Applications of Non-Parametric Kernel Smoothing Estimators in Monte Carlo Risk Assessments

Allan Francis Trapp II
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/etd

Part of the Agricultural Science Commons, Agriculture Commons, Agronomy and Crop Sciences Commons, Public Health Education and Promotion Commons, and the Statistics and Probability Commons

Recommended Citation
Trapp II, Allan Francis, "Applications of Non-Parametric Kernel Smoothing Estimators in Monte Carlo Risk Assessments" (2012). Graduate Theses and Dissertations. 12776.
https://lib.dr.iastate.edu/etd/12776

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Applications of non-parametric kernel smoothing estimators in Monte Carlo risk assessments

by

Allan Francis Trapp II

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:
Philip Dixon, Major Professor
Alicia Carriquiry
Max Morris
Mark Widrlechner
Huaiqing Wu

Iowa State University
Ames, Iowa
2012
DEDICATION

I would like to dedicate this thesis to my mother, friends, and my girlfriend. Thanks mom for preparing many delicious and nutritious (tastes just like chicken) meals when I was going batty and did not have the time. Thank you dear friends who let me spike volleyballs all over (in-bounds and out-of-bounds) when researching became frustrating. Thank you Megan Batchelder a “wily curves” coyote. (For the record, I’m Road Runner, “Beep beep!”) Whenever I needed some spice with my extra servings of research, you were there. And to Dominque Trumbo, my PC (PC = Peace Corps) volunteering friend in Africa, your phone/message cheerleading and love saw me through to the end. You motivated, listened, and consoled me. Thank you!
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF TABLES</td>
<td>vi</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>vii</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>ix</td>
</tr>
<tr>
<td>CHAPTER 1. INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>CHAPTER 2. SCHEDULING VIABILITY TESTS FOR SEEDS IN LONG-TERM STORAGE</td>
<td>5</td>
</tr>
<tr>
<td>BASED ON A BAYESIAN MULTI-LEVEL MODEL</td>
<td></td>
</tr>
<tr>
<td>2.1 Introduction</td>
<td>6</td>
</tr>
<tr>
<td>2.2 Data Description</td>
<td>8</td>
</tr>
<tr>
<td>2.3 Modeling Seed Viability for Many Seed Lots</td>
<td>9</td>
</tr>
<tr>
<td>2.3.1 Quadratic Random Coefficients Model</td>
<td>10</td>
</tr>
<tr>
<td>2.3.2 Avrami Formulation of the Seed-Viability Model</td>
<td>14</td>
</tr>
<tr>
<td>2.4 Seed-Testing Predictions</td>
<td>14</td>
</tr>
<tr>
<td>2.5 Comparison of Avrami and Quadratic Models</td>
<td>15</td>
</tr>
<tr>
<td>2.6 Selection and Evaluation of a $t^Q_i$ Quantile</td>
<td>17</td>
</tr>
<tr>
<td>2.6.1 ROC curve development</td>
<td>17</td>
</tr>
<tr>
<td>2.6.2 Selection of an Optimal Quantile of $t^Q_i$ Distributions</td>
<td>20</td>
</tr>
<tr>
<td>2.6.3 Evaluation of $\alpha = 0.05$ Quantile Rule</td>
<td>21</td>
</tr>
<tr>
<td>2.7 Discussion</td>
<td>23</td>
</tr>
<tr>
<td>2.8 Acknowledgements</td>
<td>24</td>
</tr>
<tr>
<td>CHAPTER 3. MULTIVARIATE KERNEL REGRESSION-BASED RISK</td>
<td>25</td>
</tr>
<tr>
<td>ASSESSMENT THROUGH A MODIFIED 2-DIMENSIONAL MONTE</td>
<td></td>
</tr>
</tbody>
</table>
3.1 Introduction .................................................. 26
3.2 Risk of Hemolytic Uremic Syndrome in Children Exposed to *E. coli* O157:H7 . 31
  3.2.1 Aleatoric Variables in the Dose-Response Model of HUS ................. 32
  3.2.2 Epistemic Parameter Distributions of the HUS Model .................. 34
3.3 New Approaches to Model Epistemic Parameter Uncertainty .............. 36
  3.3.1 Asymptotic MLE Distribution .................................. 36
  3.3.2 Normal Theory Distributions ................................... 38
3.4 Non-Parametric Estimation of Risk CDF ................................ 40
  3.4.1 Nadaraya-Watson Estimator Applied to CDF Estimation ............... 40
  3.4.2 Choice of Kernel: Epanechnikov Versus Uniform .................... 45
  3.4.3 Comparison of Standard Two-Dimensional Monte Carlo CDFs and Uniform-
       Based Non-Parametric CDFs ...................................... 48
3.5 Discussion .................................................. 52

CHAPTER 4. COMPARISON OF BANDWIDTH SELECTORS FOR A
MULTIVARIATE NADARAYA-WATSON KERNEL REGRESSION ES-
TIMATOR OF A CUMULATIVE DISTRIBUTION FUNCTION .......... 54
4.1 Introduction ................................................ 55
4.2 Simulation Models ........................................... 58
4.3 Optimal Bandwidth Selectors .................................. 59
  4.3.1 Golden Bandwidths .......................................... 59
  4.3.2 Optimal Penalized Average Squared Error Bandwidths .............. 61
  4.3.3 Multivariate Kernel Density Plug-In Bandwidth .................... 64
4.4 Results ...................................................... 65
4.5 Discussion .................................................. 68

CHAPTER 5. CONCLUDING THOUGHTS ........................... 70
5.1 Summary of Conclusions ...................................... 70
5.2 Topics for Further Investigation .............................. 72

APPENDIX A. CALCULATION OF $\hat{\gamma}_i^{Q}$ ..................... 73
# 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>Test Criteria and Decisions for a 50% Critical Regrow Value</td>
<td>18</td>
</tr>
<tr>
<td>Table 2.2</td>
<td>Count Results from the 5-Stratum Follow-Up Study</td>
<td>22</td>
</tr>
<tr>
<td>Table 3.1</td>
<td>Probability Bounds for the 0.025, 0.25, 0.5, 0.75, and 0.975 Risk Quantiles of Figure 3.1</td>
<td>30</td>
</tr>
<tr>
<td>Table 3.2</td>
<td>Average Squared Error ($ASE$) of Pairwise Comparisons Between Traditional and Non-Parametric $CDF$ Curves</td>
<td>52</td>
</tr>
<tr>
<td>Table 4.1</td>
<td>Canonical Bandwidths for the Uniform and Normal Kernels</td>
<td>65</td>
</tr>
<tr>
<td>Table 4.2</td>
<td>Optimal Bandwidths for a Risk Model Containing Two Epistemic Parameters</td>
<td>66</td>
</tr>
<tr>
<td>Table 4.3</td>
<td>Optimal Bandwidths for a Risk Model Containing Four Epistemic Parameters</td>
<td>68</td>
</tr>
<tr>
<td>Table 4.4</td>
<td>Optimal Bandwidths for a Risk Model Containing Seven Epistemic Parameters</td>
<td>68</td>
</tr>
<tr>
<td>Table A.1</td>
<td>$\hat{t}_i^Q$ Computations</td>
<td>73</td>
</tr>
<tr>
<td>Figure 2.1</td>
<td>(a) Plot of 3 Prevalent Types of Seed-Lot Data Patterns; (b) Shrinkage Plot</td>
<td>10</td>
</tr>
<tr>
<td>Figure 2.2</td>
<td>Comparison of Avrami and Quadratic Curves for 1 Lot</td>
<td>16</td>
</tr>
<tr>
<td>Figure 2.3</td>
<td>(a) Non-Parametric Density Estimates of $\alpha$ with Respect to Negative and Positive Populations; (b) ROC Curve</td>
<td>20</td>
</tr>
<tr>
<td>Figure 3.1</td>
<td>Two-Dimensional Monte Carlo Simulation Results of Risk of HUS in Children Under the Age of 5</td>
<td>29</td>
</tr>
<tr>
<td>Figure 3.2</td>
<td>Directed Acyclic Graph of HUS Risk Model</td>
<td>34</td>
</tr>
<tr>
<td>Figure 3.3</td>
<td>Comparison of Non-Parametrically Bootstrapped and Asymptotic Distributions of Serving Parameters</td>
<td>38</td>
</tr>
<tr>
<td>Figure 3.4</td>
<td>Comparison of Non-Parametrically Bootstrapped and Normal-Theory Based Distributions of Initial Bacterial Parameters</td>
<td>39</td>
</tr>
<tr>
<td>Figure 3.5</td>
<td>CDF Estimators: (a) Joint Distribution of Standardized Serving Epistemic Parameters. (b) Histogram of Risk Values Corresponding to Epistemic Parameter Vectors Located in the Neighborhood. (c) Comparison of Residuals for Neighborhood vs. Traditional CDF Estimators.</td>
<td>43</td>
</tr>
<tr>
<td>Figure 3.6</td>
<td>Comparison of Epanechnikov and Uniform Kernels: (a) Integrated Mean Average Squared Error ($IMASE$) as a Function of Bandwidth. (b) Squared Bias Component of $IMASE$. (c) Variance Component of $IMASE$.</td>
<td>47</td>
</tr>
<tr>
<td>Figure 3.7</td>
<td>Interpolation Graphic</td>
<td>49</td>
</tr>
<tr>
<td>Figure 3.8</td>
<td>Comparison of New vs. Old Estimator</td>
<td>51</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>---------</td>
<td>------------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Figure 4.1</td>
<td>Plot of Trimmed and Retained Observations for a Model with Two Parameters</td>
<td>63</td>
</tr>
<tr>
<td>Figure 4.2</td>
<td>Accuracy of Risk Estimates Under Six Optimal Bandwidths</td>
<td>67</td>
</tr>
</tbody>
</table>
I would like to take this opportunity to express my thanks to those who helped me with various aspects of the research and writing of this dissertation. First and foremost, thank you Philip Dixon for your guidance, patience, and support. Your kind and sage words bolstered my confidence and inspired me to be a better statistician. I would also like to thank Mark P. Widrlechner. You have made me a better writer with your amazing word smithing and awesome editing skills. And to the rest of my committee, you have taught me so much. Thank you!
CHAPTER 1. INTRODUCTION

The genesis of this dissertation comes from responses to two questions tied to risk. The first question regards the testing of seed viability, the ability of a seed to germinate. A seed is a living organism and over time it dies. The United States government has assembled a vast collection of seed lots in long-term storage facilities for many decades. The goal of storing these lots is to preserve genetic information that may be useful in the future. Genebank managers maintain these stores of seeds and must periodically test the lots to ensure they are still viable. In the event that less than 50% [85% or some other critical value] of seeds germinate in a viability test, the manager regenerates the lot. Genebank managers use their subjective discretion when scheduling viability tests. In a perfect world, they schedule every test at the right moment. In such a world, every lot has exactly one test, and the proportion of viable seeds is very near the critical germination proportion that determines the regeneration of a lot. Genebank managers do not want to risk losing a seed lot because a test is too late, and a large portion, if not all, of the seeds are dead. They do not want to waste resources because the test is too early, and a large portion, if not all, of the seeds are viable. Five years ago, Mark Widrlechner and David Kovach asked Philip Dixon and me if a statistically based, seed-viability testing schedule was possible. Chapter 2 is the result of our collaboration.

The creation of a viability-test schedule requires the use of a bevy of statistical methods and procedures. From a collection of historical seed-viability test data on long-term stored maize seed lots, we fit a random-coefficients regression model where the coefficients describe a quadratic curve of seed age versus germination percentage (Laird and Ware (1982)). A seed lot’s individuality is not lost with this model’s form. This is important because we desire a testing schedule that has unique testing ages for each seed lot. Given a critical germination level like 50%, we back-solve predicted age for each seed lot’s quadratic curve. To deduce the
uncertainty of each seed lot’s predicted test age, we frame our model in a Bayesian context (Gelman et al. (2004)). Through a Bayesian model, each seed lot has its own predicted-age distribution derived from the seed lot’s Markov Chain Monte Carlo simulations of random coefficients.

Distributions of predicted test ages permit the investigation of an optimal $\alpha$-quantile seed-test rule. In our study, the $\alpha$-quantiles of the predicted seed-age distributions generate the viability-test schedule. Use of $\alpha$-quantiles is possible because $\alpha$ affects both the risk of testing a seed lot too early and that of testing it too late. Respectively, these are the probabilities of conducting an unnecessary test and that of potentially losing a seed lot. For any seed lot of the maize collection, an $\alpha$ of 0.1 gives a smaller predicted test age than a prediction based on $\alpha = 0.9$. Consequently, we expect the proportion of unnecessary tests by using predicted test ages based on $\alpha = 0.1$ to be larger than that of predicted test ages based on $\alpha = 0.9$. Using out-of-sample (Harrell (2001)) predictions from the testing-age distributions and a data-based standard (see details in section 2.6.1), we use univariate non-parametric kernel density estimators to estimate the risks over $\alpha$ and summarize them through a receiver operating characteristic (ROC) curve. We emphasize our non-parametric approach because our optimality criterion, the cost function, requires a smooth ROC curve with continuous derivatives (Krzanowski and Hand (2009)). To select the best $\alpha$-quantile, the cost function is the natural choice of optimality criterion, because a user can assign the costs associated with each of the risks.

A new procedure that decreases the computing time of a standard two-dimensional Monte Carlo simulation of a risk is the topic of Chapters 3 and 4 (Frey (1992); Hoffman and Hammonds (1994)). Whereas Chapter 2 involves the risk of potentially losing a seed lot or conducting an unnecessary viability test, the context of Chapters 3 and 4 concerns the risk of a disease in a human population. Adapting the risk model of Pouillot and Delignette-Muller (2010), we estimate the probability that a young child exposed to *Escherichia coli* (*E. coli*) O157:H7 consumed in a frozen ground beef patty contracts hemolytic uremic syndrome (HUS), a disease that adversely affects kidneys (PubMed Health (2012)). The purpose of a two-dimensional Monte Carlo simulation of risk is to propagate aleatoric and epistemic uncertainties related
to risk (Frey (1992); Hoffman and Hammonds (1994)). Aleatoric risk variables have inherent, irreducible variability. A young child’s serving size of ground beef is a HUS-risk example because it changes across all children and is not constant. An epistemic parameter has reducible uncertainty. The mean serving size of ground beef given to young children is a HUS-risk example. As one collects more data on serving sizes, the variability of the estimated mean shrinks. Currently, the synthesis of risk through a two-dimensional Monte Carlo simulation requires a large sample of epistemic parameters (outer loop/dimension of the two-dimensional Monte Carlo simulation) and a large sample of aleatoric variables (inner loop/dimension of the two-dimensional Monte Carlo simulation).

In Chapter 3, we propose a faster quantitative risk assessment procedure, by implementing two modifications that decrease computing time. First, Pouillot and Delignette-Muller’s (2010) model has computationally intensive parametric and non-parametric bootstrap distributions of epistemic parameters. We replace these distributions with sampling distributions based on maximum likelihood estimation (MLE) theory and normal theory. We verify that the bootstrapped and corresponding sampling distributions coincide. Next, we significantly reduce the amount of simulations of aleatoric variables and use a multivariate non-parametric kernel regression estimator of the conditional cumulative distribution functions (CDFs) of risk. The traditional two-dimensional Monte Carlo simulation of a risk assessment simulates a large amount of aleatoric variables to estimate the CDF of risk at each of the simulated epistemic parameter vectors. Using Nadaraya-Watson’s multivariate non-parametric estimator of risk (Nadaraya (1964); Watson (1964)), we are able to reduce the number of simulations. Details follow in section 3.4.1.

A major concern of any multivariate kernel regression estimator is the choice of bandwidth (Härdle (1991)). Chapter 4 investigates the accuracy of the CDF estimators under various bandwidth selectors. Certain bandwidth selectors have computing times that are longer than a traditional two-dimensional Monte Carlo simulation of risk. Thus, we only consider fast bandwidth selectors. Those include the penalized integrated average-squared-error bandwidth selectors and a density-based plug-in. We recommend that risk analysts use the plug-in to generate results of a two-dimensional Monte Carlo simulation, as it is the most accurate selector.
in our investigation.
CHAPTER 2. SCHEDULING VIABILITY TESTS FOR SEEDS IN LONG-TERM STORAGE BASED ON A BAYESIAN MULTI-LEVEL MODEL

Allan Trapp II, Philip Dixon, Mark P. Widrlechner, and David A. Kovach

A paper published in Journal of Agricultural, Biological, and Environmental Statistics

Abstract

Genebank managers conduct viability tests on stored seeds so they can replace lots that have viability near a critical threshold, such as 50 or 85 % germination. Currently, these tests are typically scheduled at uniform intervals; testing every 5 years is common. A manager needs to balance the cost of an additional test against the possibility of losing a seed lot due to late retesting. We developed a data-informed method to schedule viability tests for a collection of 2,833 maize seed lots with 3 to 7 completed viability tests per lot. Given these historical data reporting on seed viability at arbitrary times, we fit a hierarchical Bayesian seed-viability model with random seed lot specific coefficients. The posterior distribution of the predicted time to cross below a critical threshold was estimated for each seed lot. We recommend a predicted quantile as a retest time, chosen to balance the importance of catching quickly decaying lots against the cost of premature tests. The method can be used with any seed-viability model; we focused on two, the Avrami viability curve and a quadratic curve that accounts for seed after-ripening. After fitting both models, we found that the quadratic curve gave more plausible predictions than did the Avrami curve. Also, a receiver operating characteristic (ROC) curve analysis and a follow-up test demonstrated that a 0.05 quantile yields reasonable predictions.
2.1 Introduction

Many plant germplasm collections no longer exist in their native environments. The genetic traits of these germplasm collections are valuable for crop improvement and new-product development and provide beneficial information to researchers. For example, the Germplasm Enhancement of Maize (GEM) Program uses different maize landraces from around the world to develop germplasm less vulnerable to crop pathogens, insects, and abiotic stresses (USDA-ARS (2010a)). Institutions world-wide recognize the benefits of germplasm preservation, and the Agricultural Research Service of the United States Department of Agriculture (USDA-ARS) coordinates an extensive network of stations across the United States charged with obtaining, preserving, regenerating, distributing, enhancing, and maintaining the health of plant germplasm (USDA-ARS (2010b)). Although material may be stored in many forms, e.g., seeds, tubers, whole plants, tissues, or roots, seed storage is the most common.

Stored seeds do not live forever and eventually need regeneration (Sharrock et al. (1998)). The life expectancy of stored seeds may be predicted by a seed-viability model. Some models incorporate storage temperature and seed-moisture content to predict germination as a function of seed age. Examples include one proposed by Ellis and Roberts (1980) and its modifications (Hay et al. (2003); Mead and Gray (1999); Tang et al. (2000)). Other models, such as the Johnson-Mehl-Avrami kinetics model (Walters et al. (2005)), consider temperature and moisture to be fixed and express germination only as a function of seed age. Because we are using data from a controlled environment seed-storage facility, we focus on this second model type.

The Avrami model’s sigmoidal shape readily accounts for seed lots that maintain initially high viability levels over a long period of time and have abrupt decay rates (Walters et al. (2005)). But it poorly describes seed lots with germination values that increase during early years of storage (e.g., first 7 to 10 years) or lots with germination values well below 100% at the start of storage. When the Avrami model is fit to lots with either characteristic, the predicted viability curve may be concave up, and model predictions can be unrealistic. Examples of these poor model fits may be found in section 5.

A suitable viability model can provide guidelines for genebank managers to schedule via-
An appropriate model will flag seed lots that are reaching a specified minimum germination level so they can be tested more frequently. A good model should also accurately identify seed lots that retain moderate to high viability levels. These could be tested less frequently.

We propose that seed viability can be modeled adequately as a quadratic function of seed age. The general 3-parameter quadratic regression is

\[ y = \beta_0 + \beta_1 t + \beta_2 t^2 + \epsilon \]  

(2.1)

where \( t \) is storage time and \( y \) the observed germination percentage. We assume \( \epsilon \)’s are independent and identically distributed as normal random variables. \( \beta_0 \) represents the initial germination percentage at \( t = 0 \), while \( \beta_1 \) and \( \beta_2 \) are parameters describing how germination levels change over time. The quadratic model was efficiently fit to data on a large number of lots, each with a few observations, by assuming that the three regression coefficients for each lot follow a multivariate normal distribution (Laird and Ware (1982)).

Seed managers need to know when the germination value for a seed lot is predicted to reach a critical value. Given predictions of the regression coefficients for a particular lot, that time to reach a critical value is calculated by solving (2.1) for \( t \). Our decision rule is based on the distribution of predicted \( t \) for each seed lot. We chose a Bayesian approach to estimate these distributions because it was considerably faster than a non-parametric bootstrap and more robust to misspecification of the random effects distributions than a parametric bootstrap.

In the following pages, we argue that a multi-level model with equation (2.1) as the first-level model can be used by genebank managers to predict viability and determine viability testing times for individual seed lots. Development and support of this model is derived from maize-viability testing data. Follow-up testing has been done to verify the accuracy of our model-based predictions.
2.2 Data Description

Our maize data come from the USDA-ARS, North Central Regional Plant Introduction Station (NCRPIS) located in Ames, Iowa. There are 2,833 seed lots representing 2,314 unique accessions of maize with at least 3 viability tests conducted at different times, resulting in a total of 11,558 observations. This data set is based on regenerated and currently stored lots produced since 1948.

In any given seed lot, the expected initial germination value is not 100%. Additionally, a living seed may not germinate during a viability test. A given seed lot may have a portion of dead seeds, and some seeds may remain dormant during testing. Seed dormancy is defined as a unique ordering of blocks to seed germination that has evolved over time to adapt plants to climate patterns and the abundance or scarcity of resources, such as water, oxygen, nutrients, and light (Finch-Savage and Leubner-Metzger (2006); Holdsworth et al. (2008)). During testing, the researcher may not provide conditions that release dormancy for all seeds. And, it is often difficult to distinguish with certainty between dormant and inviable seeds. Thus, we conducted analyses on what the seed-testing literature calls “normal germination percentages.” The fate of a seed in a viability test is classified as normal germination, abnormal germination, dormant, or dead. A seed has normal germination when the resulting seedling has essential structures that indicate it will develop into a mature plant (ISTA (2009)).

Plots of trends in germination values of these 2,833 seed lots over time revealed the potential presence of after-ripening. Seed after-ripening refers to metabolic processes that must occur in otherwise mature seeds before germination can occur. After-ripening can be most readily observed when seeds are stored for a period of time at room-temperature conditions (Bewley (1997); Finch-Savage and Leubner-Metzger (2006); Leubner-Metzger (2003)). But after-ripening at a reduced rate may also occur at colder temperatures (e.g. 4 °C) for some species (Chantre et al. (2009); Steadman et al. (2003); Widrlechner (2007)), sometimes even at temperatures slightly below the freezing point (Sivakumar et al. (2006); Wang et al. (2004)). In other words, a slow after-ripening process in cold storage may lead to a gradual increase in normal germination percentage, which is eventually counteracted by a long-term decline in
overall viability.

We have data on normal seed germination percentage and seed age. Seed age was measured in whole-year increments starting at 0.5 years. Seed germination was recorded in increments of a whole percent based on tests conducted on 200-seed samples. In each viability test, 4 groups of 50 seeds were tested on 4 separate towels grouped together. Consequently, all 200 seeds of a test were not subjected to a completely homogeneous test environment.

Much of the seed for this analysis has been stored under similar conditions over time. All 2,833 seed lots were dried, packed in clear, moisture-impermeable containers, and stored in a room held at 4°C and 25% relative humidity for the last 15 years. Due to historical changes in seed-storage conditions, some older seed lots were not continuously stored under these conditions, but instead were subjected to higher relative humidity levels. Also, different personnel conducted these germination tests over the past 61-year span. When considering the inconsistent storage conditions for older seeds and the execution of tests, we expect some overdispersion in the data.

The number of seed tests for each seed lot ranged from 3 to 7. 1,180 lots were tested 3 times; 914 lots were tested 5 times; 519 lots were tested 4 times; and the remaining 220 were tested 6 or 7 times. Interval length between any 2 consecutive tests of a seed lot varied. The median length between successive tests was 6 years, with 50% between 5 and 7 years, reflecting past practices at the NCRPIS. There was 1 seed lot where 32 years passed before it was tested again.

2.3 Modeling Seed Viability for Many Seed Lots

In this section, we outline two viability models that have different shapes of seed-viability curves. The first model is a parabola. Germination is described through a quadratic form of seed age. Parabolic viability curves can describe seeds lots with after-ripening. The second model relates seed age and germination through Johnson-Mehl-Avrami kinetics (Walters et al. (2005)). These curves monotonically decrease over time and do not accommodate after-ripening. Both models include seed-lot specific random effects for each of the parameters in the respective model.
2.3.1 Quadratic Random Coefficients Model

Motivation for our model in equation (2.1) comes from the data patterns.

![Graph of seed age vs germination]

Figure 2.1: (a) Plot of 3 Prevalent Types of Seed-Lot Data Patterns: The points on the plot represent data from 3 separate seed lots. Curves are drawn by using ordinary least squares estimates of the parameters in equation (2.1). (b) Shrinkage Plot: This plot shows how curves from one maize seed lot shrink to the overall curve (thick solid curve) of the maize seed collection when information about a seed lot is left out of the model. The first 3 testing observations were used to generate the thin solid curve; the first 4 were used to create the dashed curve; the first 5 were used to create the dotted line; all observations were used to create the dotted/dashed line.

Many of the seed lots had germination patterns that followed 1 of 3 patterns (Figure 2.1(a)). Declines in seed viability for “high-viability” lots were relatively small or unobserved during the duration of testing. Declines in viability were observed in the “traditional” and “after-ripening” patterns. In a “traditional” seed lot, germination monotonically decreases over time. In an “after-ripening” lot, germination increases during the early years of storage, then decreases.

Figure 2.1(a) illustrates two important features of seed germination: the initial germination at time 0 is not always 100%, and germination values may increase over time. Similar to a log-
linear model of viability (Ellis and Roberts (1980)) and variants thereof (Hay et al. (2003); Mead and Gray (1999); Tang et al. (2000)), our model allows for different initial ($t = 0$) germination values. Unlike many other models (Ellis and Roberts (1980); Hay et al. (2003); Mead and Gray (1999); Tang et al. (2000); Walters et al. (2005)), our model allows for after-ripening. Figure 2.1(a) illustrates the flexibility of our quadratic model (the smooth curves) to conform to the 3 common data patterns.

We have viability data from thousands of maize seed lots, but few germination tests per seed lot. Biologically, we expect that seed-germination curves for different seed lots from a maize collection stored under common conditions will share some viability characteristics over time. Instead of fitting individual curves to each seed lot, which requires at least 3 observations per seed lot, we fit a multi-level model to the entire maize collection of 2,833 seed lots. This provides seed-lot-specific viability curves from which we estimate a seed-lot-specific testing age.

Using multi-level notation (Singer and Willet (2003)), our model is:

**Level I:**

$$y_{i,j} = \beta_{0,i} + \beta_{1,i} t_{i,j} + \beta_{2,i} t_{i,j}^2 + \epsilon_{i,j}$$  \hspace{1cm} (2.2)

$$\epsilon_{i,j} \sim N \left( 0, \frac{1}{\tau^2} \right)$$  \hspace{1cm} (2.3)

**Level II:**

$$\beta_{0,i} = \beta_0 + \zeta_{0,i}$$  \hspace{1cm} (2.4)

$$\beta_{1,i} = \beta_1 + \zeta_{1,i}$$  \hspace{1cm} (2.5)

$$\beta_{2,i} = \beta_2 + \zeta_{2,i}$$  \hspace{1cm} (2.6)

$$\begin{bmatrix} \zeta_{0,i} \\ \zeta_{1,i} \\ \zeta_{2,i} \end{bmatrix} \sim \text{MVN}(0, \Sigma) \quad \text{or} \quad \begin{bmatrix} \beta_{0,i} \\ \beta_{1,i} \\ \beta_{2,i} \end{bmatrix} \sim \text{MVN}(\beta, \Sigma)$$  \hspace{1cm} (2.7)
where
\[ i = \text{the seed lot, } 1 \leq i \leq 2,833, \]
\[ j = \text{the testing occasion, } 1 \leq j \leq n_i, \text{ and } n_i \text{ is the number of tests for seed lot } i, \]
\[ y_{i,j} = \text{the germination level for the } i^{th} \text{ seed lot on testing occasion } j, \]
\[ t_{i,j} = \text{the seed age of seed lot } i \text{ on testing occasion } j, \]
\[ \tau^2 = \text{the precision of the conditionally independent and identically distributed } \epsilon_{i,j}\text{'s,} \]
\[ \beta_0 = \text{the mean initial germination for the entire maize seed collection,} \]
\[ \beta_1 = \text{the mean of the linear coefficient of the entire maize seed collection,} \]
\[ \beta_2 = \text{the mean of the quadratic coefficient of the entire maize seed collection,} \]
\[ \beta' = [\beta_0, \beta_1, \beta_2], \text{ and} \]
\[ \Sigma = \text{a 3 by 3 unstructured covariance matrix.} \]

This hierarchical formulation garners information from all maize seed lots and uses it to generate a shrunk curve for each lot. Figure 2.1(b) shows how shrinkage occurs in our data. For a seed lot that has 6 actual germination tests, the thin solid curve was produced from the first 3 tests. If a line were fit only to these points (e.g. using ordinary least squares), the curve would be concave up. This curve would be implausible because the model would indicate that germination rises above 100% after year 10. With a multi-level model, information from the maize collection shrinks these data effects into a curve that resembles the overall curve (thick solid line). However, as we incorporate more information about the lot into our model, the successive curves tend to follow the lot data more closely. This lot has 100% germination at tests 3, 4, and 5. When the number of observations increases, the predicted viability curves begin to flatten and approach the seed-lot-specific’s observed values.

### 2.3.1.1 Bayesian Estimation of the Quadratic Multi-Level Model

The parameters in equations (2.2), (2.3), and (2.7) were estimated through a Bayesian approach (Gelman and Hill (2007)). The decision rule that we develop later is based on quantiles of the distribution of a non-linear function of the seed-lot-specific parameters in
equation (2.2). More details on these age predictions are covered in the Seed-Testing Predictions section. Computations under the Bayesian approach were substantially faster than under a non-parametric bootstrap.

Since we had no a priori knowledge of the prior distributions of the parameters $\tau^2$, $\beta$, or $\Sigma$, diffuse priors were specified. Note that the matrix in equation (10) is an inverse variance-covariance matrix.

\[
\tau^2 \sim \Gamma(0.001, 0.001) \tag{2.8}
\]

\[
\beta \sim \text{MVN} \left( \begin{bmatrix} 0 \\ 100 & 0 \\ 0 & 0 \\ 0 & 0 & 100 \end{bmatrix} \right) \tag{2.9}
\]

\[
\Sigma^{-1} \sim \text{Wishart} \left( \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, 3 \right). \tag{2.10}
\]

We evaluated the sensitivity of the posterior distributions to the choice of prior distribution for $\tau^2$ in (2.8). We considered three prior distributions: $\Gamma(0.01, 0.01)$, $\Gamma(0.001, 0.001)$, and $\text{Uniform}(0, 10000)$ (Gelman (2006)). The posterior distributions of $\tau^2$ coincided under (2.8) and $\text{Uniform}(0, 10000)$ priors, which are more diffuse than a $\Gamma(0.01, 0.01)$ prior. Also, the medians of the posterior distributions of all parameters in the Bayesian model closely resembled the estimates from a random coefficient model fitted by restricted maximum likelihood. This indicates that the prior assumptions, (2.8)-(2.10), minimally influenced the posterior distributions of the parameters (Gelman et al. (2004)). We report results using $\Gamma(0.001, 0.001)$ as the prior of $\tau^2$.

Estimation of the parameters was carried out through Markov Chain Monte Carlo (MCMC) methods in WinBUGS (Lunn et al. (2000)). A burn-in period of 40,000 MCMC draws was selected by visually inspecting parameter trace plots and considering the effective number statistic, $n_{\text{eff}}$. To verify appropriate mixing of draws from the posterior distributions, three chains were initialized at disparate regions of the parameter space. Then we looked for $\hat{R}$ statistics near a value of 1 and visually inspected trace plots again (Gelman et al. (2004)). In
total, there were 10,000 draws from each posterior distribution after burn-in, but we thinned each chain by a factor of 15 to avoid computer memory issues due to the large number of model parameters.

### 2.3.2 Avrami Formulation of the Seed-Viability Model

Our multi-level quadratic model is not a commonly used seed-viability model. We compared it to a multi-level form of a commonly used viability model described by Walters et al. (2005). They modeled seed viability over time through Johnson-Mehl-Avrami kinetics (Williams et al. (1993)). This model is customarily fit in its linearized form:

\begin{equation}
\ln(-\ln(y_{i,j})) = m_i \ln(t_{i,j}) + c_{i,0} + \epsilon_{i,j}
\end{equation}

where \( c_{i,0} \) is the initial germination and \( m_i \) is the slope on the complimentary log-log scale. \( t_{i,j} \) and \( y_{i,j} \) are defined in section 2.3.1. Since there were seed tests that yielded 0% or 100% germination, we modified them with Berkson’s empirical adjustment (Berkson (1953)). Specifically, 0% germination values were replaced with \( 1/(2(200)) = 0.0025 \), and 100% germination values were replaced with 0.9975.

We assumed that \( \epsilon_{i,j} \sim N(0, 1/\tau^2) \) in equation (2.11). As in equation (2.7), we fitted a multivariate normal distribution with an unstructured covariance matrix to \( [m_i, c_{i,0}]' \). Lastly, diffuse priors were assumed for the precision, the coefficients vector, and the variance of coefficients. These prior assumptions mirrored those outlined in section 2.3.1.1. Procedures to determine appropriate burn-in and mixing were also the same.

### 2.4 Seed-Testing Predictions

After fitting the quadratic and Avrami models, we formulated seed-lot specific predictions of the age when viability drops below a critical amount, \( CRV \). \( CRV \) is the lowest acceptable level of viability as specified by a genebank manager. Selection of a \( CRV \) germination value determines the test schedule for a seed lot. In this paper, we selected \( CRV = 0.50 \). The critical germination rate could easily be modified to meet other genebank standards.
Given a CRV and parameter estimates from either of the models in sections 2.3.1 or 2.3.2, we inverted equations (2.2) and (2.11) and solved for seed age. For our quadratic model, the predicted testing age of lot $i$ at a posterior draw of the vector $[\hat{\beta}_{0,i}, \hat{\beta}_{1,i}, \hat{\beta}_{2,i}]'$ is

$$\hat{t}_t^Q = \frac{-\hat{\beta}_{1,i} \pm \sqrt{\hat{\beta}_{1,i}^2 - 4\hat{\beta}_{2,i}(\hat{\beta}_{0,i} - CRV)}}{2\hat{\beta}_{2,i}}. \tag{2.12}$$

The $Q$ superscript signifies that $\hat{t}_t^Q$ is a prediction from our quadratic viability model. The predicted curvature/concavity (evidenced through sign of $\hat{\beta}_{2,i}$), prediction of initial germination ($\hat{\beta}_{0,i}$), and location of the predicted curve’s apex all affect which root will be used as the predicted testing age for seed lot $i$. Also, there will be instances where no root exists. Appendix A details when the positive or negative discriminant is used and how we handled situations where the roots do not exist. The inversion of equation (2.11) at some posterior draw of the $[\hat{m}_i, \hat{c}_{i,0}]'$ vector has a single root given by:

$$\hat{t}_t^A = e^{\frac{\ln(-\ln(CRV)) - \hat{c}_{i,0}}{\hat{m}_i}}. \tag{2.13}$$

The superscript $A$ signifies that $\hat{t}_t^A$ is a prediction from the Avrami model.

Sampling distributions of $\hat{t}_t^Q$ and $\hat{t}_t^A$ do not have analytical solutions, but they can be estimated by substituting realizations from either the posterior distribution of $[\beta_{0,i}, \beta_{1,i}, \beta_{2,i}]$ into (2.12) or the posterior distribution of $[m_i, c_{i,0}]$ into (2.13). These posterior distributions provide a range of reasonable seed ages that may be used to schedule future viability tests.

### 2.5 Comparison of Avrami and Quadratic Models

We compare the Avrami and quadratic models by assessing their predictions. A prediction that a seed lot will not reach 50% viability until a seed age of 800 years is contrary to a wealth of past experience for crop seeds (Nagel and Börner (2010)). We use the medians of the $\hat{t}_t^Q$ and $\hat{t}_t^A$ distributions as point predictions of the time to 50% viability for the quadratic and Avrami models because they are “typical” values of the non-symmetrical posterior distributions.

Predictions of time to 50% viability from the Avrami model fail our criteria. In that model, over 94% of the lots have a test age, at $CRV=0.50$, $>200$ years, with 75% having predictions $>864$ years and 23% with predictions $>10,000$ years. 1 lot has a predicted test age of 9,439,000
years. These predictions are unrealistic and impractical. In contrast, 75% of the predictions of time to 50% viability using the quadratic model are < 65 years.

To understand why predictions from the Avrami model are so large, we investigated predictions of individual seed lots with 6 or 7 historical tests. Figure 2.2 shows Avrami and quadratic curves using median parameter estimates for a seed lot that has 6 test occasions. The predicted germination values of the Avrami curve are calculated through equation (2.13). The corresponding median $t^Q_i$ and $t^A_i$ are 46.5 and 296.9 years, respectively. 183 of the 220 lots with 6 or 7 previous seed tests have similar concave-down data patterns as in Figure 2.2. Under the Avrami model, the predicted death rate, or the first derivative of germination with respect to age, for this seed lot decreases as age increases. A concave-up shape is observed. The data do not suggest this pattern. They show that the death rate increases as time increases. A quadratic model captures this essential concave-down seed feature. The Avrami model does not accurately represent the death rates with concave-down curvatures. As a result, many of

![Figure 2.2: Comparison of Avrami (solid line) and Quadratic (dashed line) Curves for 1 Lot: Medians of the posterior distributions of the coefficients were used to create the curves. The quadratic curve captures the germination trend better than the Avrami curve does.](image-url)
its test-age predictions for seed lots are extreme. Further analyses in this paper are conducted using only the quadratic model.

2.6 Selection and Evaluation of a $t_{Q_i}^Q$ Quantile

The purpose of fitting a model is to create practical predictions of when we expect a seed lot to drop below a CRV level. However, managers want to regrow a seed lot before it reaches the CRV. Otherwise, there may not be a large enough portion of viable seeds to reproduce the lot and preserve its genetic profile. When one tests seeds at their predicted age, one of two negative consequences may happen. If the observed germination is above the CRV, money will have been spent to test seeds unnecessarily. If the observed germination is below the CRV, viability may no longer be acceptable. The costs of these consequences are not equal. The cost of an additional test is much less than the cost of unacceptable viability. Assuming our quadratic model is appropriate, predictions based on the posterior medians of $t_{Q_i}^Q$ are equally likely to be less than or greater than the true age when CRV occurs. Instead of using medians, we suggest using a lower quantile. Choice of which lower quantile is investigated through a receiver operating characteristic (ROC) curve that summarizes the true positive and the false positive rates over a range of quantiles.

2.6.1 ROC curve development

We need to evaluate decision rules based on the posterior distributions of the $t_{Q_i}^Q$'s. This is difficult because the true age at which a lot reaches a CRV, the “gold standard,” is unknown. It is impossible to continuously record the true viability of a seed lot. We can only test all seeds once. In place of a gold standard, for a given seed lot $i$, we compare a prediction of its $t_{Q_i}^Q$ distribution to its status at the last test, $t_{\text{last},i}$.

The data-based standard is appropriate here because the decision to regrow a seed lot will be based on the observed germination percentage at the last test. For this section, predictions were generated from the quadratic model refitted to the data set excluding observations at the last test for each lot. Comparisons between these predictions and $t_{\text{last},i}$ are an out-of-sample (Harrell (2001)) assessment of a model’s ability to predict future performance. If an observed
germination at $t_{\text{last},i}$ is less than CRV, the correct decision is a predicted $\hat{t}_i^Q$ that is earlier than $t_{\text{last},i}$. In other words, for a lot that is below CRV at $t_{\text{last},i}$, we want a prediction that schedules a germination test before that last testing age. If observed germination at $t_{\text{last},i}$ is above CRV, then the correct decision is a predicted $\hat{t}_i^Q$ that is later than $t_{\text{last},i}$. We want a prediction that schedules a germination test later than the most recent test.

Table 2.1: Test Criteria and Decisions for a 50% Critical Regrow Value: $t_\alpha$ represents the $\alpha$-quantile of a $t$ posterior distribution.

<table>
<thead>
<tr>
<th>Decision based on $t_\alpha$</th>
<th>Predicted test age was before last observed germination test</th>
<th>Observed germination at last germination test</th>
<th>Predicted test age was after last observed germination test</th>
</tr>
</thead>
<tbody>
<tr>
<td>False Positive ($FP$)</td>
<td>100%</td>
<td>True Negative ($TN$)</td>
<td>Correct prediction</td>
</tr>
<tr>
<td>Incorrect prediction</td>
<td>$\uparrow$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>50%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seed lot age: $t_0$</td>
<td></td>
<td></td>
<td>$t_z$</td>
</tr>
<tr>
<td>True Positive ($TP$)</td>
<td>$\uparrow$</td>
<td>False Negative ($FN$)</td>
<td>Incorrect prediction</td>
</tr>
<tr>
<td>Correct prediction</td>
<td>$\uparrow$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The positive population, $P$, is those seed lots that needed to be regrown before $t_{\text{last},i}$ because their germination values at $t_{\text{last},i}$ are smaller than the CRV. There were 147 seed lots in $P$. These seed lots should have been regrown before $t_{\text{last},i}$. Table 2.1 outlines the two decisions associated with individuals in $P$. A true positive, $TP$, occurs if a seed lot from $P$ has a predicted age that is before its $t_{\text{last},i}$, i.e., the model prediction is “regrow before $t_{\text{last},i}$.” A false negative, $FN$, is the case where a seed lot from $P$ has a predicted test age after its $t_{\text{last},i}$, i.e. the model predicts “no need to regrow before $t_{\text{last},i}$.”
The negative population, $N$, is those seed lots that do not need to be regrown before $t_{last,i}$ because their germination values at $t_{last,i}$ are greater than the CRV. There were 2,686 sampled lots in $N$. A false positive, $FP$, is a member of $N$ where its predicted test age is earlier than its $t_{last,i}$. But a viability test has already been conducted at a later seed age, $t_{last,i}$, and the germination was greater than the CRV, so the prediction is misleading and cannot be acted upon. The final class, true negative or $TN$, describes where a seed lot from $N$ has a predicted test age after its $t_{last,i}$. Table 2.1 lays out all possible classifications that result from a prediction.

We use an $\alpha$ quantile of the posterior distribution of $\hat{t}^Q_i$, calculated by omitting germination at each seed lot’s last test data, as the prediction for seed lot $i$. Different choices of $\alpha$ lead to different sets of predictions. Therefore, the counts in Table 2.1 change. For example, predictions based on a large $\alpha$ quantile, such as $\alpha = 0.9$, will have more $FN$’s than would predictions based on an $\alpha = 0.2$ quantile. Paraphrased, if we had not “known” germination at $t_{last,i}$ and used $\hat{t}_{0.9,i}$’s instead of $\hat{t}_{0.2,i}$’s as the subsequent test ages, more seed lots would have dropped below CRV germination levels when tested. Those seed lots possibly would have been lost. Conversely, if $\hat{t}_{0.2,i}$ ages were chosen, then we would commit more $FP$’s. We would have unnecessarily spent money on tests, in retrospect.

We summarize the relationship between the choice of $\alpha$ and the true positive and false positive rates using an ROC curve, estimated by non-parametric smoothing (Krzanowski and Hand (2009)). This provides a smooth, differentiable estimate of the ROC curve and facilitates estimation of optimal $\alpha$-quantile predictions (to be discussed in the next section). For each $\alpha$, we counted the number of $TP$’s in population $P$ and the number of $FP$’s in population $N$ and calculated the corresponding rates, $tpr(\alpha)$ and $fpr(\alpha)$. Because a kernel smoother has problems near the boundaries, $\alpha = 0$ and $\alpha = 1$, we estimated the density on the logit($\alpha$) scale. We used a Berkson (1953) correction of $1/(2(2,001))$ for the empirical $\alpha_i$’s that were 0, i.e. all draws in lot $i$’s posterior distributions of $\hat{t}^Q_i$ were greater than the seed lot’s last test age. This happened with 1,057 lots. The density with regard to logit($\alpha$) in the negative and positive populations, $\hat{f}_N(x)$ and $\hat{f}_P(x)$, respectively, was estimated by using a biweight kernel smoother (Silverman (1998)). To estimate $fpr(\alpha)$ for a given $\alpha$, we integrated $\hat{f}_N(x)$ for logit($x$) $\leq$ logit($\alpha$). For the
same $\alpha$, $tpr(\alpha)$ was calculated by integrating the density $\hat{f}_P(x)$.

Figure 2.3: (a) Non-Parametric Density Estimates of $\alpha$ with Respect to Negative (Solid Line) and Positive (Dashed Line) Populations: Density curves pertain to the logit transformation of $\alpha$. Areas shaded to the right represent the false positive and true positive rates corresponding to an $\alpha = 5\%$ threshold. (b) ROC Curve: Point on the ROC curve corresponds to $fpr$ and $tpr$ shaded area in (a).

The points on the ROC curve in Figure 2.3(b) represent $(fpr(\alpha), tpr(\alpha))$’s for various $\alpha$ thresholds. The area under the ROC curve is 0.87. Our $\alpha$ quantile rule can discriminate between the populations $N$ and $P$ (Krzanowski and Hand (2009)). The point on Figure 2.3(b) corresponds to a threshold of 5%.

2.6.2 Selection of an Optimal Quantile of $t_i^Q$ Distributions

The ROC curve summarizes the error rates for various $\alpha$ choices. However, the curve fails to account for the costs associated with misclassification and the relative proportion of observations from $N$ and $P$. These are crucial components in selecting an optimal threshold. The estimated relative proportion of lots in $P$ is rather small at $147/2,833$. Consequently, we applied the cost function

$$C(\alpha) = q \times (1 - tpr) \times C(N|P) + (1 - q) \times fpr \times C(P|N)$$

(2.14)
as the criterion in choosing an optimal threshold. The relative proportion of seed lots in $P$, $q$, is estimated from our data. In equation (2.14), costs associated with an $FN$ and an $FP$ are represented as $C(N|P)$ and $C(P|N)$, respectively. We believe that the cost of potentially losing a seed lot, an $FN$, is considerably higher than the cost of conducting a premature test. We consider cost functions where $C(N|P)/C(P|N) > 1$.

Under the criterion in equation (2.14), the $\alpha$ that minimizes the cost of misclassification corresponds to the point on the ROC curve with a derivative of

$$\frac{(1 - q)C(P|N)}{qC(N|P)}.$$  \hfill (2.15)

Provided the slope is well-defined at a given $\alpha$, the derivative of the ROC curve at $\alpha$ is equal to the ratio of $P$ and $N$ densities evaluated at that $\alpha$ (Krzanowski and Hand (2009)).

A researcher must choose a suitable cost ratio to determine an optimal $\alpha$ quantile. We estimate that the expense associated with potentially losing a lot to that of possibly conducting a premature test is approximately 30. For a $C(N|P)/C(P|N) = 30$ (derivative of 0.61), we have an optimal threshold of $\alpha = 0.05$. However, if one believes that the cost of premature testing equals the cost of losing a seed lot ($C(N|P)/C(P|N) = 1$), then $\alpha = 0.99$ (derivative of 18.27) is suitable. In contrast, one may specify a high cost ratio of 100. This corresponds to an $\alpha = 0.002$ (derivative of 0.18). In subsection 2.6.1, it was noted that the densities of $P$ and $N$ are not unimodal. This results in derivatives that may not be well-defined. Graphical exploration suggests that unique derivatives do exist for cost ratios of 1, 30, and 100.

### 2.6.3 Evaluation of $\alpha = 0.05$ Quantile Rule

In this section, we investigate the effectiveness of predictions based on the $\alpha = 0.05$ quantiles of the $t^Q_i$'s. A stratified sample of 125 seed lots was taken from the original 2,833 lots and was retested in 2009. We calculated the observed $TP$ and $FP$ rates and compared them to their corresponding estimated rates from section 2.6.1. Since a cost ratio of 30 is associated with an $\alpha$ of 0.05, we expected to see more premature tests, $FP$'s, than late tests (i.e., tests with germination values below 50%), $FN$'s.

We stratified the maize seed lots into 5 strata to better understand the decision rule's
An early implementation of our model was used to assign seed lots to strata. *All* data points were used to generate predictions. Lots with predicted viability of 50% before the year 2000 were assigned to stratum 1. Seed lots with predicted viability around 50% between 2007 and 2012 were allocated to stratum 2. Strata 3 and 4 were composed of seed lots with predicted test times between 2012 and 2018. Stratum 3 included only lots with 3 or 4 previous tests whereas stratum 4 included lots with 5 or more previous tests. Stratum 5 was limited to seed lots with predicted test times at or beyond 2033. 25 seed lots were randomly sampled from each stratum.

Table 2.2: Count Results from the 5-Stratum Follow-Up Study: The number of false negative and false positive cases are in columns $FN$ and $FP$, respectively. The number of sampled lots from the negative and positive populations are in columns $N$ and $P$, respectively. The Population column shows the number of lots from the original 2,833 that are in the corresponding stratum. The Retest column indicates that we sampled 25 lots from each stratum.

<table>
<thead>
<tr>
<th>Stratum</th>
<th>Size of the Population in 2009</th>
<th>Measured Germination in 2009</th>
<th>Number of Misclassifications</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>550 25</td>
<td>15 10</td>
<td>15 0</td>
</tr>
<tr>
<td></td>
<td>181 25</td>
<td>24 1</td>
<td>20 0</td>
</tr>
<tr>
<td></td>
<td>208 25</td>
<td>24 1</td>
<td>18 1</td>
</tr>
<tr>
<td></td>
<td>164 25</td>
<td>25 0</td>
<td>2 0</td>
</tr>
<tr>
<td></td>
<td>308 25</td>
<td>25 0</td>
<td>1 0</td>
</tr>
</tbody>
</table>

In the retest data, $N$ and $P$ populations were defined with respect to observed germination values of the 125 seed lots in 2009. Lots were assigned to $N$ if their observed germination values were $>50\%$ in the retest. For a lot in $N$, an $FP$ would result when the predicted age, $\hat{t}_{\alpha,i}$, would fall below the actual age of the lot in 2009. Table 2.2 displays the number of $FP$ and $FN$ cases for each of our sampled strata. The empirical $fpr$’s, calculated as $FP/N$, are rather high, especially for strata 1, 2, and 3 (1, 0.83, and 0.75, respectively). In our study, only 1 $FN$
occurred. The adjusted overall $fpr$ is 0.55. The adjusted overall $fnr$ is 0.035.

For $\alpha = 0.05$, we predicted an $fnr$ of 0.097 and an $fpr$ of 0.39. This corresponds to the point (0.39, 0.90) in Figure 2.3 when using the criterion in equation (2.15). Our observed $fnr$ is better than our predicted rate, but the predicted $fpr$ is more optimistic than the observed value (difference is 0.16). Still, the $\alpha = 0.05$ decision rule was able to identify nearly all lots that fell below 50% viability (11 out of the 12 $P$ lots).

2.7 Discussion

We have shown that the Avrami seed viability model (Walters et al. (2005)) fails to model patterns where germination increases over the initial years of storage. Biologically, after-ripening or an initial failure to break seed dormancy may result in such increases. To model these seed phenomena simply, we fitted a 3-parameter quadratic curve of germination at specified seed ages. Presumably, lots of our maize collection may decay in similar patterns over time. Thus, we fitted a Bayesian multi-level model. Curves of seed lots with only 3 previous tests were shrunk to a general maize viability curve, but curves of seed lots with a large number of historical tests were more individualistic. Since the goal was to estimate when to retest seed lots, we generated predictions of time to reach a critical germination percentage from our quadratic multi-level model and compared them to predictions based on the Avrami multi-level model. We found that our predictions were much more realistic than those based on the Avrami model.

From a physiological perspective, there is no reason to assume a symmetric viability curve. The pattern of increasing germination values over initial storage may not mirror the pattern of decreasing germination values over the later storage years. However, this is not a major concern. The goal is to fit a model that provides reasonable predictions of test ages. Our retest study shows that predicted test ages based on the 5% quantile of $\hat{t}_q^Q$ for 113 of the 125 retested lots are smaller than the actual ages when germination falls below 50%. This suggests that our prediction model is on the safe side. The observed $fpr$ is larger than the model-based estimate of $fpr$ (0.55 > 0.39) for our $\alpha = 0.05$ rule. These retest results suggest that if there is bias in the predictions, then it is towards premature predictions.
Predictions in this paper were based on a CRV of 50% and an $\alpha$ quantile of 5%. “Standard” values used by various genebanks can easily replace these quantities. If lot regeneration requires a more complete representation of a sample to ensure the conservation of its genetic profile, then a genebank manager may assign a larger CRV. If a genebank manager wishes to have a smaller proportion of premature tests, then one can choose an $\alpha > 5%$. Our model is flexible in meeting managers’ needs.

A key component of our model is its ability to pool information across similar seed lots. Although it was created for a collection of maize lots, our multi-level model should easily generalize to a collection of similar lots of other plant species. Our model is not limited to a single crop.

Lastly, our model’s predictions are dynamic. Every new viability-test data point will provide information on the current viability status of an individual lot and on the characteristics of the entire collection. Consequently, the entire collection’s viability test schedule can be updated with every successive test. Our populations, $N$ and $P$, change over time. Thus, it is appropriate that predictions change with each new viability assessment.

### 2.8 Acknowledgements

Journal paper of the Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa, Project No. 1018, was supported by USDA-ARS Research Project 3625-21000-053, Hatch Act and State of Iowa funds. Mention of commercial brand names does not constitute an endorsement of any product by the U.S. Department of Agriculture or cooperating agencies. We would also like to thank Dr. Candice Gardner for her input.
CHAPTER 3. MULTIVARIATE KERNEL REGRESSION-BASED RISK ASSESSMENT THROUGH A MODIFIED 2-DIMENSIONAL MONTE CARLO SIMULATION

Allan Trapp II and Philip Dixon
A paper to be submitted to Risk Analysis

Abstract

The two-dimensional Monte Carlo simulation is an important tool for quantitative risk assessors. Its framework easily propagates aleatoric and epistemic uncertainties related to risk. Aleatoric uncertainty concerns the inherent, irreducible variability of a risk factor. Epistemic uncertainty concerns the reducible uncertainty of a fixed risk factor. Human weight is an example of an aleatoric uncertainty while the mean of human weights is an epistemic uncertainty.

The traditional application of a two-dimensional Monte Carlo simulation in a risk assessment requires many Monte Carlo samples. In a common case, a risk assessor samples 10,000 epistemic factor vectors. For each vector, the assessor generates 10,000 vectors of aleatoric factors and calculates risk. The purpose of heavy aleatoric simulation is to estimate a cumulative frequency distribution, CDF, of risk conditional on an epistemic vector. This approach has $10^8$ calculations of risk and is computationally slow. We propose a more efficient method that reduces the number of simulations in the aleatoric dimension. For each vector of epistemic factors, we pool together the risk values of epistemic vectors close to it and estimate the conditional CDF by using the multivariate Nadaraya-Watson estimator. If a researcher uses numerically intensive sampling, like the bootstrapping of epistemic sample quantities (e.g., mean and variance), then we may further decrease computing speed and simulate the epistemic values from either an asymptotic maximum likelihood distribution or a normal-theory based distribution. For the
context of our exposition, we examine the risk of hemolytic uremic syndrome in young children exposed to *Escherichia coli* O157:H7 in frozen ground beef patties. We demonstrate that our method replicates the results of the traditional two-dimensional Monte Carlo risk assessment. Furthermore, for this problem, we find that our method is computationally three times faster than the traditional method.

### 3.1 Introduction

Risk assessors in microbiology, ecology, human health, engineering, mathematics, and statistics have been estimating risk related to potentially hazardous processes for many years. For example, an engineer may estimate the risk of an accident at an intersection (Busschaert et al. (2011)), or a public health researcher might examine the risk of a particular disease in a human population exposed to contaminated drinking water (Frey (1992)). In this paper, we examine the risk of hemolytic uremic syndrome (HUS) in young children exposed to *Escherichia coli* (*E. coli*) O157:H7 consumed in frozen ground-beef patties (Pouillot and Delignette-Muller (2010)). Within each scenario, risk analysts must address some common, pervading issues.

The first is the representation of risk factors. Naturally, in order to estimate risk, an analyst needs “data” on risk factors. This information comes from two sources: field data, which may be experimental or observational in nature, and/or expert knowledge. Commonly, field-data risk factors are represented through probabilistic densities (Helton and Johnson (2011); Jakeman et al. (2010); Pouillot and Delignette-Muller (2010)). The representation of expert knowledge in a risk assessment is more ambiguous and is often contested because an expert’s knowledge is subjective. Nonetheless, expert knowledge plays a crucial role in risk analysis because some risk variables cannot be directly observed in the field, or information on the variable is very difficult to gather due to measurement limitations and/or excessive cost. Common methods to represent and quantify expert-informed variables include interval analysis (Jaulin (2001); Kearfott and Kreinovich (1996); Moore (1966); Neumaier (1990)), possibility theory (Baudrit et al. (2008); Dubois (2006); Dubois and Prade (2006); Zadeh (1978)), fuzzy set theory (Baudrit et al. (2008); Dubois and Prade (1980); Zadeh (1965)), evidence theory (Halpern and Fagin (1992); Shafer (1976); Wasserman (1990)), and probability theory (Cullen and Frey (1999); Fontana
et al. (2009); Frey (1992); Frey and Rhodes (1998); Helton and Johnson (2011); Hoffman and Hammonds (1994); Nauta (2000); Nayak and Kundu (2001); Parry and Winter (1981); Pouillot and Delignette-Muller (2010); Stigler (1986)). Helton and Johnson (2011) describe the merits of each method. We take a traditional approach and represent these factors through probability theory, because it facilitates the analysis of uncertainty and variability in the results.

Another concern of a risk assessment is the propagation of epistemic uncertainty (uncertainty) and aleatoric uncertainty (variability) in a risk assessment (Frey (1992); Helton and Johnson (2011); Hoffman and Hammonds (1994); Jakeman et al. (2010)). If the uncertainty in a risk factor is reducible when there is more information about it, then it is an epistemic parameter. The uncertainty of the parameter arises because of a researcher’s lack of knowledge about its true, fixed value. These parameters are similar to sample statistics. For example, if a person takes a larger sample of serving sizes of frozen meat patties, then the person will be able to more precisely pinpoint the population mean serving size of all frozen meat patties. Others denote this epistemic parameter as an uncertainty (Frey and Rhodes (1998); Nauta (2000); Vicari et al. (2007)) or a Type B variable (Hoffman and Hammonds (1994)). The second type of uncertainty, an aleatoric variable, arises when a risk factor is inherently random (Helton and Johnson (2011); Jakeman et al. (2010)). In other words, an aleatoric variable has variation that is not reducible through further sampling. For example, the amount of log-transformed colony forming units (CFUs) of \textit{E. coli} O157:H7 on a frozen ground beef patty changes from patty to patty. It is not a constant, fixed value. Other researchers call this variability (Frey and Rhodes (1998); Nauta (2000); Vicari et al. (2007)) or a Type A variable (Hoffman and Hammonds (1994)).

The two-dimensional Monte Carlo simulation is a standard method that incorporates epistemic and aleatoric uncertainties into a risk assessment (Busschaert et al. (2011); Frey (1992); Frey and Rhodes (1998); Hoffman and Hammonds (1994); Mokhtari and Frey (2005); Nauta (2000); Pouillot and Delignette-Muller (2010); Vicari et al. (2007)). The simulation steps are as follows:

1. Execute steps 2–3 for \(i = 1, 2, \ldots, n_e\). This is the epistemic dimension, also known as the outer loop (Frey (1992)).
2. Randomly generate $\theta_i$ vectors from the function $f(\theta_i)$. $\theta_i$ represents the vector of $p$ epistemic parameters in iteration $i$.

3. Execute steps 4–5 for $j = 1, 2, ..., n_a$. This is the aleatoric dimension, or inner loop.

4. For each $\theta_i$, randomly generate $X_{i,j}$ from the function $g(X_{i,j}|\theta_i)$. $X_{i,j}$ represents the vector of $q$ aleatoric variables in iteration $j$ of $i$.

5. Calculate the risk value, $R_{i,j} = h(X_{i,j}, \theta_i)$.

In the outer loop, vectors of epistemic parameters, $\theta_i$, are generated from the function $f$. $f(\theta_i)$ is a joint-probability density function (pdf) of the $\theta_i$ components. For this paper’s simulation, components of $\theta_i$ are independent of each other, and the joint pdf is a product of pdf’s where an individual density takes one of four forms: an expert-informed distribution such as the triangular or PERT distribution (Vose (2000)), a Bayesian posterior distribution (Gelman et al. (2004)), a parametric or non-parametric bootstrap distribution (Efron and Tibshirani (1993)), or an estimated sampling distribution. Pouillot and Delignette-Muller (2010) describe an application of the first three distribution forms. We discuss the last form in this paper. The function of the inner loop (step 4), $g$, is also a joint pdf of aleatoric variables conditional on some or all of the $\theta_i$ components. In step 5, $h$ represents the risk function and may take a product, linear, or non-linear combination of inputs. These inputs are $X_{i,j}$ and may or may not contain some of the components of $\theta_i$.

Due to the epistemic and aleatoric uncertainties of a two-dimensional risk assessment, we cannot simply present results through a single cumulative distribution function (CDF) or a single point estimate with confidence limits. A traditional Monte Carlo risk assessment has thousands of CDF’s. We summarize them through a plot of a few CDF-based step functions. Outer lines mimic probability bands for the estimate of the true CDF, while the middle line is like a typical, median CDF (Frey (1992); Pouillot and Delignette-Muller (2010)). Figure 3.1 is an example of five CDF-based lines generated by using data from the risk of HUS in children under the age of 5 who were exposed to beef patties contaminated with $E. \text{coli}$ O157:H7. These were calculated using $n_e = 5000$ and $n_a = 1000$. The lines, from top to bottom, are the 0.025, 0.25, 0.50, 0.75, and 0.975 CDF bounds for each $b$-quantile of risk, $b \in [0,1]$. $b$ values are
Figure 3.1: Two-Dimensional Monte Carlo Simulation Results of Risk of HUS in Children Under the Age of 5

probabilities from the vertical axis. Presentation of results generated by using this quantile-quantile method is visually attractive because it is uncluttered (i.e., all 5000 curves are not presented). Furthermore, researchers can easily determine 50% or 95% probability limits of a $b$-quantile of risk. For example, the interpretation of the HUS risk values corresponding to the 0.025 and 0.975 CDF lines at a probability of 0.5 ($b = 0.5$) is: “For a collection of 100 children under 5 years of age who consumed $E. coli$ O157:H7 contaminated frozen beef patties, there is a 95% chance that the median number of children infected with HUS is between 0 and 7.” Table 3.1 provides a brief summary of four other $b$–quantiles (row values). Headings correspond to the five plotted CDF curves in Figure 3.1.

Operationally, to create Figure 3.1, there are two stages of quantile estimation. We outline Frey (1992) and Pouillot and Delignette-Muller’s (2010) procedure. First, there is the $b$-quantile estimation for each of the $n_e$ CDF’s, i.e., quantile estimation of each collection of $R_{i,j} | \theta_i$. This is quantile estimation across the aleatoric dimension of the risk assessment. $b$ in Figure 3.1 takes on values between 0 and 1 in increments of 0.01. Next, for each collection of $n_e$ $b$-quantiles, one estimates the 0.025, 0.25, 0.50, 0.75, and 0.975 quantiles. This is quantile estimation across
Table 3.1: Probability Bounds for the 0.025, 0.25, 0.5, 0.75, and 0.975 Risk Quantiles of Figure 3.1

<table>
<thead>
<tr>
<th>$b-$quantiles</th>
<th>0.025</th>
<th>0.25</th>
<th>0.50</th>
<th>0.75</th>
<th>0.975</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.025</td>
<td>0.000004</td>
<td>0.000019</td>
<td>0.000041</td>
<td>0.000077</td>
<td>0.000234</td>
</tr>
<tr>
<td>0.25</td>
<td>0.000125</td>
<td>0.000598</td>
<td>0.001201</td>
<td>0.002317</td>
<td>0.007141</td>
</tr>
<tr>
<td>0.50</td>
<td>0.001009</td>
<td>0.004958</td>
<td>0.010288</td>
<td>0.020468</td>
<td>0.071279</td>
</tr>
<tr>
<td>0.75</td>
<td>0.010036</td>
<td>0.050971</td>
<td>0.010758</td>
<td>0.208331</td>
<td>0.592765</td>
</tr>
<tr>
<td>0.975</td>
<td>0.390553</td>
<td>0.972040</td>
<td>0.999895</td>
<td>1.000000</td>
<td>1.000000</td>
</tr>
</tbody>
</table>

the epistemic dimension of the risk assessment. The result is the plot of 5 pseudo $CDF$s in Figure 3.1. We add “pseudo” because none of the curves actually pertains to an estimate of a conditional $CDF$.

An important criticism of the two-dimensional Monte Carlo simulation is its computational inefficiency (Frey (1992)). The total number of iterations, $n_e \times n_a$, and the sizes of $\theta_i$ and $X_{i,j}$, $p$ and $q$, respectively, multiplicatively increase computational effort. To reduce this effort, one may argue for small $n_e$ and $n_a$. But this is not recommended because accurate estimates of risk in the extrema (i.e., tail probabilities) require large $n_e$ and $n_a$ (Christophi and Modarres (2005)). Another potentially compounding issue concerns the bootstrap procedure. If a researcher performs either a parametric or non-parametric bootstrap like Pouillot and Delignette-Muller (2010) do under a modified version of our HUS-risk model, then computing time will increase. For either a parametric or a non-parametric bootstrap of a component of $\theta_i$, Pouillot and Delignette-Muller (2010) do not directly simulate values of the epistemic parameter value. Instead, they compute an epistemic parameter value based on a bootstrapped sample of the aleatoric variable that the epistemic parameter summarizes. The aleatoric variable corresponding to the epistemic parameter of interest is bootstrapped, not the epistemic parameter. This bootstrapping approach adds a third loop of Monte Carlo simulations. Furthermore, the estimation of the epistemic parameter from the bootstrapped sample may be time-consuming. As an example, Pouillot and Delignette-Muller (2010) calculate maximum-likelihood estimates (MLE’s) from the bootstrapped aleatoric sample. Maximum-likelihood estimation will increase computing time of a simulation when it must be performed repeatedly.
We propose a new risk-assessment method that reduces the computational intensiveness of the two-dimensional Monte Carlo simulation. In place of the naive \(CDF\) estimator in the inner loop, which requires a large \(n_a\), such as 1000, we estimate the conditional \(CDF\)’s by using the Nadaraya-Watson multivariate kernel regression estimator of uncertain parameter iterates on simulated risk values \(n_a = 25\). Next, we apply statistical inference theory to circumvent Pouillot and Delignette-Muller’s (2010) Monte Carlo simulations of bootstrapped epistemic parameter iterates. In the following sections, we will introduce the problem through the context of a microbial risk assessment about \(E.\ coli\) O157:H7, explain the new approach to handle bootstrapped epistemic parameters, discuss the application of the non-parametric Nadaraya-Watson multivariate kernel regression estimator to a risk assessment model, and conclude with a discussion about the accuracy and increased computational speed of our new \(CDF\) estimator.

### 3.2 Risk of Hemolytic Uremic Syndrome in Children Exposed to \(E.\ coli\) O157:H7

Razzaq (2006) reported that hemolytic uremic syndrome (HUS) is an emerging health concern in humans as it is related to the consumption of \(E.\ coli\) O157:H7 contaminated foods. HUS is most common in young children and may lead to acute renal failure (Liu et al. (2007); PubMed Health (2012); Razzaq (2006)). In 2005, a French outbreak of \(E.\ coli\) O157:H7 in frozen ground beef patties occurred, and ten children under the age of 5 were diagnosed with HUS (Delignette-Muller and Cornu (2008)). Delignette-Muller and Cornu (2008) collected data on the incident and investigated the risk factors associated with HUS in children under 5 years of age. The development of our new risk assessment method will use their data and assume the same single-hit dose-response model. The distributional assumptions of the aleatoric variable and epistemic parameters will mostly mimic those found in Pouillot and Delignette-Muller (2010). We will describe each variable in turn and indicate where we deviate from their model.
3.2.1 Aleatoric Variables in the Dose-Response Model of HUS

The probability that a child under 5 years of age is diagnosed with HUS, \( R \), after being exposed to \( D \) colony forming units (CFUs) of \( E. coli \) O157:H7 per serving is modeled as

\[
R = 1 - (1 - \rho)^D
\]  

(3.1)

where \( \rho \) is the probability that HUS will result after ingestion of 1 CFU of \( E. coli \) O157:H7. The value of \( \rho \) was estimated by Delignette-Muller and Cornu (2008) through a Bayesian framework. They concluded that a minimum value of \( \rho \) was \( 10^{-4} \); a most likely value was \( 1.2 \times 10^{-3} \); and a maximum value was \( 10^{-2} \) (Pouillot and Delignette-Muller (2010)). Note that \( \rho \) is an epistemic parameter because it is a constant specific to the population of children under the age of 5.

On the other hand, \( D \) is an aleatoric variable that changes from child to child. Deviating from Pouillot and Delignette-Muller (2010), we represent \( D \) as a continuous dose taking values between zero and infinity. For our study, the probability of observing a serving size greater than 250 grams is practically zero. The dose portion of the HUS-risk model is

\[
D \sim \text{Truncated } \mathcal{N}(\text{Mean } = \lambda, \text{Variance } = \lambda)
\]  

(3.2)

\[
\lambda = S \times 10^{C-Z}
\]  

(3.3)

\[
Z = \begin{cases} 
\text{Uniform}(0, 0.9) & A = 1 \\
\text{Uniform}(0.2, 1.4) & A = 2 \\
\text{Uniform}(1.2, 2.8) & A = 3 
\end{cases}
\]  

(3.4)

where \( S \) is the serving size of a frozen ground beef patty (in grams, g), \( C \) is the \( \log_{10} \)-concentration of \( E. coli \) O157:H7 on the patty before cooking (\( \log_{10} \text{CFU} \)), and \( Z \) is the decimal reduction of \( E. coli \) O157:H7 (\( \log_{10} \text{(CFU/g)} \)) related to a patty prepared rare (\( A = 1 \)), medium (\( A = 2 \)), or well-done (\( A = 3 \)). For a Poisson-distributed \( D \), Delignette-Muller and Cornu (2008) have verified that the exposure model in equations (3.2) and (3.3) is reasonable based on empirical serving sizes and empirical initial \( \log_{10} \)-concentrations of bacteria from surveys and their experimental data. We model \( D \) with a truncated normal distribution because results presented in later sections require a continuous dose. Also, the truncated normal distribution is a reasonable approximation of the Poisson distribution (Devyatov (1969)). Originally, \( Z \) was
treated as a fixed value for each cooking preference and was estimated from laboratory experiments. In Pouillot and Delignette-Muller (2010), the authors relaxed this assumption because people may use different cooking methods and have different concepts of rare, medium, and well-done. The ranges of the uniform distributions in equation (3.4) are based on Delignette-Muller and Cornu (2008)’s derivations.

The values of $S$, $C$, and $A$ inherently differ between young children as each child will eat a different amount of ground beef with a different initial contamination and a different cooking style. Thus, the variables are aleatoric, and, following Pouillot and Delignette-Muller (2010), we assume the probabilistic structure

$$S \sim \text{Weibull}(\gamma_s, \beta_s)$$

(3.5)

$$C \sim N(\mu_c, \sigma_c^2)$$

(3.6)

$$A \sim \text{Multinomial}(p_1, p_2, 1 - (p_1 + p_2)).$$

(3.7)

For simulation purposes, the variables $S$, $C$, and $A$ are statistically independent. Thus, the components of $\lambda$ in equation (3.3) are also statistically independent, and the distribution of $\lambda$ is a product of pdf’s derived from the distributions described in (3.4), (3.5), (3.6), and (3.7). Thus, $g$, the function in step 4 of the 2-dimensional Monte Carlo simulation described in the introduction, is a numerically integrated product of the truncated normal ($\lambda, \lambda$) pdf and the joint pdf of $S$, $10^C$, and $10^{-Z}$. The purpose of this loop is to numerically integrate out $S$, $10^C$, and $10^{-Z}$ to elicit the distribution of HUS risk, $R$. 
3.2.2 Epistemic Parameter Distributions of the HUS Model

![Diagram of HUS Risk Model](image)

Figure 3.2: Directed Acyclic Graph of HUS Risk Model: Ellipses represent epistemic parameter inputs. Squares represent aleatoric variable inputs. $D \sim \text{truncated normal}(\lambda, \lambda)$, where $\lambda = S \times 10^{C-Z}$ and $R$ is the risk of HUS.

Bearing in mind that a risk assessment should account for aleatoric and epistemic uncertainty (Frey (1992); Helton and Johnson (2011); Hoffman and Hammonds (1994)), we now describe how we account for the uncertainty in the epistemic parameters $\rho$, $p_1$, $p_2$, $\gamma_s$, $\beta_s$, $\mu_c$, and $\sigma^2_c$. A directed acyclic graph (Figure 3.2) describes the relationship between the inputs and risk. The epistemic parameters, denoted by ellipses, are the “outer loop” of Pouillot and Delignette-Muller’s (2010) HUS-risk model. The aleatoric variables, denoted by squares, are the “inner loop” of the HUS-risk model. Notationally, hats on epistemic parameters will denote random realizations of the corresponding parameter’s distribution.

In probabilistic risk assessments, a Program Evaluation and Review Technique (PERT) distribution, a posterior distribution, an asymptotic MLE distribution, or a normal-theory
based distribution may represent epistemic parameter uncertainty. As described in section 3.2.1, the parameter \( \rho \) is an expert-informed epistemic parameter. Direct measures of \( \rho \) do not exist. Consequently, we use a PERT distribution (Vose (2000)) to describe \( \rho \)’s uncertainty.

\[
\hat{\rho} \sim \text{PERT}(\text{min} = 10^{-4}, \text{mode} = 1.2 \times 10^{-3}, \text{max} = 10^{-2})
\]  

(3.8)

A PERT distribution is a unimodal beta distribution with a modified support of \((\text{min}, \text{max})\). The shape parameters are

\[
\alpha_1 = \frac{4(\text{mode}) + \text{max} - 5(\text{min})}{\text{max} - \text{min}}
\]

(3.9)

\[
\alpha_2 = \frac{5(\text{max}) - \text{min} - 4(\text{mode})}{\text{max} - \text{min}}.
\]

(3.10)

The parameter values of (3.8), (3.9), and (3.10) were extrapolated from a Bayesian dose-response model (Pouillot and Delignette-Muller (2010)).

A Bayesian posterior distribution represents the uncertainty of cooking preference epistemic probabilities, \((\hat{p}_1, \hat{p}_2, 1 - \hat{p}_1 - \hat{p}_2)\). In the Bayesian context, the posterior distribution of an epistemic parameter integrates prior knowledge of the epistemic parameter and pertinent data. Pouillot and Delignette-Muller (2010) recommend using a flat, uninformative conjugate prior. For categorical data, such as cooking preferences (rare, medium, and well-done), this is the Dirichlet distribution with shape parameters \( \alpha_1 = 1, \alpha_2 = 1, \) and \( \alpha_3 = 1 \). According to Delignette-Muller and Cornu’s (2008) survey of 144 households, the number of French children who ate frozen ground-beef patties rare, medium, and well-done is 14, 59, and 71, respectively. Based on this multinomial data of counts, the posterior distribution of cooking preference probabilities is

\[
\begin{bmatrix}
\hat{p}_1 \\
\hat{p}_2 \\
1 - (\hat{p}_1 + \hat{p}_2)
\end{bmatrix} \sim \text{Dirichlet}(15, 60, 72).
\]

(3.11)

See Gelman et al. (2004) for details on Bayesian inference.

For the simulation of vectors \((\hat{\gamma}_s, \hat{\beta}_s)\) and \((\hat{\mu}_c, \hat{\sigma}^2_c)\), Pouillot and Delignette-Muller (2010) proposed bootstrapping. Specifically, they used a parametric bootstrap for \((\hat{\gamma}_s, \hat{\beta}_s)\) values and a non-parametric bootstrap for \((\hat{\mu}_c, \hat{\sigma}^2_c)\) values. We note that their bootstrapping methods, both non-parametric and parametric, are computationally costly in two ways. First, they
simulate aleatoric variable values from either the observed data or MLE-fitted distribution. This introduces another loop of Monte Carlo simulations. Secondly, from the bootstrapped aleatoric values, they derive MLE’s of the epistemic parameters. This estimation procedure is time-consuming because an analyst needs to perform it \( n_e \) times for a maximum of \( p \) epistemic parameters.

To overcome these limitations, we can use statistical theory about MLE’s and normal theory to expedite Pouillot and Delignette-Muller’s (2010) bootstrapping procedures. In other words, we can directly simulate epistemic parameter values from either an MLE’s limiting distribution, a normal distribution, or a gamma distribution. The next section details our accelerated sampling scheme.

### 3.3 New Approaches to Model Epistemic Parameter Uncertainty

#### 3.3.1 Asymptotic MLE Distribution

Since we modeled aleatoric variables with probability density functions (pdf’s), we can exploit results from statistical inference and directly simulate epistemic parameters from pdf’s. This means that we do not have to carry out a time-consuming simulation procedure like Pouillot and Delignette-Muller’s (2010) parametric bootstrap of the serving size variable, \( S \).

If a researcher uses maximum likelihood estimation of an aleatoric variable’s distributional parameters, then we may simulate the epistemic parameter(s) from an asymptotic pdf of the MLE(s), provided the asymptotic pdf is close to the true sampling distribution of the MLE’s.

If we let \( X_i \overset{iid}{\sim} f(x|\theta) \) for \( i = 1, 2, ..., n \) and \( \tilde{\theta} \) be the MLE of \( \theta \), under certain regularity conditions of \( f(x|\theta) \),

\[
\tilde{\theta} \overset{d}{\sim} \mathcal{MVN}(\theta, \mathcal{I}^{-1}(\theta))
\]  

(3.12)

where \( \mathcal{I}(\theta) \) is the expected information matrix. Consequently, for a sufficiently large sample of data pertaining to an aleatoric variable, the distribution of the MLE(s) is approximately [multivariate] normal (\( \mathcal{MVN} \)) with the mean vector equal to the true epistemic parameter values, \( \theta \), and the variance-covariance matrix equal to the inverse of the expected information matrix. Since the variance of the asymptotic distribution in (3.12) depends on \( \theta \), we must
estimate it from either the expected or the observed information (Efron and Hinkley (1978)).

When using expected information, the derivation of $\mathbf{I}(\theta)$ in (3.12) may be rather difficult as it often involves expectations of nonlinear data functions. For example, the expected information matrix of the epistemic parameters of the serving size variable (3.5) involves $E(S^{\gamma_s})$, $E((\ln(S))S^{\gamma_s})$, and $E((\ln(S))^2S^{\gamma_s})$. Although closed forms exist for these expectations, this is not always the case. Stated differently, an analytic solution may not always exist for $\mathbf{I}^{-1}(\theta)$. Researchers often circumvent this issue and use observed information (Efron and Hinkley (1978); Greene (2003); Pawitan (2001)).

Observed information is defined as

$$I(\tilde{\theta}) = -\left. \frac{\delta^2 \ln L(\theta)}{\delta \theta \delta \theta'} \right|_{\theta = \tilde{\theta}}$$

where $L(\theta)$ is the likelihood function. The expectation of (3.13) leads to expected information. Efron and Hinkley (1978) have compared $I(\tilde{\theta})^{-1}$ and $\mathbf{I}(\tilde{\theta})^{-1}$ variance estimators in a univariate MLE setting. They concluded that observed information was the better estimator of variance. Others recommend using it in the estimation of the variance-covariance matrix of multivariate MLEs (Greene (2003); Pawitan (2001)).

In our study, we represent the distribution of $(\hat{\gamma}_s, \hat{\beta}_s)$ through the limiting multivariate normal distribution in (3.12) by using observed information. To check the validity of this representation, we compare the asymptotic distribution to a non-parametrically bootstrapped distribution of the parameters. We simulate 1000 bootstrapped samples and estimate $\gamma_s$ and $\beta_s$ via maximum likelihood. Figure 3.3 shows the results. Gray dots are the bootstrapped $(\hat{\gamma}_s, \hat{\beta}_s)$ estimates. Black lines are the contours of the estimated limiting distribution of $(\hat{\gamma}_s, \hat{\beta}_s)$. We see that the density of points matches the contours rather well and conclude that the asymptotic distribution is close to the sampling distribution of $(\hat{\gamma}_s, \hat{\beta}_s)$. The asymptotic multivariate normal distribution is a reasonable sampling distribution for the serving size’s epistemic parameters.
3.3.2 Normal Theory Distributions

The other bootstrapped aleatoric variable in Pouillot and Delignette-Muller’s (2010) risk assessment is the initial log\(_{10}\)-concentration of \textit{E. coli} O157:H7 on a frozen beef patty, \(C\). They performed a non-parametric bootstrap to estimate \(\mu_c\) and \(\sigma^2_c\). Furthermore, they modeled \(C\) through a normal distribution. By assuming normality of \(C\), we may directly sample estimates of \(\mu_c\) and \(\sigma^2_c\) from normal-theory based sampling distributions. We need not only rely on results from the previous section.

For \(X_i \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)\) and \(i = 1, 2, ..., n\),

\[
\bar{X} \sim \mathcal{N} \left( \mu, \frac{\sigma^2}{n} \right) \tag{3.14}
\]

and

\[
S^2 \sim \Gamma \left( \frac{n-1}{2}, \frac{2\sigma^2}{n-1} \right) \tag{3.15}
\]

where \(\bar{X} = \sum_{i=1}^{n} X_i / n\), \(n\) is the sample size, \(S^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2 / (n-1)\), and \(\Gamma\) is the gamma distribution as parameterized by Casella and Berger (2002). As in (3.12), the true epistemic
parameters, $\mu$ and $\sigma^2$ in this case, are unknown. We replace their values in the distributions of (3.14) and (3.15) with the empirical MLEs of $\mu$ and $\sigma^2$, $\hat{\mu}$ and $\hat{\sigma}^2$. Note that $\sigma^2$ in the distributions of (3.14) and (3.15) is not the unbiased variance estimator.

The aleatoric variable of log$_{10}$-concentration of initial bacterial, $C$, follows a normal distribution (3.6). Using (3.14) and (3.15), it follows that

$$
\hat{\mu}_c \sim \mathcal{N}\left(\hat{\mu}_c, \frac{\hat{\sigma}^2}{n_c}\right)
$$

(3.16)

and

$$
\hat{\sigma}^2_c \sim \Gamma\left(\frac{n_c - 1}{2}, \frac{2\hat{\sigma}^2}{n_c - 1}\right)
$$

(3.17)

where $n_c$ is the sample size of $C$ ($n_c = 18$ in our sample).

![Figure 3.4: Comparison of Non-Parametrically Bootstrapped (bars) and Normal-Theory Based Distributions (solid lines) of Initial Bacterial Parameters: Non-parametrically bootstrapped data generated by Pouillot and Delignette-Muller (2010).](image)

Since we deviate from Pouillot and Delignette-Muller’s (2010) risk model, we compare our distributions of $\hat{\mu}_c$ and $\hat{\sigma}^2_c$ to their non-parametrically bootstrapped distributions, respectively. We use 1000 bootstrap samples. Figure 3.4 shows that the normal-theory based distributions (solid smooth lines) match rather well to the non-parametric bootstrap distribution (bars of
We conclude that our epistemic distributions of $\mu_c$ and $\sigma_c$ closely resemble the non-parametrically bootstrapped distributions of the respective epistemic parameters. (3.14) and (3.15) are easily extended for multivariate estimators. For readers interested in the multivariate extension, Johnson and Wichern (2002) describe the sampling distribution of the sample mean vector ($\mathcal{MVN}$ distribution) and the sample variance-covariance matrix (Wishart distribution).

### 3.4 Non-Parametric Estimation of Risk CDF

As discussed in the introduction, results from an analysis incorporating aleatoric and epistemic uncertainties involve estimated CDF’s of risk conditional on a set of epistemic parameter values, $\theta_i$. We call this conditional CDF, $F_i$. One may estimate $F_i$ through a naive estimator such as the average of $n_a$ indicator functions. $F_i$’s generated from this approach are statistically unbiased but potentially imprecise if $n_a$ is not large enough. Hence, Frey (1992) recommends using a large $n_a$ for two-dimensional Monte Carlo simulation. This multiplicatively increases the computing time as $n_a$ increases. Our objective is to reduce this time but still have an accurate and precise estimator of $F_i$. To achieve our goal, we reduce $n_a$ down to 25 and use a Nadaraya-Watson estimator of $F_i(r)$ for $r \in [0, 1]$. At the median, the variance of the naive CDF estimator for $n_a = 25$ is 25 times smaller than that of the naive estimator with $n_a = 1$. At any risk value, we expect the naive CDF estimator to be within 0.2 units of the true $F_i$. The Nadaraya-Watson estimator of $F_i$ is a weighted average of CDF’s in a neighborhood around $\theta_i$ and significantly reduces the variance of the naive estimator.

#### 3.4.1 Nadaraya-Watson Estimator Applied to CDF Estimation

For some $\hat{\theta}_i$ and $n_a = 25$, the traditional $F_i$ estimation method would be an imprecise estimator of $F_i(r)$. Only 25 probability values are possible with so small an $n_a$. In order to hold $n_a = 25$ and obtain a more precise estimator of $F_i(r)$ beyond a crude 25-jump estimator, we propose incorporating the risk values associated with $\hat{\theta}_j$s that are in the neighborhood of $\hat{\theta}_i$ to estimate $F_i(r)$. 
The intuition behind the neighborhood idea is best illustrated through an example. To permit a clearer presentation of our \( F_i \) estimation procedure, first, we simplify the HUS-risk model described in (3.1), (3.2), and (3.3). We treat \( D, S, \gamma_s, \) and \( \beta_s \) as the only uncertain model inputs and fix the values of \( C, Z, \) and \( \rho \) to 0.232, 0.8, and 0.0012, respectively. For \( F_i \), we choose \( \hat{\theta}_i = (\hat{\gamma}_s, \hat{\beta}_s) = (2.19, 83.35) \), the actual MLE vector of the serving data. \( n_e \) is set to 1000, and \( n_a \) is 25. Estimation results are presented in figure 3.5.

Our estimation procedure works because it capitalizes on the similarities of risk values associated with epistemic parameter vectors that are near \( \hat{\theta}_i \). In figure 3.5(a), we plot the 1000 simulated epistemic parameter pairs on a standardized scale. Silverman (1998) and Hārdle et al. (2004) suggest standardizing. Mathematical justifications for standardization will be discussed later. In the center of the plot is the light-colored standardized epistemic point of interest, \( \hat{\theta}_i \). We deem a standardized \( \hat{\theta}_j \) close to \( \hat{\theta}_i \) as long as it lies within the light-colored, 0.87-unit circle (i.e. a Euclidean difference less than or equal to 0.87). The radius value of 0.87 "balances" the bias and precision of our CDF estimator. Details follow in section 3.4.3. We contend that the risk values associated with \( \hat{\theta}_j \)s are much like the realizations of \( F_i \), because the \( \hat{\theta}_j \)s are near the \( \hat{\theta}_i \).

The residual plot in Figure 3.5(c) demonstrates the effectiveness of the neighborhood-based CDF estimator versus the traditional CDF estimator (\( n_a = 1000 \)). Through numerical integration, we have access to the true \( F_i \) at any value of risk. Given a CDF estimator’s simulated risk values and corresponding probabilities, we calculate the absolute differences between the probabilities of the estimator and the true \( F_i \) at each simulated risk value. A good estimator has absolute residuals near zero for all risk values. From the Figure 3.5, we see that our estimator performs quite well. In fact, it fits better than the straight-forward Monte Carlo simulation of the CDF where \( n_a = 1000 \). 88.5% of the traditional CDF estimator’s residuals (black-colored line) are larger than those of the neighborhood-based CDF estimator (light-colored line). Conditioning on the epistemic parameters and integrating the mean squared error (\( IMSE \)) over the risk values in the plot, the root \( IMSE \) of the traditional estimator is 0.00103 while our estimator has a value of 0.000967. Even though the traditional estimator is unbiased, our estimator is still better. The squared bias of our estimator is \( 3.37 \times 10^{-9} \), very
small. The variance of the traditional estimator is 1.13 times larger than that of our estimator.

Unlike the traditional CDF estimator for some $F_i$, our estimator shares risk values associated with epistemic parameter vectors near the target, $\hat{\theta}_i$. There are two-fold benefits from this. First, the precision of our estimator tends to be better as the previous example demonstrates. Secondly, our estimation procedure is faster. In a full risk assessment using the simplified model, we would generate $1000 \times 25 = 25000$ risk values in the two-dimensional Monte Carlo. With the traditional estimation procedure, one would generate $1000 \times 1000 = 1000000$ risk values. Our two-dimensional Monte Carlo is much less computationally intensive.

Like the traditional risk estimation procedure, our neighborhood estimator of the CDF does not require the specification of a parametric relationship between $\hat{\theta}_j$ and risk. It is a non-parametric model for the collection of ordered pairs $\left\{ \left( s_{\hat{\theta}_j}, \sum_{l=1}^{n_a} I\left\{ R_{j,l}/n_a \leq r \right\} \right) \right\}_{j=1}^{n_e}$ where $s_{\hat{\theta}_j}$ is the standardized value of the $p-$dimensional parameter vector $\hat{\theta}_j$ with components $(\hat{\theta}_{j,1}, \hat{\theta}_{j,2}, \ldots, \hat{\theta}_{j,p})$. The general form of our estimator is

$$\hat{m}_h(r, s_{\hat{\theta}_i}) = n^{-1} \frac{\sum_{j=1}^{n_e} K_h(s_{\hat{\theta}_i} - s_{\hat{\theta}_j}) \sum_{l=1}^{n_a} I\{R_{j,l} \leq r\}/n_a}{\sum_{j=1}^{n_e} K_h(s_{\hat{\theta}_i} - s_{\hat{\theta}_j})}$$

(3.18)

where $K_h(s_{\hat{\theta}_i} - s_{\hat{\theta}_j}) = \frac{1}{h}K\left( \frac{s_{\hat{\theta}_i,1} - s_{\hat{\theta}_j,1}}{h}, \frac{s_{\hat{\theta}_i,2} - s_{\hat{\theta}_j,2}}{h}, \ldots, \frac{s_{\hat{\theta}_i,p} - s_{\hat{\theta}_j,p}}{h} \right)$ is some multivariate kernel function with a bandwidth $h$ for all standardized epistemic parameter inputs (Härdle (1991)); $p$ is the number of epistemic parameters; and $\sum_{l=1}^{n_a} I\{R_{j,l} \leq r\}/n_a$ is the crude estimator of the CDF, $F_j$, at some risk value, $r$ (early parts of this subsection refer to this as the unbiased, naive estimator of $F_j(r)$). We will reference this as $\hat{F}_j(r)$. Equation (3.18) is the Nadaraya-Watson estimator (Nadaraya (1964); Watson (1964)), and, in general terms, it can be viewed as a weighted average of the crude CDF estimators whose standardized epistemic vectors lie within a hypersphere with radius equal to $h$. For our simple CDF estimation in Figure 3.5, $K_h(\bullet) = \frac{1}{0.87^2 \pi} I\left\{ \sqrt{s_{\hat{\gamma}_s}^2 + s_{\hat{\beta}_s}^2} \leq 0.87 \right\}$, a bivariate uniform kernel (Härdle (1990)). With this kernel, we see that $\hat{m}_h(r, s_{\hat{\theta}_i})$ reduces to the proportion of risk values in the neighborhood that are less than $r$.

In (3.18), we point out two important characteristics of the kernel function. It is radial-symmetric where the bandwidths are not of a general matrix form, $H$ (Härdle (1991); Härdle
Figure 3.5: CDF Estimators: (a) Joint Distribution of Standardized Serving Epistemic Parameters. (b) Histogram of Risk Values Corresponding to Epistemic Parameter Vectors Located in the Neighborhood. (c) Comparison of Residuals for Neighborhood vs. Traditional CDF Estimators.
et al. (2004); Wand and Jones (1995)). A radial-symmetric kernel permits correlations between predictors. Consequently, we can account for epistemic parameter estimates, like the probabilities of cooking preference, in (3.11) of our model. The bandwidth matrix is simple, $H = hI$, meaning we can use the same bandwidth for all epistemic parameters. Thus, we need only worry about selecting one bandwidth. Bandwidth selection is an important topic and will be discussed in chapter 4.

“Pre-whitening”/standardizing the predictor variables, $\hat{\theta}_j$, permits a simple bandwidth matrix (Silverman (1998)). To generate the results of this paper, we standardize $\hat{\theta}_j$. For multivariate vectors, such as $(\hat{\gamma}_s, \hat{\beta}_s)$, we subtract the mean vector and apply a Mahalanobis transformation. This standardization is

$$s\hat{\theta}_j = \Sigma^{-1/2}(\hat{\theta}_j - \mu)$$

(3.19)

where $\Sigma^{-1/2}$ is the inverse of the Cholesky decomposition of the variance-covariance matrix of $\hat{\theta}_j$, and $\mu$ is the mean vector of $\hat{\theta}_j$. Univariate epistemic parameters, such as $\hat{\rho}$ and $\hat{\mu}_c$, are Wald transformed. A Wald transformation is

$$s\hat{\theta}_j = \frac{\hat{\theta} - \mu}{\sigma}$$

(3.20)

where $\mu$ is the mean of the epistemic parameter, $\hat{\theta}_j$, and $\sigma$ is its standard deviation. As a reminder, we represent standardized epistemic parameters with a preceding subscript $s$.

Equation (3.18) does not define a specific kernel function. On the basis of mean squared error ($MSE$), the choice of kernel is relatively unimportant when estimating a univariate density (Härdle (1990, 1991); Härdle et al. (2004)). In fact, Härdle (1990) recommends that one choose a computationally efficient kernel. Consequently, we consider the uniform kernel. Nonetheless, for a non-negative kernel-based estimator of a density, the Epanechnikov kernel is the optimal kernel under mean integrated squared error ($MISE$) (Epanechnikov (1969); Sacks and Ylvisaker (1981)). In the following section, we examine the behavior of the Nadaraya-Watson estimator in (3.18) with respect to the uniform and Epanechnikov kernels.
3.4.2 Choice of Kernel: Epanechnikov Versus Uniform

We evaluate the effectiveness of the Epanechnikov-based estimator and the uniform-based estimator through the minimization of a performance criterion. As outlined in the next few paragraphs, we find that the conditional mean of averaged squared error integrated over risk values, \( IMASE \), is a reasonable performance criterion for our estimators. It accommodates our estimator’s epistemic \( \hat{s}\theta_j \) inputs; its aleatoric \( \hat{F}_j(r) \) inputs; and its user-specified \( r \) inputs. To understand the importance of these considerations, we detail the estimators’ forms.

Kernel regression estimators work because they pool information on responses whose predictors are in a “small” neighborhood. For the ordered pairs, \( \{(\hat{s}\theta_j, \hat{F}_j(r))\}_{j=1}^{ne} \), response weights may vary according to the proximity of \( \hat{s}\theta_j \) to some target \( \hat{s}\theta_i \). For example, the Epanechnikov kernel, \( K^{epan}(\cdot) \), allocates more weight to an indicator if its epistemic parameter vector is nearer the target epistemic parameter (Epanechnikov (1969)). In contrast, the uniform kernel, \( K^{unif}(\cdot) \), assigns the same weight to each indicator whose \( \hat{s}\theta_j \) is in the neighborhood (or hypersphere) around the target \( \hat{s}\theta_i \) (Härdle (1990)). The kernel forms are

\[
K^{epan}(\hat{s}\theta_i - \hat{s}\theta_j) = \frac{1}{hp} D_1 (1 - u_{i,j}) I\{u_{i,j} \leq 1\}, \tag{3.21}
\]

\[
K^{unif}(\hat{s}\theta_i - \hat{s}\theta_j) = \frac{1}{hp} D_2 I\{u_{i,j} \leq 1\}, \tag{3.22}
\]

where \( u_{i,j} = \sum_{k=1}^{p} (\hat{s}\theta_{i,k} - \hat{s}\theta_{j,k})^2 / h^2 \) (squared Euclidean distance), and \( D_1 \) and \( D_2 \) are normalizing constants of the kernel. The constants ensure that the kernels integrate to unity. Substituting these kernels into the Nadaraya-Watson estimator of (3.18) gives

\[
\hat{m}^{epan}_h (r, \hat{s}\theta_i) = \sum_{j=1}^{ne} \left[ \frac{(1 - u_{i,j}) I\{u_{i,j} \leq 1\}}{\sum_{j=1}^{ne} (1 - u_{i,j}) I\{u_{i,j} \leq 1\}} \right] \hat{F}_j (r) = \sum_{j=1}^{ne} w^{epan}_{i,j} (h) \hat{F}_j (r), \tag{3.23}
\]

\[
\hat{m}^{unif}_h (r, \hat{s}\theta_i) = \sum_{j=1}^{ne} \left[ I\{u_{i,j} \leq 1\} \right] \hat{F}_j (r) = \sum_{j=1}^{ne} w^{unif}_{i,j} (h) \hat{F}_j (r), \tag{3.24}
\]

where \( \hat{m}^{epan}_h (r, \hat{s}\theta_i) \) is the Epanechnikov-based CDF estimator, and \( \hat{m}^{unif}_h (r, \hat{s}\theta_i) \) is the uniform-based CDF estimator. The quantities, \( w^{epan}_{i,j} (h) \) and \( w^{unif}_{i,j} (h) \), correspond to the square-bracketed terms in equations (3.23) and (3.24), respectively. They are functions of the epistemic parameters and bandwidth, \( h \).
From equations (3.23) and (3.24), we observe that both estimators are a function of epistemic, aleatoric, and user-specified quantities. Respectively, these are $u_{i,j}$, $\hat{F}_j(r)$, and $h$ and $r$ (and $h$ and $r$ are blocked together). Some criteria that account for the epistemic and aleatoric components of the estimators include average squared error ($ASE$), integrated squared error ($ISE$), and conditional mean averaged squared error ($MASE$) (Härdle et al. (2004)). Härdle (1990) has shown that all three criteria asymptotically have the same minimum.

We compare $\hat{m}_{epan}^{h}(r, s, \hat{\theta}_i)$ to $\hat{m}_{unif}^{h}(r, s, \hat{\theta}_i)$ using $MASE$:

$$d_C(h) = E_R[n_e^{-1} \sum_{i=1}^{n_e} (\hat{m}_h(r, s, \hat{\theta}_i) - m_h(r, s, \hat{\theta}_i))^2].$$

(3.25)

We prefer this criterion because $\hat{F}_i(r) = \sum_{l=1}^{n_e} I \{R_{j,l}/n_a \leq r\}$, the conditional expectation of our response, is analytically tractable; it is the actual CDF of $r$ at $s, \hat{\theta}_i$. Since we estimate conditional CDFs, $d_C(h)$ is more interpretable as it is in terms of the true conditional CDFs.

We can also examine the estimator's bias (3.27) and variance (3.28). From this point on, we will suppress the conditioning to simplify notation.

$$d_C(h) = n_e^{-1} \sum_{i=1}^{n_e} Bias^2(\hat{m}_h(r, s, \hat{\theta}_i)) + Var(\hat{m}_h(r, s, \hat{\theta}_i))$$

(3.26)

where

$$Bias^2(\hat{m}_h(r, s, \hat{\theta}_i)) = \left( E_{R|s, \theta_j}[\hat{m}_h(r, s, \hat{\theta}_i)] - F(r|s, \hat{\theta}_i) \right)^2$$

$$= \left( \sum_{j=1}^{n_e} w_{i,j}(h)F(r|s, \hat{\theta}_j) - F(r|s, \hat{\theta}_i) \right)^2$$

(3.27)

and

$$Var(\hat{m}_h(r, s, \hat{\theta}_i)) = Var_{R|s, \theta_j}[\hat{m}_h(r, s, \hat{\theta}_i)]$$

$$= \sum_{j=1}^{n_e} (w_{i,j}(h))^2 \left\{ F(r|s, \hat{\theta}_j)(1 - F(r|s, \hat{\theta}_j)) \right\}. $$

(3.28)
Before we can quantify $d_{C}(h)$, we recognize that there are two unresolved issues. First, equation (3.26) involves $F(r|s_{\hat{\theta}})$, an unknown quantity. Second, it does not reflect the total variance and bias associated with our estimator because we evaluate $d_{C}(h)$ at a single value of risk, $r$. To address the first issue, we estimate $F(r|s_{\hat{\theta}})$ through a large number of Monte Carlo simulations, $n_{a} = 10000$. As noted in the introduction of section 3.4, this traditional
The estimator is unbiased and precise for a large enough $n_a$. The bias and variance contributions of the Monte Carlo estimates of the $F(r|\hat{\theta}_j)$’s to the quantities in (3.27) and (3.28) is negligible. The conditional CDF estimator is a good estimator. The solution to the second issue is to numerically integrate $d_C(h)$ over the values of $r$, $r \in [0, 1]$, to account for the $d_C(h)$ values at all values of risk.

Figure 3.6 illustrates how the $d_C(h)$, squared bias, and variance change across values of $h$ for the Epanechnikov-based (solid line) and uniform-based (dashed line) kernel regression estimators of risk in the HUS model. The uniform-based estimator uniformly minimizes the variance. Paraphrased, it has the minimum variance at any value of $h$. As expected, the Epanechnikov-based estimator uniformly minimizes the squared bias. When combining the squared bias and variance, the Epanechnikov-based estimator has the smallest $d_C(h)$ value, 0.0013, at $h = 2.55$. The minimum $d_C(h)$ value of the uniform-based estimator is 0.0002 units higher than that of the Epanechnikov-based estimator which is located at $h = 2.3$. This suggests that the choice between these kernels is relatively unimportant. Following Härdle’s (1990) recommendation, we implement the uniform-based Nadaraya-Watson CDF estimator because it is computationally more efficient.

### 3.4.3 Comparison of Standard Two-Dimensional Monte Carlo CDFs and Uniform-Based Non-Parametric CDFs

By using the HUS risk model described in subsections 3.2.1, 3.2.2, 3.3.1, and 3.3.2, we compare risk-assessment results between the traditional two-dimensional Monte Carlo simulation and our modified two-dimensional Monte Carlo simulation with kernel-based Nadaraya-Watson estimators of conditional CDF’s. Similar to Figure 3.1, we present our risk-assessment results graphically with a quantile-quantile presentation of 5 CDF-based curves. We linearly interpolate simulated risk values to estimate quantiles.
Figure 3.7 demonstrates the procedure to interpolate the $b$-quantile, $F^{-1}(b | \hat{\theta}_j)$. For every $b = F_i(r)$, we first find the sandwiching CDF values $F_i(r_{j-1})$ and $F_i(r_j)$ where $r_j$ are the ranked risk values associated with the epistemic parameter vector $\hat{\theta}_j$. Then, we use the slope of the line connecting these two CDF points, denoted by $m$, to interpolate the $r$ value between $r_{j-1}$ and $r_j$. $r$ is the interpolated $b$-quantile of the CDF associated with $\hat{\theta}_j$. R’s quantile function has an option for linear interpolation, type = 4 (R Development Core Team (2012)). We use it for all quantile estimation steps of the risk assessment.

A last consideration before running our risk assessment is the choice of our CDF estimator’s bandwidth. Recall in section 3.4.1 that we used a bandwidth of 0.87. This was not arbitrarily chosen. In fact, this value minimizes the asymptotic mean squared error (AMSE) criterion corresponding to the multivariate Nadaraya-Watson regression estimator for this problem (Härdle
The AMSE in the context of our problem is
\[
\mu_2^2(K^{\text{unif}})h^4 \left[ \frac{\nabla_m^T(s \theta_i) \nabla f(s \theta_i)}{f_{s \theta}(s \theta_i)} + \frac{1}{2} \text{tr} \{ \mathcal{H}_m(s \theta_i) \} \right]^2 + \frac{1}{n_a h^p} \left\| K^{\text{unif}} \right\|_2^2 \frac{F(r|s \theta_i) - F^2(r|s \theta_i)}{n_a f_{s \theta}(s \theta_i)}
\]
(3.29)
where \( \mu_2^2(K^{\text{unif}}) = \left( \int s \theta^2 K^{\text{unif}}(s \theta) ds \theta \right)^2 \) and \( |K^{\text{unif}}|_2^2 = \int (K^{\text{unif}}(s \theta))^2 ds \theta \), and \( f_{s \theta}(s \theta_i) \) is the joint pdf of standardized epistemic parameters evaluated at \( s \theta_i \). \( \nabla f(s \theta_i) \) is the gradient of the joint pdf of standardized epistemic parameters at \( s \theta_i \). \( \nabla_m(s \theta_i) \) and \( \mathcal{H}_m(s \theta_i) \) are, respectively, the gradient and Hessian corresponding to the true CDF of risk at standardized epistemic parameters, \( s \theta_i \).

The value of \( h \) that minimizes AMSE in (3.29) is
\[
h_{\text{opt},i}(r) = n_e^{-1/(p+4)} \left( \frac{p |K^{\text{unif}}|_2^2 (F(r|s \theta_i) - F^2(r|s \theta_i))}{n_a f_{s \theta}(s \theta_i)} \right)^{1/(p+4)} \times \left( \frac{f_{s \theta}^2(s \theta_i)}{\mu_2^2(K^{\text{unif}}) [2 \nabla_m^T(s \theta_i) \nabla f(s \theta_i) + f_{s \theta}(s \theta_i) \text{tr} \{ \mathcal{H}_m(s \theta_i) \}]_2^2} \right)^{1/(p+4)}.
\]
(3.30)
Unlike the bandwidths in the conditional IMASE criterion of section 3.4.2, \( h_{\text{opt},i}(r) \) are specific to \( s \theta_i \) and \( r \). Integration of the right-hand side of equation (3.30) with respect to these inputs would generalize results. However, analytical integration is difficult because the form of (3.30) is complex, and numerical integration is slow. Consequently, we randomly sample five of the observed \( s \theta_i \)'s and average \( h_{\text{opt},i}(r) \)s. This accounts for some of the variation in the \( s \theta_i \)s and is quick. We do not integrate out \( r \) but select an \( F_i \)-specific value of risk. This is the estimated median corresponding to \( \hat{F}(\bullet|s \theta_i) \). We select this \( r \) because the variance of a CDF estimator is largest at the true median.

Equation (3.30) also requires other distribution-based quantities, such as the gradient and Hessian of \( F_i \) and \( f_{s \theta}(s \theta_i) \). The values of the joint pdf of epistemic parameters and the corresponding gradient are known (see equations (3.8), (3.11), (3.12), (3.16), and (3.17)). We estimate \( F(r|s \theta_i) \) through a large number of Monte Carlo simulations, \( n_a = 6000 \), and numerically approximate the gradient and Hessian of \( F(r|s \theta_i) \). For our simple approximation of the gradient and Hessian, we estimate 14 additional CDF’s in the neighborhood of a \( F(r|s \theta_i) \) (i.e., we separately ± a small step for each component of \( s \theta_i \) and estimate the corresponding CDF). We have chosen \( n_a = 6000 \) and five random \( s \theta_i \)'s to keep the number of additional
Monte Carlo simulations in the aleatoric direction less than 500000. In total, our bandwidth selection procedure requires 450000 ($5 \times 15 \times 6000$) evaluations of risk.

![Comparison of CDF Estimators](image)

**Figure 3.8: Comparison of New vs. Old Estimator**

The remaining elements of (3.30) are the kernel, $n_a$, and $n_e$. These inputs are not arbitrarily decided. Choice of the multivariate uniform kernel, $K^{unif}$, was discussed in section 3.4.2. In the HUS-risk model, $\|K^{unif}\|^2_2$ and $\mu_2^2(K^{unif})$ have values of $105/(16\pi^3)$ and $1/81$, respectively, for $p = 7$. Our choice of $n_a$ was discussed in the introduction of section 3.4. The choice of $n_e$ was determined from exploratory analyses. Silverman (1998) discussed the relation of $n_e$ to the accuracy of multivariate kernel density estimators and used $n_e = 10700$. With $n_e = 10700$, the five CDF-based curves in Figure 3.8 (which are generated from $n_e = 5000$ epistemic iterates) change very little. Furthermore, we prefer and apply $n_e = 5000$ herein because computations are four times faster than those based on $n_e = 10700$.

Figure 3.8 and Table 3.2 show that our results from our modified two-dimensional Monte Carlo risk assessment are similar to those of a traditional two-dimensional Monte Carlo risk assessment. In Figure 3.8, our risk assessment curves ($h_{opt} = 1.28$) are in black, and the
Table 3.2: Average Squared Error (ASE) of Pairwise Comparisons Between Traditional and Non-Parametric CDF Curves

<table>
<thead>
<tr>
<th>CDF Curve:</th>
<th>0.025</th>
<th>0.25</th>
<th>0.50</th>
<th>0.75</th>
<th>0.975</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASE:</td>
<td>$5.60 \times 10^{-5}$</td>
<td>$1.37 \times 10^{-5}$</td>
<td>$4.06 \times 10^{-5}$</td>
<td>$6.37 \times 10^{-5}$</td>
<td>$2.91 \times 10^{-5}$</td>
</tr>
</tbody>
</table>

3.5 Discussion

We have shown three ways that a risk assessor may expedite a two-dimensional Monte Carlo simulation of risk. For a sufficient amount of data on an aleatoric variable such as serving size, a researcher can simulate the corresponding epistemic parameters’ MLEs directly from their limiting multivariate normal distribution. We demonstrate that a distribution of MLE’s of serving size epistemic parameters obtained from 1000 non-parametric bootstrapped samples resembles the limiting distribution of the MLE’s. For an aleatoric variable modeled with a normal distribution such as initial log_{10}-bacterial concentration, a researcher can sample the mean and variance epistemic parameters from normal-theory based sampling distributions of $\bar{X}$ and $S^2$. Distributions of $\bar{X}$ and $S^2$ derived from non-parametric bootstraps of the 18 initial log-bacterial concentrations resemble the normal-theory based sampling distributions. Lastly, in place of simulating a large number of aleatoric variables (e.g., $n_a = 1000$), we demonstrated that $n_a = 25$ is sufficient when non-parametrically estimating the conditional CDFs of the risk assessment.

If simulating the aleatoric dimension of the Monte Carlo simulation is complex and time-consuming, then our model is preferred. Suppose that $n_e = 1000$ is a sufficient size in the epistemic dimension of our HUS risk-simulation study. The traditional two-dimensional Monte
Carlo simulation would set \( n_a \) equal to 1000. Thus, 1,000,000 simulations in the aleatoric dimension would be required. In our test, this resulted in a computing time of 3.09 seconds. \( n_a \) is 25 in our modified two-dimensional Monte Carlo simulation. Our approach runs 25,000 + 450,000 (bandwidth estimation) = 475,000 simulations in the aleatoric dimension, 47.5% the number of simulations in the traditional approach. Computing time for our approach is 2.17 seconds, about two-thirds the time of a traditional simulation of risk.

We also examine how sampling epistemic parameters from MLE-limiting or Normal-based sampling distributions impacts computing time. When using Pouillot and Delignette’s (2010) risk-assessment model with bootstrapped epistemic parameters, \( n_e = 1000 \), and \( n_a = 1000 \), computing time is 35 seconds. When using our model with \( n_e = 5000 \) and \( n_a = 25 \), the computing time is 10 seconds. Our approach is about three times faster.

While we have shown that one can efficiently execute a two-dimensional Monte Carlo risk assessment, the optimal bandwidth selection procedure outlined in section 3.4.3 does not lead to precise estimates of a global optimal bandwidth. A sample size of five is arguably too small. Furthermore, the criterion in (3.29) is anything but user-friendly. We would prefer a global and less derivative-intensive bandwidth-selection procedure that does not involve unknowns. The next chapter of this dissertation investigates global optimal bandwidth-selection procedures associated with the multivariate Nadaraya-Watson kernel regression estimator.
CHAPTER 4. COMPARISON OF BANDWIDTH SELECTORS FOR A MULTIVARIATE NADARAYA-WATSON KERNEL REGRESSION ESTIMATOR OF A CUMULATIVE DISTRIBUTION FUNCTION

Allan Trapp II and Philip Dixon

Abstract

Bandwidth selection is an important topic of non-parametric multivariate kernel regression estimation because bandwidth, $h$, primarily determines the estimator’s accuracy. In general, researchers pick an $h$ that balances the estimator’s bias and variance. They minimize criteria such as average squared error (ASE), penalized ASE, or asymptotic mean integrated squared error (AMISE) to select an “optimal” $h$. A review of the optimal bandwidth selection literature related to multivariate kernel-regression estimation shows that there is still ambiguity about the best bandwidth selector. We compare the effects of five penalized-ASE bandwidth selectors and an AMISE bandwidth plug-in on the average accuracy of a multivariate Nadaraya-Watson kernel-regression estimator of a cumulative distribution function, $CDF$, of hemolytic uremic syndrome (HUS) risk in young children exposed to Escherichia coli O157:H7 in ground beef patties. We consider these six bandwidth selectors because they compute relative quickly, and researchers generally desire fast results. Simulating different amounts of data ($n_e = 1000, 3000,$ and 5000) from each of three HUS-risk models of varying complexity, we find that none of the selectors consistently results in the most accurate $CDF$ estimator. However, if the goal is to produce accurate quantile-quantile risk assessment results (Pouillot and Delignette-Muller (2010)), then the AMISE-based selector performs best.
4.1 Introduction

Often, a researcher relates predictors, $X_i$, to a response, $Y_i$. A general formulation of this relationship (Fan and Gijbels (2000); Härdle et al. (2004); Hart and Yi (1998); Köhler et al. (2011)) is

$$Y_i = m(X_i) + \epsilon_i, \quad i = 1, 2, \ldots, n_e$$

(4.1)

where $m(X_i) = E(Y|X = X_i)$, and $\epsilon_i$ is an independent additive error with a mean of zero and a variance of $\sigma^2(X_i)$. A parametric treatment of $m$ would restrict the relationship to some mathematical form (e.g., a linear function of the predictors) and concede that “All models are wrong, but some are useful” (Box (1979)). In contrast, a non-parametric approach would allow $m$ to vary as a function of the data. Non-parametric methods circumvent the restrictive functional forms of parametric statistics because they are model-free (Fan and Gijbels (2000); Herrmann (2000)).

There are several non-parametric techniques of relating $X_i$ to $Y_i$. Common methods include kernel regression (Härdle (1990)), local polynomial estimation (Wand and Jones (1995)), smoothing splines (Eubank (1988)), or wavelets (Donoho and Johnston (1994)). We concentrate on the multivariate kernel regression estimator of $m$. For the context of this estimator, we use the risk simulation of chapter 3 and the non-parametric estimator of $m$ in equation (3.18).

A kernel regression estimator does have some shortcomings. When there are four or more predictors, the curse of dimensionality tends to make kernel regression impractical (Härdle and Müller (2000)). In the context of an eight-variable kernel density estimator, one needs about 10000 data points to attain an acceptable level of accuracy (specifically, a value of 0.01 under average squared error criterion) (Silverman (1998)). Our study uses simulated risk data and avoids this issue because we set $n_e$ to large values ranging from 1000 to 5000. Other non-parametric estimator considerations include the choice of kernel and bandwidth selector. Härdle et al. (2004) showed that the choice of kernel is practically irrelevant with respect to efficiency. In chapter 3, we chose the computationally-fast uniform kernel. Härdle (1991) argued that the main issue for multivariate non-parametric kernel regression is the choice of bandwidth, $h$.

Researchers have written much about “good” bandwidth selectors. Each of them used
various criteria to select an optimal bandwidth. Härdle (1990) considered mean squared error, \(MSE\), integrated squared error, \(ISE\), and average squared error, \(ASE\). In the case of a univariate Nadaraya-Watson estimator (Nadaraya (1964); Watson (1964)), all three criteria asymptotically have the same optimal bandwidth (Marron and Härdle (1986)). The \(MSE\) criterion is

\[
MSE(h) = \int_0^1 \left\{ \hat{m}_h(r, s \hat{\theta}_i) - m(r, s \hat{\theta}_i) \right\}^2 dR \quad (4.2)
\]

where \(\hat{m}_h(r, s \hat{\theta}_i)\) is the Nadaraya-Watson estimator with a uniform kernel (see equation (3.24)), and the integral is calculated for a fixed \(s \hat{\theta}_i\). Thus, \(MSE(h)\) is not a global measure of discrepancy. For our problem, we calculate the integral with respect to risk. The \(ISE\) criterion is

\[
ISE(h) = \int \left\{ \hat{m}_h(r, s \hat{\theta}_i) - m(r, s \hat{\theta}_i) \right\}^2 \, ds \hat{\theta}_i \quad (4.3)
\]

where the integral is taken with respect to the \(p\)-dimensional epistemic parameters. This measure is global because it applies to all possible values of \(s \hat{\theta}_i\). The \(ASE\) criterion is

\[
ASE(h) = \frac{1}{n_e} \sum_{i=1}^{n_e} \left\{ \hat{m}_h(r, s \hat{\theta}_i) - m(r, s \hat{\theta}_i) \right\}^2 \, w(s \hat{\theta}_i) \quad (4.4)
\]

where \(n_e\) is the number of \(s \hat{\theta}_i\)s, and \(w(s \hat{\theta}_i)\) is a weighting function that assigns less weight to \(s \hat{\theta}_i\)s in sparse regions of the epistemic parameter space. We incorporate the weighting function because researchers typically want an optimal bandwidth for the interior of the predictors’ support, and without a weight function, squared deviations of predictors \((s \hat{\theta}_i)\) near the boundary of the support dominate the value of \(ASE(h)\) (Herrmann (2000)). \(w(s \hat{\theta}_i)\) trims away the near-boundary deviations. We prefer (4.4) because it is a global measure of discrepancy, unlike \(MSE\) and its asymptotic equivalent in equation (3.29). Furthermore, it is simpler than the \(ISE\) criterion which requires integration over a \(p\)-dimensional space.

Researchers commonly use penalized versions of the \(ASE\) optimality criterion for the bandwidth selection of Nadaraya-Watson estimators. For the multivariate estimator, Härdle and Marron (1985) proposed a bandwidth based on leave-one-out cross validation. This is equivalent to Craven and Wahba’s (1979) generalized cross-validation (Härdle and Müller (2000)). Using the univariate Nadaraya-Watson estimator, Härdle et al. (2004) investigated bandwidths
derived from penalized \textit{ASE} criteria. The penalizing functions in their investigation are Shibata’s model selector (Shibata (1981)), generalized cross-validation (Craven and Wahba (1979)), Akaike’s Information Criterion (Akaike (1970)), Finite Prediction Error (Akaike (1974)), and Rice’s $T$ (Rice (1984)). Each of these penalizing functions assigns a different relative weight to the variance and bias of the non-parametric estimator of $m$. For example, Shibata’s model selector emphasizes bias reduction while Rice’s $T$ emphasizes variance reduction (Härdle et al. (2004)).

Beyond penalized-\textit{ASE} bandwidth selectors, we have recourse to other bandwidth selectors because we are estimating a conditional cumulative distribution function, \textit{CDF}. We know that the conditional \textit{CDF} is the integral of a conditional probability density function, \textit{pdf}. We argue that a bandwidth selector of a non-parametric estimator of a conditional \textit{CDF} need not come from the criteria derived for the Nadaraya-Watson estimator. Instead, one may use a bandwidth selector of the non-parametric estimator’s \textit{pdf} of predictors. Li and Racine (2008) showed that the results of their non-parametric estimators of conditional \textit{CDF}s are unaffected by the use of the \textit{pdf}-derived bandwidth selectors. They used the conditional \textit{pdf}-based bandwidth selector discussed by Hall et al. (2004). We consider the density-based bandwidth selector that Härdle and Müller (2000) described because it is data-driven and quickly computed. Details follow in subsection 4.3.3.

Other bandwidth selectors for \textit{CDF} estimation use bootstrapping, Bayesian posteriors, and conditional kernel regression \textit{AMSE} plug-ins. Cao-Abad and González-Manteiga (1993) and González-Manteiga et al. (2004) discussed smoothed and wild bootstrapped bandwidths. Zhang et al. (2006) used a Bayesian posterior distribution to select a bandwidth. We developed a plug-in bandwidth estimator for the Nadaraya-Watson estimator in section 3.4.3 (see equation (3.30)). We do not consider these bandwidth selectors further because each one has at least one limitation related to not being user-friendly, requiring a pilot bandwidth, or being too time consuming (Köhler et al. (2011)).

This paper’s objective is to compare and contrast the effects of computationally quick global bandwidth selectors on the overall accuracy of the $n_e$ Nadaraya-Watson \textit{CDF} estimators in equation (3.24) and the related non-parametric quantile-quantile estimators described in
section 3.4.3 (the black line in Figure 3.8 is an example of such an estimator). We investigate optimal bandwidths based on the five penalized regression-based ASE criteria and the density-based AMISE criterion described in Härdele and Müller (2000). By using our risk assessment model in chapter 3 and two simpler variants, we show for 1000, 3000, and 5000 that none of the considered bandwidth selectors consistently results in the most accurate Nadaraya-Watson estimator of a conditional CDF, on average. However, we do find that the density-based bandwidth plug-in generates an appropriate bandwidth when one uses a non-parametric estimate of the quantile-quantile function. In the following sections, we present the three risk models, describe the applications of the penalized ASE criteria and density-based AMISE criterion to bandwidth selection, compare bandwidths, end with a discussion about the shortcomings of penalized-ASE bandwidth selectors, and the recommend that researchers use the density-based AMISE bandwidth selector.

4.2 Simulation Models

Each of the three models from which we simulate has a different number of epistemic parameters, \( p \). We conjecture that a model’s complexity affects the behavior of the estimators under various bandwidth selectors. We examine models with \( p = 2, 4, \) or 7 epistemic parameters for \( n_a = 25 \).

The simplest model has been described in subsection 3.4.1. The model is a modified version of (4.5), (4.6), (4.7), and (4.8). We set the value of \( C \) to 0.232 and let \( D, S, \gamma_s, \) and \( \beta_s \) be the only model inputs. Thus, the two epistemic parameters are shape and scale of serving size. Their distribution is the multivariate normal described in subsection 3.3.1.

The four-parameter model lets \( C, \mu_c, \) and \( \sigma^2_c \) vary. The model is

\[
R = 1 - (1 - 0.0012)^D
\]  

where

\[
D \sim \text{Truncated N}(\text{Mean} = \lambda, \text{Variance} = \lambda)
\]

\[
\lambda = S \times 10^{C-0.8}
\]  

(4.5)
\[ S \sim \text{Weibull}(\gamma_s, \beta_s) \quad (4.8) \]

\[ C \sim \mathcal{N}(\mu_c, \sigma_c^2). \quad (4.9) \]

The distributions of the epistemic parameters \( \gamma_s, \beta_s, \mu_c, \) and \( \sigma_c^2 \) are outlined in subsections 3.3.1 and 3.3.2. The seven-parameter model is the full risk-assessment model of the previous chapter (see (3.1), (3.2), (3.3), (3.4), (3.5), (3.6), (3.7), (3.8), (3.9), (3.10), (3.11), (3.12), (3.16), and (3.17)).

### 4.3 Optimal Bandwidth Selectors

#### 4.3.1 Golden Bandwidths

Selection criteria quantify the discrepancy between a true function and its estimator at some bandwidth, \( h \). Appropriate criteria account for an estimator’s form and its inputs. In this chapter, we focus on \( \hat{m}_h^\text{unif}(r, s\hat{\theta}_i) \) and the non-parametric quantile-quantile estimator. These estimators are practically inverse functions of each other. As a result, each estimator has a different criterion.

For \( \hat{m}_h^\text{unif}(r, s\hat{\theta}_i) \), the inputs are \( s\hat{\theta}_i, r, \) and \( h \) where probability is the output. The ASE criterion is an insufficient discrepancy measure because it applies to a specific value of risk, \( r \). Thus, a global measure is

\[
IASE(h) = \int \frac{1}{n_e} \sum_{i=1}^{n_e} \left\{ \hat{m}_h^\text{unif}(r, s\hat{\theta}_i) - F_i(r|s\hat{\theta}_i) \right\}^2 w(s\hat{\theta}_i) dr \quad (4.10)
\]

where \( F_i(r|s\hat{\theta}_i) \) is the true conditional CDF. A generalization of (4.4) and (4.10) would replace \( h \) with a bandwidth matrix, \( H \). However, standardization of epistemic parameters permits the use of a scalar bandwidth, \( h \) (Härdle et al. (2004); Silverman (1998)).

For each \( b \in \{0, 0.01, ..., 1\} \) and \( B' = \{0.025, 0.25, 0.5, 0.75, \) and \( 0.975\} \), we let the non-parametric quantile-quantile estimator be \( Q \left( B' | \{ \hat{m}_h^{-1}(s\hat{\theta}_i | b) \} \right) \) where \( Q \) is the interpolated quantile function. Given a value of \( b \), the expression shows that \( B' \)-quantiles (five risk values) are taken from the set of \( n_e \) interpolated \( b \)-quantiles (Figure 3.7 demonstrates interpolation). The inputs of the quantile-quantile estimator are probabilities, \( B' \) and \( b, s\hat{\theta}_i, \) and \( h \), while risk values are the outputs. For every \( b' \in B' \), we plot each of the 101 quantile-quantile
risk values against the corresponding $b$ and have a $CDF$–like step function as presented in Figure 3.4.3. These are not CDFs, and we cannot summarize closeness through an average of squared (vertical) probability differences between $Q \left( B \mid \left\{ m_h^{-1}(s\theta_1|b) \right\} \right)$ and the true quantile-quantile function, $Q \left( B \mid \left\{ m_h^{-1}(s\hat{\theta}_1|b) \right\} \right)$. An appropriate measure of closeness is the average of squared horizontal distance between $Q \left( B \mid \left\{ m_h^{-1}(s\hat{\theta}_1|b) \right\} \right)$ and $Q \left( B \mid \left\{ m_h^{-1}(s\hat{\theta}_1|b) \right\} \right)$. We use

$$QASE(h) = \frac{1}{101} \sum_b \frac{1}{5} \sum_{b' \in B'} \left( Q \left( B' \mid \left\{ m_h^{-1}(s\hat{\theta}_1|b) \right\} \right) - Q \left( B' \mid \left\{ m_h^{-1}(s\hat{\theta}_1|b) \right\} \right) \right)^2. \quad (4.11)$$

Unlike equation (4.10), a weighting function is not included. With a sufficiently large $n_e$, $QASE(h)$ is a robust discrepancy measure.

Bandwidths corresponding to the minima of the criteria in equations (4.10) and (4.11) are the golden bandwidths of the estimators, $\hat{m}^{unif}_h(r,s\hat{\theta}_1)$ and $Q \left( B' \mid \left\{ m_h^{-1}(s\hat{\theta}_1|b) \right\} \right)$, respectively. On average, they lead to the most accurate non-parametric estimators of $m_h^{unif}(r,s\hat{\theta}_1)$ and $Q \left( B' \mid \left\{ m_h^{-1}(s\hat{\theta}_1|b) \right\} \right)$, respectively. Since these are the gold standards, we derive them for each of the nine model and $n_e$ combinations. For a given estimator, we estimate the true function with $n_a = 10000$ Monte Carlo simulations at equally spaced $h$’s (e.g., 0.5, 0.51, ...). If the estimator is $\hat{m}^{unif}_h(r,s\hat{\theta}_1)$, then we numerically integrate out the risk. The golden bandwidth is the minimum from a univariate Gaussian kernel regression of the bandwidths on the criterion values. An appropriate bandwidth comes from a visual comparison of the step criterion function to the smoothed estimator. We estimate curves with R’s locpoly function in the KernSmooth package (Wand (2011)).

Golden bandwidths are computationally costly because a researcher needs to run our proposed risk-assessment procedure for a comb of different bandwidth values (e.g., 45 values). If a research used golden bandwidths, then our procedure is useless because it would be more computationally intensive than the traditional two-dimensional Monte Carlo approach. In this chapter, we concentrate on relatively quick, data-driven bandwidth procedures. The next two subsections describe five penalized $ASE$ selectors and the density-based plug-in.
4.3.2 Optimal Penalized Average Squared Error Bandwidths

The previous subsection outlined two distinct optimality criteria for the non-parametric estimators $\hat{m}_h^{unif}(r, s_{\hat{\theta}_i})$ and $Q(B'^{-1}(s_{\hat{\theta}_i} | b))$. The criteria explored in this section are empirical estimates of $\hat{m}_h^{unif}(r, s_{\hat{\theta}_i})$’s optimality criterion in equation (4.10). Thus, the optimal bandwidths of this section should approximate the golden bandwidth of $\hat{m}_h^{unif}(r, s_{\hat{\theta}_i})$.

An estimate of the $ASE$ criterion in equation (4.4) requires a good estimator of the unknown quantity $m(r, s_{\hat{\theta}_i})$. In our problem, $m(r, s_{\hat{\theta}_i})$ is a conditional CDF evaluated at $r$. We replace it with $\hat{F}_i(r) = \frac{\sum_{j=1}^{n_e} I\{R_{i,j} \leq r | s_{\hat{\theta}_i}\}}{n_e}$ because that is an unbiased estimator of $F_i(r)$, for $r \in [0, 1]$. This replacement generates the estimate $\hat{ASE}(h, r)$, where

$$\hat{ASE}(h, r) = \frac{1}{n_e} \sum_{i=1}^{n_e} \left\{ \hat{m}_h^{unif}(r, s_{\hat{\theta}_i}) - \hat{F}_i(r) \right\}^2 w(s_{\hat{\theta}_i}).$$

(4.12)

Härdle et al. (2004) called (4.12) the *resubstitution estimate*. For the same reason as outlined in section 4.3.1 about $\hat{m}_h^{unif}(r, s_{\hat{\theta}_i})$, we adapt (4.12) by numerically integrating over the risk values. The new estimator is

$$\hat{ASE}(h) = \int \frac{1}{n_e} \sum_{i=1}^{n_e} \left\{ \hat{m}_h^{unif}(r, s_{\hat{\theta}_i}) - \hat{F}_i(r) \right\}^2 w(s_{\hat{\theta}_i}) dr.$$  (4.13)

Equation (4.13) is a data-driven discrepancy measure of the conditional Nadaraya-Watson CDF estimator, and its computational effort is equal to that of estimating the conditional CDFs.

We must consider penalized versions of (4.13) because the direct implementation of the modified criterion is downward biased (Härdle (1990)). Given that we search for the $h$ that minimizes (4.13), if we use $\hat{ASE}(h)$, then the minimizing $h$ will always be 0 because $\hat{m}_h^{unif}(r, s_{\hat{\theta}_i})$ is a function of $\hat{F}_i(r)$. Penalizing functions that have a first-order Taylor series expansion of $1 + 2u + O(u^2)$, for $u \to 0$, asymptotically correct for this bias (Härdle et al. (2004)). The estimator of the penalized $ASE$ criterion is

$$\hat{pASE}(h) = \int \frac{1}{n_e} \sum_{i=1}^{n_e} \left\{ \hat{m}_h^{unif}(r, s_{\hat{\theta}_i}) - \hat{F}_i(r) \right\}^2 \Xi\left(\frac{1}{n_i}\right) w(s_{\hat{\theta}_i}) dr$$

(4.14)

where $\Xi\left(\frac{1}{n_i}\right)$ is the penalizing function, and $n_i$ is the number of epistemic parameters in a hypersphere of radius $h$ around $s_{\hat{\theta}_i}$. Penalizing functions that satisfy the first-order Taylor
series expansion condition include Shibata’s model selector (Shibata (1981))

\[ \Xi_S \left( \frac{1}{n_i} \right) = 1 + \frac{2}{n_i}, \]  

(4.15)
generalized cross-validation (Craven and Wahba (1979))

\[ \Xi_{GCV} \left( \frac{1}{n_i} \right) = \left( 1 - \frac{1}{n_i} \right)^{-2}, \]

(4.16)
Akaike’s information criterion (Akaike (1970))

\[ \Xi_{AIC} \left( \frac{1}{n_i} \right) = \exp \left( \frac{2}{n_i} \right), \]

(4.17)
finit prediction error (Akaike (1974))

\[ \Xi_{FPE} \left( \frac{1}{n_i} \right) = \frac{n_i + 1}{n_i - 1}, \]

(4.18)
and Rice’s T (Rice (1984))

\[ \Xi_T \left( \frac{1}{n_i} \right) = \left( 1 - \frac{2}{n_i} \right)^{-1}. \]

(4.19)

Although the values of \( \hat{p} ASE(h) \) under each of these penalty functions are asymptotically equivalent as \( n_e \to \infty \), they lead to different optimal bandwidths if the size of \( n_e \) is small.

The last ambiguity of \( \hat{p} ASE(h) \) is the selection of the weighting function, \( w(\hat{s} \theta_i) \). This function trims away \( \left( \hat{m}_{unif}(r, \hat{s} \theta_i) - \hat{F}_i(r) \right)^2 \) deviations whose \( \hat{s} \theta_i \)'s are near the boundary of the epistemic parameter space. Typically, these deviations dominate the value of the criteria in (4.14). This is undesirable because researchers desire an optimal bandwidth corresponding to \( \hat{s} \theta_i \)'s in the interior of the epistemic parameter space (Herrmann (2000)). Thus, \( w(\hat{s} \theta_i) = 0 \) for all \( \hat{s} \theta_i \) near the boundary and 1 everywhere else.

We use R’s depth function from the depth package (Genest et al. (2009)) to determine near-boundary \( \hat{s} \theta_i \)'s (Rousseeuw and Struyf (1988)). A \( \hat{s} \theta_i \) is near the boundary if its depth value is less than some critical value. We choose the critical value such that 10% of the \( n_e \) epistemic parameter vectors have depths less than it. Figure 4.1 demonstrates that depth catches the boundary points (with the exception of one point). True near-boundary points lie outside the orange circle. The depth rule classifies a near-boundary point with an open circle and classifies an interior point with a filled circle. For parameters U1 and U2, we see that nearly all of the true near-boundary points are classified as such.
Figure 4.1: Plot of Trimmed (Open Points) and Retained (Filled Points) Observations for a Model with Two Parameters: U1 and U2 are the Mahalanobis standardized (see equation (3.19)) shape and scale serving parameters.
4.3.3 Multivariate Kernel Density Plug-In Bandwidth

Whereas the preceding section’s criterion theoretically relates to optimal bandwidths of \( \hat{m}_{h}^{\text{unif}}(r, s) \hat{\theta}_i \), this section’s asymptotic mean integrated squared error, AMISE, plug-in bandwidth has no direct ties to either criterion in (4.10) and (4.11). Nonetheless, the “integral” relationship of a CDF and pdf motivates the consideration of this multivariate density selection criterion. Furthermore, Li and Racine (2008) have used a multivariate pdf—based criterion to determine an optimal bandwidth of a non-parametric CDF estimator (Li and Racine (2008)).

Härdle and Müller (2000) and Scott (1992) showed that the asymptotic mean integrated squared error of a multivariate kernel density estimator is

\[
AMISE(h) = \frac{1}{4} \mu_2^2(K^{\text{unif}}) \int \left[ h^2 \text{tr}\left\{ \mathcal{H}_f(s) \right\} \right]^2 ds \hat{\theta}_i + \frac{1}{n_e h^p} \| K^{\text{unif}} \|_2^2 (4.20)
\]

where \( \mathcal{H}_f \) is the Hessian of the joint pdf of the epistemic parameters. The definitions of \( \mu_2^2(K^{\text{unif}}) \) and \( \| K^{\text{unif}} \|_2^2 \) follow the statement of equation (3.29). In our models, \( p = 2, 4, \) or 7.

The form of (4.20) is not very convenient as it involves the integral of a sum of second partial derivatives. Nonetheless, by using a Gaussian kernel, Scott (1992) derived a rule-of-thumb optimal \( h \) for the multivariate normal density with a diagonal variance-covariance matrix. In the notation of our risk model, the optimal bandwidth is

\[
\eta_e^{-1/(p+4)}. \tag{4.21}
\]

Supposing that the true \( f \) is similar to a multivariate normal density, Härdle and Müller (2000) described an optimal plug-in bandwidth that applies (4.21) and canonical kernels (Mar

Ron and Nolan (1988)). The multivariate normal condition is plausible because three of the seven standardized epistemic parameters are normal, and graphical exploration shows that the others are approximately so. The purpose of canonical kernels is to relate AMISE criteria for two separate kernel functions. More simply, given the reference optimal bandwidth in (4.21), one may derive the optimal \( h \) for any kernel. In the context of our problem, the plug-in/optimal bandwidth of the uniform kernel is

\[
h^U = \frac{\delta^U_0}{\delta^0} \eta_e^{-1/(p+4)}. \tag{4.22}
\]
Table 4.1: Canonical Bandwidths for the Uniform and Normal Kernels

| $n$ | $||K||_2^2$ | $\mu_2^2(K)$ | $\delta_0^U$ | $\delta_0^G$ | $\delta_0^U / \delta_0^G$ |
|-----|-------------|-------------|-------------|-------------|----------------|
| 1   | $\frac{1}{2}$ | $\frac{1}{9}$ | $(\frac{2}{9})^{1/5}$ | $(2\sqrt{\pi})^{-1/5}$ | .9534          |
| 2   | $\frac{1}{\pi}$ | $\frac{1}{16}$ | $(\frac{16}{\pi})^{1/6}$ | $(2\sqrt{\pi})^{-1/3}$ | 2              |
| 3   | $\frac{3}{4\pi}$ | $\frac{1}{25}$ | $(\frac{75}{4\pi})^{1/7}$ | $(2\sqrt{\pi})^{-3/7}$ | 2.2201         |
| 4   | $\frac{2}{\pi^2}$ | $\frac{1}{36}$ | $(\frac{72}{\pi^2})^{1/8}$ | $(2\sqrt{\pi})^{-1/2}$ | 2.4137         |
| 5   | $\frac{15}{8\pi^2}$ | $\frac{1}{49}$ | $(\frac{785}{8\pi^2})^{1/9}$ | $(2\sqrt{\pi})^{-5/9}$ | 2.6072         |
| 6   | $\frac{6}{\pi^3}$ | $\frac{1}{64}$ | $(\frac{384}{\pi^3})^{1/10}$ | $(2\sqrt{\pi})^{-3/5}$ | 2.7482         |
| 7   | $\frac{105}{16\pi^3}$ | $\frac{1}{81}$ | $(\frac{8505}{16\pi^3})^{1/11}$ | $(2\sqrt{\pi})^{-7/11}$ | 2.8969         |
| 8   | $\frac{24}{\pi^4}$ | $\frac{1}{100}$ | $(\frac{2400}{\pi^4})^{1/12}$ | $(2\sqrt{\pi})^{-2/3}$ | 3.0365         |

where $\delta_0^U$ and $\delta_0^G$ are the canonical bandwidths of the uniform and Gaussian kernel, respectively. Their values are

$$\delta_0^U = \left\{ \frac{||K||_2^2}{\mu_2^2(K)} \right\}^{1/(p+4)} \tag{4.23}$$

and

$$\delta_0^G = \left\{ \frac{1}{(2\sqrt{\pi})} \right\}^{p/(p+4)}. \tag{4.24}$$

Table 4.1 provides necessary values to calculate $h^U$ for any $p$ between 1 and 8.

### 4.4 Results

Under nine scenarios, we examine the proximities of Shibata’s, generalized cross-validation ($GCV$), Akaike information criterion ($AIC$), finite prediction error ($FPE$), and Rice’s $T$ optimal bandwidths and the density-based plug-in to the golden bandwidths of $\hat{m}_{h}^{\text{unif}}(r,s\hat{\theta}_i)$ and $Q \left( B' \| \{ \hat{m}_{h}^{-1}(s\hat{\theta}_i|b) \} \right)$. For these six optimal bandwidths, the three variants of the HUS-risk model (see section 4.2), and $n_e = 1000$, we also graphically compare the 0.3, 0.5, and 0.7 quantile estimates of the five CDF-like curves of Figure 3.8 to those of the two-dimensional Monte Carlo-based quantiles ($n_a = 10000$). This graphical comparison provides a visual accuracy measure of our estimated points under any one of the six optimal bandwidths. The patterns observed in these nine plots hold for $n_e = 3000$ and 5000; hence, we do not show them.

Determination of the best bandwidth selector comes from the comparisons of differences between each of the six bandwidth selectors and the golden bandwidth of an estimator.
This is possible because the shapes of the IASE\( (h) \) (4.10) and QASE\( (h) \) (4.11) curves are approximately u-shaped. As an optimal bandwidth approaches an estimator’s golden bandwidth, the accuracy of the estimator evaluated at the optimal bandwidth gets better. Given nine scenarios, we highlight which of the six optimal bandwidths is closest to a golden bandwidth. In Tables 4.2, 4.3, and 4.4, the best bandwidth selector of \( \hat{m}_{h}^{\text{unif}}(r, s \hat{\theta}_i) \) has a bandwidth value highlighted in blue. The best bandwidth selector of \( Q\left(B’|\left\{ \hat{m}_{h}^{-1}(s \hat{\theta}_i|b) \right\}\right) \) has a bandwidth value highlighted in red.

Table 4.2: Optimal Bandwidths for a Risk Model Containing Two Epistemic Parameters

| \( n_e \) | \( q \left( \theta'|\left\{ \hat{m}_{h}^{-1}(s \hat{\theta}_i|b) \right\}\right) \) | Plug-In | Shibata | GCV | AIC | FPE | Rice’s \( T \) | \( m_{h}^{\text{unif}}(r, s \hat{\theta}_i) \) |
|---------|---------------------------------|--------|--------|-----|-----|-----|-------------|----------------|
| 1000    | 0.5640                          | 0.6325 | 1.0740 | 1.0740 | 1.0740 | 1.0740 | 1.1280      |
| 3000    | 0.3310                          | 0.5266 | 0.5865 | 0.5888 | 0.5865 | 0.5888 | 0.5880 | 0.9470      |
| 5000    | 0.2505                          | 0.4837 | 0.9435 | 0.9424 | 0.9435 | 0.9424 | 0.9439 | 0.8385      |

For \( p = 2 \), Table 4.2 shows that the five criterion report very similar optimal bandwidths for \( n_e = 1000, 3000, \) and 5000. When \( n_e = 1000 \), the penalized bandwidth selectors are 0.054 away from \( \hat{m}_{h}^{\text{unif}}(r, s \hat{\theta}_i) \)'s golden bandwidth. Oddly, when \( n_e = 3000 \), they are about 0.36 away from \( \hat{m}_{h}^{\text{unif}}(r, s \hat{\theta}_i) \)'s golden bandwidth and about 0.1 away for \( n_e = 5000 \). For all investigated \( n_e \), the plug-in estimator is closest to \( Q\left(B’|\left\{ \hat{m}_{h}^{-1}(s \hat{\theta}_i|b) \right\}\right) \)'s golden bandwidth. Differences between the plug-in bandwidth and \( Q\left(B’|\left\{ \hat{m}_{h}^{-1}(s \hat{\theta}_i|b) \right\}\right) \)'s golden bandwidth range between 0.06 and 0.24.

When examining the accuracy of selected 0.3, 0.5, and 0.7-quantile estimates of the five \( CDF \)-like curves under each of the six optimal bandwidths (top row of plots in Figure 4.2) in the 2-parameter model, it is unclear whether or not one of the six optimal bandwidths results in the most accurate estimator. Ideally, we desire an optimal bandwidth that produces estimates (points labeled 1 = Plug-in, 2 = Shibata, 3 = GCV, 4 = AIC, 5 = FPE, and 6 = Rice’s \( T \)) that are consistently closest to the corresponding two-dimensional Monte Carlo quantile estimates (gray lines). For a given line-type (e.g. a dashed line), the traditional two-dimensional Monte Carlo estimate is the gray line while the colored line is the modified two-dimensional Monte Carlo estimate under various bandwidths.

For \( p = 4 \), Table 4.3 shows that the best bandwidth selector for \( \hat{m}_{h}^{\text{unif}}(r, s \hat{\theta}_i) \) is Shibata’s
Figure 4.2: Plot of 0.3, 0.5, 0.7-Quantile Estimates of the CDF-Like Curves Under Six Optimal Bandwidths: Model type (2, 4, or 7-parameter models) varies across the rows of the matrix plot while 0.3, 0.5, or 0.7-quantile of the 0.025, 0.25, 0.5, 0.75, and 0.975 CDF-like curves varies across the columns of the matrix plot. A black line demonstrates how a 0.3, 0.5, or 0.7 quantile estimate of the 0.025 CDF-like curve in either the 2, 4, or 7-parameter model changes as a function of bandwidth value. Red line = risk estimates of 0.25 CDF-like curve. Green line = risk estimates of 0.5 CDF-like curve. Blue line= risk estimates of 0.75 CDF-like curve. Aqua line = risk estimates of 0.975 CDF-like curve. Gray lines = “true” (two-dimensional Monte Carlo-based) estimates. 1 = Plug-In. 2 = Shibata. 3 = GCV. 4 = AIC. 5 = FPE. 6 = Rice’s T.
for $n_e = 1000$, Rice’s $T$ for $n_e = 3000$, and $GCV$ for $n_e = 5000$. Again, the density-based plug-in bandwidth is the best selector for $Q \left( B' \mid \{ \hat{h}^{-1}_{\hat{m}}(s \hat{\theta}_i \mid b) \} \right)$ where differences between it and the golden bandwidth range between 0.1 and 0.15. The middle row of plots in Figure 4.2 shows that the plug-in-based estimator is also the most accurate. 1 is always the closest point to the corresponding gray line for all three quantiles of the five CDF-like curves.

For $p = 7$, Table 4.4 reports that the $AIC$ bandwidth selector is closest to $\hat{m}_{h}^{unif}(r, s \hat{\theta}_i)$’s golden bandwidth when $n_e = 1000$ or 5000. However, Shibata’s is best when $n_e = 3000$. For the most part, the plug-in bandwidth is closest to $Q \left( B' \mid \{ \hat{h}^{-1}_{\hat{m}}(s \hat{\theta}_i \mid b) \} \right)$’s golden bandwidth. The differences between the two bandwidths range from 0.02 to 0.07. The bottom row of plots in Figure 4.2 shows that the plug-in bandwidth relatively produces the most accurate estimates.

4.5 Discussion

When estimating a conditional $CDF$, $\hat{m}_{h}^{unif}(r, s \hat{\theta}_i)$, the results show that there is no consistently superior bandwidth selector. Generalized cross-validation and finite prediction error bandwidths are the best for the two-parameter $CDF$ estimator, $\hat{m}_{h}^{unif}(r, s \hat{\theta}_i)$. However, any of the penalized selectors are appropriate because their bandwidth values are very close (see Table 4.2) and their accuracy with respect to estimating the curves in Figure 3.8 is practically the same (see top row of plots in Figure 4.2). For $p = 4$, generalized-cross validation performs
pretty well as it is the second most accurate estimator of risk quantiles in Figure 4.2. For \( p = 7 \), Shibata’s and AIC bandwidths are preferred.

If \( p \) is too large and \( n_e \) is not large enough, then the GCV, FPE, and Rice’s \( T \) bandwidth selectors are impractical because they are incalculable for the true, golden bandwidth (Köhler et al. (2011)). The \( \hat{p}_{ASE}(h) \) of these selectors are incalculable for small \( h \). \( h \) determines the size of an estimator’s neighborhood, and small neighborhoods have small \( n_i \) values (i.e., not many neighbors). The GCV and FPE functions are undefined for \( n_i = 1 \) while Rice’s \( T \) is undefined for \( n_i = 2 \). For a given \( h \), if any of the \( n_e \) estimators have an \( n_i \) equal to either 1 or 2, then GCV, FPE, or Rice’s \( T \) \( \hat{p}_{ASE} \) is undefined. Sometimes, the smallest value of \( h \) where \( \hat{p}_{ASE} \) is calculable occurs after the golden \( h \). Thus, these criteria cannot produce optimal bandwidths close to the true, golden bandwidth.

When discussing the best data-driven bandwidth of \( Q \left( B' \left| \left\{ \tilde{m}_{h^{-1}(\hat{\theta}_i|b)} \right\} \right. \right) \), out of our set of bandwidth estimators, we prefer the density-based plug-in, as it was the closest bandwidth selector to the golden bandwidth in 8 of the 9 scenarios and it produces the most accurate estimates of the selected quantiles of the CDF-like curves in two of the three models (see Figure 4.2). Non-parametric conditional CDF estimators of chapter 3 should use the density-based plug-in bandwidth.
CHAPTER 5. CONCLUDING THOUGHTS

5.1 Summary of Conclusions

In this dissertation, I have answered two risk-related questions.

1. Can a statistically informed seed-viability test schedule be devised that accounts for the risk of potentially losing a seed lot and the risk of potentially conducting a premature test?

2. Can the computing speed of a two-dimensional Monte Carlo simulation of risk be improved?

Chapter 2’s response to the first question has four components. First, my coauthors and I fitted a multi-level Bayesian model by using a quadratic seed-viability curve (non-traditional) of germination against seed age. Then, for each seed lot, we extrapolated the seed age at which the lot would reach a critical germination level \((CRV)\). We compared these predictions to those from the Avrami based, Bayesian multi-level model (Walters et al. (2005)) and concluded that the quadratic curve produces better predictions. The fit of the Avrami multi-level Bayesian model to the data was poor. For the third component, we defined the \(\alpha\)-quantile testing rule and non-parametrically estimated the risks of potentially losing a seed lot, \(fnr(\alpha)\)’s, and the risks of conducting an unnecessary viability test, \(fpr(\alpha)\)’s. Lastly, using the cost function criterion and user-specified costs of \(fnr\) and \(fpr\), we determined an optimal testing age by which genebank managers could schedule viability tests.

In a follow-up study, we tested the effectiveness of our \(\alpha\)-quantile testing rule by using \(\alpha = 0.05\). 125 stratified seed viability tests were conducted, and the observed germination percentages were compared to their predicted germination percentages. From the follow-up
tests, the risk of conducting an early test ($fpr$) was 0.55, and the risk of potentially losing a seed lot ($fnr$) was 0.035. The estimated $fpr$ and $fnr$ were 0.39 and 0.097, respectively. There is empirical evidence that our scheduling method produces testing ages that are earlier than the true critical testing age ($fpr$: 0.55 > 0.39). Premature tests may be the result of the restrictive symmetrical form of the quadratic viability curve (see second paragraph of section 2.7 for details). As a solution, we suggest increasing the value of $\alpha$.

Both Chapters 3 and 4 respond to the second question. A risk assessor may expedite results of a two-dimensional Monte Carlo risk assessment in two ways. First, if a risk assessment uses either a parametric or non-parametric bootstrapped distribution of epistemic parameters as implemented by Pouillot and Delignette-Muller (2010), then the risk analyst may directly simulate values from an appropriate sampling distribution. The bootstrapped distributions are time-consuming because they require additional Monte Carlo simulation and $n_e$ maximum likelihood estimates. For the HUS-risk data, my coauthor and I showed that the bootstrapped distributions of the epistemic parameter estimators related to serving size and initial log$_{10}$ concentration of bacteria closely resembled the estimators’ sampling distributions. If a risk analyst uses maximum likelihood estimators of epistemic parameters such as those of serving size, then the analyst may directly simulate epistemic parameter values from a limiting normal distribution. If an aleatoric variable is normally distributed such as initial log$_{10}$ concentration of bacteria on a frozen ground beef patty, then the analyst may directly simulate epistemic parameter values from a sampling distribution based in normal theory.

Next, we proposed that analysts use a small number of simulations in the aleatoric dimension ($n_a = 25$ instead of $n_a = 1000$) and fit multivariate Nadaraya-Watson kernel estimators of the conditional CDF’s, $F_i$’s. We showed in Figure 3.5 that the non-parametric CDF estimator may produce a better, more accurate estimator of the CDF than a traditional Monte Carlo CDF estimator. More importantly, the results from the modified two-dimensional Monte Carlo simulation generated from Nadaraya-Watson estimators of CDF’s closely resembles those from a traditional two-dimensional Monte Carlo simulation (see Figure 3.8).

We pointed out at the end of Chapter 3 that $Q\left(B'\left|\hat{\gamma}_i\left(s,\hat{\theta}_i|b\right)\right.\right]$ is sensitive to the value of bandwidth, $h$, and argued that a fast, less variable bandwidth estimator is necessary. Chapter
4 compares the accuracy of $\hat{m}_{h}^{\text{unif}}(r, s\hat{\theta}_{i})$ and $Q\left(B' \mid \{\hat{m}_{h}^{-1}(s\hat{\theta}_{i}|b)\}\right)$ estimators under six fast bandwidth selectors. We concluded that there is no best bandwidth selector for $\hat{m}_{h}^{\text{unif}}(r, s\hat{\theta}_{i})$. Also, in the context of our risk models, the shapes of the penalized integrated $ASE$ criteria were not always parabolic and concave-up like the true $IASE(h)$ curve. Our investigation suggests that $n_c$’s size may limit the effectiveness of penalized bandwidth selectors. Furthermore, for all values of $b$, the density-based plug-in estimator leads to relatively accurate estimation of $Q\left(B' \mid \{\hat{m}_{h}^{-1}(s\hat{\theta}_{i}|b)\}\right)$ (the gray step functions of Figure 3.8). It is the best bandwidth selector of $Q\left(B' \mid \{\hat{m}_{h}^{-1}(s\hat{\theta}_{i}|b)\}\right)$.

5.2 Topics for Further Investigation

In chapter 2, we related a seed lot’s germination values to a quadratic function of its observed testing ages because the curve accounted for incomplete initial germination and allowed for after-ripening. We recognize that these symmetrical curves are inconsistent with the biology of seed viability behavior. Thus, future research should investigate modifications to existing viability curves, like those from Ellis and Roberts (1980) or Walters et al. (2005).

When non-parametrically estimating the conditional $CDF$’s, $\hat{m}_{h}^{\text{unif}}(r, s\hat{\theta}_{i})$s, we used the multivariate Nadaraya-Watson estimator. Future research should investigate the performance of our modified two-dimensional Monte Carlo risk assessment method when using the multivariate local linear regression estimator of the conditional $CDF$ (Wand and Jones (1995)). H"ardle et al. (2004) report that the Nadaraya-Watson estimator is inaccurate near the boundary of the support. The multivariate local linear regression estimator fits better than the Nadaraya-Watson estimator near the boundary. Additionally, other fast, data-driven optimal bandwidth selectors exist for a multivariate local linear regression estimator (Fan and Gijbels (2000); H"ardle et al. (2004); Ruppert et al. (1995)).
Table A.1: \( \hat{t}_i^{Q} \) Computations: In our example, the \( CRV \) (critical value) corresponds to 0.50, but it may be adjusted to any desired value. The “negative” and “positive” in the discriminant column refer to the root that is used. For curves C7, C8, C10, and C12, the predicted test time is infinity, but we capped it at 10,000 years; hence, there is a column for the value of \( \hat{t}_i^{Q} \). Curve C9 has a test time upon the arrival of the seed lot (\( \hat{t}_i^{Q} = 0 \)). \( f^* = \hat{\beta}_{0,i} + \hat{\beta}_{1,i}x^* + \hat{\beta}_{2,i}(x^*)^2 \), where \((x^*, f^*)\) is the location of the curve’s vertex.

<table>
<thead>
<tr>
<th>Curve Type</th>
<th>( x^* )</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>( x^* &gt; 0 ) negative ( \hat{\beta}_{0,i} &lt; CRV ) ( f^* &gt; CRV ) negative</td>
</tr>
<tr>
<td>C2</td>
<td>( x^* &lt; 0 ) negative ( \hat{\beta}_{0,i} &gt; CRV ) ( f^* &gt; CRV ) negative</td>
</tr>
<tr>
<td>C3</td>
<td>( x^* &gt; 0 ) negative ( \hat{\beta}_{0,i} &gt; CRV ) ( f^* &lt; CRV ) negative</td>
</tr>
<tr>
<td>C4</td>
<td>( x^* &lt; 0 ) negative ( \hat{\beta}_{0,i} &lt; CRV ) ( f^* &lt; CRV ) negative</td>
</tr>
<tr>
<td>C5</td>
<td>( x^* &gt; 0 ) negative ( \hat{\beta}_{0,i} &gt; CRV ) ( f^* &lt; CRV ) negative</td>
</tr>
<tr>
<td>C6</td>
<td>( x^* &lt; 0 ) negative ( \hat{\beta}_{0,i} &lt; CRV ) ( f^* &lt; CRV ) negative</td>
</tr>
<tr>
<td>C7</td>
<td>( x^* &gt; 0 ) positive ( \hat{\beta}_{0,i} &gt; CRV ) ( f^* &gt; CRV ) negative</td>
</tr>
<tr>
<td>C8</td>
<td>( x^* &lt; 0 ) positive ( \hat{\beta}_{0,i} &lt; CRV ) ( f^* &lt; CRV ) negative</td>
</tr>
<tr>
<td>C9</td>
<td>( x^* &gt; 0 ) positive ( \hat{\beta}_{0,i} &gt; CRV ) ( f^* &lt; CRV ) negative</td>
</tr>
<tr>
<td>C10</td>
<td>( x^* &lt; 0 ) positive ( \hat{\beta}_{0,i} &lt; CRV ) ( f^* &lt; CRV ) negative</td>
</tr>
<tr>
<td>C11</td>
<td>( x^* &gt; 0 ) positive ( \hat{\beta}_{0,i} &lt; CRV ) ( f^* &lt; CRV ) negative</td>
</tr>
<tr>
<td>C12</td>
<td>( x^* &lt; 0 ) positive ( \hat{\beta}_{0,i} &gt; CRV ) ( f^* &lt; CRV ) negative</td>
</tr>
</tbody>
</table>

Knowing whether one adds or subtracts the discriminant in (2.12) is determined by a seed lot’s viability curve. A quadratic viability curve may be completely characterized by three
seed features. We quantify these features with three numerical estimates. The value of the quadratic term, $\hat{\beta}_{2,i}$, estimates the shape/convexity of a lot’s curve. $\hat{\beta}_{0,i}$ estimates the initial germination of a lot. Last, we estimate the location of a curve’s vertex. Biologically, this should be the age at which peak germination occurs (a local maximum). However, there are cases where the vertex is a local minimum of a curve. Examples include when a lot’s historical viability tests strictly increase over time or when they remain stagnant over time and vary little.

Applying derivatives to equation (2.2), we estimate the horizontal coordinate of the vertex as

$$x^* = -\frac{\hat{\beta}_{1,i}}{2\hat{\beta}_{2,i}}.$$ 

Table A.1 enumerates the cases when one takes the negative discriminant of equation (2.12); when $\hat{t}_Q$ takes a value of 0 (e.g., original lot has large portion of dead seed); or when it is 10,000 (historical data do not provide evidence of a decline in viability). Lots that may or may not exhibit after-ripening and have a concave-down curve are estimated with the C1 and C3 curves. These curves differ in their predicted initial germination. Lots with consistently high germination rates are estimated with C7 and C8.
BIBLIOGRAPHY


