RICE UNIVERSITY

An Old Dog Learns New Tricks: Novel Applications of Kernel Density Estimators on Two Financial Datasets

by

Matthew Cline Ginley

A THESISSubmitted
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE
Doctor of Philosophy

APPROVED, THESIS COMMITTEE:

Katherine B. Ensor, Chair
Professor of Statistics
Rice University

David W. Scott, Co-Chair
Noah Harding Professor of Statistics
Rice University

Brian Akins
Assistant Professor of Accounting
Rice University

Houston, Texas
December, 2017
ABSTRACT

In our first application, we contribute two nonparametric simulation methods for analyzing Leveraged Exchange Traded Fund (LETF) return volatility and how this dynamic is related to the underlying index. LETFs are constructed to provide the indicated leverage multiple of the daily total return on an underlying index. LETFs may perform as expected on a daily basis; however, fund issuers state there is no guarantee of achieving the multiple of the index return over longer time horizons. Most, if not all LETF returns data are difficult to model because of the extreme volatility present and limited availability of data. First, to isolate the effects of daily, leveraged compounding on LETF volatility, we propose an innovative method for simulating daily index returns with a chosen constraint on the multi-day period return. By controlling for the performance of the underlying index, the range of volatilities observed in a simulated sample can be attributed to compounding with leverage and the presence of tracking errors. Second, to overcome the limited history of LETF returns data, we propose a method for simulating implied LETF tracking errors while still accounting for their dependence on underlying index returns. This allows for the incorporation of the complete history of index returns in an LETF returns model. Our nonparametric methods are flexible—easily incorporating any chosen number of days, leverage ratios, or period return constraints, and can be used in combination or separately to model any quantity of interest derived from daily LETF returns.

For our second application, we tackle binary classification problems with extremely low class 1 proportions. These “rare event” problems are a considerable challenge, which is magnified when dealing with large datasets. Having a miniscule count of class 1 observations motivates the implementation of more sophisticated methods to minimize forecasting bias towards the majority class. We propose an alternative approach to established up-sampling or down-sampling algorithms driven by kernel density estimators to transform the class labels to continuous targets. Having effectively transformed the problem from classification to regression, we argue that under the assumption of a monotonic relationship between predictors and the target, approximations of the majority class are possible in a rare events setting with the use of simple heuristics. By significantly reducing the burden posed by the majority class, the complexities of minority class membership can be modeled more effectively using monotonically constrained nonparametric regression methods. Our approach is demonstrated on a large financial dataset with an extremely low class 1 proportion. Additionally, novel features engineering is introduced to assist in the application of the density estimator used for class label transformation.
## Contents

### Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of Figures</td>
<td>iii</td>
</tr>
<tr>
<td>List of Tables</td>
<td>vi</td>
</tr>
</tbody>
</table>

### I LETF Returns Dataset

1 Dynamics of LETF Returns and A New Measure of Volatility
   1.1 Universe of LETFs                                                 | 3    |
   1.2 LETF Returns                                                      | 4    |
   1.3 Volatility Drag and Convexity                                     | 8    |
   1.4 Problems with Sample Standard Deviation                           | 13   |
   1.5 Define Shortfall from Maximum Convexity                           | 17   |
   1.6 Fundamental Motivation for SMC                                   | 26   |
   1.7 Statistical Motivation for SMC                                   | 28   |

2 Simulation of Leveraged ETF Volatility Using Nonparametric Density Estimation
   2.1 Overview and Discussion of LETF Returns Literature               | 33   |
   2.2 Visualizing LETF Returns and Measuring Realized Volatility       | 34   |
   2.3 Simulation of Tracking Errors Using Nonparametric Density Estimation | 36   |
   2.4 Constrained Simulation Using Nonparametric Density Estimation   | 46   |
   2.5 Case Study: Simulating LETF Returns with Underlying Index Return Constraints and Calculating Nonparametric p-values for Volatility Statistics | 54   |
   2.6 Closing Remarks                                                  | 56   |

### II Firm Failure Events Dataset

3 Framing Firm Failure and Preliminary Model Evaluation
   3.1 Framing the Firm Failure Data Problem                             | 59   |
   3.2 Preliminary Firm Failure Model Evaluation                        | 63   |
   3.3 Expanding the Firm Failure Data Problem                          | 72   |
4 Approximations to Large Scale, Rare Events Datasets for Use with Monotonically Constrained Classifiers

4.1 Overview .................................................. 78
4.2 Rare Events Literature .................................. 80
4.3 Dataset Modifications ..................................... 82
4.4 Approximation of Rare Events Datasets Using Nonparametric Density Estimation .......... 101
4.5 KDE Bandwidth Selection .................................. 112
4.6 Rare Events Classifier Approximation ...................... 120
4.7 Closing Remarks ............................................ 132

Bibliography ............................................ 133

A Chapter 3 Appendix ........................................ 138
A.1 Response Variable Reference .............................. 138
A.2 Predictor Variables Reference ............................. 146
A.3 Box Plots of PFAILURE by Predictor Variable Deciles .............. 154
A.4 Monthly Failure Prediction Summary Table (July 2009 - December 2014) .............. 159
A.5 Bailout Events ............................................ 160

B Chapter 4 Appendix ........................................ 161
B.1 Class Conditional Histograms Using Datasets MTA and MME, All Predictors .............. 162
B.2 Scatter Plots of Dataset MME Using KDE Transform (with Tuning Parameter \( \lambda = 0.00001 \)), All Predictors ............................................. 170
# List of Figures

1.1 Histograms of Tracking Error for Small Tracking Error Group .......................... 6  
1.2 Histograms of Tracking Error for Large Tracking Error Group .......................... 7  
1.3 Leveraged Daily vs ............................................................. 11  
1.4 Leveraged Daily vs ............................................................. 13  
1.5 Quantile-Quantile Plots for Daily SPXL Returns ............................................. 16  
1.6 ACF and PACF Plots for Squared Daily SPXL Returns ....................................... 16  
1.7 Index Returns vs ........................................................................ 20  
1.8 Index Returns vs ........................................................................ 21  
1.9 Various Return Series over 252 Day Periods of High and Low Volatility ............... 23  
1.10 Two Day Samples with Long Side SMC Greater Than Short Side SMC .................. 25  
1.11 SN2 vs .................................................................................. 30  
1.12 Rankings for Long, Short LETF Tickers by SN2 and SMC ................................. 31  

2.1 Scatter Plots of 252 Day Returns for hypothetical LETF “+3x SPX” and SPXL .......... 35  
2.2 Densities of 252 Day Returns for LETF Tickers RUSS and TECL and their respective indexes compounded daily with -3x leverage ................................................................. 38  
2.3 Densities of 252 Day SMC for LETF Tickers RUSS and TECL and their respective indexes compounded daily with -3x leverage ................................................................. 38  
2.4 Lag Testing Results Summary for LETF Ticker SPXL with 252 Day Periods .......... 47  
2.5 21 Day SMC Densities for SPX ........................................................................ 53  
2.6 252 Day SMC Densities for SPX ........................................................................ 54  
2.7 Simulated 22 Day SMC Density for LETF Ticker TNA, with all underlying index returns sequences compounding to 8.92% .......................................................... 55  

3.1 Firm-Month Observations as of Year Start (MTA Dataset) ................................. 63  
3.2 Campbell Table 1 [Campbell et al., 2011] .......................................................... 64  
3.3 Campbell Table 2 [Campbell et al., 2011] .......................................................... 66  
3.4 Actual and Predicted Failure Rates (%) By Year (MTA Dataset) .......................... 67  
3.5 Campbell Figure 2 [Campbell et al., 2011] .......................................................... 67  
3.6 Actual and Predicted Failure Rates (%) By Month (MTA Dataset, Testing Data Only, Rolling Basis) .......................................................... 70  
3.7 Box Plots of PFAILURE by PFAILURE Deciles ................................................ 72  
3.8 Estimated Density for PFAILURE (Non-Failure group) ...................................... 73  
3.9 Estimated Density for PFAILURE (Failure group) ............................................. 73
3.10 Actual and Predicted Failure Rates (%) By Month (Testing Data Only, Rolling Basis), with Bailout Events ................................................................. 76

4.1 Class Conditional Histogram for Earnings Yield (log scale) ........................................ 87
4.2 Class Conditional Histogram for Quantity ANI / MarketEquity (log scale) ..................... 87
4.3 Class Conditional Histogram for Quantity CashAndShortTermInvestments / BookLiabilities (log scale) .......................................................... 91
4.4 Class Conditional Histogram for Quantity CashAndShortTermInvestments / MarketEquity (log scale) ........................................................ 91
4.5 Class Conditional Histogram for Quantity BookEquity over MarketEquity (log scale) ....... 94
4.6 Class Conditional Histogram for Quantity ABE / MarketEquity (log scale) .................... 94
4.7 Monotonicity of MME Dataset (2014 only) ............................................................... 98
4.8 Monotonicity of MME Dataset (2014 only) with Annotations ........................................ 98
4.9 MME Dataset (all years), Excess Return and Capitalization Predictors ............................ 99
4.10 MME Dataset (all years), Profitability and Book Value Predictors .................................. 100
4.11 Distribution of Latent States At Various Scott Rule Multipliers ..................................... 118
4.12 Linear Regression Coefficients From Transformed Dataset At Various Scott Rule Multipliers 119
4.13 Linear Regression Coefficients From Transformed Dataset At Various Scott Rule Multipliers (Zoomed In) .............................................................. 119
4.15 Model Slices at Various Sample Quantiles (Excess Return Predictor, Logit Scores) ........ 124
4.16 Model Slices at Various Sample Quantiles (Excess Return Predictor, Probabilities) ......... 124
4.17 Model Slices at Various Sample Quantiles (Capitalization Predictor, Logit Scores) ........... 125
4.18 Model Slices at Various Sample Quantiles (Capitalization Predictor, Probabilities) .......... 125
4.19 Test Set Results, Various Methods (July 2007 - December 2014, Monthly Re-estimation) ... 129
4.20 Test Set Results, Various Methods (July 2007 - December 2014, Monthly Re-estimation) ... 129

A.1 Box Plots of PFAILURE by NIMTAAVG Deciles .................................................. 155
A.2 Box Plots of PFAILURE by TLMTA Deciles ...................................................... 155
A.3 Box Plots of PFAILURE by CASHMTA Deciles ................................................... 156
A.4 Box Plots of PFAILURE by EXRETAvg Deciles .................................................. 156
A.5 Box Plots of PFAILURE by SIGMA Deciles ...................................................... 157
A.6 Box Plots of PFAILURE by RSIZE Deciles ...................................................... 157
A.7 Box Plots of PFAILURE by MB Deciles ......................................................... 158
A.8 Box Plots of PFAILURE by PRICE Deciles ..................................................... 158

B.1 Class Conditional Histogram for Variable NIMTAAVG, MTA Dataset ......................... 162
B.2 Class Conditional Histogram for Variable ANIMEAVG, MME Dataset ......................... 162
B.3 Class Conditional Histogram for Variable TLMTA, MTA Dataset ............................. 163
B.4 Class Conditional Histogram for Variable TLME, MME Dataset .............................. 163
B.5 Class Conditional Histogram for Variable CASHMTA, MTA Dataset .......................... 164
B.6 Class Conditional Histogram for Variable CASHME, MME Dataset .......................... 164
B.7 Class Conditional Histogram for Variable EXRETAVG, MTA Dataset .......................... 165
B.8 Class Conditional Histogram for Variable EXRETAVG, MME Dataset .......................... 165
B.9 Class Conditional Histogram for Variable SIGMA, MTA Dataset ................................ 166
B.10 Class Conditional Histogram for Variable SIGMA, MME Dataset .............................. 166
B.11 Class Conditional Histogram for Variable RSIZE, MTA Dataset ................................. 167
B.12 Class Conditional Histogram for Variable RSIZE, MME Dataset ................................. 167
B.13 Class Conditional Histogram for Variable MB, MTA Dataset .................................... 168
B.14 Class Conditional Histogram for Variable ABEME, MME Dataset ............................... 168
B.15 Class Conditional Histogram for Variable PRICE, MTA Dataset .................................. 169
B.16 Class Conditional Histogram for Variable PRICE, MME Dataset ............................... 169
B.17 Scatter Plot for Variable ANIMEAVG and Latent State \( \pi \) ........................................ 170
B.18 Scatter Plot for Variable ANIMEAVG and Latent State \( \pi \) (probit link) ......................... 170
B.19 Scatter Plot for Variable TLME and Latent State \( \pi \) ............................................. 171
B.20 Scatter Plot for Variable TLME and Latent State \( \pi \) (probit link) .............................. 171
B.21 Scatter Plot for Variable CASHME and Latent State \( \pi \) ........................................ 172
B.22 Scatter Plot for Variable CASHME and Latent State \( \pi \) (probit link) ......................... 172
B.23 Scatter Plot for Variable EXRETAVG and Latent State \( \pi \) ........................................ 173
B.24 Scatter Plot for Variable EXRETAVG and Latent State \( \pi \) (probit link) ....................... 173
B.25 Scatter Plot for Variable SIGMA and Latent State \( \pi \) ........................................ 174
B.26 Scatter Plot for Variable SIGMA and Latent State \( \pi \) (probit link) .............................. 174
B.27 Scatter Plot for Variable RSIZE and Latent State \( \pi \) ........................................ 175
B.28 Scatter Plot for Variable RSIZE and Latent State \( \pi \) (probit link) .............................. 175
B.29 Scatter Plot for Variable ABEME and Latent State \( \pi \) ........................................ 176
B.30 Scatter Plot for Variable ABEME and Latent State \( \pi \) (probit link) ......................... 176
B.31 Scatter Plot for Variable PRICE and Latent State \( \pi \) ........................................ 177
B.32 Scatter Plot for Variable PRICE and Latent State \( \pi \) (probit link) .............................. 177
## List of Tables

1.1 Underlying Indexes for Demonstration Set of LETFs .................................. 4  
1.2 Demonstration Set of Selected Pairs of LETFs ........................................... 4  
1.3 252 Day Samples with nearly identical observed SN2 .................................. 27  
1.4 252 Day Samples with nearly identical observed SN2 .................................. 27  
1.5 SN2 and SMC (252 Day Periods) Summary Statistics for Long LETF Tickers .......... 31  
1.6 SN2 and SMC (252 Day Periods) Summary Statistics for Short LETF Tickers .......... 31  

2.1 Final Lag Selections for All Demonstration LETFs ........................................ 45  
2.2 Execution Times from 100 iterations per parameter set (seconds) ....................... 52  

3.1 Summary Statistics (MTA Dataset 1963 - 2008) ........................................... 64  
3.2 Summary Statistics (MTA Dataset 1963 - 2008, Failures Only) .......................... 64  
3.3 Estimated Logistic Regression Coefficients (MTA Dataset 1963 - 2008) ................. 66  
3.4 Monthly Failure Prediction Summary (MTA Dataset, Testing Data Only, Rolling Basis) ... 71  
3.5 MTA Dataset Pre and Post Bailout Events Summary ....................................... 75  
3.6 MTA Dataset Pre and Post Bailout Events Changes ....................................... 75  

4.1 Bandwidth Validation Metrics Using Full Sample 1986-2014 .......................... 118  

A.1 Monthly Failure Prediction Summary (MTA Dataset, Testing Data Only, Rolling Basis) ... 159  
A.2 Bailout Event Detail (excludes regional banks with firm caps. under $1Bill.) ............... 160
Part I

LETF Returns Dataset
Chapter 1

Dynamics of LETF Returns and A New Measure of Volatility

[Ginley, 2015]
1.1 Universe of LETFs

The first Exchange Traded Fund (ETF) began trading in 1993, but by 2015 there are over 1,400 ETFs from over 15 different issuers, with over $1 trillion invested in ETFs globally [Simpson, 2011]. The existing universe of ETFs covers all major asset classes such as equities, debt, commodities, and real estate, and many ETFs are referenced as the underlying asset of actively traded options. We will focus on a specific subset of ETFs called Leveraged ETFs, which feature a leverage ratio different than 1:1 as part of their fund structure.

The mandate for LETFs is to provide the daily return of an underlying asset (most often an equity index), at the multiple indicated by the leverage ratio. As of 2015 there are fund offerings featuring leverage ratios of 2:1, -2:1, 3:1, and -3:1. Going forward, we will simply refer to leverage by its multiple (i.e. 2, -2, 3, -3). For example, if the underlying index returned 1% on the day, then LETFs with leverage multiples of 2, -2, 3, and -3 would have a mandate to return 2%, -2%, 3%, and -3% on the day, respectively. Positive values indicate long market exposure, while negative multiples indicate short exposure, or betting that the market will decline. Funds that feature negative leverage multiples are often times referred to as Inverse ETFs, but we will not make this distinction because the returns dynamics are the same except for the obvious negative sign on the leverage term.

In addition to the included leverage, providing investors with the ability to easily acquire short exposure is an important selling point for LETF issuers. Before these LETFs started trading in 2006 leveraging one’s assets or betting against the market required utilizing expensive margin and collateral accounts or navigating the challenges involved with trading options [Yates, 2007]. Investors have since grown to appreciate the conveniences offered by the LETF structure and as of 2015 there are over 100 funds with a total notional value over $20 billion listed on exchanges worldwide, with Direxion, ProShares, Credit Suisse, and Barclays issuing the majority of them [ETF, 2016]. In Tables 1.1 and 1.1, we have listed our wide ranging demonstration set of 10 equity indexes referenced as the underlyer by 10 pairs of long and short triple levered ETFs (10 funds with a +3 leverage multiple and 10 with a -3 multiple) issued by Direxion and ProShares. Our set includes established U.S. market indexes (S&P 500, Dow Jones Industrial Average, Russell 2000, NASDAQ 100), sector specific indexes (NYSE ARCA Gold Miners, Dow Jones U.S. Financials, S&P Technology Select), and international equity indexes (MSCI Developed Markets, MSCI Emerging Markets, and Market Vectors Russia). All the daily returns data for the 10 indexes and 20 LETFs in the demonstration set are sourced from Bloomberg and the LETF leverage and fees constants are sourced in their fund prospectuses provided by their respective issuers [Dir, 2016, Pro, 2016, Bloomberg, 2014].

As their mandate is to provide a multiple of the daily return of the underlying asset, LETFs are intended for short term hedging of market risk. However, some investors have created buy and hold strategies that incorporate long term positions in LETFs [Fisher, 2014]. A primary motivation for our work was what we perceive to be a lack of understanding of the risks involved with holding LETF positions. As we shall demonstrate later on, the daily compounding of leveraged returns produces much greater volatility than a simple leveraged investment in the index without daily compounding.
Table 1.1: Underlying Indexes for Demonstration Set of LETFs

<table>
<thead>
<tr>
<th>Index</th>
<th>Issuer</th>
<th>Leverage</th>
<th>LETF</th>
<th>LETF Name</th>
<th>LETF Start</th>
<th>Obs.</th>
</tr>
</thead>
<tbody>
<tr>
<td>DJUSFN</td>
<td>ProShares</td>
<td>+3</td>
<td>FINU</td>
<td>UltraPro Financials</td>
<td>2012-07-13</td>
<td>599</td>
</tr>
<tr>
<td>DJUSFN</td>
<td>ProShares</td>
<td>-3</td>
<td>FINZ</td>
<td>UltraPro Short Financials</td>
<td>2012-07-13</td>
<td>391</td>
</tr>
<tr>
<td>GDM</td>
<td>Direxion</td>
<td>+3</td>
<td>NUGT</td>
<td>Daily Gold Miners Bull 3x Shares</td>
<td>2010-12-09</td>
<td>1013</td>
</tr>
<tr>
<td>GDM</td>
<td>Direxion</td>
<td>-3</td>
<td>DUST</td>
<td>Daily Gold Miners Bear 3x Shares</td>
<td>2010-12-09</td>
<td>1013</td>
</tr>
<tr>
<td>INDU</td>
<td>ProShares</td>
<td>+3</td>
<td>UDOM</td>
<td>UltraPro Dow30</td>
<td>2010-02-12</td>
<td>1221</td>
</tr>
<tr>
<td>INDU</td>
<td>ProShares</td>
<td>-3</td>
<td>SDOH</td>
<td>UltraPro Short Dow30</td>
<td>2010-02-12</td>
<td>1221</td>
</tr>
<tr>
<td>IXT</td>
<td>Direxion</td>
<td>+3</td>
<td>TECL</td>
<td>Daily Technology Bull 3x Shares</td>
<td>2008-12-18</td>
<td>1510</td>
</tr>
<tr>
<td>IXT</td>
<td>Direxion</td>
<td>-3</td>
<td>TECS</td>
<td>Daily Technology Bear 3x Shares</td>
<td>2008-12-18</td>
<td>1510</td>
</tr>
<tr>
<td>MVRSX</td>
<td>Direxion</td>
<td>+3</td>
<td>RUSL</td>
<td>Daily Russia Bull 3x Shares</td>
<td>2011-05-26</td>
<td>897</td>
</tr>
<tr>
<td>MVRSX</td>
<td>Direxion</td>
<td>-3</td>
<td>RUSZ</td>
<td>Daily Russia Bear 3x Shares</td>
<td>2011-05-26</td>
<td>886</td>
</tr>
<tr>
<td>MXEA</td>
<td>Direxion</td>
<td>+3</td>
<td>DZKR</td>
<td>Daily Developed Markets Bull 3x Shares</td>
<td>2008-12-18</td>
<td>1510</td>
</tr>
<tr>
<td>MXEA</td>
<td>Direxion</td>
<td>-3</td>
<td>DPXK</td>
<td>Daily Developed Markets Bear 3x Shares</td>
<td>2008-12-18</td>
<td>1509</td>
</tr>
<tr>
<td>MXEF</td>
<td>Direxion</td>
<td>+3</td>
<td>EDC</td>
<td>Daily Emerging Markets Bull 3x Shares</td>
<td>2008-12-18</td>
<td>1510</td>
</tr>
<tr>
<td>MXEF</td>
<td>Direxion</td>
<td>-3</td>
<td>EDZ</td>
<td>Daily Emerging Markets Bear 3x Shares</td>
<td>2008-12-18</td>
<td>1510</td>
</tr>
<tr>
<td>NDX</td>
<td>ProShares</td>
<td>+3</td>
<td>TQQQ</td>
<td>UltraPro QQQ</td>
<td>2010-02-12</td>
<td>1221</td>
</tr>
<tr>
<td>NDX</td>
<td>ProShares</td>
<td>-3</td>
<td>SQQQ</td>
<td>UltraPro Short QQQ</td>
<td>2010-02-12</td>
<td>1221</td>
</tr>
<tr>
<td>RTY</td>
<td>Direxion</td>
<td>+3</td>
<td>TNA</td>
<td>Daily Small Cap Bull 3x Shares</td>
<td>2008-11-06</td>
<td>1539</td>
</tr>
<tr>
<td>RTY</td>
<td>Direxion</td>
<td>-3</td>
<td>TZA</td>
<td>Daily Small Cap Bear 3x Shares</td>
<td>2008-11-06</td>
<td>1539</td>
</tr>
<tr>
<td>SPX</td>
<td>Direxion</td>
<td>+3</td>
<td>SPXL</td>
<td>Daily S&amp;P 500 Bull 3x Shares</td>
<td>2008-11-06</td>
<td>1539</td>
</tr>
<tr>
<td>SPX</td>
<td>Direxion</td>
<td>-3</td>
<td>SPXS</td>
<td>Daily S&amp;P 500 Bear 3x Shares</td>
<td>2008-11-06</td>
<td>1539</td>
</tr>
</tbody>
</table>

Table 1.2: Demonstration Set of Selected Pairs of LETFs

1.2 LETF Returns

We decompose daily LETF returns as the time series

\[ R_{LETF,t} = \beta_{LETF} R_{Index,t} - MgmtFee_{LETF} + \epsilon_{LETF,t} \]

where \( \beta_{LETF} \) is the specified leverage multiple (constant), \( MgmtFee_{LETF} \) is the specified rate of daily management expenses (constant), \( \epsilon_{LETF,t} \) is the tracking error on observation day \( t \) (random variable), and \( R_{Index,t} \) is the total return of underlying index on day \( t \) (random variable).

The leverage multiple and annualized management expenses are indicated in a given LETF’s fund prospectus. As mentioned, as of 2015 there are commercially available LETFs with leverage multiples of 2, -2, 3, and -3. All of the management fees indicated on the LETFs in our demonstration set were listed as 0.95% annually, or a daily rate of approximately 0.375 basis points [SPX, 2016b] [SPX, 2016a] [Pro, 2016]. Fees for other LETFs from Direxion and ProShares or other fund issuers are the same or comparable [ETF, 2016].

The daily index return and LETF return usually refer to a total return, or price return in addition to daily dividend yield. Fortunately, we do not concern ourselves with the calculations involved or the dates of LETF share splits because the Bloomberg Terminal can directly provide a daily returns time series. All
of our work involves only the daily returns series and the leverage and fees constants indicated in fund prospectuses.

We include Figures 1.1 and 1.2 to illustrate the sampling distributions for the tracking errors ($\epsilon_{LETF,t}$) for our demonstration set of funds. The 20 funds have been split into a large tracking error group and a small tracking error group for presentation purposes. The distinction is based on having a min (or max) that is less than -0.025 (or greater than 0.025), or a sample standard deviation greater than 0.01. While we hope this distinction highlights the scale of possible tracking errors experienced by LETFs, we do not use this distinction anywhere in our modeling. It is merely used as a rule of thumb for arranging the plots over multiple pages.

Overall, it appears as if the LETFs in our set do a reasonable job of hitting their respective performance mandates. With the exception of LETF Tickers RUSS and FINZ, all of the distributions have a sample mean of -0.0002 or less (-0.02% or less). For the small tracking error group, 4 out of 10 have a sample mean of 0. The LETFs with the longest returns history are SPXL and SPXS, the long and short pair tracking the S&P 500 index issued by Direxion, for which we have data going back to November 2008. It is interesting to note that for the 3 pairs of LETFs in the large tracking error group that represent the MVRSX, MXEA, and MXEF indexes, tracking errors in excess of +/- 5% are not infrequent. One might suppose that errors of this magnitude would be considered unacceptable by the investing public, but that is not the case. As of early 2015 all of the LETFs in our demonstration set are still actively traded.
Figure 1.1: Histograms of Tracking Error for Small Tracking Error Group
Figure 1.2: Histograms of Tracking Error for Large Tracking Error Group
1.3 Volatility Drag and Convexity

For the remainder of this section let us ignore specific instruments, fund management fees, and tracking errors, and concern ourselves only with the effects of leveraged compounding of daily returns. Before proceeding with LETF return dynamics, we need to define the concept of volatility drag. Traditionally, this refers to the dynamic where the geometric return for the period is less than the arithmetic return [Bouchey and Nemtchinov, 2013]. This inequality,

\[
\prod_{j=1}^{p} (1 + R_j) - 1 < \sum_{j=1}^{p} R_j
\]  

(1.2)

can be easily understood by example. Say an index experiences a 1% loss followed by a 1% gain, then the final compound return (geometric) observed will be negative, while the arithmetic return is clearly 0,

\[
((1 - 0.01) * (1 + 0.01)) - 1 = -0.0001 < 0 = -0.01 + 0.01
\]  

(1.3)

This condition can flip so that the geometric return is greater than the arithmetic return.

\[
\prod_{j=1}^{p} (1 + R_j) - 1 > \sum_{j=1}^{p} R_j
\]  

(1.4)

When this relationship holds, we call it “convexity”. This is the result of a sequence of repeated gains (or losses) that compound to a return that is greater than the arithmetic return. The convexity dynamic can also be understood as repeatedly “buying high” and increasing your exposure during a sequence of positive returns, or repeatedly “selling low” and decreasing your exposure during a sequence of negative returns [Tuzun, 2012]. Let us change our example to be 2 consecutive days of 1% gains or 1% losses.

\[
((1 + 0.01) * (1 + 0.01)) - 1 = 0.0201 > 0.02 = 0.01 + 0.01
\]  

(1.5)

\[
((1 - 0.01) * (1 - 0.01)) - 1 = -0.0199 > -0.02 = -0.01 + -0.01
\]  

(1.6)

For longer time series, for convexity to be observed, it is not required that all returns are gains (or losses), only that sufficiently many of them are the same sign so that the cross products of the polynomial expansion on the left hand side are positive, and thus our inequality for the dynamic of convexity would hold.

Now we will extend these two conditions so that they are more suitable when used in the context of LETFs. Let \( D \) be the following function ...

\[
D(R|\beta) = \left[ \prod_{j=1}^{p} (1 + \beta R_j) - 1 \right] - \left[ \beta(\prod_{j=1}^{p} (1 + R_j) - 1) \right]
\]  

(1.7)

The first term is the \( p \) day return produced by leveraged daily returns–what is achieved with the common LETF structure ignoring fees and tracking error. The second term is the leveraged period return–what is achieved if an investor leveraged an initial capital investment in an index by a factor of \( \beta \), and ignored margin (borrowing) costs and tracking error. Our rationale for why the second term is not the leveraged arithmetic return \( \beta \sum R_j \) will be clarified after we consider the expansion of both products into polynomials.
\[ D(R|\beta) = \left[ \prod_{j=1}^{p} (1 + \beta R_j) - 1 \right] - \left[ \beta \left( \prod_{j=1}^{p} (1 + R_j) - 1 \right) \right] \] 
(1.8)

\[ = \sum_{g=0}^{p} \left( \sum_{h=1}^{p} \left( (\beta R_j)^{I_A(p,g,h,j)} \right) \right) - \beta \sum_{g=0}^{p} \left[ \sum_{h=1}^{p} \left( R_j^{I_A(p,g,h,j)} \right) \right] \] 
(1.9)

\[ = \sum_{g=0}^{p} \beta^g \left[ \sum_{h=1}^{p} \left( R_j^{I_A(p,g,h,j)} \right) \right] - \sum_{g=0}^{p} \beta \left[ \sum_{h=1}^{p} \left( R_j^{I_A(p,g,h,j)} \right) \right] \] 
(1.10)

\[ = \sum_{g=0}^{p} (\beta^g - \beta) \left[ \sum_{h=1}^{p} \left( R_j^{I_A(p,g,h,j)} \right) \right] \] 
(1.11)

\[ = \sum_{g=2}^{p} (\beta^g - \beta) \left[ \sum_{h=1}^{p} \left( R_j^{I_A(p,g,h,j)} \right) \right] \] 
(1.12)

where \( I_A(p,g,h,j) \) is the indicator function for set \( A \), which we are using as a convenience to hide the powers of all the \( R_j \) within each individual term in the summation. The specific values are not important for our discussion. We only need to discuss the coefficient on the inner summation containing \( \beta \). If we were using the leveraged arithmetic return \( \beta \sum_{j=1}^{p} R_j \) in our derivation, the \( (\beta^g - \beta) \) coefficient would instead be only \( \beta^g \).

In the traditional case when there is no change in leverage the interpretation of the volatility drag condition that is based on the arithmetic return must be used. Clearly, setting \( \beta = 1 \) would make the coefficient terms in our interpretation all zeros, thus the final value of \( D \) would always be a noninformative zero.

But in the presence of leverage \((|\beta| \geq 1)\), we prefer our interpretation because the arithmetic return has no real world significance. In our function \( D \), we use \( \beta (\prod_{j=1}^{p} (1 + R_j) - 1) \) instead because this is an achievable return, where an investor earns the compound return for the period on a capital base that was leveraged \( \beta \) times. We feel this is a more meaningful calculation for determining volatility drag versus convexity.

Also, if instead the observed values for the individual \( R_j \) variables were replaced with the geometric mean \( \tilde{R}_g = \prod_{j=1}^{p} (1 + R_j)^{1/p} - 1 \), a leveraged arithmetic return would not remain the same while our leveraged period return would produce the same value. This secondary reason for our preference will soon be a point of emphasis.

With our function \( D \) properly defined, we state our revised criteria for determining volatility drag versus convexity when observing a series of leveraged returns (i.e. assuming \(|\beta| > 1\)):

\[ D(R|\beta) < 0 \implies \text{volatility drag} \] 
(1.13)

\[ D(R|\beta) > 0 \implies \text{convexity} \] 
(1.14)

We must note that \( D \) is technically bounded from below at -1. In other words, the effect of volatility drag cannot be in excess of -100%, because we cannot lose more than 100% of our investment value. Mathematically, this would be expressed by censoring the individual \( (1 + \beta R_j) \) terms and the entire \( \beta (\prod_{j=1}^{p} (1 + R_j) - 1) \) term at 0. This lower bound still holds when \( \beta < -1 \) in our analysis of LETFs, because funds with a negative leverage multiple are still purchased assets with a Net Asset Value (NAV) and a share price, as opposed to shorting a stock where positions are initiated with sell transactions and losses have no theoretical lower bound.
There is also a conditional upper bound to our function $D$ dependent on the compound period return that is theoretical and not driven by real world constraints. This is because convexity is maximized when the daily return for all $p$ days is the geometric mean for the period. Let $\mathbf{R}$ be a $p$ dimensional vector (daily sequence) of returns, $\mathbf{1}_p$ be a $p$ dimensional vector of all 1s, $R = \prod_{j=1}^{p}(1 + R_j) - 1$ be the compounded return, and $\bar{R}_g = (1 + R)^{1/p} - 1$ be the geometric mean. We can derive the upper bound for $D,$

\[
(1 + \beta \bar{R}_g)^p - 1 \geq \prod_{j=1}^{p}(1 + \beta R_j) - 1
\]  

(1.15)

\[
[(1 + \beta \bar{R}_g)^p - 1] - [\beta R] \geq [\prod_{j=1}^{p}(1 + \beta R_j) - 1] - [\beta R]
\]

(1.16)

\[
[(1 + \beta \bar{R}_g)^p - 1] - [\beta((1 + \bar{R}_g)^p - 1)] \geq [\prod_{j=1}^{p}(1 + \beta R_j) - 1] - [\beta(\prod_{j=1}^{p}(1 + R_j) - 1)]
\]

(1.17)

\[
D(\bar{R}_g \mathbf{1}_p | \beta) \geq D(\mathbf{R} | \beta)
\]

(1.18)

It is very important to understand both input vectors $\bar{R}_g \mathbf{1}_p$ and $\mathbf{R}$ provide the same compound period return $R = \prod_{j=1}^{p}(1 + R_j) - 1,$ but clearly represent different ways to arrive at that return. This implies that regardless of the observed values of $R_j$ that compound to the observed value of $R,$ we can calculate the maximum convexity that is possible with a leveraged, daily compounded fund structure knowing only $R$ (and the parameters $p$ and $\beta$). Not coincidentally, this condition that maximizes convexity is equivalent to saying there was no volatility during the period (the return for each day is a constant $\bar{R}_g,$ i.e. no randomness). This point will become useful when we define our new volatility statistic in the next section.
Figure 1.3: Leveraged Daily vs. Leveraged Periodic Returns (21 Days)
In Figure 1.3 we visualize the return profile of leveraged returns, with and without daily compounding for various leverage multiples, which are indicated along the right edge of the plots. The column labeled “SHORT” plots returns where the sign on the leverage multiple is negative, and the “LONG” column is for positive multiples. The parameter $p$ is fixed at 21 days. Each function represents a return profile for a hypothetical leveraged instrument, given the $x$ value as the return of the hypothetical underlying index over a 21 day period, $x := \prod_{j=1}^{21} (1 + R_j) - 1$.

The functions marked as periodic leverage ($y_{\text{periodic}}$, colored green) are simply this $x$ value multiplied by the indicated leverage multiple $\beta$ (and censored at -1, as we cannot lose more than 100%). For the functions marked as daily leverage ($y_{\text{daily}}$, colored red), we compound the geometric mean return implied by the value of $x$ (leveraged $\beta$ times) over 21 days.

$$y_{\text{periodic}} = \beta x$$
$$y_{\text{daily}} = (1 + \beta \bar{R}_g)^{21} - 1$$
$$\bar{R}_g = (1 + x)^{1/21} - 1$$

$y_{\text{daily}}$ is the maximum convexity curve given $x$. From the figure, we see the curvature of the $y_{\text{daily}}$ curves increase as the magnitude of $\beta$ increases. For positive values of $x$, curves defined by larger (in magnitude) leverage multiples dominate, and for negative values of $x$, curves defined by smaller leverage multiples dominate. More generally, for a given series of returns, both volatility drag and convexity intensify as the magnitude of the leverage multiple increases. This might be better explained with a reference to the $(\beta^g - \beta)$ term in the function $D$ defined earlier. If $D(R|\beta) > 0$, then increasing $\beta$ will increase the value of $D$, and we say convexity has intensified. If $D(R|\beta) < 0$, then increasing $\beta$ will decrease the value of $D$, and we say volatility drag has intensified.

Now that we have a better understanding of the choice of $\beta$ on the maximum convexity curve, let us consider the impact of $p$, the length of the returns vector $R$ (or the number of days in the sample period). Similar to increasing $\beta$, the curvature also increases as $p$ increases, given that the compound return $R$ is held constant. However, the result of this change in $p$ is bounded as $p$ goes to infinity for a given compound return $R$ and geometric mean return $\bar{R}_g$.

$$\lim_{p \to \infty} \left(1 + \beta \bar{R}_g\right)^p - 1 = \lim_{p \to \infty} \left(1 + \beta\left([1 + R]^\frac{1}{p} - 1\right)\right)^p - 1 = (1 + R^\beta) - 1$$

Figure 1.4 illustrates this result of taking the limit of $p$ days to infinity for the maximum convexity curve for a given leverage multiple, across various time horizons and leverage multiples. We see that curves defined by greater values of $p$ (the geometric mean return implied by $p$ is compounded with leverage over $p$ days) dominate for each choice of leverage multiple (across both positive and negative values of $x := \prod_{j=1}^{21} (1 + R_j) - 1$).

$$y_{\text{periodic}} = \beta x$$
$$y_{\text{daily,5}} = (1 + \beta \bar{R}_g)^5 - 1 = (1 + \beta((1 + x)^{1/5} - 1))^5 - 1$$
$$y_{\text{daily,10}} = (1 + \beta \bar{R}_g)^{10} - 1 = (1 + \beta((1 + x)^{1/10} - 1))^{10} - 1$$
$$y_{\text{daily,\infty}} = \lim_{p \to \infty} \left(1 + \beta \bar{R}_g\right)^p - 1 = (1 + R^\beta) - 1$$
1.4 Problems with Sample Standard Deviation

Going all the way back to the Markowitz’ benchmark paper on portfolio theory, it has been commonly accepted to model the volatility of asset returns as the standard deviation of the returns distribution [Markowitz, 1952]. Accordingly, sample standard deviation is used to measure the realized volatility of a \( p \) day sample of returns \( s = \left[ \frac{1}{p-1} \sum_{j=1}^{p} (R_j - \bar{R})^2 \right]^{\frac{1}{2}} \).

The major criticism of sample standard deviation and a motivation for our approach is that the assumptions of independence (and identical distribution) and normality required for interpreting results do not hold
with respect to returns data. Consider the following traditional assumptions used to model daily, log returns:

\[ R_1, R_2, \ldots, R_p \] daily returns \tag{1.27}

\[ X_j = \log(1 + R_j) \] log, daily returns \tag{1.28}

\[ X_1, X_2, \ldots, X_p \sim \text{normal i.i.d.} \tag{1.29} \]

\[ Y = \sum_{j=1}^{p} X_j \] log, \( p \) day return \tag{1.30}

\[ \text{Var}[Y] = p \ast \text{Var}[X] \] given i.i.d. \tag{1.31}

\[ s_Y = \sqrt{p} \ast s_X = \sqrt{p} \sqrt{\frac{1}{p-1} \sum_{j=1}^{p} (x_j - \bar{x})^2} \] \tag{1.32}

\[ \approx \sqrt{\sum_{j=1}^{p} (x_j - \bar{x})^2} \] \tag{1.33}

If the assumption of i.i.d. normal data holds, then \( s_Y \) is an estimator of the standard deviation of the \( p \) day log returns distribution, and we could say that roughly 68\% of observations will fall within \( \pm s_Y \) of the mean. However, it is now commonly understood that asset returns data are not i.i.d. normal. We provide the normal Quantile-Quantile plots for the returns series of LETF ticker SPXL (the +3x LETF tracking the S&P500) in Figure 1.5 as our evidence against a normal distribution. Many of the sample quantiles of the returns data, especially noting the quantiles corresponding to the tails of the distributions, do not match the corresponding quantiles of a normal distribution. Figure 1.5 includes the ACF and PACF plots of the same data, which we present as evidence that the data is not independent. Note that lags 1 and 5 break the red lines for the 95\% confidence interval in both the ACF and PACF plots.

If we accept this evidence, then we must revise the derivation above to account for some non-normal returns distribution and the lack of independence. We will allow for the assumption that the daily data are identically distributed, however (although if this is not true either, alternatives for scaling have been proposed elsewhere [Hamidieh and Ensor, 2010]).

\[ X_1, X_2, \ldots, X_p \sim f_X \] identically distributed, but dependent \tag{1.34}

\[ Y = \sum_{j=1}^{p} X_j \] log, \( p \) day return \tag{1.35}

\[ \text{Var}[Y] = p \ast \text{Var}[X] + \sum_{i \neq j} \text{Cov}[X_i, X_j] \] \tag{1.36}

\[ s_Y = \sqrt{p \ast s_X^2 + \sum_{i \neq j} s_{X_i, X_j}^2} \neq \sqrt{p} \ast s_X \] \tag{1.37}

If the true process that is said to generate LETF returns is not normal and does not generate data independently (or at least without correlation), then the interpretation of the quantity \( \sqrt{p} \ast s_X \) is nearly meaningless. It is not an estimator of the standard deviation of the \( p \) day log return \( Y \), and even if we did calculate a proper estimate we can no longer claim that 68.27\% of observations will fall within \( \pm 1 \) standard deviation of the mean. We are left with a nonparametric interpretation that is based on the literal formula for the sample standard deviation estimator—the 2-norm of the centered observation vector of log returns.
\( (\sqrt{p} \ast s_X = \|x - \bar{x}\|_2) \). By nature of the squaring this statistic does not produce any sort of intuitive, real-world quantity. By contrast, we consider Maximum Drawdown, another commonly used statistic, as providing an intuitive, real-world quantity (a rate of return) that is immediately understood when trying to communicate risk. Without the necessary assumptions that make \( p \ast s_X^2 \) a true estimator of the population variance for some normal distribution of returns, and without a strong intuition for interpreting the 2-norm of returns samples absent any theoretical foundation, our goal is to create an improved statistic for measuring the volatility of LETF returns.
Theoretical Quantiles
Sample Quantiles

Figure 1.5: Quantile-Quantile Plots for Daily SPXL Returns

Geometric Returns
Log Returns

ACF
PACF

Figure 1.6: ACF and PACF Plots for Squared Daily SPXL Returns
1.5 Define Shortfall from Maximum Convexity

The lack of normality and independence in the distributions of LETF returns provide us with a golden opportunity to propose an alternative statistic for measuring their realized volatility based on the idea of volatility drag. We feel our proposed measure has an interpretation that is more intuitive than that of standard deviation and because it allows for greater variability (when compared to standard deviation on the same dataset), we feel that it is more informative as well. It is a hybrid summary of returns and volatility where the values are returns but their interpretation is to quantify the level of risk associated with the returns. Let us return to the concept of maximum convexity that was defined earlier in Section 1.3 and use it as the foundation for defining our new statistic.

For a given compound return \( R \) on an asset over a period of \( p \) days, the hypothetical leveraged, daily compounded return derived from that asset’s daily returns \( R_j \)’s has an upper bound, which we shall denote as \( R_{MAX} \). This return is achieved if the geometric mean of the return series is observed every day, such that the value is dependent on the choice of leverage multiple \( \beta \), but the geometric mean of the asset’s return series is not.

\[
R_{Index} = \prod_{j=1}^{p} (1 + R_j) - 1 \quad (1.38)
\]

\[
\bar{R}_{Index} = (1 + R_{Index})^{\frac{1}{p}} - 1 \quad (1.39)
\]

\[
R_{MAX} = (1 + \beta \bar{R}_{Index})^p - 1 \quad (1.40)
\]

where \( R_1, R_2, ..., R_p \) is a series of \( p \) daily total returns on an index, and \( \beta \) is the indicated leverage multiple. We say that given the leverage multiple, \( R_{MAX} \) is the maximum return that could be achieved by a series of \( p \) daily returns that compounds to \( R_{Index} \), and therefore also maximizes the degree of dconvexity as was defined in the previous section.

\[
R_{MAX} = (1 + \beta \bar{R}_{Index})^p - 1 \geq \prod_{j=1}^{p} (1 + \beta R_j) - 1, \forall R : \prod_{j=1}^{p} (1 + R_j) - 1 = R_{Index} \quad (1.41)
\]

\[
\Rightarrow R_{MAX} - \beta \bar{R}_{Index} \geq \left[ \prod_{j=1}^{p} (1 + \beta R_j) - 1 \right] - \beta R_{Index} \quad (1.42)
\]

The interpretation of maximum convexity is that the effect of compounding gains on gains (or losses on losses) over the \( p \) day period could produce a return up to \( R_{MAX} \), which by definition is greater than or equal to \( \beta R_{Index} \) (the result of earning the index return on a leveraged capital base).

We are now ready to define our volatility statistic that was designed specifically for LETFs, Shortfall from Maximum Convexity, which we shall abbreviate as SMC. This new statistic is defined to be the geometric excess return of the maximum return \( R_{MAX} \) with respect to the LETF return observed over the \( p \) day period.

For a series of \( p \) daily LETF returns \( R_{LETF,1}, R_{LETF,2}, ..., R_{LETF,p} \), the series of \( p \) daily total returns on the corresponding underlying index \( R_1, R_2, ..., R_p \), and a stated leverage multiple of \( \beta \) on the LETF, the
SMC for the period is

\[
R_{LETF} = \prod_{j=1}^{p} (1 + R_{LETF,j}) - 1
\]  (1.43)

\[
R_{MAX} = (1 + SMC)(1 + R_{LETF}) - 1
\]  (1.44)

\[
SMC(R_{LETF}, R_{Index}|\beta) = \frac{(1 + R_{MAX})}{(1 + R_{LETF})} - 1
\]  (1.45)

where \(R_{MAX}\) is defined as above using the given index returns along with the stated leverage multiple for the LETF.

SMC values are interpreted as the performance shortfall from the hypothetical return \(R_{MAX}\). To help demonstrate, we have included Figure 1.7, a visual example of calculating SMC for a single 252 day period (tick marks on axes and stated statistics are in decimal form). The single green point is a hypothetical 252 day observation when a +3x LETF returned 50% and the underlying index returned 35%. The blue curve represents \(R_{MAX}\) at the given index return, and the purple line represents the given index return multiplied by +3 (and bounded at -1, or -100%). For our hypothetical observation with an index return of 35%, the corresponding point on the \(R_{MAX}\) curve is 145.77%, which results in a final SMC value of 63.85%. Ignoring future movement of the underlying index, the LETF would have to return 63.85% to match the maximum return that was achievable for this 252 day period.

Figure 1.8 is essentially the same as Figure 1.7, but using real data from the ±3x LETF ticker pairs with underlying indexes SPX and RTY (S&P500 and Russell 2000) over 252 day periods. We make a special note that visuals such as this one were instrumental in our LETF analysis and played a significant role in defining SMC. The specific choices of indexes are not important, only that we selected a pair where it is commonly accepted that one is riskier than the other. In our case, we view the S&P500 as less risky than the Russell 2000. Each green point marks an observed return for the specified LETF (as indicated by index ticker and side) and number of days, and as before the blue curve represents \(R_{MAX}\) at a given index return and the purple line represents the index return multiplied by +3 (long side) or -3 (short side). The key observation is that the green points seem to disperse away from the blue curve as we consider an LETF (on the same side) based on a riskier index. Our intuition was that greater dispersion is indicative of greater volatility, and SMC was defined specifically to measure that dispersion.

Our definition of SMC is not yet final as we must properly account for sample periods when the observed LETF return is greater than \(R_{MAX}\). We will observe this result in the repeated presence of abnormally large, positive tracking errors. In our data set containing 20 LETFs this was observed in about 1% of all 252 day periods. Because of this possibility, we will include the use of a max function in a revised definition for SMC.

\[
SMC(R_{LETF}, R_{Index}|\beta) = \max\{0, \frac{(1 + R_{MAX})}{(1 + R_{LETF})} - 1\}
\]  (1.46)

Alternatively, we could add the maximum observed tracking error to the \(\beta\bar{R}_{Index}\) term before calculating \(R_{MAX} = (1 + \beta\bar{R}_{Index})^p - 1\), or simply allow for negative values of SMC. These options would generate results that are different than what is presented at the end of this section (average SMC values would increase slightly if incorporating the maximum tracking error, or decrease slightly if allowing negative values), otherwise applications of our new statistic would remain unchanged. We choose to censor at 0 because it is simpler than computing a maximum tracking error and lets us avoid any confusion that may result from interpreting negative values.

SMC not only captures the impact of volatility drag produced by the daily compounding of leveraged
returns, but also the impact produced by the compounding of tracking errors and fees. We demonstrate this and the dynamic nature of the relationship between $R_{MAX}$ and $R_{LETF}$ with 2 examples of calculating SMC over 252 day periods in Figure 1.9. Following the color scheme from previous figures, the blue line represents $R_{MAX}$ at each day, the purple line is the leverage multiple times the index return up to that date, and the green line is the observed LETF return. An orange line is also included, which is the underlying index return compounded daily at the given leverage multiple, or what the LETF return should be without fees or tracking errors.

The scenario in the top panel is for the period ending 12/31/2013, for the +3x LETF ticker UDOW (underlyer Dow Jones Industrials). We included this scenario to demonstrate the advantage of leveraged, daily compounding of returns experienced in a low volatility environment. The green line for the observed return on UDOW finishes just below the blue line for $R_{MAX}$, and we observe a low SMC of 0.0515. This means that UDOW came close to achieving the maximum potential return over the period given the leverage multiple of +3 and the performance of the index, and because TZA outperformed the 3 times the Dow Jones return, we have observed the convexity dynamic. By the end of the period we see some separation between the orange line (theoretical LETF without fees or tracking errors) and green line, which tells us that fees and tracking error were partially responsible for the performance shortfall.
$SMC = \frac{(1 + R_{MAX})}{(1 + R_{LETF})} - 1$

$SMC = 0.6385$

$R_{LETF} = 0.5$

$R_{MAX} = 1.4577$

$3 \times R_{Index} = 1.05$

Figure 1.7: Index Returns vs. LETF Returns for Hypothetical Index and $+3x$ LETF over 252 Day Periods
Figure 1.8: Index Returns vs. LETF Returns for LETF Ticker Pairs on SPX, RTY over 252 Day Periods
The scenario in the bottom panel is for the period ending 11/30/2009, for the -3x LETF ticker TZA (underlyer Russell 2000). This scenario is included to demonstrate the disadvantage of leveraged, daily compounding of returns experienced in a high volatility environment. Here, the green line finishes further below the blue line than in the first scenario, and we observe a high SMC of 1.8651. Because the purple line (-3 times the Russell 2000 performance) ends the period above the green line, we have observed the dynamic of volatility drag. TZA vastly underperformed the maximum potential return, and because of the minimal separation between orange line and green line (observed TZA performance), we can say that the shortfall was caused by the volatility drag associated with daily, leveraged compounding and not necessarily the accumulated drag associated with fees or tracking error.

We close our introduction of SMC with a discussion on the perceived asymmetry of results when using the statistic with a long and short LETF pair over the same period. At first glance, it may appear as though the statistic is biased to produce greater values with short side LETF returns data ($\beta < -1$). It occurs in the overwhelming majority of periods available in our data set, but we show this does not have to be true because of the counter examples that do exist. We are specifically referring to the following relationship, where we consider a pair of long and short LETFs on the same underlying index with equal but opposite leverage multiples, and no tracking errors or fees.

$$SMC(\beta R, R|\beta) \geq SMC(-\beta R, R| -\beta)$$ (1.47)

where $R$ is a $p$ dimensional vector of daily total returns on an index.

We initially attempted a proof based on Jensen’s inequality showing that short side SMC is guaranteed to be greater. The set up for that proof proceeds as follows. For a given leverage multiple $\beta$, and a series of $p$ daily total returns on an index $R_1, R_2, ..., R_p$, let $R = \prod_{j=1}^p (1 + R_j) - 1$ and $\bar{R} = (1 + R)^\frac{1}{p} - 1$. Also, require that $-1 \leq \beta R_j$ and $-1 \leq -\beta R_j$, because these terms represent LETF returns (without fees or tracking errors), and cannot be less than -100%. Also, the basis for our inequality is the result of changing the sign of $\beta$, so assume $\beta \geq 1$ (and $-\beta \leq -1$). Now we rearrange terms.

$$\frac{(1 + \beta \bar{R})^p}{\prod_{j=1}^p (1 + \beta R_j)} - 1 \geq \frac{(1 - \beta \bar{R})^p}{\prod_{j=1}^p (1 - \beta R_j)} - 1$$ (1.48)

$$\prod_{j=1}^p \frac{(1 - \beta R_j)^{1/p}}{(1 + \beta R_j)^{1/p}} \geq \frac{(1 - \beta \bar{R})}{(1 + \beta \bar{R})}$$ (1.50)

Let $X_j = 1 + R_j$ to guarantee non-negativity with our terms, and note that $\bar{X}_g = \prod_{j=1}^p X_j^{1/p} = 1 + \bar{R}$. We proceed by substituting terms.

$$\prod_{j=1}^p \frac{(1 - \beta X_j + \beta)}{(1 + \beta X_j - \beta)}^{1/p} \geq \frac{(1 - \beta \bar{X}_g + \beta)}{(1 + \beta \bar{X}_g - \beta)}$$ (1.51)

Next, let $f(x) = (1 - \beta x + \beta)/(1 + \beta x - \beta)$, and re-write the inequality in terms of $f$.

$$\prod_{j=1}^p f(X_j)^{1/p} \geq f(\bar{X}_g) = f(\prod_{j=1}^p X_j^{1/p})$$ (1.52)
Figure 1.9: Various Return Series over 252 Day Periods of High and Low Volatility
On the domain of $f$ allowed by our constraints for $R_j$ and $\beta$, $f$ is a convex function, and through Jensen’s inequality we can claim that

$$\frac{1}{p} \sum_{j=1}^{p} f(X_j) \geq f\left(\frac{1}{p} \sum_{j=1}^{p} X_j\right) \quad (1.53)$$

Obviously, this statement does not help our cause. If we could prove that it holds using geometric means opposed to arithmetic means, it would assert that long side SMC is always greater than or equal to short side SMC, which is not true. Perhaps more importantly, we cannot prove that it holds with geometric means. Given that $f$ is decreasing and the geometric mean of a non-negative random variable is less than or equal to the arithmetic mean, we could claim that

$$f\left(\prod_{j=1}^{p} X_j^{1/p}\right) \geq f\left(\frac{1}{p} \sum_{j=1}^{p} X_j\right) \quad (1.54)$$

$$\frac{1}{p} \sum_{j=1}^{p} f(X_j) \geq \prod_{j=1}^{p} f(X_j)^{1/p} \quad (1.55)$$

Unfortunately, neither statement helps our proof, and a log transform does not make things easier, as the transformed right hand side would not permit the use of Jensen’s inequality. At this point, we reconsidered our pursuit of a theoretical proof of the inequality and exhaustively searched the data for counter examples.

Using the entirety of our data set of 10 underlying indexes and including leverage multiples of $\pm 1, \pm 2, \pm 3$, we have 95,913 sample periods of 21 days, and 93,603 sample periods of 252 days. There were 25 instances (23 periods of 21 days, 2 periods of 252 days) when long side SMC was greater than the short side SMC (i.e. $SMC(\beta R, R | \beta) > SMC(\beta R, R | -\beta)$). Within the 25 instances, there was no evidence of necessary or sufficient conditions that might help explain the direction of the inequality. There were instances of both less than and greater than relationships in the presence of positive and negative index returns, positive and negative long and short LETF returns (i.e. $\beta R$ and $-\beta R$), and other boolean conditionals.
Figure 1.10: Two Day Samples with Long Side SMC Greater Than Short Side SMC

We can plot the region where long side SMC is greater than short side SMC, and we do this for all possible 2 day sample periods using $\beta = 2$ and $\beta = 3$ (“2x Leverage” and “3x Leverage”, respectively) in Figure 1.10. The x axis represents all possible returns for the first day in the sample, and the y axis represents the return on the second day. The red lines represent the effective boundary for daily returns allowed by the respective leverage multiples. The shaded regions represent samples that would produce a long side SMC that is greater than the short side SMC, and equality occurs along the curved boundary of the region. Technically, equality occurs along the main diagonal ($x = y$) as well, where 0 is observed on both the long and short side, but this detail is omitted for clarity. By comparing the left and right subplots, we see that the curved boundary disperses from the origin as the leverage multiple decreases, such that there is no region that exists where long side SMC is greater than short side SMC with $\beta = 1$, because in that case long side SMC is always 0 by definition. The curved boundary shifts closer towards the origin as the leverage multiple increases. As the number of days in the sample increases beyond 2, we have an analogous multidimensional surface for the boundary of equality of long and short side SMC that continues to disperse further from the origin.

Keeping these properties in mind, it becomes apparent as to why short side SMC is almost always greater in practice, given the nature of asset returns distributions. For example, for 252 day periods and $\beta = 3$, we would need to observe a sample with one day experiencing a nearly -16% return, and all other 251 days less than or equal to 0%. That kind of 1 day loss is an incredibly rare occurrence, and you would still need a strong negative trend across all other days in the sample period. Ultimately, the definition of the SMC statistic does not guarantee greater values when using a negative leverage multiple, but the support of existing index returns distributions is such that there is very low probability mass over the regions where the converse is true.
1.6 Fundamental Motivation for SMC

We have properly defined our new statistic for measuring realized volatility and now we demonstrate how it is fundamentally different from the quantity $\sqrt{p} \cdot s$, where $p$ is the chosen number of days in the sample period and $s$ is the maximum likelihood estimator (biased) of standard deviation. Going forward we will refer to this quantity as the sample 2-norm, or SN2, because of our earlier proof where we demonstrated this quantity is not actually an estimator of the standard deviation of the $p$ day log return when using correlated data. We will use log returns when calculating SN2 so that the $p$ day return is additive and the mean return represents the log of the geometric mean as opposed to arithmetic mean (although in practice not using log returns does not weaken our evidence presented later). For a $p$ day vector of LETF returns, we write the definition for SN2.

$$
\bar{R}_{LETF} = \prod_{j=1}^{p}(1 + R_{LETF,j})^{\frac{1}{p}} - 1
$$

$$
SN2(R_{LETF}) = \sqrt{p} \cdot \sqrt{\frac{1}{p} \sum_{j=1}^{p} (\log(1 + R_{LETF,j}) - \log(1 + \bar{R}_{LETF}))^2}
$$

$$
SN2(R_{LETF}) = \sqrt{\sum_{j=1}^{p} (\log(1 + R_{LETF,j}) - \log(1 + \bar{R}_{LETF}))^2}
$$

Another motivating factor for our choice of using log returns when calculating SN2 can be seen when we re-write the definition for SN2 and SMC side by side.

$$
SN2(R_{LETF}) = \sqrt{\sum_{j=1}^{p} [\log((1 + \bar{R}_{LETF})/(1 + R_{LETF,j}))]^2}
$$

$$
SMC(R_{LETF}, R_{Index}|\beta) = \exp \left[ \sum_{j=1}^{p} [\log((1 + \beta \bar{R}_{Index})/(1 + R_{LETF,j}))] \right] - 1
$$

After rearranging terms, we see that SN2 is the 2-norm of a $p$ day sample of log, geometric excess returns. SMC (without censoring) is the exponentiation of the sum of a $p$ day sample of log, geometric excess returns, but using $\beta \bar{R}_{Index}$ instead of $\bar{R}_{LETF}$ when computing the excess returns. By using log returns when calculating SN2, we are also making a fair comparison with SMC in addition to the data transformation reasons mentioned earlier.

Both SMC and SN2 have a sample distribution that is bounded on the left at 0 and extends to $+\infty$ on the right. Technically, the distribution for SMC includes $+\infty$, which is observed if the LETF return for the given sample period is $-100\%$. For both statistics, a value of 0 is observed when the LETF achieves the sample mean return for every day in the period, with the interpretation being that there was no realized volatility observed (the underlying random process produced a constant return). For SN2 the sample mean return is that of the LETF, but in the case of SMC, the sample mean return refers to the mean of the underlying index returns for the same sample period (multiplied by the stated leverage multiple of the LETF).

We acknowledge the exception that must be made for the negative SMC values that are possible before censoring because of tracking error. For those samples, censoring SMC to 0 means that because of a beneficial sequence of tracking errors, there was no shortfall from the maximum return. Also, we acknowledge that computing SMC requires the underlying index returns in addition to LETF returns.
SN2 is asymmetric with respect to the leverage parameter $\beta$, but the difference is less than the difference that results when using the same inputs with SMC (and ignoring tracking error such that for daily returns $R_{LETF} = \beta R_{Index}$). If we were not using log returns with SN2, it is trivial to check that output values are identical when using positive and negative leverage multiples ($\left[\sum_{j=1}^{p}(\beta R_j - \bar{\beta} R)^2\right]^{1/2} = \left[\sum_{j=1}^{p}(-\beta R_j + \bar{\beta} R)^2\right]^{1/2}$). With log returns, however, the daily returns are not equal ($\log(1 + \beta R_j) \neq \log(1 - \beta R_j)$), and there will be minor differences in the resulting values.

SN2 is the 2-norm of a sample of centered, log returns, and due to the squaring and square root operations, has no real-world meaning. Functionally, it measures the distance of the observed returns vector from the observed mean return. SMC is the geometric excess return of the maximum return $R_{MAX}$ with respect to the observed LETF return. It tells us the underperformance of the LETF relative to a hypothetical underlying index that provides constant returns (and still compounds to the total that was observed).

To better appreciate the different interpretations of these two statistics, let us refer to some examples from our dataset. In Table 1.3 we have two 252 day samples with nearly identical observed SN2s. Notice the substantial difference in observed SMC, however. For the TNA sample ending 2/2/2009, the underlying index (Russell 2000) lost about 10%. Based on this return the hypothetical best performance of TNA for that same period would be the $R_{MAX}$ of -28.6%. The observed return on TNA for that sample period was extremely close, resulting in a very low observed SMC of 2.27%. Contrast this with the sample period ending 12/31/2008 that observed a nearly identical SN2 as the TNA sample (0.4220 compared to 0.4148). Here the index returned 20% which implies the hypothetical best performance of TNA was 72.01%. TNA performed much worse than this maximum (51.84%) resulting in an observed SMC of 13.28%.

SN2 is the 2-norm of a sample of centered, log returns, and due to the squaring and square root operations, has no real-world meaning. Functionally, it measures the distance of the observed returns vector from the observed mean return. SMC is the geometric excess return of the maximum return $R_{MAX}$ with respect to the observed LETF return. It tells us the underperformance of the LETF relative to a hypothetical underlying index that provides constant returns (and still compounds to the total that was observed).

To better appreciate the different interpretations of these two statistics, let us refer to some examples from our dataset. In Table 1.3 we have two 252 day samples with nearly identical observed SN2s. Notice the substantial difference in observed SMC, however. For the TNA sample ending 2/2/2009, the underlying index (Russell 2000) lost about 10%. Based on this return the hypothetical best performance of TNA for that same period would be the $R_{MAX}$ of -28.6%. The observed return on TNA for that sample period was extremely close, resulting in a very low observed SMC of 2.27%. Contrast this with the sample period ending 12/31/2008 that observed a nearly identical SN2 as the TNA sample (0.4220 compared to 0.4148). Here the index returned 20% which implies the hypothetical best performance of TNA was 72.01%. TNA performed much worse than this maximum (51.84%) resulting in an observed SMC of 13.28%.

<table>
<thead>
<tr>
<th>LETF</th>
<th>Leverage</th>
<th>Index</th>
<th>EndDate</th>
<th>SN2</th>
<th>SMC</th>
<th>$R_{Index}$</th>
<th>$R_{LETF}$</th>
<th>$R_{MAX}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>TNA</td>
<td>+3</td>
<td>RTY</td>
<td>2009-02-02</td>
<td>0.4148</td>
<td>0.0227</td>
<td>-0.0990</td>
<td>-0.2860</td>
<td>-0.2698</td>
</tr>
<tr>
<td>TNA</td>
<td>+3</td>
<td>RTY</td>
<td>2008-12-31</td>
<td>0.4220</td>
<td>0.1328</td>
<td>0.2000</td>
<td>0.5184</td>
<td>0.7201</td>
</tr>
</tbody>
</table>

Table 1.3: 252 Day Samples with nearly identical observed SN2

Table 1.4 provides another pair of samples with nearly identical observed SN2s, but this time we consider 252 day sample periods. For the EDC sample ending 9/21/2012, the return was about 40% when the maximum return implied by the underlying index (MSCI Emerging Markets) performance was 53.64%, resulting in an observed SMC of 9.66%. That value is much lower than the observed SMC of 73.80% that we have for the sample period ending 5/6/2010. For that sample, the EDC return of 45.82% was substantially lower than the $R_{MAX}$ of 153.45%.

<table>
<thead>
<tr>
<th>LETF</th>
<th>Leverage</th>
<th>Index</th>
<th>EndDate</th>
<th>SN2</th>
<th>SMC</th>
<th>$R_{Index}$</th>
<th>$R_{LETF}$</th>
<th>$R_{MAX}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>EDC</td>
<td>+3</td>
<td>MXEF</td>
<td>2012-09-21</td>
<td>0.8212</td>
<td>0.0966</td>
<td>0.1540</td>
<td>0.4010</td>
<td>0.5364</td>
</tr>
<tr>
<td>EDC</td>
<td>+3</td>
<td>MXEF</td>
<td>2010-05-06</td>
<td>0.8250</td>
<td>0.7380</td>
<td>0.3639</td>
<td>0.4582</td>
<td>1.5345</td>
</tr>
</tbody>
</table>

Table 1.4: 252 Day Samples with nearly identical observed SN2
For both pairs of sample periods, the observed SN2, or 2-norm of the centered, log returns was nearly identical. Using this statistic we would interpret the sample periods for each pairing to have exhibited the same realized volatility of returns. However, within each pair we see substantial differences in shortfall from maximum convexity. In other words, the sample period with greater observed SMC has a greater degree of underperformance with respect to the sample (event) that observes the underlying index providing a constant daily return (and still compound to the same return for the period).

SMC is valued as a rate of return so the significance of the different observations (within the pairs) can be easily understood by an investor. If only considering SN2, not only do the observed values lack any kind of real-world interpretation, but there is virtually no measurable difference in realized volatility within the paired sample periods. We claim that there is a difference in the realized volatilities, and that SMC intuitively quantifies it.

1.7 Statistical Motivation for SMC

Now that we have made a qualitative argument for measuring the realized volatility of LETF returns, let us take a hard look at the data. We survey the sampling distributions of SMC and SN2 across all 20 LETFs in our data set, and using sample statistics provide a quantitative basis for the advantages of our new statistic.

Figure 1.11 provides box plots of the sample distributions for SMC and SN2 for every long and short LETF using 252 day sample periods. Looking at the plots for LETFs with underlyer GDM (NYSE ARCA Gold Miners Index) and those with underlyer INDU (Dow Jones Industrials), we see that there is considerable variety within the data set. We observe that SMC sampling distributions exhibit greater skewness compared to the respective SN2 distribution, as indicated by the length of the start of the right whisker to the end of the outlier points relative to the length of the rest of the box plot. The median, and thus lower half of each SMC box plot is to the left of the median of the respective SN2 box plot, and on an absolute basis. Our interpretation of these differences in box plots is that SN2 over-values the intensity of low volatility sample periods and under-values the intensity of high volatility sample periods compared to SMC.

In the previous section we discussed the inequality of SMC and SN2 when calculated on the same series of returns with equal but opposite leverage multiples, and we do see evidence of those relationships in the box plots. The distributions for SN2 across sides for the same underlying index are very similar, with the only visible difference being the right tail of the plots for index MVRSX. The differences in plots across sides are much greater with SMC. The short side box plots for SMC sampling distributions generally have greater ranges, longer right tails, and greater median values compared to the plots for the long side LETF on the same index. According to SMC we have evidence that LETF pairs on the same index have different risk profiles. For example, SPXL appears to be more risky than SPXS. This is notable because the box plots presented use real LETF returns, which includes a separate observation series for each long side and short side tracking error, and for LETFs with underlying index MVRSX, MXEA, and MXEF those errors were shown to be rather significant. But our claims in the previous sections assumed we observe LETF returns without noise \( R_{LETF} = \beta R_{Index} \), so we interpret this to mean that the observed values for both volatility statistics are mostly unaffected by tracking error. This is a point of emphasis in our next section when we use our proposed simulation method with LETF returns.

Range, variance, skewness, and kurtosis of all the SN2 and SMC sample distributions for all LETF tickers are presented in Tables 1.5 and 1.6. The measures of skewness and kurtosis used are the standardized third and fourth central moments, respectively. The greater value for each pair of statistics is highlighted
in red and we interpret the greater values as evidence that the given statistic provides a more granular level of detail about the realized volatility of the given LETF ticker. When considering range and variance this interpretation should be obvious–a statistic with a wider range of outcomes or greater variability of outcomes (with respect to the mean) provides more information about the underlying random process. A greater absolute value of skewness implies longer, fatter tails on one side of the distribution. We say absolute value of skewness to mean taking the absolute value of the final value for skewness as to allow for negative or positive skew in our comparisons, and this is not to be confused with calculating the third absolute moment. A greater value for kurtosis implies greater probability mass in the center and tails of the distribution. In both instances, a greater value implies more mass in the tails and thus greater variability of outcomes.

Across all long and short side LETF tickers the skewness and kurtosis values for SMC are greater than the SN2 values, except for skewness of the pair FINU/FINZ (underlying index Dow Jones US Financials), where both tickers had a negatively skewed sample distribution for SN2. The variance and range values for SMC are greater than SN2 for most short side tickers, and the results are mixed for long side tickers. Based on these statistics, our opinion is that SMC provides greater detail about the intensity of realized volatility than SN2 when observing right tail events, and overall SMC is a more informative statistic to use for short side LETF tickers.

We conclude this section with a brief application example where we rank (separately) the the set of long side and short side LETF tickers by their sample mean SN2 and SMC using 252 day sample periods. The rankings are displayed in Figure 1.12 with rank #1 having the greatest sample mean for the given statistic, and rank #10 having the least. The rank orders that do not agree show two asterisk characters on the tickers at that order number.

There are two sets of two rank orders that do not agree in both the long side and the short side. In the group of long side LETF tickers, ranks #2 and #3, and then ranks #7 and #8, do not agree. On the short side, ranks #3 and #4, and also ranks #7 and #8, do not agree. Similar to our observation about the box plots for SN2 displaying minimal differences between long and short side LETF tickers on the same underlying index, here we note that the rankings by SN2 do not change between sides. The SMC rankings have EDC, RUSL, TNA at #2, #3, #4 on the long side, but on the short side we have RUSS, TZA, EDZ. The #7 and #8 rank orders are the LETFs with the same underlying indexes across long and short sides when using SMC.

Although the sample means for both statistics are plotted on the same axis, we remind the reader that the interpretations of the values are not the same. Short side LETF ticker DUST, for example, has a mean SN2 of about 1.05 and a mean SMC of about 1.3. SN2 is the 2-norm of the centered, log returns and the mean value of 1.05 offers a general indication of the variability of returns observed in the 252 day sample periods. SMC is the shortfall of the observed LETF return from a maximum return, so the mean value of 1.3 tells us that the LETF ticker DUST underperforms the maximum return achievable given the index return by 130% on average over 252 day sample periods. We feel that the added intuition that comes with measuring realized volatility with SMC makes those rankings more easily understood and more useful when designing trading strategies for these instruments.
Figure 1.11: SN2 vs. SMC for all LETF Ticker Pairs over 252 Day Periods
<table>
<thead>
<tr>
<th>LETF</th>
<th>Index</th>
<th>Obs</th>
<th>Range(SN2)</th>
<th>Range(SMC)</th>
<th>Var(SN2)</th>
<th>Var(SMC)</th>
<th>Skew(SN2)</th>
<th>Skew(SMC)</th>
<th>Kurt(SN2)</th>
<th>Kurt(SMC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FINZ</td>
<td>DJUSFN</td>
<td>361</td>
<td>0.0978</td>
<td>0.1013</td>
<td>0.0009</td>
<td>0.0002</td>
<td>-0.4852</td>
<td>0.4622</td>
<td>1.6312</td>
<td>1.9498</td>
</tr>
<tr>
<td>NUGT</td>
<td>GDM</td>
<td>762</td>
<td>0.7218</td>
<td>0.9325</td>
<td>0.0423</td>
<td>0.0632</td>
<td>0.2626</td>
<td>0.4749</td>
<td>1.7779</td>
<td>1.9498</td>
</tr>
<tr>
<td>UDOM</td>
<td>INDU</td>
<td>970</td>
<td>0.3475</td>
<td>0.1452</td>
<td>0.0152</td>
<td>0.0018</td>
<td>0.6913</td>
<td>0.8978</td>
<td>1.8986</td>
<td>2.1961</td>
</tr>
<tr>
<td>TECL</td>
<td>IXT</td>
<td>1259</td>
<td>0.4895</td>
<td>0.7585</td>
<td>0.0370</td>
<td>0.0339</td>
<td>0.0663</td>
<td>0.5625</td>
<td>1.6694</td>
<td>2.0975</td>
</tr>
<tr>
<td>RUSL</td>
<td>MVRSX</td>
<td>646</td>
<td>0.7468</td>
<td>1.0860</td>
<td>0.0426</td>
<td>0.0669</td>
<td>1.2437</td>
<td>1.5168</td>
<td>3.6069</td>
<td>4.4666</td>
</tr>
<tr>
<td>DZK</td>
<td>MXEA</td>
<td>1259</td>
<td>0.6932</td>
<td>0.9325</td>
<td>0.0423</td>
<td>0.0632</td>
<td>0.2626</td>
<td>0.4749</td>
<td>1.7779</td>
<td>1.9498</td>
</tr>
<tr>
<td>EDC</td>
<td>MXEF</td>
<td>1259</td>
<td>0.7150</td>
<td>1.0721</td>
<td>0.0338</td>
<td>0.0459</td>
<td>0.4742</td>
<td>0.7596</td>
<td>2.0482</td>
<td>2.9446</td>
</tr>
<tr>
<td>TQQQ</td>
<td>NDX</td>
<td>970</td>
<td>0.3592</td>
<td>0.1506</td>
<td>0.0145</td>
<td>0.0024</td>
<td>0.5799</td>
<td>0.7724</td>
<td>1.8429</td>
<td>2.0746</td>
</tr>
<tr>
<td>TNA</td>
<td>RTY</td>
<td>1288</td>
<td>0.8963</td>
<td>0.8412</td>
<td>0.0434</td>
<td>0.0202</td>
<td>0.4335</td>
<td>1.4309</td>
<td>2.1737</td>
<td>5.8182</td>
</tr>
<tr>
<td>SPXL</td>
<td>SPX</td>
<td>1288</td>
<td>0.7020</td>
<td>0.4550</td>
<td>0.0239</td>
<td>0.0052</td>
<td>0.5260</td>
<td>1.4177</td>
<td>2.4170</td>
<td>4.8865</td>
</tr>
</tbody>
</table>

Table 1.5: SN2 and SMC (252 Day Periods) Summary Statistics for Long LETF Tickers

<table>
<thead>
<tr>
<th>LETF</th>
<th>Rank(SN2)</th>
<th>Rank(SMC)</th>
<th>Var(SN2)</th>
<th>Var(SMC)</th>
<th>Skew(SN2)</th>
<th>Skew(SMC)</th>
<th>Kurt(SN2)</th>
<th>Kurt(SMC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FINZ</td>
<td>1</td>
<td>2</td>
<td>0.014</td>
<td>0.0018</td>
<td>-0.7110</td>
<td>-0.0770</td>
<td>1.8092</td>
<td>3.0532</td>
</tr>
<tr>
<td>DUST</td>
<td>2</td>
<td>3</td>
<td>0.0442</td>
<td>0.5799</td>
<td>0.2558</td>
<td>0.7119</td>
<td>1.8178</td>
<td>2.0514</td>
</tr>
<tr>
<td>SDOW</td>
<td>3</td>
<td>4</td>
<td>0.0144</td>
<td>0.0084</td>
<td>0.6652</td>
<td>0.8569</td>
<td>1.8727</td>
<td>2.1631</td>
</tr>
<tr>
<td>TQOO</td>
<td>4</td>
<td>5</td>
<td>0.0183</td>
<td>0.0399</td>
<td>0.0818</td>
<td>1.8960</td>
<td>1.8787</td>
<td>7.5079</td>
</tr>
<tr>
<td>DZK</td>
<td>5</td>
<td>6</td>
<td>0.0369</td>
<td>0.2485</td>
<td>1.0722</td>
<td>1.7026</td>
<td>3.1842</td>
<td>4.6711</td>
</tr>
<tr>
<td>EDC</td>
<td>6</td>
<td>7</td>
<td>0.0350</td>
<td>0.0237</td>
<td>-0.0366</td>
<td>0.5442</td>
<td>1.6074</td>
<td>2.6469</td>
</tr>
<tr>
<td>RUSL</td>
<td>7</td>
<td>8</td>
<td>0.0340</td>
<td>0.0577</td>
<td>0.3779</td>
<td>1.0513</td>
<td>2.0376</td>
<td>4.0228</td>
</tr>
<tr>
<td>TZZA</td>
<td>8</td>
<td>9</td>
<td>0.0459</td>
<td>0.1466</td>
<td>0.4235</td>
<td>1.6939</td>
<td>2.1804</td>
<td>7.5162</td>
</tr>
<tr>
<td>TQOO</td>
<td>9</td>
<td>10</td>
<td>0.0239</td>
<td>0.0304</td>
<td>0.5282</td>
<td>2.0606</td>
<td>2.5240</td>
<td>9.9243</td>
</tr>
</tbody>
</table>

Table 1.6: SN2 and SMC (252 Day Periods) Summary Statistics for Short LETF Tickers

Figure 1.12: Rankings for Long, Short LETF Tickers by SN2 and SMC
Chapter 2

Simulation of Leveraged ETF Volatility Using Nonparametric Density Estimation

[Ginley et al., 2015]
2.1 Overview and Discussion of LETF Returns Literature

The class of securities commonly known as LETFs, or Leveraged Exchange Traded Funds, began trading on the New York Stock Exchange in 2006. LETFs are managed by their issuers with a mandate to provide a multiple of the daily return of the underlying asset, and were developed to provide professional investors a convenient means for the short term hedging of market risk. Since their issuance, their popularity in the investment world has risen steadily and now there are investors promoting buy and hold strategies that feature long term positions in LETFs [Fisher, 2014, Scott and Watson, 2013]. The lack of understanding of the risks associated with these strategies and more generally LETF return dynamics was a primary motivation for our work. In this chapter our objective is to offer improved tools for analyzing the risks involved with investing in LETFs.

Not long after LETFs started trading, researchers began to examine the leveraged returns they provide and the associated daily compounding. Jarrow [Jarrow, 2010] and Avellaneda and Zhang [Avellaneda and Zhang, 2010] were some of the first to do so. The work by both sets of authors is very similar. They clearly prove that the multi-day return on an LETF is not the same as the return on an (unleveraged) ETF with the same underlying index and relate the difference to the ETF’s volatility and borrowing costs. Avellaneda and Zhang also provide a strong collection of empirical results to reinforce their proof. In a similar fashion, Lu, Wang, and Zhang [Lu et al., 2012] provide empirical results that extensively quantify the underperformance of LETFs with respect to their benchmark over multi-day periods. They claim generally that with a leverage multiple of 2, or we say “2x leverage”, one can expect an LETF to deliver its targeted return for periods up to a month, and then they quantify expected deviations for longer horizons.

The work by Dobi and Avellaneda [Dobi and Avellaneda, 2012] takes the discussion of LETFs a step further. They highlight the total return swap exposures that are rebalanced daily by LETF issuers to provide the leveraged return target as a cause of LETF underperformance. We strongly agree with their comment that LETF underperformance “... is especially true for longer holding periods and for periods of high volatility”. Tuzun [Tuzun, 2012] expands on the ramifications of the swap exposures that are held by the LETF issuers. He notes the similarities between daily rebalancing necessitated by LETFs and portfolio insurance strategies from the 1980s. This work is written from the perspective of market regulators, but we found it helpful because it emphasized the effects of consecutive gains or consecutive losses on LETF performance. With LETFs a period of consecutive gains produces leveraged gains on leveraged gains, resulting in a compound return that is much greater than would be experienced without leverage. This scenario should not be difficult to understand, but in the case of consecutive losses with leverage, the authors demonstrate that this also results in improved performance (relatively). In other words, relative to the same index exposure without leverage, the magnitude of the cumulative loss experienced with leverage is less. This is because with every consecutive negative return, exposure is decreased before the next loss, and with leverage the decrease of exposure is greater.

Cooper [Cooper, 2010] provides a detailed review of volatility drag and LETF dynamics in an attempt to prove that there exists a leverage ratio greater than one that is optimal for buy and hold investing in any equity index. We are doubtful of his claims because of the assumption of no tracking errors. We show later this would be poor judgment with the LETFs that we analyze. Also, Bouchey and Nemtchinov [Bouchey and Nemtchinov, 2013] show that underperformance of high volatility stocks compared to low volatility stocks is mostly the result of volatility drag. There is a parallel that could be considered between their thesis and the performance of LETFs with high volatility underlying indexes compared to LETFs with low volatility underlying indexes, but we do not investigate it here.
Qualitatively speaking, we agree with the above mentioned authors that LETFs frequently underperform over time horizons greater than one month, underperformance is greater in times of high volatility, and that underperformance is intrinsically related to volatility drag. But, we view our contributions as an empirical framework that complements their analytical models and we do not attempt to reproduce the results on which their conclusions are based. Instead, our intention is to provide methods that can help LETF investors better account for these dynamics. The remainder of this chapter is organized as follows.

In Section 2.2 we review the LETF returns dataset used throughout this chapter, present a visual motivation for our contributions, and describe two statistics for measuring the realized volatility of LETF returns — sample standard deviation and Shortfall from Maximum Convexity (SMC). SMC is a variant of volatility drag that specifically accounts for the effects of leverage. We defined SMC and documented the advantages of using it to measure realized volatility in Chapter 1, which is summarized in Section 2.2.2.

In Section 2.3 we propose a method for simulating LETF tracking errors. The proposed method allows us to produce simulated LETF returns derived from index returns and LETF tracking errors that are simulated separately. In turn, we can make use of the entirety of their respective datasets, greatly improving the robustness of our density estimates used in simulation.

In Section 2.4 we exploit this separation by proposing a method for simulating index returns with a constrained total period return. Constraining the underlying index performance allows for greater nuance with modeling LETF volatility. With this method, a practitioner can formulate more customized LETF volatility summaries by including a specific returns forecast for the underlying index.

Finally, we demonstrate how to combine the two simulation methods with a brief case study in Section 2.5 and review our contributions in Section 2.6.

2.2 Visualizing LETF Returns and Measuring Realized Volatility

The demonstration set of returns from paired LETFs and underlying equity indexes is the same as the set used in Chapter 1, and the specific instruments included are listed in Section 1.1 in Tables 1.1 and 1.1. No other sources of data are referenced in this chapter.
2.2.1 Visualizing LETF Returns

![Graph showing scatter plots of 252 day returns for hypothetical LETF “+3x SPX” and SPXL](image)

Figure 2.1: Scatter Plots of 252 Day Returns for hypothetical LETF “+3x SPX” and SPXL

Our contributions can best be motivated through visualization, and so we begin with Figure 2.1. There are two scatter plots of 252 day compound LETF returns (y axis) against the matching 252 day compound index returns (x axis) from the same periods. The right plot is for +3x LETF Ticker SPXL, the left plot is for the hypothetical LETF “+3x SPX” created by daily, leveraged compounding of S&P 500 returns, and then for both the underlying index is the S&P 500. The diagonal blue line in each plot corresponds to $y = 3x$, and is provided to indicate what return would be achieved by a leveraged investment in the index without daily compounding. It has been said time and again that “a picture is worth 1000 words,” but in our case we could say 10,000 words, the approximate length of this chapter.

Three extraordinarily important observations need to be made. First, there are many more points representing a much wider range of outcomes in the left plot, which is derived from the entire daily returns series of the S&P 500 (21,591 intervals of 252 days), compared to the right plot that includes only the SPXL returns series (1288 intervals of 252 days). Second, there is an interesting crescent shape with a hard upper bound apparent in the left plot, where the overwhelming majority of points are clustered near the upper bound. Third, again in the left plot, points seem to disperse from the upper bound in a way that maintains some degree of curvature such that the range of dispersion is not constant. These characteristics were readily observable when the same plot was generated using other LETF and index pairs.

Figure 2.1 illustrates the advantages of having more data. The right plot was limited to the returns data available for SPXL, which extends back to 2008. On the other hand, the left plot used all the data available for the S&P 500, which extends back to the 1920s. Except for the Market Vectors Russia Index (MVRSX) there are at least 16 years of data for every index that we consider in this chapter. Given the dependence of an LETF’s daily returns on the matching returns of the underlying index, it follows naturally that we
should develop schemes for modeling LETF returns that can incorporate and control for the variation in the complete history of index returns data. Our methods presented in Sections 2.3 and 2.4 address these two motivations, respectively.

The potential enhancements offered by our methods are especially significant given the fact that North American equity indexes have generally been in an upward trend since March 2009. In other words, the majority of available LETF returns were observed when the index returns that they are derived from were positive or compound to a positive value over periods of months or years. This can be seen in Figure 2.1, where the right plot has no points with an x value less than 0 (the underlying index return), but there are many points that satisfy this condition in the left plot. Obviously, any quantitative LETF analysis would benefit from the ability to include data from downward trending markets and isolate the resulting effects.

2.2.2 Measuring Realized Volatility

We will measure realized volatility with Shortfall from Maximum Convexity (SMC), a novel statistic that was designed specifically for application to LETF returns [Ginley, 2015].

\[
\text{SMC}(R_{\text{LETF}}, R_{\text{Index}} | \beta_{\text{LETF}}) = \exp \left[ \sum_{j=1}^{p} \left( \log(1 + \beta_{\text{LETF}} \tilde{R}_{\text{Index}}) - \log(1 + R_{\text{LETF},j}) \right) \right] - 1 \quad (2.1)
\]

\[
\tilde{R}_{\text{Index}} = \prod_{j=1}^{p} (1 + R_{\text{Index},j})^{\frac{1}{p}} - 1 \quad (2.2)
\]

where \( \beta_{\text{LETF}} \) is the indicated leverage multiple for LETF (constant), \( R_{\text{Index}} \) is the matching \( p \) dimensional vector of daily returns for Index, and \( \tilde{R}_{\text{Index}} \) is the geometric mean of the elements of \( R_{\text{Index}} \). What differentiates SMC from the more traditional sample standard deviation is subtle — a mean return of \( \beta_{\text{LETF}} \tilde{R}_{\text{Index}} \) is used instead of \( \tilde{R}_{\text{LETF}} \), and the sum of daily differences is exponentiated instead of taking the square root of the sum of squared, daily differences.

These changes were inspired by the geometric details of the plotting arrangements used in Figure 2.1. SMC is simply the vertical distance between a data point and the point on the convex upper bound that is located at the same x value (calculating the exact form of points on the upper bound is trivial). The interpretation of this measure is the geometric excess return of a hypothetical, maximum return derived from the index return, with respect to the LETF return. We use SMC because the interpretation of sample standard deviation becomes problematic with non-normal, dependent data, and when applied to LETF returns SMC provides more information, statistically speaking. These arguments are covered in much greater detail within the original proposal for SMC, and we refer interested readers there for more information.

2.3 Simulation of Tracking Errors Using Nonparametric Density Estimation

A naive approach for modeling LETF returns would be to simply assume no tracking errors and approximate daily returns as the underlying daily index return with leverage and net of management fees. These approximated returns, after compounding, could be used in the denominator when calculating SMC in place of a true LETF return. This naive approach is simple and allows for the incorporation of the entire history of index returns data; however, it ignores the impact of LETF tracking errors on returns and realized
volatilities.

This impact becomes significant over time, as demonstrated in Figures 2.2 and 2.3 for two LETFs. First, density curves of 252 day returns of LETF Ticker RUSS (top) and LETF Ticker TECL (bottom) are plotted against 252 day returns of their respective underlying indexes, but compounded daily with a leverage multiple of -3. Note the separation between the density curves, which is due to the cumulative effect of tracking errors because the index returns were calculated with daily compounding and leverage. Second, density curves of 252 day SMC are plotted using the same data and plotting arrangements. Again, note the separation between the density curves, and that the separation is much greater when visualizing realized volatility compared to visualizing returns.

In light of this evidence, it is our opinion that approximating LETF returns by assuming no tracking errors would provide for inaccurate analyses, and in response we propose the nonparametric method for simulating LETF tracking errors described in this section. Our method estimates filtering and smoothing densities using only data from the days when LETFs were trading. The proposed method effectively allows for simulating LETF returns derived from index returns and LETF tracking errors that are simulated separately, thus incorporating the entire history of index returns and LETF returns data in an extremely efficient manner.

2.3.1 Framing Tracking Error Simulation as Filtering

When discussing any statistical analysis of daily LETF returns, over a period of \(k\) days, the following length \(2k\) random vector of daily asset returns is observed:

\[
[R_{Index,1}, \ldots, R_{Index,k}, R_{LETF,1}, \ldots, R_{LETF,k}]
\]  

where \(R_{Index,t}\) is the daily index return on day \(t\) for the underlying index ticker \(Index\), and \(R_{LETF,t}\) is the daily LETF return on day \(t\) for the specified LETF ticker \(LETF\). We assume that daily LETF returns and their matching daily index returns are related through the following decomposition:

\[
R_{LETF,t} = \beta_{LETF} R_{Index,t} - \text{MgmtFee}_{LETF} + \epsilon_{LETF,t}
\]  

where \(\beta_{LETF}\) is the indicated leverage multiple for \(LETF\) (constant), \(\text{MgmtFee}_{LETF}\) is the indicated rate of daily management expenses for \(LETF\) (constant), and \(\epsilon_{LETF,t}\) is the tracking error on day \(t\) for \(LETF\) (random variable). Implicitly, then, we are assuming that the following random vector captures the same information as the vector that was stated originally:

\[
[R_{Index,1}, \ldots, R_{Index,k}, \epsilon_{LETF,1}, \ldots, \epsilon_{LETF,k}]
\]
Figure 2.2: Densities of 252 Day Returns for LETF Tickers RUSS and TECL and their respective indexes compounded daily with -3x leverage

Figure 2.3: Densities of 252 Day SMC for LETF Tickers RUSS and TECL and their respective indexes compounded daily with -3x leverage

From this view of the LETF and index returns processes we develop a simulation method that is based
on the following factorization of the joint density for these variables:

\[ f(R_{\text{Index},1}, \ldots, R_{\text{Index},k}, \epsilon_{\text{LETF},1}, \ldots, \epsilon_{\text{LETF},k}) = \prod_{j=1}^{k} g(\epsilon_{\text{LETF},j}|R_{\text{Index},j}) \ast f(R_{\text{Index},1}, \ldots, R_{\text{Index},k}) \]  

(2.6)

The density \( g(\epsilon_{\text{LETF},j}|R_{\text{Index},j}) \) is a filter as used in state space modeling, where (by necessity) the index returns process is designated as the observation process and the LETF returns process is designated as the state process. This runs counter to a more intuitive arrangement of designating LETF returns as functions of the state (index returns) with observation noise (tracking error).

To handle any dependence that may exist within a sequence of index return variables, we do not factorize the density \( f(R_{\text{Index},1}, \ldots, R_{\text{Index},k}) \). Modeling all of the index return variables jointly captures their complete \( k \) day dependence structure. To account for dependence of LETF tracking error data and inter-dependence of tracking errors and index returns, in our filtering density we will additionally include \( \ell \) lagged index returns and \( \ell \) lagged tracking errors as given, where \( \ell \) is an integer valued parameter that will be selected in a later subsection. This arrangement reflects our view that LETF tracking errors have a simpler dependence structure compared to index returns (in practice \( 2\ell << k \)), and that LETF tracking errors are dependent on the underlying index returns. We do not consider index returns to be dependent on the respective LETF returns or tracking errors; therefore, nowhere in our approach do we attempt to model the statistical relationship of \( \epsilon_{\text{LETF}}|R_{\text{Index}} \).

To include the additional \( 2\ell \) lagged variables, the joint density factorization is revised as follows:

\[ f(R_{\text{Index},1}, \ldots, R_{\text{Index},\ell+k}, \epsilon_{\text{LETF},1}, \ldots, \epsilon_{\text{LETF},\ell+k}) = \prod_{j=\ell+2}^{\ell+k} g(\epsilon_{\text{LETF},j}|R_{\text{Index},j}, R_{\text{Index},j-\ell}, \ldots, R_{\text{Index},j-1}, \epsilon_{\text{LETF},j-\ell}, \ldots, \epsilon_{\text{LETF},j-1}) \ast g(\epsilon_{\text{LETF},1}, \ldots, \epsilon_{\text{LETF},\ell+1}|R_{\text{Index},1}, \ldots, R_{\text{Index},\ell+1}) \ast f(R_{\text{Index},1}, \ldots, R_{\text{Index},\ell+k}) \]  

(2.7)

There are three important changes to highlight. First, the filtering density \( g(\epsilon_{\text{LETF},j}|R_{\text{Index},j}, \ldots) \) is now conditioned on \( \ell \) lagged index return variables and \( \ell \) lagged LETF return variables (in addition to \( R_{\text{Index},j} \)). Second, the joint density of index returns and LETF tracking errors (left-hand side), and unconditional joint density of index returns (final term of right-hand side), both include the additional lagged variables such that the random vector being modeled is now of length \( (\ell + k) \). Third, we have introduced a conditional density for smoothing the \( (\ell + 1) \) tracking error variables, \( g(\epsilon_{\text{LETF},1}, \ldots, \epsilon_{\text{LETF},\ell+1}|R_{\text{Index},1}, \ldots, R_{\text{Index},\ell+1}) \).

This additional term is necessary to initiate, or “burn-in”, the sequence of tracking error variables, providing the first set of \( \ell \) lags that are required by the first iteration of the filtering density.

Also, the decision to smooth the first \( (\ell + 1) \) variables instead of the first \( \ell \) variables and then starting filtering at \( j = (\ell + 1) \) deserves further comment. Our reasoning is implementation related, so the dataset needs to be transformed only once. As written, the conditional densities differ only by which variables are considered as given. Otherwise both densities model the same \( (2\ell + 2) \) variables — a sequence of \( (\ell + 1) \) index returns and the matching sequence of \( (\ell + 1) \) tracking errors. Having both densities model the same variables allows for the same observation matrix being used for both of the resulting kernel density estimates, as shown in the next subsection of derivations. If only the first \( \ell \) variables were smoothed, and iterations of the filtering density began with day \( (\ell + 1) \), the dataset would need to be transformed twice. One transform
would be required to implement an estimate of the density with $2\ell$ variables used for smoothing the initial lagged errors and another to implement an estimate of the density with $(2\ell + 2)$ variables used for filtering the remaining errors. There is one exception, when $\ell = 0$, a case which results in smoothing and filtering densities that are identical and the joint density factorization reverts to the initially stated form with only two terms, $\prod_{j=1}^{k} g(\epsilon_{LETF,j}|R_{Index,j}) \ast f(R_{Index,1},...,R_{Index,k})$.

However, to avoid the additional bias introduced by smoothing the first $\ell$ tracking errors (conditioning on index returns forward in time relative to the tracking errors constitutes look-ahead bias), we will discard the first $\ell$ terms when using our simulation method, making any samples generated have an effective length of $k$. The $(\ell + 1)$st term is not discarded because the associated look-ahead bias is the same as filtering (relative to this tracking error only index returns at the same time step or backward, but not forward, are given). We acknowledge that the filtering density $g(\epsilon_{LETF,j}|R_{Index,j},...)$ also constitutes look-ahead bias. In real time we observe the daily closing index price and the daily closing LETF price (which imply daily returns) simultaneously, so the information provided by $R_{Index,j}$ should not be available for filtering $\epsilon_{LETF,j}$. But we accept this bias for the increased precision of our estimate of the filtering density that it allows, thus improving our modeling of the realized volatility of LETF returns. Our method is not intended for a live setting and we do not calculate the expectation of any random variables (e.g. $E[\epsilon_{LETF,j}|R_{Index,j},...]$) for the purposes of predicting tracking errors or LETF returns.

### 2.3.2 Derivation of Kernel Density Estimates

Given a random process generating daily LETF returns $\{R_{LETF,t}, t \in T\}$, a random process generating the daily underlying index returns $\{R_{Index,t}, t \in T\}$, the target leverage multiple of the LETF ($\beta_{LETF}$), the rate of daily management expenses incurred by the LETF ($\text{MgmtFee}_{LETF}$), and a chosen number of lags $\ell$ indicating the degree of dependence, we construct kernel density estimates of the smoothing density $g(\epsilon_{LETF,1},...,\epsilon_{LETF,\ell+1}|R_{Index,1},...,R_{Index,\ell+1})$ and filtering density $g(\epsilon_{LETF,j}|R_{Index,j},...)$ as follows.

Note, in this subsection we use slightly different notation so that the implementation details of the kernel density estimates can be documented more clearly. The selection of $\ell$ will be discussed in the following subsection that outlines the tracking error simulation routine.

1. Calculate the implied tracking errors from the observed data, and transform the data to allow for fitting the desired density estimates.

Assume the following relationship for decomposing daily LETF returns (as stated previously):

$$R_{LETF,t} = \beta_{LETF}R_{Index,t} - \text{MgmtFee}_{LETF} + \epsilon_{LETF,t}$$

Using the decomposition, observed LETF returns ($r_{LETF}$), and observed index returns ($r_{Index}$), calculate the implied tracking error observations ($e_{LETF}$). All of the observation vectors are length $T$.

Take the log transform of observed index returns log$(1 + r_{Index})$ and arrange them into a $(T-\ell) \times (\ell+1)$ observation matrix of rolling $(\ell+1)$ day periods. Denote this matrix as $\mathbf{X}$. In the same way, take the log transform of observed tracking errors log$(1 + e_{LETF})$ and arrange them into a $(T-\ell) \times (\ell+1)$ observation matrix of rolling $(\ell+1)$ day periods. Denote this matrix as $\mathbf{Y}$.

To form the final observation matrix used for fitting the kernel density estimate, combine $\mathbf{X}$ and $\mathbf{Y}$ in column-wise fashion. Denote this as $\mathbf{Z}$, a $(T-\ell) \times (2\ell+2)$ matrix. For notational convenience in the
following steps, let $n = T - \ell$ and $p = 2\ell + 2$, so that observation matrix $\mathbf{Z}$ is $n \times p$.

$$
\mathbf{Z} = [\mathbf{X} \ \mathbf{Y}] 
$$

(2.9)

$$
\begin{bmatrix}
\log(1 + r_{\text{Index},1}) & \ldots & \log(1 + r_{\text{Index},\ell+1}) & \log(1 + e_{\text{LETF},1}) & \ldots & \log(1 + e_{\text{LETF},\ell+1}) \\
\log(1 + r_{\text{Index},2}) & \ldots & \log(1 + r_{\text{Index},\ell+2}) & \log(1 + e_{\text{LETF},2}) & \ldots & \log(1 + e_{\text{LETF},\ell+2}) \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
\log(1 + r_{\text{Index},T-\ell}) & \ldots & \log(1 + r_{\text{Index},T}) & \log(1 + e_{\text{LETF},T-\ell}) & \ldots & \log(1 + e_{\text{LETF},T}) 
\end{bmatrix}
$$

(2.10)

Using log returns ($\log(1 + r)$) and log tracking errors ($\log(1 + e)$) is extremely important because of the fact that log values do not suffer from asymmetry or a hard left bound at 0 (now transformed to $-\infty$), making the transformed data less challenging to model.

2. Fit a kernel density estimate for $f(\mathbf{z})$, using a jointly independent multivariate normal kernel and observation matrix $\mathbf{Z}$.

$$
\hat{f}(\mathbf{z}) = \frac{1}{n} \sum_{i=1}^{n} \phi_p(\mathbf{z} | \hat{\mathbf{z}}_{(i)}, \text{diag}(\mathbf{h}^2))
$$

(2.11)

where $\phi_p(\mathbf{z} | \mu, \Sigma)$ is the $p$ dimensional multivariate normal density function evaluated at $\mathbf{z}$ using mean vector $\mu$ and covariance matrix $\Sigma$, $\hat{\mathbf{z}}_{(i)}$ is the $i$th row vector of observation matrix $\mathbf{Z}$, and $\text{diag}(\mathbf{h}^2)$ is the diagonal matrix formed from element-wise squaring of the $p$ dimensional vector of bandwidths. Elements indexed $1 : (\ell + 1)$ correspond to bandwidths for index returns, and elements $(\ell + 2) : p$ correspond to bandwidths for tracking errors.

The selection of $\mathbf{h}$ is dependent on the LETFs (and underlying indexes) being analyzed. Our selections for the demonstration set of LETFs will be detailed later in this section.

Note that in using a diagonal matrix based on the vector of $h_{ij}$s, the selected bandwidths for single dimensions, we are not ignoring the dependence structure that exists between the variables. The dependence structure will be estimated through the addition of the $n$ multivariate kernels centered on the observation data points. This point might be easier to understand algebraically, because although the kernel in the density estimate can be decomposed as the product of univariate kernels, it is not the same as an estimate that is the product of independent univariate density estimates.

$$
\frac{1}{n} \sum_{i=1}^{n} \phi_p(\mathbf{z} | \hat{\mathbf{z}}_{(i)}, \text{diag}(\mathbf{h}^2)) = \frac{1}{n} \sum_{i=1}^{n} \left[ \prod_{j=1}^{p} \phi(z_j | \hat{z}_{i,j}, h^2_j) \right] \neq \prod_{j=1}^{p} \left( \frac{1}{n} \sum_{i=1}^{n} \phi(z_j | \hat{z}_{i,j}, h^2_j) \right) = \prod_{j=1}^{p} \hat{f}(z_j)
$$

(2.12)

where $\hat{z}_{i,j}$ is the $j$th element of observation vector $\hat{\mathbf{z}}_{(i)}$.

By capturing the dependence structure of the data through the addition of multivariate kernels that use a diagonal covariance matrix, we are achieving an effective compromise between modeling power and complexity. Selecting bandwidths for dimensions individually incorporates some information about the respective variable, yet making $p$ separate selections separately is not difficult. The alternative choices of using a single bandwidth for all dimensions or selecting an entire covariance matrix are generally considered too simple or too difficult, respectively [Scott, 1992].
3. Factor and rearrange terms of the density estimate for \( f(z) \) to produce an estimate of the conditional density \( g(z_{(r+2):p}|z_{1:(r+1)}) \).

\[
g(z_{(r+2):p}|z_{1:(r+1)}) = f(z)/f(z_{1:(r+1)}) \quad (2.13)
\]

where \( z_{1:(r+1)} \) is the subvector of elements indexed \( 1:(r+1) \) from vector \( z \), or the elements that represent index returns whose observations were sourced from the matrix \( X \), and \( z_{(r+2):p} \) is the subvector of elements indexed \( (r+2):p \) from vector \( z \), or the elements that represent tracking errors whose observations were sourced from the matrix \( Y \). Equation (2.13) could be restated as \( g(y|x) = f(x,y)/f(x) \); however, the following notation will be used for the sake of brevity and consistency with Step 4.

\[
g(z_y|z_x) = f(z)/f(z_x) \quad (2.14)
\]

In this way, when \( x \) and \( y \) are written as subscripts, let \( z_x = z_{1:(r+1)} \) and \( z_y = z_{(r+2):p} \), and by extension \( h_x^2 = h_{1:(r+1)}^2 \) and \( h_y^2 = h_{(r+2):p}^2 \).

Continue by inserting the appropriate density estimates.

\[
\hat{g}(z_y|z_x) = \hat{f}(z)/\hat{f}(z_x)
\]

\[
\quad = \left[ \frac{1}{n} \sum_{i=1}^{n} \phi_p(z|\hat{z}_{(i),x}, \text{diag}(h_x^2)) \right] / \left[ \frac{1}{n} \sum_{i=1}^{n} \phi_{\ell+1}(z_x|\hat{z}_{(i),x}, \text{diag}(h_x^2)) \right]
\]

Because of our choice of a jointly independent multivariate normal kernel, \( \phi_p(z|\ldots) \) can be factored as follows:

\[
\phi_p(z|\hat{z}_{(i),x}, \text{diag}(h_x^2)) = \phi_{\ell+1}(z_x|\hat{z}_{(i),x}, \text{diag}(h_x^2)) \ast \phi_{\ell+1}(z_y|\hat{z}_{(i),y}, \text{diag}(h_y^2))
\]

(2.17)

Insert this factorization and rearrange terms to produce the final density estimate.

\[
\quad = \left[ \frac{1}{n} \sum_{i=1}^{n} \phi_{\ell+1}(z_y|\hat{z}_{(i),y}, \text{diag}(h_y^2)) \ast \phi_{\ell+1}(z_x|\hat{z}_{(i),x}, \text{diag}(h_x^2)) \right] / \left[ \frac{1}{n} \sum_{i=1}^{n} \phi_{\ell+1}(z_x|\hat{z}_{(i),x}, \text{diag}(h_x^2)) \right]
\]

(2.18)

\[
\quad = \sum_{i=1}^{n} \phi_{\ell+1}(z_y|\hat{z}_{(i),y}, \text{diag}(h_y^2)) \ast \alpha_i(z_x|\hat{z}_{(i),x}, \text{diag}(h_x^2))
\]

(2.19)

\[
\alpha_i(z_x|\hat{z}_{(i),x}, \text{diag}(h_x^2)) = \phi_{\ell+1}(z_x|\hat{z}_{(i),x}, \text{diag}(h_x^2)) / \left[ \sum_{i=1}^{n} \phi_{\ell+1}(z_x|\hat{z}_{(i),x}, \text{diag}(h_x^2)) \right]
\]

(2.20)

\( \hat{g}(z_y|z_x) \) is the estimate of the smoothing density that will be used to simulate the first \( (r+1) \) tracking errors of a new sample.
4. Factor and rearrange terms of the density estimate for \( f(z) \) to produce an estimate of the conditional density \( g(z_p|...) \).

\[
g(z_p|z_{-p}) = \frac{f(z)}{f(z_{-p})} \tag{2.21}
\]

where \( z_{-p} \) is the vector \( z \) without the last element. Equation (2.21) could be restated as \( g(y_{\ell+1}|x,y_{-(\ell+1)}) = f(x,y)/f(x,y_{-(\ell+1)}) \); however, the original notation will be maintained for the sake of brevity.

Continue by inserting the appropriate density estimates.

\[
\hat{g}(z_p|z_{-p}) = \hat{f}(z)/\hat{f}(z_{-p})
\]

\[
= \left[ \frac{1}{n} \sum_{i=1}^{n} \phi_p(z|\hat{z}_{(i)}, \text{diag}(h^2)) \right] \left[ \frac{1}{n} \sum_{i=1}^{n} \phi_{p-1}(z_{-p}|\hat{z}_{(i),-p}, \text{diag}(h^2_{-p})) \right]
\tag{2.22}
\]

As in Step 3, factor \( \phi_p(z|...) \), but adjusting for the desired elements of \( z \).

\[
\phi_p(z|\hat{z}_{(i)}, \text{diag}(h^2)) = \phi(z_p|\hat{z}_{i,p}, h^2_p) \cdot \phi_{p-1}(z_{-p}|\hat{z}_{(i),-p}, \text{diag}(h^2_{-p}))
\tag{2.23}
\]

Insert this factorization and rearrange terms to produce the final density estimate.

\[
\hat{g}(z_p|z_{-p}) = \left[ \frac{1}{n} \sum_{i=1}^{n} \phi(z_p|\hat{z}_{i,p}, h^2_p) \cdot \phi_{p-1}(z_{-p}|\hat{z}_{(i),-p}, \text{diag}(h^2_{-p})) \right] \left[ \frac{1}{n} \sum_{i=1}^{n} \phi_{p-1}(z_{-p}|\hat{z}_{(i),-p}, \text{diag}(h^2_{-p})) \right]
\tag{2.24}
\]

\[
\hat{g}(z_p|z_{-p}) = \sum_{i=1}^{n} \left[ \phi(z_p|\hat{z}_{i,p}, h^2_p) \cdot \phi_{p-1}(z_{-p}|\hat{z}_{(i),-p}, \text{diag}(h^2_{-p})) \right]
\tag{2.25}
\]

\[
\alpha_i(z_{-p}|\hat{z}_{(i),-p}, \text{diag}(h^2_{-p})) = \phi_{p-1}(z_{-p}|\hat{z}_{(i),-p}, \text{diag}(h^2_{-p})) / \left[ \sum_{i=1}^{n} \phi_{p-1}(z_{-p}|\hat{z}_{(i),-p}, \text{diag}(h^2_{-p})) \right]
\tag{2.26}
\]

\[
\hat{g}(z_p|z_{-p}) \text{ is the estimate of the filtering density that will be used to simulate tracking errors (}\ell+2\text{) through } k \text{ of a new sample.}
\]

2.3.3 Simulation Routine

With the final density estimates ready, LETF tracking errors can be simulated using the following routine.

Given: length (\( \ell + k \)) input vector of log index returns \( u \), length \( p \) bandwidth vector \( h \), number of daily lags \( \ell \), and length \( p \) observation vectors \( \hat{z}_{(1)}, ..., \hat{z}_{(n)} \)

1. Sample an index value \( i \in 1, ..., n \) from a multinomial distribution with weight vector \( \alpha \), where

\[
\alpha_i = \phi_{\ell+1}(u_{1;\ell+1}|\hat{z}_{(i),x}, \text{diag}(h^2_x)) / \left[ \sum_{i=1}^{n} \phi_{\ell+1}(u_{1;\ell+1}|\hat{z}_{(i),x}, \text{diag}(h^2_x)) \right]
\]
2. Use index value $i$ from step 1 to select the $i$th observation vector $\hat{z}_{(i)}$ and sample $v_{1:(\ell+1)}$, an $(\ell + 1)$ length vector, from the multivariate normal distribution with density $\phi_{\ell+1}(\hat{z}_{(i)}, y, \text{diag}(h^2_p))$.

Perform steps 3 and 4 for $j = (\ell + 2), ..., (\ell + k)$

3. Sample an index value $i \in 1, ..., n$ from a multinomial distribution with weight vector $\alpha_i$, where $\alpha_i = \phi_{p-1}(u_{(j-\ell):j}, v_{(j-\ell):j}, -p, \text{diag}(h^2_p))/\sum_{i=1}^n \phi_{p-1}(u_{(j-\ell):j}, v_{(j-\ell):j}, -p, \text{diag}(h^2_p))$, and $[u_{(j-\ell):j}, v_{(j-\ell):j}]$ is the length $(p - 1)$ vector formed by concatenating the indicated elements of $u$ and $v$.

4. Use index value $i$ from step 3 to select the $i$th observation vector $\hat{z}_{(i)}$ and sample $v_j$ from the multivariate normal distribution with density $\phi(|\hat{z}_{i,p}, h^2_p)$.

5. Combine the simulated tracking errors $v_{(\ell+1):(\ell+k)}$ with the input index returns $u_{(\ell+1):(\ell+k)}$ according to our LETF returns decomposition.

$$w_t = \beta_{LETF} \ast (\exp(u_t) - 1) - \text{MgmtFee}_{LETF} + (\exp(v_t) - 1) \quad (2.28)$$

where $w$ is the final, length $k$ output vector of simulated LETF returns. Elements indexed $1 : \ell$ in vectors $u$, $v$, and $w$ are discarded.

When simulating data for the intended application, the sampled index returns to be used as inputs are first arranged as a matrix of rolling periods of $(\ell + k)$ days (and transformed to log returns). Regular choices of $k$ are 21 or 252. The simulation routine can be executed in parallel because there is no dependence on the input data across rows. Each row of the matrix is input as vector $u$ separately and the density sampling steps can execute on each row simultaneously. Step 5 can obviously execute in parallel with basic matrix algebra. The simulation routine is also put to use for estimating the lag parameter $\ell$, but the input data is not arranged into rolling periods initially. With lag estimation this step occurs after sampling using the simulated data in order to avoid simulating multiple tracking errors based on the same index returns.

### 2.3.4 Lag and Bandwidth Selection for Simulation of LETF Tracking Errors

Estimating the lag parameter $\ell$ proceeds as a brute force sweep of possible selections 0 to $N$ (for some reasonable $N$), using the simulation routine and a Kolmogorov-Smirnov test for evaluation of fit. For each LETF and index combination and choice of $\ell$ to be considered, $K$ iterations of tracking error simulation are performed (for some large $K$). The entire observation vector of daily index returns (log transformed) for the days which LETF returns are available is supplied as the input vector $u$. The routine is executed without the use of rolling periods, but this transformation is made after Step 5 has completed, such that the simulated LETF returns are arranged into a matrix of rolling periods of $k$ returns. The compound return is calculated for each row producing a final simulated sample of $k$ day LETF returns. These two steps are repeated with the actual LETF returns to produce the reference sample of $k$ day LETF returns, and then the samples are compared using a two-tailed Kolmogorov-Smirnov test. Final lag selection is based on the resulting $p$-values collected from $K$ iterations of the simulation routine and Kolmogorov-Smirnov test.

The K-S test is comparing a sample of $k$ day LETF returns derived from observed index returns and
simulated LETF tracking errors against a sample of \( k \) day LETF returns derived from observed index returns and observed LETF tracking errors (i.e. the observed LETF returns). Each iteration uses only the index returns observed on days when the paired LETF was trading to ensure that the two samples are derived from the same sequence of data. The null hypothesis in a K-S test states that the two samples being compared are drawn from the same distribution [Conover, 1999]. Our goal is to simulate tracking errors in a way that the K-S test of the LETF returns samples cannot reject the null hypothesis at sufficiently high confidence levels. Based on the results of the K-S tests, our goal translates to selecting the value of \( \ell \) that produces greater \( p \)-values compared to other selections. This generic mandate could be interpreted in many different ways, and we settled on selecting the value of \( \ell \) with the greatest proportion of \( p \)-values that were \( \geq 0.01 \).

For some LETFs there were multiple possible \( \ell \) selections where all resulting \( p \)-values were \( \geq 0.01 \). In these cases ties were broken by selecting the least value of \( \ell \) to avoid any possible overfitting. A \( p \)-value that is \( \geq 0.01 \) indicates that iteration of simulation produced a sample that could not be rejected at a confidence level \( \geq 0.99 \), the threshold of confidence we are in effect selecting to delineate as sufficiently high.

In Table 2.1, we present our final lag selections for the demonstration set of LETFs. Column headers prefixed with “21D” or “252D” refer to results produced when using \( k = 21 \) day or \( k = 252 \) day periods, respectively, and the notation \( P[p \geq 0.01] \) refers to the proportion of \( p \)-values that were \( \geq 0.01 \). We used 10,000 iterations for testing each LETF. We feel that our selection criteria resulted in satisfactory lag parameter selections for all LETFs when using 21 day periods (except for FINZ), but the results are mixed when using 252 day periods, with 9 of 20 LETFs (FINZ, DUST, TECL, TECS, RUSL, RUSS, DZK, EDC, EDZ) failing to offer values of \( P[p \geq 0.01] \) greater than 0.9. There were many lag selections for each LETF that performed well when using 21 day periods, so we simplified the estimation task and present only the lag selection that provided the best results with 252 day periods (the values shown in columns with a “21D” prefix may not actually be the best). The minimum and median \( p \)-values were not incorporated into any decision making, but are included here for reference.

<table>
<thead>
<tr>
<th>Index</th>
<th>Leverage</th>
<th>LETF</th>
<th>Lags</th>
<th>21D</th>
<th>21D</th>
<th>21D</th>
<th>252D</th>
<th>252D</th>
<th>252D</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>21D</td>
<td>21D</td>
<td>21D</td>
<td>252D</td>
<td>252D</td>
<td>252D</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>P[p &gt;= 0.01]</td>
<td>Minimum</td>
<td>Median</td>
<td>P[p &gt;= 0.01]</td>
<td>Minimum</td>
<td>Median</td>
</tr>
<tr>
<td>DJUSFN</td>
<td>+3</td>
<td>FINU</td>
<td>3</td>
<td>1.0000</td>
<td>0.7125</td>
<td>0.9963</td>
<td>1.0000</td>
<td>0.0354</td>
<td>0.6315</td>
</tr>
<tr>
<td>DJUSFN</td>
<td>-3</td>
<td>FINZ</td>
<td>3</td>
<td>0.6034</td>
<td>0.0000</td>
<td>0.0242</td>
<td>0.0511</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>GDM</td>
<td>+3</td>
<td>NUGT</td>
<td>5</td>
<td>1.0000</td>
<td>0.1592</td>
<td>0.4248</td>
<td>0.9986</td>
<td>0.0053</td>
<td>0.0644</td>
</tr>
<tr>
<td>GDM</td>
<td>-3</td>
<td>DUST</td>
<td>3</td>
<td>1.0000</td>
<td>0.0466</td>
<td>0.3639</td>
<td>0.5112</td>
<td>0.0000</td>
<td>0.0107</td>
</tr>
<tr>
<td>INDU</td>
<td>+3</td>
<td>UDOW</td>
<td>3</td>
<td>1.0000</td>
<td>0.9962</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.5142</td>
<td>1.0000</td>
</tr>
<tr>
<td>INDU</td>
<td>-3</td>
<td>SDOW</td>
<td>5</td>
<td>1.0000</td>
<td>0.9993</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.4442</td>
<td>0.9994</td>
</tr>
<tr>
<td>IXT</td>
<td>+3</td>
<td>TECL</td>
<td>4</td>
<td>0.9996</td>
<td>0.0059</td>
<td>0.5625</td>
<td>0.5942</td>
<td>0.0000</td>
<td>0.0450</td>
</tr>
<tr>
<td>IXT</td>
<td>-3</td>
<td>TECS</td>
<td>5</td>
<td>0.9953</td>
<td>0.0006</td>
<td>0.3224</td>
<td>0.5057</td>
<td>0.0000</td>
<td>0.0107</td>
</tr>
<tr>
<td>MVRSX</td>
<td>+3</td>
<td>RUSL</td>
<td>3</td>
<td>0.9997</td>
<td>0.0086</td>
<td>0.7218</td>
<td>0.1339</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>MVRSX</td>
<td>-3</td>
<td>TECS</td>
<td>4</td>
<td>1.0000</td>
<td>0.0183</td>
<td>0.7212</td>
<td>0.5420</td>
<td>0.0000</td>
<td>0.0180</td>
</tr>
<tr>
<td>MXE A</td>
<td>+3</td>
<td>DZK</td>
<td>3</td>
<td>0.9999</td>
<td>0.0067</td>
<td>0.6241</td>
<td>0.0268</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>MXE A</td>
<td>-3</td>
<td>DPK</td>
<td>4</td>
<td>1.0000</td>
<td>0.2803</td>
<td>0.9546</td>
<td>0.9429</td>
<td>0.0000</td>
<td>0.1643</td>
</tr>
<tr>
<td>MXE F</td>
<td>+3</td>
<td>EDC</td>
<td>3</td>
<td>1.0000</td>
<td>0.0411</td>
<td>0.3948</td>
<td>0.0455</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>MXE F</td>
<td>-3</td>
<td>EDZ</td>
<td>4</td>
<td>1.0000</td>
<td>0.2243</td>
<td>0.9398</td>
<td>0.6698</td>
<td>0.0000</td>
<td>0.0403</td>
</tr>
<tr>
<td>NDX</td>
<td>+3</td>
<td>TQQQ</td>
<td>3</td>
<td>1.0000</td>
<td>0.9997</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.6649</td>
<td>1.0000</td>
</tr>
<tr>
<td>NDX</td>
<td>-3</td>
<td>SQQQ</td>
<td>3</td>
<td>1.0000</td>
<td>0.9999</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.5509</td>
<td>0.9983</td>
</tr>
<tr>
<td>RTY</td>
<td>+3</td>
<td>TNA</td>
<td>3</td>
<td>1.0000</td>
<td>0.9582</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.0911</td>
<td>0.9682</td>
</tr>
<tr>
<td>RTY</td>
<td>-3</td>
<td>TZA</td>
<td>4</td>
<td>1.0000</td>
<td>0.9986</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.0121</td>
<td>0.8987</td>
</tr>
<tr>
<td>SPX</td>
<td>+3</td>
<td>SPXL</td>
<td>5</td>
<td>1.0000</td>
<td>0.9912</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.0605</td>
<td>1.0000</td>
</tr>
<tr>
<td>SPX</td>
<td>-3</td>
<td>SPXS</td>
<td>5</td>
<td>1.0000</td>
<td>0.9912</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.1100</td>
<td>1.0000</td>
</tr>
</tbody>
</table>
Our selections for the kernel bandwidth vector $\mathbf{h}$ are based on the multivariate “Scott Rule” [Scott, 1992], defined as:

$$\hat{h}_j = \hat{\sigma}_j \ast n^{-1/(p+4)}$$

(2.29)

where $\hat{h}_j$ is the bandwidth for $j$th dimension of density estimate, $\hat{\sigma}_j$ is the estimated standard deviation of $j$th dimension, $n$ is the number of $p$ dimensional observations used for fitting density estimate, and $p$ is the number of dimensions of the sample space.

We use $s^2$, the traditional unbiased estimator of $\sigma^2$, and scale down our final selections of $\hat{h}_j$. After extensive testing of our tracking error simulation routine with the demonstration set of 20 LETFs, we concluded that the elements of $\hat{h}$ corresponding to index returns (elements 1 : $(\ell + 1)$) needed to be scaled down by a factor of 100, and the elements of $\hat{h}$ corresponding to LETF tracking errors (elements $(\ell + 2) : p$) needed to be scaled down by a factor of 100,000. We provide visual justification for this heuristic in Figure 2.4. For the final bandwidth selections, we have:

$$\hat{h}_j = s_j \ast n^{-1/(p+4)} \ast 0.01, \text{ for } j = 1, ..., (\ell + 1)$$

(2.30)

$$\hat{h}_j = s_j \ast n^{-1/(p+4)} \ast 0.000001, \text{ for } j = (\ell + 2), ..., p$$

(2.31)

$$s_j = \left[ \frac{1}{n-1} \sum_{i=1}^{n} (z_{i,j} - \bar{z}_j)^2 \right]^{\frac{1}{2}}$$

(2.32)

$$\bar{z}_j = \frac{1}{n} \sum_{i=1}^{n} \hat{z}_{i,j}$$

(2.33)

Figure 2.4 shows the lag testing results from 10,000 iterations of LETF Ticker SPXL using 252 day periods for each lag parameter value 0 through 8. In all simulations, the bandwidths were scaled down as indicated above, producing in our opinion a reasonable increase in K-S $p$-values (using any of the statistics shown) as the lag value increases from 0 to 8. These results are somewhat representative of the results experienced when testing the other LETFs. Alternatively, when using larger factors to scale the bandwidths, the simulation routine would fail on some iterations because the estimates for the smoothing and filtering densities did not have sufficient support. We experienced this type of result when testing FINZ, TECS, RUSL, and RUSS with the factors indicated above, forcing us to create an exception for these LETFs where we decrease the factor used to scale the tracking error portion of the bandwidth vector to 10,000. When using smaller factors to scale the bandwidths, significant increases in $p$-values were generally not observed at any lag, or not observed until 8 or more lags were included. In the latter cases, these results appeared to us as overfitting.

2.4 Constrained Simulation Using Nonparametric Density Estimation

The nonparametric method for constrained simulation we propose in this section was devised as part of the larger body of work on LETF volatility documented throughout this chapter, but the method is truly general in nature. It could be helpful in any setting where the practitioner would like to simulate multivariate data with a constrained sample sum (or average). For this reason the second and third subsections are presented in the context of using generic multivariate data, without references to financial data or financial constructs.
Figure 2.4: Lag Testing Results Summary for LETF Ticker SPXL with 252 Day Periods

2.4.1 Extending the LETF Returns Model with an Index Return Constraint

In order to constrain the compound underlying index return for the period, the joint density of our LETF returns model from Section 2.3 must be revised as follows:

\[
 f(R_{Index,1}, \ldots, R_{Index,\ell+k-1}, \epsilon_{LETF,1}, \ldots, \epsilon_{LETF,\ell+k}|R_0) \\
= \prod_{j=\ell+2}^{\ell+k} g(\epsilon_{LETF,j}|R_{Index,j}, R_{Index,j-\ell}, \ldots, R_{Index,j-1}, \epsilon_{LETF,j-\ell}, \ldots, \epsilon_{LETF,j-1}) \\
\times g(\epsilon_{LETF,1}, \ldots, \epsilon_{LETF,\ell+1}|R_{Index,1}, \ldots, R_{Index,\ell+1}) \times f(R_{Index,1}, \ldots, R_{Index,\ell+k-1}|R_0) \tag{2.34}
\]

where \( R_0 = \prod_{j=\ell+1}^{\ell+k} (1 + R_{Index,j}) - 1 \), the total \( k \) day compounded return of the underlying index for days \((\ell+1)\) through \((\ell+k)\), and the density function \( f(R_{Index,1}, \ldots, R_{Index,\ell+k-1}|R_0) \) is the joint density of a sequence of \( \ell \) unconstrained index return variables (the lags) and a sequence of \((k-1)\) index return variables constrained to compound to \( R_0 \). This form of the full joint density for index returns and LETF tracking errors (left-hand side) can be used to simulate sequences of LETF returns constrained such that the sampled sequence of index returns compounds to the given total period return \( R_0 \).

It needs to be emphasized that this modification is implemented using the log form of index returns as was the case with all estimators and simulation work featured in Section 2.3. When index return variables are modeled as \( X = \log(1 + R_{Index}) \), the density being estimated becomes \( f(X_1, \ldots, X_{\ell+k-1}|X_0 = \sum_{j=\ell+1}^{\ell+k-1} X_j) \), the joint density of a sequence of \( \ell \) unconstrained variables (the lags) and a sequence of \((k-1)\) variables constrained to sum to \( X_0 \). The estimate for this form of the constrained joint density is much easier to derive and more generally useful in applications outside of finance.
2.4.2 Derivation of Kernel Density Estimate

Given any $p$ dimensional random vector $X$ with $\ell \in [0, \ldots, (p - 2)]$ unconstrained dimensions (lags) and $n$ observations of $X$ arranged as an $(n \times p)$ observation matrix $\mathbf{X}$, our nonparametric density estimate is constructed as follows:

1. Fit a kernel density estimate for $f(x)$ using a jointly independent multivariate normal kernel and observation matrix $\mathbf{X}$.

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \phi_p(x | \hat{x}_{(i)}, \text{diag}(h^2)) \quad (2.35)$$

where $\phi_p(x | \mu, \Sigma)$ is the $p$ dimensional multivariate normal density function evaluated at $x$ using mean vector $\mu$ and covariance matrix $\Sigma$, $\hat{x}_{(i)}$ is the $i$th row vector of observation matrix $\mathbf{X}$, and $\text{diag}(h^2)$ is the diagonal matrix formed from element-wise squaring of the $p$ dimensional vector of bandwidths.

Note that the selection of $h$ is ultimately application dependent, so we cannot provide advice for any kind of general case other than recommending what has already been published about using the multivariate normal density as a kernel [Scott, 1992]. The selections for our applications will be detailed later in a later subsection.

2. Perform a change of variables and derive the corresponding kernel density estimate.

Let $Y_1 = X_1$, $Y_2 = X_2$, $Y_3 = X_3$, ..., $Y_p = \sum_{j=\ell+1}^{p} X_j$. Or using matrix notation, define random vector $\mathbf{Y}$ as follows:

$$\mathbf{Y} = \mathbf{T} \mathbf{X} \quad (2.36)$$

where $\mathbf{T}$ is a $(p \times p)$ matrix

$$\mathbf{T} = \begin{bmatrix}
1 & \cdots & 0 & 0 & 0 & \cdots & 0 \\
0 & \ddots & 0 & 0 & 0 & \cdots & 0 \\
0 & \cdots & 1 & 0 & 0 & \cdots & 0 \\
0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\
0 & \cdots & 0 & 0 & 1 & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & \cdots & 0 & 1 & 1 & 1 & \cdots & 1
\end{bmatrix}$$

where the first $\ell$ elements of the bottom row of $\mathbf{T}$ are 0 to allow for the unconstrained dimensions of $\mathbf{X}$.

To derive the kernel density estimate for the transformed variables, invoke the following theorem for transforming multivariate normal random vectors [Mardia et al., 1979]:

If $\mathbf{X} \sim \phi(x | \mu, \Sigma)$ and $\mathbf{Y} = A\mathbf{X} + \mathbf{c}$, then $\mathbf{Y} \sim \phi(y | A\mu + \mathbf{c}, A\Sigma A') \quad (2.38)$

where $A$ is a constant matrix, $A'$ is the transpose of that matrix, and $\mathbf{c}$ is a constant vector. Apply the
theorem to the kernel from the result of Step 1 to produce the estimate of the joint density for $Y$.

$$
\hat{f}(y) = \frac{1}{n} \sum_{i=1}^{n} \phi_p(y | T \hat{x}_i, T \text{diag}(h^2) T')
$$

(2.39)

3. Factor the kernel within $\hat{g}(y)$ into the product of an unconditional univariate kernel for $y_p$ and a multivariate kernel for $y_1, y_2, ..., y_{p-1}$ conditional on $y_p$.

$$
\phi_p(y | \ldots) = \phi(y_p | \ldots) \phi_{p-1}(y_p | y_p, \ldots)
$$

(2.40)

where $y_{-p}$ is the vector $y$ without the $p$th element.

For the kernel factor corresponding to the univariate density for $Y_1$, re-apply the theorem cited in Step 2 to construct the following univariate density kernel:

$$
\phi(y_p | \ldots) = \phi(y_p | \sum_{j=\ell+1}^{p} \hat{x}_{i,j}, \sum_{j=\ell+1}^{p} h_j^2)
$$

(2.41)

To construct $\phi_{p-1}(y_p | y_p, \ldots)$ invoke another theorem for transforming multivariate normal random vectors [Mardia et al., 1979]:

If $X \sim \phi_p(x | \mu, \Sigma)$, then $X_{-1} | X_1 \sim \phi_{p-1}(x_{-1} | \mu_{-1} + \Sigma_{2,1}(\sigma_1^2)^{-1}(x_1 - \mu_1), \Sigma_{2,2} - \Sigma_{2,1}(\sigma_1^2)^{-1}\Sigma_{1,2})$

(2.42)

$$
\Sigma = \begin{bmatrix}
\sigma_1^2 & \Sigma_{1,2} \\
\Sigma_{2,1} & \Sigma_{2,2}
\end{bmatrix}
$$

(2.43)

where $\sigma_1^2$, $\Sigma_{1,2}$, $\Sigma_{2,1}$, and $\Sigma_{2,2}$ taken together constitute a partition of $\Sigma$. Apply this theorem to the conditional kernel.

$$
\phi_{p-1}(y_p | \ldots) = \phi_{p-1}(y_p | \hat{x}_{i,-p}, \hat{\mathcal{H}})
$$

(2.44)

$$
\hat{x}_{i,-p} = \hat{x}_{i,-p} + h_{-p}^2 \left( \sum_{j=\ell+1}^{p} h_j^2 \right)^{-1} (y_p - \sum_{j=\ell+1}^{p} \hat{x}_{i,j})
$$

(2.45)

$$
\hat{\mathcal{H}} = \text{diag}(h_{-p}^2) - h_{-p}^2 \left( \sum_{j=\ell+1}^{p} h_j^2 \right)^{-1} h_{-p}^2
$$

(2.46)

where $\hat{x}_{i,-p}$ is the $i$th observation vector without the $p$th element, and $h_{-1}^2$ is the bandwidth vector $h$, without the $p$th element, squared element-wise, and transposed.

4. Factor and rearrange terms of the density estimate for $f(y)$ to produce an estimate for the conditional density $g(y_{-p} | y_p)$.
Start by factoring the density for \( f(y) \) and inserting the appropriate density estimates.

\[
g(y_p|y_p) = f(y)/f(y_p) \quad (2.47)
\]

\[
\hat{g}(y_p|y_p) = f(y)/\hat{f}(y_p) \quad (2.48)
\]

\[
\hat{g}(y_p|y_p) = \frac{1}{n} \sum_{i=1}^{n} \phi_p(y|T\hat{x}_i, T\text{diag}(h^2)^T)/[\frac{1}{n} \sum_{i=1}^{n} \phi(y_p| \sum_{j=\ell+1}^{p} \hat{x}_{i,j}, \sum_{j=\ell+1}^{p} h_j^2)] \quad (2.49)
\]

Continue by inserting the factors of \( \phi_p(y|...) \) that were derived in Step 3 and rearranging terms to produce the final density estimate.

\[
\hat{g}(y_p|y_p) = \frac{1}{n} \sum_{i=1}^{n} [\phi(y_p| \sum_{j=\ell+1}^{p} \hat{x}_{i,j}, \sum_{j=\ell+1}^{p} h_j^2)] \phi_{p-1}(y_p|\hat{x}_i, -p, \hat{H})/ \quad (2.50)
\]

\[
\hat{g}(y_p|y_p) = \sum_{i=1}^{n} \phi_{p-1}(y_p|\hat{x}_i, -p, \hat{H}) \phi_i(y) \sum_{j=\ell+1}^{p} \hat{x}_{i,j}, \sum_{j=\ell+1}^{p} h_j^2)] \quad (2.51)
\]

\[
\alpha_i(y_p| \sum_{j=\ell+1}^{p} \hat{x}_{i,j}, \sum_{j=\ell+1}^{p} h_j^2) = \phi(y_p| \sum_{j=\ell+1}^{p} \hat{x}_{i,j}, \sum_{j=\ell+1}^{p} h_j^2)/ \quad (2.52)
\]

Because there was no effective change in the \( X_1, ..., X_{p-1} \) to \( Y_1, ..., Y_{p-1} \) transformations, define \( x_0 = y_p - \sum_{j=\ell+1}^{p} x_j \) to denote the given value for the sum of the random vector and the density estimate can be written entirely in terms of \( X \). The final form of the density estimate:

\[
\hat{g}(x_{-p}|x_0) = \sum_{i=1}^{n} [\phi_{p-1}(x_{-p}|\hat{x}_i, -p, \hat{H}) \phi_i(x_0) \sum_{j=\ell+1}^{p} \hat{x}_{i,j}, \sum_{j=\ell+1}^{p} h_j^2)] \quad (2.53)
\]

\[
\alpha_i(x_0| \sum_{j=\ell+1}^{p} \hat{x}_{i,j}, \sum_{j=\ell+1}^{p} h_j^2) = \phi(x_0| \sum_{j=\ell+1}^{p} \hat{x}_{i,j}, \sum_{j=\ell+1}^{p} h_j^2)/ \quad (2.54)
\]

### 2.4.3 Simulation Routine

With the final density estimates ready, constrained data can be simulated using the following routine.

Given: length \( p \) bandwidth vector \( h \), vector sum \( \sum_{j=\ell+1}^{p} x_j \) constraint \( x_0 \), and length \( p \) observation vectors \( \hat{x}_{(1)}, ..., \hat{x}_{(n)} \)

1. Sample an index value \( i \in 1, ..., n \) from a multinomial distribution with weight vector \( \alpha \), where

\[
\alpha_i = \phi(x_0| \sum_{j=\ell+1}^{p} \hat{x}_{i,j}, \sum_{j=\ell+1}^{p} h_j^2)/[\sum_{i=1}^{n} \phi(x_0| \sum_{j=\ell+1}^{p} \hat{x}_{i,j}, \sum_{j=\ell+1}^{p} h_j^2)].
\]

2. Use index value \( i \) from step 1 to select the \( i \)th observation \( \hat{x}_{(i)} \) and sample \( x_{-p} \), a \( (p-1) \) dimensional vector, from the multivariate normal distribution with density \( \phi_{p-1}(\cdot|\hat{x}_{(i)}, -p, \hat{H}) \), where \( \hat{x}_{(i)} = ... \)
\[ \dot{x}_{(i),-p} + h_{-p}^2 (\sum_{j=\ell+1}^p \dot{x}_{i,j})^{-1} (x_0 - \sum_{j=\ell+1}^p \dot{x}_{i,j}), \text{ and } H = \text{diag}(h_{-p}^2) - h_{-p}^2 (\sum_{j=\ell+1}^p h_{j}^2)^{-1} h_{-p}^2. \]

3. Use elements \((\ell + 1)\) through \((p - 1)\) of the the sampled \((p - 1)\) dimensional vector \(x_{-p}\) from Step 2 to arrive at the \(p\)th (implied) sample value \(x_p = x_0 - \sum_{j=\ell+1}^{p-1} x_j.\)

Because Steps 1 and 3 are computationally trivial, the computational complexity of the simulation routine is no worse than that of the chosen implementation for the sampled multivariate normal distribution. Our testing and development was performed using the \texttt{mvrnorm()} implementation available in the R package \texttt{MASS}. Package version 7.3-27, R version 3.0.1.

Table 2.2 is presented to better illustrate what real time performance a user can expect with the routine (based on our choice of \texttt{mvrnorm()}). The routine was executed 100 times each using various sizes of the observation matrix \((n \times p)\), and for the sake of completeness included various values for number of iterations \(k\). For all simulations the returns data for the S&P500 index were used with a constrained return of 0 and our final bandwidth selection (described in the next subsection). As the iterations of the routine can be executed in parallel, theoretically, the number of iterations \(k\) should have no impact on the final runtime. The tests were performed on a Dell server with 2 Intel Xeon 2.5 GHz processors (6 cores each, 64 GB memory total) running a virtualized instance of Red Hat Enterprise Linux Server 6.5 (Kernel 2.6, 64 bit).

As expected there are significant increases in execution time at the largest selections for \(n\) and \(p\), while \(k\) does not have much of an impact. With an average execution time near 3 minutes when using a \((100,000 \times 1,000)\) observation matrix we hope that interested readers will not consider runtime as a drawback against using our routine in practice.
Table 2.2: Execution Times from 100 iterations per parameter set (seconds)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th>mean</th>
<th>min</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,000</td>
<td>10</td>
<td>1,000</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>1,000</td>
<td>10</td>
<td>10,000</td>
<td>0.05</td>
<td>0.05</td>
<td>0.10</td>
</tr>
<tr>
<td>1,000</td>
<td>10</td>
<td>100,000</td>
<td>0.70</td>
<td>0.67</td>
<td>0.84</td>
</tr>
<tr>
<td>1,000</td>
<td>100</td>
<td>1,000</td>
<td>0.05</td>
<td>0.05</td>
<td>0.06</td>
</tr>
<tr>
<td>1,000</td>
<td>100</td>
<td>10,000</td>
<td>0.51</td>
<td>0.42</td>
<td>0.57</td>
</tr>
<tr>
<td>1,000</td>
<td>100</td>
<td>100,000</td>
<td>3.79</td>
<td>3.63</td>
<td>4.01</td>
</tr>
<tr>
<td>1,000</td>
<td>1,000</td>
<td>1,000</td>
<td>4.50</td>
<td>4.17</td>
<td>4.67</td>
</tr>
<tr>
<td>1,000</td>
<td>1,000</td>
<td>10,000</td>
<td>13.95</td>
<td>13.72</td>
<td>14.29</td>
</tr>
<tr>
<td>1,000</td>
<td>1,000</td>
<td>100,000</td>
<td>108.81</td>
<td>108.24</td>
<td>110.41</td>
</tr>
<tr>
<td>10,000</td>
<td>10</td>
<td>1,000</td>
<td>0.05</td>
<td>0.04</td>
<td>0.05</td>
</tr>
<tr>
<td>10,000</td>
<td>10</td>
<td>10,000</td>
<td>0.10</td>
<td>0.09</td>
<td>0.16</td>
</tr>
<tr>
<td>10,000</td>
<td>10</td>
<td>100,000</td>
<td>0.84</td>
<td>0.68</td>
<td>0.96</td>
</tr>
<tr>
<td>10,000</td>
<td>100</td>
<td>1,000</td>
<td>0.24</td>
<td>0.23</td>
<td>0.31</td>
</tr>
<tr>
<td>10,000</td>
<td>100</td>
<td>10,000</td>
<td>0.71</td>
<td>0.68</td>
<td>0.77</td>
</tr>
<tr>
<td>10,000</td>
<td>100</td>
<td>100,000</td>
<td>3.99</td>
<td>3.93</td>
<td>4.15</td>
</tr>
<tr>
<td>10,000</td>
<td>1,000</td>
<td>1,000</td>
<td>11.24</td>
<td>11.07</td>
<td>12.23</td>
</tr>
<tr>
<td>10,000</td>
<td>1,000</td>
<td>10,000</td>
<td>20.67</td>
<td>20.35</td>
<td>21.76</td>
</tr>
<tr>
<td>10,000</td>
<td>1,000</td>
<td>100,000</td>
<td>115.92</td>
<td>114.91</td>
<td>117.25</td>
</tr>
<tr>
<td>100,000</td>
<td>10</td>
<td>1,000</td>
<td>0.46</td>
<td>0.40</td>
<td>0.53</td>
</tr>
<tr>
<td>100,000</td>
<td>10</td>
<td>10,000</td>
<td>0.51</td>
<td>0.48</td>
<td>0.58</td>
</tr>
<tr>
<td>100,000</td>
<td>10</td>
<td>100,000</td>
<td>1.16</td>
<td>1.05</td>
<td>1.27</td>
</tr>
<tr>
<td>100,000</td>
<td>100</td>
<td>1,000</td>
<td>2.41</td>
<td>2.33</td>
<td>2.63</td>
</tr>
<tr>
<td>100,000</td>
<td>100</td>
<td>10,000</td>
<td>2.77</td>
<td>2.73</td>
<td>2.91</td>
</tr>
<tr>
<td>100,000</td>
<td>100</td>
<td>100,000</td>
<td>6.14</td>
<td>6.01</td>
<td>6.28</td>
</tr>
<tr>
<td>100,000</td>
<td>1,000</td>
<td>1,000</td>
<td>80.94</td>
<td>79.40</td>
<td>91.49</td>
</tr>
<tr>
<td>100,000</td>
<td>1,000</td>
<td>10,000</td>
<td>89.85</td>
<td>88.91</td>
<td>102.12</td>
</tr>
<tr>
<td>100,000</td>
<td>1,000</td>
<td>100,000</td>
<td>185.86</td>
<td>184.00</td>
<td>197.93</td>
</tr>
</tbody>
</table>

2.4.4 Bandwidth Selection for Constrained Simulation of Index Returns

Our selection here is similar to the final rule presented in Section 2.3. Again we begin with the multivariate “Scott Rule” [Scott, 1992], defined as it was previously.

In the case of simulating only index returns, we simplify the initial rule because the $p$ dimensional row vectors of the observation matrix used to fit the density estimate are rolling $p$ day intervals of the same series of daily data. Therefore, a single value from the average of the $\hat{\sigma}_j$ terms will be used because individually they will be nearly identical. Also, we use $s^2$, the traditional unbiased estimator of $\sigma^2$.

$$h = \left[ \frac{1}{n} \sum_{j=1}^{p} s_j \right] n^{-1/(p+4)} \cdot 1_p$$

(2.55)

$$s_j = \left[ \frac{1}{n-1} \sum_{i=1}^{n} (x_{i,j} - \bar{x}_j)^2 \right]^{1/2}$$

(2.56)

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{i,j}$$

(2.57)

where $1_p$ is a $p$ dimensional vector of 1s. After extensive testing of our simulation routine with the 10 indexes included in the demonstration set, we concluded that our choice of $h$ needed to be scaled down by a factor of 10. We provide visual justification for this heuristic in Figures 2.5 and 2.6.

In the top half of Figure 2.5, the density for the sample distribution of SMC (colored blue) is plotted
using rolling 21 day intervals from the observed history of the S&P500 restricted to where the compound return for the period was between 1% and 2% ($0.01 \geq r \geq 0.02$). In the bottom half, the densities for the sample distribution of SMC are plotted using simulated data with (Adjustment Factor = 10, colored red) and without (Adjustment Factor = 1, colored green) scaling down the bandwidth. For the simulations, the period return constraint was set to 1.5% ($x_0 = \log(1 + 0.015)$), the midpoint of the data interval used to generate the density in the top half of the plot. Figure 2.6 follows the same layout and color scheme as Figure 2.5, but this time using rolling 252 day intervals and different target returns. Only periods where the compound return on the S&P500 was between 10% and 15% ($0.1 \geq r \geq 0.15$) were used in the top half, and the return constraint for simulated data in the lower half was 12.5% ($x_0 = \log(1 + 0.125)$). For all SMC calculations (using observed or simulated data), a leverage multiple of +3 is included as if we were approximating LETF Ticker SPXL, where $(R_j = 3R_{SPX,j} - \text{MgmtFee}_{SPXL})$ is the series of daily returns used for testing. Shortfall from Maximum Convexity was used as the test statistic to stay consistent with the rest of our work, but an adjustment factor of 10 appeared equally appropriate when standard deviation was used.

In the figures shown (and including many other cases not shown), it was obvious to us that the bandwidths derived from our simplified Scott Rule were producing simulated data that looked exaggerated when plotted against the comparable subset of observed data. We decided on a final scaling factor of 10 through trial and error, but we think the results shown in Figures 2.5 and 2.6 speak for themselves. Our final bandwidth selection used in the following application examples is then $\hat{h} = 0.1 * \left( \frac{1}{p} \sum_{j=1}^{p} s_j \right) * n^{-1/(p+4)} * 1_p$.

![Figure 2.5: 21 Day SMC Densities for SPX](image-url)
2.5 Case Study: Simulating LETF Returns with Underlying Index Return Constraints and Calculating Nonparametric $p$-values for Volatility Statistics

We conclude with a demonstration of our three contributions in one example. Our intention is to capture the significance, or rarity, of the volatility experienced during the chosen period as measured by Shortfall from Maximum Convexity. A $p$-value for the observed SMC value will be calculated with respect to a SMC sample distribution generated by our LETF return model. In the experimental setting described here a $p$-value of 1 means the lowest possible volatility was observed (if the period in question was repeated, greater volatility would be experienced 100% of the time), and 0 means the highest possible volatility was observed (if the period in question was repeated, greater volatility would be experienced 0% of the time).

The absolute low point of the U.S. equity markets during the 2008-2009 recession occurred in early March 2009, so we will analyze the realized volatility for that month (22 days) as experienced by LETF Ticker TNA. The underlying index, the Russell 2000, returned 8.92% for the month (the market low point was early in the month, and the rebound was underway by month end). The log form of this value is used as the input total period return of our constrained simulation routine from Section 2.4 to sample 100,000 sequences of 22 days of Russell 2000 returns. For each of the 100,000 index return samples, our tracking error simulation routine is executed to sample sequences of TNA tracking errors. The data are combined with a +3 leverage multiple (and daily management fee) to produce a final sample of 22 day TNA returns sequences.

As displayed by the row for Ticker TNA in Table 2.1, the lag parameter $\ell$ for the simulations was set to 3. The history of daily returns for the Russell 2000 index dating back to 1/2/1979 was arranged into
a data matrix of 9,151 observations of rolling 25 day intervals (22 days plus 3 lags) to derive the kernel density estimate for constrained simulation. The history of daily returns for the LETF Ticker TNA dating back to 11/11/2008 was combined with the matching Russell 2000 daily returns into a data matrix of 1,536 observations of 8 dimensions (4 day sequence of index returns, 4 day sequence of LETF tracking errors) to derive the kernel density estimate for tracking error simulation. All observed values were log transformed before simulation, and our bandwidth selections described in Sections 2.3 and 2.4 were applied accordingly.

Our realized volatility statistic SMC was calculated for each 22 day sample and the density of the final SMC sampling distribution is plotted below in Figure 2.7. The red line marks the actual SMC value of 0.1527 that was observed for March 2009. The resulting $p$-value for this observation with respect to the simulated sample is 0.0053, which indicates that the proportion of simulated SMC values that were less than or equal to the observed value is 0.9947.

Our constrained simulation routine for index returns guarantees all the 22 day index return samples compound to the observed index return of 8.92%. Given this fact, the range of outcomes represented by the SMC sample is entirely the result of index return volatility and tracking error. Based on the output of our simulation methods we conclude that the realized volatility for TNA in March 2009 was greater than the overwhelming majority of outcomes that may be observed while the Russell 2000 experiences an 8.92% increase over the same period.

Figure 2.7: Simulated 22 Day SMC Density for LETF Ticker TNA, with all underlying index returns sequences compounding to 8.92%
2.6 Closing Remarks

Leveraged ETF returns are unique due to the dynamics produced by daily compounding, and we hope to have illuminated some of the challenges involved with modeling these returns over time horizons longer than one day. We have presented a nonparametric framework for modeling daily LETF returns that features a simulation method for LETF tracking errors and a method for simulating index returns with a constrained period return. The two methods allow for incorporating additional index returns data and controlling for index performance, respectively. We feel that these contributions can improve the public understanding of the relationship between LETF return volatility and index returns.

Our methods can model almost any dependence structure that may exist in LETF returns data and each can be used separately if desired. When used together, the two methods complement each other nicely, enabling powerfully customized risk summaries of LETF investing strategies. Finally, the constrained simulation method is not limited to our particular index returns dataset. It can be used to sample any generic multivariate data with a constraint on the sum of the random vector elements.
Part II

Firm Failure Events Dataset
Chapter 3

Framing Firm Failure and Preliminary Model Evaluation
3.1 Framing the Firm Failure Data Problem

This chapter reviews the firm failure model presented in “Predicting Financial Distress and the Performance of Distressed Stocks”, by John Campbell, Jens Hilscher, and Jan Szilagyi [Campbell et al., 2011]. Going forward we will refer to that paper by the lead author’s last name, Campbell. From the work of Campbell and other authors that came before, our research has been motivated by how they framed the problem of predicting bankruptcy, or “failure” in a more general sense, of exchange listed firms. The framing of the prediction of firm failure as a data problem serves us nicely because of the large amounts of data available, established baseline of performance against which we can evaluate our results, and inherent difficulty of the dataset (less than 0.01% of observations in class 1) that may prove amenable to more sophisticated methods.

Let us distinguish our work from that of Campbell by highlighting that our concern lies with developing improved methods for rare events binary classification, and more specifically improved non-parametric methods for this setting. Our featured modeling application presented later in Chapter 4 is our implementation of the firm failure model and we hope to achieve improved prediction accuracy for that data problem, not to provide empirical support for the theory underlying the dynamics of firm failure. On the other hand, the authors of the papers we cite in this chapter tend to discuss their failure models in the context of economic theory, and how this knowledge is validated vis-a-vis model validation with an emphasis on the statistical significance of their chosen predictor variables.

The remainder of this chapter contains a review of relevant firm failure literature in Section 3.1.1, a summary of the Campbell model and our dataset we will use for implementing it in Section 3.1.2, a thorough review of the test results from our implementation in Section 3.2, and a discussion of extending the Campbell dataset in Section 3.3. A comprehensive technical reference of our exact reconstruction of the firm failure dataset as present in Campbell is also included in Appendices A.1 and A.2.

3.1.1 Review of Firm Failure Literature

In the financial literature, firm failure has been studied extensively for decades, covering sub-topics such as bankruptcy risk, default risk, and the correlations of these risks with equity returns. The oldest work that we cite is Beaver (1966), where six groupings of firm financial ratios are evaluated for their explanatory power with respect to failure [William H. Beaver, 1966]. Three of these groupings include predictors that are similar to what is used in our approach: net income ratios, debt to total asset ratios, and liquid asset to total asset ratios. Another similarity with Beaver’s work is that the model target variable is not simply a bankruptcy or default indicator, but a general purpose failure indicator that measures the union of bankruptcy, default, preferred dividend non-payment, or overdrawn bank account events. Expanding the set of events beyond just bankruptcy (or default) increases the count of minority class observations that will be included in the resulting dataset, and in a rare events problem setting this can decrease the difficulty.

Another historically significant work on failure modeling is Altman (1968) that formally defines the “Z-Score” measure of distress risk [Altman, 1968]. This measure is the output of a linear discriminant analysis (LDA) of a binary indicator of bankruptcy with predictors that are similar to those used by Beaver. Ohlson (1980) described what could be considered a successor measure “O-Score” based on logistic regression instead of LDA and using a slightly expanded set of predictors [Ohlson, 1980]. This transition to logistic regression is supported by Lo (1986) because of the fact that logistic regression can be applied to a larger class of underlying distributions compared to LDA [Lo, 1986].

Another commonly used model of distress risk known as “Distance to Default” is derived from Merton
(1974) [Merton, 1974]. This approach is fundamentally different than the others based on financial ratios in that it assumes a specific structure for the market value of equity— as a European call option on the value of the firm’s assets. By interpreting listed equities as call options on the respective firm assets, the options would be considered out-of-the-money (no intrinsic value) if the underlying firm assets are worth less than the face value of firm debt. In other words, stock should be considered worthless if the firm owes more than the value of its assets. If the options were to expire in this state (a 1 year horizon is commonly assumed as a model parameter) the model assumes a bankruptcy event would result. Distance to default is then the distance (in standard deviations) that the estimated firm asset value would have to fall to reach the face value of debt (according to an asset volatility estimate based on equity volatility). This distance is translated to a probability of default by taking one minus the normal c.d.f. of the distance, such that larger distances result in smaller default probabilities. Crosbie and Bohn (2003) give a full treatment of the exact implementation used by Moody’s KMV, a ratings agency that has championed this measure [Crosbie and Bohn, 2003]. Vassalou and Maria (2004) go beyond default prediction with Distance to Default and assess its relationship with equity returns [Vassalou and Xing, 2004].

With Shumway (2001), we begin to see a renewed interest in the free form models based on financial ratios because of the demonstrated success when switching from a static model to a hazard model [Shumway, 2001]. With this new model specification where firms are observed at every year in the sample, and the estimation sample used to train a logistic regression is arranged as firm-years, Shumway achieved superior results over the older, static ratio models and the option structure models derived from Merton. Beaver, McNichols, and Rhie (2004) performed an extended review of bankruptcy prediction, also implementing a hazard model using logistic regression, but their focus was to frame a discussion of any secular changes in the bankruptcy relationship instead of achieving improvements relative to older approaches [Beaver et al., 2004]. They concluded that there was significant explanatory power with their model and that the results were in fact robust throughout the period. Two more significant contributions to failure modeling can be found in Chava and Jarrow (2004) [Chava and Jarrow, 2004]. The time period was changed from years to months, and industry effects were included, with both changes improving forecasting.

Campbell, Hilscher, and Szilagyi (2008) continued with the monthly observations, but did not find evidence in support of industry effects [Campbell et al., 2008]. Most importantly, they not only described a bankruptcy model using financial ratios, but also defined a general purpose firm failure indicator of bankruptcy, default, or exchange delisting for performance as the basis for a model of failure. As noted with Beaver’s 1966 work, this generalization of failure beyond bankruptcy creates more minority class observations, which in turn allows for learning a more sophisticated relationship. The 2011 follow-up effort by the same authors, the primary reference for the variables used in our approach, added a 12 month ahead variant of their failure model and an in-depth analysis of the relationship between their model forecasted distress risk and stock returns [Campbell et al., 2011]. We largely agree with choices and assumptions made by Campbell and his co-authors, and with our preliminary results matching theirs accept their conclusions. Our motivations are to further improve the approach for forecasting with an emphasis on accuracy among high risk firms, and without concern for interpreting historical trends.

3.1.2 Firm Failure Model and Dataset Summary

Following Campbell we model the probability of a firm failing over the next month using 8 market/accounting measures. Generically, they can be thought of as profitability, leverage, liquidity, equity return, equity return volatility, market capitalization, book value, and stock price. Using notation, for firm i at month t, the 8
variables are defined as

\[ NIMTA_{i,t} = \frac{\text{NetIncome}_{i,t}}{\text{MarketEquity}_{i,t} + \text{BookLiabilities}_{i,t}} \] (3.1)

\[ TLMTA_{i,t} = \frac{\text{BookLiabilities}_{i,t}}{\text{MarketEquity}_{i,t} + \text{BookLiabilities}_{i,t}} \] (3.2)

\[ CASHMTA_{i,t} = \frac{\text{CashAndShortTermInvestments}_{i,t}}{\text{MarketEquity}_{i,t} + \text{BookLiabilities}_{i,t}} \] (3.3)

\[ EXRET_{i,t} = \log(1 + R_{i,t}) - \log(1 + R_{S&P500,t}) \] (3.4)

\[ SIGMA_{i,t} = \left( \frac{252}{N - 1} \sum_{k \in \{t-1, t-2\}} r_{i,k}^2 \right)^{\frac{1}{2}} \] (3.5)

\[ RSIZE_{i,t} = \log\left( \frac{\text{MarketEquity}_{i,t}}{\text{MarketValue}_{S&P500,t}} \right) \] (3.6)

\[ MB_{i,t} = 0.9 \times \text{BookEquity}_{i,t} + 0.1 \times \text{MarketEquity}_{i,t} \] (3.7)

\[ PRICE_{i,t} = \log(\text{ClosingPrice}_{i,t}) \] (3.8)

where \( R_{i,t} \) is the monthly stock return for firm \( i \) in month \( t \), \( R_{S&P500,t} \) is the monthly return on the S&P500 index, \( r_{i,k} \) is the daily stock return for firm \( i \) on day \( k \), and \( \text{MarketValue}_{S&P500,t} \) is the sum of the market values for all firms in the S&P500 index in month \( t \).

We can estimate the probability of a firm failing over the next month by applying logistic regression as follows:

\[ P_{FAILURE_{i,t+1}} = P(Y_{i,t+1} = 1|X_{i,t}) = 1/(1 + e^{-(-\alpha - \beta X_{i,t})}) \] (3.9)

where \( Y_{i,t+1} = 1 \) is the binary indicator of failure for firm \( i \) in month \( t + 1 \), \( X_{i,t} \) is the vector of 8 predictor variables defined previously for firm \( i \) available as of the close of month \( t \), \( \beta \) is the vector of logistic regression coefficients, and \( \alpha \) is the model intercept. For convenience of writing and presentation, going forward we may refer to the quantity \( P(Y_{i,t+1} = 1|X_{i,t}) \) as \( P_{FAILURE} \).

The event of firm failure, \( [Y_{i,t+1} = 1] \), is defined by any of the following: exchange delisting due to performance, Chapter 7 or 11 bankruptcy filing, or rating D or SD (“selective default”) from any major ratings agency. Failure is taken to be observed at the first occurrence of any of these underlying events, in the case of multiple. When we say predictor variables are available as of a specific month, this accounts for an appropriate lag for variables derived from quarterly financials (minimum of 2 months). In Section 3.2.1 we will also discuss a second model of the same form with a 13 month horizon, where we estimate a separate coefficients vector \( \beta \) to use with forecasting \( P(Y_{i,t+12} = 1|Y_{i,t+11} = 0, X_{i,t}) \).

Before proceeding, we need to clarify a minor change of notation. Campbell uses an awkward convention for time indexes, where the event probability being modeled is written as \( P(Y_{i,t} = 1|X_{i,t-1}) \) and \( P(Y_{i,t+j-1} = 1|X_{i,t}) \) for the 1 month and \( j \) month forecast horizons, respectively. For the 12 month model, then, they write \( P(Y_{i,t+11} = 1|X_{i,t-1}) \). To eliminate confusion we write the event probabilities being modeled as \( P(Y_{i,t+1} = 1|X_{i,t}) \) and \( P(Y_{i,t+j} = 1|X_{i,t}) \), for the 1 month and \( j \) month horizon, respectively. With our notation, the current, or most recent vector of predictor variables that is known shall be designated time \( t \).

The dataset of the 8 variables defined above are modified further before proceeding with model fitting. The first 7 variables are “winsorized”, or thresholded at their 5% (and 95%) quantiles, such that all observed values below (and above) are replaced with the respective quantile. \( PRICE \) is thresholded at log($0.01)
and \( \log(15.00) \). Missing values of \( \text{CASHMTA} \), \( \text{SIGMA} \), and \( \text{MB} \) are allowed and plugged with their cross-sectional mean. \( \text{NIMTA} \) and \( \text{EXRET} \) are converted into weighted averages, \( \text{NIMTAAVG} \) and \( \text{EXRETAVG} \), using lagged values from the prior year, with the weight of each successive lagged value cut in half.

The bankruptcy filing data was obtained from Dr. Sudheer Chava from Georgia Tech, and all other data was obtained from Wharton Research Data Services (WRDS) [Alanis et al., 2014, Chava et al., 2011, Chava and Purnanandam, 2010, Cha, 2014]. That service is a portal for many data providers, of which the Center for Research in Security Prices (CRSP) and Compustat (managed by S&P) are perhaps the most important. CRSP provides closing prices, shares outstanding, and daily stock returns, in addition to delisting events [CRS, 2014]. Compustat provides quarterly firm fundamentals such as net income, book value of liabilities, etc., in addition to S&P credit ratings from 1986 onward [Com, 2014, Com, 2012]. There is a third dataset that we make use of from WRDS, the Fixed Income Securities Database (FISD) provided by Mergent, which has ratings coverage from all the major agencies for publicly offered corporate bonds [Mer, 2014].

All together, we have an almost identical dataset as compared to Campbell. The only exception is ratings coverage for 1963 - 1986 on forms of corporate debt that are not publicly offered bonds. Our dataset additionally includes the years 2009 - 2014. In subsection 3.2.1, we deliberately exclude these additional years where necessary to make a fair comparison with Campbell’s results, but in all subsequent sections the years are included.

Our final reconstruction of the Campbell dataset will be referred to as the \( \text{MTA} \) dataset, after the Market value of Total Assets quantity that normalizes the raw input variable used in constructing the \( \text{NIMTA} \), \( \text{TLMTA} \), and \( \text{CASHMTA} \) predictors. Our \( \text{MTA} \) dataset covers 1963 - 2014 and includes roughly 2.4 million firm-month observations and 2,024 failure events. The overall sample average failure rate is then 0.08\%, or a more relevant statistic might be the average of monthly failure rates, which is slightly less at 0.06\%. Our sample is drawn from the population of firms with U.S. exchange listed equity. There is no minimum market capitalization threshold. Figure 3.1 displays the number of firm-months available at the start of each year.
3.2 Preliminary Firm Failure Model Evaluation

3.2.1 Results Using Reconstructed Dataset

This section uses only firm-months from 1963 - 2008 as training data for model estimation to match Campbell. Throughout this document, unless otherwise noted, data and results are generated from the probability of failure model using a 1 month horizon. The figures displaying results from Campbell (3.2, 3.3, 3.5), are screenshots produced from the original PDF file.

Comparison of Sample Summary Statistics

Table 3.1 and Figure 3.2 Panel A show summary statistics for the 1963 - 2008 sample, failure months and non-failures included. Table 3.2 and Figure 3.2 Panel B show summary statistics for failure months only. The cell contents without a percent sign are denominated as ratio multiples (columns RSIZE and MB) or US dollars (PRICE). The row for “Std. Dev. Difference” displays the distance between the whole sample means and failure group sample means in units of whole sample standard deviations.

Even though we share the same count of failure month observations, by the fact that differences do exist in summary statistics for this group, it should be evident our samples are not the same. Overall, the samples seem extremely similar with only minor differences.
Table 3.1: Summary Statistics (MTA Dataset 1963 - 2008)

<table>
<thead>
<tr>
<th>Variable</th>
<th>NIMTA</th>
<th>TLMTA</th>
<th>CASHMTA</th>
<th>EXRET</th>
<th>RSIZE</th>
<th>SIGMA</th>
<th>MB</th>
<th>PRICE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>-0.02%</td>
<td>43.62%</td>
<td>8.68%</td>
<td>-1.00%</td>
<td>1.46</td>
<td>54.98%</td>
<td>2.03</td>
<td>10.10</td>
</tr>
<tr>
<td>Median</td>
<td>0.53%</td>
<td>41.04%</td>
<td>4.63%</td>
<td>-0.79%</td>
<td>0.27</td>
<td>45.49%</td>
<td>1.53</td>
<td>12.61</td>
</tr>
<tr>
<td>St. Dev.</td>
<td>2.25%</td>
<td>28.17%</td>
<td>9.96%</td>
<td>11.80%</td>
<td>2.79</td>
<td>33.02%</td>
<td>1.54</td>
<td>5.38</td>
</tr>
<tr>
<td>Observations: 2,030,922</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.2: Summary Statistics (MTA Dataset 1963 - 2008, Failures Only)

<table>
<thead>
<tr>
<th>Variable</th>
<th>NIMTA</th>
<th>TLMTA</th>
<th>CASHMTA</th>
<th>EXRET</th>
<th>RSIZE</th>
<th>SIGMA</th>
<th>MB</th>
<th>PRICE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>-4.17%</td>
<td>74.22%</td>
<td>7.82%</td>
<td>-11.41%</td>
<td>0.14</td>
<td>121.68%</td>
<td>2.40</td>
<td>1.45</td>
</tr>
<tr>
<td>St. Dev. Difference</td>
<td>1.85</td>
<td>1.09</td>
<td>0.09</td>
<td>0.88</td>
<td>0.47</td>
<td>2.02</td>
<td>0.24</td>
<td>1.61</td>
</tr>
<tr>
<td>Observations: 1,756</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 3.2: Campbell Table 1 [Campbell et al., 2011]

Comparison of Estimated Logistic Regression Coefficients

Estimated coefficients from a standard logistic regression model are presented in Table 3.3 and Figure 3.3 for 1 and 12 month forecast horizons. As mentioned earlier, Campbell used a “\(t + 0\)” notation for a 1 month forecast horizon and a “\(t + j − 1\)” notation for extended, \(j\) month horizons. To match their table, we labeled our columns as 0 and 12 months, respectively.

The values in parentheses with asterisks shown immediately below the estimated coefficient values are estimation Z-statistics meant to quantify the significance of each predictor variable (the asterisks provide a qualitative translation of the number in parentheses—more asterisks corresponds with greater significance). Between our results and Campbell’s results, these statistics look extremely different. This is because they modified their calculation to better account for the lack of independence in firm month observations. We generally agree with their assessments of variable significance and these calculations will be ignored.

The text of Campbell clearly indicated that “Pseudo-\(R^2\)” refers to McFadden’s \(R^2\), or one minus the ratio of the log likelihood of the model over the log likelihood of an intercept-only model. This statistic is somewhat analogous to traditional \(R^2\) in a linear regression setting, with values between 0 and 1 that are meant to capture the explanatory power of the model (with respect to a null model of only using an intercept). As for “Accuracy Ratio”, we were unclear as to what Campbell was referencing, and instead we present Area Under the Curve (AUC), the well-known measure of model fit (we used a superscript “!” to acknowledge this difference between the tables). AUC is an exhaustive summary of the explanatory power of the model that accounts for false positive and false negative classification errors (range of 0 to 1, higher is better). Given the similarities in values, however, we would not be surprised if they are just referring to AUC with a more casual sounding name, as they were with McFadden’s \(R^2\).
Our firm-month observation count and failure count are extremely close for the 1 month horizon. We have slightly more failures when modeling the 12 month horizon, and slightly more firm-months. This discrepancy may need further review if we are to focus more on the longer horizons. The Pseudo-$R^2$ and AUC appear very similar for both the 1 month and 12 month horizons.

Overall, the 1 month coefficients we have appear to be similar, but with notable differences in $NIMTAAVG$ and $PRICE$. The same can be said of our 12 month coefficients, but with notable differences in $RSIZE$ and $PRICE$. If $PRICE$ was considered statistically significant in the 12 month model, it may require further review given the implication of a change in sign. It is difficult to believe that the probability of 12 month failure should increase with share price. Similarly, the 1 month coefficient for $RSIZE$ is highlighted by Campbell as having the wrong sign (positive), and our result matches this incorrect outcome. We have strong evidence in Figure A.6 from Appendix A.3 that probability of failure has a decreasing relationship with $RSIZE$. We will address this issue in greater detail when we revise all the predictors in Chapter 4 Section ??.

**Comparison of Actual and Predicted Failures Over Time**

Both Figures 3.4 and 3.5 show annualized results for the same period (firm-months from 1972 through 2008), but with our results (Figure 3.4) we also include results based on out-of-sample forecasts for the years 2009 - 2014. We mark this difference in training data and testing data with the red line. The sample data technically includes 1963 - 1971 in addition to the years shown, but we omitted that early period to match Campbell’s visual. After further review, we found no model predictable failures during that time, and it seems reasonable to think that is why Campbell left them out. This simply means for the firm-months in 1963 - 1972 where we had predictor observations there were no failures in the dataset, not that there were no failures in our failure data sources.

The black line “Actual Failures” represents the observed failure rate as a percentage for the indicated year, as calculated by the number of firm failures observed that year, divided by the number of firm-years (number of firm-months divided by 12) times 100. The blue line “Predicted Failures” is the mean $PFAILURE$ forecast over all firms for that year (using firm-years in the denominator as was used when calculating the annual “Actual Failures” rate). By averaging all forecasts for individual firms, we have an implicit global failure rate forecast for the year. The blue lines for model predicted failure rate appear almost identical, with matching peaks in 1991, 2001, and 2008. The black lines for observed failure rate share the peak in 2001, however, there are differences from the early 1980s - mid 1990s.
<table>
<thead>
<tr>
<th>Lag (months)</th>
<th>0</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>NIMTAAVG</strong></td>
<td>-21.15</td>
<td>-18.84</td>
</tr>
<tr>
<td></td>
<td>(-17.04)**</td>
<td>(-16.48)**</td>
</tr>
<tr>
<td><strong>TLMTA</strong></td>
<td>3.50</td>
<td>1.52</td>
</tr>
<tr>
<td></td>
<td>(29.92)**</td>
<td>(18.91)**</td>
</tr>
<tr>
<td><strong>CASHMTA</strong></td>
<td>-1.77</td>
<td>-2.25</td>
</tr>
<tr>
<td></td>
<td>(-6.99)**</td>
<td>(-9.60)**</td>
</tr>
<tr>
<td><strong>EXRETAvg</strong></td>
<td>-8.00</td>
<td>-8.23</td>
</tr>
<tr>
<td></td>
<td>(-16.43)**</td>
<td>(-17.74)**</td>
</tr>
<tr>
<td><strong>SIGMA</strong></td>
<td>2.35</td>
<td>1.57</td>
</tr>
<tr>
<td></td>
<td>(22.70)**</td>
<td>(20.37)**</td>
</tr>
<tr>
<td><strong>RSIZE</strong></td>
<td>0.08</td>
<td>-0.06</td>
</tr>
<tr>
<td></td>
<td>(3.04)**</td>
<td>(-3.42)**</td>
</tr>
<tr>
<td><strong>MB</strong></td>
<td>0.03</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>(2.89)**</td>
<td>(4.94)**</td>
</tr>
<tr>
<td><strong>PRICE</strong></td>
<td>-0.48</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>(-18.06)**</td>
<td>(1.11)</td>
</tr>
<tr>
<td>Constant</td>
<td>-10.43</td>
<td>-9.67</td>
</tr>
<tr>
<td></td>
<td>(-30.56)**</td>
<td>(-41.12)**</td>
</tr>
</tbody>
</table>

**Observations** 2,030,922 1,943,826  
**Failures** 1,756 2,254  
**Pseudo-$R^2$** 0.328 0.115  
**Area Under Curve** 0.960 0.858  

Table 3.3: Estimated Logistic Regression Coefficients (MTA Dataset 1963 - 2008)

Figure 3.3: Campbell Table 2 [Campbell et al., 2011]
Figure 3.4: Actual and Predicted Failure Rates (%) By Year (MTA Dataset)

Figure 3.5: Campbell Figure 2 [Campbell et al., 2011]
3.2.2 Monthly Prediction Test Set Results 2009 - 2014

This section uses only the live, out-of-sample prediction results for each month from July 2007 - December 2014. July 2007 was chosen as the starting point because of the observed peak in the S&P 500 index in June 2007, making our test set cover almost an entire peak to peak cycle of the index. This should help minimize any bias that could be introduced during (strictly) rising or falling markets because of the significant usage of the S&P 500 in constructing our predictor variables. “Testing Data Only, Rolling Basis” means that for each month, all data up until (but excluding) the month shown is used to estimate the model, and then we generate $P_{\text{FAILURE}}$ forecasts for all firms for the given month.

Figure 3.6 can be thought of as zooming in on July 2007 - December 2014 in Figure 3.4, except that the y axis displays monthly failure forecasts, and all $P_{\text{FAILURE}}$ forecasts were generated for the rolling set of testing data only. To clarify, along the x axis “Month” indicates the month of observing failures, and so then refers to market/fundamentals data as of the previous month end. For example, the far left point on the blue line corresponds to July 2007. At this point, we are predicting the probability of failure occurring in that month, and the $P_{\text{FAILURE}}$ forecasts were generated using market/fundamentals data that were available as of 6/30/2007. However, in estimating the model used for generating forecasts at this month, only market/fundamentals data through 5/31/2007 and failure events through June 2007 were included to ensure that model forecasts did not have knowledge of observed failures in July 2007 (which would not have been available in a live setting).

The black line “Actual Failures” represents the observed failure rate as a percentage for the indicated month, as calculated by the number of firm failures that divided by the number of firms times 100. The blue line “Predicted Failures” is the mean $P_{\text{FAILURE}}$ forecast over all firms for that month, multiplied by 100. By averaging all forecasts for individual firms, we have an implicit global failure rate forecast for the month. The predicted failure rate does a reasonable job of tracking the actual failure rate except for the depth of the recession in the first half of 2009.

In Table 3.4 we present a brief summary of the monthly failure predictions and related Tickers from first 24 months of our test set, July 2007 - June 2009. The same table for the remaining months in our test set, July 2009 - December 2014, can be found in Appendix A.4. “Month” indicates the month of observing failures, and all details regarding model estimation mentioned above for the x axis in Figure 3.6 hold here as well. “Fail.Rate” represents the observed failure rate as a percentage for the indicated month, as calculated by “Failures” divided by “Firms” times 100. This column matches exactly the black line “Actual Failures” from Figure 3.6. “Pred.Rate” is the mean $P_{\text{FAILURE}}$ forecast over all firms for that month. By averaging all forecasts for individual firms, we have an implicit global failure rate forecast for the month. This column matches exactly the blue line “Predicted Failures” from Figure 3.6.

“Pred.DC10” is the $P_{\text{FAILURE}}$ value that represents the cutoff for the top decile of the forecast distribution for that month. “Not.DC10.Tickers” is the comma separated list of firm tickers that failed in the given month, but whose $P_{\text{FAILURE}}$ forecast was not in the top decile. These can be thought of as false negatives when using the top decile value for a group of predictions at the same month as a relative threshold for making a discrete failure predictions. This relative threshold is used to match some of the results from Campbell where they measured the performance of stocks in the top and bottom decile of monthly $P_{\text{FAILURE}}$ distributions. “Top.DC10.Tickers” is the comma separated list of firm tickers that failed in the given month, and whose $P_{\text{FAILURE}}$ forecast was in the top decile. These can be thought of as true positives when using the same top decile of forecasts as a threshold for class predictions. The number shown in “Top.DC10.Tickers” is the count of firms in the top decile, and is included for convenience. For
the moment we are not concerned with the absolute performance of the model, but instead these results will be used later on as a baseline for comparing the results of our improved failure model.
Figure 3.6: Actual and Predicted Failure Rates (%) By Month (MTA Dataset, Testing Data Only, Rolling Basis)
<table>
<thead>
<tr>
<th>Month</th>
<th>Firms Failures</th>
<th>Fail Rate Pred.Rate Pred.DC10</th>
<th>Not.DC10.Tickers</th>
<th>Top.DC10.Tickers</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007-JUL</td>
<td>5254</td>
<td>0.0000%</td>
<td>0.0183%</td>
<td>SXT, RRI</td>
</tr>
<tr>
<td>2007-AUG</td>
<td>5248</td>
<td>0.1524%</td>
<td>0.0240%</td>
<td>PTBQT, QFABQ, SUTMQ, AHMIQ, ASTTY, HMBNQ</td>
</tr>
<tr>
<td>2007-SEP</td>
<td>5301</td>
<td>0.0566%</td>
<td>0.0353%</td>
<td>COH</td>
</tr>
<tr>
<td>2007-OCT</td>
<td>5299</td>
<td>0.0189%</td>
<td>0.0396%</td>
<td>MVGRQ, SCOXQ</td>
</tr>
<tr>
<td>2007-NOV</td>
<td>5259</td>
<td>0.0189%</td>
<td>0.0510%</td>
<td>NRVHQ</td>
</tr>
<tr>
<td>2007-DEC</td>
<td>5199</td>
<td>0.0385%</td>
<td>0.0664%</td>
<td>3WMANQ, DFCLQ</td>
</tr>
<tr>
<td>2008-JAN</td>
<td>5193</td>
<td>0.0385%</td>
<td>0.0791%</td>
<td>SXT, RRI</td>
</tr>
<tr>
<td>2008-FEB</td>
<td>5192</td>
<td>0.0578%</td>
<td>0.0763%</td>
<td>PTBQT, QFABQ, SUTMQ, AHMIQ, ASTTY, HMBNQ</td>
</tr>
<tr>
<td>2008-MAR</td>
<td>5245</td>
<td>0.1345%</td>
<td>0.0756%</td>
<td>RELY, TMRQ, DDDC, ADST, 3986B, GAXIQ, IMGE.1</td>
</tr>
<tr>
<td>2008-APR</td>
<td>5230</td>
<td>0.1530%</td>
<td>0.0923%</td>
<td>NEST, 3AEIM, VRSOQ, PRNTRQ, JRCCQ, REDEQ, OTODQ, 3NHXR</td>
</tr>
<tr>
<td>2008-MAY</td>
<td>5212</td>
<td>0.0384%</td>
<td>0.0976%</td>
<td>DLPX, MCELQ</td>
</tr>
<tr>
<td>2008-JUN</td>
<td>5199</td>
<td>0.0389%</td>
<td>0.0943%</td>
<td>SIX, DESCQ</td>
</tr>
<tr>
<td>2008-JUL</td>
<td>5136</td>
<td>0.0974%</td>
<td>0.1114%</td>
<td>IDMCQ, MODTQ, BJT, CYDS, PILQ</td>
</tr>
<tr>
<td>2008-AUG</td>
<td>5136</td>
<td>0.1359%</td>
<td>0.1206%</td>
<td>LIPD, POLXF, FCSEQ, LEHMQ, WAMUQ, PSITY, BLHI</td>
</tr>
<tr>
<td>2008-SEP</td>
<td>5150</td>
<td>0.1526%</td>
<td>0.1813%</td>
<td>SYBRQ, AGIXQ, WRSQ</td>
</tr>
<tr>
<td>2008-OCT</td>
<td>5091</td>
<td>0.1571%</td>
<td>0.3184%</td>
<td>LFGRQ, CTFTYQ, DWNFQ, PPC, BGPTQ, APBL, VSUNQ, SPMD</td>
</tr>
<tr>
<td>2008-DEC</td>
<td>5000</td>
<td>0.1000%</td>
<td>0.4747%</td>
<td>PGICQ, INGQ, CNSTQ, BFCF, ARTEQ</td>
</tr>
<tr>
<td>2009-JAN</td>
<td>4962</td>
<td>0.1209%</td>
<td>0.4960%</td>
<td>ACEL, NRTLQ, SSCC, MEG, AUMN, CODE</td>
</tr>
<tr>
<td>2009-FEB</td>
<td>4926</td>
<td>0.1015%</td>
<td>0.4874%</td>
<td>VITRY, MWYQ, TMQPQ, INVC, DHRT</td>
</tr>
<tr>
<td>2009-MAR</td>
<td>4939</td>
<td>0.1215%</td>
<td>0.4462%</td>
<td>CHMT, GGP, RVHLOQ, NASMQ, EVOQ, NBFQAQ</td>
</tr>
<tr>
<td>2009-APR</td>
<td>4911</td>
<td>0.1222%</td>
<td>0.3175%</td>
<td>RFP, ASYTQ, EMMIS, NXST, DSUPQ</td>
</tr>
<tr>
<td>2009-MAY</td>
<td>4877</td>
<td>0.1845%</td>
<td>0.2308%</td>
<td>ALSE, AXLL, CSARQ, SAH, TCLRY, ADLS, TPUTQ, PMII</td>
</tr>
<tr>
<td>2009-JUN</td>
<td>4779</td>
<td>0.1255%</td>
<td>0.1717%</td>
<td>SNSP, AVSR</td>
</tr>
</tbody>
</table>

Table 3.4: Monthly Failure Prediction Summary (MTA Dataset, Testing Data Only, Rolling Basis)

### 3.2.3 Full In-Sample Model Summary 1963 - 2014

The figures presented in this section use the entire MTA dataset, 1963 - 2014, as training data. After estimation, $P_{FAILURE}$ values were generated for all firm-months as in-sample forecasts.

Figure 3.7 features box plots of $P_{FAILURE}$ deciles. Within each decile of $P_{FAILURE}$ there is a pair of box plots, one for the failure and non-failure groups. The box plots provide a visual summary of the distribution of $P_{FAILURE}$ within each decile. To review, a box represents the 25%, 50%, and 75% quantiles of $P_{FAILURE}$, and the vertical lines extend out to $\pm 1.5$ times the height of the box (i.e. inter-quartile range). At the top of the plotting area, the raw number of failures and the failure rate within each decile are displayed. Given that we are using the full 1963 - 2014 sample, each decile represents roughly 240,000 observations, which is the denominator used to calculate the failure rates. It is clear that all but the top decile of observations result in $P_{FAILURE}$ values close to 0.

Additional univariate detail can be found in Appendices A.3 and B.1. The former compiles replicates of this same box plot figure, but where $x$ axis is switched from deciles of $P_{FAILURE}$ to deciles of a predictor variable (one figure for each predictor). The latter compiles overlapping class conditional histograms for each predictor in our MTA dataset.

Figures 3.8 and 3.9 illustrate density estimates of the in-sample $P_{FAILURE}$ forecasts. We present a separate density plot for the failure and non-failure groups. This is because of the extreme difference in scale on the y axis between the two plots. The x axes are identical. Densities are scaled as if the non-failure and failure group sample sizes were the same (i.e. both densities integrate to 1).

For reference, the non-failure group sample represents 99.92% of the full sample, of which, 98.37% produced a $P_{FAILURE} < 1%$. The failure group sample represents 0.08% of the full sample, of which, 59.14% produced a $P_{FAILURE} \geq 1%$. Using $P_{FAILURE} \geq 1%$ as a threshold, the failure rate within that sub-sample is 2.96%. That failure rate is 37 times greater than the failure rate of the full sample, providing evidence that the base model can provide some degree of separation between the failure and non-failure group.
3.3 Expanding the Firm Failure Data Problem

Part of the initial appeal to the Campbell model was how as of the 2011 publication, the included dataset ended as of December 2008, when the Lehman Crash was unfolding and the broader equity indexes were about to bottom out. The evaluation of any kind of bankruptcy or failure model is perhaps most interesting or revealing in a bear market, or crash scenario, and the 2008 - 2009 mortgage meltdown was the worst in a generation. There was an immediate, instinctual desire to know more about how the Campbell model performed during this period and the years following when most of the global economy was recovering. This evaluation period could represent a “make-or-break” moment for the model. Performing well could be interpreted as evidence in favor of the validity of the model and help further establish its legacy within the literature, and vice versa.

The forecasting results covering the tail end of the dataset in Campbell (final months of 2008) did not appear promising. We confirmed this disappointment with our reconstruction that included data through 2014. This poor forecasting performance is evident in Figures 3.4 and 3.5, where the blue line for predicted failures spikes in 2008 - 2009, but then only a modest increase in the black line for observed failures. The out-of-sample forecasts by month presented in Figure 3.6 look even worse. There appears to be significant divergence between actual and predicted failure rates for November 2008 - April 2009. After this period, however, the monthly errors between actual and predicted failure rates appear much more reasonable.
Figure 3.8: Estimated Density for PFAILURE (Non-Failure group)

Figure 3.9: Estimated Density for PFAILURE (Failure group)
Upon further review of our sources, we came across a related effort by Hilscher and Wilson (2015) expanded further the set of failure events so that bailout events initiated by various U.S. Government programs were included [Hilscher and Wilson, 2015]. We were unable to acquire their source data, but we found a similar offering from the Bailout Tracker project at Propublica [Kiel, 2016]. The dataset we acquired was extensive, and included a diverse range of capital flows and individual bailout programs. We settled on including purchases and loans made under the “Capital Purchase Program”, “Automotive Industry Financing Program”, and “Systematically Significant Failing Institutions” as failure events.

These data were aggregated with all the other original failure events in the construction of an updated MTA dataset. A listing of the bailout events we included can be found in Appendix A.5. They are mostly banks under the “Capital Purchase Program”, with a handful of notable exceptions like AIG, well-known investment banks, and well-known auto manufacturers. There are 183 additional bailout events not present in the table on regional banks with capitalizations under $1 Billion. The changes to our MTA Dataset resulting from the addition of bailout events are summarised in Tables 3.5 and 3.6. By adding the bailouts, there was a net increase of 235 failure observations and thus a slight increase in the sample mean failure rate, from 0.084% to 0.0943%. The monthly breakdown reveals that a majority of the added failure events occurred in the crash period of late 2008 into early 2009, and that there were actually 7 failures that are no longer counted because of newly added bailouts on the same firm that predate the originally recorded failure.

Figure 3.10 is an exact reproduction of 3.6, but with the notable exception of using the updated MTA dataset with bailout events for both recording the actual failure rate and estimating a logistic regression model to generate the predicted failure rate. The updated results are dramatically different for the crash period, so much so that they may appear unrecognizable because of the change in scale on the y axis. There is now an extremely large jump in the observed failure rate in December 2008 and a small increase in the predicted rate with respect to the previous figure made without bailout events. It seems clear the forecast error is greater here than it would be when calculated without bailouts, but this outcome is somewhat satisfying. We have been able to verify our prior expectation of a substantial increase in failure events during the post-Lehman Crash period and that the model anticipated the increase to some degree of success. We make this qualified claim of success even though the observed failure rate dwarfs the predicted rate in December 2008, for example, because the forecasted rate was roughly 5 times the test mean forecast.

The anticipation of the bailout events indicated by the model forecasts during the post-Lehman Crash period increased our appreciation of the model. It seems that the Campbell approach to framing firm failure as a data problem is extensible if the conditions justify doing so and robust to various market regimes. Unfortunately, any extension to new failure data may introduce further complications because of sampling biases. With the bailout events further validation should be performed with respect to the amounts received and capitalization (or liquid or tangible assets). Firms that received small capital inflows might very well have not failed without it. Also, timeliness of the bailout data collection was possibly an issue. Propublica has indicated that not all bailout activity was made public on the recorded transaction date [Hilscher and Wilson, 2015]. We included all bailout events without a lag, but for the purposes of making model forecasts in a live setting this issue would need clarification.

After the addition of the bailout data, the results from the post-Lehman Crash period now mirror to some degree the results from the end of the internet bubble era at the turn of the millennium. As shown in Figure 3.4, this period also has a spike in the observed failure rate that is anticipated to some degree by the predicted rate. The presence of this negative bias between observed and predicted failure rates during these
two big market downturns may be evidence of a systemic risk component not being captured by firm failure model. It is reasonable to suggest that observing greater than expected failure rates during downturns should lead us to consider the possibility that when many firms are experiencing failure simultaneously, the true aggregate distress risk will increase beyond what is expected for each firm in isolation. Given that the firm failure model is driven only by firm level financial ratios and assumes independence of individual firm-month observations, we know that any contributions to risk that are sourced by data dependencies will be ignored.

<table>
<thead>
<tr>
<th>Description</th>
<th>Obs. Fails. Fail.Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>MTA Dataset (Without Bailout Events)</td>
<td>2,410,293 2024 0.0840%</td>
</tr>
<tr>
<td>... Add Bailout events, remove obs. as appropriate</td>
<td>-15,484 +235</td>
</tr>
<tr>
<td>MTA Dataset (With Bailout Events)</td>
<td>2,394,809 2259 0.0943%</td>
</tr>
</tbody>
</table>

Table 3.5: MTA Dataset Pre and Post Bailout Events Summary

<table>
<thead>
<tr>
<th>Month</th>
<th>Failures</th>
<th>Bailouts</th>
<th>Non-Bailouts</th>
<th>Bankruptcies</th>
<th>Defaults</th>
<th>Delistings</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008-OCT</td>
<td>+8</td>
<td>+8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2008-NOV</td>
<td>+37</td>
<td>+37</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2008-DEC</td>
<td>+106</td>
<td>+106</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2009-JAN</td>
<td>+57</td>
<td>+57</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2009-FEB</td>
<td>+16</td>
<td>+16</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2009-MAR</td>
<td>+6</td>
<td>+6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2009-APR</td>
<td>+3</td>
<td>+3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2009-MAY</td>
<td>+3</td>
<td>+3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2009-JUN</td>
<td>+1</td>
<td>+2</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
<td>0</td>
</tr>
<tr>
<td>2009-JUL</td>
<td>+2</td>
<td>+2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2009-AUG</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
<td>0</td>
</tr>
<tr>
<td>2009-SEP</td>
<td>+1</td>
<td>+1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2009-OCT</td>
<td>+1</td>
<td>+1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2011-APR</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>2011-JUN</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>2011-NOV</td>
<td>-2</td>
<td>0</td>
<td>-2</td>
<td>0</td>
<td>0</td>
<td>-2</td>
</tr>
<tr>
<td>2012-FEB</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>TOTAL</td>
<td>+235</td>
<td>+242</td>
<td>-7</td>
<td>0</td>
<td>-2</td>
<td>-5</td>
</tr>
</tbody>
</table>

Table 3.6: MTA Dataset Pre and Post Bailout Events Changes
Figure 3.10: Actual and Predicted Failure Rates (%) By Month (Testing Data Only, Rolling Basis), with Bailout Events
Chapter 4

Approximations to Large Scale, Rare Events Datasets for Use with Monotonically Constrained Classifiers
4.1 Overview

The challenges offered by a rare events classification problem are daunting. The challenges offered by a large-scale rare events problem with an extreme imbalance in classes will humble even a well-experienced statistician. In these situations when majority class counts are in the millions and minority class counts are in the thousands, the problem at hand is no different than finding a needle in a haystack. Massive effort can be expended on an improved model and yet produce no legitimate advantage over random guessing, and even the act of making that assessment of poor results can be time consuming and difficult. If the true relationship is linear or approximately linear, logistic regression could be a viable option. Estimation is fast and forecasts satisfy the implied requirement of taking the form of a probability (class label forecasts are nearly meaningless when evaluating rare events because of the imbalance).

If the true relationship is not linear, a logical alternative with a big dataset would be to estimate a nonparametric model and let the data speak for itself. This may be a terrible approach in a rare events setting, however. Consider for a moment that the dataset might represent a disproportionately large sample of noise (majority class) and a residual sample of signal (minority class). The relative lack of signal, or minority class data, means that any true relationship will be weakly evidenced and model output will be highly variable. The relative lack of signal also implies that the act of model estimation, which by necessity requires the majority class information in addition to the minority class information to learn the underlying relationship, will be inefficient. Taken together we can also claim that all nonparametric classification models requiring substantial computation are unusable with large-scale, rare events datasets because testing and development become wildly impractical. It is unreasonable to spend days or perhaps weeks waiting for algorithms to execute only to produce inconsistent (if not completely inaccurate) results.

In our approach to non-linear, large-scale, rare events classification we will avoid all of these problems by exploiting two dataset qualities: low-dimensionality and monotonicity. With the duly noted constraints of non-linearity and big data, the first quality immediately suggests a solution based on kernel density estimators (KDEs). Avoiding the “curse of dimensionality” associated with kernel densities ensures that selecting a bandwidth is a very manageable decision because of theoretically sound “plug-in” rules available in the literature. In combination with a selection rule and a product kernel (or diagonal smoothing matrix) assumption, kernel density estimators can be implemented very efficiently through parallelization, and with the large dataset kernels will be very capable of learning the structure relating the majority and minority classes. This application of kernel densities performed admirably for us in terms of traditional evaluation metrics, but it was performing poorly in the tails of the sample space, as KDEs are known to do. This issue made us consider an appeal to monotonicity, or somehow enforcing a monotonic constraint on the estimated relationship of the classifier output (minority class probability) with respect to the predictors.

Let us return to using the needle in a haystack metaphor to explain how monotonicity can help us with a rare events classification problem. Imagine a very large 3-dimensional pile of hay in a room oriented such that moving east, north, and up are considered increases along an imaginary \(x\), \(y\), and \(z\) axis, respectively. Also, pretend that we have strong theoretical and empirical support for assuming that the true probability of finding a needle is monotonically non-decreasing with respect to all of the room dimensions. With this additional insight to the problem, our effective proposal is to sample the proportion of needles to pieces of hay at each observed needle location, and then train a regression on these sample proportions only (with the help of a link function), effectively ignoring all the exact locations of the pieces of hay after the sampling is completed. The sampling step, performed at needle locations (minority points) only, can be viewed as approximating the information contained by the full set of known hay and needle locations (full dataset).
For the subsequent regression method to make effective forecasts, we emphasize the specific requirement that it enforce the proper monotonicity constraints during estimation. In training a model on an approximation of the dataset, there will surely be a loss of information in the form of under-sampling at some regions of sample space. Under the assumption of monotonicity, we claim that this loss of information is not problematic because the model surface can be interpolated between the sampled locations that are present in the regression training set with minimal bias. More importantly, with forecasting at points beyond the empirical support (extrapolating), the monotonicity constraints will ensure that any decrease in sample density does not result in an invalid model. In other words, with a non-decreasing constraint, at input points beyond the right edge of the support, the model will generate forecasts that are greater than or equal to a forecast made at the sample maximum. Likewise, at inputs beyond the left edge of the support, the model will generate forecasts that are less than or equal to a forecast made at the sample minimum. Matching claims can easily be made for the non-increasing case by switching the use of the two inequalities.

Our core proposal is to perform the sampling of minority class probabilities as we alluded to in our metaphor with a kernel density estimator. This action is identical to generating in-sample forecasts at minority class points with a kernel smoother classifier, or naive bayes classifier without the assumption of independence and Gaussianity. The most basic variant of our approach calls for a Gaussian product kernel construction and sampling minority class probabilities at minority class points only. We will show how this sample of probabilities effectively approximates a large-scale rare events dataset, transforming the classification problem to a much more manageable problem of regression on a very small dataset. We also offer variations to this approximation where additional sample locations are included in regions outside of the minority class support. All variants of our proposed transformation can work with any monotonically constrained regression method, but we have a preferred trees ensemble that will be covered later along with a test using real data.

Our rare events dataset transformation proposal can be thought of as a probabilistic alternative to piecewise, divide and conquer approaches that avoids data fragmentation problems [Maalouf and Trafalis, 2011a]. Instead of partitioning the sample space based on some thresholds applied to the input variables, we will take a probabilistic view of the full sample space that allows for an approximation to the relationship between the classes. Partitioning requires decisions on how many partitions and what partitions to use, but our approach requires selecting only one bandwidth vector and we show in practice how even that decision can be reduced to selecting a single scalar multiple of existing “plug-in” rules. Overall, we are confident that the advantages provided by our proposal are worth the effort. In exchange for making a small set of reasonable assumptions, deciding on a single scalar, and a minor computational load, we open the door to a wide class of powerful nonparametric regression methods that would otherwise be inaccessible such as those with monotonicity constraints. The end result is a quick way to “zoom in” on the most informative regions of the sample space, where the minority class observations are located, and learn a relationship that is more accurate in out-of-sample forecasting.

The remainder of this chapter contains the following sections. Next, we review the rare events literature that inspired our approach in Section 4.2. A comprehensive review of a real world financial dataset used for testing our approach presented in Section 4.3. The specification for our dataset transformation based on kernel density estimators is covered in Section 4.4, and kernel bandwidth selection is covered in Section 4.5. Section 4.6 includes a demonstration of our full rare events classifier approximation and test results using a large financial dataset. The chapter concludes with a brief set of closing remarks in Section 4.7.
4.2 Rare Events Literature

A great starting point for rare events research is the work of King and Zeng from 2001, “Logistic Regression in Rare Events Data” [King and Zeng, 2001]. The authors describe the downward bias that is present with logistic regression when using an imbalanced dataset, offer a correction, and more generally do a fine job with illuminating statistical challenges brought on by rare events data. Another broadly useful rare events resource is “Rare events and imbalanced datasets: an overview” by Maalouf and Trafalis from 2011 [Maalouf and Trafalis, 2011a]. A detailed listing of common problems with rare events modeling are outlined, such absolute rarity of data, relative rarity of data, and impropriety of classification metrics. The firm failure dataset we use to demonstrate our method later in Section 4.6 is a perfect example of the absolute rarity issue. No matter how many decades of financial history are available, within the world of exchange listed equities there have only been so many firms that experienced failure events. Another interesting consideration of rare events modeling is brought up in a work by Chawla, et al [Chawla et al., 2002]. They claim that often the cost of misclassifying a minority example as from the majority is much higher than the cost of doing the reverse. With our firm failure dataset, it turns out this is not the case. Misclassifying a (true) failure event as a non-failure event can be costly, but misclassifying a (true) non-failure event as a failure event could be much more costly. With the latter misclassification, if the model output motivated a short position on the firm in question, portfolio losses could be severe because they are unbounded.

Chawla, et al also present a categorization of rare events methods into two broad classes that our survey of the literature agrees with: modified cost weighting of the training data or re-sampling the training data [Chawla et al., 2002]. Both categories are addressing the primary difficulty with rare events data, that there is a lack of information for learning the minority class. Re-weighting strategies accept the class samples as is but increase the significance allotted to minority examples during model estimation to compensate for this lack of information, while re-sampling increases the size of the minority class (or decreases size of the majority) to compensate prior to model estimation.

For the general case of transforming any chosen classifier into one that is cost-sensitive, or one that incorporates re-weighting, Domingos presents an aptly titled framework “MetaCost” [Domingos, 1999]. Then in specific instances there are re-weighting extensions for traditional classifiers like logistic regression, such as “rare events-weighted kernel logistic regression” [Maalouf and Trafalis, 2011b]. Keeping in mind the consideration of scale with our firm failure dataset, this kind of logistic regression extension is not appealing because of the required computational load. However, if a linear kernel restriction is acceptable, there is a specialized variant of weighted kernel logistic regression that achieves performance gains without sacrificing predictive accuracy by combining truncated Newton methods with the iteratively re-weighted least squares estimation algorithm [Maalouf and Siddiqi, 2014]. Very recently, another logistic regression variant for big data has caught our attention that provides a unique take on re-weighting whereby the likelihood is approximated with the assistance leverage scores [Wang et al., 2017]. The leverage scores function as weights that indicate which observations have the greatest impact on the resulting evaluation of the likelihood.

Rare events methods based on linear models such as logistic regression are by definition monotonically constrained, another model assumption we have made with our firm failure dataset. Depending on the dataset and the specific method chosen, this can be problematic and require removing predictors or restricting tuning parameter choices because many of the linear classifier extensions are not constrained in a specific direction (non-increasing or non-decreasing). For example, when fitting vanilla logistic regression on our firm failure dataset, we see the wrong sign (minus) on two coefficients for features that are engineered to have an increasing relationship with minority class probability. Without a formal mechanism to constrain
for direction, linear model validation will present difficult choices in these situations.

Methods for learning imbalanced classes that rely on re-sampling are numerous, can combine both over-sampling of the minority class with under-sampling of the majority class, and often times are ensembled through bagging or boosting. “Synthetic Minority Over-sampling Technique”, or SMOTE, is a well-known rare events learning method and unique for over-sampling the minority class through the creation of synthetic training samples located between existing minority class samples [Chawla et al., 2002]. SMOTE is demonstrated to provide superior performance using Area Under the ROC curve (AUROC) compared to repeated sampling of existing training data. More recently, we have found evidence in favor of re-sampling ensembles that make use of Random Forests [Dai and Hua, 2016]. Random Forests and other decision tree ensembles can incorporate re-weighting in addition to re-sampling, so they uniquely belong in both of our broader rare events categories [Chen et al., 2004].

There are two notable works that share much in common with our kernel density driven approximation approach. The first, “ROSE: A Package for Binary Imbalanced Learning”, reviews a software implementation that is somewhat similar to the SMOTE algorithm, but the re-sampling in ROSE is rooted in kernel density estimation [Lunardon et al., 2014a]. As we will describe later for our own approach, ROSE also utilizes a Gaussian kernel restricted to a diagonal covariance matrix where the diagonal elements are some multiple of a preferred plug-in rule [Menardi and Torelli, 2014, Lunardon et al., 2014b]. The second work that bears a close resemblance to ours is “Isotonic Classification Trees”, where “isotonic” is synonymous with monotonically constrained [Van De Kamp et al., 2009]. The method is structured as post-processing for standard classification trees. A monotonic relationship between the model output (minority class probability) and predictors is enforced by appropriately adjusting and pruning leaf nodes of a previously estimated classification tree. Later in Section 4.6, we will demonstrate how to use a monotonically constrained regression tree as the post-processing to follow our kernel based approximation to complete a rare events classifier. Both the isotonic classification trees and our primary demonstration are oddly similar in that they call for a second stage, or post-processing, that effectively produces a monotonic classification tree.

Rare events research is not just limited to models and estimation algorithms. There is also an on-going discussion of how to evaluate classifier performance with rare events datasets. This open question is primarily motivated by the fact that traditional metrics like mean squared error (MSE) are biased towards the majority class. A null model that predicts a constant minority class probability of 0 at all inputs will appear to be a strong choice when evaluated by MSE, while some of the best and most sophisticated methods ever developed will only appear slightly better, if at all. Alternatives such as AUROC, variations of Precision and Recall, Brier Curves, Lift, Log Loss (or Cross Entropy or Binomial Deviance), and Probability Calibration are all much more informative than MSE [Dai and Hua, 2016] [Caruana and Niculescu-Mizil, 2004] [Hernandez-Orallo et al., 2011] [Hernández-Orallo et al., 2012] [Maalouf and Trafalis, 2011a]. Our preferred metrics are AUROC because of how it captures the importance of the false positive rate with our firm failure forecasting problem, and Probability Calibration because of how it captures the long run accuracy of classifier forecast probabilities (especially among forecasts greater than 1%) [Fawcett, 2006] [Caruana and Niculescu-Mizil, 2004] [DeGroot and Fienberg, 1983]. When we evaluate our approach in Section 4.6.2 these two metrics will be used in addition to Area Under the Precision Recall curve (AUPR) [Davis and Goadrich, 2006]. This last metric is included to help comparability with other rare events methods that cite Precision and Recall metrics, but we do not put too much emphasis on it because of the bias [Powers, 2007].
4.3 Dataset Modifications

The nonparametric approach to modeling firm failure that we propose later in this chapter begins with the modifications to the dataset and features reengineering we will describe in this section. Without these revisions, our kernel density based transformation would form an incoherent probability model because of the truncated support of the variables, monotonic regression would not be possible because of bimodalities, and test results could be biased because of plugged missing values. The preparatory steps outlined here may not be glamorous, but they do satisfy some necessary conditions prior to model estimation.

Differences aside, we construct our dataset in a fashion that mirrors many of the works we cited in our review of the literature in Chapter 3. We include the years 1986 through 2014, for a total of 1,935,224 observations, of which 2,083 are failures. Those figures leave us with a class 1 sample mean proportion of 0.11%. Observations are defined to represent one unique firm at a specific month in time. The raw data is sourced almost entirely from COMPUSTAT and CRSP [CRS, 2014] [Com, 2014] [Com, 2012] [Mer, 2014] [Kiel, 2016] [Cha, 2014].

We began our research efforts with the failure model predictors and target indicator as defined originally by Campbell, Hilscher, and Szilagyi [Campbell et al., 2011]. We will reference this dataset going forward by the initialism $MTA$, a reference to market value of total assets, the quantity that is used to normalize the first three predictors. Each firm-month observation is comprised of eight measures of market observables and accounting fundamentals, which can be generically referred to as profitability, leverage, liquidity, equity return, equity return volatility, market capitalization, book value, and stock price. The model target is a binary indicator of observing a failure event for firm $i$ in month $t$. More complete detail on predictor and failure indicator construction can be found in Chapter 3 Section 3.1.2.

Failure and non-failure class conditional histograms for each predictor in the $MTA$ dataset are included in Appendix B.1. Immediately below the histograms for each predictor is the same figure, but for our re-engineered variables using a slightly modified failure dataset that will be referenced by the initialism $MME$, which stands for “monotonic market equity”. The $MME$ dataset is the result of our efforts to prepare for monotonically constrained modeling and the first three predictors listed will now use the market value of equity only for normalization. Besides the specific changes in the construction of predictors, the $MME$ dataset differs from the $MTA$ dataset by removal of missing values instead of plugging them. We make this change because we are only concerned with forecasting future failure events, not explaining historical failures. Missing raw data for our predictors is almost entirely a problem restricted to historic data from decades past, so it makes some sense to plug them if explaining the history of failure is a priority. In more recent years, and going forward with in new months, missing data is extremely rare. Because this issue is longer relevant, we see no reason to risk introducing bias into our model or test results with the use of plugged values.

The other difference between the $MTA$ and $MME$ datasets is the removal of the winsorization step. This action will no longer be performed on any of the features in the $MME$ dataset that we use to estimate our final model. Thresholding the tails of predictor variables at the commonly chosen 5th and 95th quantile is a quick and easy way to eliminate the influence of outliers, but in doing so a point mass is created at the chosen quantile. Especially in the case of slowly decaying tail behavior, these point masses may contain significant density relative to the adjacent regions just outside the chosen quantile. We refer the reader to Figures B.9 and B.15 for predictors $SIGMA$ and $PRICE$, respectively (in Appendix B.1) for benign examples of this result. For predictors $CASHMTA$ and $MB$, however, winsorization produces undesirable consequences during model estimation. We will discuss these issues in greater detail later when proposing
As an alternative to winsorization, we will control for outliers through use of the log function. Incorporating the log function with every predictor can be thought of as an “absolute” method to control outliers that is not dependent on the variable’s empirical distribution, as opposed to the “relative” method of winsorization that is based on the quantiles of a distribution. All predictors except for EXRET, RSIZE, and PRICE, which already make use of the log function, will be modified as appropriate on a case by case basis. Repeated use of the log function on every predictor provided a secondary benefit, by allowing for much more manageable selection of the kernel density bandwidth. It is well established that financial data has fat tails, and our dataset is no exception. Without the log transformation, the fat tails may have been insurmountable with the kernel density estimator because of the need for impractically large bandwidths.

The benefits of eschewing winsorization are two fold. One, beyond merely controlling for outliers, we are also transforming the tails of the distribution so that they decay naturally without creating a left or right edge to the support. Two, without depending on empirical distributions, data storage requirements are greatly reduced. In our dataset the winsorization is applied to lagged values of NIMTA and EXRET before calculating their weighted average terms, NIMTAAVG and EXRETAVG, respectively. Assuming one wishes to avoid look-ahead bias, this sequence effectively requires an extra copy of the dataset for every month of backtesting because of the changing winsorization thresholds. This is not a trivial requirement at roughly 1GB of disk space per instance and 28 years of data. When incorporating the log function into the construction of every variable instead of winsorizing, only one copy of the dataset is necessary, which can be thresholded by month at every new iteration of testing.

4.3.1 Evaluation and Re-engineering of Failure Dataset Features

We have covered the global changes of removing observations with missing values and the winsorization step on all predictors, and now we will provide a detailed evaluation of each one individually and write out our revised construction. Our primary motivation is to establish each variable with support on the set of real numbers and oriented so that there appears to be an increasing (or non-decreasing) relationship with respect to the observed failure odds (implied by our model output, the failure indicator). Full support on the real line is for the application of a kernel density estimator, and a non-decreasing orientation is for the subsequent implementation of the monotonically constrained regression (not necessary, but having all variables share the same directional relationship simplifies software implementation and interpretation of results).

A secondary motivation for re-engineering the failure dataset features is to maximize the impact of stock market information as reflected by the MarketEquity term. More specifically, we will change the denominator of the first three predictors in the MTA dataset (those that bear the namesake for our reference). NIMTA, TLMTA, and CASHMTA will be changed so that they are normalized by MarketEquity only, and not the total assets quantity MarketEquity + BookLiabilities. Campbell, et al specifically noted that normalizing variables with MarketEquity + BookLiabilities instead of BookEquity + BookLiabilities improved model performance because of the more timely information incorporated through stock market prices [Campbell et al., 2008]. This is easy to understand given that our observations are firm-months, and the BookEquity and BookLiabilities terms are only updated quarterly upon the firm’s filing of a new SEC form 10-Q. Substituting MarketEquity for BookEquity incorporates a renewed valuation of firm equity at every month instead of every third month. We are taking this intuition to its logical conclusion, choosing to rely only on the always fresh MarketEquity component and ignoring BookLiabilities, which will become stale. Eliminating the liabilities term from the normalization quantity of the first three predictors does not
remove that information from our MME dataset completely, however. The influence of liabilities will still be measured by the leverage predictor, which is based on the same raw input, BookLiabilities.

One last motivation for features re-engineering arose indirectly because of our voluntary restraint to not add any additional features. In this light, our only route towards an improved measurement of the data problem was to improve the existing features. We chose to deny ourselves the addition of new features so that any improvement in our final results could be attributed to our framework alone, and not the addition of new information (which may also have lacked the proper theoretical justification). There was another consideration that encouraged us to reject the addition of more features. Kernel density estimators suffer from the “curse of dimensionality” [Scott, 1992]. With their usage, parsimonious modeling is something to ascribe to for more than the sake of generally avoiding increased model complexity. As more features are added, the sample will have more and more empty regions, which in turn makes bandwidth selection increasingly difficult because of the lack of structural continuity throughout the global sample space.

**Profitability**

The quarterly measurable NIMTA that is used to construct NIMTAAVG is significant and appropriately results in a negative coefficient when included in a linear model of failure, but we think that it can be improved by removing the BookLiabilities term from the denominator. With this change, the resulting quarterly ratio becomes the traditionally known earnings yield. We are not arguing for the modified ratio because of its established usage, but because the inclusion of firm liabilities in the denominator masks the more meaningful changes to the market value of equity. Furthermore, the liabilities term is only updated quarterly, while stock prices are updated throughout all trading days. Without the liabilities term, the influence of market equity on the earnings yield ratio will be not only be more significant but more timely as well.

A second motivation for our revisions to the profitability predictor is to control for bimodality. Exhibiting multiple, separate modes of behavior with respect to the model output constitutes a direct violation of a monotonicity assumption that needs to be corrected. Adjustments are necessary regardless of the complexity of the chosen model to be implemented if the given model assumes monotonicity (i.e. bimodality needs to be corrected for when using linear models as well as the constrained non-parametric classes we will consider in our work). For profitability, our analysis suggests that observed odds of failure increases away from the sample mode (or mean), while at the center of the distribution, there appears to be a region of typical values where failure is minimized.

We support our observations with Figure 4.1, a class conditional histogram of annualized earnings yield ($4 \times \text{NetIncome}/\text{MarketEquity}$). The $x$-axis in the figure has been labeled as a percent on a log scale to reinforce the variation observed in our dataset. It is obvious that failure odds are increasing as earnings yield trends towards greater negative values. It is not so obvious, however, that there is a similar trend towards greater positive values, but this behavior can be quickly established with sample statistics. Nearly 70% of observations in our sample have positive earnings, while this region contains not quite 20% of the failures. That results in a 0.03% failure rate across all firm-months with positive earnings. If we look at only observations with an annualized earnings yield greater than 20%, we have restricted ourselves to only 6% of the observations in the full sample, yet maintained 6% of the failures. As defined by a 20% cutoff, the right tail of annualized earnings yield exhibits a mean failure rate of 0.1%, more than triple what is observed among all observations with positive earnings.

To control for this change from a decreasing to an increasing relationship with respect to observed failure
odds towards the right tail, we will apply the absolute value function to $NetIncome$ before dividing by $MarketEquity$. Our intuition is that the true failure odds within the region of earnings yields closest to 0 is at its lowest and then increases towards both tails. We shall properly name this modified $NetIncome$ as $ANI$ and visualize it with the class conditional histogram found in Figure 4.2. By “folding” net income with the use of absolute value, we have taken the union of the two regions where the observed failure odds is greatest. The result is a measure of profitability (after normalizing with firm market value) that over its entire support exhibits an increasing relationship with respect to failure odds. As an alternative to taking the absolute value, thresholding all $NetIncome > 0$ at 0 was also considered. This could form the basis for a monotonic predictor of failure, but the information that allows for learning a relationship where failure odds are increasing towards the right tail would be lost.

To complete our revision, we take the log of the resulting ratio to restore the support to the real line and increase separation between the classes. To ensure positivity of the ratio beforehand, we plug $NetIncome$ as $\$1,000$ when observing 0 (the smallest positive increment in our data source). Our final, renamed quarterly measure of profitability is $ANIME$, and before model estimation a weighted average ($ANIMEAVG$) is constructed using the same lags and weights as with the original quarterly measure $NIMTA$ and its matching weighted average $NIMTAAVG$. Figure B.2 in Appendix B.1 presents a class conditional histogram for $ANIMEAVG$. We formally write the definition of the quarterly $ANIME$ here.

$$\text{ANI}_{i,t} = |NetIncome_{i,t}| + 1000 \cdot I_{\{0\}}(NetIncome_{i,t}) \quad (4.1)$$

$$ANIME_{i,t} = \log\left(\frac{\text{ANI}_{i,t}}{MarketEquity_{i,t}}\right) \quad (4.2)$$

The advantage of using the absolute value of net income can be better understood when you consider a firm that is trending towards greater risk (increased financial insolvency). Negative earnings would be expected, however, the firm may produce an anomalous quarter of strong results because of financial statement manipulation, asset liquidation, abrupt decrease in payroll by downsizing, or some other unrepeatable activity. When net income is transformed by an absolute value function, extreme positive earnings are measured the same as extreme negative earnings. The anomalous quarter will no longer result in misleading forecasts from our model, which will still conclude that the firm exhibits greater failure risk. The effectiveness of this change is improved further when used in conjunction with the quarterly weighted average (after making revisions to the quarterly profitability predictor, we do not change the construction of a quarterly average before model estimation). Having 3 quarters of negative earnings and 1 quarter of (large) positive earnings, when averaged (without using absolute value), would appear to be less risky. The single quarter with positive earnings will increase the quarterly earnings average, which will result in a lower failure risk forecast. When the absolute value function is applied to quarterly earnings, the resulting average in this example will not decrease, thus appearing more risky, because of the single quarter with (large) positive earnings.

**Leverage**

From the perspective of linear modeling, $TLMTA$ is already significant and the estimated coefficient results in the expected sign: increasing leverage is associated with greater failure risk. The intention was not to change the composition of this predictor. We only wished to transform the support from $(0, 1)$ to $\mathbb{R}$, in a fashion similar to what was being done with the other predictors. The obvious choice to make was to apply the logit function (or inverse logistic transform), $\log(TLMTA_{i,t}/(1 - TLMTA_{i,t}))$, but then the laws of the log function allowed for an unintended simplification of terms that results in removing $BookLiabilities$ from
the denominator. Thus, predictor $TLMTA$ has been revised and renamed appropriately as $TLME$.

$$TLME_{i,t} = \log\left(\frac{\text{BookLiabilities}_{i,t}}{\text{MarketEquity}_{i,t}}\right)$$ (4.3)

To ensure the simplified liabilities to equity ratio is greater than 0 before taking the log, observations of the numerator BookLiabilities that are 0 or negative are thresholded to $\$1,000$, the smallest positive increment measured by our data source [Com, 2014].
Figure 4.1: Class Conditional Histogram for Earnings Yield (log scale)

Figure 4.2: Class Conditional Histogram for Quantity ANI / MarketEquity (log scale)
Liquidity

The liquidity predictor $CASHMTA$ appears significant and with a negative coefficient as expected when included in a linear model of failure with winsorization. Unfortunately, thresholding the tails also introduces a slight bimodality. This dynamic is evident in Figure B.5 in Appendix B.1. The empirical failure odds appear greatest at the left edge and right edge of the sample distribution, and what we see at the right edge is a direct result of winsorizing. Our concern is that this outsized mass is resulting in unintended consequences during model estimation, such as upward bias in the model surface towards the right tail. This bias would reduce the influence of the larger trend present of decreasing liquidity coinciding with increasing failure. For a linear model especially, this bias would be problematic and justify further investigation of the predictor.

A much more significant criticism of this predictor is that the explanatory power is muted by the offsetting relationships of the raw input $CashAndShortTermInvestments$ when normalized by firm equity compared to firm liabilities. As originally defined, $CASHMTA$ seems to be a poor predictor of failure as evidenced by the lack of separation between classes in Figure B.5 in Appendix B.1. The class conditional histogram demonstrates that the failure class support and non-failure class support for liquidity is almost identical. To investigate further, we separately considered the ratios $CashAndShortTermInvestments/BookLiabilities$ and $CashAndShortTermInvestments/MarketEquity$, with respect to the failure indicator. It was apparent that failure odds have a decreasing relationship with the former, but an increasing relationship with the latter. We confirmed these observations by performing univariate logistic regression of our failure indicator on (the log of) each ratio separately. For the former the estimated regression coefficient was negative, and with the latter it was positive, and in both cases the coefficient p-values were at or near 0 (highest significance). This offsetting behavior of the two ratios is illustrated by the class conditional histograms presented in Figures 4.3 and 4.4, one for each of the alternate liquidity ratios.

Our interpretation is that when using $BookLiabilities$ as the denominator in a liquidity ratio, the relationship with the odds implied by our failure indicator behaves similar to what one might expect with a traditional debt service ratio. In this case, even though we are using total liabilities in the denominator, a balance sheet aggregate that includes many more balances than those traditionally used for outstanding debt, the result is for the most part unchanged—declining liquidity is associated with an increasing likelihood of experiencing failure events. Conversely, when we tested a liquidity ratio with only $MarketEquity$ in the denominator, we observed increasing liquidity is associated with an increasing likelihood of experiencing failure events. This result appeared counter-intuitive at first. How could increasing liquidity indicate a firm is more likely to fail?

Our interpretation is that the dynamics of this alternative are dominated by changes in $MarketEquity$, not the term for cash used in the numerator, $CashAndShortTermInvestments$. In other words, perhaps this alternate ratio is better understood as literally the proportion of market value derived from a firm’s cash balances, as opposed to firm cash normalized by capitalization. The ratio of failure class to non-failure class observations increases as this ratio increases towards (and goes beyond) a value of 1, which indicates the firm holds a cash balance equal to the value of the firm itself. This increasing relationship with respect to failure odds may reflect that the market no longer views the firm as a “going concern”, but instead the firm is effectively valued as a holding company of cash (and other liquid assets).

Regardless if our interpretation is correct, we must note a potential problem with using a market capitalization weighted liquidity ratio: greater values may also foreshadow a corporate takeover. Firms that can be purchased for less than what they hold in cash may be ripe for acquisition. Given that our firm failure framework effectively ignores reorganization activity by lumping these outcomes in the non-failure class (the
failure indicator is set to 1 only upon confirmation of a failure event), engineering a liquidity predictor in this way may contribute to drawing inaccurate conclusions from model forecasts. In these cases, additional model validation with merger and acquisitions data may be warranted.

Despite any concerns regarding acquisition outcomes, we chose the ratio with MarketEquity in the denominator as the one to include in our approach. We did so in order to maintain the timeliness factor we referenced in earlier when discussing the big picture ramifications of using the market value of equity as opposed to the book value of equity. This was a guiding principle in all of the revisions to variables that reference firm equity, and we were hesitant to proceed in a different direction for a single predictor without strong evidence to motivate doing so.

To complete the revised predictor we take the log of our chosen liquidity ratio. This added step nicely increases class separation and transforms the support of the variable to the real line. Note, we do not need to multiply the resulting quantity by $-1$ to establish an increasing relationship with respect to failure odds, because of the noted dynamics when using MarketEquity as the denominator, this relationship is already increasing. The improved separation between classes achieved after all of our revisions to the liquidity ratio can be seen in Figure B.6 in Appendix B.1. In keeping with the appropriate naming convention, our new liquidity predictor is $CASHME$.

$$CASHME_{i,t} = \log\left(\frac{\text{CashAndShortTermInvestments}_{i,t}}{\text{MarketEquity}_{i,t}}\right)$$ (4.4)

To ensure the ratio is greater than 0 before taking the log, observations of the numerator CashAndShortTermInvestments that are 0 or negative are plugged with $10$ million. This rule will lump together the region where CashAndShortTermInvestments $\leq 0$ with the region where CashAndShortTermInvestments $\geq$ MarketEquity. These two regions of the underlying sample space have significantly higher than average observed failure rates, and for the purposes of constructing a monotonic predictor of failure, they needed to overlap. It should not be surprising that firms with a 0 or negative cash balance exhibit greater failure odds, and the significance of cash balances eclipsing firm market value was noted earlier when we discussed using only MarketEquity as the denominator in a liquidity ratio. We were hesitant to remove observations with a 0 or negative cash balance, and without plugging, we would be unable to take the log to increase separation between liquidity ratio values slightly greater than 0 and those closer to 1. Alternatively, if observations of CashAndShortTermInvestments $\leq 0$ were plugged with $1,000$ (the smallest positive increment in our dataset), this region is shifted to overlap with the low failure risk region. Our choice of a $10$ million plug successfully achieves the overlap of the two high failure risk regions, which in turn results in a liquidity predictor of failure that appears monotonically non-decreasing.

**Equity Return**

$EXRET$ demonstrated strong significance in a multivariate logistic regression and the decreasing relationship with respect to observed failure odds was what we anticipated from visualizing the classes in Figure B.7 in Appendix B.1. We multiply by $-1$ to establish a non-decreasing relationship, but otherwise this variable was constructed in a satisfactory manner for our purposes with the use of log. Nothing was changed beyond the individual terms from each quarter when calculating the the weighted quarterly average $EXRETAVG$.

$$EXRET_{i,t} = -\log(1 + R_{i,t}) + \log(1 + R_{S&P500,t})$$ (4.5)
Volatility

The volatility predictor, $SIGMA$, demonstrates statistical significance in a linear model with a positive sign on the estimated coefficient. This non-decreasing relationship with respect to observed failure odds is illustrated in Figure B.9 in Appendix B.1, and the revised predictor we propose is illustrated in Figure B.10. Unchanged, $SIGMA$ was in most respects suitable for inclusion but we take the log to ensure support across the entire real line. A secondary benefit in greater separation between the two classes is achieved as well.

$$SIGMA_{i,t} = \log \left( \left( \frac{\sum_{k \in \{t, t-1, t-2\}} r_{i,k}^2}{c_t + c_{t-1} + c_{t-2} - 1} \right)^{\frac{1}{2}} \right)$$  \hspace{1cm} (4.6)
Figure 4.3: Class Conditional Histogram for Quantity CashAndShortTermInvestments / BookLiabilities (log scale)

Figure 4.4: Class Conditional Histogram for Quantity CashAndShortTermInvestments / MarketEquity (log scale)
Market Capitalization

The predictor for capitalization is the lone feature that is ineffective when used in a multivariate logistic regression. The regressor appears significant, but the sign on the estimated coefficient is positive. A quick glance at the class conditional histogram in Figure B.11 in Appendix B.1 will confirm that in fact $RSIZE_i$ exhibits some non-increasing relationship with respect to observed failure odds. If we were to restrict ourselves to a linear model, then, this predictor would require further investigation or removal.

We are not restricted to a linear model, and instead assume a more general monotonic relationship. This condition is already satisfied, but to match the non-decreasing direction of the other predictors, we multiply $RSIZE_i$ by $-1$. Otherwise, it is constructed in a satisfactory manner for our purposes with some use of log.

$$RSIZE_i = -\log\left(\frac{\text{MarketEquity}_{i,t}}{\text{MarketValue}_{S&P500,t}}\right)$$ (4.7)

Book Value

The original book value predictor $MB$ was the most problematic and questionable inclusion in a linear model of failure. It appears significant in a logistic regression, but upon further review this input is flawed because of its bimodal relationship with the odds implied by the model output. We must draw attention to the class conditional histogram for $MB$ in Figure B.11 in Appendix B.1. The artificial point masses created during winsorization at the left and right edge of the distribution suggest that failure odds increase towards high and low quantiles of $MB$. This behavior is a violation of monotonicity and needs to be corrected before inclusion in a linear model or our nonparametric approach.

Revisions to $MB$ will proceed in a fashion similar to our revisions to profitability. In Figure 4.5 we present a class conditional histogram of book to market ratio ($\frac{\text{BookEquity}_{i,t}}{\text{MarketEquity}_{i,t}}$). We chose to work with book value of equity in the numerator (the reciprocal of a market to book ratio) because of the difficulties associated with the observed values of $\text{BookEquity} \leq 0$ being placed in a denominator. The $x$-axis in the figure has been labeled on a log scale to reinforce the observed variation, while using market to book increments to allow for quoting of the displayed tick values as valuation multiples (e.g. the sample mode is a firm that trades roughly at a multiple of 2 times book value). Note that even though we are now working with the book to market ratio, there is still bimodality present. There is increasing failure odds as this ratio trends towards greater negative values and greater positive values. The region near the center of the sample has the lowest observed failure odds.

We will again make use of the absolute value function to transform the raw input ($\frac{\text{BookEquity}_{i,t}}{\text{MarketEquity}_{i,t}}$) into a variable that results in a monotonic relationship with respect to failure odds. Following in the footsteps of our properly named $ANI$ variable, we follow suit by naming the transformed $\text{BookEquity}_{i,t}$ as $ABE_i$. With the book value predictor, however, Figure 4.5 suggests that the “folding” of the variable might be improved if the shifted data were to begin at a ratio value of 1 instead of 0. In this manner, book to market ratios that were observed as 0 are shifted to 1, -0.1 to 1.1, -0.2 to 1.2, and so forth. This shift is equivalent to adding $\text{MarketEquity}_{i,t}$ to the absolute value of $\frac{\text{BookEquity}_{i,t}}{\text{MarketEquity}_{i,t}}$ for values of $\text{BookEquity}_{i,t} \leq 0$. The resulting transform is demonstrated in Figure 4.6, and here we write the definition formally.

$$ABE_{i,t} = |\text{BookEquity}_{i,t}| + \text{MarketEquity}_{i,t} \times \mathbb{1}_{(-\infty, 0)}(\text{BookEquity}_{i,t})$$ (4.8)

$$ABEME_{i,t} = \log\left(\frac{ABE_{i,t}}{\text{MarketEquity}_{i,t}}\right)$$ (4.9)
The combined use of the absolute value function and adding \textit{MarketEquity} to observations where \textit{BookEquity} \leq 0 results in shifting that region to overlap with the other region exhibiting greater failure odds, where \textit{BookEquity} \geq \textit{MarketEquity}. The union of the two high failure regions results in a book value predictor that appears to have an increasing relationship. Finally, we take the log of the revised ratio to increase separation and restore the support to the real line. Figure B.14 in Appendix B.1 presents a class conditional histogram for our final predictor, \textit{ABEME}.
Figure 4.5: Class Conditional Histogram for Quantity BookEquity over MarketEquity (log scale)

Figure 4.6: Class Conditional Histogram for Quantity $ABE / MarketEquity$ (log scale)
Stock Price

The stock price predictor as originally constructed is statistically significant in the context of a multivariate linear model, and we qualitatively agree with the sign on the resulting coefficient: decreasing stock price appears to be associated with greater failure risk. \( PRICE \) already enjoyed support across the real line because of the included the use of the log function. Our visuals suggest this predictor exhibits a monotonic non-increasing relationship with observed failure odds, but as we proceeded with \( EXRET \) and \( RSIZE \), we multiply by \(-1\) to redirect the relationship as non-decreasing.

\[
PRICE_{i,t} = -\log(ClosingPrice_{i,t})
\]  

Additionally, let us note that stock prices used to calculate \( PRICE \) are no longer capped at $15. This heuristic was included originally as an alternate winsorization of sorts, but with no replacement rule for the left edge of the distribution. Regardless of the intention, we remove this step because it is no longer necessary or desirable with our nonparametric approach. We prefer all variables to have full support on the real line.

4.3.2 Monotonicity of Transformed Failure Dataset

All of our 8 predictors now appear to be monotonic non-decreasing with respect to the failure indicator, and are defined to have a theoretical support over the real line. These were our requirements before proceeding with model estimation, while linearity was not included. Our non-parametric approach will let the data speak for itself, constrained only by our assumption of monotonicity.

We will now diagram our failure data problem to draw emphasis to the strong degree of monotonicity and non-linearity exhibited by our features. Figure 4.7 is a scatter plot of \( EXRETAVG \) and \( RSIZE \) using only data from 2014, the last year in our sample. We advise the reader to keep in mind these are our revised predictors that have been multiplied by \(-1\), meaning that increasing \( x \) values correspond to larger negative excess returns and increasing \( y \) values correspond to smaller relative firm capitalizations. We observe that among the smallest \( RSIZE \) values (largest firms) there is very little variation in \( EXRETAVG \), but this is no longer true with larger \( RSIZE \) values (smaller firms). This dynamic is wholly predictable given that the quarterly \( EXRET \) values are defined to be excess returns with respect to the S&P 500 and our sample is limited to U.S. listed equity. The largest firms by capitalization have the greatest influence on the capitalization weighted index, and through that influence will never be able to significantly deviate from it. Meanwhile greater volatility of returns is expected with smaller capitalization stocks.

We are most interested in the locations of the relatively small number of red failure points (20, out of roughly 52,000 total). It is clear they are all clustered in the top right of the plot, and the number of surrounding non-failure points significantly decreases at increasing \( x \) and \( y \) values. We are observing empirically with our re-engineered features and will later assume theoretically with model selection is that failure risk appears monotonically non-decreasing across all dimensions, and yet definitely not linear. The importance of this fact cannot be understated. This insight was crucial to the re-engineering of features and our approach overall.

To expand on our definition of monotonic data earlier, and assist with describing our use of kernels in Section 4.4, let us now look at Figure 4.8. This is the same plot, only there has been annotations added. Going forward, when discussing the joint right tails of all of the predictors we will refer to this region as the “right corner” of the sample space. With the additional consideration of the failure indicator along
with the predictors and yet without assuming any model, we assert the right corner is the region with the
greatest failure risk. As we move closer to the right corner (or more precisely, the value of any dimension of
any hypothetical firm-month input vector increases) we expect failure risk to increase or remain unchanged.
The converse to these three assessments holds as well. The region defined by the joint left tails of all the
predictors will be referred to as the “left corner” of the sample space, we assert the left corner is the region
of least failure risk, and as we move closer to the left corner we expect failure risk to decrease or remain
unchanged.

Qualitatively speaking, the empirical support of a multi-dimensional class can be thought of as the region
delineated by a convex hull surrounding the observations. With the assumption of monotonicity across all
dimensions with respect to the class label, the empirical support can be extended to include the region of all
points that lie within an appropriate hyper-cube that is defined by either the dimension-wise minimum or
maximum point of the class. The dimension-wise min (or max) point simply means the hypothetical point
defined by taking the sample minimum (or maximum) of each univariate sample distribution separately (and
may be an actual observation in some cases). For our failure class we have established a monotonic non-
decreasing relationship across all dimensions, meaning the “appropriate hyper-cube” that more completely
defines the support of the failure class includes all points that dominate the dimension-wise minimum point
(any vector where every dimension is greater than or equal to the dimension-wise min).

If we temporarily restrict ourselves to the 2-dimensional sub-sample displayed in Figure 4.8, we see the
class 1 support explicitly delineated with red arrows. At any point within the region contained by the arrows,
we know that a forecasted probability of failure must be greater than 0 (discounting any assumption of model
noise). Outside of this region, the forecasted probability may be in fact 0. If at least 1 failure event has been
observed, then we can say the failure class support is some non-empty set, and it must contain the region
we defined as the right corner.

To provide further re-enforcement and better demonstrate the massive size of the failure dataset, we
present in Figure 4.9 what is effectively the same plot as in Figures 4.7 and 4.8. With this iteration,
however, we are using the entire sample and a colorized density surface for the non-failure class (direct
plotting of every observation in the non-failure class would be textbook over-plotting). The general shape
of the non-failure class is unchanged. There is much less variation in $EXRETAVG$ for the smaller values
of $RSIZE$ compared to larger ones. The general clustering of failure points in the right corner is also
unchanged, but in displaying every failure event ever to be observed, we see much greater variation within
the right corner. There is also now some overlap of failure observations throughout the non-failure class, but
there is noticeably fewer failures towards the left corner.

When plotting the full sample, the intensity of monotonicity exhibited may be less than when plotting
individual years such as 2014, but we are still confident that the assumption holds. Regardless of what
figure is referenced, we are also confident that the relationship between predictors and the odds implied by
the failure indicator is not linear. Failure risk will be extremely low and nearly constant throughout the
left corner and even middle of the sample where the non-failure class density is greatest, but then rapidly
increase in the right corner. Let us recall a criticism against a linear model of failure from the beginning of
this section on the basis of the $RSIZE$ coefficient. We noted that logistic regression results in an estimated
coefficient with a positive sign, which contradicts the decreasing relationship with respect to failure odds
that is apparent in the univariate distribution of this predictor (before our transformations). This criticism
is supported perhaps more strongly by the bivariate plots we have discussed here in Figures 4.7 and 4.9. It
is clear that failure odds increase as firm capitalizations decrease and excess equity returns decrease (both
quantities are with respect to the S&P 500 index), but that relationship is not linear. In the next two sections we will describe and then demonstrate our non-parametric approach that leverages the assumption of monotonicity to efficiently capture the non-linear dynamics exhibited by $EXRETAVG$, $RSIZE$, and the remaining 6 predictors of failure.
Figure 4.7: Monotonicity of MME Dataset (2014 only)

Figure 4.8: Monotonicity of MME Dataset (2014 only) with Annotations
Figure 4.9: M/M/E Dataset (all years), Excess Return and Capitalization Predictors
Figure 4.10: MME Dataset (all years), Profitability and Book Value Predictors
4.4 Approximation of Rare Events Datasets Using Nonparametric Density Estimation

The data problem we face is binary classification and we have an extremely low minority class proportion (less than 0.01%). Otherwise referred to as rare events, this problem is challenging enough as is for all the reasons highlighted in the literature we discussed earlier in Section 4.1. At first one may be tempted to implement methods that are more sophisticated than logistic regression to better learn the limited amount of signal (minority class information) available. Unfortunately, that could take weeks of computation for a single fitting, because of the roughly 700,000 observations in our thresholded dataset (or 2.4 million for the raw dataset). Clearly, we are working with “big data” as well. One might think, then, that we are confronting the worst of both worlds from big and small data. To overcome this enormous challenge, we will first take advantage of the low dimension of the sample space (only 8 predictors), and then later in Section 4.6 make an appeal to the principle of monotonicity to further improve our position.

In a rare events situation, our intuition was revealed after stepping back from the whole picture. We realized that we should be focusing on only the information we know to be certain, and not wasting time with what we don’t know much about. With the firm failure dataset, the information provided by the locations of firm failure (class 1 points) should be the focus of our attention. Conversely, we will try to minimize or perhaps ignore completely the regions of the sample space where there are few or no class 1 observations.

We will proceed by estimating conditional probabilities \( P[Y_i = 1 | X_i = x_i] \) at the minority class points in the training set and then set these estimates as the targets in a regression. With only 8 variables (our failure dataset does not suffer from “the curse of dimensionality”), a kernel density estimator is the perfect tool for efficient estimation of the conditional probabilities. This unique application requires an appropriately chosen kernel bandwidth to preserve the majority class information we need for modeling, but we will demonstrate later this decision is not difficult and that the result—keeping only minority class points and discarding the majority class—is well worth it. After the application of the density estimator (and a link function), we will find ourselves in a regression setting with small data, an advantageous position that allows more powerful learning methods compared to rare events classification with big data. This can be thought of as analogous to reducing the effective number of variables with the help of a principal components transform. Instead, with our “rare events kernel density transformation”, we are trying to make the problem more manageable by reducing the number of observations and converting from a discrete to a continuous target.
4.4.1 Specification of KDE Transform

Before we start the specification of our dataset transformation based on kernel density estimators, let us quickly review some basic notation.

\[ \frac{X}{\Sigma}, \frac{X}{\Sigma} \] random variable and vector (Latin/Greek) (4.11)

\[ \frac{x}{\sigma}, \frac{x}{\sigma}, \frac{x}{\Sigma} \] observed scalar, vector, and matrix (Latin/Greek) (4.12)

\[ X, y \] training data matrix and targets vector (4.13)

\[ x_*, y_* \] test point and target values (4.14)

\[ X_*, y_* \] test data matrix and targets vector (4.15)

\[ n, d \] constants for dimensions of training data matrix (4.16)

\[ 0, 1 \] \( d \) dimensional vector of all 0s and 1s (4.17)

\[ f : \mathbb{R}^d \rightarrow [0, \infty), g : [0, 1] \rightarrow \mathbb{R} \] generic density function and link function (4.18)

\[ P, E \] probability and expectation operators (4.19)

\[ \phi(x|\mu, \Sigma), \Phi(x|\mu, \Sigma) \] multivariate Gaussian density and distribution functions (4.20)

To repeat, the goal of our transformation is to allow for the application of a more sophisticated, non-linear learning method for forecasting firm failure. Before transformation, the underlying data problem can be summarized as rare events binary classification, with firm failure, or class 1, being the minority class, and with big data but only a small number of predictors. We can write this problem using notation as follows.

\[ Y_i | X_i = x_{(i)} \sim Bernoulli(\pi_i) \] Bernoulli likelihood with latent class 1 probability \( \pi \) (4.21)

\[ x_{(i)} \in \mathbb{R}^d, y_i \in \{0, 1\} \] \( i \)th observation vector and training target (4.22)

\[ n \approx 1 \text{ million}, d \approx 10 \] big data, small number of predictors (4.23)

\[ n_0 = \sum_{i=1}^{n} I\{y_i\}, n_1 = \sum_{i=1}^{n} I\{1\}(y_i) \] number of class 0 and class 1 observations (4.24)

\[ n = n_0 + n_1, n \approx n_0, n_0 >> n_1 \] class 1 observations are rare events (4.25)

Our transformation is based primarily on two additional assumptions beyond the standard Bernoulli likelihood for \( Y | X \). The first is to assume independently distributed Gaussian noise on the predictors. Our predictors are now defined to be proper random vectors because of the assumed noise, \( X_i = x_{(i)} + \Xi, \Xi \sim \text{Gaussian}(0, \sigma^2 1) \), for some vector \( \sigma^2 \) that will be chosen later. The noise is independent in two regards–from the inputs (homoscedastic), and between dimensions (diagonal covariance matrix on the noise distribution). Changing either decision could be considered in future work.

It could be said that this step in our overall transformation constitutes an “errors in variables” approach. The qualitative interpretation for our firm failure dataset is such that all firm fundamental (model predictor) observations now represent a realization of some underlying random process of the firm. For example, a given value of \( \text{CASHMTA} \) represents one possibility of cash holdings for the given firm and month, and if we were to go back in time and repeat this period we may observe a slightly different balance. Even if a practitioner’s motivations were different than ours, assuming noise on the inputs is sensible and perhaps beneficial because it allows for accounting irregularities or misstatements, and other market idiosyncracies that should not be considered as signal by the failure model. Uncertainty in the inputs will translate to more uncertainty in
forecasts and allow for more reliable conclusions.

The assumption of independent Gaussian noise on the predictors is functionally equivalent to assuming that the joint density of the sample can be estimated using a kernel density estimator with a Gaussian kernel. Written this way, \( f(x) = \frac{1}{n} \sum_{i=1}^{n} \phi(x|x_{(i)}, h^21^\prime) \), it should be clear that making a decision on the noise variance \( \sigma^2 \) is the same as selecting a bandwidth \( h^2 \) for the kernel density estimator. Regardless of the interpretation, it is important to remember that using a multivariate kernel with a diagonal covariance matrix is not the same as using a product of univariate density estimates. We will still capture the multivariate dependence structure while assuming a simplified covariance matrix within each kernel.

In combination with the observed class labels, the assumption of a kernel density estimator on the sample of predictors provides everything we need to estimate conditional class 1 (or class 0) probabilities. One could say we have “induced” a realization of all the latent states, \( p \), without formally estimating a classifier. This state of affairs is possible because although the class labels will factor in to the resulting calculations, their distribution can be safely ignored when fitting the kernel density estimator (i.e. choosing a bandwidth). The chosen kernel and bandwidth will be the same for class 0 and class 1 observations. The derivation of the estimator for the latent states is similar to what you see with a “naive bayes” calculation, but when using a kernel density estimator instead of assuming Gaussianity, and also when using a multivariate kernel instead of assuming independence. In notation, we describe our latent state estimator as follows.

\[
\pi_* = P[Y = 1|X_* = x_*] \text{ conditional probability of class 1} \tag{4.26}
\]

\[
X_i = x_{(i)} + \Xi, \Xi \sim \text{Gaussian}(0, h^21^\prime) \text{ Gaussian noise with covariance matrix } h^21^\prime \tag{4.27}
\]

\[
\pi_* = x_*, h = \frac{f(x_*)|Y = 1)P[Y = 1]}{f(x_*)} \text{ latent state} \tag{4.28}
\]

\[
P[Y = 1] = \frac{n_1}{n} \text{ observed class 1 proportion} \tag{4.29}
\]

\[
(x_{(i)}, y_i = 1) : 1 \leq i \leq n_1 \text{ observations 1 to } n_1 \text{ of training data are class 1} \tag{4.30}
\]

\[
\hat{f}(x|Y = 1) = \frac{1}{n_1} \sum_{i=1}^{n_1} \phi(x|x_{(i)}, h^21^\prime) \text{ KDE for class 1 sample} \tag{4.31}
\]

\[
\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \phi(x|x_{(i)}, h^21^\prime) \text{ KDE for full sample} \tag{4.32}
\]

We can simplify terms to write our final estimator.

\[
\hat{\pi}_* = \hat{P}[Y = 1|X_* = x_*, h] = \frac{\hat{f}(x_*)|Y = 1)\hat{P}[Y = 1]}{\hat{f}(x_*)} \tag{4.33}
\]

\[
= \frac{\left( \frac{1}{n_1} \sum_{i=1}^{n_1} \phi(x_*, x_{(i)}, h^21^\prime) \right) \frac{n_1}{n}}{\frac{1}{n} \sum_{i=1}^{n} \phi(x_*, x_{(i)}, h^21^\prime)} \tag{4.34}
\]

\[
= \frac{\sum_{i=1}^{n_1} \phi(x_*, x_{(i)}, h^21^\prime)}{\sum_{i=1}^{n} \phi(x_*, x_{(i)}, h^21^\prime)} \tag{4.35}
\]

\[
= \frac{\sum_{i=1}^{n_1} \exp[-\frac{1}{2}(h^{-2})'(x_* - x_{(i)})^2]}{\sum_{i=1}^{n} \exp[-\frac{1}{2}(h^{-2})'(x_* - x_{(i)})^2]} \tag{4.36}
\]

\[
= \left( 1 + \frac{\sum_{i=n_1+1}^{n} \exp[-\frac{1}{2}(h^{-2})'(x_* - x_{(i)})^2]}{\sum_{i=1}^{n_1} \exp[-\frac{1}{2}(h^{-2})'(x_* - x_{(i)})^2]} \right)^{-1} \tag{4.37}
\]

For those familiar with the theory connecting histograms to kernel density estimators, this final expression...
can be interpreted as the ratio of histogram bar heights, but calculated using Gaussian density curves instead of histogram blocks (and without binning, or rounding of observed data). In Line 4.36, we are simply dividing the bar height at a point evaluated with only the failure class by the full sample bar height at the same point. In Line 4.5 terms are rearranged to eliminate computing the failure kernel sum twice.

Line 4.5 in our derivation elicits two important comments. First, we removed the common terms outside of the exponential kernel, and because the Gaussian density covariance matrix is the diagonal matrix of squared bandwidths, we are able to rearrange terms within the exponential kernel. This change significantly improves the computational complexity and memory requirements of the final software implementation. With the arrangement seen here, we avoid computing all of the off diagonal terms normally required with the evaluation of the exponential kernel. Also, note that the summation at the core of every kernel density estimator is fully parallelizable, where each kernel evaluation can be executed independently before all values are summed together. Taken together, these advantages can be leveraged so that the final estimated probabilities are generated efficiently even with large datasets.

Second, by using the Gaussian kernel (i.e. we have infinite support) and the same bandwidth in the numerator and denominator of our conditional density estimator, we can easily claim that for all $\mathbf{x}, \hat{P}[Y = 1|\mathbf{X} = \mathbf{x}, \mathbf{h}] \in (0, 1)$. It is preferable that the range of this estimator not include either boundary value for our underlying application of modeling ex ante failure risk. At the start of a given time period, we can never say with complete certainty that a firm will or will not fail (failure risk of 1 or 0, respectively). Thus, we are preserving this aspect of the data required by our application domain for any subsequent regression method chosen. It could make sense to allow these boundary probabilities in a fully Bayesian approach, whereby a distribution of probabilities is modeled at each observation, but that is not relevant for our currently defined framework so we leave this consideration for future work.

The second of the primary assumptions that we make to enable our approach is that we can approximate a complete classifier of a rare events dataset with a regression of the induced latent states (transformed with an appropriate link function) on the predictors, but using only latent states and predictors at the minority class points. The locations of majority class points will not be used directly in estimating the regression. We are claiming that the information provided by the majority class will be incorporated via the latent states at the minority class points. To bring our principal components analogy full circle, our claim is that the transformed dataset comprised of latent states estimated at failure points only retains sufficient information for learning a useful model, just as a dataset transformed using PCA might retain sufficient information in the first handful of principal components. Moreover, both our kernel based transformation and PCA transformations would not necessarily imply an approximation if latent states were generated at all points (majority included) and all principal components were retained, respectively.

With a large dataset, and by using a nonparametric method to induce class 1 probabilities, we will be able to sufficiently preserve the local structure present in the dataset where it matters most, at or near points from the minority class. Conversely, we will be ignoring any structure present in the regions of the sample space that are dominated by the majority class, and given our rare events setting, this applies to a substantially large proportion of the space. Furthermore, with our assumption of a monotonic increasing relationship between distress risk and the predictors, we can say that we are actually just ignoring (or underrepresenting) a single, large region of the sample space that represents the corner formed by the minimum of each predictor
variable. We can now write our much more manageable problem setting using notation.

\[ z_i = g(\hat{\pi}_i) = g(\hat{P}[Y_i = 1|X_i = x(i)]) \] transformed target values \hfill (4.38)

\[ Z_i = r(x(i)) + \Upsilon, E[\Upsilon] = 0 \] unknown regression function \( r \) with additive noise \( \Upsilon \) \hfill (4.39)

\[ x(i) \in \mathbb{R}^d, z_i \in \mathbb{R} \] ith observation vector and transformed training target \hfill (4.40)

\[ (x(i), y_i = 1, z_i) : 1 \leq i \leq n \] ignore \( z_i \) at all class 0 observations \hfill (4.41)

\[ n_1, d \text{ constants for dimensions of transformed data matrix} \] \hfill (4.42)

\[ n_1 \approx 1000, d \approx 10 \text{ small data, few variables} \] \hfill (4.43)

\[ \hat{y}_* = g^{-1}(\hat{z}_*) \text{ test point class 1 forecast} \] \hfill (4.44)

With this transformation from classification to regression we are focusing on what we know about class 1, while still taking some account of what we know about class 0 through the denominator in our latent state estimator. Ultimately, the vast majority of raw data are class 0, and so this class can be thought of as noise that is corrupting our ability to learn the signal, or class 1. It is interesting to note that if we set our estimator bandwidth \( h = 0 \), our approach degenerates to a “reverse anomaly detection” model of strictly using class 1 observations. On the other hand, if we set our estimator bandwidth \( h = \infty \), our approach degenerates to a null model that forecasts a constant at all points throughout the sample space, the sample mean class 1 probability \( \frac{n_1}{n} \). In this light the bandwidth vector represents the degree to which the practitioner wants to emphasize the minority class. When \( h = 0 \), we have full emphasis on the minority class (no information from the majority class is allowed), and when \( h = \infty \) there is no information available from either class such that we produce only a deterministic conclusion of the sample mean everywhere.

We have some final points to make about the homoscedastic noise assumption before proceeding. In the context of the assumed kernel density estimator alone, the noise we assume is constant variance and independent of the predictors. However, this assumption results in heteroscedasticity of the transformed regression targets \( z_i \). Given a bandwidth selection, every \( z_i \) has a unique distribution because of the unique sums of kernel density terms involved, and these distributions are very much dependent on the locations of observed predictors. If there was a need, one could very easily sample directly from the distribution of each \( z_i \) (or \( \pi_i \)) by using our assumed kernel density estimator to simulate new data at each observed \( x(i) \), and then for each draw re-estimate the transformed target \( z_i \).

So in the context of our dataset transformation in sequence with a regression method, the homoscedastic noise assertion is not entirely true. There will be heteroscedasticity present in the transformed regression targets, and a lack of complete accounting for this observation noise model by the regression method would count as an additional reason for referring to our transformation plus regression framework as an approximation to a true classifier. The first and most obvious reason for referring to our framework as an approximation would be the discarding of induced latent states at majority class points when fitting the regression. One might be able to make an argument that full recovery of the original classification problem is possible if all non-failure points were included in the regression and the nonparametric, heteroscedastic observation noise was completely accounted for, but we leave that proof for future work. Overall, this transformation is incredibly simple to perform. With a diagonal smoothing matrix, every latent state estimation is \( O(n * d) \), and estimating latent states only at minority class points \( n_1 \ll n \) is then \( O(n_1 * n * d) \). The result is a much more manageable dataset with observation count equal to minority class observations, but one that still captures most of the relevant information present in the original dataset. To maximize the informational
contained in the transformed sample and minimize the impact of noise or artifacts that could be introduced by an approximation, we will select a monotonically constrained regression method afterward. These benefits are achieved with only minimal amounts of computation and the additional decision of selecting a bandwidth vector that is the same dimension of the sample space. We shall demonstrate later in Section 4.5 that there are solid general purpose, “plug-in” rules that work without further consideration in addition to other cross-validation or optimization based rules.

4.4.2 Assumed Probability Model

Before moving on to bandwidth selection we need to take a step back and more properly define our kernel density based transformation as an approximation of a model of the assumed latent states (class 1 probabilities). This will not only improve our grasp of the approach, but also allow for the enumeration of other variations of the transform that may prove more effective.

Let us begin with an extended restatement of the assumptions placed on the data from Section 4.4.1 using notation.

\[ Y \sim \text{Bernoulli}(p) \quad (4.45) \]
\[ f(y) = p \cdot \mathbb{I}_{(1)}(y) + (1 - p) \cdot \mathbb{I}_{(0)}(y) \quad (4.46) \]
\[ P[Y = 1] = p \quad (4.47) \]
\[ X_i = x(i) + \Xi \quad (4.48) \]
\[ \Xi \sim \text{Gaussian}(0, h^2 1^T) \quad (4.49) \]
\[ f(x|h), f(x|h, Y = 1), h : \text{sample KDE, class 1 KDE, bandwidth vector} \quad (4.50) \]
\[ \pi = P[Y = 1|X = x, h] = \frac{f(x|h, Y = 1)P[Y = 1]}{f(x|h)} \quad (4.51) \]

The Bernoulli observation model for \( Y \) in combination with the noise assumption on \( X \) allows us to quickly arrive at an estimator for the conditional probability of observing class 1 (i.e. the latent state). This estimator of the latent states \( \pi \) was originally used only to generate new target values for regression, but here we would like to acknowledge the estimand formally as a random variable with a density that can be manipulated. Let \( \Pi := E[Y|X] = P[Y = 1|X] \), and now use this new variable to write an alternate conditional probability model for \( Y \).

\[ Y_*|\Pi_* = \pi, X_* = x_* \sim \text{Bernoulli}(\pi) \quad (4.52) \]
\[ f(y_*|\pi, x_*), h = \pi \cdot \mathbb{I}_{(1)}(y*) + (1 - \pi) \cdot \mathbb{I}_{(0)}(y*) \quad (4.53) \]
\[ P[Y_* = 1|\Pi_* = \pi, X_* = x_* , h] = \pi_\ast \quad (4.54) \]

It should be self evident that \( \Pi \) is redundant. All that we have done is written the conditional probability law of \( Y \) in terms of \( \pi \) instead of the Bayesian arrangement of density estimators that defines \( \pi \). It is true that \( P[Y = 1|\Pi_* = \pi, X_* = x_* , h] = P[Y = 1|X_* = x_* , h] \), which is to say that conditional on \( \Pi \), there is no additional knowledge about the distribution of \( Y \). This setup is not without purpose. By assuming the existence of the latent state \( \Pi \), we can also assume the existence of some unknown marginal density
f(\pi | x, h). With this unknown density and \( f(y_*|\pi_*, x_*, h) \) we can derive \( f(y_*|x_*, h) \) as follows.

\[
f(y_*|\pi_*, x_*, h)f(\pi_*|x_*, h) = (\pi_*I_{\{1\}}(y_*) + (1 - \pi_*)I_{\{0\}}(y_*)) * f(\pi_*|x_*, h)
\]
\[
f(y_*|x_*, h) = \pi_* * f(\pi_*|x_*, h)I_{\{1\}}(y_*) + (1 - \pi_*) * f(\pi_*|x_*, h)I_{\{0\}}(y_*)
\]
\[
\int f(y_*|\pi_*|x_*, h)d\pi_* = \left[ \int \pi_* * f(\pi_*|x_*, h)d\pi_* \right]I_{\{1\}}(y_*)
\]
\[
\left[ \int (1 - \pi_*) * f(\pi_*|x_*, h)d\pi_* \right]I_{\{0\}}(y_*)
\]
\[
f(y_*|x_*, h) = E[\Pi_*|X_* = x_*, h]I_{\{1\}}(y_*) + E[(1 - \Pi_*)|X_* = x_*, h]I_{\{0\}}(y_*)
\]
\[
= E[\Pi_*|X_* = x_*, h]I_{\{1\}}(y_*) + (1 - E[\Pi_*|X_* = x_*, h])I_{\{0\}}(y_*)
\]

We now have an alternate definition for the conditional probability of observing class 1, \( P[Y_* = 1|X_* = x_*, h] = E[\Pi_*|X_* = x_*, h] \). Abstracting the latent state of \( Y \) into the random variable \( \Pi \) allowed us to re-write the objective of a binary classification problem as a regression problem instead. We have changed the game, so to speak, and in our situation that is a massive advantage. As noted at the beginning of this section, large binary classification datasets are extremely challenging to model. With the new objective \( E[\Pi_*|X_* = x_*, h] \), however, we can solve our problem with a monotonically constrained regression method in conjunction with an approximation of the majority class. In using this alternate line of attack on binary classification, we hope to achieve a substantial decrease in workload along with improved model estimation.

Based on the above derivations let us briefly discuss variations of our kernel density based transformation originally described in Section 4.4.1.

A1. Estimate \( E[\Pi_*|X_* = x_*, h] \) by training a regression with latent states at all dataset points.

This variant is mentioned only for reference purposes. Using all latent states would mean we are no longer approximating the objective, just transforming it from classification to regression with noisy inputs. This variant is possible, but probably a poor choice because calculating the latent states for the full dataset is not insignificant computation. More importantly, if you are going to include all observed points in the model training data, you are measuring all the information necessary to calculate class 1 probabilities through full accounting of both classes so the transformation provides no benefit.

A2. Estimate \( E[\Pi_*|X_* = x_*, h] \) by training a regression with latent states at class 1 points only.

This is the approach presented earlier in Section 4.4.1. We mention it here to provide contrast and a complete interpretation. Our view is that the ignored points from the majority class provide little to no information beyond what was measured by their impact on the generated latent state targets at the class 1 points. More precisely, we will ignore regions outside of the minority class support while paying ample attention to the most informative points in the dataset–the failure points, and so even though failure is extremely rare our approach is maximizing their value. This variation of our transformation is the simplest as it represents the most extreme approximation of \( E[\Pi_*|X_* = x_*, h] \).

A3. Extend Variant A2 with a heuristic for including induced latent states at class 0 points.

This variant falls somewhere in between including all latent states or only latent states at class 1 points in the model training data. An obvious choice for included class 0 points would be the set of points that determine the empirical minimum value for each input dimension. With our failure dataset, this would be the “safest” or least risky non-failure firm-month in each dimension (while ignoring the dimension
of time). The latent states at these points will most likely be close to 0, and with their inclusion any monotonically constrained regression method should be able to interpolate downward to them from the originally included failure points.

Alternatively, one could pursue some grid strategy whereby you estimate latent states at evenly spaced points that are far from the minority class centroid and extend towards the left corner of the sample space. Given that class 0 is the most probable outcome (by far) almost everywhere, there is not much reason to be restricted to observed non-failure points only. Meanwhile, because class 1 is so rare, latent states at points observed in this class should always be included.

A4. Develop some strategy for resampling from the kernel density estimator at randomly chosen failure and non-failure points, and performing the chosen regression method on each sample, and then aggregating all the regressions.

This would be similar to many established rare events approaches that were highlighted in Section 4.2 calling for down-sampling the majority class, or up-sampling the minority class, but using regression instead of classification.

The question one is attempting to answer when choosing between Variant A2 and A3 (or 4.5), is “Will there be upward bias in the estimated regression model if it is trained using latent states sampled only at the observed failure points?” If the chosen regression method is nonparametric and monotonically constrained (i.e. local structure is emphasized and relationships showing increasing class 1 probability towards the left corner are disallowed), that might not necessarily be true. At points within the observed support of the minority class (the region indicated by failure points included in the training set at a given point in time), given the extremely large number of majority class points that will be acknowledged in the kernel estimator and general amount of overlap between the two classes, upward bias should be minimal and will converge to 0 as $h \to \infty$. There could be bias when forecasting at points close to and beyond the edge of the observed minority class support in the left corner of the sample space, however. Therefore, any heuristic employed to correct for perceived upward bias should place emphasis on that region.

To assist with the development of a heuristic that accounts for information lost from not including latent states from all class 0 points in the regression, we can take our derivation of the probability model one step further. If we assume the existence of some unknown marginal densities for the latent state $\pi$ that are also class conditional, the objective $E[\Pi_*|X_*=x_*,h]$ can be partitioned into separate models of the majority and minority class. Once partitioned, we can develop approximations of the majority class in isolation, allowing for full model validation while using the same minority class component throughout.

This can be easily understood by writing out the split-class probability model using notation. Let $f(\pi|x,h,Y=1)$ and $f(\pi|x,h,Y=0)$ be the class 1 and class 0 marginal densities for the latent state, and
then we decompose our regression objective as follows.

\[
E[\Pi_*|X_* = x_*, h] = \int \pi_* \ast f(\pi_*|x_*, h) d\pi_*
\]

\[
= \int \pi_* \ast \left[ f(\pi_*|x_*, h, Y_* = 1)P[Y_* = 1|\ldots ] + f(\pi_*|x_*, h, Y_* = 0)P[Y_* = 0|\ldots ] \right] d\pi_* \tag{4.60}
\]

\[
= \left[ \int \pi_* \ast f(\pi_*|x_*, h, Y_* = 1) d\pi_* \right] P[Y_* = 1|\ldots ] + \left[ \int \pi_* \ast f(\pi_*|x_*, h, Y_* = 0) d\pi_* \right] P[Y_* = 0|\ldots ] \tag{4.61}
\]

\[
= E[\Pi_*|X_* = x_*, h, Y_* = 1]P[Y_* = 1|\ldots ] + E[\Pi_*|X_* = x_*, h, Y_* = 0]P[Y_* = 0|\ldots ] + E[\Pi_*|X_* = x_*, h, Y_* = 0](1 - P[Y_* = 1|\ldots ]) \tag{4.62}
\]

From Line 4.64, it is clear that we have now partitioned our original regression objective into two–one expectation for each class. This result also necessitates the introduction of class probability terms, \( P[Y_* = 1|\ldots ] \) and \( P[Y_* = 0|\ldots ] \) (or \( (1 - P[Y_* = 1|\ldots ]) \)), which function as partition weights. The dots notation allows for using either unconditional or conditional probabilities for \( Y_* \). The terms \( x_* \) and \( h \) are viewed as given and could be used without invalidating the derivation. The usual caveat that increasing complexity is not necessarily desirable applies. Sensible estimates for these probabilities might be the unconditional sample mean, a linear logistic model, or a kernel density estimator. If using the density estimator with a Gaussian kernel and the same bandwidth \( h \) chosen for our transform, this would mean re-using the latent states.

With this split-class model, we will be training one regression with latent states estimated at all class 1 points (only), and then training some other regression or approximation using some subset of points from class 0. It is extremely important to recognize that this split model is possible because the latent state \( \Pi \) is an abstraction of ex ante failure risk, not ex post. All firms will have some \( \pi > 0 \) because there is always some risk in the market. Firms that do not fail (observing \( [Y_* = 0] \)), still had some risk of failure at the beginning of the observed month. Conversely, firms that fail (observing \( [Y_* = 1] \)), did not have an ex ante risk of 1 (i.e. guaranteed failure) because good fortune in the form of an angel investor, lax lender, or competitor’s mistake is always possible. The distributions for latent state variables do not degenerate with the passage of time. The variables are designated according to a specific month start (and firm), and measure the information available at that moment no matter what might be observed in the future.

The class conditional distributions underlying the pair of regression objectives, \( f(\pi_*|x_*, h, Y_* = 1) \) and \( f(\pi_*|x_*, h, Y_* = 0) \), can be thought of as distributions of firm risk levels that are possible because of any market outcome. The distribution for \( \Pi \) conditioned on \( Y = 1 \) is an abstraction of all outcomes that push a company towards failure: natural disaster, fraud, botched product launches, secular decline of industry, etc. The distribution for \( \Pi \) conditioned on \( Y = 0 \) is an abstraction of all outcomes that push a company away from failure: increasing revenues, decreasing costs of raw materials, succesful roll out of new products, improved macro-economic conditions, etc. This line of interpretation permits us to read the last line of our derivation, line 4.64, as if we are asking, “if we assume this firm will fail, what \( \Pi \) can we expect?” and “if we assume this firm will not fail, what \( \Pi \) can we expect?”. Model forecasts are generated as if we temporarily assume failure or non-failure to assess all the possibilities that could result in each event (without knowing the outcomes), and then the two assessments (expectations) are aggregated as a weighted average.
We will now briefly describe variants of our framework that make use of the split-class model. Note, each entry represents an extension to the same numbered variant listed earlier when discussing regressions that estimate a single objective. Variant B1 is an extension of A1 and so forth. Also, we do refer to specific estimators of \( P[Y = 1 | \ldots] \) in some of the variants here that will be evaluated later on in Section 4.6, but more generally all the variants we describe may be implemented with another choice.

B1. Estimate \( E[\Pi_* | X_* = x_*, h, Y_* = 1] \) and \( E[\Pi_* | X_* = x_*, h, Y_* = 0] \) using latent states generated at all failure and non-failure points, respectively.

This variant is once again mentioned as a hypothetical reference only, and the same drawbacks still apply (because sampling the latent states at all dataset points is not efficient). The split-class model offers advantages that make this variant more intriguing, however. You now have the opportunity to implement entirely different regression methods for the class 0 and class 1 data, where the separate regressions could each be considered anomaly detection.

Anomaly detection approaches are based on training a model using only the data from a single class. A model trained using only majority class data would be standard anomaly detection. Using this model, the anomalous location of a test point would suggest it is from the minority class. On the other hand, a model trained using only minority class data would be “reverse” anomaly detection. Here, model forecasts indicating that a test point has a standard or not unusual location would suggest it is from the minority class.

B2. Estimate \( E[\Pi_* | X_* = x_*, h, Y_* = 1] \), using induced latent states at failure points only and ignore \( E[\Pi_* | X_* = x_*, h, Y_* = 1] \).

This variant is the simplest approximation of the full model, and is exactly the same as Variant A2 mentioned earlier, but with a different interpretation. Here, we do not assume that \( E[\Pi_* | X_* = x_*, h, Y_* = 1] \approx E[\Pi_* | X_* = x_*, h] \), and instead assume \( P[Y = 1 | \ldots] = 1 \) and \( P[Y = 0 | \ldots] = 0 \). Thus, we are assuming failure, and from that perspective we are approximating the full model and generating forecasts for the latent states.

For any estimator of \( P[Y = 1 | \ldots] \) other than a constant 1, some kind of information from the majority class will begin to be incorporated from the chosen \( E[\Pi_* | X_* = x_*, h, Y_* = 0] \) \( P[Y = 0 | \ldots] \) terms. Because of this fact, those alternatives will be considered separately under Variant B3.

B3. Estimate \( E[\Pi_* | X_* = x_*, h, Y_* = 1] \) using latent states at all failure points, and estimate \( E[\Pi_* | X_* = x_*, h, Y_* = 0] \) using some heuristic for sampling latent states at non-failure points (or some grid of evenly spaced points positioned away from the failure class points).

The key difference between this and Variant A3 is that latent states sampled from non-failure points (or a grid) are incorporated in a separate regression for class 0 instead of being lumped in to a single dataset for training a single regression. This separation can also be useful without using any sampled latent states in addition to those generated at failure points for the class 1 model. For any estimate of \( P[Y = 1 | \ldots] \) that is not a constant 1, estimating a null model for \( E[\Pi_* | X_* = x_*, h, Y_* = 0] \) to be 0 everywhere requires no effort but will help minimize bias in the left corner of the sample space.

For example, let us consider estimating \( P[Y = 1 | \ldots] \) to be a simple step function with respect to the dimension-wise minimum values of class 1 (or some other reference threshold). This step function outputs a probability of 1 for all locations within the observed class 1 support, and then a probability
of 0 for all locations outside this support. According to this step function estimator for $P[Y = 1|...]$, points within the support of the minority class are forecasted according to $E[\Pi_*|X_* = x_*, h, Y_* = 1]$, and then points outside the minority class support are assumed 0. This could also be thought of as a piece-wise model, such that forecasts at test points located outside of the minority class support, where the majority class dominates, are censored to 0. Censoring the forecasts of test points located there saves substantial computation compared to a full classification model fitting, and may even improve forecasting performance metrics depending on the density of the test set. This might be the best variant overall because of the minimal effort beyond the proposed, basic transformation, but which also controls for upward bias.

More complicated step functions, or linear models, etc. may also be considered for the estimator of $P[Y = 1|...]$ in combination with the constant 0 model for $E[\Pi_*|X_* = x_*, h, Y_* = 0]$ (or some other approximation). In all cases, the resulting class 1 probability forecasts will be less than or equal to those produced using Variant B2 because of the more explicit inclusion of the majority class.

B4. Develop some strategy for resampling latent states from the kernel density estimator at randomly chosen failure points and repeatedly estimating $E[\Pi_*|X_* = x_*, h, Y_* = 1]$. Repeat the strategy with non-failure points and $E[\Pi_*|X_* = x_*, h, Y_* = 0]$, and then aggregate each set of regressions separately before combining with the partition weights $P[Y_* = 1|...]$ and $P[Y_* = 0|...]$ to produce the final model.

This option extends Variant 4.5 so that the resampled latent states are used in separate regressions for class 1 and class 0. If the $P[Y_* = 0|...]$ and $P[Y_* = 1|...]$ terms are set to be the unconditional sample proportions, in terms of a re-sampling strategy with a Gaussian product kernel, this variant is identical to the ROSE rare events method [Lunardon et al., 2014a]. The key difference is that we are sampling the classifier surface implied by the Gaussian noise and then using those points to train a regression model, whereas with ROSE the sampling maintains the binary class labels and uses the data to perform classification. In any case, because of the numerous decisions required by a resampling strategy and estimating $P[Y = 1|...]$ this variant is not appealing.

As we concluded before with the first set of transform variants, we will ignore Variant B1 and 4.5 and base the decision between Variant B2 and B3 on any perceived upward bias with special consideration given to the left corner of the sample region. Variant B2 ignores this consideration completely, while those lumped under Variant B3 attempt to alleviate bias with varying degrees of input from the majority class. The implemented class 0 model in Variant B3 might then be viewed as an explicit bias correction term to the primary model for class 1. Likewise, the inclusion of additional latent states sampled at non-failure points in Variant A3 could also be viewed as a form of bias correction, but implicitly in this case as the changes are made before model estimation.

Later on in Section 4.6 we will demonstrate the efficacy of estimating a monotonically constrained regression model of the transformed latent states along with a full presentation of testing results. That presentation will include results from 4 of the variants described here: A2 (or B2), B3 with a null, constant 0 model for class 0, and A3 and B3 both including latent states at a grid of points in the regions outside of the observed minority class support. Across all variants, at a minimum we include latent states estimated at failure points in a monotonically constrained regression method. This guarantees our approach will emphasize the most important quantity of information contained in the training data—where evidence of class 1 truly exists (or where evidence of a probability of observing class 1 greater than 0 truly exists). With rare events datasets, class 1 points are more informative than class 0 points. The transformation variants presented here work in
part because they maximize the information extracted from class 1 points while minimizing the computation spent on class 0 points.

4.5 KDE Bandwidth Selection

Bandwidth selection for kernel density estimators is an open-ended problem generally, but it is easily interpreted as enforcing some desirable level of smoothness on the model surface. Depending on the dataset and motivations, that desired smoothness property may vary substantially. Fortunately, bandwidth selection has been shown to be amenable to various “rules of thumb” that estimate a bandwidth with respect to some reasonable optimization criterion chosen by the original author [Scott, 1992]. Also fortunate for our situation, is that with rare events problems bandwidth selection is constrained in practice. On the low end, you end up seeing latent states hitting the maximum boundary of 1, which is unreasonable and problematic for subsequent model estimation. Then, as bandwidths become larger, an oversmoothed state is reached where the model begins to converge to the full sample mean. Later in this subsection, we will describe our approach to bandwidth selection that makes use of both an optimized selection rule and our practical constraints.

To proceed, we first need to define the two concepts—monotonically dominating pairs of points (or “dominating pairs”) and deviances from monotonicity. Given any two points from a sample, we say the points constitute a monotonically dominating pair if for every dimension one point has a value that is greater than or equal to the corresponding value of the other point. Using notation, we would write that a dominating pair is a tuple \((x_A, x_B) : x_{Aj} \leq x_{Bj}, \forall j\). Dominating pairs are notable and useful because they allow us to empirically validate the monotonicity of an estimated model surface. Simply put, for a model to be considered monotonic non-decreasing, the estimated model output value at the dominant point must be greater than or equal to the output at the subordinated point (or \(\hat{y}_A \leq \hat{y}_B\)), for all unique dominating pairs in the sample (and conversely, \(\hat{y}_A \geq \hat{y}_B\), for a monotonic non-increasing model).

If this property does not hold for a dominating pair, then we say it constitutes a break from monotonicity, which can be measured as the difference between model outputs \((\hat{y}_A - \hat{y}_B)\). In defining formally deviances from monotonicity, we will set to 0 the difference for any dominating pair that satisfies the specified direction of monotonicity. When there is no break from monotonicity observed within a dominating pair, for our purposes we prefer our measurement to show 0 and not some differential that informs how steep the model is increasing or decreasing. Ultimately, we are not concerned with that kind of information, and are happy to let the data speak for itself through nonparametric methods. With notation, non-decreasing monotonicity deviances are written as \(\max(\hat{y}_A - \hat{y}_B, 0)\), and non-increasing monotonicity deviances are written as \(\max(\hat{y}_B - \hat{y}_A, 0)\). Note that monotonicity deviances are defined with respect to a specific direction of monotonicity, but the definition of dominating pairs is indifferent. For brevity, going forward we will only discuss non-decreasing monotonicity, as any variable that is considered monotonic non-increasing with respect to the model output can be multiplied by -1 to form a monotonic non-decreasing relationship.

To summarize the degree to which a model violates the property of monotonicity, a natural choice would be to take the mean of monotonicity deviances at all dominating pairs in the sample. Likewise, we will be using the mean of squared monotonicity deviances to exaggerate the impact of larger deviances. In pursuit of an optimal bandwidth, we will then seek to minimize this mean squared monotonicity deviance. Optimality is always a relative condition, and for us it refers to the kernel density estimator bandwidth that results in a sample of latent states where violations of monotonicity have been minimized or eliminated.
A consequence of the definition for our specific latent state estimator, as defined in Section 4.4, is that all monotonicity violations will be eliminated as all the individual bandwidth dimensions go to $+\infty$. As was noted along with the definition, this is because at larger bandwdiths our estimator converges to a constant, the sample mean, at all points throughout the sample space. In order to formalize a decision procedure for our unique bandwidth selection problem, then, we need to penalize the growth in the bandwidth vector so that we decide on some optimal bandwidth that is less than $+\infty$. In the same vein as other classic mean-variance optimization problems such as ridge regression, we chose the 2-norm of the bandwidth vector as the penalty term. Using notation, we can now write our bandwidth selection rule an optimization problem.

Let $\mathcal{M}$ be the set of all dominating pairs of minority class points, $m_0 = |\mathcal{M}|$ be the number of dominating pairs, $d$ be the number of predictors in the estimation set, $\hat{\pi} = \hat{\mathbb{P}}[Y = 1|X = \mathbf{x}, \mathbf{h}]$ be our latent state estimator defined on Line in Section 4.4, and $\lambda$ be a tuning parameter. $\hat{h}$ is our selected bandwidth that is the solution to the following:

$$
\hat{h} = \arg\min_{h \in (0, +\infty)^d} \left\{ \frac{1}{m_0} \sum_{m \in \mathcal{M}} \max(\hat{\pi}_{mA} - \hat{\pi}_{mB}, 0)^2 + \lambda \|h\|_2 \right\}
$$

(4.65)

We must emphasize that the set $\mathcal{M}$ of dominating pairs upon which our selection rule is based is derived from minority class points only. That is, we will be validating monotonicity at class 1 points only (dominating pairs where both points referenced are from the minority class). In a rare events problem, with such a large proportion of majority class points, we have little to no information about the true underlying relationship. We are resigned to a few, brief glimpses of the minority class, and as we have said before only at these locations can we conclude the class 1 probability is greater than 0. In a similar fashion, only at class 1 points would it be computationally efficient to validate monotonicity. When two minority class points form a dominating pair, we know that the classifier output must be greater than 0 at both points, and that the output must be greater at the dominant point relative to the subordinated one. No matter the resulting deviance between every dominating pair of minority class points, we gain valuable information concerning how close (or far) the estimated sample respects monotonicity. Validation of monotonicity at class 1 points also helps support our proposed rare events approximation whereby the eventual constrained regression method is trained using only latent states estimated at class 1 points. The computational burden of training the regression is reduced by not only having to process far fewer observations (minority class count instead of full sample count), but also by working with the pre-validated dataset that should require fewer iterations of the estimation algorithm used.

Alternatively, imagine the futility involved with validation between dominating pairs where the subordinated point is from the majority class, or both points in the pair are from the majority. Because of the sizable count of the latter in a large rare events datasets, it would be sensible to think that monotonically constrained classification is a waste of time and effort. The researcher must be wary as to what real value may be derived from estimating a microscopic increase in class 1 probability between two close majority points that are far from any minority class point. Regardless of the class of the dominant point in a dominating pair, when the subordinated point is from the majority, the true class 1 probability there may be 0 (due to the nature of rare events). This fact does not justify an increase in left boundary for the model output at the dominant point. Classifier output probabilities are restricted to be in $(0, 1)$, and the presence of a subordinated point from the majority class does not provide evidence in favor of shrinking that interval. From both situations, we see that exhaustively validating large rare events datasets for monotonicity by examining all dominating pairs would be tedious and ineffective.
This argument may not hold if “dominating n-tuples”, or dominating sequences of more than two points, are considered, but any such effort would be computationally intensive. Furthermore, the result of all monotonicity violations being eliminated as bandwidth dimensions go to $+\infty$ would still hold if dominating $n$-tuples were considered, meaning that the justifications for our selection rule are still sound even though we do not consider these extended relationships (although the final bandwidth value might change if we were to do so). Finally, we acknowledge we are ignoring dominating pairs where the dominant point is from the majority class but the subordinated point is from the minority. Useful information could be gained from the monotonicity deviances between these points. In developing our approach we were satisfied with the final results when using only latent states estimated at class 1 points, so we never felt the need to include non-failure points in the bandwidth selection process. Extensions to our approach that include these extra dominating pairs could be a direction of future work, especially if the pairs were to be incorporated in the kind of density based resampling that was suggested in Variants and .

As written in Equation 4.65, the solution to the optimization problem is a $d$-dimensional strictly positive bandwidth vector. In practice, this is too general because it allows bandwidth vectors that exhibit a different ordinality than that of the sample variances of the predictors. Our kernel based transformation was developed in order to holistically capture the a rare events dataset with continuous outputs, not to inject domain knowledge or adjust the relative weight of individual predictors. Ultimately, we want the noisy dataset as is, with only moderate, uniform smoothing to capture all local structure that is present. With these objectives in mind, if one predictor from the dataset has greater variance than another, we will require the relationship between bandwidths for these predictors to match as well. To adhere to this requirement and further simplify matters, we will then approximate the problem in Equation 4.65 by admitting only solutions that are some multiple of our preferred plug-in rule for Gaussian product kernels, called the “Multivariate Scott Rule” [Scott, 1992]. This selection rule minimizes the asymptotic mean integrated squared error for a density estimator constructed using a Gaussian product kernel with respect to the true density. The key input to the Multivariate Scott Rule is some variance estimator for each dimension, so any multiple of this rule will be satisfactory for our purposes.

Denote the Multivariate Scott Rule bandwidth estimator (for Gaussian product kernels) as $\hat{h}_{SR} = \hat{\sigma} n^{-1/(d+4)}$, where $\hat{\sigma}$ is the diagonal of the sample covariance matrix for the estimation set and $n$ is the observation count [Scott, 1992]. Then, let $\nu > 0$ be our base rule multiplier, and re-write $\hat{\pi}$ in terms of $\nu \hat{h}_{SR}$ as the expression $\hat{P}[Y = 1|X = x, \nu \hat{h}_{SR}]$. Our bandwidth rule shall be revised to denote $\hat{h} = \nu \hat{h}_{SR}$, the solution to the following 1-dimensional projection of the $d$-dimensional optimization problem defined earlier:

$$\hat{h} = \nu \hat{h}_{SR} : \nu = \arg \min_{\nu \in (0, \infty)} \left\{ \frac{1}{m_0} \sum_{m \in M} \max(\hat{\pi}_{m_A} - \hat{\pi}_{m_B}, 0)^2 + \lambda \left\| \nu \hat{h}_{SR} \right\|_2 \right\}$$  \hspace{1cm} (4.66)

We can simplify this selection rule even further because of practical constraints. For smaller values of the multiplier $\nu$, some of the latent state estimates will be a hard 1, which is problematic for the usual link functions (logit, probit, etc.) that will be used with any regression method that we implement following the kernel based transformation. We demonstrated in Section 4.4.1 that our latent state estimator as defined will result in values within the interval $(0, 1)$, but because of hardware limitations this is not observed in practice. With our failure dataset, $\hat{\pi} = 1$ was always observed at some input locations when $\nu \leq 1.5$. For larger values of $\nu$, we eventually arrive at a state where no monotonicity breaks are present, meaning that the resulting bandwidth would be considered oversmoothed for our purposes. With our failure dataset, this state was always reached at values of $\nu \geq 4.6$. Keeping these two observations in mind, in practice we can
use the revised bounds of \( \nu \in (1.6, 4.5) \) for the simplified optimization problem in Equation 4.66.

In Figure 4.11 we illustrate density curves for the sample of latent states that results at various base rule multipliers. This figure was created using our full failure dataset (1986-2014), but in accordance with our dataset transformation proposal latent states were estimated at failure class points only. We observe that as the multiplier increases from 1.6 (purple curve) to 5.6 (red curve) we see a less pronounced peak near 0 and an increase in total mass in the right tail that extends to 1. The multiplier values correspond to the following (in increasing order): the practical lower bound at which there are no latent states equal to 1, our final bandwidth selection which is the solution to our optimization problem when \( \lambda \approx 0.00001 \) (more detail will follow shortly), the practical upper bound at which there are no monotonicity breaks, and then a moment matching choice at which the sample variance of the estimated latent states is roughly equal to the sample variance of the full sample binary output variable. This last multiplier was included as a visual reference only. It was never considered for further testing because of poor preliminary results, but it could be thought of as a reasonable absolute maximum rule if an over-smoothed bandwidth was desired.

Figure 4.11 nicely summarizes the ramifications of increased smoothing, such that the kurtosis of the sample distribution of latent states increases at greater multipliers (or bandwidths). This effect becomes extremely significant with our rare events setting because there will be fewer extreme values present when learning the true relationship that underlies the right tail. With using greater multiples of a base bandwidth rule, there will be decreased granularity everywhere, but the strongest impact will be seen within regions with the greatest class 1 probability (right corner). Conversely, because of the sheer number of observations with low class 1 probability (all points outside the right corner), increased smoothing from greater multiples will have less of an impact on learning the relationship that underlies the peaked region near 0.

For additional clarity on the ramifications of increased smoothing, let us refer to Table 4.1, where we provide an array of validation metrics for each multiplier \( \nu \in \{1.6, 1.7, \ldots, 5.6\} \) using the full sample of our firm failure dataset. At each base rule (Multivariate Scott Rule) multiplier value shown in column “SR.mult”, we generated latent state estimates at every class 1 point in our full dataset, and then calculated the various metrics on the resulting latent state sample (in addition to displaying the 2-Norm of the bandwidth vector used in the second column). The table rows with a value in the first column of 1.6, 3.5, 4.1, and 5.6 correspond to the respectively labeled density curves in Figure 4.11.

The 3rd through 6th columns are self-explanatory: the latent state sample mean, variance, kurtosis, and maximum. The multiplier value \( \nu = 5.6 \) was used as our cutoff in Table 4.1 and Figure 4.11 because that multiplier results in a latent state sample variance that is roughly the same as the full sample variance of the original binary class label. As mentioned earlier, this could be thought of as an over-smoothed bandwidth. As an alternative for an over-smoothed cutoff point, a change in the direction of latent state sample kurtosis could be used. Note that the increase in kurtosis with greater bandwidths begins to reverse at our greatest multipliers. From then on the distribution of latent states begins to flatten out like the uniform distribution, and kurtosis begins to decrease, because there are now frequent deviations (significant mass from the mean). Qualitatively, this would be counter-productive to learning the true relationship that underlies the right tail, as the surplus density mass in the right tail would introduce upward bias.

The 7th through 9th columns of Table 4.1 include the count of monotonicity breaks, the proportion of monotonicity breaks to dominating pairs, and the mean squared monotonicity deviance for the resulting latent state sample. The proportion of breaks starts around 4.5% and then steadily decreases to 0 at a multiplier value of \( \nu = 4.1 \). Greater multiplier values are over-smoothing with respect to our optimization criteria of mean squared monotonicity deviance, and in practice we found no reasons to consider multipliers
beyond our over-smoothed rule at when equal sample variance is achieved.

For the 10th through 12th columns of Table 4.1, we are referring to the estimation of a linear regression (with probit link function) on the latent state sample. For column “coef.dist”, we are calculating the 2-Norm of the difference between resulting regression coefficients from using the estimated latent states and resulting regression coefficients from performing probit regression on the full sample. Similar to considering the sample moments, this column is included to provide some understanding of the difference between the latent state sample and the full sample. Column “coef.int” is simply the intercept coefficient that resulted from regression on the latent states. The estimated intercept from a probit regression using the full sample is roughly −2.6, and we see a similar result from our latent state regressions around a multiplier of 2.9.

Column “coef.mono” is the count of regression coefficients that are non-negative. With our modified version of the firm failure dataset, all features were engineered to be monotonic non-decreasing with respect to the binary failure indicator. We are expecting to see all 8 predictors having a non-negative coefficient at some level of smoothing, and we do at multiplier values \( \nu \geq 3.7 \). Given that the class of non-decreasing linear models is a subset of non-decreasing non-parametric models, we interpret the finding of a bandwidth multiplier threshold above which always result in a linear model with all non-negative coefficients as evidence in support of our assertion that the true firm failure relationship is in fact monotonic with respect to all 8 predictors. Additionally, we interpret the evidence as favorable towards learning some non-decreasing non-parametric relationship using some multiplier values \( \nu < 3.7 \). The superset non-decreasing non-parametric is a much larger set than non-decreasing linear, so a constrained non-parametric model should be able to properly accommodate more noise in the estimation set (i.e. less smoothing). Conversely, if there were no bandwidth multiplier threshold above which always resulted in a linear model with all non-negative coefficients, this would not necessarily rule out non-decreasing non-parametric relationships. It would just make that assumption harder to justify because of the removal of the linear subset, and thus require stronger supporting evidence.

This argument based on non-negative linear regression coefficients is captured in Figures 4.12 and 4.13. We are visualizing the progression of regression coefficients that result when training a linear model (via a probit link function) with the full failure dataset transformed using multiplier values \( \nu \in \{1.6, 40.0\} \), with the exception of the points at \( x = 0 \), which correspond to the probit regression coefficients that result from using the full dataset. Both figures use the same data and multipliers and scaled on the same \( x \) axis, only the \( y \) axis is different. In Figure 4.12 we see that as the multiplier increases the Intercept coefficient converges towards the full sample mean of 0.001075 transformed using the probit link function, or −3.069, and then all the other coefficients converge towards 0. In Figure 4.13, we confirm that with multiplier values \( \nu \geq 3.7 \), all regression coefficients are non-negative. After smoothing away some of the noise present in the dataset, we have arrived at an estimation sample that is in fact monotonic non-decreasing under a linear assumption, providing us justification for using less smoothing under a less restrictive non-parametric assumption.

As our optimization problem penalizes for increasing bandwidths, we must decide on the tuning parameter \( \lambda \) which reflects how much weight we assign to the penalty component. Naturally, we recommend the use of cross validation on the training set to select \( \lambda \). For datasets with independent observations, this should be straight-forward, but with our failure dataset comprised of firm-month observations over the years 1986-2014, predictor vectors that make extensive reference to the S&P 500 large cap equity index, and an extremely low observed class 1 proportion, we made some specific choices for our cross validation procedure. Our test set is October 2014 through December 2014 (7.25 years, or 87 months), because the S&P 500 peaked in the third quarter of 2007. With this delineation, we will be testing over a nearly full index level peak-to-peak cycle
(“nearly” because the index continued to climb in the years after the final month of our dataset). Testing over a full equity market cycle should help avoid bias in our results that could be introduced if only bear or bull market periods were considered. We have then partitioned our training set (January 1986 - September 2007) into 3 folds using the two S&P 500 level peaks that occurred within the training period (in 1990 Q2 and 2000 Q1) to minimize bias in the same way. The first fold is January 1986 through June 1990 (4.5 years, or 54 months), the second fold is July 1990 through March 2000 (9.75 years, or 117 months), and the third is April 2000 through September 2007 (7.5 years, or 90 months). To arrive at our final selection of $\lambda$ that will be used with the test set, we will take the $\lambda$ value that results in the lowest mean Area Under the Receiver Operating Characteristic curve (AUROC) statistic from 1) training with fold 1 to test fold 2 and 2) training with fold 2 to test fold 3. AUROC is chosen in favor of mean squared error because the latter is a poor choice with rare events data, where a null model that predicts 0 at all points performs well according to this statistic [Powers, 2007]. Alternatively, mean log loss can be used, which was tested with our dataset and resulted in a bandwidth that was nearly the same as when we performed cross validation using AUROC.

Figure 4.14 displays the cross validated AUROC statistics that we produced with a range of tuning parameter values $\lambda \in [0.000001, 0.1]$. The bars correspond to ranges of $\lambda$ that would approximately result in one of the multiplier values $\nu \in \{1.6, 2.0, 2.5, 3.0, 3.5, 4.1\}$ that was considered, while the points correspond to the exact matches. For solving our simplified optimization problem in Equation 4.66 with a 1-dimensional grid search over the restricted range of the base rule multiplier $\nu \in [1.6, 4.5]$, the cross validation is actually driven by selecting a discretized range of values for $\nu$. For each multiplier value, forecasts are generated and then we calculate AUROC, and repeat these steps at every fold of training data. After averaging the matched AUROC statistics from the different folds, we solve for the range of $\lambda$ that would imply the selected values of $\nu$. Our results were generated using montonic-BART as the regression method that followed our kernel based transformation, and from Figure 4.14 it is clear where we will set the tuning parameter $\lambda \approx 0.00001$ for live testing with the test set. This approximate value of $\lambda$ is supported by the multiplier value $\nu = 3.5$, which was evaluated using the full sample previously in Table 4.1. There we can see that $\nu = 3.5$ eliminates the overwhelming majority of monotonicity breaks, still results in a maximum latent state near 1.0, and sample kurtosis that is about half of the maximum achievable with the transform at $\nu = 5.3$. 
<table>
<thead>
<tr>
<th>SR.mult</th>
<th>bw2N</th>
<th>mean</th>
<th>var</th>
<th>kurt</th>
<th>max</th>
<th>mono.breaks</th>
<th>breaks.prop</th>
<th>devsq.mean</th>
<th>coef.dist</th>
<th>coef.int</th>
<th>coef.mono</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>1.5894</td>
<td>0.3853</td>
<td>0.0758</td>
<td>6.0500</td>
<td>0.0000</td>
<td>1.4566</td>
<td>0.0000</td>
<td>1.3672</td>
<td>0.0000</td>
<td>1.20</td>
<td>1.20</td>
</tr>
<tr>
<td>1.7</td>
<td>2.0075</td>
<td>0.5147</td>
<td>0.0520</td>
<td>8.8339</td>
<td>0.0000</td>
<td>5.1770</td>
<td>0.0000</td>
<td>9.3117</td>
<td>0.0000</td>
<td>1.45</td>
<td>1.45</td>
</tr>
<tr>
<td>1.8</td>
<td>2.1256</td>
<td>0.3113</td>
<td>0.0244</td>
<td>8.6115</td>
<td>0.0000</td>
<td>4.3850</td>
<td>0.0000</td>
<td>6.2102</td>
<td>0.0000</td>
<td>1.65</td>
<td>1.65</td>
</tr>
<tr>
<td>1.9</td>
<td>2.2437</td>
<td>0.1216</td>
<td>0.0053</td>
<td>10.7345</td>
<td>0.0000</td>
<td>3.5660</td>
<td>0.0000</td>
<td>4.9071</td>
<td>0.0000</td>
<td>1.81</td>
<td>1.81</td>
</tr>
<tr>
<td>2.0</td>
<td>2.3617</td>
<td>0.0297</td>
<td>0.0025</td>
<td>13.0151</td>
<td>0.0000</td>
<td>2.7510</td>
<td>0.0000</td>
<td>2.5380</td>
<td>0.0000</td>
<td>1.95</td>
<td>1.95</td>
</tr>
<tr>
<td>2.1</td>
<td>2.4798</td>
<td>0.0086</td>
<td>0.0024</td>
<td>15.7197</td>
<td>0.0000</td>
<td>2.1200</td>
<td>0.0000</td>
<td>1.5899</td>
<td>0.0000</td>
<td>2.08</td>
<td>2.08</td>
</tr>
<tr>
<td>2.2</td>
<td>2.5979</td>
<td>0.0075</td>
<td>0.0020</td>
<td>18.6224</td>
<td>0.0000</td>
<td>1.5667</td>
<td>0.0000</td>
<td>0.9939</td>
<td>0.0000</td>
<td>2.18</td>
<td>2.18</td>
</tr>
<tr>
<td>2.3</td>
<td>2.7160</td>
<td>0.0068</td>
<td>0.0018</td>
<td>21.8993</td>
<td>0.0000</td>
<td>1.1119</td>
<td>0.0000</td>
<td>6.2302</td>
<td>0.0000</td>
<td>2.28</td>
<td>2.28</td>
</tr>
<tr>
<td>2.4</td>
<td>2.8341</td>
<td>0.0061</td>
<td>0.0015</td>
<td>25.5851</td>
<td>0.0000</td>
<td>0.7856</td>
<td>0.0000</td>
<td>39.2892</td>
<td>0.0000</td>
<td>2.35</td>
<td>2.35</td>
</tr>
<tr>
<td>2.5</td>
<td>2.9522</td>
<td>0.0051</td>
<td>0.0013</td>
<td>29.6989</td>
<td>0.0000</td>
<td>0.5560</td>
<td>0.0000</td>
<td>24.7899</td>
<td>0.0000</td>
<td>2.42</td>
<td>2.42</td>
</tr>
<tr>
<td>2.6</td>
<td>3.0703</td>
<td>0.0050</td>
<td>0.0011</td>
<td>33.8437</td>
<td>0.0000</td>
<td>0.4069</td>
<td>0.0000</td>
<td>15.5452</td>
<td>0.0000</td>
<td>2.48</td>
<td>2.48</td>
</tr>
<tr>
<td>2.7</td>
<td>3.1884</td>
<td>0.0047</td>
<td>0.0009</td>
<td>38.1976</td>
<td>0.0000</td>
<td>0.2800</td>
<td>0.0000</td>
<td>9.6451</td>
<td>0.0000</td>
<td>2.53</td>
<td>2.53</td>
</tr>
<tr>
<td>2.8</td>
<td>3.3064</td>
<td>0.0044</td>
<td>0.0008</td>
<td>42.6742</td>
<td>0.0000</td>
<td>0.2044</td>
<td>0.0000</td>
<td>5.9131</td>
<td>0.0000</td>
<td>2.58</td>
<td>2.58</td>
</tr>
<tr>
<td>2.9</td>
<td>3.4245</td>
<td>0.0041</td>
<td>0.0007</td>
<td>47.3222</td>
<td>0.0000</td>
<td>0.1430</td>
<td>0.0000</td>
<td>3.6223</td>
<td>0.0000</td>
<td>2.61</td>
<td>2.61</td>
</tr>
<tr>
<td>3.0</td>
<td>3.5426</td>
<td>0.0039</td>
<td>0.0006</td>
<td>52.1617</td>
<td>0.0000</td>
<td>0.0940</td>
<td>0.0000</td>
<td>2.2519</td>
<td>0.0000</td>
<td>2.65</td>
<td>2.65</td>
</tr>
<tr>
<td>3.1</td>
<td>3.6607</td>
<td>0.0037</td>
<td>0.0005</td>
<td>57.1630</td>
<td>0.0000</td>
<td>0.0610</td>
<td>0.0000</td>
<td>1.4430</td>
<td>0.0000</td>
<td>2.68</td>
<td>2.68</td>
</tr>
<tr>
<td>3.2</td>
<td>3.7788</td>
<td>0.0035</td>
<td>0.0005</td>
<td>62.2762</td>
<td>0.0000</td>
<td>0.0460</td>
<td>0.0000</td>
<td>0.9397</td>
<td>0.0000</td>
<td>2.70</td>
<td>2.70</td>
</tr>
<tr>
<td>3.3</td>
<td>3.8969</td>
<td>0.0033</td>
<td>0.0005</td>
<td>67.4498</td>
<td>0.0000</td>
<td>0.0310</td>
<td>0.0000</td>
<td>0.6058</td>
<td>0.0000</td>
<td>2.73</td>
<td>2.73</td>
</tr>
<tr>
<td>3.4</td>
<td>4.0150</td>
<td>0.0032</td>
<td>0.0005</td>
<td>72.6266</td>
<td>0.0000</td>
<td>0.0210</td>
<td>0.0000</td>
<td>0.3832</td>
<td>0.0000</td>
<td>2.75</td>
<td>2.75</td>
</tr>
<tr>
<td>3.5</td>
<td>4.1331</td>
<td>0.0031</td>
<td>0.0005</td>
<td>77.7459</td>
<td>0.0000</td>
<td>0.0150</td>
<td>0.0000</td>
<td>0.2345</td>
<td>0.0000</td>
<td>2.77</td>
<td>2.77</td>
</tr>
<tr>
<td>3.6</td>
<td>4.2511</td>
<td>0.0029</td>
<td>0.0005</td>
<td>82.7599</td>
<td>0.0000</td>
<td>0.0110</td>
<td>0.0000</td>
<td>0.1387</td>
<td>0.0000</td>
<td>2.78</td>
<td>2.78</td>
</tr>
<tr>
<td>3.7</td>
<td>4.3692</td>
<td>0.0028</td>
<td>0.0005</td>
<td>87.6518</td>
<td>0.0000</td>
<td>0.0070</td>
<td>0.0000</td>
<td>0.0767</td>
<td>0.0000</td>
<td>2.80</td>
<td>2.80</td>
</tr>
<tr>
<td>3.8</td>
<td>4.4873</td>
<td>0.0027</td>
<td>0.0005</td>
<td>92.4392</td>
<td>0.0000</td>
<td>0.0040</td>
<td>0.0000</td>
<td>0.0359</td>
<td>0.0000</td>
<td>2.82</td>
<td>2.82</td>
</tr>
<tr>
<td>3.9</td>
<td>4.6054</td>
<td>0.0026</td>
<td>0.0005</td>
<td>97.1620</td>
<td>0.0000</td>
<td>0.0020</td>
<td>0.0000</td>
<td>0.0120</td>
<td>0.0000</td>
<td>2.83</td>
<td>2.83</td>
</tr>
<tr>
<td>4.0</td>
<td>4.7235</td>
<td>0.0025</td>
<td>0.0005</td>
<td>101.8655</td>
<td>0.0000</td>
<td>0.0010</td>
<td>0.0000</td>
<td>0.0018</td>
<td>0.0000</td>
<td>2.84</td>
<td>2.84</td>
</tr>
<tr>
<td>4.1</td>
<td>4.8416</td>
<td>0.0025</td>
<td>0.0005</td>
<td>106.5833</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0126</td>
<td>0.0000</td>
<td>2.86</td>
<td>2.86</td>
</tr>
</tbody>
</table>

\(^{1}\text{Scott Rule Multiplier } \nu. \quad ^{2}\text{2-Norm of Bandwidth Vector.} \quad ^{3}\text{Latent State Sample Mean. Full sample } \bar{y} = 0.001075. \quad ^{4}\text{Latent State Sample Variance. Full sample } s_y^2 = 0.001076 \quad ^{5}\text{Latent State Sample Kurtosis. Full sample } \bar{y}_4/s_y^4 = 9271 \quad ^{6}\text{Latent State Sample Maximum} \quad ^{7}\text{Count of Monotonicity Breaks.} \quad ^{8}\text{Proportion of Monotonicity Breaks to Monotonicity Pairs} \quad ^{9}\text{Mean of Squared Monotonicity Deviances } (\times 10^5) \quad ^{10}\text{Euclidean Distance between Linear Regression Coefficients estimated from transformed dataset and probit link and Probit Regression Coefficients estimated from full dataset} \quad ^{11}\text{Intercept Coefficient estimated from transformed dataset and probit link.} \quad ^{12}\text{For reference, full sample Probit Regression Intercept Coefficient is } -2.46. \\

Table 4.1: Bandwidth Validation Metrics Using Full Sample 1986-2014

Figure 4.11: Distribution of Latent States At Various Scott Rule Multipliers
Figure 4.12: Linear Regression Coefficients From Transformed Dataset At Various Scott Rule Multipliers

Figure 4.13: Linear Regression Coefficients From Transformed Dataset At Various Scott Rule Multipliers (Zoomed In)
For those that are critical of our bandwidth selection algorithm, direct cross validation of AUROC or log loss on the training set without any references to the mean squared monotonicity deviance, 2-Norm of the bandwidth vector, or tuning parameter would be a simpler alternative. We tested this option, and it produced a slightly smaller bandwidth and slightly worse test set results. Given the decrease in forecasting performance and that our proposed algorithm is fully parallelizable, we argue the increased complexity is worth it. In either case, it is important to note there would be no accounting of in-sample fit; only out-of-sample fit is considered. This is obvious for direct cross validation AUROC, and with our selection algorithm both quantities in the optimization objective are penalties that constrain model fit, not terms that measure model fit. The mean squared monotonicity deviance measures the property of monotonicity, similar to other penalty terms that enforce smoothness across space and time, and the 2-Norm of the bandwidth measures the magnitude of the model relationship.

MSE, AUROC, mean log loss, etc. from in-sample forecasts are not helpful for multiple reasons. One, from the rich backlog of financial literature that motivated our work, we already accept that a firm failure relationship exists with our predictors and do not need to re-establish one or reproduce the interpretations of past authors. Two, our primary use case is forecasting, so it is reasonable to let cross validation (out-of-sample results within the training set) drive model selection without a term for in-sample fit in the optimization problem. Three, in the context of rare events binary classification with temporally dependent data, model fit of the latent surface will be difficult and subjective, if not irrelevant.

Having settled on our choice of kernel bandwidth, and before we move on to combining our kernel based transform with regression in Section 4.6, let us mention the collection of scatter plots in Appendix B.2. Using \( \lambda = 0.00001 \) on the full, transformed dataset (latent states at all failure points in the full sample), we have plotted each predictor against the latent state and probit linked latent state. The plots for all predictors indicate some dependence of \( y \) on \( x \) that is monotonic in nature. It also appears that in the plots using the (raw) latent state, there is significant heteroscedasticity in the relationship such that variance in the latent state is greater at greater predictor values. This dynamic does appear muted in the plots using the probit linked latent states, however. Accommodations for heteroscedasticity could be made in the form of an adaptive kernel bandwidth, but we leave this consideration for future work.

### 4.6 Rare Events Classifier Approximation

Having completely defined our dataset transformation, we are now ready to combine it with a regression to finish our classifier approximation. Let us summarize this workflow with a brief outline.

1. Collect data and construct failure indicator, predictors, etc.
2. Using training data only, cross validate the KDE bandwidth
3. Using training data only, estimate latent states (generate KDE forecasts) at minority class points
4. Using estimated latent states only (after applying a link function like the logit, probit, etc.), train a monotonically constrained regression model
5. Generate final forecasts at test points from the regression (and after applying the link function inverse). These forecasts are in effect class 1 probabilities from an approximated monotonic classifier.
This workflow may appear complicated, but in reality we will achieve great expressive power with minimal computation because the KDE construction based on a diagonal smoothing matrix. In our experience all steps can be completed in less than 15 minutes on a multi-core server. Our primary claims of efficiency are based on two key points—KDE calculations can be parallelized and the regression will then be estimated using a greatly reduced observation count. The final step for generating forecasts is also extremely efficient because we avoid having to calculate latent states at test points (not insignificant given that there have been months with nearly 8,000 listed firms), and with many Bayesian regression methods model estimation and test set forecasting are the same operation. When using these kinds of methods that are estimated via sampling, steps 4 and 5 from our outline are combined into one.

![Tuning Parameter vs AUROC](Figure 4.14: Cross Validated AUROC Using Training Data (January 1986 - September 2007))

Our preferred choice of monotonically constrained regression method to use in combination with our kernel based transformation is an ensemble of decision trees known properly as Bayesian Additive Regression Trees (BART) [Chipman et al., 2010]. The monotonically constrained variant (MonoBART) delivers as promised, but as of this writing has not been implemented for the classification setting or enhanced with parallelized estimation algorithms for large datasets [Chipman et al., 2016]. From a practical perspective, given that our dataset transformation (by coincidence) addresses these shortcomings, the two methods could be seen as natural compliments that produce a powerful, efficient rare events classifier in combination. Also a fortunate coincidence is that BART and MonoBART are the kind of regressions just mentioned that combine estimation and forecasting into a single sampling operation, thus reducing the overall workflow required.

From a more theoretical perspective, we may develop a better understanding of why the KDE driven transformation is a perfect match for MonoBART. BART models operate on the spacings between training data points, operating as if the sample space is a multi-dimensional grid of rectangular tiles. Our dataset
transformation greatly reduces the number of points, which in turn exponentially reduces the number of gridlines read by the BART estimation algorithm allowing for massive improvements in iterations required for convergence of the sampler. Then, during estimation, the monotonicity assumption allows for meaningful interpolation between sampled latent states to fill in the gaps of missing information that may have resulted from learning a model on an approximation of the full dataset. In another light, if one considers each latent state to be a summarization of surrounding region, using our approximation with BART models could be thought of as a “Sparse BART” variation, analogous to the “inducing point” approaches to sparse Gaussian Process models [Bauer et al., 2016] [Bijl et al., 2016].

4.6.1 Approximated Classifier Demonstration

For demonstration purposes we applied our kernel based transformation to the full MME failure dataset (1986 - 2014) using the bandwidth selection we described in Section 4.5, and with the resulting approximated dataset fit a MonoBART regression with default hyperparameters and 2,000 posterior draws. While we may not be able to succinctly write the final specification for the MonoBART model, we can visualize it and compare what we see to a standard linear model estimated on the same dataset but without our approximation. We present this comparison in Figures 4.15 and 4.16 from the marginal perspective of the excess return predictor, and in Figures 4.17 and 4.18 from the marginal perspective of the capitalization predictor.

All of these figures are arranged in similar fashion. Along the $y$ axis, Figures 4.15 and 4.17 plot the logit score ($y \in (-\infty, \infty)$) and Figures 4.16 and 4.18 plot the class 1 probability ($y \in (0, 1)$). The top sub-plots labeled LogisticMME presents a logistic regression (trained without any approximations), the bottom sub-plots labeled MonoBARTKDE present a monotonically constrained BART regression (trained on our kernel driven approximation of the dataset). The colored curves are slices of the model surface at various sample quantiles, where all of the other predictors not referenced are held constant at the indicated quantile. For example, the purple curves in Figures 4.15 and 4.16 were generated such that all predictors besides $EXRETAVG$ were held constant at their respective 99th quantiles. The slices shown designate the posterior mean of the 2,000 model draws, and the test set of points that defines these slices can be thought of as some hypothetical, large, comprehensive collection of listed equities measured on 12/31/2014 ($y$ values are January 2015 forecasts). The vertical black lines denote the 10th and 90th quantiles for the predictor represented along the $x$ axis ($EXRETAVG$ or $RSIZE$).

In the top sub-plot of Figure 4.15 we clearly see linear model slices, and the non-parametric fit in the bottom sub-plot is substantially different yet still respects the monotonic constraints. Monotonicity is easily validated within the $EXRETAVG$ dimension by the fact that none of the slices exhibit a negative first derivative at any point, and between all the remaining dimensions not pictured by the fact that none of the slices intersect and reverse rank at any point (i.e. the 99th quantile slice dominates the 95th quantile slice, and so on). Unlike the linear model, all the MonoBARTKDE slices flatten out towards the edges of the support, and there is greater separation between the 99th quantile slice and the ones below it. Also, with the MonoBARTKDE we see almost no separation between the 50th quantile and 10th quantile slices. The MonoBARTKDE slices hit a maximum logit score near 5, while this output value is never possible with the linear model. The MonoBARTKDE slices do not change to the left of the black line for the 10th quantile of $EXRETAVG$, and dominate their matching slice from the linear model in this region.

In Figure 4.16 the full impact of the differences between the models just mentioned becomes apparent by viewing the class 1 probabilities. All of the MonoBARTKDE slices converge close to 1 towards the right edge
of the support, but this only happens for the 99th quantile slice of the linear model. The separation between the MonoBARTKDE slice for the 99th quantile and the lower quantiles appears substantial for most of the figure, while this is only true with the linear model slices towards the right edge. Finally, note that when viewing class 1 probabilities, there is almost no difference between the 10th quantile and 99th quantile slices with the linear model, even though this was not readily apparent when viewing logit scores.

All together, these observations demonstrate that the MonoBARTKDE approach has delivered on our expectations. Our full model has learned a more dynamic relationship in the higher risk regions (but one that is not forever increasing towards the right edge of the observed support), and a less dynamic relationship in the lower risk regions. As such, the region of strongest disagreement between the linear model and our MonoBARTKDE approach is the high risk region beyond the 90th quantile line of EXRETAVG. Conversely, below this black 90th quantile line the slices for the 50th quantile and 10th quantile are very similar, which we also interpret as evidence in favor of our approach. Little if any model detail has been lost in the extensive lower risk region of the sample space, which was largely ignored through our deliberate under-sampling (there are by definition few failure observations here, and we are estimating latent states only at failure points). If the elevated levels of the 90th, 95th, and 99th quantile slices in the low risk region are interpreted as upward bias, the variants on our approach described in Section 4.4.2 are available for fine tuning. Regardless if any variant is chosen, the entire model can be implemented efficiently because of the parallelizable nature of kernel density estimators and the significantly reduced estimation set used with the MonoBART.

Finally, we must highlight Figures 4.17 and 4.18 because in the case of the linear model, the capitalization predictor RSIZE results in a negative coefficient, which evaluates poorly. Our motivating literature strongly agrees with our own empirical findings that failure odds are non-decreasing with respect to decreasing capitalization (remember, we have flipped the orientation on our predictor) [Campbell et al., 2011]. This relationship is most likely not linear, and with our non-parametric approach we confirm the appropriate non-decreasing trend. In Figure 4.17 we see the incorrect non-increasing relationship in the top sub-plot for the linear model, and in the bottom we the non-decreasing relationship that results from our MonoBARTKDE approach. We must note that the resulting non-parametric relationship is weak (creating only a minor increase in the model surface near the 90th quantile line for RSIZE), but having better awareness of changing dynamics in high risk regions is the core motivation of our research and we view this specific case as a success. This more correct view is perhaps better appreciated when looking at Figure 4.18 where minority class probabilities are plotted along the y axis. Here, with the linear model, increasing RSIZE incorrectly results in significant decreases in the model output for the 99th quantile slice. Conversely, with MonoBARTKDE, we see a significant increase in model output for the 99th quantile slice and minor increases for the 95th and 90th quantile slices.
Figure 4.15: Model Slices at Various Sample Quantiles (Excess Return Predictor, Logit Scores)

Figure 4.16: Model Slices at Various Sample Quantiles (Excess Return Predictor, Probabilities)
Figure 4.17: Model Slices at Various Sample Quantiles (Capitalization Predictor, Logit Scores)

Figure 4.18: Model Slices at Various Sample Quantiles (Capitalization Predictor, Probabilities)
4.6.2 Approximated Classifier Evaluation

We evaluate five models, including our proposed approximated classifier MonoBARTKDE which consists of our KDE based transformation of the training set in sequence with a monotonic BART regression model. The other four models tested: TradBARTKDE, the same approximated classifier as MonoBARTKDE but using traditional BART instead of the monotonically constrained version; KernSmoothMME a simple non-parametric kernel smoother with the same Gaussian kernel and bandwidth as our dataset transformation; LogisticMME, a vanilla logistic regression; LogisticMTA, a vanilla logistic regression using an older version of the firm failure dataset suggested in the literature that does not include our modifications [Campbell et al., 2011]. The four models besides MonoBARTKDE were chosen to provide robustness checks on the top level design decisions that were made throughout our research project. Modification of the firm failure dataset is the difference between LogisticMTA and LogisticMME. Moving from linear models to non-parametric models with the application of kernels is the difference between LogisticMME and KernSmoothMME. Revising the use of kernels to be a pre-processing step for a BART model is the difference between KernSmoothMME and TradBARTKDE. The switch from unconstrained, traditional BART to monotonic BART is the difference between TradBARTKDE and MonoBARTKDE. If our decision making was in fact prudent, we should see some improvement in results as we move from evaluation of the LogisticMTA model to the MonoBARTKDE model.

For bandwidth selection with MonoBARTKDE, cross validation was performed once before the first test month with training data as described earlier in Section 4.5. The resulting tuning parameter value, $\lambda \approx 0.00001$, was then used to re-select the base rule multiplier $\nu$ at every month of test data. TradBARTKDE used the same transformed dataset that was used with MonoBARTKDE at every month. KernSmoothMME used the same bandwidth selection that was used to transform the dataset before estimating the BART models (final model forecasts from KernSmoothMME used the same bandwidth that was used for estimating latent states before BART estimation with the MonoBARTKDE and TradBARTKDE models). All BART regressions were performed using the default hyper-parameters and $2,000,000$ posterior draws.

The testing set of the firm failure dataset we used for evaluation covers July 2007 - December 2014. July 2007 was chosen as the starting point because of the observed peak in the S&P 500 index in June 2007, making our test set cover almost an entire peak to peak cycle of the index. This should help minimize any bias that could be introduced during (strictly) rising or falling markets because of the significant usage of the S&P 500 in constructing our predictor variables. For all models, at each month we re-estimate the model using all data up until the given month, and then generate forecasts for every firm. For every monthly set of forecasts, an AUROC and AUPR is calculated. Additionally, Probability Calibration is calculated using the forecasts from all test months together instead of reviewing each month forecast individually. With a median of only 4 failure events observed per month, this evaluation metric would not be meaningful on a monthly basis.

Figure 4.19 presents our test results according to monthly AUROC (greater AUROC indicates a better model). There is not much differentiation between the methods but overall KernSmoothMME appears to be the best with the highest upper quartile, median, and lower quartile. MonoBARTKDE appears competitive with a similar upper quartile and median, but a lower quartile that is below that of the linear models. The linear models are somewhat in the middle of the pack, with lower medians but higher lower quartiles compared to the other methods. The TradBARTKDE appears to be the worst across the board. All the methods appear similar in terms of lower outliers, but KernSmoothMME appears to have a single, very bad month (any outlier point with an AUROC less than 0.5 is worse than random prediction).
Figure 4.20 presents our test results according to monthly AUPR (greater AUPR indicates a better model). Unfortunately, again we do not see much differentiation between the methods and KernSmoothMME appears to be the best overall. MonoBartKDE is competitive in terms of a similar median AUPR, but oddly enough the LogisticMTA method has a better upper quartile. LogisticMME results appear very similar to MonoBARTKDE, and again TradBARTKDE appears to be the worst choice in the group.

While there is only minor variation between the methods according to AUROC and AUPR, in neither setting did MonoBARTKDE come out on top. These results could be interpreted as motivating one of the variants to our approach listed in Section 4.4.2 with the hopes of improving model performance in the low risk regions (left corner) of the sample space. There may not be many failures in the low risk regions, but there is a substantial proportion of the overall dataset. If MonoBARTKDE has upward bias in the low risk regions, then, forecasting performance as measured by AUROC and AUPR would suffer.

All is not lost, though, as we perform one last round of model evaluation using a lightly customized version of Probability Calibration. The traditional definition calls for a smoothing window defined by a fixed number of forecast points. Instead we will used a fixed set of forecast bins or intervals, so that all forecasts within an interval, no matter how many, will be aggregated into a summary statistic. As it represents the most interesting, challenging, and rewarding part of the classification problem, we are most concerned with the right tail of the forecast distribution. To conduct our evaluation accordingly, we allow for extra visual detail in the tail, which we roughly consider to be all forecasts greater than 1% (less than 10% of test set forecasts). We settled on 1% intervals for forecasts less than 10%, and 10% intervals for forecasts greater than 10%. The jump to intervals of 10% at forecasts above 10% was necessitated because of how few points available at these levels of risk (the rarest of rare events).

Figures 4.21 through 4.24 present our model evaluation using our interpretation of Probability Calibration for all of the various methods except for TradBARTKDE. All four figures are arranged in identical fashion. The summarized intervals are depicted as horizontal eye beams and a point. The width of the eye beam denotes the range of the forecast interval being summarized. All forecasted probabilities within that range are included in (only) that summary. The point denotes the mean of all forecasts within the range denoted by the interval, and height (y value) of every eye beam denotes the observed failure proportion among all forecasts included in the interval. The blue line on the main diagonal represents where you would see every interval with a perfectly calibrated model. An ideal model would have all intervals straddling the blue line, and the point on the line exactly. With an ideal model, for example, one could be very confident that given a forecast probability of say 0.5, about half of all firm months with this level of risk will result in a failure event.

Figure 4.21 contains a Probability Calibration plot for the LogisticMTA model. The first handful of intervals line up along the blue diagonal well, but beyond we see intervals off the blue line in the intervals within (0.06, 0.1] and the upper most (0.2, 0.3], (0.3, 0.4], and (0.4, 0.5] intervals. In Figure 4.22 we have the calibration for the LogisticMME model. Even though this is also a summary of logistic regression results, they are noticeably different because our revised MME dataset was used. Forecasts are now present in the (0.5, 0.6], (0.6, 0.7], and (0.8, 0.9] intervals. The LogisticMME model may then be considered more expressive than the LogisticMTA model, but we are still seeing most of the intervals above (0.4, 0.05] below the blue diagonal line (with a small exception for the (0.09, 0.1] interval). In both cases, the linear model appears to be too aggressive in producing high forecasts.

Figure 4.23 contains the Probability Calibration plot for the KernSmoothMME model. We see the (0.01, 0.02] and (0.02, 0.03] are almost calibrated properly, but the intervals above and below are off signif-
Finally, we turn our attention to Figure 4.24 with the calibration for the MonoBARTKDE model. The (0.5,0.6] and (0.9,1.0] intervals appear problematic, and this is also true to a lesser extent with the (0.08,0.09] and (0.09,0.1] intervals. On the positive side we see some encouraging signs in the (0.1,0.2], (0.2,0.3], (0.3,0.4], (0.4,0.5], (0.6,0.7], and (0.7,0.8] intervals. These 6 intervals are not quite on the blue diagonal, but all together this evidence may suggest the MonoBARTKDE model is better calibrated than the others. An extended test set would help discern the asymptotic calibration of these models at the higher risk intervals. With the (0.5,0.6] interval produced by our MonoBARTKDE model, for example, perhaps this appears poorly calibrated because there weren’t any observations at this level of risk in the test set, and not because of any deficiency with the model.
Figure 4.19: Test Set Results, Various Methods (July 2007 - December 2014, Monthly Re-estimation)

Figure 4.20: Test Set Results, Various Methods (July 2007 - December 2014, Monthly Re-estimation)
Figure 4.21: Probability Calibration, LogisticMTA (July 2007 - December 2014, Monthly Re-estimation)

Figure 4.22: Probability Calibration, LogisticMME (July 2007 - December 2014, Monthly Re-estimation)
Figure 4.23: Probability Calibration, KernSmoothMME (July 2007 - December 2014, Monthly Re-estimation)

Figure 4.24: Probability Calibration, MonoBARTKDE (July 2007 - December 2014, Monthly Re-estimation)
4.7 Closing Remarks

Rare events classification with large datasets is an extremely difficult machine learning problem, but in the presence of low dimensionality and monotonicity kernel density estimators can form the basis for an efficient approximation that is analogous to a principal components transform. Our approximation quickly transforms the classification problem to a regression problem with a greatly reduced observation count. We demonstrated how the approximation can be used with monotonically constrained BART models, resulting in a sophisticated, monotonic classifier. This pairing of methods is especially unique because currently there is no implementation of monotonically constrained BART for classification. We mean to say that our approximation has opened the door for using MonoBART for classification, at all, not just speed up or make more practical using it with a large dataset. The more typical case of improvements in execution speed also apply to using our proposed approximation with traditional BART. The traditional (unconstrained) BART implementation that offers classification is far too slow to handle a large dataset like the firm failure dataset we used for testing in this paper. Stand alone traditional BART classification, even with options enabled for parallelized estimation, was only partially complete after 8 hours, but the TradBARTKDE model we tested (that uses BART regression) completed estimation and forecasting in under 10 minutes for most test months.

In terms of next steps, some of the variants discussed in Section 4.4.2 are worthy of future consideration and should improve the forecasting accuracy in regions of low risk regions. The other most immediate topic of future work would be some adaptive bandwidth selection to better account for heteroscedasticity. We anticipate this kind of model extension would go further towards improving the probability calibration of MonoBARTKDE. More extensive model evaluation, with more recent data from 2014 - 2017 and more competing methods would also be a worthwhile pursuit. The final test results from MonoBARTKDE are admittedly underwhelming, but we counter that our method could be more useful in an investment management setting where probability calibration is considered most important. More generally, in combination with other methods in a large multi-class ensemble, MonoBARTKDE would be a worthwhile addition because of its unique awareness in the right tail of the forecast distribution.
Bibliography


[Dir, 2016] (2016). Direxion Funds Leveraged & Inverse ETFs.


Appendix A

Chapter 3 Appendix

A.1 Response Variable Reference

A.1.1 Specifications

Here we highlight relevant passages from Campbell and provide our commentary. Unless otherwise noted, all block quotes come from that paper.

Failure events

We define failure to be the first of the following events: chapter 7 or chapter 11 bankruptcy filing, de-listing due to performance related reasons, and a default or selective default rating by a rating agency.

Ch. 7 bankruptcy, Ch. 11 bankruptcy, D credit rating, SD credit rating (“selective default” rating), and delisting events with code 552 are considered failure events. The description for delisting code 552 is “Delisted by current exchange - price fell below acceptable level” [CRS, 2014]. The set of delisting codes may need to be expanded depending on how we interpret “delisting due to performance related reasons”. For now, we proceed while using only delisting events with code 552 for the simple reason that including others introduces significantly more failure observations for the years 1963 - 2008 in our model implementation, relative to the figures displayed in Campbell.

In the 2008 paper also written by Campbell, Hilscher, and Szilagyi, “In Search of Distress Risk”, which effectively covers the same model as the one presented in the 2011 paper, there were some additional comments regarding the scope of the bankruptcy dataset and which delisting codes were deemed relevant [Campbell et al., 2008].

The bankruptcy indicator we use is taken from Chava and Jarrow (2004); it includes all bankruptcy filings in the Wall Street Journal Index, the SDC database, SEC filings, and the CCH Capital Changes Reporter.

... failure indicator, which equals one if a firm files for bankruptcy, is delisted for financial reasons, or receives a D rating, ...
Typical financial reasons to delist a stock include failures to maintain minimum market capitalization or stock price, file financial statements, or pay exchange fees. Nonfinancial reasons to delist a stock include mergers and minor delays in filing financial statements.

It may be fair to conclude that Campbell, et al. reconsidered the delisting codes they included as a failure event in the 2011 version of the model compared to the 2008 version. In the 2008 paper they only mention D ratings for defaults, but in the 2011 paper they specifically mention D or SD rating codes. The difference is minor, but it is very clear they have enumerated different rating codes, so it seems possible they have also enumerated different delisting codes. In the 2011 paper, they use the phrase “performance related reasons”, but in the 2008 paper they write “financial reasons” and provide an additional footnote. As for bankruptcies, it is clear that both papers only consider Ch. 7 and Ch. 11 filings.

Conditionality of the extended horizon models

Treatment of firms that experience M&A activity and treatment of missing data

At first reading, the interpretation of the models with extended horizons (12 months and 36 months) was confusing. The confusion was caused by the following quote that states these models are for the event of failure after 12 months (or 36 months), given no failure up until that point.

Since investors will care not only about modeling financial distress over the next month but will also be interested in the determinants of failure in the future, we consider different prediction horizons. We estimate the probability of failure 12 months in the future, given that the firm has not failed over the next 12 months and we do the same for 36 months.

As was the case with interpreting the delisting event codes, we found some relevant quotes in the 2008 paper, “In Search of Distress Risk” [Campbell et al., 2008]. The use of conditioning the extended horizon models on the event of no failure up until the indicated forecast month is intentional and is motivated by their treatment of M&A activity as missing data. As explained earlier, all of the time indexes have been revised to reflect our notation where we write forecast horizons as \( t + 1 \), \( t + 12 \), and \( t + j \) (Campbell, et al. used \( t \), \( t + 12 \), and \( t + j \)).

In particular, we assume that the probability of bankruptcy in \( j \) months, conditional on survival in the data set for \( j - 1 \) months, is given by

\[
P(Y_{i,t+j} = 1|Y_{i,t+j-1} = 0, X_{i,t}) = \frac{1}{1 + \exp(-\alpha - \beta X_{i,t})} \tag{A.1}
\]

Variation in the parameters with the horizon \( j \), and exit from the data set through mergers and acquisitions, only make this problem worse.

Also, many firms exit the data set for other reasons between dates \( t \) and \( t + j \).
model in a live setting. When you generate live, out-of-sample forecasts for the extended horizon models at a specific month, it is impossible to know whether or not the firms being forecast do not fail up through the extended horizon. In our work towards an improved model, we might want to take this problem as a motivation to consider an unconditional model of failure over extended horizons.

The problem Campbell, et al. refer to is that of computing cumulative probabilities of failure using their model (only binary classification allowed, failure/not failure) given that firms can exit the dataset through M&A and other kinds of restructuring, or because the data is simply missing; none of which constitute failure or not failure. They are effectively saying that they consider data changes due to M&A as missing data, and they estimate all of their models (1 month, 12 month, etc.) using only data where they concretely know the firm failed or did not fail (i.e. there was no failure event observed AND valid predictor observations up until the forecast horizon). This choice should mean that forecasts produced from their model are biased upward (biased towards more failures), because based on the data the authors included in the final estimation, failures represented a larger portion of the dataset (relative to including observations of predictor variables with unknown failure outcomes as 0s).

Without considering expanding the scope of the model, the other alternative of including observations of predictor variables with unknown failure outcomes as 0s (whether because of missing data or M&A), should bias the predictions downwards (biased towards fewer failures, so that firms generally appear less risky). This is not ideal either, because some distressed companies that would otherwise fail end up being bought or restructured. Just because these firms did not result in a failure event does not mean they were not at risk of failure, so in this light the authors choice is reasonable.

A.1.2 Implied Specifications and Unresolved Issues

Repeat Failures

Is there a time cutoff for recognizing failure of restructured firms? If a firm fails in a given month, and restructures, and goes on to experience another failure event, is this considered a separate firm (left in the dataset) or do we remove it?

Without any guidance within Campbell, this issue is then decided implicitly by the ID conventions underlying CRSP and COMPUSTAT (the data sources for the predictor variables). If a firm fails, restructures, and then receives a new identifier in both of the primary data sources, it will re-enter the model and allow for a subsequent failure event to occur. This must happen in both CRSP and COMPUSTAT because of the linking table used to join data from the two data sources. If only one source issues a new identifier for a restructured firm, the fact that the other identifier (which is associated with a past failure event) remains prevents the firm from re-entering the dataset used for estimation.

A.1.3 Definitions

Firm Failure $Y_{i,t+1} \in \{0,1\}$

Firm Failure, One Year Forward $Y_{i,t+12} \in \{0,1\}$

Firm Failure, Three Years Forward $Y_{i,t+36} \in \{0,1\}$

Binary indicator of failure for firm $i$ in month $t + j$ for some horizon $j$. If the firm fails during the specified month, the indicator is set to 1, otherwise 0.
Using a logistic regression model, we write the probability of the firm failing at 1 month and 13 months forward as:

\[ P(Y_{i,t+1} = 1 | X_{i,t}) = \frac{1}{1 + \exp(-\alpha - \beta X_{i,t})} \]  
(A.2)

\[ P(Y_{i,t+12} = 1 | Y_{i,t+11} = 0, X_{i,t}) = \frac{1}{1 + \exp(-\alpha - \beta_{12} X_{i,t})} \]  
(A.3)

where \( X_{i,t} \) is the vector of predictor variables for firm \( i \) available as of the close of month \( t \), and \( \beta, \beta_{12} \) are the vectors of regression coefficients for the indicated horizon. Campbell also presents a model of the same form with a 36 month horizon, but for now we will focus on 1 and 12 months forward. When we say predictor variables are available as of month \( t \), this accounts for an appropriate lag for variables taken from quarterly financials. This will be covered in more detail in the section for independent variables.

As noted in Specification A.1.1, for the longer term forecasts it must also be true that there was no failure in the months between the month of the predictor observations and the forecast horizon. For this reason the model of \( Y_{i,t+j} \) is written such that it is conditional on the event \( Y_{i,t+j-1} = 0 \). Campbell et. al implicitly assume that \( Y_{i,t} = 0 \Rightarrow Y_{i,k} = 0, \forall k < t \), so this conditionality is sufficient.

Also, by nature of Specification A.1.1, Campbell, et al. are also assuming that \( X_{i,t+j-1} \) has been observed (not missing) when they condition on \( Y_{i,t+j-1} = 0 \). The absence of a failure event is not the only information assumed, but additionally it must be true that the firm has not left the predictors dataset. A firm may leave the dataset for a variety of reasons that do not constitute a failure event (alternative bankruptcy chapters or delisting codes, restructuring, M&A, or simply missing).

This is all to say that Campbell, et al. remove these extraneous outcomes from the dataset when constructing their model of firm failure with forecast horizons beyond 1 month. For the sake of simplicity the authors chose to ignore these scenarios and focus only on firms that explicitly fail or do not fail with respect to their stated collection of failure events.

A.1.4 Data Sources

In Campbell, all failure event data is sourced from Kamakura Risk Information Services (KRIS), a proprietary dataset maintained by Kamakura Corporation [Campbell et al., 2011]. Campbell’s co-author, Jens Hilscher, is a senior research fellow with Kamakura. We will assume that this dataset provides comprehensive coverage of all U.S. equity exchange delisting events, U.S. corporate bankruptcy filings, and North American corporate default rating announcements for the sample period, 1963 - 2008, and our ultimate goal is to reproduce it using the following sources.

Bankruptcy events from Dr. Sudheer Chava (Georgia Tech) [Cha, 2014]

- stated years available: 1964 - 2014
- earliest record date: 1964.03.05
- latest record date: 2014.12.31
- firm ID field: permno
- records: 2991
Source A.1.4 appears comprehensive for the stated years, but were not given any measure of population coverage. Dr. Chava co-authored a paper on bankruptcy modeling with Robert A. Jarrow in 2004, who is a principal of the firm that provided the proprietary failure event dataset to Campbell, et al [Chava and Jarrow, 2004]. That working relationship may imply that we are using the same bankruptcy dataset. If/when Dr. Chava updates this data for 2015, we could update the entirety of our results as well. All other data sources used in our implementation have already been updated for 2015 except this one. This source does not include the year 1963, but in practice this is irrelevant because of limitations of the data sources used to construct the predictor variables. The earliest failure event when we have observed predictors in the final, 1 month horizon model occurs in 1972 (and judging by Campbell’s visuals, this was true with their implementation as well). Because of this fact we will ignore the lack of 1963 bankruptcies when we summarize the coverage of our combined dataset later on.

**Exchange delisting events from WRDS, CRSP, Stock/Events [CRS, 2014]**

- stated years available: 1925-DEC - 2014
- earliest record date: 1926.02.24
- latest record date: 2014.12.31
- firm ID field: permno
- records: 30687
- file: CRSP_delisting_192512_2014.csv.gz (0.683MB)
- content: 1 record per U.S. equity exchange delisting event

Source A.1.4 covers the NYSE, NASDAQ, AMEX exchanges, all delisting reason codes, and supporting details. It is comprehensive for all years. The delisting event codes used in our model are covered in Sections A.1.1 and A.1.5.

**Firm credit rating announcements from WRDS, COMPUSTAT, RatingsXpress, Entity Ratings [Com, 2012]**

- stated years available: 1923 - 2012-JUL
- earliest record date: 1923.05.18
- latest record date: 2012.08.06
- firm ID field: gvkey
- records: 335855
- file: COMP_entityrating_1923_2012.csv.gz (4.9MB)
- content: 1 record per S&P firm-level credit rating, outlook, or watch announcement
All rating codes (not just defaults) and supporting detail are included. Foreign currency and short term credit ratings are included. No default ratings exist before 1986 (and only partial coverage of 2012-AUG). This issue has been reviewed with multiple support personnel at COMPUSTAT and S&P, and no satisfactory response has been provided. Given the structure of the Moody’s rating scale (similar, but no "D" rating for a default, which are announced separately from ratings) and the extensive usage of all other ratings codes in S&P ratings coverage for years before 1986, it is possible that S&P used to adhere to the same default treatment as Moody’s. Before 1986, when a firm defaulted, perhaps S&P announced the default outside of any ratings downgrade? Then, starting in 1986, it seems as if S&P created the D rating code to simplify the default announcement process.

Monthly firm credit ratings from WRDS, COMPUSTAT, North America, Ratings [Com, 2014]
- stated years available: 1973 - 2014
- earliest record date: 1978.11.30
- latest record date: 2014.12.31
- firm ID field: gvkey
- records: 2905928
- content: 1 record per month end date and firm with an active S&P firm-level credit rating, outlook, or watch

All rating codes and supporting detail are included. Foreign currency and short term credit ratings are included. This source is almost identical to Source A.1.4, but arranged in a firm-month as opposed to dated announcement format. With this data, the exact day of the month when a rating change was announced is unknown to us, but because the firm failure model is monthly and not daily, this loss of precision is not a problem. Given this arrangement, it is trivial to generate “default events” by finding the earliest months for all D and SD ratings present (with respect to each firm that ever received such a rating). The note from Source A.1.4 about missing default ratings prior to 1986 applies here as well.

Issue credit rating announcements from WRDS, Mergent FISD, Bond Ratings [Mer, 2014]
- years: 1950 - 2014-OCT
- earliest record date: 1950.03.08
- latest record date: 2014.10.31
- firm ID field: cusip
- records: 2389727
- file: FISD_bondrating_1950_201410.csv.gz (21.3MB)
- content: 1 record per S&P/Moody’s/Fitch/Duff issue-level credit rating announcement
Source A.1.4 covers credit rating announcements on U.S. publicly offered bonds by Fitch IBCA, Moody’s, S&P and Duff & Phelps. All rating codes and supporting detail are included. Coverage appears comprehensive for the stated years (with respect to all relevant ratings announced by the 4 agencies). This ratings data does not cover bonds that were privately placed, bank debt, or any other component of what might be incorporated into a firm-level credit rating other than publicly offered bonds.

In union, sources A.1.4 - A.1.4 come very close to matching the proprietary dataset used in Campbell, which we assumed to be comprehensive. It is reasonable to claim our sources for delisting events and bankruptcy filings are comprehensive, leaving only defaults as an area of weakness. For the full sample period, we have publicly offered bond ratings from S&P, Moody’s, Fitch, and Duff & Phelps, and after 1986 we also have firm-level ratings from S&P. That makes ratings on corporate debt that are not publicly offered bonds (i.e. bank debt, private placements, etc.) for 1963 - 1985 as the only sub-population without any representation in our failure sample data. Given our reading of a follow-up paper by Campbell’s co-author and a separate work about ratings disagreements between agencies, we are not concerned about potential gaps in our default rating coverage [Hilscher and Wilson, 2015, Akins, 2015]. The former paper uses 1986 as the starting point in their dataset (no doubt a reference to COMPUSTAT coverage), and the latter indicates that there can be disagreements between ratings agencies, but most often S&P is the first to declare default (they have the most aggressive criteria).

**Firm federal bailout announcements from ProPublica Bailout Tracker [Kiel, 2016]**

- years: 2008-OCT - 2014
- earliest record date: 2008-OCT
- latest record date: 2009-MAY
- firm ID field: (none) firm names and addresses were joined to COMPUSTAT gvkeys and CRSP permnos
- records: ?
- file: bailout-events.csv (1.5MB)
- content: 1 record per government bailout transaction, transfer, or payment (key is govt program, event category, firm name, date)

Source A.1.4 is motivated by a 2015 paper by John Campbell’s co-author, Jens Hilscher, where the firm failure model is re-visited [Hilscher and Wilson, 2015]. In that paper, notably published after the 2008-2009 recession, the act of receiving any government bailout funding is considered a firm failure event in addition to the delistings, bankruptcies, and defaults that were outlined earlier. The ProPublica Bailout Tracker website contains comprehensive coverage of these events, which was purchased and delicately incorporated to our existing recreation of Campbell’s failure dataset. The addition of this data source is covered in greater detail in Section 3.3.

**(OPTIONAL) Firm default events from Moody’s data service [Moo, 2014]**

**(OPTIONAL) Issue default events from Altman/NYU default database [Alt, 2014]**

Sources A.1.4, and A.1.4 have not yet been acquired, but represent logical future additions to improve the depth and breadth of our dataset. They are alternate sources of default events, not necessarily default
ratings. Both sources record default events without regard to the firm-level or issue-level credit ratings before or after the event of default. In combination with Sources A.1.4, A.1.4, and A.1.4, these potential additions could produce a more comprehensive set of defaults.

The Moodys default events could be partially scraped from Moodys freely available data service in their annual default reports [Vazza and Kraemer, 2015, Ou et al., 2015]. All default events are listed by date and company name in long form credit market reports that include extensive supporting detail. There is one pdf report each year from 1994 - 2014. Our existing dataset is only lacking in default event coverage for years before 1986 (when default ratings begin to appear in the Compustat/S&P ratings), meaning the added effort to collect the Moody’s data is probably not worth our time. Meanwhile, the Altman/NYU default database covers all years starting with 1971. That database contains roughly 3,400 observations on defaulted bond issues and a one-off export of the data could be purchased at the academic rate of $350. If we were to consider purchasing this data, we would want to preview what is available to determine if it would truly expand our default event coverage for years before 1986.

A.1.5 Mappings

\( Y_{i,t} = 1 \) if any of the following hold:

1. Firm \( i \) has any bankruptcy event dated month \( t \) in source A.1.4
2. Firm \( i \) has a delisting event coded 552 dated month \( t \) in source A.1.4
3. Firm \( i \) has a credit rating code D or SD (for local currency long term or short term rating) dated month \( t \) in source A.1.4 from 1986-JAN to 2012-JUL
4. Firm \( i \) has a credit rating code D or SD (for local currency long term or short term rating) dated month \( t \) in source A.1.4 from 2012-AUG to 2014
5. Firm \( i \) has a bond with credit rating code D, DD, or DDD from S&P, Fitch, or Duff&Phelps, dated month \( t \) in source A.1.4 from 1963 to 2014-OCT
6. Firm \( i \) has a bailout transaction coded as a “Purchase” or “Loan” dated month \( t \) in source A.1.4

Delisting code 552 is “Delisted by current exchange - price fell below acceptable level”, and is included because it satisfies the quoted specifications from Campbell [Campbell et al., 2011]. We may decide to test the inclusion of other delisting codes to match the 2008 version of the failure model by Campbell, et al., which seems to have an expanded definition of failure with respect to exchange delisting. When producing these preliminary results, however, including other delisting codes beyond 552 resulted in a significantly greater failure event count than what was presented in the 2011 paper (our main reference).

All of the DD and DDD rating codes present in Source A.1.4 show Fitch or Duff&Phelps as the issuing ratings agency. We reviewed these ratings and found that they almost always coincided with a D rating from S&P on the same date for the same issue, so it is safe to assume these ratings indicate a default. For example, Columbia Energy Group has multiple bonds with S&P rating code D on July 31, 1991, and for every bond we also see a Fitch rating code DDD on the same date. Moody’s credit ratings are also present in Source A.1.4, but because their system does not necessarily change the current rating in the event of a default announcement, we do not consider any specific rating code from Moody’s as indication of a default event [Moo, 2014]. This point was mentioned briefly in the notes for Source A.1.4. Moody’s rating system
does not have any rating code that is analogous to a D, SD, DD, or DDD rating code from the other ratings agencies. Moody’s announces default events separately from ratings changes, and their descriptions for rating codes CCC, CC, and C allude to the fact that a firm could be in default while rated at of any these codes [Moo, 2014].

\[ Y_{i,t} = 0 \] for all other firm-months included in the logistic regression. Inclusion in the regression is an important consideration. In Section A.1.3 we highlighted the fact that all months after a failure month for a single firm are removed, and in Section A.1.1 we elaborated on the fact that for horizons beyond 1 month, the model is conditioned on firms not failing up through the month before the forecast horizon. In practice, for extended horizons, these considerations are identical. When removing all monthly observations for a single firm after a failure month, pairs of predictor and failure indicator observations where there was a failure event between the month of the predictors and the forecast horizon month of the failure indicator will have been removed.

### A.2 Predictor Variables Reference

#### A.2.1 Specifications

Here we highlight relevant passages from Campbell and provide our commentary. Unless otherwise noted, all block quotes come from that paper.

**Observation matrix**

*We use monthly failure event data that runs from January 1963 to December 2008.*

*We construct the following eight measures of financial distress, three accounting-based measures and five market-based measures. We construct all our measures using quarterly and annual accounting data from COMPSTAT and daily and monthly data from CRSP.*

The final observation matrix will be arranged so that each row represents a unique firm-month, and there will be 8 columns for the 8 variables. The 3 accounting-based measures will be sourced from quarterly and annual COMPSTAT data. The 5 market-based measures will be sourced from daily and monthly CRSP data.

**Lagged observations**

*All of our measures are lagged so that they are observable at the beginning of the month over which we measure whether or not the firm fails. The three accounting measures are based on quarterly data and we assume that it is available two months after the end of the accounting quarter. Market data is measured at the end of the previous month.*

*For example, we use data up to December 1990 to estimate the coefficients on the eight variables in our model, calculate failure probabilities, and then sort stocks into portfolios in January 1991.*

For the base model with a 1 month horizon, we will include predictor observations dated December 1962 to November 2008, which are joined to the failure indicator at 1 month forward (January 1963 to December 2008, respectively). The month end date indicated in the key of each firm-month observation
dictates the availability of source data for calculating the variables. Market data (CRSP) that is dated up through the stated month end of the firm-month key is available. Accounting data (COMPUSTAT) that is dated up through the most recent quarter end, which must be at least 2 months prior, is available. To help demonstrate, consider firm-month observations for the following months (from the perspective of month end, for a firm that uses standard March 31, June 30, September 30, December 31 quarter ends):

- **2009-OCT**: data considered observable as of 2009-10-31 are used to construct predictors of failure occurring in 2009-NOV. The most recent quarter available for constructing the accounting measures is 2009-Q2, because as of 2009-10-31, the most recent quarter end date we will acknowledge is on or before 2009-08-31.

- **2009-NOV**: data considered observable as of 2009-11-30 are used to construct predictors of failure occurring in 2009-DEC. The most recent quarter available for constructing the accounting measures is 2009-Q3, because as of 2009-11-30, the most recent quarter end date we will acknowledge is on or before 2009-09-30.

- **2009-DEC**: data considered observable as of 2009-12-31 are used to construct predictors of failure occurring in 2010-JAN. The most recent quarter available for constructing the accounting measures is 2009-Q3, because as of 2009-12-31, the most recent quarter end date we will acknowledge is on or before 2009-10-31.

**Censoring outliers**

To control for outliers ... we replace any value below the 5th percentile with the 5th percentile value and replace values above the 95th percentile with the 95th percentile value.

Each of the seven explanatory variables is winsorized using a 5/95 percentile interval in order to eliminate outliers.

After all variables have been calculated from raw inputs, replace outliers accordingly before model fitting. It is clear we do not winsorize the independent variable \( PRICE \) because of the second comment (taken from the appendix, and we know there are 8 variables), and the way that variable is defined.

**Missing values**

For a firm-month observation to be included in the estimation sample we must observe leverage (TLMTA), profitability (NIMTA), excess return (EXRET), and market capitalization (RSIZE). We do not require a valid measure of SIGMA and replace it with its cross-sectional mean when this variable is missing. We use a similar procedure for missing lags of (NIMTA) and (EXRET) in constructing the weighted average measures (NIMTAAVG) and (EXRETAVG). We also replace missing values of cash (CASHMTA) and market-to-book (MB) with the respective cross-sectional means.

To be included in the final sample, we need to observe 5 variables: TLMTA, NIMTA, EXRET, RSIZE, and PRICE. Missing values of SIGMA, CASHMTA, and MB are allowed and plugged with their cross-sectional means. Also, missing lags for calculating the average measures NIMTAAVG and EXRETAVG are allowed and plugged with a “similar procedure” with respect to using the cross-sectional
mean. Campbell’s co-author, Jens Hilscher, was asked about this in an email and he responded, "Basically we either replace with lagged values (when available) or with cross sectional means.” We interpret this as instructions to look back one lag for a value, and if that lag is missing (there are two consecutive lags with missing values), use the cross-sectional mean.

A.2.2 Implied Specifications and Unresolved Issues

Order of operations during dataset preparation

Do we winsorize or plug missing values first?

By using the sample summary statistics in Table 1 and regression coefficients in Table 2 of Campbell as a benchmark, it was clear that the proper order is to winsorize all variables first. After this step, the sample summary statistics are calculated (only using observations without missing values). Then, missing values of $SIGMA$, $CASHMTA$, and $MB$ are plugged, missing lags of $NIMTA$ and $EXRET$ are plugged, and average measures $NIMTAAVG$ and $EXRETA AVG$ are calculated before proceeding with estimating the model and presenting the resulting coefficients.

Sample Thresholds and Exclusions

Is there a market cap threshold or any exclusions for sectors, industries, etc.?

There are no such restrictions mentioned in Campbell and we proceed under the assumption there are none. We may want to consider some restrictions in our implementation, and/or revise the usage of winsorized variables and plugged values.

A.2.3 Definitions

For all formulas, subscript $i$ refers to the firm and subscript $t$ refers to the month.

Profitability $NIMTAAVG_{i,t}$

We measure profitability as the ratio of net income (losses) over the previous quarter to the market value of total assets ($NIMTA$). ... the market value of total assets [is] the sum of book value of total liabilities and market equity ...

$$NIMTA_{i,t} = \frac{\text{NetIncome}_{i,t}}{\text{MarketEquity}_{i,t} + \text{BookLiabilities}_{i,t}} \quad (A.4)$$

We construct a measure of average profitability over the previous four quarters ($NIMTAAVG$). Since losses over the most recent quarter will be more informative than losses four quarters ago, we place more weight on more recent observations. Thus ($NIMTAAVG$) is a geometrically weighted average level of profitability where the weight is halved each quarter.

$$NIMTAAVG_{i,t} = \frac{1 - \phi^3}{1 - \phi^{12}} (NIMTA_{i,t} + \phi^3 NIMTA_{i,t-3} + \phi^6 NIMTA_{i,t-6} + \phi^9 NIMTA_{i,t-9}) \quad (A.5)$$

$$\phi = 2^{-\frac{1}{2}} \quad (A.6)$$

That specific value of $\phi$ implies that the weight is halved each quarter.
Leverage $TLMTA_{i,t}$

Our measure of leverage is total liabilities divided by market total assets ($TLMTA$).

$$TLMTA_{i,t} = \frac{\text{BookLiabilities}_{i,t}}{\text{MarketEquity}_{i,t} + \text{BookLiabilities}_{i,t}} \quad (A.7)$$

Liquidity $CASHMTA_{i,t}$

We measure short-term liquidity using cash holdings scaled by market total assets ($CASHMTA$).

$$CASHMTA_{i,t} = \frac{\text{CashAndShortTermInvestments}_{i,t}}{\text{MarketEquity}_{i,t} + \text{BookLiabilities}_{i,t}} \quad (A.8)$$

Excess Return $EXRETAvg_{i,t}$

We add the firm’s equity return ($EXRET$) which is the stock’s excess return relative to the S&P 500 index return.

$$EXRET_{i,t} = \log(1 + R_{i,t}) - \log(1 + R_{S&P500,t}) \quad (A.9)$$

where $R_{i,t}$ is the monthly stock return for firm $i$ in month $t$ and $R_{S&P500,t}$ is the monthly return on the S&P500 index.

In a similar spirit [to (NIMTAAVG)] we also construct a measure of average returns over the last 12 months ($EXRETAvg$) which also places relatively more weight on more recent returns.

$$EXRETAvg_{i,t} = \frac{1 - \phi}{1 - \phi^{12}} (EXRET_{i,t} + \phi^{11}EXRET_{i,t-1} + ... + \phi^{11}EXRET_{i,t-11}) \quad (A.10)$$

$$\phi = 2^{-\frac{4}{3}} \quad (A.11)$$

That specific value of $\phi$ implies that the weight is halved each quarter, not that the monthly weights are halved every subsequent month. The weights still decay by half every quarter (when summed). This translates to meaning the sum of the weights of the second set of months (subscripts $t-3,t-4,t-5$) is equal to half the sum of the weights of the first set of months (subscripts $t,t-1,t-2$).

Volatility $SIGMA_{i,t}$

Volatility ($SIGMA$) is a measure of the stock’s standard deviation over the previous three months.

Our measure of equity return volatility is the annualized 3-month return standard deviation centered around zero.

We eliminate cases where too few observations are available to construct a valid measure of volatility and set ($SIGMA$) to missing if there are fewer than five non-zero return observations over the three months window.
\[ SIGMA_{i,t-1} = \left( \frac{252}{N-1} \sum_{k \in \{t-90, \ldots, t\}} r_{i,k}^2 \right)^{1/2} \]  

(A.12)

where \( r_{i,k} \) is the daily stock return for firm \( i \) on day \( k \).

**Capitalization** \( RSIZE_{i,t} \)

Relative size (\( RSIZE \)) is the firm’s equity capitalization relative to the S&P500 index, which we measure by taking the log of the ratio.

\[ RSIZE_{i,t} = \log\left( \frac{\text{MarketEquity}_{i,t}}{\text{MarketValue}_{\text{S&P500},t}} \right) \]  

(A.13)

where \( \text{MarketValue}_{\text{S&P500},t} \) is the sum of the market values for all firms in the S&P500 index in month \( t \).

**Book Value** \( MB_{i,t} \)

We calculate the firm’s ratio of market equity to book equity (\( MB \)).

\[ MB_{i,t} = \frac{\text{MarketEquity}_{i,t}}{\text{BookEquity}_{adj,i,t}} \]  

(A.14)

We adjust BE by the difference between market equity (\( ME \)) and BE ...

\[ MB_{i,t} = \frac{\text{MarketEquity}_{i,t}}{0.9 \times \text{BookEquity}_{i,t} + 0.1 \times \text{MarketEquity}_{i,t}} \]  

(A.15)

To adjust for negative levels of \( \text{BookEquity} \) we replace those observations with $1 before calculating the (\( MB \)) ratio.

**Share Price** \( PRICE_{i,t} \)

We add the log of the stock price, which we cap at $15 (\( PRICE \)). Variation above $15 does not seem to affect failure probability and so the measure is capped at that level.

\[ PRICE_{i,t} = \log(\text{ClosingPrice}_{i,t}) \]  

(A.16)

### A.2.4 Data Sources

All predictor variables in Campbell are constructed using data from CRSP and COMPUSTAT. We have full access to these sources as well, so we should be able to reproduce their dataset, with a notable exception for the impact of having different failure event data. In Sections A.1.3 and A.1.5 we mention that when a firm fails, all subsequent firm-months are removed. Without matching failure event data, we will be removing different sets of post-failure firm-months, even though we have access to the same sources for constructing the predictors.

- stated years available: 1925-DEC - 2014
- earliest record date: 1960.01.31
- latest record date: 2014.12.31
- firm ID field: permno
- records: 3,806,440
- file: CRSP_monthlystock_1960_2014.csv.gz (115.5MB)
- content: 1 record per listed U.S. equity security and month listed

Source A.2.4 covers the NYSE, NASDAQ, AMEX exchanges, and features month end closing price, month end shares outstanding, and monthly return (with and without dividends), in addition to supporting details. This data is used in calculating all of the predictor variables except for SIGMA, which uses daily data from CRSP. Source A.2.4 is comprehensive for all years. To reduce size of the exported file, our data begins with 1960, which would be the earliest year necessary when constructing an observation of the predictor variables for 1961-MAR. That month is the earliest month possible given that quarterly accounting data from Compustat cuts off before January 1961 (we would allow quarterly accounting data from 1961-JAN as of 1961-MAR, which would then require 11 months of lags going back through data from 1960 for calculating EXRETAVG).


- stated years available: 1925.12.31 - 2014
- earliest record date: 1960.01.04
- latest record date: 2014.12.31
- firm ID field: permno
- records: 78,803,160
- content: 1 record per listed U.S. equity security and trading day listed

The details here are the same as with Source A.2.4, but the data has daily granularity. This source is only used for calculating SIGMA. To make FTP transfers easier, we extracted the data in 3 separate files.


- stated years available: 1925-DEC - 2014
- earliest record date: 1960.01.31
• latest record date: 2014.12.31
• firm ID field: (none)
• records: 660
• file: CRSP\_monthlysp500\_1960\_2014.csv.gz (24KB)
• content: 1 record per month

Source A.2.4 includes closing level, closing total market value, and monthly return of the S&P 500 index for every month. This data is used in calculating EXRET and RSIZE.

Quarterly fundamentals data from WRDS, Compustat, North America, Fundamentals Quarterly [Com, 2014]
• stated years available: 1961-JAN - 2014
• earliest record date: 1961.01.31
• latest record date: 2014.12.31
• firm ID field: gvkey
• records: 1,435,025
• file: COMP\_quarterlyfund\_1961\_2014.csv.gz (57.9MB)
• content: 1 record per corporate entity and fiscal or calendar quarter

Source A.2.4 covers all North American companies that file distinct 10K or 10Q statements with the SEC, and features net income, total liabilities, cash and short-term investments, and book value of common equity, in addition to other fundamentals and supporting details. This data is used in calculating NIMTA, TLMTA, CASHMTA, and MB. Source A.2.4 is comprehensive for all quarters when firms filed quarterly statements (coverage decreases noticeably before 1979, a cutoff point that Compustat uses to separate what it considers “historical” data from “current” data [Com, 2014]). Because the scope of this dataset is beyond what we need (strictly those firm-quarters when there was listed equity outstanding), when exporting from WRDS, we use the following filters: Consolidation “C”, Industry “INDL”, Data “STD”, Population “D”, Quarter “Fiscal” & “Calendar”, Currency “USD”, Company “Active” & “Inactive”. Additionally, we require both the fiscal and calendar quarter fields to be specified, meaning that we remove quarters that represent restatements when a company changes fiscal calendars. More detail about the specific contents of Compustat can be found in the WRDS help sections [Com, 2014].

Annual pension details from WRDS, Compustat, North America, Pension Annual [Com, 2014]
• stated years available: 1973-JAN - 2014
• earliest record date: 1973.01.31
• latest record date: 2014.12.31
• firm ID field: gvkey
The details here are the same as with Source A.2.4, but only data related to pensions is included, and it is presented annually. This source is only used for calculating MB, and the required field is not available in Compustat’s quarterly pension data.

**Identifier links from WRDS, CRSP, CRSP/Compustat Merged, Linking Table [CRS, 2014]**

- firm ID field: permno, gvkey
- records: 88,192
- file: CRSP.ccmlink_ALL.csv.gz (6.6MB)
- content: 1 record per linked permno and gvkey

This data is necessary to join the CRSP and Compustat datasets. The details included delineate when companies restructure and/or merge (Compustat) and when stock tickers IPO, delist, and/or combine (CRSP). For our implementation, we only consider links of type “LC” and “LU”, and link primary codes “P” and “C”. More detail about the challenges involved with joining CRSP and Compustat can be found in the WRDS help sections [CRS, 2014].

### A.2.5 Mappings

Below we list the raw variables collected from each source. The exact variable descriptions from WRDS help are included in quotes [CRS, 2014, Com, 2014].

1. Variables from Source A.2.4
   - `prc` (“Price”)
   - `shrout` (“Number of Shares Outstanding”)
   - `retx` (monthly “Holding Period Return without Dividends”)

2. Variables from Source A.2.4
   - `retx` (daily “Return without Dividends”)

3. Variables from Source A.2.4
   - `sprtrn` (“Return on S&P Composite Index”)
   - `totval` (S&P500 “Total Market Value”)

4. Variables from Source A.2.4
   - `NIQ` (“Net Income (Loss)”)  
   - `LTQ` (“Liabilities - Total”)
5. Variables from Source A.2.4

- **PRBA** ("Postretirement Benefit Asset")

The raw variables we collect are then mapped to accounting and market quantities below. All quantities (left hand side) shown here were used in the variable definitions covered in Section A.2.3. The mappings are outlined in Campbell.

\[
\text{NetIncome} = NIQ \times 1,000,000 \\
\text{BookLiabilities} = (LTQ + MIBQ) \times 1,000,000 \\
\text{CashAndShortTermInvestments} = CHEQ \times 1,000,000 \\
\text{MarketEquity} = |prc| \times \text{shrout} \times 1000 \\
R = retx \quad \text{(monthly return from Source A.2.4)} \\
r = retx \quad \text{(daily return from Source A.2.4)} \\
R_{S&P500} = sprtn \\
\text{MarketValue}_{S&P500} = totval \times 1000 \\
\text{BookEquity} = (SEQQ + TXDITCQ + PRBA - PSTKQ) \times 1,000,000 \\
\text{ClosingPrice} = |prc|
\]

### A.3 Box Plots of PFAILURE by Predictor Variable Deciles

This appendix uses the entire 1963 - 2014 sample of firm months as training data. After estimation, **PFAILURE** values were generated for all firm-months as in-sample forecasts. For each of the 8 predictor variables we present a figure containing a paired set of **PFAILURE** box plots at each predictor decile, one each for the non-failure and failure group. The trend and separation of the box plots is then a demonstration of the predictor’s ability to separate the failure group from the non-failure group. Additionally, at the top of the plotting area, the raw number of failures and the failure rate within each decile are displayed. Each decile represents roughly 240,000 observations, which is the denominator used to calculate the failure rates. For the deciles of the variable **PRICE** in Figure A.8, we combine all observations in deciles 6 through 10 together. As noted in Section A.2.3, this is because Campbell capped all share prices at $15 before taking the log function, so there are no separable decile values above log($15).
Figure A.1: Box Plots of PFAILURE by NIMTAAVG Deciles

Figure A.2: Box Plots of PFAILURE by TLMTA Deciles
Figure A.3: Box Plots of PFAILURE by CASHMTA Deciles

Figure A.4: Box Plots of PFAILURE by EXRETAVG Deciles
Figure A.5: Box Plots of PFAILURE by SIGMA Deciles

Figure A.6: Box Plots of PFAILURE by RSIZE Deciles
Figure A.7: Box Plots of PFAILURE by MB Deciles

Figure A.8: Box Plots of PFAILURE by PRICE Deciles
### A.4 Monthly Failure Prediction Summary Table (July 2009 - December 2014)

<table>
<thead>
<tr>
<th>Month</th>
<th>Firms Failures</th>
<th>Fail. Rate</th>
<th>Pred. Rate</th>
<th>Pred. DC10</th>
<th>Not. DC10</th>
<th>Top. DC10</th>
<th>Top. DC10 Tickers</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009-JUL</td>
<td>4764</td>
<td>0.0843%</td>
<td>0.1247%</td>
<td>0.2948%</td>
<td>HW, PCBI, FRP, BWTRQ</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009-AUG</td>
<td>4728</td>
<td>0.1381%</td>
<td>0.1302%</td>
<td>0.2159%</td>
<td>USS</td>
<td>CBCGQ, SYNX, BZ, TLCVF, CIT, CVGI</td>
<td></td>
</tr>
<tr>
<td>2009-SEP</td>
<td>4761</td>
<td>0.0210%</td>
<td>0.0839%</td>
<td>0.1520%</td>
<td>ABMC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009-OCT</td>
<td>4754</td>
<td>0.0613%</td>
<td>0.0841%</td>
<td>0.1150%</td>
<td>NXXIQ, CHIQ, NCS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009-NOV</td>
<td>4739</td>
<td>0.1477%</td>
<td>0.0841%</td>
<td>0.1127%</td>
<td>TKOI</td>
<td>ADVIQ, INVHY, DCGNQ, ALTUQ, EXXI, QGP</td>
<td></td>
</tr>
<tr>
<td>2009-DEC</td>
<td>4670</td>
<td>0.1285%</td>
<td>0.0831%</td>
<td>0.1076%</td>
<td>TRAC</td>
<td>MERC, HABC, HIST, ACHIPQ, DUNRQ</td>
<td></td>
</tr>
<tr>
<td>2010-JAN</td>
<td>4653</td>
<td>0.1496%</td>
<td>0.1729%</td>
<td>0.3622%</td>
<td>ביא</td>
<td>YTCW, TSQ, ASNYQ</td>
<td></td>
</tr>
<tr>
<td>2010-FEB</td>
<td>4642</td>
<td>0.0215%</td>
<td>0.0433%</td>
<td>0.0502%</td>
<td>ACAS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2010-MAR</td>
<td>4642</td>
<td>0.0215%</td>
<td>0.0521%</td>
<td>0.0631%</td>
<td>UIS</td>
<td>CBCGQ, SYNX, BZ, TLCVF, CIT, CVGI</td>
<td></td>
</tr>
<tr>
<td>2010-APR</td>
<td>4677</td>
<td>0.0855%</td>
<td>0.0519%</td>
<td>0.0505%</td>
<td>EPGRQ, WGNIQ, USCM, BRKIQ</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2010-MAY</td>
<td>4658</td>
<td>0.0214%</td>
<td>0.0332%</td>
<td>0.0403%</td>
<td>PSBC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2010-JUN</td>
<td>4630</td>
<td>0.0432%</td>
<td>0.0433%</td>
<td>0.0520%</td>
<td>ACAS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2010-JUL</td>
<td>4632</td>
<td>0.1511%</td>
<td>0.0566%</td>
<td>0.0629%</td>
<td>PCGR, FSNNMQ, CSNT, BLIAQ, MSII, JILIC, WAVE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2010-AUG</td>
<td>4608</td>
<td>0.0217%</td>
<td>0.0558%</td>
<td>0.0631%</td>
<td>DISK</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2010-SEP</td>
<td>4692</td>
<td>0.1918%</td>
<td>0.0553%</td>
<td>0.0590%</td>
<td>ENTN, RDDYQ</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2010-OCT</td>
<td>4677</td>
<td>0.0855%</td>
<td>0.0519%</td>
<td>0.0505%</td>
<td>EPGRQ, WGNIQ, USCM, BRKIQ</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2010-NOV</td>
<td>4658</td>
<td>0.0214%</td>
<td>0.0332%</td>
<td>0.0403%</td>
<td>PSBC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2010-DEC</td>
<td>4630</td>
<td>0.0432%</td>
<td>0.0433%</td>
<td>0.0520%</td>
<td>ACAS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011-JAN</td>
<td>4652</td>
<td>0.1075%</td>
<td>0.0449%</td>
<td>0.0550%</td>
<td>GNBH, TIQC, DIGA1, ANPMF, TSTRQ</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011-FEB</td>
<td>4632</td>
<td>0.0864%</td>
<td>0.0432%</td>
<td>0.0479%</td>
<td>AMBC, PPHIQ, RZTIQ, GIGIQ</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011-MAR</td>
<td>4658</td>
<td>0.0859%</td>
<td>0.0516%</td>
<td>0.0671%</td>
<td>ISCIQ, MACE, MACC, PINN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011-APR</td>
<td>4609</td>
<td>0.0859%</td>
<td>0.0516%</td>
<td>0.0671%</td>
<td>ISCIQ, MACE, MACC, PINN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011-MAY</td>
<td>4586</td>
<td>0.0214%</td>
<td>0.0332%</td>
<td>0.0403%</td>
<td>PSBC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011-JUN</td>
<td>4530</td>
<td>0.0432%</td>
<td>0.0433%</td>
<td>0.0520%</td>
<td>ACAS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011-JUL</td>
<td>4532</td>
<td>0.1511%</td>
<td>0.0566%</td>
<td>0.0629%</td>
<td>PCGR, FSNNMQ, CSNT, BLIAQ, MSII, JILIC, WAVE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011-AUG</td>
<td>4522</td>
<td>0.0217%</td>
<td>0.0558%</td>
<td>0.0631%</td>
<td>DISK</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011-SEP</td>
<td>4544</td>
<td>0.0901%</td>
<td>0.0620%</td>
<td>0.0534%</td>
<td>AMTC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011-OCT</td>
<td>4544</td>
<td>0.0217%</td>
<td>0.0332%</td>
<td>0.0403%</td>
<td>PSBC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011-NOV</td>
<td>4530</td>
<td>0.0432%</td>
<td>0.0433%</td>
<td>0.0520%</td>
<td>ACAS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011-DEC</td>
<td>4532</td>
<td>0.1511%</td>
<td>0.0566%</td>
<td>0.0629%</td>
<td>PCGR, FSNNMQ, CSNT, BLIAQ, MSII, JILIC, WAVE</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table A.1: Monthly Failure Prediction Summary (MTA Dataset, Testing Data Only, Rolling Basis)
### A.5 Bailout Events

<table>
<thead>
<tr>
<th>Ticker</th>
<th>Name</th>
<th>SIC</th>
<th>Cap.</th>
<th>Prog.Code</th>
<th>EventDate</th>
<th>PFAILURE</th>
</tr>
</thead>
<tbody>
<tr>
<td>BAC</td>
<td>Bank of America Corp</td>
<td>6020</td>
<td>$182,345mill.</td>
<td>2008-10-28</td>
<td>0.1079%</td>
<td>48.59%</td>
</tr>
<tr>
<td>WFC</td>
<td>Wells Fargo &amp; Co</td>
<td>6020</td>
<td>$124,185mill.</td>
<td>2008-10-28</td>
<td>0.0646%</td>
<td>26.75%</td>
</tr>
<tr>
<td>GS</td>
<td>Goldman Sachs Group Inc</td>
<td>6020</td>
<td>$52,287mill.</td>
<td>2008-11-14</td>
<td>0.1305%</td>
<td>93.19%</td>
</tr>
<tr>
<td>BBT</td>
<td>Bank of New York Mellon Corp</td>
<td>6020</td>
<td>$37,339mill.</td>
<td>2008-10-28</td>
<td>0.0718%</td>
<td>29.60%</td>
</tr>
<tr>
<td>STT</td>
<td>State Street Corp</td>
<td>6020</td>
<td>$24,558mill.</td>
<td>2008-10-28</td>
<td>0.0882%</td>
<td>60.21%</td>
</tr>
<tr>
<td>AXP</td>
<td>American Express Co</td>
<td>6020</td>
<td>$21,516mill.</td>
<td>2009-01-09</td>
<td>0.3554%</td>
<td>36.38%</td>
</tr>
<tr>
<td>FNC</td>
<td>FNC Financial Services Group Inc.</td>
<td>6020</td>
<td>$18,362mill.</td>
<td>2008-12-31</td>
<td>0.1468%</td>
<td>57.51%</td>
</tr>
<tr>
<td>COF</td>
<td>Capital One Financial Corp.</td>
<td>6141</td>
<td>$17,084mill.</td>
<td>2008-11-14</td>
<td>0.1314%</td>
<td>58.42%</td>
</tr>
<tr>
<td>STI</td>
<td>SunTrust Banks Inc.</td>
<td>6120</td>
<td>$21,609mill.</td>
<td>2008-11-14</td>
<td>0.1751%</td>
<td>52.79%</td>
</tr>
<tr>
<td>NTRS</td>
<td>Northern Trust Corp</td>
<td>6331</td>
<td>$4,667mill.</td>
<td>2009-01-29</td>
<td>0.1334%</td>
<td>65.70%</td>
</tr>
<tr>
<td>RF</td>
<td>Regions Financial Corp</td>
<td>6020</td>
<td>$7,674mill.</td>
<td>2008-11-14</td>
<td>0.3335%</td>
<td>38.94%</td>
</tr>
<tr>
<td>M&amp;T</td>
<td>M&amp;T Bank Corp</td>
<td>6020</td>
<td>$7,080mill.</td>
<td>2008-12-23</td>
<td>0.1681%</td>
<td>1.82%</td>
</tr>
<tr>
<td>KEY</td>
<td>KeyCorp</td>
<td>6020</td>
<td>$6,140mill.</td>
<td>2008-11-14</td>
<td>0.3104%</td>
<td>30.54%</td>
</tr>
<tr>
<td>FITB</td>
<td>Fifth Third Bancorp</td>
<td>6020</td>
<td>$5,521mill.</td>
<td>2008-12-31</td>
<td>0.7671%</td>
<td>4.78%</td>
</tr>
<tr>
<td>ASB</td>
<td>Associated Banc-Corp</td>
<td>6020</td>
<td>$2,817mill.</td>
<td>2008-11-21</td>
<td>0.0739%</td>
<td>20.64%</td>
</tr>
<tr>
<td>DSN</td>
<td>Discover Financial Services Inc</td>
<td>6141</td>
<td>$2,753mill.</td>
<td>2009-03-13</td>
<td>0.2221%</td>
<td>42.89%</td>
</tr>
<tr>
<td>CYN</td>
<td>City National Corp</td>
<td>6020</td>
<td>$2,561mill.</td>
<td>2008-11-21</td>
<td>0.0399%</td>
<td>17.67%</td>
</tr>
<tr>
<td>FHN</td>
<td>First Horizon National Corp</td>
<td>6020</td>
<td>$2,401mill.</td>
<td>2008-11-14</td>
<td>0.2591%</td>
<td>5.71%</td>
</tr>
<tr>
<td>TCB</td>
<td>TCF Financial Corp</td>
<td>6020</td>
<td>$2,323mill.</td>
<td>2008-11-14</td>
<td>0.0607%</td>
<td>91.11%</td>
</tr>
<tr>
<td>FULT</td>
<td>Fulton Financial Corp</td>
<td>6020</td>
<td>$1,946mill.</td>
<td>2008-12-23</td>
<td>0.1606%</td>
<td>73.86%</td>
</tr>
<tr>
<td>BPOP</td>
<td>Popular Inc</td>
<td>6020</td>
<td>$1,763mill.</td>
<td>2008-12-05</td>
<td>0.4314%</td>
<td>10.48%</td>
</tr>
<tr>
<td>FMER</td>
<td>FirstMerit Corp</td>
<td>6020</td>
<td>$1,687mill.</td>
<td>2009-01-09</td>
<td>0.2215%</td>
<td>25.45%</td>
</tr>
<tr>
<td>WLC</td>
<td>Wilmington Trust Corp</td>
<td>6020</td>
<td>$1,647mill.</td>
<td>2008-12-12</td>
<td>0.1348%</td>
<td>77.09%</td>
</tr>
<tr>
<td>WAFD</td>
<td>Washington Federal Inc.</td>
<td>6035</td>
<td>$1,549mill.</td>
<td>2008-11-14</td>
<td>0.0567%</td>
<td>10.11%</td>
</tr>
<tr>
<td>SUSQ</td>
<td>Susquehanna Bancshares Inc</td>
<td>6020</td>
<td>$1,316mill.</td>
<td>2008-12-12</td>
<td>0.1621%</td>
<td>74.65%</td>
</tr>
<tr>
<td>SIVB</td>
<td>SVB Financial Group</td>
<td>6020</td>
<td>$1,313mill.</td>
<td>2008-12-12</td>
<td>0.1080%</td>
<td>14.65%</td>
</tr>
<tr>
<td>WABC</td>
<td>Westamerica Bancorporation</td>
<td>6020</td>
<td>$1,234mill.</td>
<td>2009-02-13</td>
<td>0.0620%</td>
<td>91.56%</td>
</tr>
<tr>
<td>NPBC</td>
<td>National Penn Bancshares Inc</td>
<td>6020</td>
<td>$1,231mill.</td>
<td>2008-12-12</td>
<td>0.0956%</td>
<td>92.65%</td>
</tr>
<tr>
<td>FNB</td>
<td>F.N.B. Corp</td>
<td>6020</td>
<td>$1,183mill.</td>
<td>2009-01-09</td>
<td>0.2477%</td>
<td>91.08%</td>
</tr>
<tr>
<td>TRMK</td>
<td>Trustmark Corp</td>
<td>6020</td>
<td>$1,176mill.</td>
<td>2008-11-21</td>
<td>0.0714%</td>
<td>47.46%</td>
</tr>
<tr>
<td>ONB</td>
<td>Old National Bancorp</td>
<td>6020</td>
<td>$1,139mill.</td>
<td>2008-12-12</td>
<td>0.0675%</td>
<td>20.06%</td>
</tr>
<tr>
<td>WNY</td>
<td>Whitney Holding Corp.</td>
<td>6020</td>
<td>$1,122mill.</td>
<td>2008-12-19</td>
<td>0.1126%</td>
<td>8.05%</td>
</tr>
<tr>
<td>PVTB</td>
<td>PrivateBancorp Inc</td>
<td>6020</td>
<td>$1,091mill.</td>
<td>2009-01-30</td>
<td>0.1020%</td>
<td>51.23%</td>
</tr>
<tr>
<td>SBNY</td>
<td>Signature Bank</td>
<td>6020</td>
<td>$1,048mill.</td>
<td>2008-12-12</td>
<td>0.0457%</td>
<td>64.11%</td>
</tr>
<tr>
<td>PBP</td>
<td>First BanCorp (Puerto Rico)</td>
<td>6020</td>
<td>$1,031mill.</td>
<td>2009-01-16</td>
<td>0.2572%</td>
<td>27.75%</td>
</tr>
<tr>
<td>UMPQ</td>
<td>Umpqua Holdings Corp</td>
<td>6020</td>
<td>$1,023mill.</td>
<td>2008-11-14</td>
<td>0.0737%</td>
<td>67.38%</td>
</tr>
<tr>
<td>CATY</td>
<td>Cathay General Bancorp</td>
<td>6020</td>
<td>$1,016mill.</td>
<td>2008-12-05</td>
<td>0.2292%</td>
<td>84.90%</td>
</tr>
<tr>
<td>CIT</td>
<td>CIT Group Inc</td>
<td>6172</td>
<td>$954mill.</td>
<td>2008-12-31</td>
<td>1.8970%</td>
<td>14.65%</td>
</tr>
<tr>
<td>CMB</td>
<td>General Motors Financial Co Inc</td>
<td>6141</td>
<td>$855mill.</td>
<td>2008-12-29</td>
<td>0.4341%</td>
<td>14.43%</td>
</tr>
</tbody>
</table>

1 Standard Industrial Classification codes. 6000s (Division for Finance, Insurance, and Real Estate).
2 Federal bailout program codes: P, Purchase Transaction. CPP (Capital Purchase Program).
3 SSFI (Systematically Significant Failing Institutions), AIFP Automotive Industry Financing Program.
4 Observed as of the close of the last business day of the previous month.
5 Quantile (%) of PFAILURE over all forecasts for the same month.

Table A.2: Bailout Event Detail (excludes regional banks with firm caps. under $1Bill.)
Appendix B

Chapter 4 Appendix
B.1 Class Conditional Histograms Using Datasets $MTA$ and $MME$, All Predictors

Figure B.1: Class Conditional Histogram for Variable $NIMTAAVG$, $MTA$ Dataset

Figure B.2: Class Conditional Histogram for Variable $ANIMEAVG$, $MME$ Dataset
Figure B.3: Class Conditional Histogram for Variable TLMTA, MTA Dataset

Figure B.4: Class Conditional Histogram for Variable TLME, MME Dataset
Figure B.5: Class Conditional Histogram for Variable \textit{CASHMTA}, \textit{MTA} Dataset

Figure B.6: Class Conditional Histogram for Variable \textit{CASHME}, \textit{MME} Dataset
Figure B.7: Class Conditional Histogram for Variable $EXRETAVG$, $MTA$ Dataset

Figure B.8: Class Conditional Histogram for Variable $EXRETAVG$, $MME$ Dataset
Figure B.9: Class Conditional Histogram for Variable $SIGMA$, $MTA$ Dataset

Figure B.10: Class Conditional Histogram for Variable $SIGMA$, $MME$ Dataset
Figure B.11: Class Conditional Histogram for Variable $RSIZE$, $MTA$ Dataset

Figure B.12: Class Conditional Histogram for Variable $RSIZE$, $MME$ Dataset
Figure B.13: Class Conditional Histogram for Variable MB, MTA Dataset

Figure B.14: Class Conditional Histogram for Variable ABEME, MME Dataset
Figure B.15: Class Conditional Histogram for Variable $PRICE$, MTA Dataset

Figure B.16: Class Conditional Histogram for Variable $PRICE$, MME Dataset
B.2 Scatter Plots of Dataset *MME* Using KDE Transform (with Tuning Parameter $\lambda = 0.00001$), All Predictors

![Figure B.17: Scatter Plot for Variable ANIMEAVG and Latent State $\pi$](image1)

![Figure B.18: Scatter Plot for Variable ANIMEAVG and Latent State $\pi$ (probit link)](image2)
Figure B.19: Scatter Plot for Variable $TLME$ and Latent State $\pi$

Figure B.20: Scatter Plot for Variable $TLME$ and Latent State $\pi$ (probit link)
Figure B.21: Scatter Plot for Variable CASHME and Latent State π

Figure B.22: Scatter Plot for Variable CASHME and Latent State π (probit link)
Figure B.23: Scatter Plot for Variable $EXRETAVG$ and Latent State $\pi$ 

Figure B.24: Scatter Plot for Variable $EXRETAVG$ and Latent State $\pi$ (probit link)
Figure B.25: Scatter Plot for Variable $SIGMA$ and Latent State $\pi$

Figure B.26: Scatter Plot for Variable $SIGMA$ and Latent State $\pi$ (probit link)
Figure B.27: Scatter Plot for Variable RSIZE and Latent State $\pi$

Figure B.28: Scatter Plot for Variable RSIZE and Latent State $\pi$ (probit link)
Figure B.29: Scatter Plot for Variable \textit{ABEME} and Latent State \(\pi\)

Figure B.30: Scatter Plot for Variable \textit{ABEME} and Latent State \(\pi\) (probit link)
Figure B.31: Scatter Plot for Variable $PRICE$ and Latent State $\pi$

Figure B.32: Scatter Plot for Variable $PRICE$ and Latent State $\pi$ (probit link)