greenhouse gas emissions reductions. Environmental and Resource Economics, 44, 351–377.

Newbold, S.C., Griffiths, C., Moore, C., Wolverton, A., and Kopits, E. 2010. The ‘‘social cost of carbon’’ made simple. Working Paper Series Working Paper # 10-07, National Center for Environmental Economics, U.S. Environmental Protection Agency, Washington, DC

Nordhaus, W.D. (2013). The climate casino: Risk, uncertainty, and economics for a warming world. Yale University Press.

Nordhaus, W.D. 1977a. Strategies for the control of carbon dioxide. Cowles Foundation discussion paper no. 443.

Nordhaus, W.D. 1977b. Economic growth and climate: the carbon dioxide problem. The American Economic Review 67(1):341-346.

Nordhaus, W.D. 1994. Managing the global commons: the economics of climate change. Cambridge, MA: MIT press.

Nordhaus, W.D. 2007. The challenge of global warming: Economic models and environmental policy. New Haven, Connecticut: Yale University.

Nordhaus, W.D. 2008. A question of balance: Weighing the options on global warming policies. Yale University Press.

Nordhaus, W.D. and Boyer, J.G. 2000. Warming the World: Economic Models of Global Warming, Cambridge: The MIT Press.

Nordhaus, W.D., Sztorc, P. 2013. DICE 2013R: Introduction and User's Manual. (http://www.econ.yale.edu/~nordhaus/homepage/documents/DICE_Manual_103113r2.pdf)

Parmesan, C., and Yohe, G. 2003. A globally coherent fingerprint of climate change impacts across natural systems. Nature, 421(6918), 37-42.

Pearl, J. 1988. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann: San Mateo, CA.

Peck, S., and Teisberg, T. 1993. Global warming uncertainties and the value of information—an analysis using CETA. Resource and Energy Economics, 15(1), 71–97.

Peterson, S. 2006. Uncertainty and economic analysis of climate change: a survey of approaches and findings. Environmental Modeling and Assessment, 11(1), 1–17.

Pizer, W.A. 1999. The optimal choice of climate change policy in the presence of uncertainty. Resource and Energy Economics, 21, 255–287.

Rao, M. 2005. More on a new concept of entropy and information. Journal of Theoretical Probability, 18(4), 967-981.

Rao, M., Chen,Y., Vemuri, B.C., and Wang, F., 2004. Cumulative residual entropy: a new measure of information. IEEE Trans. Inform. Theory 50(6), 1220–1228

Reilly, T. 2000. Sensitivity analysis for dependent variables. Decision Sciences 31(3), 551-572.

Richard, S.F. 1975. Multivariate risk aversion, utility independence and separable utility functions. Management Science, 22(1), 12-21.

Schmidt, M. G. W., Lorenz, A., Held, H., and Kriegler, E. 2011. Climate targets under uncertainty: challenges and remedies. Climatic Change: Letters, 104(3), 783–791

Shannon, C.E. 1948. A mathematical theory of communication. Bell Syst. Tech. J., 27, 379-423 and 623–656.

Smith, A.E., Ryan, P.B., and Evans, J.S. 1992. The effect of neglecting correlations when propagating uncertainty and estimating the population distribution of risk. Risk Analysis 12(4), 467-474.

Smith, J.E. 1993. Moment methods for decision analysis. Management science, 39(3), 340-358.

Sokolov, A., Stone, P., Forest, C., Prinn, R., Sarofim, M., Webster, M., Paltsev, S., Schlosser, C.A., Kicklighter, D., Dutkiewicz, S., Reilly, J., Wang, C., Felzer, B., Melillo, J., and Jacoby, H. 2009 Probabilistic forecast for 21st century climate based on uncertainties in emissions (without policy) and climate parameters. J Climate 22 (19):5175–5204

Stern . N., Peters, S., Bakhshi, V., Bowen, A., Cameron, C., Catovsky, S., Crane, D., Cruickshank, S., Dietz, S., Edmondson, N., Garbett, S., Hamid, L., Hoffman, G., Ingram, D., Jones, B., Patmore, N., Radcliffe, H., Sathiyarajah, R., Stock, M., Taylor, C., Vernon, T., Wanjie, H., and Zenghelis, D. 2006. The Economics of Climate Change. Cabinet Office – HM Treasury. Cambridge University Press.

Stern, N. 2008. The economics of climate change. The American Economic Review, Papers and Proceedings of the One Hundred Twentieth Annual Meeting of the American Economic Association, 98 (2), 1-37.

Sutcu, M., and Abbas A.E. 2014. Approximating Joint Probability Distribution with Dependence Trees Using Cumulative Residual Entropy. 34th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering (forthcoming)

Thomas, M.U. 1979. Technical Note—A Generalized Maximum Entropy Principle. Operations Research, 27(6), 1188-1196.

Tol, R.S.J. 1997. On the optimal control of carbon dioxide emissions: an application of FUND. Environmental Modeling & Assessment, 2(3), 151-163.

Tol, R.S.J. 2002a. New estimates of the damage costs of climate change, Part I: Benchmark estimates. Environmental and Resource Economics 21 (1): 47-73.

Tol, R.S.J. 2002b. New estimates of the damage costs of climate change, Part II: Dynamic estimates. Environmental and Resource Economics 21 (1): 135-160.

United States Environmental Protection Agency (EPA) http://www.epa.gov/

Walther, G.R., Post, E., Convey, P., Menzel, A., Parmesan, C., Beebee, T. J.C., Fromentin, J.M., Guldberg, O.H., and Bairlein, F. 2002. Ecological responses to recent climate change. Nature, 416(6879), 389-395.

Wang, F., Vemuri, B. C., Rao, M., and Chen, Y. 2003. A new & robust information theoretic measure and its application to image alignment. In Information Processing in Medical Imaging. 388-400. Springer Berlin Heidelberg.

Wang, T., and Dyer, J.S. 2012. A copulas-based approach to modeling dependence in decision trees. Operations Research, 60(1), 225–242.

Weber, M. 1987. Decision making with incomplete information. European Journal of Operational Research, 28(1), 44-57.

Webster, M., Sokolov, A. P., Reilly, J. M., Forest, C. E., Paltsev, S., and Schlosser, A. 2012. Analysis of climate policy targets under uncertainty. Climatic Change, 112, 569–583.

World Nature Organization (WNO) http://www.wno.org/