ADVANCEMENTS IN THE ELICITATION, AGGREGATION, AND FORECASTING OF PROBABILITY DISTRIBUTIONS UNDER TIME CONSTRAINTS

A Dissertation
Presented to the Faculty of
Claremont Graduate University
and
San Diego State University

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
in
Computational Science - Statistics

by
Jonathan Louis Wilson
Spring 2013
This dissertation has been duly read, reviewed, and critiqued by the Committee listed below, which hereby approves the manuscript of Jonathan Louis Wilson as fulfilling the scope and quality requirements for meriting the degree of Doctor of Philosophy.

Kristin Duncan, Chair
Department of Mathematics and Statistics, San Diego State University
Assistant Professor

Richard Levine
Department of Mathematics and Statistics, San Diego State University
Professor

Barbara Bailey
Department of Mathematics and Statistics, San Diego State University
Assistant Professor

John Angus
Institute of Mathematical Sciences, Claremont Graduate University
Professor

Allon Percus
Institute of Mathematical Sciences, Claremont Graduate University
Associate Professor
Copyright © 2013
by
Jonathan Louis Wilson
All Rights Reserved
ABSTRACT OF THE THESIS

Advancements In the Elicitation, Aggregation, and Forecasting of Probability Distributions Under Time Constraints
by
Jonathan Louis Wilson
Doctor of Philosophy in Computational Science - Statistics
Claremont Graduate University and San Diego State University, 2013

Opinion pooling involves the process of combining the beliefs of multiple individuals into a single consensus which best represents the beliefs of the group. Instances where these methods are often applied occur when historical data is difficult or even impossible to collect. The elicitation of individual beliefs, aggregation into a consensus, and forecasting from the consensus distribution are all important processes in the field of opinion pooling. This thesis develops new techniques in each of these areas, increasing the possibility for more accurate decision making under uncertainty.

Currently accepted elicitation techniques have yet to incorporate technological advancements made in the field of visual computing. To help experts exhibit their beliefs more accurately, new visually aided computer software is developed and tested. Individual opinions are expressed as probability distributions constructed with help from a moderator. An experiment is conducted to test whether the newly developed software aids in the ability to better predict the outcome of unknown events across different probability distributions.

Constructing a consensus probability distribution relies not only on accurate individual elicitations, but also on the formulaic algorithm used to combine the multiple probability distributions. Popular techniques used today to combine these distributions suffer from some debilitating drawbacks. Investigating the problem through probability distance and divergence measures, new consensus pooling operators are proposed as opposed to the currently used Kullback-Leibler distance measure.
Finally, new theory on applying opinion pooling methodology under time constraints is developed. Often decisions need to be made before a specific deadline arrives. In these instances, opinions are not only influenced by available evidence, but also by the amount of time remaining before the decision must be made. With very little research completed in this area, new algorithms and theory are proposed to better incorporate the time dimension into the opinion pooling framework. A thorough example of elicitations and predictions made through prediction markets implements the newly developed theory.
DEDICATION

Dedicated to my loving wife Emily.
It ain’t what you don’t know that gets you into trouble.
It’s what you know for sure that just ain’t so.

– Mark Twain
ACKNOWLEDGEMENTS

First, I would like to thank my committee chair Dr. Kristin Duncan. My introduction into this type of research began over five years ago when Kristin suggested I complete a unique project involving risk analysis for her Bayesian statistics course. Little did I know at the time that this suggestion would set me down a long road ahead with many new research ideas explored along the way. Working together on research for SPAWAR, we were able to implement our ideas in new and creative ways. I am grateful for all the insights and discussions we had over the years, and it is certain this work would not be where it is today with her helpful guidance.

I would also like to give special thanks to Dr. Richard Levine who has had a major influence on the choices in my academic and professional career. As my Masters level studies at SDSU were concluding, I remember discussing with Rich my future academic and professional options upon graduation. I knew that I wanted to continue my academic studies somewhere, and that I wanted to continue to learn and implement new computational statistical methods rather than just focusing on developing statistical theory. By the end of our conversation he had me convinced that entering the brand new joint doctoral program with SDSU and Claremont Graduate University was the best option. Upon completion, I wholeheartedly agree with that assessment and am proud to have completed the program.

Additionally, I would like to thank my committee members Dr. Barbara Bailey, Dr. John Angus, and Dr. Allon Percus for their insights over the years. Each has provided significant impact and guidance to the work presented in this dissertation as well as numerous outside learnings.

Finally, I would like to thank Dr. Jose Castillo and the Computational Science Research Center in providing a rich academic environment and for help in funding the costs associated with the joint SDSU/CGU Ph.D. program.
**TABLE OF CONTENTS**

<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>iv</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>viii</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>xii</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>xiv</td>
</tr>
<tr>
<td>1 INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>2 ELICITATION</td>
<td>3</td>
</tr>
<tr>
<td>2.1 Context of Elicitation</td>
<td>3</td>
</tr>
<tr>
<td>2.2 Types of Elicitation</td>
<td>6</td>
</tr>
<tr>
<td>2.3 Evaluating Elicitations Through Scoring Rules</td>
<td>8</td>
</tr>
<tr>
<td>2.4 Elicitation Experiment</td>
<td>10</td>
</tr>
<tr>
<td>2.4.1 Initial Experimental Design</td>
<td>11</td>
</tr>
<tr>
<td>2.4.2 Follow Up Experiment</td>
<td>14</td>
</tr>
<tr>
<td>2.4.3 Follow Up Experiment Results</td>
<td>15</td>
</tr>
<tr>
<td>2.5 Analysis of Competing Hypotheses</td>
<td>19</td>
</tr>
<tr>
<td>3 Combining Probability Distributions</td>
<td>26</td>
</tr>
<tr>
<td>3.1 Overview</td>
<td>26</td>
</tr>
<tr>
<td>3.2 Axiomatic Approaches</td>
<td>27</td>
</tr>
<tr>
<td>3.2.1 LinOP</td>
<td>29</td>
</tr>
<tr>
<td>3.2.2 LogOP</td>
<td>29</td>
</tr>
<tr>
<td>3.2.3 Power Mean</td>
<td>31</td>
</tr>
</tbody>
</table>
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 2.1</td>
<td>Linguistic probabilities and translations. [36]</td>
<td>4</td>
</tr>
<tr>
<td>Table 2.2</td>
<td>Questionnaire for Experiment 1.</td>
<td>13</td>
</tr>
<tr>
<td>Table 2.3</td>
<td>ACH matrix for stadium example.</td>
<td>24</td>
</tr>
<tr>
<td>Table 2.4</td>
<td>Results for ACH extensions example.</td>
<td>24</td>
</tr>
<tr>
<td>Table 3.1</td>
<td>The change in LogOP when pooling (0,0,0.30,0.6,0.95) through the replacement of zero values.</td>
<td>31</td>
</tr>
<tr>
<td>Table 3.2</td>
<td>Divergence and distance measures investigated.</td>
<td>36</td>
</tr>
<tr>
<td>Table 3.3</td>
<td>Requirements for the different types of metrics.</td>
<td>37</td>
</tr>
<tr>
<td>Table 3.4</td>
<td>Closed form solutions for divergence measures.</td>
<td>50</td>
</tr>
<tr>
<td>Table 4.1</td>
<td>Consensus operators analyzed using experiment data.</td>
<td>58</td>
</tr>
<tr>
<td>Table 4.2</td>
<td>Pooled distribution (P) from $Q_1 = [0.50, 0.50]$ and $Q_2$ under equal weighting.</td>
<td>60</td>
</tr>
<tr>
<td>Table 4.3</td>
<td>Ordered levels of uncertainty for distance measures when pooling $Q_1 = [0.50, 0.50]$ and $Q_2 = [x, 1-x]$ under equal weighting.</td>
<td>62</td>
</tr>
<tr>
<td>Table 4.4</td>
<td>Pooled distribution (P) when $Q_1 = Q_2 = Q_3 = Q_4 = [0.50, 0.50]$, $Q_5 = [1, 0]$ under equal weighting.</td>
<td>62</td>
</tr>
<tr>
<td>Table 4.5</td>
<td>Pooled distribution (P) when $Q_1 = [0.50, 0.50]$, $Q_2 = Q_3 = Q_4 = [0.99, 0.01]$, $Q_5 = [1, 0]$ under equal weighting.</td>
<td>63</td>
</tr>
<tr>
<td>Table 4.6</td>
<td>Operator rankings from Experiment 1.</td>
<td>67</td>
</tr>
<tr>
<td>Table 4.7</td>
<td>Operator rankings from Experiment 2.</td>
<td>69</td>
</tr>
<tr>
<td>Table 4.8</td>
<td>Operator rankings from Experiment 3.</td>
<td>71</td>
</tr>
<tr>
<td>Table 4.9</td>
<td>Consensus operator percentiles across all three experiments.</td>
<td>72</td>
</tr>
<tr>
<td>Table 5.1</td>
<td>Comparison of bounded and unbounded results where $X_{31} = 12000$, $L = 11400$, $H = 13000$, $\mu = 0$ and $\sigma = 75$.</td>
<td>97</td>
</tr>
</tbody>
</table>
Table 5.2. Comparison of unbounded, fixed bounds, and time dependent bounds
where $X_{31} = 12000$, $L = 11400$, $H = 13000, \mu = 0$ and $\sigma = 75$. .................. 101

Table 6.1. A hypothetical windowing procedure for $t = 2$, where the range is
repeatedly calculated over randomly chosen blocks until the distribution
$W_2$ is created. .......................................................... 132
LIST OF FIGURES

Figure 2.1. Screen shot of the ELICIT GUI tool developed in R .................. 12
Figure 2.2. Elicited distributions from Experiment 1 - Question 1 .............. 16
Figure 2.3. Elicited distributions from Experiment 1 - Question 2 .............. 17
Figure 2.4. Elicited distributions from Experiment 1 - Question 3 .............. 18
Figure 2.5. Marginal posterior distributions for $p$ and proposal densities used
for importance sampling. ........................................................................ 25
Figure 3.1. Calculating the constant $c$ by (3.6) when pooling (0, 0.06, 0.11). ... 34
Figure 4.1. Results of first opinion ($p_1$) when pooling (0.5, 0.5) and (x,1-x) for
each consensus operator. ........................................................................ 61
Figure 4.2. Example of LinOP and the incorrect averaging of parameters when
$Q_1 = \text{Beta}(1,3)$ and $Q_2 = \text{Beta}(3,1)$. .................................................. 64
Figure 4.3. Consensus operator results when pooling two normal distributions. ... 65
Figure 5.1. General expectations of the probability on the outcome of a dichoto-
mous event as a deadline approaches. ...................................................... 76
Figure 5.2. A demonstration of information flow on the probability of termina-
tion positions. ......................................................................................... 78
Figure 5.3. Examples of parameter influence from equation (5.6) when $\tau = 100$. .... 85
Figure 5.4. Random Walk versus Bounded Random Walk from equation (5.5) ........ 87
Figure 5.5. Examples of parameter influence on equation (5.8), where $L = 95$
and $H = 105$. ......................................................................................... 89
Figure 5.6. An example showing similar bounded random walks calculated
from both methods. .................................................................................. 91
Figure 5.7. An example of a large overcorrection after the random walk escapes
the bounds. ......................................................................................... 92
Figure 5.8. A diverging random walk from equation 5.5. ................................ 93
Figure 5.9. Histograms of the terminating values of 10,000 simulated random walks from both methods. ............................................................. 95

Figure 5.10. Histogram of 50,000 simulations for the bounded stock market example. .... 96

Figure 5.11. Example of a random walk with time dependent bounds. .................... 100

Figure 5.12. Histogram of 50,000 random walk simulations with time dependent bounds. ............................................................................ 101

Figure 5.13. Car preference example with $S$ as a 3x3 identity matrix. ...................... 108

Figure 5.14. Parameter results for $\alpha$ from 25,000 simulations after 5 time steps. ....... 113

Figure 5.15. Parameter results for $\alpha$ from 1,000 simulations after 200 time steps. ........ 114

Figure 5.16. Cumulative expected probabilities for each hypothesis from a single simulation with time dependent evidence weights.......... 115

Figure 5.17. Parameter results for $\alpha$ from 2,000 simulations after 100 time steps with time dependent evidence weights. ................................. 116

Figure 5.18. Cumulative expected probabilities for each hypothesis from a single simulation with time dependent ACH matrix evaluation values. .......... 118

Figure 5.19. Parameter results for $\alpha$ from 1,000 simulations after 100 time steps with time dependent ACH matrix evaluation values. ................................. 118

Figure 5.20. Cumulative expected probabilities for each hypothesis from a single simulation with both time dependent weights and time dependent ACH matrix evaluation values. ........................................................ 119

Figure 5.21. Parameter results for $\alpha$ from 1,000 simulations after 100 time steps with both time dependent weights and time dependent ACH matrix evaluation values. ........................................................ 120

Figure 6.1. The Intrade closing prices & volume for the prediction market on Obama health care reform. ............................................................ 123

Figure 6.2. The 55 DJIA time series, where all data to the left of the vertical bar is removed from the windowing procedure. ........................................ 127

Figure 6.3. Calibration curve for the DJIA prediction market where trades are pooled on 5 point intervals, and the diameter of the circle is proportional to sample size of trades executed. ........................................ 128
Figure 6.4. A histogram of all values of the Dow expressed as daily percentage changes on one minute intervals ($X_{it}$)......................................................... 130

Figure 6.5. The Window distributions from bootstrapping for 380, 215, 90, and 5 minutes remaining with red fits from a kernel density estimation using a reflection boundary. ......................................................... 131

Figure 6.6. Calibration curve for the DJIA prediction market using model predictions pooled at 5 point intervals. ................................................................. 135

Figure 6.7. The Dow, Intrade, and model fits for 11/30/09. ........................................... 135

Figure 6.8. The Dow, Intrade, and model fits for 02/22/10. ........................................... 136

Figure 6.9. The Dow, Intrade, and model fits for 12/07/09. ........................................... 136

Figure 6.10. The Dow, Intrade, and model fits for 01/21/10........................................... 137

Figure 6.11. The accuracy rates for all Intrade contracts and the new statistical model over all days at different time points. ........................................... 138
CHAPTER 1
INTRODUCTION

Across numerous fields, important decisions are often made under imperfect information. A decision-maker (DM), the person who actually makes the final decision, usually consults with others who he believes might possess some valuable knowledge beyond his own. The reasoning is that DMs can make a more informed decision after incorporating the opinions of other experts, hopefully leading to a higher probability of success. In many instances, quantitative data is extremely scarce and subject matter experts are the only source of available information for a DM. For instance, when trying to locate terrorists or investigate the reliability of a nuclear reactor, there is little to no historical data available for a typical statistical analysis in which repeated trials of an event can be recorded, analyzed, and predicted. When experts’ beliefs are all that is obtainable, those opinions essentially serve as data, and are used to determine the best possible decision.

Opinion pooling emerged as a legitimate research field after World War II with numerous contributions from the RAND Corporation [18]. Between WWII and the Vietnam war, numerous think tanks received unprecedented amounts of grant awards dedicated to expert research. While the aggregation of opinions was not new at the time, the science of studying uncertainty of opinions and the mathematical, psychological, and statistical aspects began to take root. With the rise of the Cold War, risk analysis became a high priority in the 1950s and 1960s with the Nuclear Regulatory Commission (NRC) which relied heavily on subjective expert opinions. Eventually NASA used some valuable opinion techniques from RAND and the NRC to help with space programs. Over the past twenty years, with the advent of greater computing power, the mathematical and statistical power in opinion pooling has strengthened.
Opinion pooling can be broken down into two distinct processes. The first is the gathering of information through the elicitation process, where experts are asked to quantify their opinions through probability distributions. With multiple techniques available to help experts in the elicitation process, it is often the case that experts will elicit significantly different distributions depending on the method chosen for the elicitation. It is key that experts’ beliefs are accurately transferred into probability distributions to ensure the final decision is made under correct assumptions. Once the experts’ opinions are accurately elicited, the second process begins with identifying the best methods to pool the opinions into a consensus. The improvement of both processes should provide DMs with greater capabilities to make more informed and better overall decisions.

The dissertation is organized as follows. Chapter 2 begins by describing the elicitation process and the challenges therein. To help minimize known biases, a GUI visualization tool is constructed in R for use by the experts. Two experiments are conducted to test if the new elicitation tool can help experts overcome these biases. Chapter 3 contains derivations of new pooling methods for optimally combining probability distributions using distance measures from the information theory field. Different statistical techniques for combining multiple elicited opinions are explored through axiomatic approaches as well as probability distance measures. Chapter 4 contains an analysis of the different pooling mechanisms using the two data sets collected from the elicitation experiments in Chapter 2 as well as a dataset provided by an outside survey competition. The evaluation serves as a basis that shows that the currently accepted methods do not always outperform the newly constructed methods. In Chapter 5 research is presented involving the problem of how the addition of time constraints affect experts’ opinions. Specifically, as a deadline approaches, how does expert opinion change due solely to the time constraints. Chapter 6 describes how prediction markets are being used as a new elicitation technique, and how the methodology derived in Chapter 5 can be applied to analyze the overall accuracy of a specific prediction market. Finally a summary in Chapter 7 discusses the new findings and benefits that should come from this research.
CHAPTER 2
ELICITATION

2.1 CONTEXT OF ELICITATION

An important initial step of opinion pooling involves the elicitation of experts’ beliefs in a quantitative form. Current popular processes are fairly simple for eliciting categorical distributions, where experts are simply asked for probability estimates across the different hypotheses. Typically experts ignore any variability around the probabilities given for each category or element of the distribution. However, for continuous distributions experts are asked to think more critically about their responses, and are typically asked to provide parameter estimates with variability surrounding those estimates. To accomplish this, experts are asked for specific quantiles corresponding to their beliefs and a distribution is fit to minimize the difference from their elicitations and a true predetermined probability distribution.

It is often difficult for experts to quantitatively express how they feel about a certain hypothesis. Table 2.1 [36] shows the difficulties in trying to numerically match verbal descriptors to probabilities. Variation in opinions can be due not only to different thoughts about what might occur, but also on how the opinion is elicited into a distribution.

This elicitation process is a very important step in the opinion pooling process. Inaccurate elicitation can lead to results which do not represent the true opinions of the experts. Proper elicitation techniques have been studied in detail ([18],[20]), however there are still large areas of research to be completed in the field. While at first it might seem to be a fairly easy task for an expert to answer what the likelihood of an event might be, significant biases are known to occur. One of the most dominant being overconfidence; where experts significantly underestimate the variance of the likelihood of the event. In general terms,
<table>
<thead>
<tr>
<th>Rank</th>
<th>Phrase</th>
<th>N</th>
<th>Mean</th>
<th>Median</th>
<th>SD</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Highly probable</td>
<td>187</td>
<td>0.89</td>
<td>0.90</td>
<td>0.04</td>
<td>0.60-0.99</td>
</tr>
<tr>
<td>2</td>
<td>Very likely</td>
<td>185</td>
<td>0.87</td>
<td>0.90</td>
<td>0.06</td>
<td>0.60-0.99</td>
</tr>
<tr>
<td>3</td>
<td>Very probable</td>
<td>187</td>
<td>0.87</td>
<td>0.89</td>
<td>0.07</td>
<td>0.60-0.99</td>
</tr>
<tr>
<td>4</td>
<td>Quite likely</td>
<td>188</td>
<td>0.79</td>
<td>0.80</td>
<td>0.10</td>
<td>0.30-0.99</td>
</tr>
<tr>
<td>5</td>
<td>Usually</td>
<td>187</td>
<td>0.77</td>
<td>0.75</td>
<td>0.13</td>
<td>0.15-0.99</td>
</tr>
<tr>
<td>6</td>
<td>Good chance</td>
<td>188</td>
<td>0.74</td>
<td>0.75</td>
<td>0.12</td>
<td>0.25-0.95</td>
</tr>
<tr>
<td>7</td>
<td>Predictable</td>
<td>146</td>
<td>0.74</td>
<td>0.75</td>
<td>0.20</td>
<td>0.25-0.95</td>
</tr>
<tr>
<td>8</td>
<td>Likely</td>
<td>188</td>
<td>0.72</td>
<td>0.75</td>
<td>0.11</td>
<td>0.25-0.99</td>
</tr>
<tr>
<td>9</td>
<td>Probable</td>
<td>188</td>
<td>0.71</td>
<td>0.75</td>
<td>0.17</td>
<td>0.01-0.99</td>
</tr>
<tr>
<td>10</td>
<td>Rather likely</td>
<td>188</td>
<td>0.69</td>
<td>0.70</td>
<td>0.09</td>
<td>0.15-0.99</td>
</tr>
<tr>
<td>11</td>
<td>Pretty good chance</td>
<td>188</td>
<td>0.67</td>
<td>0.70</td>
<td>0.12</td>
<td>0.25-0.95</td>
</tr>
<tr>
<td>12</td>
<td>Fairly likely</td>
<td>188</td>
<td>0.66</td>
<td>0.70</td>
<td>0.12</td>
<td>0.15-0.95</td>
</tr>
<tr>
<td>13</td>
<td>Somewhat likely</td>
<td>187</td>
<td>0.59</td>
<td>0.60</td>
<td>0.18</td>
<td>0.20-0.92</td>
</tr>
<tr>
<td>14</td>
<td>Better than even</td>
<td>187</td>
<td>0.58</td>
<td>0.60</td>
<td>0.06</td>
<td>0.45-0.89</td>
</tr>
<tr>
<td>15</td>
<td>Rather</td>
<td>124</td>
<td>0.58</td>
<td>0.60</td>
<td>0.11</td>
<td>0.10-0.80</td>
</tr>
<tr>
<td>16</td>
<td>Slightly more than half the time</td>
<td>188</td>
<td>0.55</td>
<td>0.55</td>
<td>0.06</td>
<td>0.45-0.80</td>
</tr>
<tr>
<td>17</td>
<td>Slight odds in favor</td>
<td>187</td>
<td>0.55</td>
<td>0.55</td>
<td>0.08</td>
<td>0.05-0.75</td>
</tr>
<tr>
<td>18</td>
<td>Fair chance</td>
<td>188</td>
<td>0.51</td>
<td>0.50</td>
<td>0.13</td>
<td>0.20-0.85</td>
</tr>
<tr>
<td>19</td>
<td>Tossup</td>
<td>188</td>
<td>0.50</td>
<td>0.50</td>
<td>0.00</td>
<td>0.45-0.52</td>
</tr>
<tr>
<td>20</td>
<td>Fighting chance</td>
<td>186</td>
<td>0.47</td>
<td>0.50</td>
<td>0.17</td>
<td>0.05-0.90</td>
</tr>
<tr>
<td>21</td>
<td>Slightly less than half the time</td>
<td>188</td>
<td>0.45</td>
<td>0.45</td>
<td>0.04</td>
<td>0.05-0.50</td>
</tr>
<tr>
<td>22</td>
<td>Slight odds against</td>
<td>185</td>
<td>0.45</td>
<td>0.45</td>
<td>0.11</td>
<td>0.10-0.99</td>
</tr>
<tr>
<td>23</td>
<td>Not quite even</td>
<td>180</td>
<td>0.44</td>
<td>0.45</td>
<td>0.07</td>
<td>0.05-0.60</td>
</tr>
<tr>
<td>24</td>
<td>Inconclusive</td>
<td>153</td>
<td>0.43</td>
<td>0.50</td>
<td>0.14</td>
<td>0.01-0.75</td>
</tr>
<tr>
<td>25</td>
<td>Uncertain</td>
<td>173</td>
<td>0.40</td>
<td>0.50</td>
<td>0.14</td>
<td>0.08-0.90</td>
</tr>
<tr>
<td>26</td>
<td>Possible</td>
<td>178</td>
<td>0.37</td>
<td>0.49</td>
<td>0.23</td>
<td>0.01-0.99</td>
</tr>
<tr>
<td>27</td>
<td>Somewhat unlikely</td>
<td>186</td>
<td>0.31</td>
<td>0.33</td>
<td>0.12</td>
<td>0.03-0.80</td>
</tr>
<tr>
<td>28</td>
<td>Fairly unlikely</td>
<td>187</td>
<td>0.25</td>
<td>0.25</td>
<td>0.11</td>
<td>0.02-0.75</td>
</tr>
<tr>
<td>29</td>
<td>Rather unlikely</td>
<td>187</td>
<td>0.21</td>
<td>0.20</td>
<td>0.10</td>
<td>0.01-0.75</td>
</tr>
<tr>
<td>30</td>
<td>Not very probable</td>
<td>187</td>
<td>0.20</td>
<td>0.20</td>
<td>0.12</td>
<td>0.01-0.60</td>
</tr>
<tr>
<td>31</td>
<td>Unlikely</td>
<td>188</td>
<td>0.18</td>
<td>0.16</td>
<td>0.10</td>
<td>0.01-0.45</td>
</tr>
<tr>
<td>32</td>
<td>Not much chance</td>
<td>186</td>
<td>0.16</td>
<td>0.15</td>
<td>0.09</td>
<td>0.01-0.45</td>
</tr>
<tr>
<td>33</td>
<td>Seldom</td>
<td>188</td>
<td>0.16</td>
<td>0.15</td>
<td>0.08</td>
<td>0.01-0.47</td>
</tr>
<tr>
<td>34</td>
<td>Barely possible</td>
<td>180</td>
<td>0.13</td>
<td>0.05</td>
<td>0.17</td>
<td>0.01-0.60</td>
</tr>
<tr>
<td>35</td>
<td>Faintly possible</td>
<td>184</td>
<td>0.13</td>
<td>0.05</td>
<td>0.16</td>
<td>0.01-0.50</td>
</tr>
<tr>
<td>36</td>
<td>Improbable</td>
<td>187</td>
<td>0.12</td>
<td>0.10</td>
<td>0.09</td>
<td>0.01-0.40</td>
</tr>
<tr>
<td>37</td>
<td>Quite unlikely</td>
<td>187</td>
<td>0.11</td>
<td>0.10</td>
<td>0.08</td>
<td>0.01-0.50</td>
</tr>
<tr>
<td>38</td>
<td>Very unlikely</td>
<td>186</td>
<td>0.09</td>
<td>0.10</td>
<td>0.07</td>
<td>0.01-0.50</td>
</tr>
<tr>
<td>39</td>
<td>Rare</td>
<td>187</td>
<td>0.07</td>
<td>0.05</td>
<td>0.07</td>
<td>0.01-0.30</td>
</tr>
<tr>
<td>40</td>
<td>Highly improbable</td>
<td>181</td>
<td>0.06</td>
<td>0.05</td>
<td>0.05</td>
<td>0.01-0.30</td>
</tr>
</tbody>
</table>
experts tend to think they know a lot more than they actually know. It is the job of the person eliciting the experts’ opinions to help minimize bias as much as possible. While some administrative techniques have provided help for DMs ([43], [42]), the overconfidence bias continues to exist no matter what elicitation method is used. To help overcome the overconfidence bias, very little has been done to utilize the advancements of computer technology. Specifically, visualization feedback of probability distributions for experts have yet to be widely introduced. Investigating the literature over the past few decades leads to only a handful of papers involving graphical elicitation ([54], [55], [25]), with most being applied to very specific fields of interest. Nevertheless, the results are promising. While previous research has focused on either single probability point estimates, or discrete probability mass functions, researchers advocate providing the elicitee with (computer) graphics that will help them in “visualizing and manipulating the relevant numerical parameters” [20].

The importance of this task is specifically noted in the literature with the following comment coming from O’Hagan [42] regarding elicitation techniques for parametric distributions.

“...Despite the large quantity of research that has been reviewed here, there are striking gaps. Ideally, methods would have been examined empirically by using them to quantify people’s opinions both in experimental tests and in real applications. Regrettably, some of the methods have never been used for either purpose, and the great majority have only been used in a single set of trials. This makes it difficult for the area of elicitation to evolve - without a reasonable amount of empirical work it is hard to identify which elicitation methods work well and which are best forgotten. ... There is now a much greater need for user-friendly software that elicits prior distributions. Ideally such software would offer a choice in the assessment tasks that are used for any particular elicitation, so that users could vary tasks and compare the elicited distribution. Then the
software would help advance research into assessment methods as well as providing a practical means of quantifying expert opinion in real applications.”

In 2008 O’Hagan realized the need for visualization software, and created an initial R-software package [52] entitled SHELF [43]. The package allows users to fit a beta, lognormal, or gamma distribution from elicited quantiles. However, the package is still based upon the old methods of anchoring experts to their elicited quantiles, then showing a fitted distribution post-elicitation, followed by asking the expert if they are content with the fitted distribution or would they like to repeat the process over again. Allowing the expert to directly modify the parameters in a continuous manner should prove to be beneficial over this type of elicitation.

To help experts from being overconfident, new visualization techniques are put forth in the dissertation through the use of an R software GUI package. The developed GUI, entitled ELICIT, reverses the standard algorithm of asking experts for quantiles and showing a fitted plot. Instead experts manually adjust a fitted plot, and use quantiles as real-time feedback. It is hypothesized that allowing experts to change the structure of the distribution through direct control over the parameters along with immediate visual feedback should prove beneficial in terms of accuracy, as it has been shown to do for some discrete distributions [54]. To validate the benefits of new elicitation techniques, experts’ opinions are elicited and analyzed across various topics using traditional elicitation methods and compared against these new methods using the developed software.

### 2.2 Types of Elicitation

There are an array of different elicitation processes available to retrieve an expert’s beliefs quantitatively. Whether a single probability estimate or a discrete or continuous probability distribution, it is assumed that each expert will be given the same type of elicitation for the problem at hand.
The simplest form of elicitation asks an expert what is the likelihood of an event happening. For instance experts might be asked, “What is the probability it will rain tomorrow?” Many studies have asked numerous experts similar questions to evaluate whether experts can correctly identify probabilities of unknown events. To evaluate if the experts are assigning the correct probabilities, resulting relative frequencies from events which have repeated trials can be compared to the expert’s elicited probabilities. The results can be plotted in a calibration curve, where one axis is the subjective probability judgment, and the other is the observed relative frequency. (For an example of a calibration curve see Figure 6.6). Numerous calibration studies have seen the same patterns of overconfidence, where experts tend to push their opinions towards 0 or 1 more often than the actual relative frequencies [35].

While many studies have investigated single probability elicitation techniques, utilizing these single estimates by themselves is not often done. These single estimates can lead to elicited probability distributions by simply concatenating the compliment probability. Instead of a single point estimate, a distribution containing two probabilities is typically used in the opinion pooling algorithms. Distributional theory can be applied to choose the best method of combining the probability mass functions. Therefore there is little focus on single probability estimates, as by default they can be transformed into probability mass functions.

Ignored thus far is the fact that the decision maker pooling the probabilities of experts has not quantified the uncertainty within each expert. Just because two experts elicit a 40% chance of event happening does not mean that they both believe there is an equal chance of the event occurring 50% of the time as well. To capture this uncertainty, experts must be asked not only on their point estimates, but also their variances around those estimates. This leads to the elicitation of a complete probability distribution for each expert for each element of interest.

To elicit a probability distribution experts are typically asked to give specific quantiles of the distribution, and then a best-fitting distribution is constructed to those quantiles. Typical distributions used for probabilities include the beta distribution for an event with two outcomes or a Dirichlet distribution for a categorical event with more than two outcomes. For
continuous values experts are usually fitted similarly to the normal distribution for unbounded values and the gamma distribution for positively defined values. While these are the most popular distributions, there is no reason that experts should be confined to them, and if the expert believes a certain parametric distribution to be a better fit, the extensions to do so are not difficult.

2.3 Evaluating Elicitations Through Scoring Rules

A major difficulty arises in evaluating expert elicitations due to the fact that repeated trials are not always possible. Many times only a single outcome is observed. For example, in evaluating the probability of which city will have the next terrorist attack across five US cities, a calibration curve is impossible to construct. New methods for evaluating elicitations are needed to determine whether experts are able to correctly express their opinions. There is a trade off in terms of accuracy between belief around a point estimate while not allowing for too much variance and providing no information. Scoring rules measure a degree of association between an elicited distribution and the probability of an event. Note that scoring rules require that the true outcome is known in order to calculate the actual score.

Scoring rules for probability mass functions quantify how much an expert associated their responses with the correct outcome. Defining some notation, suppose an expert has expressed a distribution for $n$ elements, or distinct categorical hypotheses. Let $p_i$ denote the probability assigned by the expert to outcome $O_i$, and let $d_i = 1$ if $O_i$ occurs, and $d_i = 0$ if $O_i$ does not occur, subject to $\sum p_i = 1$.

When determining which scoring rule to use, the first requirement is that the rule be proper. For a scoring rule to be proper, the rule must encourage experts to elicit their true beliefs if the experts know how they will be scored. The linear scoring rule is an example of a non-proper scoring rule. Imagine an event with two possible outcomes $O_1$ and $O_2$. The linear scoring rule assigns a score of $p_1$ if outcome 1 occurs, and $p_2$ if outcome 2 occurs, where the score is simply the probability assigned to the eventual outcome. To see why this is not a
proper scoring rule imagine the experts’ true belief is \( p = [0.7, 0.3] \). Their expectation is that 70% of the time they will receive a score of 0.7 and 30% of the time they will score 0.3. Therefore the expert’s expected score would be \( 0.7^2 + 0.3^2 = 0.58 \). However if they tried to alter from their true distribution and stated \( p = [1, 0] \) then similarly their expected value would be \( 0.7*1 + 0.3*0 = 0.7 \). Having a higher expected value by revealing a false probability distribution illustrates that this rule is not a proper scoring rule. Only proper scoring rules will be investigated in this research.

One of the popular proper scoring rules for categorical distributions is the Brier Score, or probability score, and is defined as

\[
PS = \sum_{j=1}^{n} (p_j - d_j)^2.
\]

This is often used as a scoring rule in determining the accuracy of weather forecasting. Note that \( PS \) is a penalty type score where lower values imply greater accuracy. Comparisons of judgments across groups can be done by calculating \( \bar{PS} \) (the arithmetic mean of individual probability scores), and comparing it to the performance of a completely non-preferential judge, or one which elicits equal probabilities of \( 1/n \) for each element.

Another popular scoring rule is the logarithmic scoring rule defined as

\[
LS = \log \left( \sum_{j=1}^{n} p_j d_j \right).
\]

The log scoring rule has been found to be more accurate in some studies, although weather forecasters continue to prefer the probability scoring rule. Both the logarithmic scoring rule and the probability scoring rule are criticized since they do not penalize for larger misplaced probabilities within ordinal categorical variables.

Continuous probability distributions are often used in elicitation contexts, however scoring rules are not used quite as frequently as they are in the discrete case. Nevertheless, proper scoring rules are designed by extending the discrete theory. Let \( \theta \) be the quantity of
interest, with \( f(\theta) \) being the expert’s probability density function for the value that \( \theta \) will eventually take. That true value which does occur is defined as \( \theta^* \). The equivalent of the probability scoring rule is defined as

\[
PS^* = 2f(\theta^*) - \int_{-\infty}^{\infty} f(\theta)^2 d\theta,
\]

with the logarithmic equivalent being

\[
LS^* = \log[f(\theta^*)].
\]

As opposed to the discrete case, the \( LS^* \) is much more popular than \( PS^* \) since the scoring rule depends only on the value of the function at \( \theta^* \).

### 2.4 Elicitation Experiment

Elicitation techniques play an important role in the elicitation process, and care must be taken to ensure experts’ recorded beliefs exhibit the greatest possible similarity to their true beliefs. While simply asking experts what they believe to be the probability of an event is the easiest elicitation technique, and in practice is often done, helping experts think critically about their probability estimates as a distribution has been shown to be beneficial. There have been many studies investigating accuracy based on different scoring rules, but very few on changing the entire procedure of the elicitation. There have been even fewer experiments utilizing the power new graphical methods computers provide. One study by Wang [54] had experts initially eliciting their probability through two visualization methods, the first through the use of a probability wheel, which is essentially an adaptable pie chart. The second was through a scaled probability bar, which is essentially a moving bar chart. Both of these methods were found to clearly outperform direct numerical assessment. Allowing the user to define the distribution strictly through visual means improved accuracy. However, these
techniques were never built upon and tested on non-discrete data. The use of visualization techniques on continuous distributions could prove to be of as much, if not more, value.

Traditional elicitation techniques for continuous distributions begin by asking the expert for the 50th percentile, or median of the distribution. Next the user is asked for their 5th and 95th percentiles, with the possibility of more divisions such as quartiles or deciles. A cumulative distribution function (CDF) or probability density function (PDF) is then fit to these elicited estimates. A more complex method can be used by allowing the user to see a graphical representation of their distribution post elicitation and re-adjusting their quantile estimates. Biases typically seen in this process include overconfidence both in the point estimate and in the variance. Results have shown that only between 40%-70% of the supposed 95% credible intervals contain the outcome of the random variable ([2], [35]).

A new method is proposed which reverses the traditional steps of anchoring an expert to their original estimates, instead allowing the expert to adjust the location and shape of the distribution through graphical controls of the parameters. Real-time feedback will then be given on the percentiles and summary statistics of the distribution as the expert adjusts the shape of the distribution. We believe that freeing the expert from anchoring themselves to an initial median can minimize the overconfidence biases. Constructing a distribution to an interval that looks accurate to them will hopefully result in more than 40%-70% accuracy of credible intervals containing the outcome to the question at hand. A screen shot of ELICIT is seen in Figure 2.1.

Two separate experiments were conducted to test the effects of the newly designed software. The first experiment is described, followed by the second, which improves upon the first.

### 2.4.1 Initial Experimental Design

For the first experiment, one-on-one interviews with two randomly split groups of experts were conducted, one group of which received the newly designed visualization software, while the other received no visualization until the end of the elicitation. Both groups
were initially trained on the process of elicitation using questions involving both aleatory (or statistical) uncertainty, as well as the more complicated epistemic uncertainty, or lack of knowledge. The experts in the experimental group adjusted their distributions visually to a normal distribution of their liking through moving the two parameter slider bars for the mean and standard deviation. The median, quartiles, and 5th and 95th percentiles were given as immediate feedback to the expert as they adjusted the parameters. Once the expert was satisfied with their distribution, the true results were given to them. The process was repeated through similar exercises until the user felt they were comfortably trained to begin the experiment. The experts in the control group were asked to provide the mean, 5th and 95th percentiles, and quartiles directly. After fitting a distribution which aligns to their quantiles, the distribution was displayed, and they were given opportunity to adjust their quantile estimates. After both groups were trained on their respective processes the experiment was conducted with similar questions across different types of probability distributions.
The probability distributions tested included the continuous normal, beta, and gamma distributions in addition to numerous categorical variables. The categorical distribution questions provide data collection for evaluating the performances of the different pooling operators. Multiple questions were asked for each distribution type, containing both aleatory and epistemic uncertainty. The experts selected for this experiment consisted of local university professors, business people, and college students. The questionnaire consisted of 3 training/seed questions, and 12 test questions involving topics related to finance and are listed in Table 2.2.

**Table 2.2. Questionnaire for Experiment 1.**

<table>
<thead>
<tr>
<th>#</th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>What was the value of the DOW Jones on Jan 1 2008?</td>
</tr>
<tr>
<td>2</td>
<td>What was the price of an ounce of gold on Jan 1 2008? (Probability Distribution: Below 500, 501-700, 701-900, 901-1100, 1101+)</td>
</tr>
<tr>
<td>3</td>
<td>What was the national unemployment rate during January 2008?</td>
</tr>
<tr>
<td>4</td>
<td>What proportion of the popular vote did Barack Obama get for the 2008 Presidential Election?</td>
</tr>
<tr>
<td>5</td>
<td>Which company posted the highest total gross profit Q2 2010? (Microsoft, Intel, AT&amp;T, Disney, Kraft)</td>
</tr>
<tr>
<td>6</td>
<td>How many vetoes did President George W. Bush issue during his presidency?</td>
</tr>
<tr>
<td>7</td>
<td>What will be the national unemployment rate for December 2010?</td>
</tr>
<tr>
<td>8</td>
<td>During December 2010 how many days will the DOW finish more than ± 100 points? (Probability Distribution 0-4,5-9,10-14,15-19,20+)</td>
</tr>
<tr>
<td>9</td>
<td>What will be the price of a barrel of light sweet crude oil on Dec 31 2010?</td>
</tr>
<tr>
<td>10</td>
<td>What will be the closing price of the DOW Jones on Dec 31 2010?</td>
</tr>
<tr>
<td>11</td>
<td>What will be the weekly approval rating of President Obama on Dec 31 2010?</td>
</tr>
<tr>
<td>12</td>
<td>What percentage of voters will vote for the legalization of marijuana in the California Nov. 2010 election?</td>
</tr>
<tr>
<td>13</td>
<td>In what country will Osama Bin Laden be captured/killed before Dec 31 2010? (Probability Distribution: Afghanistan, Pakistan, Iran, Any other country, Will not be found)</td>
</tr>
<tr>
<td>14</td>
<td>What will be the number of Democrats in the House of Representatives after the November elections?</td>
</tr>
<tr>
<td>15</td>
<td>Which country will President Obama visit first before Dec 31 2010? (Probability Distribution: Cuba, Iran, Iraq, Mexico, None of these listed)</td>
</tr>
</tbody>
</table>
Using the fact that under current methods around 50% of individuals capture the outcomes of interest in their confidence interval, and estimating that an accurate 95% of individuals should contain the outcome variable in a 95% confidence interval led to a sample size calculation (with \( \alpha = .05 \), \( \beta = .80 \), and accounting for the continuity correction), of 19 individuals in the visualization group and 19 in the standard/control group, for a total of 38 subjects.

The results were to be analyzed by comparing the visualization software group against the standard methods control group. A t-test on the average standard deviation across groups was to be done to see which group was able to minimize their overconfidence the most, by having larger standard deviation. Additionally, similar tests using scoring rules such as the quadratic/Brier score for discrete distributions, and the logarithmic and quadratic scores for continuous distributions were supposed to measure accuracy while accounting for the amount of information given. However, the opinions across experts deviated from one another on a scale that was much larger than expected. Some experts would answer so differently from one another that it was decided that observations across groups could not be compared accurately. After testing 19 subjects, a redesign of the experiment was deemed necessary to accurately test the hypotheses. Even using non-parametric tests, the experimental design still could easily be improved upon by controlling for inter-expert variation.

### 2.4.2 Follow Up Experiment

In the second experiment steps were taken to minimize the effects of experts having such disparate opinions. The first change was that the nature of the questions were made to be more accessible to the average subject. Rather than technical financial questions, easier topics that people experience in everyday life were used. The second change drastically improved the design by restructuring the experiment into a matched-pairs design rather than comparing two independent groups. In this second experiment, experts would first be asked three questions through a randomly selected method of the GUI software or through the quantile technique SHELF [43]. After answering the three questions each expert was asked the same
three questions again through the alternative unchosen method. Experts were unaware that they would be asked the same questions a second time, and were unable to review their first answers when answering the second time. The reasoning for this design is that the impact of the elicitation technique could still be measured without being lost in the drastically different opinions, which continued to occur.

The three questions used in this experiment were:

1. What was the average price of regular unleaded gasoline in San Diego on 01/01/2011? (Normal Distribution)
2. How many miles is it from San Diego to San Francisco (as the crow flies)? (Categorical: 0-399, 400-500, 501-599, 600-699, 700+)
3. What proportion of American Households have a household income above/below $75,000? (Beta).

For the third question experts would answer one time for the proportion of households making over $75k, and the other time making under $75k. This was done to help remove the bias of experts trying to answer similarly to how they answered during the first round. Switching to the compliment it made it harder for experts to answer in the same manner as they did during the first round. Since the answer is a distribution on a two outcome variable, assessments can easily be done by switching the parameters of the beta distribution.

2.4.3 Follow Up Experiment Results

Overall there were few practical benefits seen by utilizing the new GUI system for elicitation as opposed to asking for elicited quantiles through SHELF [43]. In fact, the subjects tended to be more confident when answering through the GUI, giving narrower standard deviations in the two continuous variable questions. However, even though the variability was smaller in the GUI system resulting in poorer overall accuracy, an overwhelming majority preferred using the GUI system, finding it much easier to comprehend the elicitation process.
For each of the three questions, multiple statistical tests are calculated to determine whether there are differences when experts use the GUI method versus SHELF for elicitation. Two tests include a paired t-test and Wilcoxon sign rank tests on whether the standard deviations changed across elicitation methods. The non-parametric Wilcoxon test is conducted due to the presence of extremely large outliers. Similar tests are done on using logarithmic scoring rules on which group was better to accurately answer the questions.

For the first question, relating to the price of regular gasoline, the elicited distributions across each method are displayed in Figure 2.2. Barring the two overconfident distributions elicited through the GUI, the distributions look fairly comparable, and a test could be done to see if indeed GUI users had larger standard deviations. Conducting a paired t-test on the differences of the standard deviations across experts results in an insignificant p-value of 0.15, with the leaning towards the SHELF method having larger standard deviations. A non-parametric Wilcoxon signed rank test also concluded in non-significant differences in the standard deviations with a p-value of 0.18. Performance-wise, when testing differences of the log-score ($LS^*$) both the t-test and Wilcoxon signed rank test returned insignificant results with p-values of 0.15 and 0.35 respectively.

![Figure 2.2. Elicited distributions from Experiment 1 - Question 1.](image-url)
Question two had very few changes in opinion with 65% of responses being exactly equal across methods. It was easy for experts to remember their responses from the first elicitation round, and they often quickly entered the same probabilities when asked through the second method. Boxplots for the responses across both methods are displayed in Figure 2.3. Due to very few changes in opinions across methods, a comparison across the methods is omitted from the analysis. However, a comparison on performance using the logarithmic scoring rule $LS$ can be done to test the relative performance of experts on each method. Since experts contained zero values for the correct response in both methods, taking the log results in $-\infty$, making a t-test incalcuable. A Wilcoxon signed rank test results in an insignificant p-value of 0.48. Similar to the first question there are no significant differences seen across elicitation methods.

Figure 2.3. Elicited distributions from Experiment 1 - Question 2.
For the third and final question, distributions were visually more concentrated when elicited through the GUI (Figure 2.4). A t-test of a difference across methods results in a close to significant p-value of 0.06, while the Wilcoxon signed rank test is significant at a p-value of 0.03. In terms of performance using the logarithmic scoring rule $L^*$, a t-test on the values is not applicable due to outliers created by poorly elicited distributions. A Wilcoxon signed rank test returns a close to significant p-value of 0.053. For this question, SHELF seemed to outperform the GUI method in terms of lowering overconfidence and raising performances of experts.

![Elicited distributions from Experiment 1 - Question 3.](image)

Across all questions there was no evidence to suggest that the new GUI method influences experts to expand their standard deviations, hence lowering the overconfidence bias. In fact, in the final question, the opposite was seen with SHELF having significantly wider distributions. However, most interestingly, of the 20 experts who participated, 18 preferred using the GUI system, with one preferring SHELF, and one having no preference. With the GUI clearly being preferred by the experts, it is unfortunate that there is no evidence of adding to greater predictability, with some discouraging results. One possible reason for
the wider distributions is when opinions are elicited using SHELF experts must create space
in between their quantiles. By forcing the numbers to be apart often on increments of 1, 5, or
10, it lead more often to a wider distribution as opposed to fitting it directly.

Evaluations on how to best pool the distributions will be analyzed in Chapter 4 after
describing the available techniques in Chapter 3. First however, Chapter 2 is concluded with a
proposal of another advanced elicitation technique.

### 2.5 Analysis of Competing Hypotheses

Analysis of competing hypotheses (ACH) is a method for systematically comparing
the likelihoods of competing hypotheses based on the available evidence. Richard Heuer [26]
developed the procedure for the CIA in the 1970s, primarily for use in intelligence matters.
Heuer’s work did not provide a mathematical basis for drawing conclusions, though he did
note that formalizing the process would be desirable. As opposed to the elicitation techniques
discussed thus far in which experts try to consolidate their opinions while simultaneously
considering all evidence, ACH requires an expert to evaluate the importance of multiple
evidence items independently. Again focus here is on the elicitation of a probability
distribution bounded between zero and one.

The multinomial-Dirichlet model is a generalization of the beta-binomial model in
which there are $N$ categories rather than the two categories success and failure. A draw from
a multinomial distribution is represented by a vector $x = (x_1, x_2, \ldots, x_N)$, where each $x_j$
is the count of observations that fall into category $j$. Let $p = (p_1, p_2, \ldots, p_N)$ denote the
probabilities of belonging to the $N$ categories so that $\sum_{j=1}^N p_j = 1$ and let $n = \sum_{j=1}^N x_j$.

A Dirichlet distribution is placed on $p$, the category probabilities of the multinomial
distribution. The parameter for the Dirichlet distribution is $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_N)$, a vector
with nonnegative entries. Letting $\alpha_0 = \sum_{j=1}^N \alpha_j$, the marginal distribution of each $p_j$ is a beta
distribution.

The multinomial and Dirichlet distributions are conjugate to one another, meaning that
when we start with a Dirichlet prior distribution on the category probabilities and update our
knowledge with multinomial data, the resulting posterior distribution on the category (or hypothesis) probabilities is Dirichlet. Specifically, when the prior distribution on \( p \) is Dirichlet(\( \alpha \)), and multinomial data \( x \) is observed, the posterior distribution on \( p \), conditioned on this data, is Dirichlet(\( \alpha_1 + x_1, \alpha_2 + x_2, \ldots, \alpha_N + x_N \)).

To pose ACH in terms of multinomial and Dirichlet distributions, we consider the final outcome, or determination of the true hypothesis to be a single draw from a multinomial distribution. Evaluating the evidence provides information about \( p \), the parameter of this multinomial distribution, which gives the probabilities of the hypotheses.

The algorithm for implementing the procedure is as follows;

1. **Construct the framework for the ACH matrix.** List hypotheses, \( H_j \), \( j = 1, \ldots, N \) as column headings and evidence items, \( E_i \), \( i = 1, \ldots, M \) as row headings. Include prior beliefs or “other evidence” as the first evidence item.

2. **Assign evidence weights.** Determine the equivalent prior sample size (\( n_{ess} \)) of the evidence as a whole. Assign weights, \( w_i \), to the evidence items indicating their strength or relative importance. Scale these weights so that \( \sum_{i=1}^{M} w_i = n_{ess} \).

3. **Relate evidence to hypotheses.** Proceeding one row at a time, rate the relative likelihood of \( E_i \) conditioned on each hypothesis by filling in the matrix with values \( x_{ij} \). One may begin by assigning \( x_{il} = 1 \) where \( H_l \) is the hypothesis under which we are least likely to observe \( E_i \). Continue by assigning the other \( x_{ij} \) values relative to \( x_{il} \). If \( E_i \) is twice as likely to be observed when \( H_j \) is true compared to when \( H_l \) is true, then \( x_{ij} = 2 \). After the initial assignment, scale these values so that \( \sum_{j=1}^{N} x_{ij} = w_i \).

4. **Compute the posterior.** The posterior distribution of \( p \), the probabilities of the hypotheses, is Dirichlet with parameter \( \alpha \) given by \( \alpha_j = \sum_{i=1}^{M} x_{ij} \), for \( j = 1, \ldots, N \). The marginal posterior distributions of the individual \( p_j \) parameters are beta(\( \alpha_j, \alpha_0 - \alpha_j \)).

Step 1 corresponds to steps 1, 2, and 3 of Heuer’s method. Step 2 is a formalization of Heuer’s call to “Analyze the diagnosticity of the evidence and arguments”. The method of assigning \( x_{ij} \) values in Step 3 is posed as in McLaughlin [40], though we note here that 0 values are permitted for \( x_{ij} \). An alternative way to elicit \( x_{ij} \) values is to interpret it as the number of “observations” of \( H_j \) with which \( E_i \) can be associated. Written out formally, we
have

\[ y \sim \text{Multinomial}(1, p) \]
\[ p \sim \text{Dirichlet}(\alpha) \]
\[ \alpha_j = \sum_{i=1}^{M} x_{ij}, \quad j = 1..N, \quad (2.2) \]

where \( y \) is the outcome we are trying to predict.

All inference is conducted using the posterior distribution. The mean of the posterior Dirichlet(\( \alpha \)) distribution is

\[ \hat{p} = \left( \frac{\alpha_1}{\alpha_0}, \ldots, \frac{\alpha_N}{\alpha_0} \right). \]

Interval estimates provide a measure of uncertainty for the point estimates above, but numerical methods are required to find quantiles of beta and Dirichlet distributions. A 95\% equal-tail credible set for a single probability \( p_j \) is obtained by finding the .025 and .975 quantiles of the beta(\( \alpha_j, \alpha_0 - \alpha_j \)) distribution. Highest posterior density sets [10] will be slightly narrower than the equal-tail sets. The width of these interval estimates is strongly affected by the choice of \( n_{ess} \).

The multinomial-Dirichlet ACH model, as presented thus far yields a Dirichlet posterior distribution on the hypotheses that is easy to compute and easy to use for conducting inference. However, it is missing the ability to handle items that give evidence against one or more hypotheses and items that are not relevant to all hypotheses. Unfortunately these extensions result in posterior distributions that are not Dirichlet and this loss of conjugacy makes inference more difficult. The use of Monte Carlo methods, however, makes inference computationally feasible, and samples from the posterior can be obtained fairly quickly.

Evidence against hypothesis \( H_k \) is expressed as seeing observations that do not fall into category \( k \). In a beta-binomial setting, where there is only one category that is “not \( H_k \)”, this is easy to handle. In a Dirichlet model, the observations need to be allocated amongst all the categories that are “not \( H_k \)” without changing the relative probabilities of these other
hypotheses and without changing our certainty about the relative probabilities of these other hypotheses. This is accomplished by treating the evidence against as a binomial random variable with probability of success $1 - p_k$. We know the observations belong to the “not $H_k$” categories, but we do not know how many observations fall into each. If an evidence item $E_i$ is against a set of hypotheses, rather than just a single hypothesis $H_k$, it is “for” the complement of this set which we will refer to as $F_i$. Then the evidence is treated as a binomial random variable with probability of success $\sum_{j \in F_i} p_j$.

Let our current knowledge of $p$ be given by a proper prior distribution $\pi(p)$, and let $E_i$ provide evidence for a set of hypotheses $F_i$, with the strength of this evidence represented by $w_i$. Then, the posterior distribution for $p$, updated with $E_i$ is given by

$$
\pi(p | E_i) \propto \left( \sum_{j \in F_i} p_j \right)^{w_i} \pi(p).
$$

(2.3)

In practice, an analyst would input plus and minus signs in the matrix to indicate evidence for and against hypotheses and use $w_i$ to represent $\sum_{j \in F_i} x_{ij}$.

When a piece of evidence is irrelevant to hypothesis $H_k$, it is as though the data associated with $H_k$ is missing. It is not appropriate to assign $x_{ik} = 0$ when $E_i$ is not relevant to $H_k$ as this would reduce the likelihood of $H_k$ relative to the other hypotheses when it should be held constant.

Let our current knowledge of $p$ again be given by the distribution $\pi(p)$, and let $E_i$ be relevant only to hypotheses in $\mathcal{R}_i$. Then, the posterior distribution for $p$, updated with $E_i$ is

$$
\pi(p | E_i) \propto \left[ \prod_{j \in \mathcal{R}_i} \frac{p_j^{x_{ij}}}{\sum_{k \in \mathcal{R}_i} p_k} \right] \pi(p).
$$

(2.4)

This posterior distribution is guaranteed to be proper if the prior distribution is proper.

When a user decides to enter an evidence item as either for, against, or irrelevant to a hypothesis, the posterior distribution for $p$ is no longer in closed form. In order to express the posterior, we partition the evidence by type into three sets. Evidence items with only
numerical values will belong to set $A$. Evidence items that contain irrelevant hypotheses (NA’s) will be labeled as set $B$. Evidence items that are just for (+) and against (-) will belong to set $C$. Items that contain both an NA, and + or - will be placed into set $C$. Evidence in set $C$ may not contain numeric values.

Using Equations (2.2), (2.3), and (2.4) the posterior distribution for any set of evidence is given by

$$
\pi(p|E) \propto \prod_{i \in A} \prod_{j=1}^{N} p_{x_{ij}}^{x_{ij}} \times \prod_{i \in B} \prod_{j \in R_i} \frac{p_{ij}}{\sum_{k \in R_i} p_{ik}} \times \prod_{i \in C} \left( \sum_{j \in F_i} p_j \right)^{w_i}
$$

where $N$ is the number of hypotheses, $R_i$ is the set of relevant hypotheses for evidence item $i$, $F_i$ is the set of hypotheses that item $i$ provides evidence for (+), and $w_i$ is the scaled weight for evidence item $i$. Importance sampling is used to sample from the distribution.

An example is now worked through to demonstrate the ACH elicitation methods. The city of San Diego in 2006 declined to provide the owners of the San Diego Chargers football team with the support they were seeking to build a new stadium and redevelop the site of Qualcomm Stadium. The Chargers organization then stated that they will definitely be moving from Qualcomm Stadium, with their contract requiring them to stay only through the end of the 2008 season. Other cities in the Southwest such as Las Vegas and San Antonio have been proposed as new locations for the Chargers. However, the Chargers have stated that they would like to stay in the San Diego area and have considered several other sites in San Diego County; Oceanside, National City, and Chula Vista. National City dropped from the running in the spring of 2007. In this section we apply the multinomial-Dirichlet ACH model to the problem of predicting the new stadium site for the Chargers based on available evidence. The evidence was obtained from news articles in the San Diego Union Tribune through the spring of 2007. The three hypotheses are Oceanside, Chula Vista, and Other.

Table 2.3 shows the scaled input of an analyst who used the extensions proposed. $n_{ess}$ was again chosen to be 10. There are four evidence items with complete numeric values belonging to set $A$ ($E_3$, $E_5$, $E_6$, $E_7$), one with an irrelevant hypothesis belonging to set $B$
(E_4), and one with evidence for and against belonging to set C (E_2). The posterior distribution for p is \( \pi(p|E) \propto p_1^{3.45}p_2^{3.02}p_3^{1.53}(p_1 + p_2) \).

**Table 2.3. ACH matrix for stadium example.**

<table>
<thead>
<tr>
<th></th>
<th>w</th>
<th>Oceanside</th>
<th>Chula Vista</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>E1: Prior beliefs or unlisted evidence</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E2: Chargers say they want to stay in San Diego</td>
<td>2.0</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>E3: Chargers want financial assistance (any city)</td>
<td>2.5</td>
<td>0.75</td>
<td>0.25</td>
<td>1.50</td>
</tr>
<tr>
<td>E4: Chargers like parking and transit in Oceanside</td>
<td>1.0</td>
<td>0.67</td>
<td>0.33</td>
<td>NA</td>
</tr>
<tr>
<td>E5: San Diego State University wants to be involved</td>
<td>0.5</td>
<td>0.14</td>
<td>0.35</td>
<td>0.01</td>
</tr>
<tr>
<td>E6: Chargers paid $200K to study sites in Chula Vista</td>
<td>2.0</td>
<td>0.09</td>
<td>1.90</td>
<td>0.01</td>
</tr>
<tr>
<td>E7: Oceanside council set aside $100K for consultants</td>
<td>2.0</td>
<td>1.80</td>
<td>0.19</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Computing \( \sum_{i \in A} x_{ij} \), the column sums for evidence items in A, gives \( \alpha = (2.78, 2.69, 1.53) \). Next, we update \( \alpha \) with evidence in B to obtain \( \alpha^* \). \( E_4 = (.67, .33, NA) \) gives the updated value \( \alpha^* = (3.30, 2.95, 1.75) \). Finally, we update \( \alpha^* \) with evidence in C to obtain \( \alpha^{**} \). \( E_2 = (+, +, -) \) gives \( \alpha^{**} = (4.36, 3.89, 1.75) \).

We now have a close Dirichlet distribution to use as an importance function for importance sampling. Table 2.4 summarizes the results of the posterior containing estimates of the mean, 95% HPD intervals, and \( p_{max} \) with a Monte Carlo sample size of 100,000. Figure 2.5 displays the true marginal posteriors along with the importance function. Results are quite similar to the previous assignment of matrix values, with Oceanside having a slightly higher posterior mean in this assignment. Since either of these special types of evidence can be interpreted as missing data, there will generally be more variance in the posterior distribution when “NA”, “+”, and “-” are used than when all matrix entries are numeric.

**Table 2.4. Results for ACH extensions example.**

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Interval</th>
<th>( p_{max} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>OC</td>
<td>0.442</td>
<td>(0.156, 0.726)</td>
<td>0.535</td>
</tr>
<tr>
<td>CV</td>
<td>0.388</td>
<td>(0.120, 0.675)</td>
<td>0.406</td>
</tr>
<tr>
<td>Other</td>
<td>0.170</td>
<td>(0.004, 0.397)</td>
<td>0.059</td>
</tr>
</tbody>
</table>
Figure 2.5. Marginal posterior distributions for $p$ and proposal densities used for importance sampling.
CHAPTER 3
COMBINING PROBABILITY DISTRIBUTIONS

3.1 OVERVIEW

Once the experts’ distributions are properly elicited, the DM’s next task is to combine the experts’ opinions into a single consensus. Obtaining a consensus opinion from a group of experts who have each put forth an opinion as a probability distribution has been the subject of much research ([6], [16], [51], [18]). Clemen and Winkler [16] categorize opinion pooling into two major fields: mathematical approaches and behavioral approaches. Mathematical approaches are further divided as either axiomatic or Bayesian techniques. Axiomatic techniques involve those methods which try to force the pooled distribution to adhere to specific mathematical properties. Bayesian techniques involve the DM updating his prior distribution, or opinion, based on the “new data” obtained from the experts. Behavioral approaches include the Delphi method, the Nominal Group Technique, and Kaplan’s approach. All behavioral methods involve interaction among experts through different techniques. A further review of behavioral approaches can be found in Clemen and Winkler [16].

Bayesian pooling methods are difficult to implement because they require deriving a likelihood function. The likelihood function needs to accurately account for three important aspects. First, it must be able to measure bias, necessitating that each expert be properly calibrated. Secondly, the precisions of experts’ estimates need to be measured through some type of certainty measure. And lastly, the interdependence among experts should be accounted for. Many different ad-hoc Bayesian methods in the literature try to simplify the problem for the DM. Genest and Zidek [23], and Clemen and Winkler [16] provide reviews on numerous Bayesian methods.
In addition to the standard axiomatic pooling operators and Bayesian methods, an alternative pooling method has been proposed. Cooke’s method [18] uses the product of the entropy (Definition 3.12) and calibration score from known seed variables as a means to determine a weighting structure in a linear pooling operator. Critics of this technique include Sandri [48] who shows through a simple example how experts with disparate opinions about a future outcome, where one expert is preferred in terms of the realistic concept of accuracy and plausibility, can result in calibration scores which are equal or in the opposite ranking one would expect. The authors note that this comes from the fact that Cooke’s method does not have an individual quality index, and issue a warning that a decision maker can be unaware they are applying a suboptimal weighting structure.

Although Bayesian methods and Cooke’s method are noted to be preferred by some authors on the subject ([16], [15]), they are almost never applied in practice, with axiomatic techniques being used with much higher frequency due to their overall simplicity while maintaining a high level of accuracy ([14], [37], [21]). Due to the axiomatic approaches receiving much of the attention, they are the primary focus of the following sections which evaluate existing pooling operators. In the following section, this dissertation puts forth numerous new axiomatic techniques outside of the typical two popular linear and logarithmic pooling operators.

### 3.2 Axiomatic Approaches

The two primary axiomatic approaches used in practice are linear opinion pooling (LinOP) and logarithmic opinion pooling (LogOP) [11] [24]. Respectively these are equivalent to the arithmetic and geometric means. In addition, some other axiomatic approaches sometimes used include the power mean, trimmed mean, and Winsorized mean. Each of these means will be described and evaluated. In addition a new proposed axiomatic approach is discussed which alleviates some of the problems with the commonly used methods.
Defining our notation, let \( \theta \) denote the event or parameter of interest to the DM and \( j \) index the elements of \( \theta \) when it is vector-valued \((j = 1, \ldots, n)\). Let \( m \) denote the number of experts submitting opinions and \( q_i(\theta) \) denote the opinion of expert \( i \) \((i = 1, \ldots, m)\). A weight, \( w_i \), is associated with each expert subject to \( \sum_{i=1}^{m} w_i = 1 \), and \( w_i \geq 0 \) for all \( i \). \( p(\theta) \) denotes the pooled opinion. When \( \theta \) has multiple elements, expert and pooled marginal distributions are denoted by \( q_i(\theta_j) \) and \( p(\theta_j) \) respectively.

Before each pooling operator is discussed in detail, the desired properties mentioned in the literature for axiomatic opinion pooling operators are defined.

**Definition 3.1**

*Unanimity* - If all experts submit probabilities of \( v \) for the \( j^{th} \) element, then the pooled estimate should be equal to \( v \) for the \( j^{th} \) element; that is, \( q_i(\theta_j) = v \) for all \( i \) implies \( p(\theta_j) = v \).

**Definition 3.2**

*Zero Preservation* - If all experts submit probabilities of zero, the pooled estimate should also be equal to zero (a special case of unanimity); that is, \( q_i(\theta_j) = 0 \) for all \( i \) implies \( p(\theta_j) = 0 \).

**Definition 3.3**

*Marginalization* - The pooled opinion is the same whether we calculate the marginal distributions separately or combine the experts’ entire distributions, and then concatenate each element. \( p(\theta) = [p(\theta_1), \ldots, p(\theta_n)] \).

**Definition 3.4**

*Boundedness* - The pooled estimate should not be lower than the smallest expert opinion, nor higher than the largest expert opinion; that is, \( \min_i[q_i(\theta_j)] \leq p(\theta_j) \leq \max_i[q_i(\theta_j)] \) for all \( j \).

**Definition 3.5**

*Monotonicity* - If any expert increases or decreases their opinion, the pooled opinion should move in the same direction; that is, \( q_i^*(\theta_j) > q_i(\theta_j) \) implies \( p^*(\theta_j) > p(\theta_j) \).
Definition 3.6

External Bayesianity - Updating the finished pooled operator is equivalent to updating each expert’s distribution and re-pooling. Let $l$ be the agreed upon likelihood function. Then

$$\frac{l_p(\theta)}{\sum l(p(\theta))} = \left[ \frac{l_{p_1}(\theta)}{\sum l(p_1(\theta))}, \ldots, \frac{l_{p_n}(\theta)}{\sum l(p_n(\theta))} \right].$$

3.2.1 LinOP

Of the two well known axiomatic approaches, LinOP is more popular amongst practitioners ([4]). LinOP is defined as the weighted arithmetic average,

$$p_{Lin}(\theta_j) = \sum_{i=1}^{n} w_i q_i(\theta_j). \quad (3.1)$$

In addition to ease of computation and interpretation, an advantage of LinOP is that it satisfies marginalization (Definition 3.3). This is particularly useful when $\theta$ consists of a vector of probabilities that add to one because it ensures that there is no need for any rescaling of the pooled estimate to obtain a distribution.

A disadvantage of LinOP is its sensitivity to aberrant experts which makes it a poor choice for selecting a consensus opinion. Outliers can lead to multimodal distributions of opinions or even dominate the decision completely. The problem of LinOP’s sensitivity to extreme values has been addressed by researchers who have sought to reduce the extremity of the outlying opinions ([31], [21]). Jose and Winkler [31] propose using trimmed and Winsorized means to rectify the influence of outliers, however each of these methods alters the actual values contained in the dataset. One of the motivations for the Hybrid mean proposed in section 3.2.4 was to design a pooling method that uses all the original data but is still able to minimize the effects of outlying observations.

3.2.2 LogOP

The primary alternative to LinOP is the logarithmic pooling operator, LogOP, which is the weighted geometric mean of the experts’ opinions. $p_{Log}(\theta)$ is referred to as the logarithmic pooling operator because it is equivalent to the weighted arithmetic mean of the logarithms of
the individual probabilities. LogOP is defined as

\[ p_{\text{Log}}(\theta_j) = k \prod_{i=1}^{n} q_i(\theta_j)^{w_i}, \]  

(3.2)

where \( k \) is the normalization constant to ensure a probability density.

Turning to LogOP often rectifies many of the failures seen in LinOP. LogOP is less influenced by large outliers as compared to LinOP meaning that the DM is more likely to reach a single decision which better represents the experts ([23],[21]). The logarithmic combination is also attractive to some because it satisfies external Bayesianity (Definition 3.6). One major flaw with this assumption is that every expert needs to interpret evidence in the same way; ensuring a constant likelihood function. Hence, deriving a likelihood function across experts remains problematic. If used correctly under the right conditions, it is expected that the LogOP will outperform LinOP [28], however these conditions can be quite strict.

A significant drawback of LogOP is the impact of zeros on the pooled value. It takes just one expert to value an element a zero to ensure the pooled estimate is zero for that element. In essence an expert that gives a zero probability acts as a veto mechanism for the marginal distribution of that element. This impossibility value of zero can drastically impact the pooled distribution. In the most extreme scenario there can be at least one zero value in each element resulting in a final pooled vector of all zeros.

Lindley claims that experts should avoid using probabilities of 0 or 1 except during logical true-false statements [39]. The result is known as Cromwell’s rule in reference to his saying, “I beseech you, in the bowels of Christ, think it possible you may be mistaken.” An example is given where a Bayesian attaches a prior probability of zero that the moon is made of green cheese. Then no matter how many armies of astronauts bring back green cheese it cannot convince him otherwise. This argument is easily extended to the LogOP’s veto power due to the similar multiplicative nature of Bayesian updating.

However, in reality many experts often denote probabilities of zero and one, as seen in the elicitation chapter. Zeros occur because some hypotheses are comparatively irrelevant to
other much more likely hypotheses. To show the need for assignments of zero Lindley’s 
example of the moon’s material is expanded upon. There are alternative hypotheses of which 
the moon could be made of blue cheese, white cheese, or even red ice cream. There are an 
infinite number of possibilities, but in reality people automatically create a subset of 
possibilities by assuming irrelevant hypotheses have probability of zero. Without doing this 
any legitimate hypothesis such as “iron” would become $\frac{1}{\infty}$ due to the inclusion of the infinite 
number of alternatives.

When zero values are allowed to be assigned to hypotheses, some practitioners turn to 
band-aid type measures to try and overcome the zero vetoing effect in LogOP by 
manipulating the data into something usable. For example, many practitioners simply add 
small values to zeros when a log computation is needed to prevent vetoing [21]. In a quick 
example, suppose a DM is asked to pool the individual probabilities of five experts with 
values of (0,0,0.3,0.6,0.95). By replacing the zero values with a small constant one can use 
LogOP to obtain a non-zero result. The values of LogOP from replacing the zero entries with 
alternate values are listed in Table 3.1. Notice that the estimate is highly sensitive to the 
choice of the replacement value. In our examination of HybOP in the next section we will 
show how the addition of a small constant influences LogOP in a fixed mathematical way that 
should not be ignored.

<table>
<thead>
<tr>
<th>Value to replace zeros</th>
<th>0.001</th>
<th>0.005</th>
<th>0.010</th>
<th>0.050</th>
<th>0.100</th>
</tr>
</thead>
<tbody>
<tr>
<td>LogOP (equal weights)</td>
<td>0.044</td>
<td>0.084</td>
<td>0.111</td>
<td>0.212</td>
<td>0.280</td>
</tr>
</tbody>
</table>

### 3.2.3 Power Mean

While the vast majority of axiomatic pooling is currently done by either LinOP or 
LogOP, Cooke [18] suggests a power mean operator, an alternative which attempts to combine 
some of the benefits of LinOP and LogOP. It is expressed as
where $k$ is a normalization factor ensuring a density, and $r$ is a tuning parameter which ranges from $-\infty$ to $\infty$. When $r = 0$ or $1$, (3.3) is equivalent to LogOP and LinOP respectively. When $r = -\infty$, the minimum value of each element will be chosen as the pooled estimate, and analogously the maximum value when $r = \infty$. Recommendations for the choice of $r$ are almost non-existent with only explanations given on what will occur as $r$ changes, but no reason behind what specific value to choose for $r$.

### 3.2.4 Hybrid Mean

LinOP and LogOP are useful, well-established pooling methods. However for data sets in which opinions are heavily skewed and contain zero values, neither provides the DM with a robust method. To overcome these drawbacks I propose a new hybrid pooling operator (HybOP), defined as,

$$M_r(j) = k \left( \sum_{i=1}^{n} w_i q_{ij}^r \right)^{\frac{1}{r}}, \quad (3.3)$$

The hybrid pooling operator (3.4) can be thought of as a function of $c$ in which the pooled value increases from LogOP to LinOP as $c$ increases, hence the hybrid name. A special case of HybOP when $c = 0$ gives LogOP. From Hoehn [27] (equal weights) and Bullen [8] (general weights) it is known that

$$\lim_{c \to \infty} \prod_{i=1}^{n} (q_i(\theta) + c)^{w_i} - c = \sum_{i=1}^{n} w_i q_i(\theta). \quad (3.5)$$

As the constant $c$ increases to infinity, HybOP (with no normalization needed), approaches LinOP. This is an interesting result given the fact that the multiplicative nature of the original operator vanishes, being replaced by an additive measure solely from location shifts of the data.
Notice that HybOP overcomes the troubles discussed earlier with LinOP and LogOP. The geometric pooling structure is maintained in essence for small $c$ allowing for better pooling of skewed distributions, while also ensuring zero values will not be destructive to the marginal distributions.

Since the probabilities submitted to the DM by the experts are fixed constants, equation (3.4) is a function strictly of the constant $c$. The problem that faces the DM is what value of $c$ gives an “optimal consensus”. While there is no objective function which can be used to pick an optimal $c$, a recommendation method is now proposed, but at the end of the day the DM should use their discretion in the final choice of $c$. The method proposed here can be used as a logical starting point with adjustments made after the results are thoroughly investigated by the DM if deemed to be necessary.

Though mathematically different from a linear combination of LogOp and LinOp, the limit in $c$ shows that HybOP serves as a middle ground between LogOP and LinOP, falling closer to LogOP when $c$ is small and closer to LinOP when $c$ is large. The shape of $p_{\text{Hyb}}(\theta_j)$ as a function of $c$ is dependent on the $q_i(\theta_j)$ values. When at least one of the $q_i(\theta_j)$ values is zero, the HybOP function must also be equal to zero value $c = 0$. When all the $q_i(\theta_j)$ are positive, the function begins at a larger value (the logarithmic mean) when $c = 0$. As a result when there is a zero entry in $q_i(\theta_j)$, there is a much larger degree of convexity as the function approaches the linear mean. The location and variance of opinions also impact the rate of convergence, but to a lesser degree than the zero vetoing.

A proposed recommendation for the choice of $c$ is to minimize the distance from where the linear mean lies on the vertical axis, the point $(0, p_{Lin}(\theta))$, to the HybOP function. The formula for this minimum distance, calculated using the Pythagorean theorem is,

$$\arg \min_c \left[ c^2 + \left( \sum_{i=1}^{n} w_i q_i - p_{\text{Hyb}}(\theta) \right)^2 \right]. \quad (3.6)$$
For sets of opinions which do not have zero values in them, the distance will be minimized with a very small constant value and thus will result in a pooled estimate which is close to the logarithmic mean. For those sets of opinions which have zeros, there is a larger constant which will minimize the distance and thus more of a “middle ground” is attained. An example of this computation is shown in Figure 3.1 when pooling the three values 0, .06 and .11. The hybrid function begins at the origin and converges to the arithmetic mean of 0.0567 as $c$ tends to $\infty$. The shortest distance from the point of the arithmetic mean on the vertical axis to the function is shown leading to the choice of $c$ at 0.0129. With this choice of $c$ the hybrid mean results in a value of 0.0358, larger than the geometric mean of zero and less than the arithmetic mean of 0.0567.

![Figure 3.1](image)

**Figure 3.1. Calculating the constant $c$ by (3.6) when pooling (0, 0.06, 0.11).**

The process for choosing $c$ when pooling opinions of a single element event is now extended to the optimization when $\theta$ is vector valued. In this case the goal is to minimize the sum of the distances across all marginal distributions. For cases when $A > 1$ the
recommended choice of $c$ is given by,

$$\arg\min_c \sum_{j: q_j \neq 0} \sqrt{c^2 + \left( \sum_{i=1}^{n} w_i q_i(\theta_j) - p_{\text{Hyb}}(\theta_j) \right)^2}$$  \hspace{1cm} (3.7)

where $q_j = \sum_{i=1}^{n} q_{ij}$. Note that marginal distributions for which all experts have proffered an opinion of zero are excluded from this computation. Due to the zero preservation property of HybOP, the pooled value for these elements is zero regardless of the choice of $c$. When more elements with all zero opinions are included in the calculation the choice of $c$ is pushed lower. Therefore, these elements are excluded so that adding irrelevant elements does not impact the choice of $c$ and thus the final pooled value. For instance, enlarging a search area on a map to include regions where all experts think the target is not likely to be should not by itself result in a change in the pooled opinion.

In (3.6) and (3.7) the choice of $c$ leads to a pooled result that is invariant to scale transformations of the data. Any rescaling of the data will result in the equivalent rescaling of the optimal constant choice. This invariance property may give a DM further reason to choose $c$ via an algorithmic method such as (3.6) or (3.7) rather than opting for a fixed value $c$ chosen in advance.

### 3.3 Distance Measures

Approaching the task of pooling the opinions of experts through probability distance measures is an additional algorithmic approach to computing the “best” pooled distribution. Divergence measures are functions which quantify the relative “distance” from one probability distribution to another. Ideally, the pooled distribution is one which minimizes the distance between all experts to the pooled distribution. However, calculating the distance between two probability distributions is not a trivial task, with many sophisticated methods available. It is a goal of this research to rectify some of the problems of the current popular methods through the use of alternative divergence measures. Table 3.2 shows all the divergence measures that are investigated.
Using the same notation as in the axiomatic opinion pooling framework, imagine a group of \( i = 1, \ldots, m \) experts submitting probability mass functions over \( j = 1, \ldots, n \) elements. The \( \theta \) notation is dropped for simplicity. The expert weights \( w_i \) are assumed to be positive and sum to one. Let \( Q_i \) denote expert \( i \)'s probability distribution and \( P \) be the optimized pooled consensus of the experts' distributions. \( P \) is an optimized result from minimizing the relative distance according to a probability distance measure function \( f \),

\[
\arg \min_{P_1, P_2, \ldots, P_n} f(Q\|P).
\] (3.8)

For multiple experts, we wish to find the optimal \( P \) which minimizes the sum of the final distribution to each expert distribution,

\[
\arg \min_{p_1, p_2, \ldots, p_n} \sum_{i=1}^{m} f(Q_i\|P).
\] (3.9)

For non-symmetric measures, depending on which direction one chooses to look at the problem, the functions in Table 3.2 can be minimized with respect to \( P \), and \( w \), (or just \( P \) assuming equal weights of \( w = 1/m \)).

**Table 3.2. Divergence and distance measures investigated.**

<table>
<thead>
<tr>
<th>Name</th>
<th>Label</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kullback Leibler divergence</td>
<td>( K(Q|P) )</td>
<td>( \sum_{j=1}^{n} q_j \ln \left( \frac{q_j}{p_j} \right) )</td>
</tr>
<tr>
<td>Hellinger distance</td>
<td>( H(Q|P) )</td>
<td>( \frac{1}{2} \sum_{j=1}^{n} \left( \sqrt{p_j} - \sqrt{q_j} \right)^2 )</td>
</tr>
<tr>
<td>Total Variation distance</td>
<td>( V(Q|P) )</td>
<td>( \frac{1}{2} \sum_{j=1}^{n}</td>
</tr>
<tr>
<td>Chi-Squared divergence</td>
<td>( C(Q|P) )</td>
<td>( \sum_{j=1}^{n} \left( \frac{p_j - q_j}{p_j} \right)^2 )</td>
</tr>
<tr>
<td>Jensen-Shannon divergence</td>
<td>( J(Q|P) )</td>
<td>( \frac{1}{2} \sum_{j=1}^{n} q_j \ln \left( \frac{2q_j}{p_j + q_j} \right) + \frac{1}{2} \sum_{j=1}^{n} p_j \ln \left( \frac{2p_j}{p_j + q_j} \right) )</td>
</tr>
<tr>
<td>Bhattacharyya distance</td>
<td>( B(Q|P) )</td>
<td>(-\log \left( \sum_{j=1}^{n} \sqrt{p_j q_j} \right) )</td>
</tr>
<tr>
<td>Triangular divergence</td>
<td>( T(Q|P) )</td>
<td>( \sum_{j=1}^{n} \left( \frac{p_j - q_j}{p_j + q_j} \right)^2 )</td>
</tr>
</tbody>
</table>
3.3.1 Desired Properties

Ideally a distance measure would be a metric, or one which satisfies non-negativity, symmetry, the triangle inequality, and the identity of indiscernibles. Each of these is now defined for a distance function $f$ on two probability functions $P$ and $Q$.

**Definition 3.7** A function $f$ is non-negative if $f(P∥Q) \geq 0$

**Definition 3.8** A function $f$ is symmetric if $f(P∥Q) = f(Q∥P)$

**Definition 3.9** A function $f$ satisfies the triangle inequality if for a third probability distribution $R$, $f(P∥Q) \leq f(P∥R) + f(R∥Q)$

**Definition 3.10** A function $f$ satisfies the identity of indiscernibles when $f(P∥Q) = 0$ if, and only if $P = Q$.

In order to be a true metric, a distance function must satisfy all four criteria. Certain combinations of the four are labeled as alternative types of metrics. Table 3.3 lists the five different types of metrics and which definitions are required to hold for each.

**Table 3.3. Requirements for the different types of metrics.**

<table>
<thead>
<tr>
<th>Desired Properties</th>
<th>Non-Negativity</th>
<th>Symmetric</th>
<th>Triangle Ineq.</th>
<th>I.I.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metric</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Pseudometric</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quasimetric</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Semimetric</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Premetric</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A well known approach to calculating distances between probability distributions is based upon the foundations of information theory [19]. The field has evolved from originally solving problems of data compression and communication theory (serving as the theory behind MP3 and .ZIP files), to involve fields such as statistical inference, medical sciences, and plagiarism. The broad term of “information” in this context is regarded as a quantifying measure of the amount of information in an outcome from a probability distribution.
**Definition 3.11** Information is denoted \( I(p(x)) = \log(1/p(x)) = -\log(p(x)) \) where \( p(x) \) is the prior probability of the outcome that occurred.

Often the logarithms are expressed in base two, or bits, however when dealing with statistical distributions, base \( e \) is often used, which converts the units to nats. For the purposes of this paper the logarithm will be defined in base \( e \) unless otherwise noted. Information values are inversely related to the likelihood of the event. Highly probable events correspond to low information, whereas infrequent events correspond to high information values.

The entropy of a distribution, also known as the expected information, measures the amount of randomness in a single probability distribution.

**Definition 3.12** Entropy is defined as \( H(X) = -\sum p(x) \log p(x) \)

It is easily verifiable between discrete uniform random variables with different possible number of outcomes, that the distribution with fewer outcomes has less entropy, or less randomness. For example a fair coin flip, which has two outcomes, has an entropy of 0.69 (1 in base 2), compared to a roll of a fair six-sided die which has an entropy of 1.79 (2.58 in base 2).

### 3.3.2 Kullback Leibler Divergence

Relative entropy, also known as Kullback-Leibler (K) divergence, is one way to quantify the distance between two distributions. In statistical terms it can be thought of as an expected logarithm of the likelihood ratio. For two distributions \( P \) and \( Q \) the K divergence is defined as

\[
K(Q \| P) = \sum_{j=1}^{n} q_j \ln \left( \frac{q_j}{p_j} \right).
\]  

(3.10)

To put K divergence into context, if we knew the true distribution of a random variable was \( Q \), the code needed to describe the random variable could be constructed with average
length $H(Q)$. However, if instead we used distribution $P$, the code would need to be $H(Q) + K(Q\|P)$ (bits for base 2) long on average to describe the random variable.

As discussed in the previous section, divergence measures exhibit faults which vary in degree from minor to the point of an absolute breakdown. An oddity for K divergence is that it is only a premetric, but yet the most popular divergence measure. Additionally, it is not a symmetric measure. In most general cases, distances are thought to be equally measured in terms of moving from point A to B versus moving from point B to A, however this is not the case in K divergence. As a result, the triangle inequality does not hold for K divergence. Immediately, one might begin to question the validity of this distance measure, since it does not satisfy the metric properties that many people believe to be necessary in terms of a defining distance.

Through the use of Lagrangian multipliers, it can be shown that the minimizing pooling operators to K divergence are the arithmetic/linear and geometric/logarithmic means [1]. These calculations are repeated now for comparison to the other newly constructed optimal pooling operators.

With multiple experts, we wish to choose the $P$ which minimizes

$$\sum_{i=1}^{m} w_i K(Q_i \| P).$$

Let

$$P_{K(Q\|P)} = \arg \min_{p_1, p_2, \ldots, p_n} \left[ w_1 \sum_{j=1}^{n} q_{1j} \ln \left( \frac{q_{1j}}{p_j} \right) + \cdots + w_m \sum_{j=1}^{n} q_{mj} \ln \left( \frac{q_{mj}}{p_j} \right) \right]$$

where $\sum p_j = 1$ and $p_j \geq 0$. By applying the method of Lagrangian multipliers, we have

$$L_{K(Q\|P)} = \left[ w_1 \sum_{j=1}^{n} q_{1j} \ln \left( \frac{q_{1j}}{p_j} \right) + \cdots + w_m \sum_{j=1}^{n} q_{mj} \ln \left( \frac{q_{mj}}{p_j} \right) \right] - \lambda \left[ \sum_{j=1}^{n} p_j - 1 \right]$$

$$\frac{\partial L_{K(Q\|P)}}{\partial p_j} = -\frac{w_1 q_{1j}}{p_j} - \cdots - \frac{w_m q_{mj}}{p_j} - \lambda$$
which we then set to 0 and solve
\[ \frac{-\sum_{i=1}^{m} w_i q_{ij}}{p_j} = \lambda \quad \text{to find} \quad p_j = \frac{-\sum_{i=1}^{m} w_i q_{ij}}{\lambda}. \]

Similarly, we set
\[ \frac{\partial L_{K(Q||P)}}{\partial \lambda} = \sum_{j=1}^{n} p_j - 1 = 0 \quad \text{and find} \quad \sum_{j=1}^{n} p_j = 1. \]

Therefore,
\[ \lambda = -\sum_{j=1}^{n} p_j = -\sum_{i=1}^{m} w_i = -1 \quad \text{and} \quad p_j = \sum_{i=1}^{m} w_i q_{ij}. \]

Since K divergence is not symmetric, looking at this measure in the opposite direction is necessary. Switching the order from the previous example, the K divergence for two distributions P and Q is defined as
\[ K(P||Q) = \sum_{j=1}^{n} p_j \ln \left( \frac{p_j}{q_j} \right). \]

With multiple experts, we wish to choose the P which minimizes
\[ \sum_{i=1}^{m} w_i K(P||Q_i). \]

Let
\[ P_{K(P||Q)} = \arg \min_{p_1, p_2, \ldots, p_n} \left[ w_1 \sum_{j=1}^{n} p_j \ln \left( \frac{p_j}{q_{1j}} \right) + \cdots + w_m \sum_{j=1}^{n} p_j \ln \left( \frac{p_j}{q_{mj}} \right) \right] \]
where \( \sum p_j = 1 \) and \( p_j \geq 0 \). By applying the method of Lagrangian multipliers, we have
\[ L_{K(P||Q)} = \left[ w_1 \sum_{j=1}^{n} p_j \ln \left( \frac{p_j}{q_{1j}} \right) + \cdots + w_m \sum_{j=1}^{n} p_j \ln \left( \frac{p_j}{q_{mj}} \right) \right] - \lambda \left[ \sum_{j=1}^{n} p_j - 1 \right] \]
\[ \frac{\partial L_{K(P||Q)}}{\partial p_j} = w_1 \left[ 1 + \ln \left( \frac{p_j}{q_{1j}} \right) \right] + \cdots + w_m \left[ 1 + \ln \left( \frac{p_j}{q_{mj}} \right) \right] - \lambda \]
\[
\sum_{i=1}^{m} w_i [1 + \ln(p_j) - \ln(q_{ij})] - \lambda
\]

which we then set to 0 and solve

\[
\sum_{i=1}^{m} w_i \ln(p_j) = \lambda - \sum_{i=1}^{m} (1 - \ln(q_{ij}))
\]

\[
\ln(p_j) = \frac{\lambda - \sum_{i=1}^{m} w_i (1 - \ln(q_{ij}))}{\sum_{i=1}^{m} w_i}
\]

\[
\ln(p_j) = \lambda - 1 + \sum_{i=1}^{m} w_i \ln(q_{ij}) \quad \text{to find} \quad p_j = e^{\lambda - 1} \prod_{i=1}^{m} (q_{ij})^{w_i}.
\]

Similarly, we set

\[
\frac{\partial L_{K(P\parallel Q)}}{\partial \lambda} = \sum_{j=1}^{n} p_j - 1 = 0 \quad \text{and find} \quad \sum_{j=1}^{n} p_j = 1 = e^{\lambda - 1} \sum_{j=1}^{n} \prod_{i=1}^{m} (q_{ij})^{w_i}.
\]

Therefore, letting \( l_{K(P\parallel Q)} \) be the normalizing constant

\[
l_{K(P\parallel Q)} = e^{\lambda - 1} = \frac{1}{\sum_{j=1}^{n} \prod_{i=1}^{m} (q_{ij})^{w_i}} \quad \text{and} \quad p_j = l_{K(P\parallel Q)} \prod_{i=1}^{m} (q_{ij})^{w_i}.
\]

The results show that \( K(Q\parallel P) \) is optimized by LinOP, the arithmetic average, and \( K(P\parallel Q) \) is optimized by LogOP, the geometric mean, both of which were described in the axiomatic section.

If one or more of the experts \( (Q_i) \) believes an element to be impossible (which occurred frequently in the elicitation experiment in Chapter 2), and elicits a probability of zero for the respective element (i.e., \( q_{ij} = 0 \)), then depending on which direction one looks at the K divergence, the influence of the elicted zero will have different effects. By continuity it is easy to see that \( 0 \log_{0}^{0} = 0 \) and \( 0 \log_{0}^{0} = 0 \) and \( p \log_{0}^{p} = \infty \).
Assuming that only some of the experts allocate a zero probability to the element, we investigate what happens to the K divergence by having $q_{ij} = 0$. Beginning in the direction of $K(Q\|P)$, where the arithmetic mean is the optimal pooling operator, the distance for the individual expert would be $0 \ln \frac{0}{p_j}$, which is zero no matter what the value of $p_j$. In this case the expert is not involved in the element distance calculation since the distance between zero and any $p_j$ will always be zero, meaning any elements which have a zero just count as “zero distance”. No impact is seen from the zero probability to the overall divergence measure since no added weight is included.

For the experts who elicited a positive probability there is an additional problem. For these experts, $p_j$ cannot equal zero or else there is a breakdown in continuity (distance of $\infty$) since $K(Q\|P) = q \ln(q/0) = \infty$ for $q > 0$. If one expert deems an element to be impossible and others do not, this probability distance measure will not allow the pooled valued to be equal to zero. This phenomenon is known as veto power, is a property of the geometric mean and will therefore break down in this direction. However, the arithmetic mean will not break down since that pooled element value $p_j$ will not result in a zero value.

The second problem is that in terms of $K(P\|Q)$ (when LogOP is optimal), the veto effect is active requiring that $P$ be absolutely continuous with respect to $Q$. That is $p_j$ must equal zero when any $q_{ij} = 0$. In this case, $p_j \ln \frac{p_j}{0}$ will be the distance calculated for the individual expert. Therefore in order to have a non-infinite distance, $p_j$ must be equal to zero, since $0 \log \frac{0}{0} = 0$. This confirms the vetoing nature of the geometric mean. The arithmetic mean will therefore break down (distance of $\infty$) in this direction since $p_j$ will not equal zero. Only a pooling mechanism with veto power, such as the geometric mean, will work in this direction.

The popularity of Kullback Leibler divergence is vast, but due to the problems described in this section, additional pooling operators are constructed in the subsequent sections. Outside of Kullback Leibler divergence, the following distance measures have not
been used to calculate optimal opinion pooling consensus operators outside of this dissertation.

### 3.3.3 Chi-Squared Divergence

Chi-squared divergence, a premetric also known as Kagan’s divergence, represents the squared difference of the two probability distributions, normed by the prevalence of the reference distribution. It is used broadly for tests of equality for two distributions, goodness of fit, and symmetry, but to the authors’ knowledge has not been mentioned as an alternative to Kullback Leibler divergence in constructing a consensus distribution. The divergence measure is non-symmetric so the pooling operator needs to be optimized in both directions. The direction of P to Q will be investigated first followed by the investigation of Q to P.

In the direction of P to Q, the Chi-square divergence for two distributions $Q$ and $P$ is defined as

$$C(P\|Q) = \sum_{j=1}^{n} \frac{(p_j - q_j)^2}{q_j}.$$ 

With multiple experts, we wish to choose the $P$ which minimizes

$$\sum_{i=1}^{m} w_i C(P\|Q_i).$$

Let

$$P_{C(P\|Q)} = \arg\min_{p_1, p_2, \ldots, p_n} \left[ w_1 \sum_{j=1}^{n} \frac{(p_j - q_{1j})^2}{q_{1j}} + \cdots + w_m \sum_{j=1}^{n} \frac{(p_j - q_{mj})^2}{q_{mj}} \right]$$ 

where $\sum p_j = 1$ and $p_j \geq 0$. By applying the method of Lagrangian multipliers, we define

$$L_{C(P\|Q)} = \left[ w_1 \sum_{j=1}^{n} \frac{(p_j - q_{1j})^2}{q_{1j}} + \cdots + w_m \sum_{j=1}^{n} \frac{(p_j - q_{mj})^2}{q_{mj}} \right] - \lambda \left[ \sum_{j=1}^{n} p_j - 1 \right]$$

and

$$\frac{\partial L_{C(P\|Q)}}{\partial p_j} = 2w_1 \frac{(p_j - q_{1j})}{q_{1j}} + \cdots + 2w_m \frac{(p_j - q_{mj})}{q_{mj}} - \lambda.$$
which we then set to zero and solve

\[
\sum_{i=1}^{m} \frac{2w_{i}p_{j}}{q_{ij}} - 2 = \lambda \quad \text{to find} \quad p_{j} = \frac{\lambda/2 + 1}{\sum_{i=1}^{m} \frac{w_{i}}{q_{ij}}}.
\]

Similarly, we set

\[
\frac{\partial L_{C(P\|Q)}}{\partial \lambda} = \sum_{j=1}^{n} p_{j} - 1 = 0 \quad \text{and find} \quad \sum_{j=1}^{n} p_{j} = 1 = \sum_{j=1}^{n} \frac{\lambda/2 + 1}{\sum_{i=1}^{m} \frac{w_{i}}{q_{ij}}}.
\]

Therefore, defining \( l_{C(P\|Q)} \) as the normalization constant

\[
\lambda/2 + 1 = \sum_{j=1}^{n} \left[ \frac{1}{\sum_{i=1}^{m} \frac{w_{i}}{q_{ij}}} \right] = l_{C(P\|Q)} \quad \text{and} \quad p_{j} = \left[ \frac{l_{C(P\|Q)}}{\sum_{i=1}^{m} \frac{w_{i}}{q_{ij}}} \right].
\]

Notice that the veto effect is active under \( C(P\|Q) \), implying any time a \( q_{ij} = 0 \), \( p_{j} \) must also be equal to zero to prevent an infinite distance measure.

Since Chi-Square divergence is not symmetric the divergence is investigated in the opposite direction. In the direction of \( Q \) to \( P \), the \( C \) divergence for two distributions \( Q \) and \( P \) is defined as

\[
C(Q\|P) = \sum_{j=1}^{n} \frac{(p_{j} - q_{j})^{2}}{p_{j}}.
\]

With multiple experts, we wish to choose the \( P \) which minimizes

\[
\sum_{i=1}^{m} w_{i}C(Q_{i}\|P).
\]

Let

\[
P_{C(Q\|P)} = \arg \min_{p_{1}, p_{2}, \ldots, p_{n}} \left[ w_{1} \sum_{j=1}^{n} \frac{(p_{j} - q_{1j})^{2}}{p_{j}} + \cdots + w_{m} \sum_{j=1}^{n} \frac{(p_{j} - q_{mj})^{2}}{p_{j}} \right]
\]

where \( \sum p_{j} = 1 \) and \( p_{j} \geq 0 \). By applying the method of Lagrangian multipliers, we define

\[
L_{C(Q\|P)} = \left[ w_{1} \sum_{j=1}^{n} \frac{(p_{j} - q_{1j})^{2}}{p_{j}} + \cdots + w_{m} \sum_{j=1}^{n} \frac{(p_{j} - q_{mj})^{2}}{p_{j}} \right] - \lambda \left[ \sum_{j=1}^{n} p_{j} - 1 \right]
\]
\[
\frac{\partial L_{C(Q\parallel P)}}{\partial p_j} = w_1 \left( 1 - \frac{q_{1j}^2}{p_j^2} \right) + \cdots + w_m \left( 1 - \frac{q_{mj}^2}{p_j^2} \right) - \lambda
\]

which we then set to 0 and solve

\[
\sum_{j=1}^{m} w_i \left( 1 - \frac{q_{ij}^2}{p_j^2} \right) = \lambda \quad \text{to find} \quad p_j = \frac{\sqrt{\sum_{i=1}^{m} w_i q_{ij}^2}}{\sqrt{\lambda - 1}}.
\]

Similarly, we set

\[
\frac{\partial L_{C(Q\parallel P)}}{\partial \lambda} = \sum_{j=1}^{n} p_j - 1 = 0 \quad \text{and find} \quad \sum_{j=1}^{n} p_j = 1 = \sum_{j=1}^{n} \frac{\sqrt{\sum_{i=1}^{m} w_i q_{ij}^2}}{\sqrt{\lambda - 1}}.
\]

Therefore, defining \( l_{C(Q\parallel P)} \) as the normalization constant

\[
\sqrt{\lambda - 1} = \sum_{j=1}^{n} \sqrt{\sum_{i=1}^{m} w_i q_{ij}^2} = l_{C(Q\parallel P)} \quad \text{and} \quad p_j = \frac{1}{l_{C(Q\parallel P)}} \sqrt{\sum_{i=1}^{m} w_i q_{ij}^2}.
\]

Note under \( C(Q\parallel P) \) there is no veto effect needed for calculating the optimal distribution. However, as opposed to K divergence those experts who denote a zero for an element still have a distance of \( p_j \) associated with them. As \( p_j \) moves away from zero, the expert’s zero value still holds some weight in the calculation of that element.

3.3.4 Total Variation Distance

Total Variation distance is a metric which is used often in engineering, numerical analysis of differential equations, and image denoising [12]. The Total Variation distance \( V \) for two distributions \( Q \) and \( P \) is defined as

\[
V(Q\parallel P) = \sup |P(x) - Q(x)|.
\]
For discrete probability distributions an equivalent function to use for minimization is

\[ V(Q\|P) = \frac{1}{2} \sum_{j=1}^{n} |p_j - q_j|. \]

With multiple experts, we wish to choose the \( P \) which minimizes

\[ \sum_{i=1}^{m} w_i V(Q_i\|P). \]

Let

\[ P_V = \arg \max_{p_1, p_2, \ldots, p_n} \left[ \frac{w_1}{2} \sum_{j=1}^{n} |p_j - q_{1j}| + \cdots + \frac{w_m}{2} \sum_{j=1}^{n} |p_j - q_{mj}| \right] \]

where \( \sum p_j = 1 \) and \( p_j \geq 0 \). By applying the method of Lagrangian multipliers, we have

\[ L_V = \left[ \frac{w_1}{2} \sum_{j=1}^{n} |p_j - q_{1j}| + \cdots + \frac{w_m}{2} \sum_{j=1}^{n} |p_j - q_{mj}| \right] - \lambda \left[ \sum_{j=1}^{n} p_j - 1 \right]. \]

Setting

\[ \frac{\partial L_V}{\partial p_j} = \frac{w_1}{2} \frac{p_j - q_{1j}}{|p_j - q_{1j}|} + \cdots + \frac{w_m}{2} \frac{p_j - q_{mj}}{|p_j - q_{mj}|} - \lambda = 0 \]

we find

\[ \sum_{i=1}^{m} w_i (p_j - q_{ij}) = 2\lambda \quad \text{or} \quad 2\lambda = \sum_{i=1}^{m} \left[ w_i 1_{(p_j > q_{ij})} - w_i 1_{(p_j < q_{ij})} \right]. \]

Therefore there is no optimal \( p_j \) as it disappears in the derivative. Since there is no unique solution for \( P \), this divergence measure is abandoned as a potential candidate for use as a consensus operator.

Through an example it is easy to see why there is no optimal \( P \). Suppose that there are two experts giving probability distributions about a binary event. The first expert denotes a 20% chance of success, while the second expert denotes a 40% chance of success. The two experts’ distributions are \( q_1 = [0.20, 0.80] \) and \( q_2 = [0.40, 0.60] \). Finding the optimal \( P \) which minimizes \( \sum_{i=1}^{m} w_i V(Q_i\|P) \), leads to a non-unique solution. Any solution of \( p = [x, (1 - x)] \)
for $x \in [0.2, 0.4]$ will result in the same minimum aggregated distance under equal expert weights. Suppose $P$ was $[0.25, 0.75]$, the aggregated distance under equal weighting is $(w_1 \times 0.05)$ for the first expert and $(w_2 \times 0.15)$ for the second expert. Under equal weighting of $\frac{1}{2}$ for each expert, the total sums to 0.10. Similarly, if $P$ is $[.22, .78]$, the aggregated distance would be $(w_1 \times 0.02)$ for the first expert and $(w_2 \times 0.18)$ for the second expert. Under equal weighting again, the total also sums to 0.10. The distributions $[0.25, 0.75]$ and $[0.22, 0.78]$ under equal weighting both have the same minimum total variation distance showing that a unique solution is not always attainable.

### 3.3.5 Hellinger Distance

Hellinger divergence (H), is a metric bounded between zero and one and named after Ernst Hellinger. H divergence is used often in density estimation, in particular for discrete data when maximum likelihood methods fail ([5], [33]). For two distributions $Q$ and $P$, the Hellinger distance is defined as

$$H(Q \parallel P) = \frac{1}{2} \sum_{j=1}^{n} (\sqrt{p_j} - \sqrt{q_j})^2.$$ 

With multiple experts, we wish to choose the $P$ which minimizes

$$\sum_{i=1}^{m} w_i H(Q_i \parallel P).$$

Let

$$P_H = \arg \min_{p_1, p_2, \ldots, p_n} \left[ \frac{w_1}{2} \sum_{j=1}^{n} (\sqrt{p_j} - \sqrt{q_{1j}})^2 + \cdots + \frac{w_m}{2} \sum_{j=1}^{n} (\sqrt{p_j} - \sqrt{q_{mj}})^2 \right]$$

where $\sum p_j = 1$ and $p_j \geq 0$. By applying the method of Lagrangian multipliers, we have

$$L_H = \left[ \frac{w_1}{2} \sum_{j=1}^{n} (\sqrt{p_j} - \sqrt{q_{1j}})^2 + \cdots + \frac{w_m}{2} \sum_{j=1}^{n} (\sqrt{p_j} - \sqrt{q_{mj}})^2 \right] - \lambda \left[ \sum_{j=1}^{n} p_j - 1 \right].$$
Setting
\[ \frac{\partial L_H}{\partial p_j} = \frac{w_1}{2} \left( \frac{\sqrt{p_j} - \sqrt{q_{ij}}}{\sqrt{p_j}} \right) + \cdots + \frac{w_m}{2} \left( \frac{\sqrt{p_j} - \sqrt{q_{mj}}}{\sqrt{p_j}} \right) - \lambda = 0 \]
we find
\[ \sum_{i=1}^{m} w_i \left( 1 - \frac{\sqrt{q_{ij}}}{\sqrt{p_j}} \right) = 2\lambda \quad \text{so that} \quad p_j = \frac{(\sum_{i=1}^{m} w_i \sqrt{q_{ij}})^2}{(1 - 2\lambda)^2}. \]

Similarly, we set
\[ \frac{\partial L_H}{\partial \lambda} = \sum_{j=1}^{n} p_j - 1 = 0 \quad \text{and find} \quad \sum_{j=1}^{n} p_j = 1 = \sum_{j=1}^{n} \frac{(\sum_{i=1}^{m} w_i \sqrt{q_{ij}})^2}{(1 - 2\lambda)^2}. \]

Therefore, defining \( l_H \) as the normalization constant, we have
\[ (1 - 2\lambda)^2 = \sum_{j=1}^{n} \left[ \left( \sum_{i=1}^{m} w_i \sqrt{q_{ij}} \right)^2 \right] = l_H \quad \text{and} \quad p_j = \frac{1}{l_H} \left( \sum_{i=1}^{m} w_i \sqrt{q_{ij}} \right)^2. \]

### 3.3.6 Bhattacharyya Distance

The Bhattacharyya distance (B) is a semi-metric since it does not satisfy the triangle inequality. B divergence is used to determine the amount of overlap between two statistical samples, and is often used in the fields of signal processing, classification, and pattern recognition ([7], [32]). For two distributions \( Q \) and \( P \) the Bhattacharyya distance is defined as
\[ B(Q\| P) = - \log \left( \sum_{j=1}^{n} \sqrt{p_j q_j} \right). \]

With multiple experts, we wish to choose the \( P \) which minimizes
\[ \sum_{i=1}^{m} w_i B(Q_i\| P), \]
or equivalently choose the $P$ which maximizes

$$\sum_{i=1}^{m} w_i \sum_{j=1}^{n} \sqrt{p_j q_{ij}}.$$ 

Let

$$P_B = \arg \max_{p_1, p_2, \ldots, p_n} \left[ w_1 \sum_{j=1}^{n} \sqrt{p_j q_{1j}} + \cdots + w_m \sum_{j=1}^{n} \sqrt{p_j q_{mj}} \right]$$

where $\sum p_j = 1$ and $p_j \geq 0$. By applying the method of Lagrangian multipliers, we have

$$L_B = \left[ w_1 \sum_{j=1}^{n} \sqrt{p_j q_{1j}} + \cdots + w_m \sum_{j=1}^{n} \sqrt{p_j q_{mj}} \right] - \lambda \left[ \sum_{j=1}^{n} p_j - 1 \right].$$

Setting

$$\frac{\partial L_B}{\partial p_j} = \frac{w_1 q_{1j}}{2 \sqrt{p_j q_{1j}}} + \cdots + \frac{w_m q_{mj}}{2 \sqrt{p_j q_{mj}}} - \lambda = 0,$$

we find

$$\sum_{i=1}^{m} \frac{w_i q_{ij}}{\sqrt{p_j q_{ij}}} = 2\lambda \quad \text{so that} \quad p_j = \frac{(\sum_{i=1}^{m} w_i \sqrt{q_{ij}})^2}{(2\lambda)^2}.$$ 

Similarly, we set

$$\frac{\partial L_B}{\partial \lambda} = \sum_{j=1}^{n} p_j - 1 = 0 \quad \text{and find} \quad \sum_{j=1}^{n} p_j = 1 = \sum_{j=1}^{n} \frac{(\sum_{i=1}^{m} w_i \sqrt{q_{ij}})^2}{(2\lambda)^2}.$$ 

Therefore, defining $l_B$ as the normalization constant, we have

$$(2\lambda)^2 = \sum_{j=1}^{n} \left[ \left( \sum_{i=1}^{m} w_i \sqrt{q_{ij}} \right)^2 \right] \quad \text{and} \quad p_j = \frac{1}{l_B} \left( \sum_{i=1}^{m} w_i \sqrt{q_{ij}} \right)^2.$$ 

Notice that the optimal $P$ for the Bhattacharyya distance is the same as the Hellinger distance. Table (3.4) below shows the optimized pooling operators for all the divergence measures discussed thus far.
Table 3.4. Closed form solutions for divergence measures.

<table>
<thead>
<tr>
<th>Divergence/Distance Measure</th>
<th>Closed Form Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K(\mathcal{Q} \parallel \mathcal{P})$</td>
<td>$\sum_{i=1}^{m} w_i q_{ij}$</td>
</tr>
<tr>
<td>$K(\mathcal{P} \parallel \mathcal{Q})$</td>
<td>$l \prod_{i=1}^{m} (q_{ij})^{w_i}$</td>
</tr>
<tr>
<td>$C(\mathcal{P} \parallel \mathcal{Q})$</td>
<td>$l(\sum_{i=1}^{m} w_i q_{ij})^{-1}$</td>
</tr>
<tr>
<td>$C(\mathcal{Q} \parallel \mathcal{P})$</td>
<td>$l\sqrt{\sum_{i=1}^{m} w_i q_{ij}^2}$</td>
</tr>
<tr>
<td>$H(\mathcal{Q} \parallel \mathcal{P})$</td>
<td>$l(\sum_{i=1}^{m} w_i \sqrt{q_{ij}})^2$</td>
</tr>
<tr>
<td>$B(\mathcal{Q} \parallel \mathcal{P})$</td>
<td>$l(\sum_{i=1}^{m} w_i / \sqrt{q_{ij}})^2$</td>
</tr>
</tbody>
</table>

### 3.3.7 Jensen-Shannon Divergence

The distance and divergence measures discussed up to this point have been shown to have closed form optimal solutions (outside of the non-unique solution for $\text{V}(\mathcal{Q} \parallel \mathcal{P})$ under equal expert weighting). This subsection and the next on Jensen-Shannon divergence and Triangular divergence have not yet been found to have closed form solutions when the applying similar standard Lagrangian multiplier methods used to calculate the closed form solutions for other measures. Jensen-Shannon divergence and Triangular divergence are first defined and discussed, followed by an explanation of a designed optimization algorithm used to solve for the optimal $\mathcal{P}$ under each measure.

Jensen-Shannon divergence ($J$), is defined as

$$J(\mathcal{Q} \parallel \mathcal{P}) = \frac{1}{2} K\left(\mathcal{Q} \parallel \frac{\mathcal{P} + \mathcal{Q}}{2}\right) + \frac{1}{2} K\left(\mathcal{P} \parallel \frac{\mathcal{P} + \mathcal{Q}}{2}\right).$$

Although $J$ divergence is calculated using $K$ divergence, $J$ divergence will not exhibit the same problems. An important property of $J$ divergence is that, while it is not itself a metric, it is a square of a metric [22]. Note choosing to minimize on $J$ divergence or the root of $J$ divergence is irrelevant as the root does not impact the choice of the optimal consensus distribution $\mathcal{P}$. $J$ divergence is symmetric in nature, meaning $J(\mathcal{Q} \parallel \mathcal{P}) = J(\mathcal{P} \parallel \mathcal{Q})$. Immediately, this seems to be more of a reliable metric in terms of what most people believe “distance” to imply. Secondly, $J$ divergence does not require that $\mathcal{P}$ be absolutely continuous with respect to $\mathcal{Q}$. For the two
K divergence measures calculated in the J divergence formula there are different effects whenever a $q_{ij} = 0$. For $K(Q\|\frac{1}{2}(P + Q))$ the distance value will always be zero. However, for $K(P\|\frac{1}{2}(P + Q))$, the distance will be $p \ln(2)$. There is therefore an increasingly quantifiable penalty associated with the pooled opinion differing from zero. This adds a much needed benefit over K divergence and if a user wants a symmetric distance which can handle zero valued opinions, K divergence is not an option. Using J divergence as an alternative allows us to calculate an optimized pooled opinion in situations where under K divergence cannot be applied.

Despite being unable to derive a closed form solution for the consensus distribution under J divergence, there are some properties which can be derived about J divergence itself. By using Jensen’s Inequality (Definition 3.13), and the fact that the entropy is non-negative and bounded by $\log(m)$ (Theorems 3.1 and 3.2), it is shown in Theorem 3.3 that $J(Q\|P)$ is non-negative with a maximum distance of $\log(2)$.

**Definition 3.13** Jensen’s Inequality: If $f$ is a real continuous strictly concave function, and $\sum w = 1$, where $w > 0$, then,

$$\sum_{i=1}^{m} w_m f(x_m) \leq \sum_{i=1}^{m} f(w_m x_m). \tag{3.12}$$

**Theorem 3.1** $H(x) \geq 0$. Where $H(x)$ follows from Definition 3.12.

**Proof 3.1** $0 \leq p(x) \leq 1$ implies $\log\left(\frac{1}{p(x)}\right) \geq 0$.

**Theorem 3.2** $H(x) \leq \log(m)$.

**Proof 3.2** By Jensen’s Inequality (Definition 3.13),

$$H(X) = -\sum_{i=1}^{m} p \log p = \sum_{i=1}^{m} p \log \frac{1}{p} \leq \log \left(\sum_{i=1}^{m} \frac{1}{p}\right) = \log(m). \tag{3.13}$$
Theorem 3.3 \( 0 \leq J(Q\parallel P) \leq \log(2) \).

Proof 3.3

\[
J(Q\parallel P) = \frac{1}{2} K\left(Q\parallel \frac{P+Q}{2}\right) + \frac{1}{2} K\left(P\parallel \frac{P+Q}{2}\right)
\]

\[
= \frac{1}{2} \sum \left[ q \log \frac{q}{.5(p+q)} + p \log \frac{p}{.5(p+q)} \right]
\]

\[
= \frac{1}{2} \sum (p+q) \left[ \frac{q}{p+q} \log \frac{2q}{p+q} + \frac{p}{p+q} \log \frac{2p}{p+q} \right]
\]

\[
= \frac{1}{2} \sum (p+q) \left[ \frac{q+p}{p+q} \log(2) + \left( \frac{q}{p+q} \log \frac{q}{p+q} + \frac{p}{p+q} \log \frac{p}{p+q} \right) \right]
\]

\[
= \frac{1}{2} \times 2 \left[ \log(2) - H\left(\frac{Q}{P+Q}, \frac{P}{P+Q}\right) \right].
\]

Since \( H(x) \geq 0 \) by Theorem 3.1, \( J(Q\parallel P) \leq \log(2) \). Since \( K(Q\parallel P) \geq 0 \), this completes the proof.

3.3.8 Triangular Divergence

Triangular divergence (T), the square of a metric, is one of the newer divergence measures to be proposed to combine probability distributions. Defined by Topsoe [53], he proves that there are strong connections between T and K divergence measures. T divergence is defined as

\[
T(Q\parallel P) = \sum_{j=1}^{n} \frac{(p_j - q_j)^2}{p_j + q_j} = \sum_{j=1}^{n} \frac{2p_jq_j}{p_j + q_j}.
\]  (3.14)

However, unlike K divergence, T divergence does not exhibit a zero vetoing structure. How T divergence compares to K and other divergence measures will be investigated in the upcoming chapter. Similarly to J divergence, applying the methods of Lagrangian multipliers does not result in a unique closed form solution. In order to solve for the optimal \( P \) an optimization algorithm is needed, and will now be described.
3.3.9 Optimization Algorithm

Using Lagrangian multipliers for J and T divergence does not lead to a closed solution for $P$. In order to find the $P$ which minimizes each distance an optimization algorithm must be used. Simulated annealing was the first method chosen to find the $P$ which minimizes the distances. This algorithm was chosen since it is designed for finding the global optimum solution of a function given a large search space. Searching for the best solution to a large number of decimal places is not needed in this area, as probabilities containing more than a couple of decimal places start to become indistinguishable in terms of actions for the DM in most instances. The simulated annealing method is an adaptation of the Metropolis-Hastings algorithm, with the pseudo-code written in context of opinion pooling under J divergence described below:

\[
\begin{align*}
&\text{SET } T \text{ (temperature)} \\
&\text{SET } p \text{ (easy starting point is } p = \text{ LinOP)} \\
&\text{SET } p_{best} = p \text{ (First point has to be best so far)} \\
&\text{WHILE } T > \text{(Some frozen } T) \\
&\quad \text{FOR LOOP} \\
&\quad \quad \text{Simulate } p_{new} = \text{ random neighbor of } p \\
&\quad \quad \quad \text{IF } J(Q|p) < J(Q|p_{best}) \text{ THEN set } p_{best} = p \\
&\quad \quad \quad \text{IF } J(Q|p_{new}) < J(Q|p) \text{ THEN set } p = p_{new} \\
&\quad \quad \quad \text{IF } J(Q|p_{new}) > J(Q|p) \\
&\quad \quad \quad \quad \text{THEN set } p = p_{new} \text{ with } \text{prob } \exp([J(Q|p) - J(Q|p_{new})]/T) \\
&\quad \quad \text{SET } T = rT \text{ (reduce temperature } r < 1) \\
&\quad \text{RETURN } p_{best}
\end{align*}
\]

Replacing J divergence with T divergence in the pseudo-code above will give the optimal global solution to the Triangular divergence measure. Depending on the probabilities elicited in $Q$, tweaking is needed for the temperature, the reduction in temperature, and the jump size in the random neighbor. If there is a small range in solutions from the other pooling operators, a faster cooling schedule with small jumps to random neighbors can be
implemented. But when there is a lot of variation in the optimal $P$ from other operators, then it is a good idea to test a larger space with a slower cooling schedule. In some problems, it can be difficult to find the correct jump size and temperature which adequately cover the space. Some trial and error is necessary to make sure the space is being covered by using a counter in the code. An additional step to ensure accurate convergence is to vary the starting points from LinOP and run the algorithm from multiple starting points.

An alternative algorithm I constructed which is significantly faster than the simulated annealing method is now proposed. Although this method is much simpler and will not guarantee a global solution as simulated annealing does under asymptotic time, the results are comparably similar under multiple tests. The pseudo-code below shows that the new search algorithm is similar to simulated annealing but is able to leverage the fact that multiple searches can be conducted simultaneously greatly increasing the speed and efficiency of the algorithm.

```
SET S (number of simulations)
SET eps (epsilon of neighbor)
SET p (easy starting point is p = LinOP)
SET psimBest=p
WHILE eps>(eps limit)
    SET psims = Simulate S random neighbors of psimBest (eps)
    Calculate J(Q|psims) distances
    SET psimNewBest = psim with min(J(Q|psims))
    IF J(Q|psimNewBest)<J(Q|psimBest)
        SET psimBest=psimNewBest
        SET eps=r*eps (where r<1)
RETURN psimBest
```

The simulation of the random neighbors is dependent on epsilon. The method chosen for this algorithm is to add uniform random noise to psimBest with a minimum of negative
eps and maximum of positive eps. A bound must be placed on each proposed neighbor, where each element must be between zero and one.

### 3.4 Applying Consensus Operators to Continuous Probability Distributions

In each of the previous sections of this chapter, the general framework consisted of each of the m experts eliciting single probability estimates for the n hypotheses, with the sum of the probabilities totaling one. There are two situations where this type of elicitation technique is not adequate for a DM. First, the experts did not put a measure of variability around each of the single probability estimates, even though an expert might have more certainty in their estimate about a particular hypothesis. If the DM requires a more thorough evaluation of a particular hypothesis, the DM may ask each expert for variability estimates around each hypothesis in the finite set of choices. By eliciting Beta distributions from each of the experts, the DM can retrieve the information he needs from the experts in order to make a more informed decision.

The second situation where the methods of this chapter can be inadequate for a DM can occur when a DM is interested in eliciting estimates for a continuous variable. For instance, instead of pooling probability estimates directly on the likelihood of an event (elicited as a probability), a DM may be interested in the experts’ opinions on the average monetary holdings of a major terrorist organization (elicited as a parametric distribution). In this case each expert is likely to give the mean and standard deviation of a normal distribution as their opinion rather than a vector of probabilities summing to one.

To accommodate for these new pooling situations, the general framework of pooling vectors of probabilities is adjusted by discretizing the continuous probability distribution into the format described earlier throughout this chapter. This follows from the general workings of Clemen and Winkler where notation for LinOP and LogOP for aggregating discrete event probabilities or continuous distributions are used interchangeably [16]. Therefore the $q_i(\theta)$ described earlier can be interchanged with $f_i(\theta)$ where $f(x)$ is the probability density function.
for whatever density is chosen by the expert to conduct the elicitation. Under this framework, this allows different experts to use different density functions, while still allowing the DM to use any of the pooling operators described in this chapter to derive the consensus distribution. For many problems, some experts may choose to use a heavily tailed distribution while others may prefer a normal distribution. Since it is the height of the density that is being pooled, the DM does not need to dictate which density function to use to the expert.

In order to apply each of the pooling operators derived from the distance measures in Table 3.2, the continuous distributions elicited from the experts are transformed into a discrete format, by converting the heights of the density into a vector of probability estimates. When pooling continuous distributions across multiple experts it is not appropriate to directly pool the parameters of the distribution. Not utilizing the structure of the density of the continuous distribution can lead to the DM making mistakes in assuming that parameters scale in the same way of the shape of the distribution. For example, pooling two Beta distributions of \((\alpha = 1, \beta = 3)\) and \((\alpha = 3, \beta = 1)\) using LinOP with equal weights will not look anything close to a Beta \((\alpha = 2, \beta = 2)\) distribution. This example is shown in more detail in the next chapter.

As a result of pooling the density values and not the parameters, the final consensus distribution need not be a probability distribution. To ensure the probability distribution is coherent (i.e. 100% probability) the distribution is integrated and normalized. Since the distributions elicited from the experts are continuous and there is often no distinct minimum or maximum value to the distribution, the range of integration needs to be defined by the DM. In practice it is recommended the low end \((\theta_L)\) be set to be the minimum of first percentiles from each of the experts’ continuous probability distributions. Similarly, the upper end of integration \((\theta_U)\) can be set to the maximum of the 99th percentiles from each of the experts continuous probability distributions. Ultimately, the DM should have control of the range of integration, adjusting depending on the problem at hand.
Also needed to calculate the normalization constant by integration is the number of subdivisions between $\theta_L$ and $\theta_U$. In the examples worked in Chapter 4, the number of subdivisions was set to 1000 for the distance measures with closed form solutions. Due to the complexity of the search algorithms when using Jensen-Shannon and Triangular divergence, the search space was set to a much smaller level of 100 subdivisions. While the number of subdivisions in the analysis in Chapter 4 is fixed, it is recommended the final number be chosen by the DM, being reconsidered before each problem. A single outlying distribution can have significant effects in the level of discreteness across the multiple continuous probability distributions.

The recommended determinations of the weights $w_i$, are similar to those when pooling in the discrete case. Nothing about changing from discrete measurements to continuous measurements should impact the relative performance of one expert against another.
CHAPTER 4
POOLING OPERATOR PERFORMANCE USING
EXPERIMENT DATA

The relative performances of the consensus pooling operators derived through information theory methods in Chapter 3 are now investigated using some simple hypothetical examples, followed by evaluations using the data from the elicitation experiments discussed in Chapter 2. If the newly derived consensus operators do not perform as adequate as the current methods (LinOP and LogOP) there would be little need to call for a change in general practices. Opinion pooling operator performance for the operators listed in Table 4.1 is measured via the log probability scoring rule (2.1). The larger the log probability score, the better the performance of the elicited distribution. Overall rankings of the consensus opinions in relation to one another and also against individual’s opinions allow for performance based comparisons across each operator. More simply, the general hypothesis to be tested is that each pooling operator is equally likely to perform the best. However, alternatively, due to the fact that some pooling operators are known to exhibit different properties, is there sufficient evidence to claim that some consensus pooling operators outperform others?

Table 4.1. Consensus operators analyzed using experiment data.

<table>
<thead>
<tr>
<th>Pooling Operator</th>
<th>Closed Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>LinOP</td>
<td>$\sum_{i=1}^{m} w_i q_{ij}$</td>
</tr>
<tr>
<td>LogOP</td>
<td>$l \prod_{i=1}^{m} (q_{ij})^{w_i}$</td>
</tr>
<tr>
<td>Chi-Square ($P \parallel Q$)</td>
<td>$l(\sum_{i=1}^{m} \frac{w_i}{q_{ij}})^{-1}$</td>
</tr>
<tr>
<td>Chi-Square ($Q \parallel P$)</td>
<td>$l \sqrt{\sum_{i=1}^{m} w_i q_{ij}^2}$</td>
</tr>
<tr>
<td>Hellinger/Bhattacharyya</td>
<td>$l(\sum_{i=1}^{m} w_i \sqrt{q_{ij}})^2$</td>
</tr>
<tr>
<td>Jensen Shannon</td>
<td>Optimization Algorithm</td>
</tr>
<tr>
<td>Triangular</td>
<td>Optimization Algorithm</td>
</tr>
</tbody>
</table>
Investigating the performance of the consensus operators using ranked values is necessitated by the extremities that many individual and consensus opinions exhibit from log probability scoring. When the correct value to a question happens to be in the extreme tail of an elicited distribution, it leads to extremely low log probability scores. If an individual elicits a probability of zero to the eventual correct outcome, the log probability score is $-\infty$, making averages across multiple individuals useless for comparative analysis. Simplifying to rank values allows for an easier understanding of the performance of an operator without allowing for a single individual to determine the overall performance with an outlying observation.

For this analysis, three elicitation experiments supply the data for testing how the pooling operators perform against one another and against individual opinions. Fifteen questions from the initial elicitation experiment, combined with six questions from the secondary ELICIT/SHELF experiment, as well as ten questions from an outside experiment “Crystal Ball Competition” [50] leads to a total of 31 questions to measure the performance of the consensus pooling operators.

In the following subsections, simple examples are worked through first to give readers an understanding of the differences across pooling operators. Results from each experiment from Chapter 2 are then analyzed separately, followed by a combined analysis, and discussion of each of the consensus opinion pooling operators. Finally, general recommendations on when to use each consensus operator are given.

4.1 Demonstration of Consensus Operators

To highlight the differences across each pooling mechanism some informal examples are worked through ranging from trivial to more complex. The first subsection works through simple examples where multiple experts elicited vector valued probability estimates with no measure of error around each hypothesis. The second subsection shows additional examples where multiple experts elicited probability density functions.
4.1.1 Examples Pooling Probability Estimates

The first example is a simple binary event with two experts. The first expert believes an even chance of the two hypotheses, while the second expert believes very strongly in the first hypothesis. The two distributions elicited are $Q_1 = [0.50, 0.50]$ and $Q_2 = [0.99, 0.01]$. The resulting optimal opinions are listed in Table 4.2. The pooled values on the probability of occurrence of the first hypothesis vary quite significantly from 0.689 to 0.971. The pooling operators which exhibit veto power $K(P\|Q)$ and $C(P\|Q)$ stray away from the others, already showing polarization. $C(Q\|P)$ maintains the highest level of uniformity, a desired effect in instances where minimizing outlying observations when experts all have the same relative overlapping knowledge base.

Under the same pretenses, imagine the second expert now is absolutely positive on the occurrence and changes his opinion from 99% to 100%, so $Q_2 = [1, 0]$. Notice the veto effect taking place for $K(P\|Q)$ and $C(P\|Q)$ in Table 4.2. Movement across the other pooling operators is not as drastic when compared to the results when $Q_2 = [.99, .01]$, with most moving hardly at all.

**Table 4.2. Pooled distribution ($P$) from $Q_1 = [0.50, 0.50]$ and $Q_2$ under equal weighting.**

<table>
<thead>
<tr>
<th>Divergence/Distance Measure</th>
<th>$Q_2 = [.99, .01]$</th>
<th>$Q_2 = [1, 0]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K(Q|P)$</td>
<td>$p_1$</td>
<td>$p_2$</td>
</tr>
<tr>
<td></td>
<td>.745</td>
<td>.255</td>
</tr>
<tr>
<td>$K(P|Q)$</td>
<td>.909</td>
<td>.091</td>
</tr>
<tr>
<td>$C(P|Q)$</td>
<td>.971</td>
<td>.023</td>
</tr>
<tr>
<td>$C(Q|P)$</td>
<td>.689</td>
<td>.311</td>
</tr>
<tr>
<td>$H(Q|P)$ and $B(Q|P)$</td>
<td>.816</td>
<td>.184</td>
</tr>
<tr>
<td>$J(Q|P)$</td>
<td>.798</td>
<td>.202</td>
</tr>
<tr>
<td>$T(Q|P)$</td>
<td>.770</td>
<td>.230</td>
</tr>
</tbody>
</table>

The same scenario is now extended to account for a variety of opinions from the second expert by pooling $Q_1 = [0.50, 0.50]$ and $Q_2 = [x, (1-x)]$, where $x \in [0, 1]$. Figure 4.1 shows the results of the first hypothesis ($p_1$) as $x$ varies from 0 to 1. No matter the opinion
of the second expert, using $C(P\|Q)$ as the distance measure will always result in the consensus opinion having the highest level of uncertainty ($p_1$ closest to 0.5). The vetoing nature of $K(P\|Q)$, and $C(P\|Q)$, can be seen easily with the sequences terminating at points with no uncertainty ([1,0] and [0,1]). Additionally, the distance measure choice is essentially irrelevant when $x$ ranges from 0.4 to 0.6, as the results are practically equivalent across all distance measures. However, when the two opinions are far from one another, the choice of distance measure has a significant effect.

![Figure 4.1. Results of first opinion ($p_1$) when pooling (0.5, 0.5) and (x,1-x) for each consensus operator.](image)

As noted before, it is a very common obstacle in opinion pooling that experts tend to be much too overconfident when eliciting opinions [42] [38]. By using the divergence measure $C(Q\|P)$ over the standard arithmetic average when using $K(Q\|P)$ the decision
maker is able to better limit the overconfidence in the pooled opinion. Table 4.3 shows the rank order in terms of maintaining a more conservative consensus opinion with a higher degree of uncertainty when pooling two opinions, with the first fixed at absolute uncertainty (0.5,0.5).

Table 4.3. Ordered levels of uncertainty for distance measures when pooling \( Q_1 = [0.50, 0.50] \) and \( Q_2 = [x, 1 - x] \) under equal weighting.

<table>
<thead>
<tr>
<th>Divergence/Distance Measure</th>
<th>Uncertainty Ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>( C(Q ∥ P) )</td>
<td>1</td>
</tr>
<tr>
<td>( K(Q ∥ P) )</td>
<td>2</td>
</tr>
<tr>
<td>( T(Q ∥ P) )</td>
<td>3</td>
</tr>
<tr>
<td>( J(Q ∥ P) )</td>
<td>4</td>
</tr>
<tr>
<td>( H(Q ∥ P) ) and ( B(Q ∥ P) )</td>
<td>5</td>
</tr>
<tr>
<td>( K(P ∥ Q) )</td>
<td>6</td>
</tr>
<tr>
<td>( C(P ∥ Q) )</td>
<td>7</td>
</tr>
</tbody>
</table>

Imagine now that we increase the number of experts eliciting parity distributions from one to four, while maintaining one expert who is certain about the first hypothesis occurring. The resulting pooled distributions are listed in Table 4.4. Again even when adding additional experts with no preferences for any hypothesis, \( C(Q ∥ P) \) maintains the lowest level of polarization. However, interestingly \( T(Q ∥ P) \) surpasses the level of uncertainty of \( K(Q ∥ P) \) implying the uncertainty rankings given in Table 4.3 are not fixed across all possibilities.

Table 4.4. Pooled distribution \((P)\) when \( Q_1 = Q_2 = Q_3 = Q_4 = [0.50, 0.50], \) \( Q_5 = [1, 0] \) under equal weighting.

<table>
<thead>
<tr>
<th>Divergence/Distance Measure</th>
<th>( p_1 )</th>
<th>( p_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( K(Q ∥ P) )</td>
<td>.600</td>
<td>.400</td>
</tr>
<tr>
<td>( K(P ∥ Q) )</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>( C(P ∥ Q) )</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>( C(Q ∥ P) )</td>
<td>.586</td>
<td>.414</td>
</tr>
<tr>
<td>( H(Q ∥ P) ) and ( B(Q ∥ P) )</td>
<td>.647</td>
<td>.353</td>
</tr>
<tr>
<td>( J(Q ∥ P) )</td>
<td>.617</td>
<td>.383</td>
</tr>
<tr>
<td>( T(Q ∥ P) )</td>
<td>.596</td>
<td>.404</td>
</tr>
</tbody>
</table>
Imagine now we increase the number of experts eliciting strong opinions in favor of the first hypothesis from one to four, maintaining only one parity expert distribution, and only one with absolute certainty. The resulting pooled distributions are listed in Table 4.5. Again 
\( C(Q\|P) \) exhibits much less polarization compared to the other pooling operators. However, unlike the previous examples, \( T(Q\|P) \) has followed the masses and now has the most parity (excluding the methods with vetoing), proving the relative uncertainty ranking for \( T(Q\|P) \) is not fixed.

**Table 4.5. Pooled distribution \((P)\) when**

\[ Q_1 = [0.50, 0.50], \]
\[ Q_2 = Q_3 = Q_4 = [0.99, 0.01], Q_5 = [1, 0] \]

**under equal weighting.**

<table>
<thead>
<tr>
<th>Divergence/Distance Measure</th>
<th>( p_1 )</th>
<th>( p_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( K(Q|P) )</td>
<td>.894</td>
<td>.106</td>
</tr>
<tr>
<td>( K(P|Q) )</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>( C(P|Q) )</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>( C(Q|P) )</td>
<td>.804</td>
<td>.196</td>
</tr>
<tr>
<td>( H(Q|P) ) and ( B(Q|P) )</td>
<td>.956</td>
<td>.044</td>
</tr>
<tr>
<td>( J(Q|P) )</td>
<td>.960</td>
<td>.040</td>
</tr>
<tr>
<td>( T(Q|P) )</td>
<td>.974</td>
<td>.026</td>
</tr>
</tbody>
</table>

### 4.1.2 Examples Pooling Continuous Probability Distributions

A few simple examples are now worked to show the process and outcomes of pooling continuous probability distributions. The first example shows why directly pooling parameters is not equivalent to pooling density values across distributions. Recall in the previous chapter there was an example mentioned where the DM is interested in pooling two Beta distributions of \( (\alpha = 1, \beta = 3) \) and \( (\alpha = 3, \beta = 1) \) where the argument was made not to pool the parameters of the distribution directly. Figure 4.2 shows the two expert densities (black and blue lines) along with the results from LinOP (green line) and the results of averaging the parameters directly (red line). Note that the results from LinOP and the results
from pooling the parameters are drastically different. The results of LinOP look similar to a mixture distribution, whereas the results from averaging the parameters do not reflect the beliefs of either expert very well. Additionally, in this example LinOP follows the generally desired properties of unanimity (Definition 3.1), and boundedness (Definition 3.4), whereas averaging the Beta parameters does not follow either property.

The next example is designed to show the general properties of pooling two continuous probability distributions, allowing for insights into the general behavior of each pooling operator listed in Table 4.1. Suppose there are two expert distributions, the first normally distribution with $\mu_1 = 12, \sigma_1 = 1$ and the second normally distributed with $\mu_2 = 8, \sigma_2 = 2$. The two expert distributions along with the results from each of the consensus operators are shown in Figure 4.3. The strong vetoing nature discussed earlier when pooling using $K(P\|Q)$ and $C(P\|Q)$, is evident. These two distributions are much narrower than the results from the other consensus operators, and using these consensus operators
operators should be done so sparingly by a DM. The other five consensus operators return fairly similar shaped distributions, but there are clear differences across the different pooling methods. The results from \( K(Q\|P) \) (LinOP) are bounded by the results from the other four remaining methods across the values of \( x \). Evaluating the subtle differences between these consensus operators via aggregation of the data collected from the experiments in Chapter 2 serves as the basis of the rest of this chapter.

![Figure 4.3. Consensus operator results when pooling two normal distributions.](image)

### 4.2 Initial Experiment Results

Recall from Section 2.4.1 the initial elicitation experiment asked 15 questions listed in Table 2.2 (10 elicited as PDFs and 5 as PMFs) regarding financial and political topics to 19 individuals. The 7 consensus operators, in addition to the 19 individual opinions make a total of 26 distinct probability distributions to rank for each of the 15 questions. The average ranks,
the standard deviation of the ranks, and the count of times each consensus operator was the
best performer of the seven are shown in Table 4.6. While the Hellinger/Bhattacharyya
operator had the best average ranking across all questions at 9.4, the Jensen-Shannon operator
was close at 9.7 with one of the two methods being the best consensus operator in almost half
of the questions. However, despite LogOP and Chi-Square \((P \parallel Q)\) not being recommended as
viable solutions due to the vetoing power in the operator, these two methods significantly
outperformed the other consensus operators on the five probability vector elicitations.
Reversely, the Hellinger/Bhattacharyya, Jensen-Shannon, and Triangular operators
outperformed the other distance measures on the ten density elicitations.

One detail that stands out immediately is the simple arithmetic average (LinOP) was
the best pooling operator for only one of the 15 questions. Given that this consensus method
is often cited as being the most popular in the opinion pooling field, it was expected to
perform relatively better than the others. However, this gives some evidence that other opinion
pooling operators may need to be considered in addition to LinOP when choosing a consensus
operator.

Notice also that LogOP and Chi-Square \((P \parallel Q)\) had larger average rankings and much
larger standard deviations compared to the other pooling operators, outside of the five PMFs
where it was actually the best. This variance in performance is due to the extreme vetoing
nature, where the pooled probability distributions spans over the small space where every
individual elicits positive probability. If that small area ends up containing the correct answer,
then the operators perform outstandingly, which is what happened in the five PMF questions.
However, for the continuous questions these two operators returned a density height of zero at
the correct answer, automatically making them the worst opinion under the logarithmic
scoring rule. The trading of excellent and poor behavior can be seen thorough the much larger
standard deviations for these two operators.

Another interesting occurrence is that for some operators the average ranks do not
appear extraordinarily better than an option chosen at random out of the group of experts.
Table 4.6. Operator rankings from Experiment 1.

<table>
<thead>
<tr>
<th>Method</th>
<th>Avg</th>
<th>S.D.</th>
<th>Best</th>
<th>Avg (10 PDFs)</th>
<th>Avg (5 PMFs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LinOP</td>
<td>11.8</td>
<td>3.0</td>
<td>1</td>
<td>10.5</td>
<td>14.4</td>
</tr>
<tr>
<td>LogOP</td>
<td>13.7</td>
<td>8.7</td>
<td>4</td>
<td>15.9</td>
<td>9.5</td>
</tr>
<tr>
<td>Chi-Square ($P\parallel Q$)</td>
<td>19.0</td>
<td>9.0</td>
<td>1</td>
<td>23.9</td>
<td>9.3</td>
</tr>
<tr>
<td>Chi-Square ($Q\parallel P$)</td>
<td>13.6</td>
<td>3.3</td>
<td>1</td>
<td>12.2</td>
<td>16.4</td>
</tr>
<tr>
<td>Hellinger/Bhattacharyya</td>
<td>9.4</td>
<td>2.4</td>
<td>4</td>
<td>8.7</td>
<td>10.8</td>
</tr>
<tr>
<td>Jensen Shannon</td>
<td>9.7</td>
<td>2.8</td>
<td>3</td>
<td>8.6</td>
<td>12.0</td>
</tr>
<tr>
<td>Triangular</td>
<td>10.4</td>
<td>2.8</td>
<td>1</td>
<td>8.5</td>
<td>13.2</td>
</tr>
</tbody>
</table>

Given that there are 26 distributions being ranked, a ranking of 13.5 would be a middle of the road opinion across each of the different questions. While most of the operators did do better than this average ranking, it was not by a large amount. Hellinger/Bhattacharyya, at an average of 9.4 was best, but was not near the top on every question certainly. Although, across all 15 questions the poorest ranking performance of Hellinger/Bhattacharyya was 12th (5 times). Jensen Shannon and Triangular also had very few poor performances across the questions, with their poorest ranks coming in at 13 and 14 respectively.

Comparing the performances of the consensus operators to individual’s distributions shows only 2 individuals had a better average then the Hellinger/Bhattacharyya operator with values of 8.2 and 8.9, implying most individuals answered some questions well and other questions not as well. The experiment did not have the same individuals doing well on all questions and some doing poorly on all questions. Hence it is a possibility that for individual questions the consensus opinions could be bounded by the performances of the individuals since no opinion pooling operator was the single absolute best for any of the 15 questions.

Regarding the initial hypothesis of testing pooling operator equality, there seems to be significant differences in terms of ranked performance. For this experiment the Hellinger/Bhattacharyya performed well along with the optimizing Jensen-Shannon and Triangular operators. Just behind those operators performances were LinOP, and Chi-Square ($Q\parallel P$). Finally, LogOP and Chi-Square ($P\parallel Q$) performed rather poorly on the continuous PDFs and would not be recommended as accurate consensus operators for these types of
distributions. However, a final decision regarding consensus operator equality is delayed until the analysis of each experiment is completed.

### 4.3 Follow Up Experiment Results

The second experiment, which conducted elicitations through ELICIT and SHELF is now analyzed in a similar manner to the previous section. Recall there were 3 questions asked to a total of 20 subjects, through 2 methods. Experts were asked to elicit a normal distribution for the first question, a vector of probabilities for the second, and a beta distribution for the third with the process repeated an additional time. Combined with the 7 pooling operators, this results in 27 probability distributions to rank across each of the 6 questions. This experiment is slightly different than the first in that individuals were asked the same questions twice. Given that the subjects answered fairly similarly across elicitation methods it would be expected to see correlated results across elicitation techniques, especially for Question 2 where there was minimal variation seen between the different elicitation methods.

Table 4.7 displays the rankings (out of 27) for performances across ELICIT and SHELF based on the log scoring rule, the average rank, and the count of times the consensus was the best opinion pooling operator. Again, the best operator is defined as the method generating the largest log probability score. Similar to the results from the first experiment, the Hellinger/Bhattacharyya operator had the best average ranking across the 27 distributions. This operator stood out from the rest in terms of performance with the best score on 3 of 6 questions. Similarly to the first experiment, Hellinger/Bhattacharyya also had the best ranked worst performance, a ranking of 10 in both of the density function elicitation questions.

Also similar to the first experiment, Jensen-Shannon performed well in terms of the count of times it was the best consensus operator. In fact, the order of average rankings across all questions was the same in the first and second experiment. Due to the relatively good performances of the other opinion pooling operators there are still significant doubts remaining that LinOP is the sole operator needed in the opinion pooling field.
Table 4.7. Operator rankings from Experiment 2.

<table>
<thead>
<tr>
<th>Method</th>
<th>ELICIT</th>
<th>SHELF</th>
<th>Avg.</th>
<th>Best</th>
</tr>
</thead>
<tbody>
<tr>
<td>LinOP</td>
<td>10</td>
<td>14</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td>LogOP</td>
<td>22</td>
<td>26</td>
<td>20</td>
<td>18</td>
</tr>
<tr>
<td>Chi-Square (P|Q)</td>
<td>26</td>
<td>27</td>
<td>23</td>
<td>25</td>
</tr>
<tr>
<td>Chi-Square (Q|P)</td>
<td>12</td>
<td>15</td>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td>Hellinger/Bhattacharyya</td>
<td>7</td>
<td>10</td>
<td>8</td>
<td>11</td>
</tr>
<tr>
<td>Jensen Shannon</td>
<td>8</td>
<td>11</td>
<td>6</td>
<td>14</td>
</tr>
<tr>
<td>Triangular</td>
<td>9</td>
<td>13</td>
<td>9</td>
<td>12</td>
</tr>
</tbody>
</table>

Investigating the performance of the operators across SHELF and ELICIT, it is easy to see that the ranked performances are fairly stable even though there was slight variation in individual opinions. Spearman’s rank correlation coefficient was calculated to be 0.547 which is significant to a p-value less than 0.001. The significant correlation across these two methods adds to the evidence that the elicitation methods are fairly stable across attempts given each individual’s internal variation.

Not surprisingly, the pooling operators LogOP and Chi-Square \(P\|Q\) performed rather poorly overall. The vetoing nature of the two consensus operators narrows the variance of the consensus distribution considerably, severely limiting the capabilities of the operators. In some cases the operators resulted in the worst and second to worst opinion of all elicited probability distributions including consensus operators.

4.4 CRYSTAL BALL COMPETITION

The final experiment takes its data from the “Crystal Ball Competition” conducted by Significance magazine in March 2010 [50]. The voluntary mail-in survey inserted into the magazine contained 10 questions about cultural events in the future. To help the subjects, if historical outcomes for the question had occurred in the past, three historical data points were provided. Interestingly, the final question in the competition asked how many entrants there would be in the competition. While answers ranged quite heavily for this question (from 6 to 17.9 million), the correct answer was 51 individuals. Combined with the 7 pooling operators,
there are a total of 58 probability distributions to rank in a similar manner to the previous experiments. The much larger sample size compared to the first two experiments provides additional insight into how the consensus operators compare relatively to numerous individual opinions.

For each of the ten questions, each individual was asked to submit two parameters to a normal distribution by providing a mean and standard deviation for the correct answer. Many questions had wildly varying parameters across individuals, as seen in the final question with one expert eliciting an expected value for of the number of respondents to the survey at 17.9 million and a standard deviation of only 70,000. Outlying distributions were more of an influence on the performance of the different pooling algorithms in this experiment compared to the previous two.

Additionally, for the questions which displayed the outcomes of previous trials, the subjects understandably answered with values similar to those historical results. However, the stability the subjects expected in the outcomes did not always materialize. Many elicited standard deviations were extremely narrow, resulting in multiple questions having the majority of probability distributions with essentially zero probability at the correct answer. It seems providing the individuals with past information added further to the problem of narrowing the range of possible elicited opinions.

The average rank, standard deviation of the ranks, and count of times the consensus operator was the best of the seven consensus operators using log probability scores are displayed in Table 4.8. Unlike the first two experiments, Hellinger/Bhattacharyya did not have the best overall ranking (15.9), but still performed significantly better than the median ranked opinion of 29.5. Chi-Square \((Q\|P)\) performed best in terms of average ranks with a value at 15.3, with LinOP finishing just above at an average ranking of 15.4. The consensus operators from LinOP and Hellinger/Bhattacharyya both performed best on two questions, with those from LogOP, Chi-Square \((Q\|P)\), and Triangular each performing best for a single question. Hellinger/Bhattacharyya had the “least poorest” performance with a maximum rank of 23
while LinOP had a maximum rankings of 25. Each of the consensus operators outside of LogOP and Chi-Square \((Q\|P)\) were better than the median rank of 29.5 for all ten questions.

### Table 4.8. Operator rankings from Experiment 3.

<table>
<thead>
<tr>
<th>Method</th>
<th>Avg</th>
<th>S.D.</th>
<th>Best</th>
</tr>
</thead>
<tbody>
<tr>
<td>LinOP</td>
<td>15.4</td>
<td>5.8</td>
<td>2</td>
</tr>
<tr>
<td>LogOP</td>
<td>50.1</td>
<td>14.2</td>
<td>1</td>
</tr>
<tr>
<td>Chi-Square ((P|Q))</td>
<td>57.2</td>
<td>1.3</td>
<td>0</td>
</tr>
<tr>
<td>Chi-Square ((Q|P))</td>
<td>15.3</td>
<td>8.3</td>
<td>1</td>
</tr>
<tr>
<td>Hellinger/Bhattacharyya</td>
<td>15.9</td>
<td>4.1</td>
<td>2</td>
</tr>
<tr>
<td>Jensen Shannon</td>
<td>17.2</td>
<td>7.1</td>
<td>0</td>
</tr>
<tr>
<td>Triangular</td>
<td>17.7</td>
<td>6.0</td>
<td>1</td>
</tr>
</tbody>
</table>

Extremely poor performances were displayed again by LogOP and Chi-Square \((P\|Q)\), despite LogOP being the best operator for one question. In multiple cases pooling via these operators resulted in a distribution with zero probability at the correct answer making the log probability score of \(-\infty\) the worst of all the opinions. The addition of more and more opinions allows for a greater opportunity for the narrowing of the consensus distribution to occur, contradicting the general notions of opinion pooling that additional opinions imply more accurate consensus opinions.

Only one of the experts was able to outperform the five consensus operators in terms of average rankings outside of the poor performing LogOP and Chi-Square \((P\|Q)\). With an average rank of 11.5, this expert significantly outperformed the other experts as well as the consensus operators. However, the five high performing consensus operators occupied ranks 2 through 6, indicating that a DM is helped by utilizing the consensus operators over any individual opinion chosen at random.

### 4.5 Pooling Operator Conclusions

When analyzing all three experiments simultaneously, general conclusions can be drawn on the performance of the seven consensus operators listed in Table 4.1. Recall the
original hypothesis was to test whether operators outside of LinOP and LogOP can outperform the two major consensus operators already in use. Table 4.9 shows the final counts of each operator performing the best across all 31 questions, the average percentile rank (right inclusive), the standard deviation of the percentiles, and the minimum and maximum percentile. The reason for evaluating on percentile ranks is to accommodate for the different number of experts across the three experiments.

Table 4.9. Consensus operator percentiles across all three experiments.

<table>
<thead>
<tr>
<th>Method</th>
<th>Best Count</th>
<th>Average</th>
<th>S.D.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>LinOP</td>
<td>4</td>
<td>64.7</td>
<td>13.0</td>
<td>38.5</td>
<td>89.7</td>
</tr>
<tr>
<td>LogOP</td>
<td>5</td>
<td>33.2</td>
<td>32.2</td>
<td>3.4</td>
<td>96.2</td>
</tr>
<tr>
<td>Chi-Square (P \parallel Q)</td>
<td>1</td>
<td>16.4</td>
<td>26.8</td>
<td>1.7</td>
<td>96.2</td>
</tr>
<tr>
<td>Chi-Square (Q \parallel P)</td>
<td>4</td>
<td>60.0</td>
<td>16.6</td>
<td>26.9</td>
<td>93.1</td>
</tr>
<tr>
<td>Hellinger/Bhattacharyya</td>
<td>9</td>
<td>70.6</td>
<td>8.5</td>
<td>57.7</td>
<td>84.6</td>
</tr>
<tr>
<td>Jensen Shannon</td>
<td>6</td>
<td>68.8</td>
<td>11.2</td>
<td>51.9</td>
<td>89.7</td>
</tr>
<tr>
<td>Triangular</td>
<td>2</td>
<td>66.6</td>
<td>10.4</td>
<td>50.0</td>
<td>87.9</td>
</tr>
</tbody>
</table>

Clearly, there are alternative operators outside of LinOP and LogOP that were able to successfully build a better consensus distributions measured in terms of log probability scoring. The consensus distributions when using Hellinger/Bhattacharyya distance and Jensen Shannon distance were often the best performing results, implying that further research should be given to these pooling operators as potential alternatives for a DM.

In terms of overall performance, all but two consensus operators performed on average better than the median percentile. Both LogOP and Chi-Square \(P \parallel Q\) performed significantly poorly with average percentiles well below the median. Although they did end up having the best consensus opinion for six of the 31 questions, these two operators are so inconsistent it is unwise to apply them to all situations blindly. For any upcoming question it is extremely difficult to determine when they will result in accurate distributions as seen by the large percentile standard deviations. The varying excellent performances and poor performances by LogOP and Chi-Square \(P \parallel Q\) is a trade-off which is not always acceptable.
if a single pooling operator must be chosen. Therefore, these two methods can not recommended as consensus pooling operators unless there is a specific reason to use them, such as all individuals are working under absolute independent knowledge.

Contrary to the two poor consensus operators, Hellinger/Bhattacharyya performed the best in terms of the average percentile, overall count, and consistency seen through the smaller standard deviation of the percentiles. In fact the minimum percentile across all questions for this operator was 57.7, only slightly below many other consensus operators overall averages. The stability across all questions for the Hellinger/Bhattacharyya operator is also an ideal advantage compared to the variability seen in the other consensus operators. Despite the outstanding results, the Hellinger/Bhattacharyya consensus operator has received no attention in the literature for use in opinion pooling. Given the excellent performance of the operator using Hellinger/Bhattacharyya distance across multiple tests, it appears this consensus operator can outperform LinOP in arriving at a more accurate probability distribution. One hypothesis for explaining the overall success of this operator is that the highest level of uniformity is maintained in the consensus distribution, helping to balance out the well known problem of overconfidence in each individual expert. However to reiterate, one should refrain from extrapolating the results. Further research is clearly needed to prove that LinOP does not perform as well as Hellinger/Bhattacharyya generally, but this research has shown that there are significant doubts that LinOP performs the best in all situations.

Alternatives to Hellinger/Bhattacharyya which also performed well on average include LinOP, Chi-Square ($Q \parallel P$), Triangular, and Jensen-Shannon operators. All of these consensus operators had fairly similar overall average percentiles in the 60’s, well above the median opinion. However, LinOP, and Chi-Square ($Q \parallel P$) were much more inconsistent compared to the Triangular, and Jensen/Shannon operators. The minimum percentile rankings for LinOP, Chi-Square ($Q \parallel P$) were 38.5 and 26.9 respectively, while Triangular, and Jensen-Shannon operators worst performances were at 51.9 and 50.0 respectively. Additionally,
Jensen-Shannon stood out in terms of the overall count of best performances with 6 compared to 4 for LinOP, 2 for Triangular, and 4 for Chi-Square ($Q \parallel P$).

While the Hellinger/Bhattacharyya consensus operator outperformed all others in these experiments, it is not proof of outright dominance for any proposed question. There were plenty of instances where the operator was outperformed by the other consensus operators. But the ability to consistently be ranked near the top with little deviation leads one to believe that Hellinger/Bhattacharyya is a viable alternative to LinOP. In the least, there are doubts raised that LinOP is the only consensus opinion needed by a DM. These new alternatives should open the door to new theoretical research determining when and where each operator should be applied.
CHAPTER 5
DECISION MAKING UNDER TIME CONSTRAINTS

The conversation now turns to an area of the decision making process which is rarely discussed in the opinion pooling literature but has significant influence on the final outcome of any decision. Decisions are not solely determined by the available information to individuals, but are influenced by the surrounding environment in which they must be made. Specifically, the amount of time remaining before the outcome eventually occurs is an extremely important variable, however it is rarely discussed in the decision making process. For example, when an individual is asked to elicit a probability regarding whether a candidate will be elected for office, they would likely give different values under the same information depending on if the election was 1 year away or 1 day away. For the election happening the following day, it is likely that there is simply not enough time for new evidence to change the final outcome, and hence the approaching deadline polarizes the elicited probability to absolute certainty near a probability of 1 or 0 depending on whether the candidate is believed to be favored or not.

Figure 5.1 shows the general expectations of the probability of occurrence for a dichotomous event as time progresses and the deadline approaches. At the origination point there are numerous time steps remaining, and it is unlikely that the certainty of the outcome has been determined. However, as the deadline approaches, the likelihood of occurrence will be pushed into certain likelihood or unlikelihood given that the number of time steps remaining will only allow for so much variation to occur.

Under fixed information, elicited probabilities should change as the amount of time remaining to the deadline decreases. The key to determining how much the elicitation value should change is a function of the variance of the information flow. When large shocks to the system occur, greater room for variability is needed, and hence less polarization in the probability of the final likelihood as the deadline nears. In situations where new information is
Figure 5.1. General expectations of the probability on the outcome of a dichotomous event as a deadline approaches.
scarce, and the variance of the information flow is small, the polarization of the elicitation value as the deadline nears should occur more quickly. In cases where the elicitation variance follows a known distribution it is possible to calculate with certainty the final range of possible outcomes. When the variance is unknown, the historical performance of the likelihood can serve as a variability estimate, or the expert may elicit their own expectation of the variance.

The impetus for the development of the research in this chapter came from investigating the specific problem of whether the stock market would terminate at a positive or negative value for the day (Chapter 6 investigates this problem fully). As the end of the day approaches, the state of the market becomes more and more important given the lack of potential mobility within the final few minutes. For instance if the market is up 0.2% in the first hour of trading versus 0.2% with 2 minutes to go in the trading day, it is natural to assume significantly different probabilities of the market finishing positive for the day. The realization came that even with no state change in the value of the stock market, it was simply the time remaining which was influencing the likelihood of the final state of the market. Simultaneously modeling this balance between the state of the market with the approaching deadline information is the focus of this chapter.

Figure 5.2 illustrates this concept through two random walks. The noise of Series 1 is normally distributed with a mean of zero and standard deviation of 1, while the noise for Series 2 has a mean of zero and a standard deviation of 7. At the outset, with 300 time steps remaining, each series has an equal probability of terminating above the zero threshold upon reaching the deadline. However given the states of the two series at 50 time steps remaining, it is more likely that Series 1 will finish above the zero threshold compared to Series 2 not solely due to its current position, but more a result of the variance structure and time remaining before the deadline. The fact that Series 1 has a smaller variance implies that it is less likely to decrease in value as quickly as Series 2 may decline. Series 1 simply does not have the time remaining to explore the same space as Series 2. In fact given the states of the two series with 50 time steps remaining, Series 1 has a 99% chance of finishing in positive
territory, whereas Series 2 only has a 59% chance. The theory and calculations behind this example are shown in the following section.

**Figure 5.2. A demonstration of information flow on the probability of termination positions.**

This chapter introduces general elicitation techniques and decision making probability calculations under a time constraint deadline. The processes are analyzed across a number of different types of decision variables and probability distributions. The notation is first described followed by multiple sections investigating decisions classified as unbounded and bounded continuous variables. These sections are followed by similar investigations into categorical variables, which are bounded by nature in terms of being elicited on the probability scale between zero and one.
5.1 UNBOUNDED ELICITATIONS

The simplest type of opinion pooling with a deadline occurs when the variable of interest is an unbounded continuous variable. In this situation, there are no barriers to prevent the outcome from terminating neither above a maximum value nor below a minimum value, or any specific value for that matter. Under a variance structure that follows a closed form probability distribution it is quite simple to calculate the probability of the outcome terminating within a certain range given a fixed amount of time remaining. Similarly, it is also simple to calculate a confidence interval on the possible termination range under these same conditions.

Defining some notation, let $X$ be the elicitation variable of interest, $t$ be the amount of time remaining before the deadline, where $T$ is the total amount of time units remaining when the initial elicitation occurs. Therefore, $t = T, T-1, \ldots, 1, 0$, and $x_t$ is the state of the variable of interest with $t$ time units remaining. An important question for the decision maker should be what is the probability of $x$ finishing between a fixed low value ($x_L$) and high value ($x_H$) when $t = 0$, or $P(x_L < x_0 < x_H | t)$ or some variation therein. Alternatively, the decision maker could fix the percent of the distribution they wanted to cover similar to a confidence interval and determine the critical values of $x_L$ and $x_H$.

The expected value and variance of an upcoming step size (from $x_t$ to $x_{t-1}$) are needed to determine how much variability will occur from point $t$ to the termination time point when $t = 0$. Note it is not required that the expected value of the step size be equal to zero. For any single step size let $f_\Delta$ be the probability distribution with a finite mean $\mu$ and variance $\sigma^2$ describing the behavior of any randomly chosen step occurring between time units, or $\Delta = x_{t-1} - x_t$. The expected value and variance of the distribution $f_\Delta$ can either be modeled from collected historical steps or elicited directly from the expert in the case of no historical data. Letting $S_t$ be the sum of the $t$ i.i.d. random observations from the probability distribution $f_\Delta$, one can calculate the probability of ending between any $x_L$ and $x_H$ from simply utilizing the standard normal distribution and the central limit theorem (Theorem 5.1).
**Theorem 5.1** Central limit theorem: Let \( y_1, y_2, \ldots, y_n \) be identically independently distributed with a finite mean \( \mu \) and variance \( \sigma^2 \), then \( \frac{\sqrt{n}(\bar{y} - \mu)}{\sigma} \xrightarrow{D} N(0, 1) \).

A great benefit of the central limit theorem (Theorem 5.1) is that the expected value of the variance step size need not be zero, nor follow a normal distribution. Many times in practice step sizes follow a distribution having expected small steps with the possibility of very large shifts. A distribution with extremely long tails can be used in these situations to represent these cases accurately.

Theorem 5.2 puts the findings into context stating that for a fixed number of time observations remaining the distribution of the change in final outcome from the current state will be normally distributed with an expected value of \( t\mu \), and variance of \( t\sigma^2 \), or the number of steps multiplied by the distribution of an individual step size. The expected value of the final expected outcome is then \( x_t + t\mu \), with a variance around that estimate of \( t\sigma^2 \).

**Theorem 5.2** \( \frac{S_t - t\mu}{\sigma\sqrt{t}} \xrightarrow{D} N(0, 1) \), where \( S_t = \sum_{i=1}^{n} y_i \).

From this, it is easy to calculate the probability of \( x_0 \) terminating between two specific values \( x_L \) and \( x_H \) given the current state of \( x_t \):

\[
P(x_L \leq x_0 \leq x_H) = P \left( \frac{x_L - (x_t + t\mu)}{\sigma\sqrt{t}} \leq z \leq \frac{x_H - (x_t + t\mu)}{\sigma\sqrt{t}} \right)
\]

\[
= \Phi \left( \frac{x_H - (x_t + t\mu)}{\sigma\sqrt{t}} \right) - \Phi \left( \frac{x_L - (x_t + t\mu)}{\sigma\sqrt{t}} \right) \quad (5.1)
\]

where \( \Phi \) is the cumulative distribution of the standard normal distribution.

Alternatively a confidence interval for \( x_0 \) can be constructed given a fixed level of confidence at 100\((1 - \alpha)\)%, where \( \alpha \) is the probability of the outcome terminating outside of the confidence interval. The 100\((1 - \alpha)\)% confidence interval for \( x_0 \) is

\[
(x_t + t\mu) \pm z_{\alpha/2}\sigma\sqrt{t} \quad (5.3)
\]
where $z_{\alpha/2}$ is the critical value from the standard normal distribution corresponding to a tail probability of $\alpha/2$.

As an example of a continuous unbounded variable, imagine an individual is asked on January 1 to predict the closing value of the Dow Jones Industrial Average (DJIA) at the end of the year. While technically bounded at zero, interpreting the DJIA as a continuous variable is acceptable due to the extremely low likelihood of the market ever reaching this lower bound. Hypothetically, suppose we wish to view the problem with new elicitations from the individual occurring every single day from January 1 to December 31, upon which the final correct answer is immediately accurately revealed. Suppose on January 1 the initial point estimate for the elicitation was 12,150. On January 2 suppose there was good news announced about upcoming weather, and the individual raises their expected value to 12,400. This process continues daily until December 1 when the individual has an elicited expected value of 13,600. At this point there are 31 days left before the outcome is revealed, meaning $t = 31$.

If we are interested in calculating the probabilities of finishing within a certain range, the expected value and variance of a daily shift need to be defined via a probability distribution. The distribution can be of any form, not being required to be normally distributed, as long as the distribution has a fixed mean and variance. With numerous data points occurring daily from January 1 to December 1 it would be easy to use density fitting methods to determine the best fit for the distribution of daily step sizes $f_\Delta$. Suppose this $f_\Delta$ is determined to be best fit by a normal distribution with $\mu = 10$ and $\sigma = 75$, and we are interested in calculating the probability that the terminating DJIA is between 13,000 and
14,000. Using equation (5.2),

\[
P(x_L \leq x_0 \leq x_H) = \Phi\left(\frac{x_H - (x_t + t\mu)}{\sigma \sqrt{t}}\right) - \Phi\left(\frac{x_L - (x_t + t\mu)}{\sigma \sqrt{t}}\right)
\]

\[
= \Phi\left(\frac{14,000 - (13,600 + 31(10))}{75 \sqrt{31}}\right) - \Phi\left(\frac{13,000 - (13,600 + 31(10))}{75 \sqrt{31}}\right)
\]

\[
= \Phi(0.2155) - \Phi(-2.1792)
\]

\[
= 0.5853 - 0.0147
\]

\[
= 0.5706
\]

meaning there is a 57.06% chance that the terminating value of the DJIA will end between 13,000 and 14,000 given the daily change step structure of \(f_{\Delta}\).

Calculating the 95% confidence interval for the terminating value given the same step structure for \(f_{\Delta}\) is done via equation (5.3), resulting in

\[(x_t + t\mu) \pm z_{\alpha/2}\sigma \sqrt{t}\]

\[(13,600 + 31(10)) \pm 1.96(75) \sqrt{31}\]

\[13,910 \pm 818.46\]

\[(13,091.54 ; 14,728.46).\]

Key to this type of analysis is an accurate fit or elicitation of the step size distribution \(f_{\Delta}\). In this example there were over 330 historical data points which results in a distribution being able to be fit quite easily and accurately. Without numerous previous data points it is recommended that the expert directly elicit the parameters of the distribution \(f_{\Delta}\). If this is not possible, then standard elicitation techniques where the individual elicits their own termination boundaries are still available and recommended. However in many instances changes in the smaller time ranges can be easier to determine rather than the sum of all the steps simultaneously, as seen in the previous stock market example.
5.2 BOUNDED ELICITATIONS

In many instances the topic of interest for an elicitation is bounded in nature. For example, all fractional rate variables are bounded between 0% and 100%, such as the percentage of Americans in favor of a specific candidate, the unemployment rate, and likelihood to vote in the next election. An alternative example of a bounded variable outside of the 0-1 range includes the retail price of an airplane ticket whose company will never sell a seat below a certain low price, nor larger than a higher price level, but charges a variety of prices over time between the two bounds. The variables described in this section may terminate at either of the high or low bounds or any value in between and are quantitative in nature. The following section will investigate cases when the outcome is strictly limited to terminating at either the high or low bound and are dichotomous in nature.

When a quantitative variable of interest is bounded, it is essential that 100% of an elicited probability distribution is included between the upper and lower bounds. Since it is impossible for the outcome to terminate outside of the bounds, there should be zero probability allocated outside of the realm of possibilities. The determination of what is or is not impossible should be absolute and not open to subjective interpretation. Often impossibility is confused with extreme unlikelihood, and bounded variables discussed in this section are hard-line with absolute zero probability of any value occurring outside of the range. Due to the introduction of this range restriction, the techniques described in the previous section cannot be applied since the sum of the individual observations from an unbounded probability distribution are by nature unbounded.

Two approaches are now discussed which can be taken to ensure that the elicitation from an expert stays with the range of the upper and lower bounds as it varies over time. The first method follows the work of Nicolau [41] where as the elicitation of an individual approaches either of the bounds, there is a reversion effect that “pulls” values back within an acceptable range. The second approach discusses newly developed methodology inspired by
the former method which instead “pushes” an approaching series away from the predefined bounds rather than “pulling” it back towards the acceptable range.

5.2.1 Incorporating Bounds via Reversion

Nicolau’s [41] work on bounded series originates from random walk theory similar to the processes described in the previous section discussing unbounded elicitations. The technique works by defining a range of acceptable values where a random walk can act as it normally would under a given variance structure. However, if the random walk happens to escape the predefined range, then the proceeding steps will be adjusted accordingly to ensure the random walk returns to the acceptable bounded range. Defining the expectation of how the bounds are set up formally,

$$E[\Delta | X_t = x] = e^k(e^{-\alpha_1(x-\tau)} - e^{\alpha_2(x-\tau)})$$ (5.4)

where $\Delta$ is still $x_{t-1} - x_t$ as in the previous section, with new parameters $\alpha_1 \geq 0, \alpha_2 \geq 0, k < 0$ being introduced.

In discrete time situations the bounded random walk process then becomes,

$$X_{t-1} = X_t + e^k(e^{-\alpha_1(X_t-\tau)} - e^{\alpha_2(X_t-\tau)}) + \sigma_t \epsilon_t$$ (5.5)

where $\epsilon_t$ is a sequence of i.i.d normal random variables with $E[\epsilon_t] = 0$ and $Var[\epsilon_t] = 1$. Here the authors allow for the step size volatility measure $\sigma^2$ to follow a GARCH structure or other time dependent nature rather than have it fixed across all $t$.

Interpreting the other parameters, the constant $\tau$ in equation (5.5) represents the centered value or center of the allowable ranged distribution. When the bounded random walk is in the “neighborhood” of $\tau$, the state of the sequence is not near either of the bounds and will behave similarly to how any unbounded random walk would act. However, as the random walk deviates from $\tau$ the elements of equation (5.5) pull the random walk back towards $\tau$. 

Under a fixed $\tau$ and variance structure $\sigma_t \epsilon_t$, focus is given to reversion factor

$$a(x) = e^k(e^{-\alpha_1(x-\tau)} - e^{\alpha_2(x-\tau)})$$

(5.6)

to explain how equation (5.5) is capable of pulling values that stray away from $\tau$ back towards an area around $\tau$. First, it is obvious that when $\alpha_1 + \alpha_2 = 0$, $a(x) = 0 \forall x$ implying $X_{t-1}$ will behave simply as a random walk. To get the reverting boundary effect to take hold, $k$, $\alpha_1$, and $\alpha_2$ should be set to ensure $a(x) \neq 0$. The function $a(x)$ should exhibit negative values when the upper bound is being approached, or positive values as the lower bound is being approached. How large or small these corrective reversions of $a(x)$ will be are determined by the parameters $\alpha_1$, $\alpha_2$, and $k$, however the initialization of these parameters can be difficult due to the lack of easy interpretation. Examining the parameters under a specific example shows how varying the parameter values can dramatically affect the reversion action in equation (5.6).

Figure 5.3. Examples of parameter influence from equation (5.6) when $\tau = 100$. 
Figure 5.3 shows how equation (5.6) can be approximately zero for a wide range of values around \( x \), but will significantly deviate once the bounds are approached which are a function of the parameters \( k \), \( \alpha_1 \), and \( \alpha_2 \). For example, when \( k = -15 \), \( \alpha_1 = 3 \), \( \alpha_2 = 3 \), \( a(x) \approx 0 \) over the interval \( I = (95, 105) \). However, if \( x \) were to approach either of the hypothetical bounds of these parameters of 95 or 105 then the random walk would be pulled back towards \( \tau \) or the area around \( \tau \) where \( a(x) \approx 0 \).

Figure 5.4 shows an example of both a random walk and a bounded random walk under the parameters of \( \tau = 100 \), \( k = -15 \), \( \alpha_1 = 3 \), \( \alpha_2 = 3 \), and \( \sigma_t = \sigma = .4 \) for 1,000 discrete step sizes. Between the bounds of 95 and 105 the random walk performs very similarly to the bounded random walk. However, upon reaching the upper bound at about step 150 the function (5.6) begins to move negatively away from zero ensuring the random walk stays within the bounds. The process happens reversely at around step 225 ensuring that \( a(x) > 0 \) and pulling the series back towards the allowable area where it then mirrors the variance structure of the unbounded random walk. Hence, equation (5.5) is capable of acting like a random walk without influence from \( a(x) \) for a wide range of values if desired, and is only pulled back towards \( \tau \) after reaching a certain threshold away from \( \tau \).

While it may appear that equation (5.6) is capable of keeping a random walk in between bounds indefinitely, the boundaries are unfortunately not absolute. Looking carefully at Figure 5.4, the bounded time series plot actually crossed the upper threshold of 105 on multiple occasions, albeit by a very small amount. This is due to the fact that the random noise generated from \( \sigma_t \epsilon_t \) can be a greater value than \( a(x) \) depending on how steep the tails of the curve are as in Figure 5.3. A large random noise component in the direction of crossing the bound, can lead to an even more drastic overcorrection in the following step (Figure 5.8). In order to absolutely prevent any venturing outside the bounds, the algorithm provided by Nicolau must be altered through additional rules; corrections which will be implemented during the discussion of the newly developed methods.
A second difficulty of Nicolau’s method is in the choosing of the initial parameter values \(k, \alpha_1,\) and \(\alpha_2\). It is very difficult for a new user to see how the choices relate to the designated boundaries without constructing multiple visualizations of the function (similar to Figure 5.4) via guess and check methods. Rather than use the parameters \(k, \alpha_1,\) and \(\alpha_2\) to fit the boundaries, it would be simpler to directly use the boundary values in the function rather than fit the parameters to those boundaries. Due to these shortcomings, a new method of bounded series is approached from a slightly different angle in the following subsection.

5.2.2 Incorporating Bounds via Resistance

As an alternative to the “pulling” reversion effect in equation (5.6), a new bounding function is now proposed using a similar foundation. As the elicited value approaches the predefined bounds there is a reflective “pushing” which takes place, keeping the final elicited value strictly within the bounds. Rather than allowing for a random walk series to take place in a designated region, and revert back into that region as the series deviates, under this new method the series encounters exponentially greater resistance the closer \(x\) ventures towards
the bounds. The basis for this new function follows from equation (5.5) with an alternate adjustment element,

\[ X_{t-1} = X_t + \frac{\alpha_L}{X_t - L} + \frac{\alpha_H}{X_t - H} + \sigma_t \epsilon_t, \]  

(5.7)

where \( \alpha_L > 0, \alpha_H > 0 \) are scaling factors similar to \( \alpha_1 \) and \( \alpha_2 \) in \( a(x) \), and \( L \) and \( H \) are the fixed quantitative values for the lower and upper bounds. Investigating equation (5.7) at a more detailed level, the focus is turned to the corrective element in the equation

\[ b(x) = \frac{\alpha_L}{X_t - L} + \frac{\alpha_H}{X_t - H}. \]  

(5.8)

Returning to the same example previously discussed, it is easy to see both directly in equation (5.8) and in Figure 5.5 that as \( x \) approaches either of the bounds of \( L = 95 \) and \( H = 105 \) that the function \( b(x) \) begins to resist the random walk pushing it back towards the acceptable range where \( b(x) \approx 0 \) for \( x \), away from the bounds. Comparing the corrective actions of \( b(x) \) to \( a(x) \) across different values of \( x \) seen Figures 5.3 and Figure 5.5 shows that they can be constructed similarly in either manner, however the need for only one non-intuitive parameter makes the elicitation of the bounding function much easier in equation (5.8).

While the bounded function via the resistance technique in equation (5.7) serves as a simpler alternative by using the bounds directly as parameters, it still suffers from the same shortcoming discussed in the reversion method function (5.6) of having the possibility of the next step venturing outside of the bounds from two different causes. First, large corrections can occur in equation (5.7) when \( X_t \) is extremely close to either \( L \) or \( H \). These corrections can be so large that the series overshoots the bound on the other side. The second possibility for venturing outside of the bounds can occur when outlying random variance realizations from an unbounded distribution occur. To prevent the series from venturing outside of the bounds, two “security measures” are implemented to ensure that either an over-correction or large random noise component does not result in a step which will exceeded the limits of the bounds. The variability originating from either option is not of major importance, so for ease
in the algorithm the new function of interest becomes

\[ X_{t-1} = X_t + c(x), \tag{5.9} \]

where

\[ c(x) = \frac{\alpha_L}{X_t - L} + \frac{\alpha_H}{X_t - H} + \sigma_t \epsilon_t = b(x) + \sigma_t \epsilon_t, \tag{5.10} \]

the total combined change from \( b(x) \) and the random noise structure. However, an alteration to \( c(x) \) is made to ensure the value of \( X_{t-1} \) is not outside of the bounds. By Windsorizing (replacing extreme values at the tails of the distribution) the potential large changes in \( c(x) \), each potential step is limited to a predefined limit of the variance structure. Letting \( F \) be the distribution of \( \sigma \epsilon \) and \( \eta \) be a chosen quantile of \( F \), then any \( c(x) < F_\eta \) is set to \( c(x) = F_\eta \) or conversely any \( c(x) > F_{1-\eta} \) is respectively set to \( c(x) = F_{1-\eta} \). When \( \eta \) is set to a small value, such as 0.005, this limits the possibility of seeing a large correction whether from a random draw of the variance structure, or a possible large correction term \( b(x) \). This in itself significantly improves the odds of remaining within the bounds, however it does not
completely eliminate the probability. To ensure this, any step wishing to venture outside of the bounds is simply not allowed, forcing it to remain within the bounds. In other words if $X_t + c(x) > H$ or $X_t + c(x) < L$ then $X_{t-1}$ is not set to $X_t + c(x)$, but rather $H - \alpha_H$ or $L + \alpha_L$ respectively. Essentially when the walk tries to venture outside the bounds, $c(x)$ is set to a value just within the bounds preventing the outlawed occurrence from becoming a reality.

The pseudocode of the algorithm is stated for clarity:

```
SET parameters
CALCULATE c(x)
    IF c(x) > High Windsorized threshold
        THEN c(x) = High Windsorized threshold
    IF c(x) < Low Windsorized threshold
        THEN c(x) = Low Windsorized threshold
    IF $X_{t} + c(x) >$ Upper Bound (H)
        THEN SET $X_{t-1} = H - \alpha_H$
    IF $X_{t} + c(x) <$ Lower Bound (L)
        THEN SET $X_{t-1} = L + \alpha_L$
    ELSE SET $X_{t-1} = X_t + c(x)$
```

Under most situations, using techniques either from equation (5.5) or equation (5.7) will yield similarly behaving random walks. Figure 5.6 shows an example where an unbounded random walk deviates quite significantly beyond the bounds, while both equations (5.5) and (5.7) behave quite similarly within the bounds. However, the two algorithms will not always yield similar results due to the lack of the range restrictions in equation (5.5). As an example for the need of range restrictions, Figure 5.7 shows a case when equation (5.5) experienced a large overcorrection. Because of the restrictions imposed on $c(x)$, it is impossible for the new method to suffer from this same type of behavior.

In an even more drastic example, under the function (5.5) a random walk can completely diverge to values of positive or negative infinity. Once the series has escaped the
Figure 5.6. An example showing similar bounded random walks calculated from both methods.
Figure 5.7. An example of a large overcorrection after the random walk escapes the bounds.
bounded area the correction term can be over-correcting in a manner so much so that over-corrections become larger and larger overshooting the bounded area each time. The process repeats until reaching extremely large values, usually only requiring a handful of time steps. As an example, Figure 5.8 shows this divergence during steps 910 to 914 where the series moved from 95.6, 93.9, 115.1, -12 trillion, and finally positive infinity. The lack of necessary preventions limits the capabilities to provide useful and accurate elicitations in opinion pooling.

Figure 5.8. A diverging random walk from equation 5.5.

Due to the lack of applicable distributional theory under both bounding algorithms, an alternative is needed to calculate a probability of the series terminating within a certain range, or constructing some type of confidence interval. By simulating numerous series and investigating the behavior it is possible to calculate the likelihood of an event while incorporating the needed range restrictions. Continuing with the same parameters from the
previous examples, 10,000 random walks of 1,000 time steps each were simulated, and while only 0.45% of the 10,000 simulated random walks terminated outside of the bounds using (5.5), 78% of the random walks under function (5.5) at some point ventured beyond the desired pre-defined bounds of 95 and 105. This shows that once a random walk jumps outside a bound it is quickly corrected to within the acceptable range, and thus the probability of terminating at a value beyond the bounds is small. However, recall that these corrections can often be a result of a major shock as seen in Figure 5.7, resulting in a new random walk that does not follow the same sort of behavior as equation (5.7).

Histograms of the terminating values from a standard random walk and the two bounded algorithms are shown in Figure 5.9. Note that the distribution of the random walk is approximately normal as shown earlier in Theorem 5.2. The bounded random walk terminating values are obviously not approximately normally distributed. The terminating values from equation (5.5) are much more platykurtic than a normal distribution with an almost uniform likelihood in the center of the distribution. This can be a result of the fact that the larger corrections force the distribution to be more uniform since the correction size is so large it almost acts as a resetting of sorts. The bounded random walk from equation (5.7) is still platykurtic compared to a normal distribution, however it can be seen that outcomes near the starting value of 100 are slightly more likely than the tail observations.

As an alternative approach, the corrections added in equation (5.7), could also be applied to equation (5.5) to ensure the strict bounds are satisfied. However, with the initial parameterization using the actual bounds as values in equation (5.5), it is believed that users will find the initial choosing of the parameters much simpler, as well as easier to comprehend.

5.2.3 Bounded Example in Opinion Pooling Context

The newly developed bounded random walk theory is now put into the framework of general elicitations, allowing a user to determine a confidence interval for the final outcome not by determining the values directly, but by simply determining the expectation and variance of an individual step size. Since the variable of interest is never allowed to venture
outside of the predetermined fixed bounds it is implied that the starting point is already between the bounds. Once between the bounds, the algorithm using equation (5.7) will keep all values within the bounds with absolute certainty. A situation where the variable may have non-fixed time dependent bounds adds complexity to the model, and is discussed in the following subsection.

An example is now presented where the starting point is within the fixed constant bounds and must remain between those bounds over all $t$. Using another similar stock market example, suppose that the current state of the Dow Jones Industrial Average is 12,000 with an expected daily change of $\mu = 0$ and $\sigma = 75$. In an unbounded situation, calculating the probability of terminating in a certain range after 31 days could be done from the methods discussed in Section 5.1. However, suppose there was an announcement from the Federal Reserve stating they will ensure price stability for the following month not allowing the DJIA to dip below 11,400 or go above 13,000 over the next 31 days. However unlikely it may be for the Federal Reserve to set an upper limit of the stock market, the example shows how
equation (5.7) may be implemented in practice with non-equidistant bounds away from the initial starting value.

Simulating allows for the estimation of the probability of ending within certain terminating values, without the need for distributional theory. For example, suppose the problem at hand is calculating the probability of the DJIA terminating between $x_L = 11,500$ and $x_H = 12,500$ given the bounds of $L = 11,400$ and $H = 13,000$. The results from simulating 50,000 random walks using the function (5.7) where $\alpha_1 = \alpha_2 = 10$ are shown in Figure 5.10. To estimate the probability of terminating between 11,500 and 12,500, a simple sum of the values within that range divided by the number of simulations ($n=50,000$) will give the approximate result. In this case $40,941/50,000 = 81.88\%$. For comparison, in the unbounded case the probability of a random walk under these same specifications terminating between 11,500 and 12,500 from equation (5.2) is 76.88\%. By introducing the bounds and eliminating the possibility of tail events, the likelihood of finishing in the acceptable range is increased by 5%.

Figure 5.10. Histogram of 50,000 simulations for the bounded stock market example.
Since the parametric distribution of terminating points is not calculable in the bounded situation, likelihood intervals also need to be estimated by simulation. The first option is similar to the construction of a confidence interval by utilizing the order statistics \( \theta_{[i]} \), where \( \theta_{[i]} \) is the \( i \)th largest simulated terminating point. Constructing an equal tailed \( 100(1 - \alpha)\% \) confidence interval is then calculated by simply taking the order statistics at \( (\theta_{[na/2]}, \theta_{[n(1-\alpha/2)]}) \). The second approach is to construct the Highest Posterior Density (HPD) interval by computing all possible credible intervals and selecting the interval with the shortest range [13]. In this example, the equal tailed 95% confidence interval is calculated to be (11,410.00, 12,802.77) while the 95% HPD interval is (11,400.11, 12,678.61). The HPD interval is 114.27 points smaller than the equal tailed interval, and shifted to the left towards smaller values due to the lack of symmetry in the simulated distribution. For comparison, in the unbounded case the 95% confidence interval calculated from equation (5.3) is (11,181.54, 12,818.46). The much smaller widths of the likelihood intervals are expected from the introduction of the bounds which increased the probability of terminating away from the tails of the distribution. Table 5.1 contains the different calculations for the example across the unbounded and bounded results.

**Table 5.1. Comparison of bounded and unbounded results where**

\( X_{31} = 12000, L = 11400, H = 13000, \mu = 0 \) and \( \sigma = 75 \).

<table>
<thead>
<tr>
<th></th>
<th>Unbounded</th>
<th>Bounded</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P(11,400 &lt; x_0 &lt; 13,000) )</td>
<td>91.63%</td>
<td>100%</td>
</tr>
<tr>
<td>( P(11,500 &lt; x_0 &lt; 12,500) )</td>
<td>76.88%</td>
<td>81.88%</td>
</tr>
<tr>
<td>Confidence Interval</td>
<td>(11,181.54, 12,818.46)</td>
<td>(11,410.00, 12,802.77)</td>
</tr>
<tr>
<td>HPD Interval</td>
<td>(11,181.54, 12,818.46)</td>
<td>(11,400.11, 12,678.61)</td>
</tr>
</tbody>
</table>

**5.2.4 Time Dependent Boundaries**

In the example of the previous subsection there were two key assumptions made in regards to the upper and lower bounds. First was the restriction ensuring the DJIA terminated between 11,400 and 13,000 after 31 days. The second restriction ensured that the DJIA was
not allowed to take a value outside those same bounds on any of the previous 30 days before the final termination point. Relaxing the second assumption and allowing the possibility of any non-termination step venturing outside of the final bounds is now discussed. For example, suppose the Federal Reserve may be willing to allow for the DJIA to venture slightly below the lower final termination bound to 11,375 when there are 7 days remaining. However at that point the Federal Reserve might feel the need to interject, buying US Treasuries and forcefully encouraging the DJIA to rise. The discussion now turns to using time dependent bounds as opposed to the fixed bounds for all $t$ discussed up to this point.

In the case of time dependent bounds the same general algorithm is updated by allowing the bounds $L$ and $H$ to become time dependent parameters $L_t$ and $H_t$. Defining each bound as a time dependent vector of possible non-constant values removes the restriction that the random walk must stay between the termination endpoints throughout the given number of time steps remaining. Determination of the functional form of the time dependent bounds is adaptable to the elicitation topic and is user controlled. In some situations a step function on the boundaries might fit the matter at hand, whereas in other situations a non-linear quadratic function may be required to achieve the desired bounding structure. Requiring a certain function be applied to all elicitation situations is not practical in the numerous fields to which opinion pooling techniques are applied. Allowing the user to determine time specific bounds is useful in practice as there are numerous different situations which require varying paces of convergence to the final bounds.

Therefore the new process of the random walk under time dependent bounds is

$$X_{t-1} = X_t + c(x)$$

(5.11)

where

$$c(x) = \frac{\alpha_L}{X_t - L_t} + \frac{\alpha_H}{X_t - H_t} + \sigma_t \epsilon_t.$$  

(5.12)
The algorithm which handles outlying observations and disallows steps beyond the bounds for time dependent boundaries is extremely similar to the previous method under fixed bounds described in Section 5.2.2, but is now defined incorporating the introduction of the time dependent bounds $L_t$ and $H_t$.

**SET parameters**

**CALCULATE** $c(x)$

IF $c(x) >$ High Windsorized threshold 
    THEN $c(x) =$ High Windsorized threshold
IF $c(x) <$ Low Windsorized threshold 
    THEN $c(x) =$ Low Windsorized threshold
IF $X_t + c(x) > H_{t-1}$ (Upper Bound at Next Step) 
    THEN SET $X_{t-1} = H_{t-1} - \alpha_H$
ELSEIF $X_t + c(x) < L_{t-1}$ (Lower Bound at Next Step) 
    THEN SET $X_{t-1} = L_{t-1} + \alpha_L$
ELSE SET $X_{t-1} =$ $X_t + c(x)$

The example involving the DJIA and the bounds introduced by the Federal Reserve discussed in the previous subsection is now reexamined with the inclusion of time dependent bounds. Recall that the DJIA, starting at 12,000 is required to terminate between 11,400 and 13,000 after 31 days due to a mandate by the Federal Reserve. The daily step size of the market follows a random walk procedure with the noise coming from a normal distribution with $\mu = 0$ and $\sigma = 75$, and the parameters of equation (5.12) set again at $\alpha_L = \alpha_H = 10$. The two time dependent boundary vectors $L_t$ and $H_t$ are constructed so the lower bound $L_t$ follows a step function increasing in value as $t$ decreases, while the upper bound follows a linear decreasing function. The dashed lines in Figure 5.11 correspond to the time dependent bounds just described. Notice also in Figure 5.11 that as the random walk moves from step 20 to step 21, the lower bound was raised beyond the value of where the non-bounded random walk would normally venture, however increased in value to the new limit $L_{21} + \alpha_L$. 
Figure 5.11. Example of a random walk with time dependent bounds.

Similar simulation methods used in previous examples can be applied to this problem to estimate probabilities of terminating above or below specific values. A histogram of terminating values from 50,000 simulations are shown in Figure 5.12, and at first glance appears quite similar in shape to the distribution of simulations seen with the fixed bounds in Figure 5.10. Comparing the probability estimates for both in Table 5.2 shows the time dependent bounds have only a slight effect on the probability of terminating between 11500 and 12500, decreasing the probability by 1.82% to 80.06% from 81.88%. Note however that these time dependent bounds are indeed capable of causing drastic changes in termination likelihoods. In this example however, the time dependent bounds are not strict immobilizers to the random walk. Imagine a case where extremely narrow bounds kept the DJIA between 11995 and 12005 until the final step where the bounds extended to 11400 and 13000. Allowing for only a single step with $\mu = 0$ and $\sigma = 75$, the probability of terminating between 11500 and 12500 would rise very close to 100%, a great bit higher from the current calculated likelihood of around 80%.
Figure 5.12. Histogram of 50,000 random walk simulations with time dependent bounds.

Table 5.2. Comparison of unbounded, fixed bounds, and time dependent bounds where $X_{31} = 12000$, $L = 11400$, $H = 13000$, $\mu = 0$ and $\sigma = 75$.

<table>
<thead>
<tr>
<th></th>
<th>Unbounded</th>
<th>Fixed Bounds</th>
<th>Time Dep. Bounds</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(11400 &lt; x_0 &lt; 13000)$</td>
<td>91.63%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>$P(11500 &lt; x_0 &lt; 12500)$</td>
<td>76.88%</td>
<td>81.88%</td>
<td>80.06%</td>
</tr>
<tr>
<td>Confidence Interval</td>
<td>(11181.5,12818.5)</td>
<td>(11410.0,12802.8)</td>
<td>(11410.0,12810.6)</td>
</tr>
<tr>
<td>HPD Interval</td>
<td>(11181.5,12818.5)</td>
<td>(11400.1,12678.6)</td>
<td>(11400.1,12682.5)</td>
</tr>
</tbody>
</table>
Using simulation techniques allows for a user to estimate the probability of the final state of the series falling within a certain range for any quantitative variable. A probability estimate can be attained at any given time point simply from the elicitation of the starting value and the elicitation of the distribution of a single time step. Additionally, probability estimates can be calculated when dichotomizing a quantitative variable into two regions above and below a certain fixed value. Similarly, these probabilities can also be calculated when categorizing a variable into multiple ordinal bins (similar to a histogram) by simply taking the relative frequency of the number of simulations resulting in each bin. However, these methods cannot be applied directly to the elicitation of a categorical variable with no ordinal structure. The following section builds on the same type of analysis for the elicitation and behavior of a strict categorical outcome under time constraints.

5.3 Categorical Elicitation Under Time Constraints

Categorical elicitation is often applied in situations when there is a lack of preexisting data. Some examples of categorical elicitations with time constraints include: tracking terrorist locations before a known attack date, predicting the relocation destination of a professional sports franchise when their current stadium lease contract expires, or which candidate will be selected for the Presidential nomination before the party convention. Whether the outcome is dichotomous or has more than two possibilities, the key modeling exercise is still to determine how an approaching deadline in conjunction with incoming information simultaneously affect the likelihood of the final outcome.

For example, when determining the likelihood of who the next winning political candidate will be for an upcoming election, the amount of time remaining before election day is an important determinant of which candidate will win, even if the information (or political landscape) is held constant across time. Under the same information, being two years from election day versus the same scenario just two hours from polls closing should yield very different probability values for which candidate will win. The possibility of new information
arriving two years before the election is almost certain, thus increasing the likelihood of any candidate becoming more or less favorable. However only two hours from the election ending, the possibility of new incoming information is much lower, with an almost negligible effect on the final outcome.

There is relatively little developed theory investigating problems of categorical prediction under time constraints in the literature. The only instance of directly applied knowledge found involving this type of problem is in the psychology literature under the label Decision Field Theory (DFT) ([45], [44], [9]). The current and latest methodology in Decision Field theory is reviewed first followed by new theory in the earlier discussed methods of Analysis of Competing Hypotheses (ACH). Through the use of time dependent information similar to that in DFT, the earlier discussed methods of ACH are expanded to handle time dependent information thus allowing a user to easily determine the likelihood of the final outcome under any given time structure.

5.3.1 Decision Field Theory

Decision Field Theory was originally developed as a means to model the psychology behind human decision making; attempting to explain why people’s decision making processes do not always follow consistent logic or patterns. The primary modeling structure behind DFT states that when an individual is contemplating a choice between different possibilities, their overall preferences are altered as they think about the different elements which make up the decision. As time progresses, the preference of a certain outcome may increase or decrease depending on how much attention is paid to different elements of the overall decision. Once a particular quantitative threshold preference value is crossed, the individual will end further contemplation and choose that option. This setup differs slightly from the structure faced in opinion pooling where a preference threshold is not always included, but often a forced deadline is imposed. While not the most prominent area of focus in DFT, there is research dedicated to the introduction of a deadline before any option passes the preference threshold. One of the more advanced techniques which builds upon the basic
utility modeled DFT theory is Multialternative Decision Field Theory (MDFT) [46]. MDFT is now detailed in full to serve as background for the development of the time dependent ACH methods discussed in the following subsection.

The main premise of DFT is that the algorithm is able to accurately assign quantitative values to a set of beliefs over time in a manner similar to the human mind. The process is done through quantitative values termed “valences”. The valence for option \( i \) at time \( t \) denoted \( v_i(t) \) represents the momentary advantage of option \( i \) when compared to other options when considering a specific attribute \( j \). The ordered combination of valences across all \( i \) options forms a valence vector \( V(t) \). For example, suppose an individual has a choice between three alternatives when choosing to buy a new car: an Audi (A), a Buick (B), or a Dodge (D). This produces the three-dimensional valence vector \( V(t) = [v_A(t), v_B(t), v_D(t)]' \).

Suppose at a random time point \( t \), \( V(t) = [0.9, 0.45, 0.3] \). This then indicates that the individual prefers option A to B and option B to D since the valence values are larger when comparing directly across choices.

Valences are made up of three subcomponents: a comparison process (C), valuations (M), and attention weights (W). The comparison process and valuations are generally taken to be constant (although not required), while the attention weights differ across time steps.

\[
V(t) = CMW(t) \tag{5.13}
\]

The valuations (M) serve as personal evaluations of each option on each attribute. An attribute is simply a type of evidence one may want to consider when making the decision. Returning to the car example, suppose the consumer is only interested in evaluating the cars on two attributes, fuel economy (E) and paint quality (Q). Let \( m_{ij} \) represent the subjective valuation of option \( i \) on attribute \( j \), where higher values imply higher valuation. The two vectors \( M_E = [m_{AE}, m_{BE}, m_{DE}]' \) and \( M_Q = [m_{AQ}, m_{BQ}, m_{DQ}]' \) are then concatenated to make the matrix
\( M = [M_E \| M_Q] \). For example, suppose a user determines the attribute valuations to be

\[
M = [M_E \| M_Q] = \begin{bmatrix}
1 & 3 \\
3 & 1 \\
0.85 & 3.2
\end{bmatrix}
\]

where the first row of \( M \) corresponds to the Audi (A), the second row to the Buick (B), and third row to the Dodge (D). This implies that the individual with these valuations prefers the fuel economy in the Buick followed by the Audi, and lastly the Dodge. However, in terms of paint quality the individual prefers the Dodge slightly over the Audi, and more above the Buick. The relative values of the attributes are comparable only within a single attribute as a regularization will be done later.

The right component of a valence is \( W_j(t) \), which is defined as the momentary attention applied to attribute \( j \). In the car example with two attributes, an individual can focus on fuel economy (E), paint quality (Q), or a mixture of the two. The relative proportion of attention applied to either E or Q will vary over time as the individual considers each of the different attributes. The attention weights form the vector \( W(t) = [W_E(t), W_Q(t)] \). Often in practice these weights are dichotomous where individuals focus 100% of their attention on a single attribute during one moment in time, however it is not necessary. Under the dichotomous attention philosophy \( W(t) \) will alternate between \([0,1]\) and \([1,0]\) depending on if focus is solely dedicated to the fuel economy attribute (E) or paint quality attribute (Q).

The product \( MW(t) \) determines the weighted value of each alternative at each time point. The \( i \)th row of \( MW(t) \) equals the weighted value of the \( i \)th option:

\[ W_E(t)m_{iE} + W_Q(t)m_{iQ}. \]

Obviously under absolute attention weights of \([1,0]\) or \([0,1]\), \( MW(t) \) reduces to simply \( m_{iE} \) or \( m_{iQ} \) respectively. The difference between MDFT and the classic weighted utility model [34] is the stochastic nature of the attention weights \( W_E(t) \) and \( W_Q(t) \). For dichotomous outcome situations, attention weights can be drawn from either a uniform distribution for non-absolute weights or a Bernoulli distribution if it is desired to limit the
weights to absolute attention. If there are more than two outcomes, attention weights can be
drawn from a Dirichlet distribution for non-absolute weights or a multinomial distribution if
absolute attention is preferred.

The left component $C$, serves as a comparison procedure needed to determine the
relative advantage or disadvantage of each option on the weighted attributes being considered
at that moment. Generally the valence for each option is produced by contrasting the weighted
value of one alternative against the average of all the others. In the car example, the valence
for option A, would be

$$v_A(t) = W_E(t)m_{AE} + W_E(t)m_{AQ} - 0.5[W_E(t)m_{BE} + W_E(t)m_{BQ} + W_E(t)m_{DE} + W_E(t)m_{DQ}]$$

Again under absolute weights of $[1,0]$ or $[0,1]$ the valence for option A reduces to

$$v_A(t) = m_{AE} - 0.5[m_{BE} - m_{DE}] \text{ or } v_A(t) = m_{AQ} - 0.5[m_{BQ} - m_{DQ}]$$

respectively. The comparison matrix $C$ can be designed by defining a square matrix of size $i \times i$ with diagonal
elements set to 1, and off diagonal elements set to $\frac{-1}{(i-1)}$.

Interpreting another way, the product of $CM$ serves as the values (normally fixed) to
which the stochastic attention weights are applied. Continuing with the calculation from a
different perspective for the previous example,

$$[CM] = \begin{bmatrix} 1 & -0.5 & -0.5 \\ -0.5 & 1 & -0.5 \\ -0.5 & -0.5 & 1 \end{bmatrix} \times \begin{bmatrix} 1 & 3 \\ 3 & 1 \\ .85 & 3.2 \end{bmatrix} = \begin{bmatrix} -0.925 & 0.900 \\ 2.075 & -2.100 \\ -1.150 & 1.200 \end{bmatrix}$$

Note that the columns of $CM$ are normalized within each attribute implying the column sums
of $CM$ should be equal to zero.

When the attention weights are multiplied against $CM$, it forms the valence vector
$V(t)$. This valence is then added to current preference valuation state to determine the overall
preference state at the next time point. The current preference state for each alternative $P(t)$
represents the nominal value or utility attributable to each possible option. The option with
the highest value $P_i(t)$ at any given time point is the most likely option. If a termination point
is artificially imposed at any given time point, the option with the largest value in $P_i(t)$ is chosen as the preferable decision.

The final parameter of MDFT is the degree to which the system remembers the previous state when moving from one time point to the next. Determining whether there is any positive or negative feedback from the previous state is modeled through a symmetric remembrance matrix $S$. Defined formally, the preference state for the following time point is

$$P(t + 1) = SP(t) + V(t + 1).$$  \hspace{1cm} (5.14)

The remembrance matrix $S$ determines the memory state of the previous preference state $P$, before incorporating the new valence state $V(t + 1)$. When the diagonal elements of $S$ are set to 1 there is a perfect memory remembrance, whereas if $S = 0$ then there is no memory state, and intermediate values produce intermediate strengths of memory. Usually in practice the diagonal elements are set very close to one and the off diagonal elements set to small negative values near zero to produce slight negative influences where strong alternatives suppress weaker alternatives.

Figure 5.13 shows the preference states over time for the three different car options under the same parameters with $S$ set to the diagonal matrix with values of 1 on the diagonal and zeros off diagonal. In this case at time $t = 400$ the preference state was for the Buick, however at time $t = 1000$ the preference state is for the Dodge. The framework allows for the possibilities of changes in individual preference choice under fixed element preferences by applying stochastic weights to the different attributes. Having the remembrance matrix $S$ change from the identity matrix would increase the noise seen in each preference state. Even slight changes in the remembrance matrix $S$ can cause significant changes to the final preference states.

In the simple case of two attributes and three outcomes it is easy to see the impact each parameter has in the MDFT model. However in reality there are often a large number of
attributes which can be related to the final outcome. For instance in the car example there are
definitely more than 2 attributes a potential consumer would evaluate before deciding to make
a large car purchase. Asking an individual to elicit all preferences across all possible attributes
however is not feasible. A remedy to this situation is to split the attributes into important
attributes $M_1$, and non-important attributes $M_2$. Those non-important attributes $M_2$ are then
dropped from the analysis, and then modeled instead by random error term. Under this
scenario equation (5.13) becomes

$$V(t) = CMW(t) = CM_1W(t) + \epsilon(t)$$  \hspace{1cm} (5.15)$$

where $\epsilon(t) = CM_1W(t)$ is treated as stochastic error or residual term.

The methodology of MDFT is actually quite similar to the methods discussed in
Section 2.5 - Analysis of Competing Hypotheses. The act of attributing limited attention to
different attributes in MDFT is similar to paying attention to different pieces of evidence

Figure 5.13. Car preference example with $S$ as a 3x3 identity matrix.
under the ACH methodology. Specifically, the valence matrix $V$ is set up in a similar way to that of the main evidence matrix $X$ in the ACH algorithm. Adding the ability for ACH methodology to handle time dependent evidence as well as incorporating attention weights through a remembrance matrix is now discussed in the following subsection.

5.3.2 Time Dependent Analysis of Competing Hypotheses

Previously discussed ACH theory is now updated to incorporate elicitation and probability calculations under time constraints. Recall from Section 2.5 that the multinomial-Dirichlet model can be applied to help the elicitation and evaluation of alternative hypotheses $H_j$ across different evidence items $E_i$. Recall that the posterior distribution under numeric valued evidence items is equivalent to a Dirichlet distribution with parameters $\alpha$ given by $\alpha_j = \sum_{i=1}^{M} x_{ij}$, subject to $\sum_{i=1}^{M} w_i = n_{ess}$ where $w_i$ are the individual evidence weights, $n_{ess}$ is the overall allotment of evidence, and $x_{ij}$ is the evidence allotment to hypothesis $j$ on evidence item $i$.

Under the standard algorithm, the initialization of the ACH evidence matrix is similar to the elicitation of the valuations ($M$) completed in MDFT. Each evidence item in ACH receives its own independent subjective quantitative evaluation just as the valuations $M$ receive in MDFT. Additionally, each evidence item in the ACH model receives a weight ($w_i$) according to how much attention should be paid to each evidence item, similar to the weights $W(t)$ in MDFT.

Three major additions to the original ACH theory are presented in this subsection. The first addition is the ability for evidence weights to become time dependent, the second is the new ability for elicitation values within a single hypothesis to become time dependent, and the third is to incorporate a stochastic nature of the attention weights similar to MDFT.

Beginning with the first addition, the need for some of the evidence weights to be time dependent across all hypotheses is fairly obvious. Across different time points evidence weights can be increasing, decreasing, or both by following a non-monotonic function over
time. Changes in evidence weights over time will be applicable to all hypotheses so the function is simply describing how much evidence should be attributed to the specific evidence item relative to other evidence items. Time dependent weighting should not be used for the purposes of increasing or decreasing the likelihood of a specific hypothesis within a given evidence item.

For example, returning to the previous ACH example from Section 2.5 of predicting which location the San Diego Chargers will relocate to either Oceanside, Chula Vista or somewhere else, imagine a new evidence item that the Chargers stated three years ago that they preferred to relocate as soon as possible. As time moves forward, this statement does not hold as much weight as it did when it was first issued, and the weight attributed to this piece of evidence should be decreasing as time progresses. On the other hand, suppose a new evidence item where the Chargers said in 2008 that budgeted money allocated towards relocation costs will expire in 2013. In 2008 this evidence item should receive little weight, but as the expiration gets closer the threat of the loss of money becomes more important increasing the evidence item’s importance across all hypotheses. Upon expiration in 2013, this evidence item becomes moot and the weight should then be drastically lowered, perhaps all the way to zero. Both of these examples show, that independent of any specific hypothesis, evidences items need to be able to receive different levels of attention, or weights, as time moves forward.

In addition to the need for time dependent evidence, the addition of time dependent valuations within a single piece of evidence is also needed. For example in the Chargers scenario, suppose an evidence item states that Chula Vista has said it will double its financial incentives to the Chargers every six months until a decision is made on where they are to relocate. As time increases the evidence item itself is static, however Chula Vista should receive a higher value in the ACH valuation matrix as time increases. Instead of a fixed numeric value a function of $t$ can be placed in the ACH valuation matrix ensuring that as time increases the Chula Vista hypothesis will receive more weight. However, as time increases the evidence item itself will remain constant in the evidence weight compared to other evidence
items. The Chula Vista hypothesis will simply receive more and more of the share of this constant weight as time moves forward.

The third and final addition to the ACH theory incorporates the stochastic attention weighting structure following similarly to MDFT. As described earlier in the section on MDFT, when a decision is being contemplated over time, individuals pay different levels of attention to different pieces of evidence. To incorporate the stochastic weighting structure into the ACH framework, the decision making process is split into incremental steps from the first step $t_T$ until the deadline $t_0$ arrives. At any given time step $t$, attention is paid to a single evidence item based on a weighted selection process. The value of $n_{\text{ess}}$ remains fixed but is broken up over time steps where each time step contributes an information gain in $\alpha$ of $n_{\text{ess}}/T$ where $T$ is the total amount of time steps. As time steps continue, additional evidence is gained until the final deadline is reached when it is established that $\sum \alpha = \sum w = n_{\text{ess}}$. For example if $n_{\text{ess}} = 18$ and there are 16 time points remaining until arriving at the deadline of $t = 0$ then at each step $\alpha$ will increase by $18/16 = 1.125$. So upon the completion of 16 distinct time steps, $\sum \alpha = n_{\text{ess}} = 18$. Due to the stochastic nature of the evidence weights, repeatedly simulating outcomes allows for the ability to estimate the final probability on hypothesis likelihood calculations.

The general algorithm for the new ACH algorithm under an imposed deadline proceeds as follows. The general construction of the ACH valuation matrix proceeds exactly as describe in Section 2.5, but with the allowance for time dependent evidence weighting and time dependent ACH matrix values. After constructing the ACH matrix, define the vector of information gain at each step size $\nu$. If the information gain is equal at all time steps then set $\nu$ equal to $n_{\text{ess}}/T$ for all $t$. Next, repeating in a loop over the $T$ time steps, calculate the correct ACH valuation matrix at time $t$ ($X_t$), with correct time dependent evidence weights and matrix values. A random selection of a single evidence item by a single draw from a multinomial distribution with probabilities $w_t/n_{\text{ess}}$. Next, scale the randomly selected evidence row to ensure that $\sum \alpha_t = \nu_t$. The calculated row will be kept in a new matrix which
holds the values of alpha at each step size in the rows of this matrix. Taking the cumulative column sums across this matrix results in the final $\alpha$ parameters of the Dirichlet distribution.

A simple example with both fixed weights and fixed evidence items across time is now given under the new ACH algorithm. Suppose there are three evidence items and four hypotheses. The first evidence item suggests the latter two hypotheses are thought to be twice as likely as the first two hypotheses, while the second evidence item dictates all hypotheses are equally likely, and the final evidence item shows sequentially increasing preferences across hypotheses. For simplicity suppose the weights chosen are equal to the row sums of $X$ so that the values do not need to be scaled, resulting in $w(t) = w = [6, 4, 10]$ and $n_{ess} = \sum w = 20$ and

$$X = \begin{bmatrix} 1 & 1 & 2 & 2 \\ 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \end{bmatrix}.$$  

Under the standard ACH theory it is easy to calculate the final Dirichlet parameters $\alpha = [3, 4, 6, 7]$, and the resulting expectation probabilities of $[0.15, 0.20, 0.30, 0.35]$. This calculation is done however with no limitation on the number of time steps an individual may consider the evidence. Suppose if there were only a handful of time points before a deadline approached, then a decision maker might pay attention to certain evidence items more than others not allowing for the “correct” distribution to emerge.

In the same example, suppose a deadline is introduced such that only 5 iterations of different evidence items can be considered before the deadline is reached. In this scenario, there should be more variability in the final expectation probabilities as opposed to having unlimited time. At each time point only one evidence item will be considered, receiving an allotment of $\sum w / 5$, or $20/5 = 4$ for each time step. For the first time step a single evidence item is selected with probability $w / \sum w = (0.3, 0.2, 0.5)$, and for this example suppose the second row is chosen. Since the evidence item has equal values across all hypotheses the allotment is also distributed equally with the restriction that the sum equal the allotment of 4,
implying $\alpha_4 = (1, 1, 1, 1)$. Suppose for the next time step evidence item 3 is selected from the random weighted draw. The matrix values of (1, 2, 3, 4) are scaled to the evidence weight allotment of 4, hence (0.4, 0.8, 1.2, 1.6). To update the parameters, these new allotments are sequentially added to the previous state so $\alpha_3$ (one step closer to the deadline) is then (1.4, 1.8, 2.2, 2.6). This process then continues for the next three steps until the deadline of $\alpha_0$ is reached. When running this entire process over numerous iterations it is expected that final $\alpha_0$ parameter vector should equal that calculated under non-time constraints of $\alpha = (3, 4, 6, 7)$. However in this example with such a short amount of time before the deadline, this will not be the case across all simulations. Figure 5.14 shows the outcomes of the final $\alpha$ vector across 25,000 simulations with 5 time steps. Note the mean and median are equal to the expected parameters, however there is significant variance around those parameters. Note that the outlying observations when $\alpha = (5, 5, 5, 5)$ can only occur when evidence item 2 is the only evidence item considered across each of the 5 steps.

![Figure 5.14. Parameter results for $\alpha$ from 25,000 simulations after 5 time steps.](image)
Increasing the number of time steps from 5 to 200, and re-running the example under the same parameters, Figure 5.15 shows that the variance is much smaller with 200 time steps as opposed to those seen in Figure 5.14 with only 5 time steps. It is therefore important for a user to consider the number of time points carefully, as it can have a significant effect on the simulation variance.

**Figure 5.15. Parameter results for $\alpha$ from 1,000 simulations after 200 time steps.**

Examples showing how the extensions of time dependent evidence and time dependent values within a single evidence item are now presented under similar parameters to the previous example. First, an example of time dependent evidence weights $w(t)$ is worked through followed by an example with time dependent evidence allotments. Continuing with the previous example, suppose the weights of $w = [6, 4, 10]$ are altered to give evidence item three less weight in the beginning and more weight towards the end as the deadline nears. To provide clarity in the example, the number of time steps is set to be $T = 100$. Therefore at any time point $t$, the weight vector will be $w = [6, 4, (T - t)/10] = w = [6, 4, 10 - (t/10)]$. Note
that at each time point $t$, $w$ will be rescaled so $\sum w = n_{ess} = 20$, however this scaling is irrelevant to the probability each evidence item will be chosen. In the beginning time steps, when $t$ is large (100), the weight for evidence item 3 will be nonexistent, implying only evidence items 1 and 2 will be considered during the first time point. However, as the deadline approaches and $t$ becomes closer and closer to zero, the convergence to the previous weight vector to $w = [6, 4, 10]$ takes hold. Under this scenario it is expected that the probabilities of evidence items 1 and 2 move together in early time steps and similarly evidence items 3 and 4 will be expected to move together since they follow the same evidence weighting. However as time progresses and the weighting for evidence item 3 becomes larger, it is expected a division will occur in each of these since hypothesis 2 is preferable to hypothesis 1 and hypothesis 4 is preferable to hypothesis 3. These expected results do occur as predicted in a single simulation shown in Figure 5.16.

![Figure 5.16. Cumulative expected probabilities for each hypothesis from a single simulation with time dependent evidence weights.](image)

The effects of lowering the relative weight of evidence item 3 in the early time steps can be seen more clearly by simulating numerous outcomes under this time dependent
weighting scheme. Recall under constant evidence weights the parameters are expected to converge to \( \alpha = [3, 4, 6, 7] \). However Figure 5.17 shows that under this new time dependent evidence weighting the parameter values of hypothesis 1 and hypothesis 4 are drawn in closer towards the other parameters for the other two hypotheses with the mean parameter values resulting in \( \alpha = [3.39, 4.00, 6.00, 6.61] \). Notice that the expectations for evidence items 2 and 3 are exactly the same as under non-time dependent evidence. This is not a coincidence, and resides from the fact that whether incorporating evidence item 3 or not the summed parameters for hypotheses 2 and 3 scale on a 2:3 ratio. Therefore the weighting of evidence item 3 is irrelevant to the marginal parameter outcomes of hypotheses 2 and 3, but results in hypothesis 1 being slightly more preferable at the expense of hypothesis 4.

![Figure 5.17. Parameter results for \( \alpha \) from 2,000 simulations after 100 time steps with time dependent evidence weights.](image)

The second example turns to investigating the impacts from incorporating time dependent valuations within an evidence item while holding the weight allocations constant. Suppose the elicited ACH matrix is altered so that evidence item 3 for hypothesis 4 is
constructed to receive a lower value during the beginning time steps, and a more important higher weighting towards the deadline. One instance of this type of allocation could be constructed by replacing $X_{3,4} = 4$ with $X_{3,4}(t) = (T - t)$, forming the new matrix,

$$X(t) = \begin{bmatrix} 1 & 1 & 2 & 2 \\ 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & (T - t) \end{bmatrix}.$$

Returning to the constant weight vector of $w = [6, 4, 10]$, if in the early time steps evidence item 3 is randomly selected (probability of 0.5), hypothesis 4 will receive lower valued parameter allocations, whereas when the deadline begins to approach hypothesis 4 will be receiving almost all of the allocated weight for this evidence item (10). When simulating an outcome from this scenario it is expected to see evidence item 4 begin with lower expected probabilities than in the previous examples, but then significantly rise in likelihood as the deadline nears. A single simulation under these parameters is shown in Figure 5.18 verifying these expectations.

The distribution of the results from extending the simulation count to 1,000 are shown in Figure 5.19. Compared to the previous examples it is easily verified that upon reaching the deadline, hypothesis 4 is receiving more of the parameter allocation at the expense of the other hypotheses, with mean parameter values at $t_0$ of $\alpha = [2.29, 2.58, 3.87, 11.26]$.

Finally the two previously discussed time dependent weights and time dependent hypothesis valuations are incorporated simultaneously with Figure 5.20 showing the results from a single simulation. Both previously discussed conclusions of seeing hypotheses 1 and 2 move together and hypotheses 3 and 4 move together with the eventual separation of hypothesis 4 are easily visualized. Evidence item 3 is hardly considered in the beginning time steps, however as the deadline approaches, hypothesis 4 becomes significantly more likely both from the fact that evidence item 3 is receiving more weight, and within that evidence item
Figure 5.18. Cumulative expected probabilities for each hypothesis from a single simulation with time dependent ACH matrix evaluation values.

Figure 5.19. Parameter results for $\alpha$ from 1,000 simulations after 100 time steps with time dependent ACH matrix evaluation values.
hypothesis 4 is receiving more of the parameter allocation. Figure 5.21 shows the distribution of the parameter results from 1,000 simulations with a mean of $\alpha = [2.88, 2.99, 4.50, 9.63]$.

![Figure 5.21. Cumulative expected probabilities for each hypothesis from a single simulation with both time dependent weights and time dependent ACH matrix evaluation values.](image)

To calculate final expected probabilities of any hypothesis or the probability any hypothesis is the most likely to be chosen can be done by simulating from the numerous non-constant $\alpha$ parameters at $t = 0$. These simulations allow for the ability to calculate any specific probability statement through earlier discussed methods in the ACH elicitation section.

This new ACH algorithm allows users to replicate the varying levels of attention given to evidence items over time while simultaneously incorporating time dependent evidence and hypothesis weighting. The ability to utilize time dependent values in both evidence and hypotheses allows for more flexibility for users to incorporate additional evidence items which may not have been used in the previous methodology.
Figure 5.21. Parameter results for $\alpha$ from 1,000 simulations after 100 time steps with both time dependent weights and time dependent ACH matrix evaluation values.
CHAPTER 6
ELICITATION AND FORECASTING VIA
PREDICTION MARKETS

Prediction markets are a relatively new elicitation and probability aggregation platform where the open market attempts to calculate the true probability of an event before a deadline approaches [14]. The distinction between prediction markets and traditional opinion polling is exemplified through the differences in asking individuals who they are planning to vote for in the next presidential election versus who they think will actually win the election [56]. While both of these questions are used by political pollsters to forecast which candidate will be elected, the responses given need not be the same. Traditional opinion polling techniques are focused on grouping individuals’ absolute preferences, whereas prediction markets are focused on eliciting individuals’ beliefs about the probability each candidate will win.

A candidate’s true probability of victory in an election depends not only on polling numbers, but additional factors such as the historical volatility of those numbers, and just as importantly, the amount of time remaining before the election. To demonstrate how volatility is important to the behavior of a prediction market, imagine that in the two weeks leading up to the election a candidate has been polling at 55% consistently. Now compare this scenario to a candidate who has varied up and down in the polls from 25% to 75% and now back to 55%. The fact that there is less volatility in the first scenario should lead one to believe that the first candidate has a higher probability of being elected even though both candidates are polling at the same current level of 55%. Likewise, a candidate polling at 60% the night before the election has a much higher probability of being elected than a candidate polling at 60% six months before the election since there is little time for voters to change their opinions. Similar to the findings in the random walk theory developed in the previous chapter, both the deadline
and variability when approaching the deadline are important variables which influence the final probability of the outcome.

Polling numbers in conjunction with volatility and time remaining can be used to elicit the probability of a candidate being elected. However, transforming a polling sample proportion into a probability of being elected is not a straightforward task. Prediction markets are designed specifically to tackle this question at the outset. Individuals participate in the prediction market by exhibiting their beliefs about which candidate will win by trading shares priced on a probability scale. Prediction markets have no interest in who participants plan to vote for, only in what they believe the probability of the candidate winning the election to be.

The first major prediction market began in earnest in 1988 with the academically run Iowa Electronic Market (IEM) [30] where students invest real money in a variety of contracts for research purposes. Alternative for-profit prediction markets have recently gained in popularity for eliciting the opinion of a group. Intrade [29], one of the more popular online trading platforms based in Europe, provides individuals the opportunity to buy and sell contracts on the future likelihoods of numerous events ranging from the popular contracts on election winners, to pop culture phenomena such as movie box office returns, and even average global temperatures.

On the website, contracts are traded in manner similar to stocks on the probability of a future event with contracts dealing with binary outcomes. For example, a market might involve whether or not a candidate will win an election, whether or not an actor will win an award, or whether or not the Dow Jones will finish positive for the day. Contracts are traded on a scale of 0-100, 100 representing the belief of absolute certainty the event will happen, and 0 leaving no doubt the event will not occur. Eventually, every contract terminates at either 0 (event did not occur) or 100 (event did occur). The termination can either occur upon success or failure of the event, or the lack of a success occurring before a predetermined deadline. Because the range is on a 0-100 scale the numerical value is usually interpreted as
the percent likelihood that the event of interest will occur. Individuals have the capability to buy and sell multiple shares with one another at any price between 0 and 100.

In March 2010, one of the more popular prediction markets asked whether or not Obama will pass a health care bill by June 30, 2010. The last traded price as of March 8th 2010 was 48.0. This can be interpreted that Intrade traders believed there was a 48.0% chance that a healthcare bill would be passed by the termination date of June 30th 2010. Traders who believed this price was too high, i.e., those who thought the chance of passage was lower than 48%, could have sold shares at this value, whereas traders who believed the price was too low could have bought shares at this value. The closing price along with the volume traded for this contract over the beginning months of 2010 are shown in Figure 6.1. Of course Obama signed into law the passage of this health care package on March 23, 2010 and the contract expired at 100.

![Figure 6.1. The Intrade closing prices & volume for the prediction market on Obama health care reform.](image)

If the health care bill did not pass before the deadline, the contract price would have expired at 0, and those traders who bought contracts would have a negative return of $0.10 \times p$ per share, where $p$ is the executed contract price. Those who sold contracts would have a positive return of $0.10 \times p$ per share. On the other hand since the health care bill passed before the June 30th deadline, the contract price expired at 100, and those traders who bought contracts had a positive return of $0.10 \times (100 - p)$ per share. Those who sold contracts, had
a negative return of $0.10 \times (100 - p)$ per share. Note that it is not required that traders hold their positions to contract termination. They are free to buy and sell existing contracts at any price before the market is closed.

Current prices of active contracts are regularly broadcast by U.S. news organizations as accurate predictors of the outcome of the event of interest. CNBC has broadcast Intrade contract prices as predictors of future elections which in turn are expected to influence the stock market [17]. As the host of the television show *Fast Money* said in January 2008 when discussing the next day’s presidential caucus in Iowa, “The polls have the leading candidates in a statistical dead heat. But, ... we don’t follow the polls. We follow the money. And the money is telling a different story.” Intrade contracts did not have a dead heat, with Obama trading at a 52% chance of winning, Clinton at 39%, and Edwards at 25%. Certainly not a runaway, but this information was seen as highly valuable to future stock prices if Obama were to win. Eventually Obama did end up winning, validating the results from Intrade.

As this new type of forecasting gains traction in terms of mainstream popularity it is important to test whether prediction markets are indeed accurate. Often simple calibration studies are not able to fully capture the importance of previous volatility and time remaining in the contract. Comparing predictions from a newly designed algorithmic model for the probability of an event to prices from a specific Intrade prediction market can serve as a better reflection of the prediction market’s accuracy.

An initial difficulty when attempting to construct this model lies in the fact that a large number of repeated trials are necessary to quantify accuracy, and most markets on Intrade are highly specific, only generating a single outcome and leaving a sample size of one. Many previous analyses on the accuracy of prediction markets utilize U.S. election data [47], with most involving presidential elections. While an election is in itself a repeated event, elections across different years, different branches of government, or different geographical regions can be extremely different from one another. There have been suggestions from other analyses of prediction markets to analyze state or even local level prediction markets to better understand
how the markets work. These less glamorous prediction markets receive much less attention with many having very few executed trades. Elections that seem to have a clear favorite also generate little trading interest, causing a liquidity problem in the contract prices [57]. Looking to alternative prediction markets outside of the national political arena might provide better data than trying to group all elections under a single umbrella, or waiting decades for enough presidential elections to occur.

6.1 Dow Prediction Market Data

Finding a prediction market which has a high level of uncertainty, trader interest, and can be quickly repeated is key to the calibration analysis. The prediction market which trades on whether the Dow Jones Industrial Average (Dow) will end positive for the day is one market which satisfies these needs to a great extent. Daily trends of the Dow are more similar on a day-to-day basis compared to elections across time and geographic regions. The Dow prediction markets also maintain a high degree of independence across days and additionally, have the greatly added benefit that data can be quickly accumulated, with a new prediction market completing daily. Luckily traders on Intrade find this prediction market interesting, thus minimizing major liquidity troubles.

Analyzing the repeated trials of the same Dow prediction market allows for insight into whether the prediction market is capable of accurately predicting the final outcome. For a more rigorous comparison, predictions from a statistical model are compared to the actual Intrade contract prices with an aim of revealing any potential biases in the prediction market. The statistical model is built specifically for this Dow prediction market in which there is a high degree of independence and repeatability, hence this specific model would not generalize easily to election prediction markets. However, if there is a high level of accuracy when comparing the Dow prediction market to the statistical model, this does provide some evidence that prediction markets in general can be accurate. Going one step further, if the Dow prediction market is shown to be accurate does this lend credence to showing that other prediction markets are accurate as well? The Dow market is unique in that repeated trials can
be quickly accumulated and generalizing the results to an event that only occurs once may not be justifiable. However, if we can show repeated prediction markets show a high degree of accuracy that might lead one to believe the expected value of any single prediction market would be accurate as well, even on an entirely different topic. Conversely, if we see inaccurate results from a more robust model in this highly repeatable and independent market example, maybe that casts doubt on the accuracy we would see in other, less “clean” markets.

For the Dow Jones prediction market, Intrade makes every executed trade publicly available for two days after the close of the market. The data set was accumulated from 11/30/09 to 2/26/10, spanning over 55 full trading days. There were a total of 1,726 executable trades completed via Intrade, with 15,293 shares exchanged. Prediction market trading volume varies significantly, correlating positively with days when there is a higher level of uncertainty in the final outcome. For instance, days when the Dow rallies 200 points in the first ten minutes of trading will correspond to light Intrade volume simply because everyone agrees that the probability is very high the Dow will finish positive that day. On trading days when the Dow is close to finishing where it began, Intrade volume tends to be higher since there are more varying opinions on whether the Dow will finish positive or negative.

As anyone following the performance of their own 401K can tell you, the stock market has been far from stable over the past few years. During the data collection time period, the Dow began at 10,309 and ended at 10,326. While the overall Dow gain was minuscule, 59% of the days ended positive. Short term gains on a daily basis were wiped out by larger negative days. Data for the Dow Jones minute-by-minute values were downloaded from the Dow Jones Averages website [3]. Figure 6.2 shows the Dow for each trading day and the different degrees of variability of the repeated time series which will need to be accounted for when constructing the model.

6.2 A Simple Analysis: Calibration

A quick method to see if the Intrade traders are able to correctly predict the outcome of the Dow is to plot a calibration curve. For the prediction market to be calibrated correctly,
Figure 6.2. The 55 DJIA time series, where all data to the left of the vertical bar is removed from the windowing procedure.
the percentage of executed trades terminating in a positive Dow day should correlate strongly to the prices traded on Intrade. For example, when investigating all Intrade contracts traded at 25, 25% of those contracts should end with the Dow positive for the day. However, since there are few trades executed at each specific price, trades in similar regions are pooled together. Figure 6.3 shows the calibration curve for prices pooled on 5 point intervals plotted at the midpoints. For example, every trade executed between 0 and 5 is pooled and plotted at 2.5 on the horizontal axis. The eventual proportion of trades terminating in a positive Dow day is shown on the vertical axis. It is interesting to see that the calibration curve seems to deviate from the ideal for Intrade prices between 20-60. This shows that the Dow ended up finishing positive more than Intrade traders believed. This could simply result from the fact that on 59% of the traded days the Dow finished positive, and a data set spanning a longer time period might show a better calibration curve.

![Figure 6.3. Calibration curve for the DJIA prediction market where trades are pooled on 5 point intervals, and the diameter of the circle is proportional to sample size of trades executed.](image)
While Figure 6.3 might suggest that traders on Intrade are not able to correctly predict the outcome of the Dow, this calibration plot ignores two very important factors. The first question that needs to be investigated further is: at what value was the Dow when these trades were executed? The trades might have been executed in line with where one would expect the probabilities to be, and random stock market swings are to blame. Secondly, at what point in time were the trades executed? Trades executed earlier in the trading day when Intrade contract price volatility is lighter might have different calibration values than those traded closer to the end of the day when larger price swings occur. Accounting for each problem separately is a futile effort. Ignoring the amount of time remaining in the day while accounting for the current position of the Dow leads to the mistake of thinking volatility in the Dow is uniform across all time periods. Clearly the Dow’s final value should have a larger expected variance over an entire day than over a two minute period. Similarly, ignoring the current position of the Dow and only using the amount of time remaining is also a mistake. At a fixed time remaining in the day, the Dow at a daily change of 250 Dow points has a much higher probability of terminating positive than the Dow at a change of -250 Dow points. A model must be created which will transfer raw Dow market data into a probability of ending positive for the day using both types of information. These modeled probabilities can then be compared against the Intrade values to gain a deeper understanding of the accuracy of the prediction market as a whole.

### 6.3 A More Advanced Model

Two major difficulties arise when building a model to predict correct Intrade contract prices. First, as discussed earlier the contract price is highly influenced by the amount of time remaining in the trading day. As the amount of time remaining approaches zero, uncertainty approaches zero as well leading to an exponential increase in the volatility of contract prices as the deadline nears. The second problem is that the time series is disjoint, meaning that each day starts over at a value of 0% gain or loss. Therefore we cannot use general time series
methods to model the contract price. To solve the problem, the random walk theory developed in Chapter 5 is relied on heavily, serving as inspiration for the algorithm described below.

The data used for the modeling includes the 55 disjoint Dow time series, all on one minute increments, for days in which Intrade contract data was collected. There are officially 390 minutes in the trading day, however the first 5 minutes of Dow movement are removed due to the extreme volatility usually only seen during this time frame as a result of the market adjusting from pre-market and after hours trading. The probability distribution of the percent change in the Dow at any minute across all days is determined to be normally distributed (Figure 6.4) with a mean of zero and can therefore be normalized by simply dividing by the standard deviation.

![Histogram of Daily Percentage Changes](image)

**Figure 6.4. A histogram of all values of the Dow expressed as daily percentage changes on one minute intervals ($X_{it}$).**

The algorithm begins by investigating the intraday variance of the Dow. In order to create the distribution for the expected variance for any given Dow trading day at different time points remaining in the day, a bootstrapping approach is taken where randomly sliced Dow time series are selected for data. The range is then taken over this selected “window” as a measure of variability. Repeating this process will result in a distribution for the window
over the specific number of minutes. For example, suppose we want a distribution for how varying a Dow time series might be over the final 60 minutes. A random block of 60 minutes is chosen from the entire collection of days and amount of time remaining. The range is then calculated over the chosen window resulting in one draw of the distribution. The process is repeated 2,000 times until a distribution is created. A plot showing the distributions over different time lengths can be seen in Figure 6.5.

![Figure 6.5. The Window distributions from bootstrapping for 380, 215, 90, and 5 minutes remaining with red fits from a kernel density estimation using a reflection boundary.](image)

A hypothetical visual example for when there are only two minutes remaining is shown in Table 6.1. The same process is then repeated for every minute, resulting in 385 different window distributions.

For any given day and time remaining, the window distribution is integrated to the amount the Dow must travel in order to change the dichotomous outcome of the Dow ending positive or negative. A change in the belief of the termination of the Dow being positive or
negative occurs when the Dow itself moves from positive to negative territory (or vice versa). Any Dow series terminating positive will correspond to a Intrade termination price of 100, whereas if the Dow finishes negative the Intrade price will terminate at 0. While the analysis could be completed in the discrete case by simply looking at the proportion of values in the empirical distribution $W_t$, a continuous fit was placed onto each $W$ distribution to utilize the values of any outlying quick large changes in the Dow.

A fit on the window distribution can be done either through a beta distribution, or non-parametric kernel density estimation using a reflection boundary [49]. Interestingly, there was little variation in the final results when comparing the parametric distributional fit to the non-parametric approach.

The multiple window distributions can be integrated to calculate expected probabilities that the Dow will finish positive for the day, given its current state and how many minutes are left in the trading day. Let $x_{it}$ be the percent change in the Dow from the beginning of day $i$ with $t$ minutes remaining. The data is normalized and inverted by letting $\theta_{it} = \Phi(x_{it}/\sigma_x)$ where $\Phi$ is the CDF for the normal distribution. This 0-1 scale can now easily be compared against Intrade prices which are on a 0-100 scale by multiplying predictions by a factor of 100. Let $W_t$ be the constructed window distribution for $t$ minutes remaining.

For any given day and time remaining, the distribution $W_t$ is integrated over 0 to $|\theta_t - 0.5|$. The upper limit of the integral $|\theta_t - 0.5|$ is the variability the Dow must travel in

### Table 6.1. A hypothetical windowing procedure for $t = 2$, where the range is repeatedly calculated over randomly chosen blocks until the distribution $W_2$ is created.

<table>
<thead>
<tr>
<th>Day</th>
<th>Minutes Remaining</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>385</td>
</tr>
<tr>
<td>1</td>
<td>384</td>
</tr>
<tr>
<td>2</td>
<td>383</td>
</tr>
<tr>
<td>3</td>
<td>382</td>
</tr>
<tr>
<td>...</td>
<td>380</td>
</tr>
<tr>
<td>55</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Day</th>
<th>Minutes Remaining</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.20 .25 .30 .40 .30 .28 ... .4</td>
</tr>
<tr>
<td>2</td>
<td>.40 .41 .42 .44 .43 .42 ... .7</td>
</tr>
<tr>
<td>3</td>
<td>.70 .90 .95 .98 .99 .98 ... .98</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>55</td>
<td>.30 .40 .21 .10 .15 .16 ... .24</td>
</tr>
</tbody>
</table>
order to change the dichotomous outcome of the Dow ending positive or negative. A change in the belief of the termination of the Dow being positive or negative occurs when $\theta_t$ crosses the 0.5 value. Any Dow series terminating at $\theta_{i0} > 0.5$ will correspond to an Intrade termination price of 100, whereas if $\theta_{i0} < 0.5$ the Intrade price will terminate at 0.

The core of the prediction lies in finding the probability of $\theta$ crossing the 0.5 threshold, or a transfer from negative to positive (or vice versa) in the Dow. Letting

$$ F(w|\theta, t) = \frac{1}{2} \left[ 1 - \int_{0}^{\theta_t-0.5} w_t(\theta) \, d\theta \right] $$

gives a method for predicting the probability of the Dow terminating positive or negative. The inclusion of the $\frac{1}{2}$ is due to the assumption that the Dow time series has an equal probability of increasing or decreasing at any given point, so only half of the eventually realized ranges will be in the direction towards 0.5. The final contract price is then determined by whether or not $\theta$ is greater than 0.5. Letting $C_{it}$ be the predicted contract price,

$$ C_{it} = \begin{cases} 
100 \times (1 - F(w|\theta, t)) & \text{for } \theta_{it} > 0.5 \\
100 \times F(w|\theta, t) & \text{for } \theta_{it} < 0.5,
\end{cases} $$

allows us to compare $C_{it}$ directly to Intrade prices.

As an example suppose the Dow is perfectly even on the day, or $x_{it} = 0$, resulting in $\theta_t=0.5$, and thus $F(w|\theta, t)=0.5$. Hence, $C_{it} = 50$, which makes sense on a practical level since a market even on the day should have a 50/50 chance of ending positive or negative for that day.

Suppose now the Dow is extremely negative at $x_{it} = -2.7\%$. Since this value rarely occurs, $\theta_{it}$ will be very small and close to zero. Calculating $|\theta_t - 0.5|$ will have a value close to 0.5, and while depending on $t$, should lead to a very small $F(w|\theta, t)$. Since $\theta_t < 0.5$, $C_{it}$ will also have a very small value. The process also aligns with common sense: when the Dow is down a large amount, the probability that it might end positive that day is small. Similarly
when the Dow is very positive the model results in predicted contract prices that are very close to 100.

### 6.4 Dow Prediction Market Accuracy Evaluation

The goal of this research was not necessarily to determine whether prediction markets or this newly created model does a better job of actually predicting the correct outcome, but rather to show that if prediction markets are compared to a more accurate reflection of reality, they can be verified as an accurate opinion pooling mechanism. Comparing the model predictions against executed prediction market contract prices shows that there is a strong relationship between the two. Using the same calibration methodology as earlier, a new calibration curve is constructed for Intrade contract prices versus the model predictions. Instead of using the percentage of trades which had positive Dow days on the vertical axis, the model predictions are binned and used. Figure 6.6 shows that the model provides a much better basis for showing that prediction markets are indeed accurate. Notice that prediction market trades executed less than 50 correspond with slightly higher model predictions, and those trades executed over 50 correspond with slightly lower model predictions. If the model is assumed to be accurate, traders seem to be consistently overconfident throughout the range of Intrade prices, something that has been seen in previous analyses of prediction markets.

Fitted plots for a few selected days are presented in Figures 6.7 to 6.10. As seen in the figures, prediction markets are able to not only maintain a level of stability around a constant market, but can quickly capture drastic moves in the Dow. The quick analysis using a calibration plot is now deemed insufficient. A deeper understanding through more advanced statistical modeling provides evidence that prediction markets are to some extent accurate.

Finally, Figure 6.11 displays accuracy rates for both the Intrade contracts and for the predictions from the model across time. The accuracy rate measures the proportion of Intrade contracts and model predictions which were associated with the correct outcome. Both Intrade contracts and model predictions were deemed successful if they had a price/prediction below
Figure 6.6. Calibration curve for the DJIA prediction market using model predictions pooled at 5 point intervals.

Figure 6.7. The Dow, Intrade, and model fits for 11/30/09.
Figure 6.8. The Dow, Intrade, and model fits for 02/22/10.

Figure 6.9. The Dow, Intrade, and model fits for 12/07/09.
Figure 6.10. The Dow, Intrade, and model fits for 01/21/10.

50 and terminated at a price of 0, or additionally those priced/predicted above 50 terminating at a price of 100. Obviously, at the end of the day both Intrade contracts and the model predictions should be at 100%. Both methods begin the day with a high accuracy rate of about 70%. Note that the newly developed model is slightly more accurate than the prediction market from the hours of 10:00 A.M. to 1:00 P.M. by about 5%. However, from 1:00 P.M. to 3:00 P.M. the Intrade contract prices are slightly more accurate than the model. Eventually the rates become indistinguishable for the final hour while quickly rising to 100%. These high rates provide additional evidence that both the model and the Intrade prediction market were accurate in determining whether the Dow will finish positive or negative on the day.

While it appears that the prediction market is accurate to some degree, there is a hint of a difference between Intrade predictions and the model predictions. There is a slight showing of overconfidence in the Intrade market, or under-confidence in the model, but the reasons for this remain to be seen. Additionally, some additional techniques that could improve the statistical model include an added momentum time series approach through a Bayesian GARCH model, a weighted window selection technique, new standardizing alternatives, and different choices for variability rather than the range. The inclusion of the GARCH model could rectify any problems from the heavy assumption that at any given time
the Dow has an equal probability of gaining or losing value at any given time. Previous research in regards to stock market data has shown that GARCH models, which incorporate heteroskedastic variance, outperform those which assume a constant variance. The inclusion of a weighted windowing structure might lead to different results if window draws were weighted more heavily towards the beginning, middle, or end of the Dow trading days.
CHAPTER 7

DISCUSSION

In total, the research presented in this dissertation compiles a thorough undertaking of the statistical methodology involved in the opinion pooling arena. The newly constructed visualization tool for improved expert elicitations, theoretical developments for optimal pooling methods via alternative distance measures, and the research on the importance of the time dimension in decision making all push the boundaries of current opinion pooling methodology.

As data storage becomes cheaper and additional data aggregation methods become more prevalent, the field of computational statistics is beginning to diverge into situations that often deal with either too much data, or oppositely, too little data. In these instances where there is little to no data available to collect, the opinions from individuals experts will often serve as the only data source. Eliciting probability distributions, aggregating these multiple distributions, and forecasting the likelihood of the event using the new techniques discussed in this dissertation can significantly help experts and DMs by leading to more accurate and informed decisions.

To help with the elicitation process, new improved visualization software was designed and tested. While the results did not show significant improvements in the performance of the elicitations, the overwhelming individual preference for the more advanced graphical capabilities in conducting elicitations gives hope that better results will come in the future as a result of these and similar improvements. As more attention is given to statistical graphical user interfaces in other fields, the capabilities should increase in the field of opinion pooling as well by utilizing much of the same new technology.

Helping to improve the aggregation methods available to a DM, multiple newly developed consensus operators were put forth using distance measures never before applied to
the field of opinion pooling. These new tools will now allow for DMs to make better decisions when aggregating the opinions of multiple experts. Through experimentation results, many of these newly derived consensus operators were shown to outperform the current existing methods. The greatest potential is reserved for the consensus operator using the Hellinger and Bhattacharyya distances. Since this operator maintains the highest relative degree of uniformity it could be helping to counteract the overconfidence seen in individual expert elicitations.

The science behind decision making under a deadline has not been fully explored by the statistical community. While the field of psychology has had some brief exploration in the area, the science behind conducting proper elicitations and forecasting opinions under a deadline scenario has not been fully explored despite the relatively high frequency that these types of decisions are made. The research put forth in this document can serve as a starting point for numerous advancements in this area. Examples of applications of this new theory were demonstrated through the implementation of a popular prediction market. With the popularity of prediction markets continuing to grow, research in this area should continue to expand rapidly.

Opinions of experts are not limited to being used in situations when there is little data available. Using prior knowledge and updating it with data from a sample has lead to a revolution in statistics over the past century through the incorporation of Bayesian methodology. In fact, when analyzing different prediction methods, algorithms which use both standard statistical data analysis methods combined with human intuition have been seen to outperform each method by itself. While there is a clear benefit to incorporating outside human knowledge into a statistical model, academic research in this area is lacking. Further theory and evidence on how to best incorporate the opinions of experts in conjunction with statistical modeling to make more informed and accurate decisions is an area statisticians have yet to conquer.
BIBLIOGRAPHY


143


[44] T. PLESKAC AND J. BUSEMEYER, Two-stage dynamic signal detection: A theory of


