Corporate Default Prediction

by

Xingwei Wu

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Engineering - Industrial Engineering and Operations Research

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:
Professor Xin Guo, Chair
Professor Ilan Adler
Professor Noureddine El Karoui

Fall 2011

Abstract

Corporate Default Prediction

by

Xingwei Wu
Doctor of Philosophy in Engineering - Industrial Engineering and Operations Research

University of California, Berkeley

Professor Xin Guo, Chair

In the literature of predicting corporate default, it is an ad-hoc process to select the predictors and different models often use different predictors. We study the predictors of U.S corporate default by Forward Stepwise and Lasso model selection methods. Out of 30 candidate default predictors that have been used in the default-predicting literature, we identify a set of eight default predictors that have strong effects in predicting default using the U.S corporate default data from 1984-2009. We compare the eight default predictors' predicting effect over the past three major economic recessions and find that the recession in early 1990 and the recent sub-prime mortgage crisis share some common default characteristics, while the recession in 2000 is different from the other two. We then present a decision-based default prediction framework where we incorporate the default forecaster's loss utility into default classification and derive an optimal decision rule for this classification problem. By combining the default forecaster's loss utility into Support Vector Machines(SVMs), we show that minimizing the utility adjusted hinge loss is consistent with minimizing utility adjusted classification loss. Our empirical classification result of the decision-based Support Vector Machines demonstrates more classification accuracy and flexibilities in meeting different default forecasters' goals in comparison to traditional statistical methods.

To My Grandparents, Parents

Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank members of my dissertation committee: Professor Ilan Adler, Professor Xin Guo and Professor Noureddine El Karoui for their helpful comments and encouragement.

Part I

Chapter 1

Introduction

Corporate credit risk is the risk of financial losses due to unexpected changes in the credit quality of a counter-party in a financial agreement. Examples of credit quality changes are agency downgrades, failure to service debt obligation, filing for bankruptcy.

At the center of corporate credit risk is the probability of default, by which we mean any type of failure to honor a financial agreement. Default is a rare event. Prior to default, it is difficult to differentiate between firms that will default and those that will not. One approach is to make probabilistic assessments of the likelihood of default. According to Moody's KMV [22], the typical firm has a default probability of around 2% in any year. However, there are considerable variations in default probabilities across firms. For example, the odds of a firm with a AAA rating defaulting are only about 0.02% annually. A single A-rated firm has odds of around 0.1% annually, five times higher than a AAA. At the bottom of the rating scale, a CCC-rated firm's odds of defaulting are 4%, 200 times the odds of a AAA-rated firm. The loss suffered by a lender or counter-party in the event of default is usually significant and is determined largely by the details of the particular contract or obligation. For example, from Moody's KMV [22], typical loss rates in the event of default for senior secured bonds, subordinated bonds and zero coupon bonds are 49%, 68%, and 81%, respectively.

Given the recent financial crisis, stressed by the incorrect assessment of the default risk and the correlation imbedded in corporate loans, bonds and derivatives, study of firms' credit risk has assumed increased importance to the financial industry and regulatory agency. The first part of this dissertation is devoted to review of major results in corporate default modeling and empirical literature. In the second part of the dissertation, we conduct empirical study on corporate default predictors using statistical model selection methods and then presents a decision-based default prediction framework that takes into accounts of the default forecaster's utility in predicting default. This new approach has the potential of providing more prediction accuracy and modeling flexibilities when compared with traditional statistical prediction methods.

Chapter 2

Single Firm Default

Single firm default prediction is about studying, modeling and understanding the default behavior of a single obligor without interacting with other firms. Four major modeling methods in the literature are: Structural Model, Reduced Form Model, Incomplete Information Model and Statistical Model.

2.1 Structural Model

The fundamental of the structural model, which goes back to Black & Scholes (1973) [14] and Merton (1974) [66], is that corporate liabilities are contingent claims on the assets of a firm. The default time in the structure model is usually characterized as the first hitting time of the firm's asset value to a given boundary determined by the firm's capital structure. The default time is in general a predictable stopping time in the structural model when asset process is modeled by a continuous Markov process.

2.1.1 Merton Model

The progenitor of structural model is established by Merton (1974), which proposed a mechanism for the default of a firm in terms of the relationship between its assets and liabilities.

Consider a firm whose asset value follows some stochastic process V_t. The firm finances itself by equity and debt. In Merton's model, it consists of one single debt obligation or zero-coupon bond with face value B and maturity T. The value at time t of equity and debt is denoted by E_t and B_t. The asset value of the firm is $V_t = E_t + B_t$, $0 \leq t \leq T$. Merton assumes that the firm cannot pay out dividends or issue new debt. Default occurs if the firm misses a payment to its debt holders, which in Merton model can occur only at the maturity T of the bond. At the maturity time T,

$$E_T = (V_T - B)^+,$$

$$B_T = \min(V_T, B) = B - (B - V_T)^+.$$

This implies that E_T, the value of the firm's equity at time T E_T, is equal to the payoff of a European call option on T, and the value of the firm's debt at maturity equals the nominal value of the liabilities minus the pay-off of an European put option on V_T with exercise price B.

In the Merton Model, it is assumed that the process V_t follows a geometric Brownian motion under the real-world measure:

$$dV_t = \mu_V V_t dt + \sigma_V V_t dW_t, \tag{2.1}$$

and the default probability is given by:

$$P(V_T \leq B) = P(\ln V_T \leq \ln B) = \Phi(\frac{\ln(B/V_0) - (\mu_V - \frac{1}{2}\sigma_V^2)T}{\sigma_V\sqrt{T}}). \tag{2.2}$$

Under the risk-neutral probability measure V_t follows $dV_t = rV_t dt + \sigma_V V_t dW_t'$. The equity value at current time given by Black-Scholes call option formula C is:

$$E_0 = C(\sigma_V, T, B, r, V_0) = V_0\phi(d_+) - e^{-rT}B\phi(d_-), \tag{2.3}$$

where r is the risk free interest rate, ϕ is cumulative distribution function of Gaussian distribution and

$$d_\pm = \frac{(r \pm \frac{1}{2}\sigma_V^2)T - \ln L}{\sigma_V\sqrt{T}}.$$

The value of the corresponding bonds at time 0 is:

$$B_0 = Be^{-rT} - P(\sigma_V, T, B, r, V_0) = V_0 - V_0\phi(d_+) + e^{-rT}B\phi(d_-). \tag{2.4}$$

The value of equity and bond (2.3) (2.4) together proves the market value identity

$$V_0 = E_0 + B_0. \tag{2.5}$$

While both equity and debt values depend on the firm's leverage ratio, equation (2.5) shows that their sum does not. This demonstrates that Modigliani & Miller (1958) [67] theorem holds also in the presence of default, which asserts that the market value of the firm is independent of its leverage. See Rubinstein (2003, 2006) [74] [75] for a discussion.

One widely applied structural forecasting model is a particular application of Merton's model that is developed by the Moody' KMV corporation, which is referred to as the KMV-Merton model. The KMV-Merton model applies the framework of Merton and recognizes that neither the underlying value of the firm nor its volatility are directly observable. Under the model's assumptions both can be inferred from the value of equity, the volatility of equity and several other observable variables by solving two nonlinear simultaneous equations. After inferring these values, the model specifies that the probability of default is the normal cumulative density function of a z-score (distance to default) depending on the firm's underlying value, the firm's volatility and the face value of the firm's debt.

The KMV-Merton model uses two important equations. The first is the Black-Scholes-Merton equation (2.3). The second relates the volatility of the firm's value to the volatility of its equity. Under Merton's assumptions the value of equity is a function of the value of the firm and time, so it follows from Ito's lemma that:

$$\sigma_E = (\frac{V}{E})N(d_1)\sigma_V. \tag{2.6}$$

The KMV-Merton model basically uses these two nonlinear equations (2.3) and (2.6) to translate the value and volatility of a firm's equity into an implied probability of default. The first step in implementing the KMV-Merton model is to estimate σ_E from either historical stock returns data or from option implied volatility data. The second step is to choose a forecasting horizon and a measure of the face value of the firm's debt. For example, it is common to use historical returns data to estimate σ_E, assume a forecasting horizon of one year (T = 1), and take the book value of the firm's total liabilities to be the face value of the firm's debt. The third step is

to collect values of the risk-free rate and the market equity of the firm. After performing these three steps, values for each of the variables in the nonlinear equations are obtained except for V and σ_V, the total value of the firm and the volatility of firm value respectively. The fourth step is to simultaneously solve equations numerically (2.3) and (2.6) for values of V and σ_V. Then, the distance to default defined as the number of standard deviations the expected asset value at maturity away from the default asset value can be calculated as:

$$DD = \frac{\ln(V/F) + (\mu - 0.5\sigma_V^2)T}{\sigma_V\sqrt{T}}. \tag{2.7}$$

The corresponding implied probability of default (Expected Default Frequency, EDF) is: $\pi_{KMV} = N(-DD)$.

The actual model that Moody's KMV uses is proprietary that it is a generalization of the Merton model that allows for various classes and maturities of debt. And instead of using the cumulative normal distribution to convert DD to default probability, Moody's KMV uses its large historical database to estimate the empirical distribution of changes in distances to default and to calculate the default probability based on the empirical distribution. Finally, KMV also makes proprietary adjustments to the accounting information it uses to calculate the face value of debt.

2.1.2 First-passage Model

In the Merton approach, firm value can dwindle to almost nothing without triggering default. This is unfavorable to bondholders. Bond indenture provisions often include safety covenants that give bond investors the right to reorganize a firm if its value falls below a given barrier. Black and Cox (1976)[13] propose a structural model where an obligor defaults when the value of its assets hits below a certain barrier. This is the so-called first-passage-time model. In this class of model, default occurs when the asset-value process crosses for the first time a default threshold B.

Suppose the default barrier D is a constant valued in $(0, V_0)$. The default time τ is a continuous random variable valued in $(0, \infty]$ given by

$$\tau = \inf\{t \geq 0 : V_t \leq D\}.$$

Under the Black-Scholes setting with asset dynamics (2.1), default probabilities are calculated as:

$$P(T) = P(M_T < D) = P(\min_{s \leq T}(ms + \sigma W_s) < \ln(D/V_0)),$$

where $M_t = \min_{s \leq T} V_s$ is the running low of firm asset value, which has a inverse Gaussian distribution.

2.1.3 Exogenous V.S. Endogenous Default Boundaries

In Merton's model, the term structure of credit spreads implied by the structural model asymptotically approaches zero with maturity time T tends to zero. This contradicts with empirical evidence that spread still exists in the very short term maturity. This discrepancy follows from three structural model properties: the firm asset value follows a geometric brownian motion which is continuous and predictable; the firm asset value grows at a positive risk-free rate; the capital structure is constant.

To address the credit spread discrepancy, different structural default models have been proposed. Longstaff and Schwartz (1995) [63] assume that the firm defaults when its asset value first V falls beneath a proportion β of debt principal value P:

$$V_{ex-boundary} = \beta P,$$

where the risk free interest rate follows Vasicek dynamics, and firm-value follows geometric Brownian motion.,

Collin-Dufresne and Goldstein (2001) [19] assume that the firm maintains a target capital structure and the exogenous default threshold changes dynamically over time, as a mean-reverting process:

$$dk_t = \lambda(y_t - v - k_t)dt.$$

In the endogenous default model, the default boundary is that which maximizes equity value. It is determined not only by debt principal, but also by the riskiness of the firm's activities, the maturity of debt issued, payout levels, default costs, and corporate tax rates. Leland and Toft (1996) [61] formulate the endogenous default model as an optimal stopping problem, and solve it for the optimal capital structure and default policy. The model specifications are as follows:

- the asset process is modeled as a geometric Brownian motion, $V_t = e^{Z(t)}$, where $Z_t = Z_0 + mt + \sigma W_t$, W_t is a standard Brownian motion and the growth rate $\mu = m + \sigma^2/2$;
- the firm generates cash flow at the rate δV_t at time t;
- the debt of the firm is modeled as a consol bond with coupon rate $C > 0$, and tax rate $\theta \in (0, 1)$, tax shield θC;
- all agents are assumed to be risk-neutral, and discount cash flows at a fixed interest rate r.

The initial value of equity to the shareholders is:

$$F(V_0, C, \tau) = E[\int_0^\tau e^{-rt}(\delta V_t + (\theta - 1)C)dt],$$

where τ is a liquidation policy chosen by the equity shareholders.

Equity shareholders would therefore choose the liquidation policy solving the optimization problem

$$S_0 = \sup_{\tau \in \Gamma} F(V_0, C, \tau).$$

The optimal liquidation time shown in Leland and Toft (1996) [61] is:

$$\tau(V_B) = \inf\{t : V_t \leq V_B\},$$

where $V_B = \frac{(1-\theta)C\gamma(r-\mu)}{r(1+\gamma)\delta}$ and $\gamma = \frac{m+\sqrt{m^2+2r\sigma^2}}{\sigma^2}$.

2.1.4 Empirical Study

There are several inconsistent conclusion regarding testing of KMV models. Kealhofer (2003) [54] conducted "power test" and "intra-cohort analysis" using default data of non-financial firm with public debt ratings from 1979 to 1990 and show that:

- KMV-Merton model is a more accurate predictor of default than agency debt ratings (S&P);
- EDFs contain all the information in S&P ratings, and correctly predicted nearly 80% percent of the sub-investment grade rating changes six months in advance and approximately 65% of the investment-grade rating change.

Using default data (total 1,449 firm defaults) for the period 1980-2003 from the Altman default database and default lists in Moody's website, Bharath (2006) [11] compare the KMV-Merton model to a similar but much simpler alternative. However, they find that:

- KMV-Merton model does not produce a sufficient statistic for the probability of default. It performs slightly worse as a predictor in hazard models and in out of sample forecasts;
- several other forecasting variables (log(Market Equity), Log(Face Debt), $1/\sigma_E$, equity excess return over past year, NI/TA) are also important predictors, and fitted hazard model values outperform KMV-Merton default probabilities out of sample;
- implied default probabilities from credit default swaps (CDS spread data of 1998 to 2003 from www.credittrade.com) and corporate bond yield spreads (bond data of 1971-1997 extracted from the Lehman Brothers Fixed Income Database) are only weakly correlated with KMV-Merton default probabilities after adjusting for agency ratings, bond characteristics, and our alternative predictor;
- the distance to default functional form is useful, while solving the simultaneous nonlinear equations required by the KMV-Merton model is not.

Leland (2004) [60] examines the default probabilities predicted by two types of structural models: exogenous default boundaries and endogenous default boundaries. He tests the ability of these models to capture the actual average default frequencies across bonds with different ratings reported in Moody's (2001) corporate bond default data, 1970-2000. When default costs and recovery rates are matched, Leland finds that

- exogenous and endogenous default boundary models fit observed default frequencies equally well. The models predict longer-term default frequencies quite accurately both for investment grade and non-investment grade bonds. Shorter-term default frequencies tend to be underestimated, suggesting that a jump component should be included in asset value dynamics.

Eom, Helwege and Huang (2004) [38] test five structural models of corporate bond pricing: Merton (1974), Geske (1977), Longstaff and Schwartz (1995), Leland and Toft (1996), and Collin-Dufresne and Goldstein (2001). Using a sample of 182 noncallable bond prices from firms with simple capital structures during the period 1986-1997, they find that:

- all the structural models tend to generate extremely low spreads on the bonds that the models consider safe and to generate very high spreads on the bonds considered to be very risky. Both Merton and Geske models share a tendency toward under-estimation of corporate bond spreads on average, but Geske model is less severe suggesting that the endogenous default boundary of Geske model is a major improvement and that the option to make coupon payments in distress helps to improve the spread prediction. Leland and Toft over-predicts spreads on average (especially shorter-maturity bond), owing largely to the assumption of a continuous coupon. The correlation between credit spread and interest rate in Longstaff and Schwartz, and Collin-Dufresne and Goldstein is not very important empirically, and their models on overage underestimate the spread;

- future research should be on raising spreads on the safer bonds without raising spreads too much for the riskiest bonds. The modeling of the coupon and the recovery rate assumption are important factors that affects the variance of the spread prediction.

2.2 Reduced Form Model

Structural models have two drawbacks when calibrated with the real world data: first, the firm's asset value process is not directly observable; second, the default event in structural model comes without any surprise (predictable default time) which implies credit spreads should be close to zero on short maturity debt which is inconsistent with historical market credit spread data.

In contrast, reduced form models are developed to avoid modeling the firm's unobservable asset value process. Reduced form models go back to Lando (1994, 1998) [55] [56], Jarrow and Turnbull (1995) [72] and Duffie and Singleton (1999) [34]. In the reduced form model, it is assumed that default occurs at a hazard rate, or intensity without any warning. The dynamics of the default intensity are usually specified under the risk neutral probability.

More specifically, fixing a probability space $(\Omega, \mathcal{F}, P)$, there is an R^d-valued process X_t, which represents macro-economical and firm-specific state variable. Let N^i denote the default process of firm i, $i = 1, ...n$, such that the default of the ith firm occurs when N^i jumps from 0 to 1. The filtration is generated collectively by the information contained in the state variables $\mathcal{F}_t = \sigma(X_s, 0 \leq s \leq t)$.

Let τ^i denote the first jump time of N^i with default intensity defined by the instantaneous default probability $\lambda_t^i = \lim_{h \downarrow 0} \frac{P(t+h \geq \tau \geq t | \mathcal{G}_t)}{h}$. Then the conditional and unconditional distributions of τ^i are given by

$$P(\tau^i > t | \mathcal{F}_t) = exp(-\int_0^t \lambda_s^i(X_s)ds), t \in [0, T],$$

$$P(\tau^i > t) = E[exp(-\int_0^t \lambda_s^i(X_s)ds)], t \in [0, T].$$

The intensity function λ_t^i is $\mathcal{F}_t$ measurable and the realized history of the process λ_t^i define a non-homogeneous poisson process N^i stopped at its first jump. The process N is called a Cox process, or doubly-stochastic point process.

2.2.1 Affine Default Intensity

Duffie and Kan (1996) [31] introduce a class of intensity models that have closed-form solutions for $q(T) = P(\tau \leq T)$. Assume that the risk factor X solves the stochastic differential equation

$$dX_t = \mu(X_t)dt + \sigma(X_t)dW_t, \tag{2.8}$$

where the coefficients are affine functions of the state variables:

- $\mu(x) = \mu_0 + \mu_1 x$, where $\mu_0 \in R^d$ is a vector of constants and μ_1 in $R^{d \times d}$ is a matrix of constants;
- $(\sigma(x)\sigma(X)^T)_{ij} = (\sigma_0)_{ij} + (\sigma_1)_{ij} x$, where $\sigma_0 \in R$ and $\sigma_1 \in R^{d \times d}$ are matrices of constants;
- $W \in R^d$ is a standard Brownian motion under risk-neutral measure Q.

Additionally the function Λ that maps the state variables X into the default intensity is also assumed affine. The default probability is given by:

$$q(T) = 1 - exp(a(T) - b(T)X_0), \tag{2.9}$$

where the functions $a(.), b(.)$ solve a system of ordinary differential equations.

The risk dynamics of (2.8) can be extended to include unexpected jumps

$$dX_t = \mu(X_t)dt + \sigma(X_t)dW_t + dJ_t,$$

where J is a pure jump process whose arrival intensity $h(X_t)$ at time t is affine in X_t.

2.2.2 Empirical Study

Shumway (2001) [78] uses a hazard rate approach to calibrate the reduced form model. He concludes that:

- the hazard model is identical to the logistic regression model, and is identical to the binomial likelihood function obtained by treating annual bankruptcy indicator variable as independent binomials;
- single-period bankruptcy classification models give biased and inconsistent probability estimates while hazard models produce consistent estimates;
- about half of the accounting ratios that have been used in previous models are not statistically significant bankruptcy predictor;
- several market-driven variables are strongly related to bankruptcy probability, including market size, past stock returns, and the idiosyncratic standard deviation of stock returns.

Chava, Jarrow (2004) [17] follow Shumway's approach and investigate the forecasting accuracy of bankruptcy hazard rate models for U.S. companies over the time period 1962-1999 (1461 bankruptcies) using both yearly and monthly observation intervals. They estimate the hazard rate model with explanatory variables:

- Altman's (1968) WC/TA, RE/TA, EBIT/TA, ME/TL, SL/TA;
- Zmijewski's Variable (1984) NI/TA, TL/TA, CA/CL;
- Shumway (2001) NI/TA, TL/TA, relative size, excess monthly return, stock volatility.

They conclude that:

- Shumway (2001) model has superior forecasting performance as opposed to Altman (1968) and Zmijewski (1984);
- it is important to include industry effects (4 digit SIC codes) in hazard rate estimation. Industry groupings are significant affect both the intercept and slope coefficients in forecasting;
- bankruptcy prediction is markedly improved using monthly observation intervals;
- accounting variables add little predictive power when market variables are already included.

More recently, Duffie, Saita and Wang (2007) [33] build a multi-period hazard model with stochastic covariates for US Industrial firms. Based on over 390,000 firm-months of data spanning 1980 to 2004, they conclude that:

- the term structure of conditional future default probabilities depends on a firm's distance to default (a volatility-adjusted measure of leverage), on the firm's trailing stock return, on trailing S&P 500 returns, and on US interest rates.

The stochastic explanatory variables they use for the default intensity are:

- distance to default;
- trailing one-year stock return;
- three-month treasury bill rate;
- trailing one-year return on the S$P 500 index.

2.3 Incomplete Information Model

Incomplete information default models build an intrinsic connection between structural models and reduced form models. By taking into accounts the incomplete information available to the market, structural models with predictable stopping time can be transformed into reduced form models with inaccessible stopping time.

2.3.1 Information with Noise

Duffie and Lando (2001) [32] introduce the notion of incomplete account information. In their model, bond investors receive information about the firm's asset process at selected times $t_1, t_2, ...$, with $t_i < t_{i+1}$ and at each observation time there is a noisy accounting report of assets, given by $\hat{V}_t$ (V_t is the true asset value.) $\log \hat{V}_t$ and $\log V_t$ are joint normal and suppose that $Y(t) = \log \hat{V}_t = Z(t) + U(t)$, where $U(t)$ is normally distributed and independent of $Z(t)$.

The information filtration $\mathcal{H}_t$ available to the market is defined by:

$$\mathcal{H}_t = \sigma(Y(t_1), ..., Y(t_n), 1_{\tau \leq s} : 0 \leq s \leq t),$$

for the largest n such that $t_n \leq t$, where $\tau = \tau(V_B)$. The density of the true asset value $Z(t)$ conditional on the noise observation Y_t and on $\tau > t$ is given by:

$$g(x|y, z_0, t) = \frac{b(x|y, z_0, t)}{\int_{\tilde{v}}^{+\infty} b(z|y, z_0, t)dz},$$

where $b(x|Y_t, z_0, t)dx = P(\tau > t, Z_t \in dx|Y_t)$ for $x > \tilde{v}$ is given by:

$$b(x|Y_t, z_0, t) = \frac{\psi(z_0 - \tilde{v}, x - \tilde{v}, \sigma\sqrt{t})\Phi_U(Y_t - x)\Phi_Z(x)}{\Phi_Y(Y_t)}$$

where $\psi(z_0, x, \sigma\sqrt{t})$ is the probability that $\min\{Z_s : 0 \leq s \leq t\} > 0$ conditional on Z starting at some given level z_0 at time 0 and ending at some level x at a given time t. The $\mathcal{H}_t$-conditional probability $p(t, s) = P(\tau > s|\mathcal{H}_t)$ of survival to some future time $s > t$ is:

$$p(t, s) = \int_{\tilde{v}}^{+\infty} (1 - \pi(s - t, x - \tilde{v}))g(x|Y_t, z_0, t)dx,$$

where $\pi(t,x)$ denotes the probability of first passage of a Brownian motion with drift m and volatility parameter σ from an initial condition $x > 0$ to a level below 0 before time t. The default intensity $\lambda_t = \lim_{h\downarrow 0} \frac{P(\tau \leq t+h|\mathcal{F}_t)}{h}$ is given by:

$$\lambda_t(\omega) = \frac{1}{2}\sigma^2 f_x(t, \tilde{v}, \omega), 0 < t \leq \tau,$$

where $f(t, ., \omega)$ is the conditional density of the $\mathcal{H}_t$ conditional distribution of Z_t.

2.3.2 Partial and Delayed Information

A work by Cétin, Jarrow and Protter (2004) [16] generalize Duffie and Lando's reduce form model by constructing an economy where the market only sees a reduction of the manager's information set. The reduced information makes default a surprise to the market, therefore there exists a default intensity process. The model specifications are as follows.

Let X be the cash balances of the firm, normalized by the value of the money market account, with the following dynamics:

$$dX_t = \sigma dW_t,$$

and $X_0 = x > 0$.

Define $g(t) = \sup\{s \leq t : X_s = 0\}$, and the random time g(t) corresponds to the last time (before t) that cash balances hit zero. Let

$$\tau_\alpha = \inf\{t > 0 : t - g(t) \geq \frac{\alpha^2}{2}, \texttt{ where } X_s < 0 \texttt{ for } s \in (g(t-), t)\}$$

be the first time that the firm's cash balances have continued to be negative for at least $\alpha^2/2$ units of time.

Default occurs the first time after τ_α that the cash balances double in magnitude. Define the default time τ to be:

$$\tau := \inf\{t > \tau_\alpha : X_t = 2X_{\tau_\alpha}\}.$$

The market does not see the firm's cash balances. Instead, until the firm has had prolonged negative cash balances for a certain time, that is, until random time τ_α, the market only knows when the firm has positive cash balances or when it has negative or zero cash balances, and whether the cash balances are above or below the default threshold $2X_{\tau_\alpha}$. Define a new process:

$$Y_t = \begin{cases} X_t & t < \tau_\alpha \\ 2X_{\tau_\alpha} - X_t & t \geq \tau_\alpha \end{cases}$$

And $\tau = \inf\{t \geq \tau_\alpha : Y_t = 0\}$. Let

$$\texttt{sign}(x) = \begin{cases} 1 & x > 0 \\ -1 & x \leq 0 \end{cases}$$

Set $\mathcal{G}_t = \sigma\{\texttt{sign}(Y_s); s \leq t\}$ be the complete and right continuous filtration which is the information set observed by the market.

Within this model, Cétin, Jarrow and Protter (2004) [16] show that τ is equivalent to

$$\tau = \inf\{t > 0 : \triangle M_t \geq \alpha\},$$

where $M_t = E[\frac{2}{\sqrt{\pi Y_t}}]$ is the Azéma's martingale with respect to $\mathcal{G}_t$. Therefore, τ is a jump time of Azéma martingale, hence it is totally inaccessible in the filtration $\mathcal{G}_t$. Define $N_t = 1_{\{t \geq \tau\}}$. By Doob-Meyer decomposition, there exists a continuous, increasing, and predictable process A_t, such that $N_t - A_t$ is a $\mathcal{G}$ martingale which has only one jump at τ of size 1. The process A_t is called $\mathcal{G}$ compensator of τ.

More recently, Guo, Jarrow and Zeng (2009) [47] rigorously define incomplete information considered in all these models with the notion of "delayed filtrations". They characterize two distinct types of delayed information, continuous and discrete: the first generated by a time change of filtrations and the second by finitely many marked point process.

2.4 Statistical Model

2.4.1 Linear Discriminant Analysis and Altman's Z-Socre

Altman (1968) [4] assesses the quality of financial ratio analysis and developes Z-score for corporate bankruptcy prediction. A set of financial and economic ratios are used for constructing the Z-score where a multiple discriminant statistical methodology is employed.

Linear Discriminant Analysis (LDA) is a statistical technique used to classify an observation into one of several pre-defined groups dependent upon the observation's characteristics. It is used primarily to classify and to predict when the dependent variable is qualitative, for instance, gender or bankruptcy.

The first step of LDA is to establish explicit group classification, e.g., bankruptcy and non-bankruptcy. After establishing the groups, a training data set of objects with labels and individual characteristics is collected for the classification purpose. LDA attempts to derive a linear combination of these characteristics (discriminant function) to best discriminates between groups. In the distress/bankruptcy prediction case, the discriminant function is of the form:

$$Z = v_1 X_1 + v_2 X_2 + ... + v_n X_n,$$

which transforms the individual characteristics variables $X_i, i = 1, 2, ..., p$ to a single discriminant score (z-value) used to classify the object. The LDA computes the discriminant coefficients $v_i, i = 1, 2, ..., p$ under the assumption that the class prior distribution has a multivariate Gaussian density.

After the initial groups are defined and firms selected, five financial ratio variables are served as the prediction variables X_i, i=1,2,...,5. They are:

- X_1: Working Capital/Total Assets;
- X_2: Retained Earnings/Total Assets;
- X_3: Earnings Before Interest and Taxes/Total Assets;
- X_4: Market Value of Equity/Book Value of Total Debt;
- X_5: Sales/Total Assets.

The final discriminant function is as follows:

$$Z = 0.012X_1 + 0.014X_2 + 0.033X_3 + 0.006X_4 + 0.999X_5. \tag{2.10}$$

An updated z-score model, ZETA model, (see Altman (2001) [6]) is created by Altman, Haldeman and Narayanan (1977) [7] to accommodate the change in the size, financial profile/accounting

of business failures. An adjustment to the basic data is made to accommodate important accounting modifications, such as: capitalization of leases, reserves, minority interests and other liabilities on the balance sheet, captive finance companies and other non-consolidated subsidiaries, goodwill and intangibles and capitalized research and development costs, capitalized interest and certain other deferred charges. A seven-variable model is constructed which not only classifies the test sample well, but also proves the most reliable in various validation procedures. The new discriminant variables are X_1: return on assets measured by the earnings before interest and taxes/total assets; X_2: stability of earnings measured by a normalized measure of the standard error of estimate around a five to ten-year trend in X_1; X_3: debt service measured by the interest coverage ratio (EBIT/total interests payment); X_4: cumulative profitability measured by the firm's retained earnings/total assets; X_5: liquidity measured by current ratio; X_6: capitalization measure by common equity/total capital; X_7: size measure by total asset.

2.4.2 Logistic Regression Model

Ohlson (1980) [69] proposes a logistic regression model for bankruptcy prediction: Let X_i denote a vector of predictors for the ith observation; let β be a vector of unknown parameters, and let $P(X_i, \beta)$ denote the probability of bankruptcy for any given $X_i, i = 1, 2, ..., p$ and β. For any specific function P, the maximum likelihood estimates of $\beta_1, \beta_2, ..., \beta_p$ are obtained by solving:

$$\max_{\beta} l(\beta),$$

where $l(\beta) = \sum_{i \in S_1} \log P(X_i, \beta) + \sum_{i \in S_2} \log(1 - P(X_i))$. S_1 is the set of bankrupt firms and S_2 is the set of non-bankrupt firms. P is chosen to be the logistic function $P = (1 + exp\{-\beta^T X_i\})$.

Ohlson identifies four basic factors as being statistically significant in affecting the probability of failure within one year. These are: 1) the size of the company: log(total assets/GNP price-level index); 2) a measure of the financial structure: total liabilities/total assets, working capital/total assets, current liabilities/current assets, an indicator variable equal one if total liabilities exceeds total assets; 3) a measure of performance: net income/total assets; 4) a measure of current liquidity: funds provided by operations/total liability.

Lau (1987) [59] uses a logistic predictive models by assuming all firms will enter one of J=5 states. He uses five financial states instead of failing/non-failing dichotomy to approximate the continuum of corporate financial health, and estimate the probabilities that a firm will enter each of the five financial states. The five financial states are 0: financial stability; 1: omitting or reducing dividend payments; 2: technical default and default on loan payments; 3: protection under Chapter 11 of the bankruptcy act; 4: bankruptcy and liquidation.

Each firm's default likelihood is predicted by K=10 explanatory variables $x_1, ..., x_{10}$; the probability p_j that a given firm will eventually enter state j is computed as

$$p_j = exp(Z_j) / \sum_{j=1}^{5} exp(Z_j),$$

where $Z_j = b_{j,1}x_1 + b_{j,2}x_2 + ... + b_{j,10}x_{10}$, j=1,2...,5.

Ten explanatory variables are: 1. loan restrictive terms; 2. industry normalized debt-to-equity ratio; 3. working-capital flow to total debt ratio; 4. trend of common stock prices; 5. industry normalized operating expenses to sales ratio; 6. distribution of common stock dividends; 7. liquidation of operating assets; 8. trend of capital expenditure; 9. trend of working-capital flow; 10. omission or reduction of dividend payments.

2.4.3 Empirical Study

Altman (1968) [4] uses a training data set consisting of 66 manufacturing corporations from 1946-1965 with 33 firms in each of the two groups: bankruptcy and non-bankruptcy. Non-bankruptcy group consists of a paired sample of manufacturing firms randomly stratified by industry and by size. Financial statement data is derived from one annual reporting period prior to bankruptcy.

The LDA model claims to be very accurate in classifying 95 % of the total sample correctly. The Type I error is proved to be only 6% while the Type II error is even lower at 3%. Altman (2000) [5] conducts three subsequent tests where he examines 86 distressed companies from 1969-1975, 110 bankrupts from 1976-1995 and 120 from 1997-1999. He finds that the Z-Score model, using a cutoff score of 2.675, is between 82% and 94% accurate.

The major findings of Altman (1968, 2000) [4] [5] are:

- the discriminant model correctly predicts bankruptcy in the sample, and is accurate in out of sample test;
- bankruptcy can be accurately predicted up to two years prior to actual failure with the accuracy diminishing rapidly after the second year;

The limitation of Altman (1968, 2000) [4] [5] is the sample data set used to construct the model, which consists of only 66 manufacturing firms from 1946-1965.

Using logistic regression model, Ohlson (1980) [69] finds that:

- previous studies appear to have overstated the predictive power of models developed and tested, especially if one employs predictors derived from statements which are released after the date of bankruptcy;
- a significant improvement in goodness-of-fit is more likely to occur by augmenting the accounting-based data with market price data.

The data set used in Ohlson (1980) [69] includes 105 bankruptcy firms and 2058 non-bankruptcy firms from 1970 to 1976.

Lau (1987) [59] uses a data set consisting of 350 firms in state 0, and 20, 15, 10, 5 firms in states 1, 2, 3, 4, respectively, selected from year 1971-1980. Lau finds that:

- instead of having a dichotomic classification of bankruptcy/non-bankruptcy state, financial distress can be modeled using five different financial states.

Recently, Campbell, Hilscher and Szilagyi (2008) [15] also use the logistic regression approach to construct their default prediction model. They adopt a broader failure indicator: bankruptcy filing, delisted for financial reasons, or receives a D rating. (January 1963- December 2003), and the whole sample universe consists of 800 bankruptcies, 1,600 failures and predictor variables for 1.7 million firm months.

In their model, market value version of explanatory variables of Shumway (2001), Chava and Jarrow (2004), lagged information and liquidity measure are added to the logistic regression: $P_{t-1}(Y_{it} = 1) = \frac{1}{1+exp\{-\alpha-\beta x_{i,t-1}\}}$. They conclude that:

- their model has greater explanatory power than Shumway (2001) and Chava and Jarrow (2004), and have meaningful empirical advantages over the bankruptcy risk scores proposed by Altman (1968) and Ohlson (1980).
- failure risk cannot be adequately summarized by a measure of KMV's distance to default.

More recently, Davydenko (2009) [29] studies whether default is triggered by low market asset values (economic distress) or by liquidity shortage (financial distress) with the logistic regression model. He studies the determinants of the timing of default using a sample of speculative-grade bond issuers with observed market values of equity, bonds, and bank debt. Default events include bankruptcy fillings, bond payment omissions, and successful distressed bond exchange offers. He uses bond, loan, and equity prices with the debt structure data to estimate monthly market values of total debt and equity. The list of defaults is based on the May 2006 issue of Moody's Default Risk Service (DRS) database. The final sample consists of 806 high-yield (junk) firms including 213 firms that defaulted 225 times and 593 non-default between 1997 and 2005. He finds that:

- low firm market asset value/debt (V/D) is the most important single determinant of default. Lack of liquidity (quick ratio) provides some additional predicative power, particularly when additional equity/debt finance is costly. These effects dominate Z-Score, other accounting-based predictors;
- at the best average default barrier $L_B = V/D = 0.72$, 30% of firms are misclassified (Type 1 or 2 error);
- other firm-specific variables also help to explain default barrier: Firm liquidity, volatility, liquidation value etc. Liquidity factor is more important when external financing (equity/debt cost) is high.

The logistic regression variables Davydenko includes are:

- market variables: Market Assets/Face Debt, Asset Volatility, log (Total Asset), Industry Distress, Risk-free rate;
- accounting variables: Quick Ratio, Cash Shortage, Cash Short if restricted/unrestricted, Coupon Rate, Payout Ratio, Debt Maturity, Replacement Cost/TA, Normalized number of issues of Bonds, Log(Time at Risk).

Chapter 3

Multi-Firm Default

Defaults of firms in the economy will cluster if there are common factors that affect individual firms' default risk. What factors cause the economy-wide default rate to change over time, and why does it vary as much as it does? Individual companies are linked via industry-specific and general economic conditions. As a result, the default events of companies are correlated. Besides a common dependence on the economic and industry environment, a sudden large variation in the credit risk of one issuer will propagate to other issuers and cause spread jumps across firms, which is called the contagious effect.

One of the most important open questions in financial markets is modeling correlation of multi-firm default. An understanding of this correlation is important to banks, traders, and other operatives in corporate-debt markets with respect for portfolio management. It is also of interests to regulators. The Basel Committee on Banking Supervision has repeatedly urged a move towards a system in which the capital adequacy requirements of banks are based on the overall credit-riskiness of the banks' aggregate portfolios.

3.1 Structural Model

3.1.1 Asset Correlation

Zhou (2001) [83] provides a basic theoretical model for default correlations based on first-passage-times. He modeled the pair-wise default correlation between two firms by the joint probability of a two-dimensional stochastic process passing a boundary. Two assumptions are made in his paper:

Assumption 1: Let V_1 and V_2 denote the total asset values of firm 1 and firm 2. The dynamics of V_1 and V_2 are given by the following vector stochastic process:

$$\begin{pmatrix} d\ln(V_1) \\ d\ln(V_2) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \Omega \begin{pmatrix} dz_1 \\ dz_2 \end{pmatrix},$$

where μ_1 and μ_2 are constant drift terms, z_1 and z_2 are two independent standard Brownian motions, and Ω is a constant 2 x 2 matrix:

$$\Omega * \Omega' = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}.$$

The coefficient, $\rho = Corr(d\ln(V_1), d\ln(V_2))$, is the correlation between the movements in the asset values of the two firms.

Assumption 2: The default of a firm is triggered by a decline in its asset value. For each firm i, there exists a time-dependent value $C_i(t)$ such that the firm continues to operate and meets its contractual obligations as long as $V_i(t) > C_i(t)$. However, if $V_i(t)$ falls below threshold level $C_i(t)$, the firm defaults. Assume $C_i(t) = e^{\lambda_i t} K_i$.

Under this setting, Zhou shows that the default rate of an individual firm is

$$P(D_i(t) = 1) = \Phi(-\frac{Z_i}{\sqrt{t}} - \frac{\mu_i - \lambda_i}{\sigma_i}\sqrt{t}) + e^{\frac{2(\lambda_i - \mu_i)Z_i}{\sigma_i}}\Phi(-\frac{Z_i}{\sqrt{t}} + \frac{\mu_i - \lambda_i}{\sigma_i}\sqrt{t}),$$

where $Z_i = \frac{\ln(V_{i,0}/K_i)}{\sigma_i}$. And the joint default probability can be calculated based on Planar Diffusions theory.

Hull, Predescu and White (2006) [49] extended the structural form approach and used correlated asset processes. It assumes the underlying Geometric Brownian motion is:

$$dW_t = \alpha_i(t)dB(t) + \sqrt{1 - \alpha_i(t)^2}dM_i(t)$$

where B(t) and $M_i(t)$ are independent brownian motion, and the variable α_i defines how sensitive the default probability of company i is to the common factor B(t). Under this setting, the correlation between the processes followed by the assets of firm i and j is $\alpha_i(t)\alpha_j(t)$ at time t.

3.1.2 Empirical Study

Zhou (2001) [83] uses 1970-1993 default rates for various rating categories from Moody's default studies, the default correlations between rating categories was calculated from Z-score implied by historical default rates. He finds that:

- default correlations are generally very small over short horizons, and first increase and then slowly decrease with time;
- high credit quality implies a low default correlation over typical horizons;
- the time of peak default correlation depends on the credit quality of the underlying firms.

On the other hand, D*ü*llmann, Scheicher and Schmieder (2007) [36] find that:

- there exists substantial time variation in asset correlations both for the market model and the sector model. The median inferred asset correlation in the market model ranges from 4% to 16% during sample period from 1996 to 2004. For the sector model, the inferred intra-sector asset correlations are only about 2 percentage points higher than the inferred asset correlations in the market model and exhibit a similar time pattern;
- upturns in the stock market tend to increase asset correlations and downturns in stock market tend to decrease asset correlations;
- a finer sector classification may lead to more precise sector correlation estimates;
- it is desirable to apply a single-factor model with borrower(size)-dependent correlations than a multi-factor model with sector-dependent correlations.

The data set used in D*ü*llmann, Scheicher and Schmieder (2007) contains monthly time series of Moody's KMV asset values for around 2000 European firms from 1996 to 2004.

In the market model(single factor), correlation is modeled by a single common risk factor defined as the returns of the aggregate portfolio of all firms in the sample:

$$X_i = \rho_i Y + \sqrt{1-\rho_i^2}\epsilon_i.$$

In the sector model(multi-factor), systematic risk factors are linked to industry sectors. The intra-sector asset correlations are defined as the median of the (individual) asset correlations in ever y sector. The inter-sector correlations are calculated as the sample correlations between the time series of two sector-index returns. Asset correlations in the sector model fundamentally differ from correlations in the market model that they are always aggregated at sector level, whereas the market model contains individual pairwise correlations with the market index:

$$X_i = \rho_{s(i)} Y_{s(i)} + \sqrt{1-\rho_{s(i)}^2}\epsilon_i,$$

where X_i is the normalized asset return of borrower i in sector $s(i)$. Y is the market asset return index and $Y_{s(i)}$ are asset return index of sector $s(i)$. ϵ_i are the independent idiosyncratic risk component.

Zhang, Zhu and Lee (2008) [81] find that

- default-implied asset correlations range from 5% to around 30% in the sample data set, depending on the grouping of the underlying borrowers;
- borrowers with higher ratings, or lower EDF values, tend to have higher asset correlations. This supports the intuition that larger firms tend to have larger systematic risk, and tend to be more closely correlated with the performance of the economy than do smaller firms;
- asset correlations manifest themselves more in default clustering during periods of deteriorating credit quality, which reflects the cyclical nature of defaults in an economy.

The data set is from Moody's KMV historical default database, consisting of 16,268 publicly traded U.S. non-financial firms from 1981 to 2006. This period has several economic cycles and high default episodes. The probabilities of default come from the Moody's KMV EDF. Because it is more difficult to track default history for small firms and missing defaults can lead to significant downward bias in realized default correlation, small firms are excluded in the study, which reduces the sample size to 5,040 firms with 524,891 firm-month observations.

Default-implied asset correlation is calculated as follows:

- group borrowers in groups i, i=1,...,n. Assume all borrowers in the same group have the same default probability. $\hat{p}_i$ and $\hat{p}_j$ are the estimated default probabilities for borrowers in group i and j. $\hat{p}_{ij}$ is the estimated joint default frequency between borrowers in group i and borrower in group j;
- for each time period, a realization of p_i, p_j, and p_{ij} can be observed. Assume each realization across time is from the same distribution;
- $\hat{p}_i$ and $\hat{p}_{ij}$ are estimated by:

$$p_i = \sum_t w_i^t \frac{D_i^t}{N_i^t}, and$$

$$p_{ij} = \sum_t w_{ij}^t \frac{D_i^t D_j^t}{N_i^t N_j^t},$$

where w_i^t, w_{ij}^t are the weight associated to the relative importance of the sample in given year t. D_t are the number of defaults in a given year t, and N_t are the total number of borrowers at the beginning of year t.

After the estimated joint default probability is calculated, the asset correlation can be implied from the bivariate normal distribution under the structural model setting.

Akhavein, Kocagil and Neugebauer (2005) [3] study intra-industry and inter-industry asset correlations following the Fitch industry categorizations. They find that:

- inter-industry asset correlations are relatively smaller than intra-industry asset correlations;
- intra-industry as well as inter-industry asset correlations vary across different industries;
- there are notable empirical differences between intra-industry and inter-industry asset correlations.

The data set includes 66,740 yearly observations comprising of 7,886 firms and 1,039 defaults from January 1970 to December 2004. Issuers are categorized by eight broad letter rating categories (AAA, AA, A, BBB, BB, B, CCC, Default). Issuers are classified into one of Fitch's twenty-five industry categorizations. Asset correlation is derived using three methodologies:

- asset correlations based on default correlation,
- asset correlations based on rating transitions,
- asset correlations based on equity market information.

3.2 Reduced Form Model

3.2.1 Conditional Independence

The reduced form model has the convenient features that conditioning on the state variables, defaults become independent events, and default correlation arises due to the common influence of these state variables.

Suppose that the default times $\tau_1, ...\tau_k$ of k given names have respective intensity process $\lambda_1, ..., \lambda_k$, and are doubly stochastic. Conditioned on the information in the driving filtration $\mathcal{F}_t = \sigma(X_s, 0 \leq s \leq t)$, where X_s represents macro-economical and firm-specific state variables that determine the respective intensities, the event times $\tau_1, ..., \tau_k$ are independent. Note the only source of correlation of the default times is via the correlation of the intensities.

The conditional independence assumption makes the computation of the joint distribution of default time simple. Consider the joint survival event $\tau_1 \geq t_1, ..., \tau_k \geq t_k$, without loss of generality, assuming $t_1 \leq t_2 \leq ... \leq t_k$. For any current time $t < t_1$,

$$P(\tau_1 \geq t_1, ..., \tau_k \geq t_k | \mathcal{F}_t) = E_t(e^{-\int_t^{t_k} \mu(s)ds}),$$

where $\mu(t) = \sum_{i, t_i > t} \lambda_i(t)$.

3.2.2 Counterparty Risk and Contagion Default

Bernanke (1983) [10] discusses and studies the contagious effects of the bank failures in the context the Great Depression and concludes that the financial industry collapse of the early 1930's has a propagate effect on the economy. Lang and Stulz (1992) [58] study the intra-industry effect of bankruptcy announcement and conclude that there is a significant contagious effect for highly leveled industries and a competitive effect for highly concentrated industries with low leverage.

Jarrow and Yu (2001) [50] argue that a default intensity that depends linearly on a set of smoothly varying macroeconomic variables is unlikely to account for the clustering of defaults around an economic recession. They introduce a concept of counter-party risk into their default risk modeling, where each firm has a unique counter-party structure that arises from its relation with other firms in the economy in which the default of a firm's counter-party might affect its own default probability.

Suppose that the filtration is generated collectively by the information contained in the state variables X_t and the default processes $N_t^i, i = 1, 2, ..., I$:

$$\mathcal{F}_t = \mathcal{F}_t^X \vee \mathcal{F}_t^1 \vee ... \vee \mathcal{F}_t^I,$$

where

$$\mathcal{F}_t^X = \sigma(X_s, 0 \leq s \leq t), \mathcal{F}_t^i = \sigma(N_s^i, 0 \leq s \leq t),$$

are the filtrations generated by X_t and N_t^i, respectively.

Define a new filtration $\mathcal{G}_t^i$ as follows. First, let the filtration generated by the default processes of all firms other than that of the ith be denoted by

$$\mathcal{F}_t^{-i} = \mathcal{F}_t^1 \vee ... \vee \mathcal{F}_t^{i-1} \vee \mathcal{F}_t^{i+1} \vee ... \vee \mathcal{F}_t^I.$$

Then, let

$$\mathcal{G}_t^i = \mathcal{F}_t^i \vee \mathcal{F}_{T^*}^X \vee \mathcal{F}_{T^*}^{-i}.$$

$\mathcal{G}_0^i$ contains complete information on the state variables and the default processes of all firms other than that of the ith, all the way up to time T^*.

Let τ^i denote the first jump time of N^i. Then the conditional and unconditional distributions of τ^i are given by

$$P(\tau^i > t | \mathcal{G}_0^i) = exp(-\int_0^t \lambda_s^i ds), t \in [0, T^*],$$

$$P(\tau^i > t) = E[exp(-\int_0^t \lambda_s^i ds)], t \in [0, T^*].$$

The intensity function λ_t^i is $\mathcal{G}_0^i$ measurable and the realized history of the process λ_t^i define a non-homogeneous poisson process N^i. Under this counter-party framework, the joint distribution of several first jump times are no longer independent conditioning on the complete history of the state variables.

To account for the cluster effect in large portfolios of defaultable securities Davis and Lo (2001) [27] introduce a contagion model , where they use binary random variables for the default state of each firm. These random variables are a function of a common set of independent identically distributed binary random variables. More specifically, let $Z_i, i = 1, ..., n$ be random variables such that $Z_i = 1$ if firm i defaults and $Z_i = 0$ otherwise. Then the number of defaults

is

$$N = Z_1 + Z_2 + ... + Z_n.$$

In Davis and Lo's model, the value of Z_i is determined as follows. For i=1,...,n and j=1,..., n with $j \neq i$ let $X_i, Y_{i,j}$ be independent Bernoulli random variables,

$$Z_i = X_i + (1 - X_i) * (1 - \prod_{j \neq i}(1 - X_j Y_{ji})).$$

The idea behind the contagion model is that firm may default due to endogenous factor ($X_i = 1$), or may be infected by default of firm j ($X_j = 1$). The random variable Y_{ji} determines whether infection takes place or not.

Let $F(n, k, p, q)$ denote the probability mass distribution of the random variable N, i.e. $F(n, k, p, q) = P(N = k)$, its distribution function F is given by

$$F(n, k, p, q) = C_k^n \alpha_{nk}^{pq},$$

where

$$\alpha^{pq} = p^k (1-p)^{n-k} (1-q)^{k(n-k)} + \sum_{i=1}^{k-1} C_i^k p^i (1-p)^{n-i} * (1 - (1-q)^i)^{k-i} (1-q)^{i(n-k)}.$$

Collin-Dufresne, Goldstein and Helwege (2003) [20] propose a reduced-form model where jump-to-default is priced which generates a market-wide jump in credit spreads. While the framework is consistent with a counter-party risk interpretation it is interpreted as updating of beliefs due to an unexpected event.

Their contagion-risk model is based on updating of beliefs and goes as follows: N firms with default intensities that are equal to λ_i^H if the economy is in state H, or λ_i^L if the economy is in state L. State H and state L represent the economic conditions. Investors do not know whether the economy is in state H or L, but form a prior $p^H(t) = P(H|\mathcal{F}_t)$, where $\mathcal{F}_t$ represents all information investors have available at time t. The $\mathcal{F}_t$-default intensity is:

$$\bar{\lambda_i}(t) = p^H(t)\lambda_i^H + (1 - p^H(t))\lambda_i^L.$$

Applying Bayes' rule, the updating process for $p^H(t)$ can be calculated as:

$$dp^H(t) = p^H(t)(1 - p^H(t)) \sum_{i=1}^{N} \frac{\lambda_i^H - \lambda_i^L}{\bar{\lambda}_i(t)} (d\mathbf{1}_{\tau_i \leq t} - \bar{\lambda}_i(t)\mathbf{1}_{\tau_i > t} dt).$$

To capture the contagion effect, Azizpour and Giesecke (2008) [9] propose a top-down reduced form model. The default events are modeled as a realization of a marked point process (T_n, D_n), where T_n represents a date with at least one default incidence and D_n is the number of defaults at T_n. T_n is assumed to be a non-explosive counting process N_t with intensity λ_t, where:

$$d\lambda_t = \kappa(c - \lambda_t)dt + \sigma\sqrt{\lambda_t} dW_t + \delta dL_t.$$

Here $\kappa > 0, c > 0, \sigma \geq 0$ and $\lambda_0 > 0$. D_n is drawn from a fixed distribution independently, and

L is a response jump process given by

$$L = \sum_{n=1}^{N} l(D_n).$$

Suppose the observation filtration $\mathcal{G}$ is generated by the market point process (T_n, D_n) and a process X of explanatory covariates. Take X to be of the form $X_t = u(Y_t, t)$, with u: $R \times R_+ \to R$ and Y is a standard-Brownian motion independent of the D_n. The covariates need not bear information about the Brownian motion W driving λ, in which case W must be filtered from the observations.

Papageorgiou and Sircar (2008) [70] propose a hybrid of top-down and bottom-up reduced-form approaches to model the default intensity. They first group the original set of firms into homogeneous groups according to their 5-years CDS spread, where they assume that firms in the same group share the same default intensity process. They then incorporate a fast mean-reverting stochastic volatility in the default intensity to help explain the fat tail of portfolio loss distribution. Their model goes as follows:

- the default intensity process shared by all firms of the homogeneous group i, $\lambda^i = (\lambda_t^i)_{t \geq 0}$ is given by
$$\lambda_t^i = X_t^i + c_i Z_y, i = 1, ..., k.$$
the idiosyncratic factors $X^1, ..., X^k$ are independent from each other and independent of the systematic factor X. The process $X^1, ..., X^k$ and Z are non-negative almost surely;

- fix a group i (i=1,...,k). The dynamics of X^i and Z are given by:
$$dX_t^i = \alpha_i(\bar{x}_i - X_t^i)dt + f_i(Y_t^i)\sqrt{X_t^i}dW_t^i,$$
$$dY_t^i = \frac{1}{\epsilon}X_t^i(\bar{y}_i - Y_t^i)dt + \frac{v_i\sqrt{2}}{\sqrt{\epsilon}\sqrt{X_t^i}}dW_t^{Y^i},$$
$$dZ_t = \alpha_z(\bar{z} - Z_t)dt + \sigma_z\sqrt{Z_t}dW_t^Z,$$
where the Wiener process W^Z is independent of W^i and W^{Y^i}.

They find that by separating firms of similar credit risk and modeling their aggregate credit risk within their homogeneous group, more accurate estimates of their default probabilities can be obtained than assuming the existence of a unique idiosyncratic risk.

3.2.3 Frailty Model

Duffie, Eckner, Horel and Saita (2008) [30] develop a new model of corporate default intensities in the presence of a time-varying latent frailty factors. The frailty model is built as follows: for a given firm i, let U_{it} be the observable default-prediction covariates that are specific to firm i. Also let V_t be observable macro-economic covariates. Finally, define an unobservable macro-economic covariate Y_t with "frailty" influence on default events, so that

$$\lambda_{it} = exp(\alpha + < \beta, V_{it} > + < \gamma, U_{it} > + Y_t).$$

Here Y_t is assumed to follow Ornstein-Uhlenbeck (OU) process such that: $dY_t = -\kappa Y_t dt + dB_t$. Because Y_t is not observable, its posterior probability distribution is estimated from the available

information set $\mathcal{F}_t$ which includes the prior history of the observable covariates $(U_s, V_s) : s \leq t$ and also includes previous observations of the periods of survival and times D of default of all firms.

Suppose that the frailty process Y is independent of the observable covariate process $W = (U, V)$, with respect to the econometrician's limited filtration $(\mathcal{F}_t)$, the likelihood for the default data set is then:

$$\begin{aligned} L(\eta, \theta | W, D) &= \int L(\eta, \theta | W, y, D) p_Y(y) dy \\ &= L(\eta | W) \int L(\theta | W, y, D) p_Y(y) dy, \end{aligned} \tag{3.1}$$

where η is the parameter describing the dynamics of the covariate process W and $\theta = (\beta, \gamma, \kappa)$. $p_Y(.)$ is the unconditional probability density of the path of the unobserved frailty process Y. $L(\theta | W, y, D)$ is calculated in Duffie, Saita, and Wang (2007) [33].

3.2.4 Empirical Study

Lucas (1995) [64] uses data from Moody's Investor Service covering twenty-four years of data from 1970 through 1993 to compute the default correlation between companies in rating classes. He concludes that:

- it is important to note that historical statistics describe only observed or realized phenomena, not the true underlying correlation. Also, since default rates are small in the higher rating categories and small over short time periods, the limited length of the time series becomes a problem;
- the empirical analysis depends upon the true underlying default probability for each rating category remaining the same over time. But the evidence reveals that default probability and correlation are not stationary. It is hard to determine whether historical fluctuations in default rates are caused by default correlation or simply by changes in default probability;
- it is very hard to say anything about default correlation among and between specific industries. Calculated intra-industry default correlation will be lower if a particular industry enjoyed generally good business results in the time period studied relative to its inherent risk. Calculated default correlation would be higher if the specific industry experienced unfavorable business conditions sometime during the time period. The true exante correlation is unobservable, and efforts need to be put to infer default correlation from other channels such as stock prices;
- default correlation is found to be generally low and it decreases as ratings increase; default correlation is also found generally to increase with time at first, and then to decrease with time.

Using month-end bond price for the period January 1973-March 1998 from Warga Fixed Income Database, Collin-Dufresne, Goldstein and Helwege (2003) [20] find:

- credit events of large firms generate a market wide increase in credit spreads and a significant "flight-to-quality" response in the Treasury market;
- the risk premium associated with jump-to-default risk for a typical investment grade firm has an upper bound of a few basis points per year, but the risk premium for contagion-risk may be considerably larger.

Defaults of firms in the economy will cluster if there are common factors that affect individual firms' default risk. The questions are: what factors cause the economy-wide default rate to change over time, and why does it vary as much as it does? Das, Freed and Geng (2006) [24] study the underlying determinants of default probabilities and provide a statistical model to model the time-variation in default probabilities and default correlations. The data set consists of issuer-level default probabilities of almost all U.S. public non-financial firms for the period 1987-2000. In this paper, default probability is calculated under the structural model of Merton:

$$PD = 1 - \phi(DTD),$$

where $\phi(.)$ is the standard normal pdf, and DTD is the "distance to default":

$$DTD = \frac{\ln(V/D) + (\mu_v - 0.5\sigma_v^2)T}{\sigma_v\sqrt{T}}.$$

In practice, DTD is often computed (as in Moody's KMV) as:

$$DTD = \frac{1 - D/V}{\sigma_v\sqrt{T}}.$$

This model identifies two of the three primary determinants of the default probability for an individual firm:

- debt ratio D/V;
- firm asset volatility σ_v;
- firm asset return μ_v.

Differentiate the Distance to Default, one has:

$$\Delta(DTD) \approx \frac{-1}{\sigma_v\sqrt{T}}\Delta(D/V) - \frac{1 - D/V}{\sigma_v^2 T}\Delta(\sigma_v\sqrt{T}),$$

indicating that at low debt levels, the DTD is more sensitive to changes in volatility than to changes in debt levels. Over the sample period, for high grade firms, the impact of the change in volatility is 3.7 times the impact of the change in debt value, and for low grade firms, it is 1.8 times.

The correlation between PDs of two firms will depend on the correlation between the underlying determinants of the default probabilities. The authors classify the sample by rating as well as by sub-periods (1/87-6/90, 7/90-12/93, 1/94-6/97, and 7/97-10/00). The median correlation between pairs of firms is calculated. Several observations in their paper are:

- the median correlation between firms for the default determinants is positive. For high grade and medium grade firms, the highest correlation is between the volatilities of firm;
- differences between the rating classes is driven mostly by differences in correlation between volatilities;
- over time for each rating class, correlations between firm returns are the most stable, while the correlations between volatilities are the highest in Period I and IV when the level of volatility itself is high across economy.

Das, Freed and Geng (2006) [24] find that:

- across bond rating category, the default probability increases by more than 100% between times of low default risk and times of high default risk;
- when default probabilities rise, so do their correlations. Correlations rise from close to zero to levels of 17-38% that are much higher than correlations between asset returns;
- the joint credit risk varies because both default probabilities and default correlations vary with economic conditions. Market-wide volatility plays an important role in determining the time-variation in joint default risk. Clustering of defaults occurs during times of high volatility because both default probabilities and correlation between defaults increase;
- both default probabilities and default correlations are related to firm asset volatility. Structural models of joint default risk should explicitly model correlations of volatilities, along with the correlations between firm returns;
- both the default intensity process of individual firms and the correlation between the intensity processes between pairs of firms can be modeled with regimes based on an aggregate economy-wide default level.

Das, Duffie, Kapadia and Saita (2007) [23] test the intensity-based approach under which firms' default times are correlated only as implied by the correlation of factors determining their default intensities. They show that:

- the data do not support the joint hypothesis of well-specified default intensities and the conditional dependence assumption;
- the conditional dependence assumption is violated in the presence of contagion or frailty;
- some evidence of default clustering exceeding that implied by the doubly stochastic model with the given intensities is found.

Azizpour and Giesecke (2008) [9] use default timing data from Moody's Default Risk Service which provides detailed issue and issuer information on industry, rating, date and type of default, and other items. The sample period is January 1970 to October 2006. An issuer is included in the data set if it is not a sovereign and has a senior rating, which is an issuer-level rating generated by Moody's from ratings of particular debt obligations. As of October 2006, the data set includes a total of 6048 firms, of which 3215 are investment grade rated issuers. A "default" is a credit event in any of the following Moody's default categories: (1) A missed or delayed disbursement of interest or principal, including delayed payments made within a grace period; (2) Bankruptcy (Section 77, Chapter 10, Chapter 11, Chapter 7, Prepackaged Chapter 11), administration, legal receivership, or other legal blocks to the timely payment of interest or principal; (3) A distressed exchange occurs where: (i) the issuer offers debt holders a new security or package of securities that amount to a diminished financial obligation; or (ii) the exchange had the apparent purpose of helping the borrower avoid default. A repeated default by the same issuer is included in the set of events if it was not within a year of the initial event and the issuer's rating was raised above Caa after the initial default.

They explore the role of contagion, i.e., the default of a firm having a direct impact on the conditional default rates of the surviving firms, channeled through the complex web of contractual relationships in the economy. They develop filtered maximum likelihood estimators and goodness-of-fit tests for point processes to measure the additional impact of contagion on default rates, over and beyond that due to firms' exposure to observable or unobservable (frailty) risk factors.

The parameter vector to be estimated is $(\theta, \gamma, \upsilon)$, where $\theta = (\kappa, c, \sigma, \delta, \lambda_0, w)$ is the intensity parameters, w represents the parameters of the weight function L, γ is the parameter of the distribution D_n, and υ is the parameter of the distribution of X. The likelihood function for the sample period [0,τ] of the data is given:

$$L_\tau(\theta, \gamma, \upsilon | N_\tau, T, D, X) = f_\tau(N_\tau, T; \theta | D, X) g(D; \gamma) p_\tau(X; \upsilon)$$

where $f_\tau(.|D, X)$ is the conditional density of the event data count N_τ and the event date vector $T = (T_1, ..., T_{N_\tau})$. $g(., \gamma)$ is the probability function of D, and $p_\tau(., \upsilon)$ is the density of the covariates path over $[0, \tau]$. The three terms can be maximized separately to give the full likelihood estimates.

Instead of estimating a parametric model for $g(., \gamma)$, D_n is described by their empirical distribution. The covariate model $p_\tau(., \upsilon)$ is made precise. An equivalent probability measure is developed to evaluate the density $f_\tau(.|D, X)$, which transforms the counting process $(N; \mathcal{F})$ into a standard $\mathcal{F}$-Poisson process. The density is then expressed in terms of a conditional expectation under the equivalent measure of the Radon-Nikodym derivative, given the observations. For U.S. firms during 1970-2006, they find that

- strong evidence that contagion represents a significant additional source of default clustering;
- contagion and frailty phenomena are found to be roughly equally important for explaining the default clustering in the data that is not captured by a traditional doubly stochastic model, in which firms' default times are correlated only as implied by the correlation of observable risk factors determining their default intensities.

Using data set containing 402,434 firm-months of data between January 1979 and March 2004, Duffie, Eckner, Horel and Saita (2008) [30] find that:

- significant evidence exists among U.S. corporates of a common unobserved source of default risk that increases default correlation and extreme portfolio loss risk above and beyond that implied by observable common and correlated macroeconomic and firm-specific sources of default risk;
- the posterior distribution of the frailty variable shows that the expected rate of corporate defaults was much higher in 1989-1990 and 2001-2002, and much lower during the mid-nineties and in 2003-2004, than those implied by an analogous model without frailty;
- an out-of-sample test for data between 1980 and 2003 indicates that a model without frailty significantly underestimates the probability of extreme positive as well as negative events in portfolios of corporate credits, while a model with frailty gives a more accurate assessment of credit risk on the portfolio level.

3.3 Copula Model

In statistics, a copula is used as a general way of formulating a multivariate distribution in such a way that various general types of dependence can be represented. More specifically, the dependence between real-valued random variables $X_1, ...X_n$ is completely described by their joint distribution function

$$F(x_1, ..., x_n) = P(X_1 \leq x_1, ..., X_n \leq x_n)$$

The idea of splitting F into a part which describes the dependence structure and a part which describe the marginal behavior leads to the concept of copula. The following lemma is the fundamental key for copula.

Lemma 1. *Let X be a random variable with distribution function F. For* $\alpha \in (0,1)$ *Define:*

$$F^{-1}(\alpha) = \inf\{x|F(x) \geq \alpha\}$$

1.For any standard-uniformly distributed $U \sim U(0,1)$*, we have* $F^{-1}(U) \sim F$*.*
2.If F is continuous, then the random variable F(X) is uniformly distributed, i.e. $F(X) \sim U(0,1)$*.*

Definition 2. *Suppose a random vector* $X = (X_1, ..., X_n)$ *have continuous marginal distributions* $F_1, ..., F_n$*, the copula C of the random vector* $(X_1, ..., X_n)$ *or the multivariate distribution F is defined as follows:*

$$\begin{aligned} F(x_1, ..., x_n) &= P(X_1 \leq x_1, ..., X_n \leq x_n) \\ &= P(F_1(X_1) \leq F_1(x_1), ..., F_n(X_n) \leq F_n(x_n)) \\ &\equiv C(F_1(x_1), ..., F_n(x_n)) \end{aligned}$$

C is a distribution function of multivariate uniform distribution.

The following is the well-known Sklar's Theorem:

Theorem 3. *Suppose* $X_1, ..., X_d$ *are random variables with continuous marginal distribution function* $F_1, ..., F_d$ *and joint distribution function F, then exists a unique copula C such that for all* $x = (x_1, ..., x_d) \in R^d$*:*

$$F(x_1, ..., x_d) = C(F_1(x_1), ..., F_d(x_d)).$$

Conversely, given any marginal distribution function $F_1, ..., F_d$ *and copula C, F defined above is a d-variate distribution function with marginal* $F_1, ...F_d$

With this recipe, multivariate distributions can be constructed by adding to the marginal distributions a copula function with a pre-described interdependence structure. On the other hand, we can also construct a copula C from any joint distribution with continuous marginal distribution $F_1, ..., F_d$.

Theorem 4. *If* $(X_1, ..., X_n)$ *has copula C and* $T_1, ..., T_n$ *are increasing continuous functions, then* $(T_1(X_1), ..., T_n(X_n))$ *also has copula C.*

3.3.1 Copula Functions

Li (2000) [62] uses a bivariate normal copula with correlation γ to capture the dependence structure between two survival times with marginal distributions F_A and F_B:

$$P(T_A < s, T_B < t) = \Phi_2(\Phi^{-1}(F_A(s)), \Phi^{-1}(F_B(t)), \gamma).$$

Suppose that the one year default probability for security A and B are $P(Z \leq Z_A)$ and $P(Z \leq Z_B)$, where Z is a standard normal random variable and Z_A and Z_B are predetermined asset levels. If ρ is the asset correlation, the joint default probability for credit A and B is calculated

as follows:

$$P(Z \leq Z_A, Z \leq Z_B) = \int_{-\infty}^{Z_A} \int_{-\infty}^{Z_B} \phi_2(x, y|\rho) dxdy = \Phi_2(Z_A, Z_B, \rho),$$

where $\phi_2(x, y|\rho)$ is the standard bivariate normal density function with correlation coefficient ρ, and Φ_2 is the bivariate cumulative normal distribution function.

Li (2000)[62] concludes that CreditMetrics uses a bivariate normal copula function with the asset correlation as the correlation parameter in the copula function. Thus, to generate survival times of two credit risks, one can use a bivariate normal copula function with correlation parameter equal to the CreditMetrics asset correlation. Note that this correlation parameter is not the correlation coefficient between the two survival times which is in general much smaller than the asset correlation. Also, using asset correlation, one can construct high dimensional normal copula functions to model the credit portfolio of any size.

Giesecke (2003) [42] uses a bivariate exponential copula to model the correlated default. The idea is to let defaults of firms be driven by firm-specific as well as economy-wide shock events. Suppose there are poisson processes N_1, N_2 and N with respective intensities $\lambda_1, \lambda_2, and \lambda$. Λ_i is the idiosyncratic shock intensity of firm i, while λ is the intensity of a macro-economic shock affecting both firms simultaneously. Define the default time τ_i of firm i by:

$$\tau_i = \inf\{t \geq 0 : N_i(t) + N(t) > 0\}.$$

Then the survive time distribution is:

$$s_i(t) = P(\tau_i > t) = P[N_i(t) + N(t) = 0] = e^{-(\lambda_i + \lambda)t}.$$

The joint survival probability is:

$$\begin{aligned} s(t,u) &= P(\tau_1 \geq t, \tau_2 \geq u) \\ &= P(N_1(t) = 0, N_2(u) = 0, N(t \vee u) = 0) \\ &= e^{-(\lambda_1+\lambda)t-(\lambda_2+\lambda)u+\lambda(t\wedge u)} \\ &= s_1(t)s_2(t)\min(e^{\lambda t}, e^{\lambda u}). \end{aligned} \tag{3.2}$$

There exists a unique function C^τ such that:

$$s(t,u) = C^\tau(s_1(t), s_2(u)).$$

Let $\theta_i = \frac{\lambda}{\lambda_i+\lambda}$, one has:

$$C^\tau(u,v) = s(s_1^{-1}(u), s_2^{-1}(v)) = \min(vu^{1-\theta_1}, uv^{1-\theta_2}).$$

The parameter vector $\theta = (\theta_1, \theta_2)$ controls the degree of dependence between the default times. If the firms default independently, then $\theta_1 = \theta_2 = 0$ and one gets $C^\tau = uv$, the product copula. If the firms are perfectly positively correlated, then $\theta_1 = \theta_2 = 1$ and $C^\tau(u,v) = u \wedge v$. Note that because of the Frechet upper bound of copula, i.e., $uv \leq C^\tau \leq u \wedge v$ the default can only be positively related.

Other copulas that have been used in the literatures are:

Student's t copula: Let $T_{\rho,v}$ be the standardized multivariate Student's t distribution with ν degrees of freedom and correlation matrix ρ. The multi-variate Student's t copula is then

defined as follows:

$$C(\mu_1, ..., \mu_n; \rho, \nu) = T_{\rho,\nu}(t_\nu^{-1}(\mu_1), ..., t_\nu^{-1}(\mu_n)),$$

where t_ν^{-1} is the inverse of the cumulative distribution function of a univariate Student's t distribution with ν degree of freedom. See Daul, Giorgi, Lindskog and McNeil (2003) [26].

Gumbel copula:

$$C(\mu_1, ..., \mu_n) = exp[-(\sum_{i=1}^{n}(-ln\mu_i)^\alpha)^{1/\alpha}],$$

where α is the parameter determining the tail dependence of the distribution. See Gumbel (1960) [45].

Clayton copula:

$$C(\mu_1, ..., \mu_n) = [\sum_{i=1}^{n}\mu_i^{-\alpha} - n + 1]^{-1/\alpha},$$

where α is the parameter determining the tail dependence of the distribution. See Clayton [18].

3.3.2 Empirical Study

Das and Geng (2004)[25] undertake an empirical examination of the joint stochastic process of default risk over the period 1987 to 2000 using copulas functions. The data set comprises 600 issuers classified into six rating classes tracked by Moody from 1987 to 2000. Moody has monthly PDs for each issuer. The default intensity is calculated as: $\lambda_{it} = -ln(1 - PD_{it})$. They find that:

- the skewed double-exponential distribution is the best choice for the marginal distribution of each issuer's hazard rate process, and combines well the normal, Gumbel, Clayton and Student's t copulas in the joint dependence relationship;
- a regime-switching model of the intensity process better represents the properties of correlated default than a jump model.

Embrechts (2009) [37] has his comments on application of copula models: "The meta-Gaussian model becomes enormously popular, in the end causing problems because the market too strongly believed in it. The Gaussian-copula is the worst invention ever for credit risk management. Currently various more realistic versions have been worked out replacing the Gauss-copula by a finite mixture of Gaussians or even an infinite mixture leading to elliptical copulas. Unfortunately, most of these models are inherently static and fail to incorporate the dynamics of markets, especially distress. The 2007 subprime crisis around the pricing of CDO is a very clear proof of this."

According to Nassim Nicholas Taleb, "People got very excited about the Gaussian copula because of its mathematical elegance, but the thing never worked. Co-association between securities is not measurable using correlation; in other words because past history is not predictive of the future, anything that relies on correlation is charlatanism."

3.4 Graphical Model

3.4.1 Voter Model

Giesecke and Weber (2004, 2006) [43] [44] introduce a Voter model in which firms interact with their business partners in a lattice-type economy.

Suppose the economy consists of a collection F of firms, an arbitrary firm $i \in F$ interacts with a collection $N(i) \subseteq F|\{i\}$ of business partners (neighbors). Define the neighborhood $N(i)$ of a firm i by $N(i) = j : |j - i| = 1$. A firm's interaction with its neighbors is symmetric, meaning $j \in N(i) \Leftarrow i \in N(j)$. The dimension d of the lattice can be interpreted as the degree of complexity of the business partner network.

They associate each firm $i \in Z^d$ with a state variable $\xi(i) \in 0, 1$, which describes the firm's liquidity state with respect to the interaction with its business partners N(i). $\xi(i) = 1$ means that firm i's liquidity reserves are stressed and might be insufficient to honor due obligations. $\xi(i) = 0$ means that firm i is financially healthy and honors its obligations to business partners timely. Assume that a transition of firm i from liquidity state $\xi(i)$ to state $1-\xi(i)$ is an unpredictable Poisson event. After a unit-exponential waiting time, a firm i adopts the liquidity state of one of its 2d business partners which is chosen with uniform probability $\frac{1}{2d}$. The evolution of firms' liquidity state over time is modeled by a continuous-time Markov process $(\eta_t)_{t\geq 0}$ with state space $X = 0, 1^{\mathbb{Z}^d}$ and transition rate c given by

$$c_{(i,\xi)} = \begin{cases} \frac{1}{2d}\sum_{j\in N(i)} \xi(j) & \texttt{if } \xi(i) = 0, \\ \frac{1}{2d}\sum_{j\in N(i)}[1 - \xi(j)], & \texttt{if } \xi(i) = 1. \end{cases}$$

The continuous-time Markov process describing the joint evolution of firm's liquidity state converges as time approaches infinity. The macro-economic business environment in the steady state is described by a random vector with given distribution and the influence of both contagion and cyclical effects are modeled on the firms.

3.4.2 Markov Random Field

Filiz, Guo, Morton and Sturmfels (2008) [39] propose a graphical Ising model for correlated defaults. The model has an intuitive graphic structure and in a very special homogenous case, the loss distribution for one period version is just a summation of binomial random variables. It can represent any given marginal distribution for single firms and pairwise correlation matrix.

Ising Model: Give an undirected graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, where each node of the graph corresponds to a binary random variable X_i. The joint probability distribution of the random variable $X := (X_1, ...X_V)$ is given by

$$p(x_1, x_2, x_3, ..x_v; \eta) = \frac{1}{Z} * exp\{(\sum_{s\in V} \eta_s x_s + \sum_{(s,t)\in E} \eta_{s,t} x_s x_t)\}$$

where $\eta_s, \eta_{s,t} \in R$ are parameters and Z is the normalization constant known.

They show that:

Theorem 5. *Assume any given set of statistics P_i,$P_{u,v}$ from some probability distribution on M binary random variables. Then, there exists a unique set of parameters η_i, η_{uv} such that the single and double node marginals implied in the Ising model match P_i,$P_{u,v}$.*

To analyze the model, they impose further structure on the Ising model. Take $M = N + S$, where nodes 1,2,...,N represent individual firms and N+1,...,N+S represent individual industry sectors, the joint probability distribution for $(X_1, ..., X_N)$ is defined as:

$$P(X_1 = x_1, ..., X_N = x_N) := \sum_{s \in \{0,1\}^S} Q(X_1 = x_1, ..., X_N = x_N, X_{N+1} = s_1, ..., X_{N+S} = s_{N+S}).$$

It is assumed that each firm belongs to a particular sector $j = 1, 2, ..., S$ such that firm nodes 1,...N are partitioned into S subsets with N_j elements, i.e. $N = \sum_{j=1}^{S} N_j$. The parameters η_i and η_{uv} as follows:

- each firm node i has a single edge connecting to its respective sector node;
- for any particular sector node j, all firm nodes that connect to it have the same node weight η_{F_j} and same edge weight η_{FS_j};
- sector nodes are allowed to have different node weights η_{S_j} and they connect to each other with different edge weights $\eta_{N+u,N+v}$.

Under this graphical structure, the probability distribution for $(X_1, ..., X_N)$ becomes:

$$\begin{aligned} P(X_1 = x_1, ..., X_N = x_N) &= \frac{1}{Z_S} \sum_{s \in \{0,1\}^S} exp(\sum_{j=1}^{S} s_j \eta_{S_j} + s_j n_j \eta_{FS_j} + n_j \eta_{F_j}) \\ & \quad exp(\sum_{(u,v): u,v \in 1,...,S} s_u s_v \eta_{N+u,N+v}), \end{aligned} \tag{3.3}$$

where $n_j := \sum_{i:\eta_{i,N+j}} x_i$ is the number of defaulting firms in sector j, and Z_S is a normalization constant.

Part II

Given the recent financial crisis, stressed by the incorrect assessment of the default risk and correlation imbedded in corporate loans, bonds and derivatives, the study of firms' credit risk has assumed increased importance to the financial industry and regulatory agency. We conduct our research on corporate default under the reduced form framework and focus on predicting firm default using historical default data and firms' fundamental, market performance and macro-economic data.

For purpose of statistical prediction of default, it is crucial to estimate and to compare the performance of different models in order to choose the best one that has the least prediction errors. However, in the literature, it is an ad-hoc process to select the predictors and different models often have different predictors. Some literature even get contradicting results. For example, Shumway (2001) [78] finds that half of the accounting ratios used in Altman (1968) [4] and Zmijewski (1984) [84] are poor predictors and several previously neglected market-driven variables (firm's market size, past stock return and the idiosyncratic standard deviation of stock return) are strongly related to default/bankruptcy probability. Chava, Jarrow (2004) also conclude that Shumway (2001) model has superior forecasting performance as opposed to Altman (1968) and Zmijewski (1984) and that accounting variables add little predictive power when market variables are already included.

These ad-hoc approaches and contradicting results in the literatures motivate us to study the predictors of corporate defaults. Given a large set of corporate default predictors $x_1, x_2, ..., x_p$, the problems raised here are:

- with a large number of predictors, how can we determine a smaller subset of default predictors that have strong effects in predicting default?
- having specified the refined set of predictors, how can we build a default predicting model that is able to take into accounts the default forecasters' utility?

In this dissertation research, the two problems are studied. We conduct empirical studies on statistical default predicting models with different predictors and find an optimal set of default predictors. We then present a new methodology to predict and to classify firm defaults.

Chapter 4

Statistical Model Selection

In this chapter, we study the predictors of U.S corporate default by forward stepwise and lasso model selection methods under the logistic regression setting. Out of 30 candidate default predictors that have been used in the default-predicting literature, we identify a set of eight default predictors that have strong effects in predicting default using the U.S corporate default data from 1984-2009. We compare the eight default predictors' predicting effect over the past three major economic recessions and find that the recession in early 1990 and the recent sub-prime mortgage crisis share some common default characteristics, while the recession in 2000 is different from the other two.

4.1 Data

The empirical analysis is based on corporate default data from Mergent Fixed Income Securities Database (FISD), COMPUSTAT and CRSP database. FISD is a comprehensive database of publicly-offered U.S. corporate debt. FISD contains issue details and default status from 1984 to present, and provides details on debt issues and the issuers. COMPUSTAT and CRSP database contains U.S firms' fundamental and market performance data.

FISD records default as occurring when a debt issue either violates a bond covenant (technical default), misses an interest or principal payment, or files for bankruptcy (either Chapter 11 reorganization or Chapter 7 liquidation). The default data set which our analysis is based on has 4238 recorded issue defaults from 1984 to 2009, of which 3619 (85%) defaults file for bankruptcy, 284 (6.7%) miss an interest payment, 37 (0.8%) miss principal payment, 19 (0.4%) violate bond covenant and 279 (7.1%) miss a default type.

We construct and clean our firm level default data by implementing the following three steps: first, we combine and integrate the defaulted issues with the same ISSUER_ID into firm level defaults in FISD database at the time of default. We treat consecutive defaults of one firm within a one year time window as one default event and use the first default date within one year as the default event date of that year.[1] This leaves us with total of 1644 defaults at the firm level. Secondly, since firms often default together with their wholly-owned subsidiaries, we manually analyze default cases within the same corporate family within one year and use the parent company's consolidated financial information to analyze the default for the corporate family. This leaves us with a total of 1424 defaults at the firm level, Thirdly, by mapping six digit cusip numbers, we link the default firm data in FISD with stock data from CRSP and then to accounting data from COMPUSTAT by the linking table of permco to gvkey. This leaves us

[1] Davydenko (2009) [29] uses a two year time window; Lando (2010) [57] uses a one month time window.

with 657 defaulted cases from 1984 to 2009. Finally, because of the too big too fail nature of the financial industry and lack of business operating measures, we restrict our default sample to non-financial firms with four digits SIC code less than 6000 or greater than 7000 and get the default sample consists of 461 defaulted firms with 483 default cases from 1984 to 2009 during which we can get CRSP and COMPUSTAT data for at least one year before defaults.

Our controlled data sample is limited to U.S junk (speculative grade) firms (rated BB+ and below by Moody's and S&P), since firms with investment grades rarely default.[2] The control sample includes 844 non-financial junk firms that did not default during 1984 to 2009. Thus, the overall sample consists of 1305 junk firms, including 461 firms that defaulted 483 times, with total 5979 firm-year observation and an average of 230 firm observations per year from January 1984 to December 2009. Table 4.1 and Figure 4.1 display the number of default firms by year.

Years	Number of junk firms	Number of defaults	Default rate
1984	2	0	0.00%
1985	3	0	0.00%
1986	6	2	33.33%
1987	5	2	40.00%
1988	8	3	37.50%
1989	26	6	23.08%
1990	39	12	30.77%
1991	51	18	35.29%
1992	88	22	25.00%
1993	110	11	10.00%
1994	145	4	2.76%
1995	192	9	4.69%
1996	274	17	6.20%
1997	324	14	4.32%
1998	386	22	5.70%
1999	453	39	8.61%
2000	452	50	11.06%
2001	450	80	17.78%
2002	436	48	11.01%
2003	423	31	7.33%
2004	425	16	3.76%
2005	393	14	3.56%
2006	358	5	1.40%
2007	331	6	1.81%
2008	303	18	5.94%
2009	296	34	11.49%
1984-2009	5979	483	8.08%

Table 4.1: Number of Defaults by Year

[2]Collin-Dufresne, Goldstein, and Helwege (2003) [20] notice that since 1937, only four firms with investment-grades from Moody's have defaulted on their bonds.

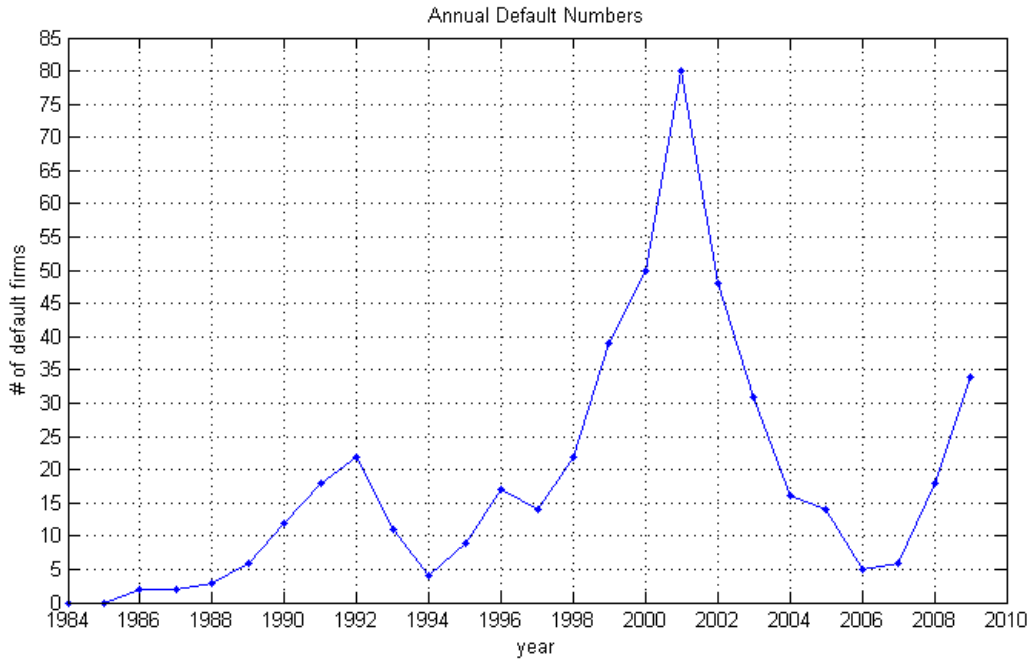

Figure 4.1: Annual Default Numbers

4.2 Econometric Model

In this section, we describe the logistic regression settings and the forward stepwise and lasso model selection methods that we use.

4.2.1 Regression Specification

We follow the empirical method in default literature and use the logistic regression model to predict one year ahead default probabilities, conditioned on the value of default predictors, both firm-specific and common. We denote:

- the firm-specific default predictors of firm i at the end of year t: x_{it}^k, $k = 1, 2, ...p$, where p is the number of firm-specific predictors;
- the common default predictors at the end of year t: z_t^l, $l = 1, 2, ...q$, where q is the number of common predictors;
- the default event of firm i at the end of year t: y_{it} with $y_{it} = 1$ if firm i defaults at year t, $y_{it} = 0$ if it does not;
- the default probability of firm i during year t: p_{it},

with $t = 1, 2, ..., T$.

With the logistic regression specification, the default probability is modeled as a linear function of the default predictors x_{it}^k, i.e.,

$$\texttt{logit}(p_{i(t+1)}) = \log(\frac{p_{i(t+1)}}{1 - p_{i(t+1)}}) = \alpha + \beta_1 x_{it}^1 + ... + \beta_p x_{it}^p + \gamma_1 z_t^1 + ... + \gamma_q z_t^q, \tag{4.1}$$

where $\alpha, \beta_i, \gamma_j$ are parameters to be estimated, $i = 1, 2, ...p, j = 1, 2, ..., q$.

This is a panel data regression. Provided that the default predictors are constructed as financial ratios and returns and that distressed non-financial firms are homogeneous, we assume that the default predictor coefficients $\beta_k, \gamma_l, k = 1, 2, ...p, l = 1, 2..., q$ are constant both cross-section and cross-time, as in the majority of empirical default prediction literature.

Note that (4.1) implies

$$p_{i(t+1)} = \frac{1}{1 + e^{\alpha + \sum_{k=1}^{p} \beta_k x_{it}^k + \sum_{l=1}^{q} \gamma_l z_t^l}}.$$

Assume that firms default independently conditioned on the default predictors x_{it}^k and z_t^l. The likelihood of all default events is

$$\prod_{t=1}^{T} \prod_{i=1}^{N_t} p_{it}^{y_{it}} (1 - p_{it})^{(1-y_{it})}. \tag{4.2}$$

Note that we only look at the non-investment grade firms, the total number of firm observations N_t at each year varies, and the panel data is unbalanced.

Taking log yields the log likelihood:

$$\begin{aligned} l(\alpha, \beta, \gamma) &= \sum_{t=1}^{T} \sum_{i=1}^{N_t} (y_{it} \log p_{it} + (1 - y_{it}) \log(1 - p_{it})) \\ &= \sum_{t=1}^{T} \sum_{i=1}^{N_t} (1 - y_{it})(\alpha + \sum_{k=1}^{p} \beta_k x_{it}^k + \sum_{l=1}^{q} \gamma_l x_t^l) \\ &\quad - \log(1 + e^{\alpha + \sum_{k=1}^{p} \beta_k x_{it}^k + \sum_{l=1}^{q} \gamma_l z_t^l}). \end{aligned} \tag{4.3}$$

To minimize the convex negative log-likelihood function $-l(\alpha, \beta, \gamma)$, one can take the first order derivative with respect to α, β, γ and solve the score equation using Newton-Raphson method.

4.2.2 Model Selection

For the purpose of predicting firm default using historical default data, it is important to estimate and to compare the performance of different models in order to choose the best one that has the least predicting errors in the sense of negative log-likelihood in the logistic regression settings. With a large number of predictors, we often would like to determine a smaller subset that exhibit the strongest predicting effects. We introduce two model selection methods for this purpose.

Forward Stepwise Selection

Recall that all subset selection finds the subset out of all $p+q$ predictors that gives the smallest prediction error by enumerating all possible subsets. However, as $p+q$ gets as large as 30, searching through all possible subsets becomes very slow and even infeasible. Instead, one needs to seek a good path through them, and forward stepwise selection is one such approach.

Forward stepwise selection starts with the intercept, and then sequentially adds into the model the predictor that most reduces the prediction error. In our logistic regression case,

the prediction error is characterized by the out of sample negative loglikeliood estimated by 10 fold cross-validation.

Forward stepwise selection is a greedy algorithm, producing a nested sequence of models. It is computationally more attractive comparing to all-subset selection, however it often leads to locally optimal solutions rather than globally optimal solutions.

The Lasso

The Lasso variable selection technique is proposed by Tibshirani (1996)[79]. The Lasso minimizes the loss function subject to the sum of the absolute value of the coefficients being less than a constant. The lasso does not focus on subsets but defines a continuous shrinking operation. Because of the nature of this L1 norm constraint it tends to produce some coefficients that are exactly 0 and hence gives interpretable models. Moreover, the Lasso enjoys some of the favorable properties of both subset selection and ridge regression in the sense that it produces interpretable models like subset selection and exhibits the stability of ridge regression.

The lasso estimate for logistic regression is defined:

$$\begin{aligned} \hat{\alpha}^{lasso}, \hat{\beta}^{lasso}, \hat{\gamma}^{lasso} &= \arg\min_{\alpha,\beta,\gamma} -l(\alpha,\beta,\gamma) \\ &\quad \text{subject to: } \sum_{k=1}^{p} |\beta_k| + \sum_{l=1}^{q} |\gamma_l| \leq t, \end{aligned} \tag{4.4}$$

where $t \geq 0$ is a tuning parameter, and $l(\alpha, \beta, \gamma)$ is in (4.3). Like in the forward stepwise selection, the tuning parameter is adaptively chose to minimize predicted negative loglikeliood estimated by 10 fold cross-validation.

The Lasso problem can also be written in the equivalent Lagrangian form

$$\hat{\alpha}^{lasso}, \hat{\beta}^{lasso}, \hat{\gamma}^{lasso} = \arg\min_{\alpha,\beta,\gamma} -l(\alpha,\beta,\gamma) + \lambda(\sum_{k=1}^{p} |\beta_k| + \sum_{l=1}^{q} |\gamma_l|).$$

Computation of the solution to equation (4.4) is a quadratic programming problem with linear inequality constraints. We adopt the coordinate descent algorithm developed by Friedman, Hastie and Tibshirani (2010) [40] to solve the lasso estimation problem.

4.3 Empirical Results

Under the logistic model specifications introduced in Section 4.2, we select and construct 30 candidate default predictors that have been often used in the default-predicting literature (Table A.2). Out of 30 candidate default predictors that have been used in the default-predicting literature, we identify a set of eight default predictors that have strong effects in predicting default using the U.S corporate default data from 1984-2009. We also compare the eight default predictors' predicting effect over the past three major economic recessions and find that the recession in early 1990 and the recent sub-prime mortgage crisis share some common default characteristics, while the recession in 2000 is different from the other two.

In order to construct default predictors, we combine yearly accounting data from COMPUSTAT with monthly equity market data from CRSP. Abbreviation and data source descriptions are in Table A.1. The 30 candidate predictors are grouped into nine categories that capture default causal factors at the individual firm and macro-economy level: Size, Capital structure, Growth, Profitability, Debt Coverage, Liquidity, Business Operations, Market Performance, and

Macro-Factor. We construct measures of firm size as: log(Sale), log(TA) and RSIZE; firms' capital structure is characterized by: ME/BD, ME/TA, TL/TA and SD/BD; Sale Gth, NI Gth and OM CH represents firm's growth traits; RE/TA, EBIT/TA and NI/TA describe firm's profitability; ICR and NI/TL gauges firm's debt coverage; WC/TA, CR, CH/TA, QR are liquidity measures; INVT/SALE, AR/SALE and SALE/TA depicts business operations; market performance is characterized by SRT and SIGMA; SPRET, T90RET, TSPRD, GDP CH, IPI CH and CPI CH are macro-level factors that are common across firms. Table 4.2 shows the summary statistics for the 30 default predictors.

		Default Firms					Non-Default Firms				
Category	Predictor	Mean	Median	Std.Dev	5%	95%	Mean	Median	Std.Dev	5%	95%
Size	log(Sale)	6.018	6.040	1.757	3.476	8.708	6.879	6.837	1.343	4.755	9.136
	log(TA)	6.264	6.103	1.562	3.654	8.945	7.109	7.003	1.192	5.316	9.258
	RSIZE(e-07)	1.599	0.444	3.841	0.052	6.768	2.208	0.903	4.014	0.170	9.049
Capital Structure	ME/BD	1.471	0.158	14.396	0.012	3.010	16.527	1.310	304.822	0.121	12.487
	ME/TA	0.308	0.104	0.696	0.010	1.152	0.759	0.525	0.934	0.065	2.163
	TL/TA	1.099	0.930	0.684	0.547	2.031	0.733	0.689	0.310	0.397	1.212
	SD/BD	0.352	0.155	0.385	0	1	0.104	0.031	0.191	0	0.511
Growth	SALE Gth	0.232	0	1.429	-0.426	1.542	0.256	0.077	2.666	-0.223	0.841
	NI Gth	-7.635	-0.913	27.998	-33.408	1.016	-0.871	0.092	25.623	-6.671	4.539
	OM CH	-0.432	-0.015	26.995	-0.387	0.496	0.263	0	12.844	-0.136	0.145
Profitability	RE/TA	-0.991	-0.306	3.567	-3.992	0.138	-0.080	0.029	0.532	-0.906	0.387
	EBIT/TA	-0.087	-0.017	0.296	-0.526	0.098	0.058	0.066	0.101	-0.090	0.177
	NI/TA	-0.290	-0.137	0.502	-1.098	0.032	-0.014	0.016	0.198	-0.226	0.109
Debt Coverage	ICR	-1.246	-0.246	5.432	-7.607	2.278	4.338	1.829	38.026	-2.297	12.569
	NI/TL	-0.228	-0.145	0.299	-0.759	0.040	0.001	0.023	0.481	-0.274	0.223
Liquidity	WC/TA	-0.167	-0.006	0.564	-1.222	0.393	0.137	0.122	0.193	-0.082	0.445
	CR	1.311	0.962	1.772	0.124	3.302	1.912	1.645	1.571	0.605	3.927
	CH/TA	0.064	0.031	0.094	0.001	0.232	0.067	0.038	0.082	0.002	0.227
	QR	0.912	0.614	1.439	0.085	2.695	1.395	1.094	1.508	0.345	3.250
Business Operations	INVT/SALE	0.107	0.072	0.163	0	0.312	0.117	0.083	1.114	0	0.273
	AR/SALE	0.158	0.133	0.156	0.009	0.363	0.232	0.139	5.980	0.014	0.314
	SALE/TA	1.174	0.960	1.088	0.140	2.927	1.062	0.876	0.845	0.204	2.632
Market Performance	SRT	-0.463	-0.625	0.640	-0.964	0.409	0.153	0.033	0.835	-0.738	1.328
	SIGMA	0.248	0.220	0.147	0.096	0.510	0.150	0.126	0.127	0.055	0.302
Macro-Factor	SPRET	0.038	0.035	0.210	-0.385	0.310	0.076	0.090	0.195	-0.234	0.310
	T90RET	0.049	0.051	0.019	0.013	0.084	0.041	0.048	0.018	0.012	0.062
	TSPRD	0.031	0.058	0.079	-0.128	0.130	0.026	0.037	0.073	-0.128	0.130
	GDP CH	0.029	0.036	0.016	0	0.048	0.030	0.031	0.013	0	0.048
	IPI CH	0.022	0.033	0.032	-0.034	0.059	0.027	0.032	0.029	-0.034	0.072
	CPI CH	0.027	0.027	0.013	0.001	0.047	0.025	0.025	0.010	0.016	0.041

Table 4.2: Summary Statistics

4.3.1 Model Selection Results

To identify a subset of default predictors from the 30 candidates that has the best predicting effect, we use the forward stepwise and lasso subset model selection method. Forward stepwise selection is a discrete selection procedure during which a predictor is either added or dropped from the model, and lasso continuous shrinks the coefficients and retains predictors that have the strongest effects. The loss function is a negative loglikeliood and we use 10 fold cross-validation to estimate out of sample negative loglikeliood. The whole data from 1984 to 2009 is used to run the model selection procedure. Figure 4.2 shows the 10 fold cross-validation model selection procedure. Forward stepwise selects 15 predictors as the best predicting model, while lasso identifies 9 predictors as the best model.

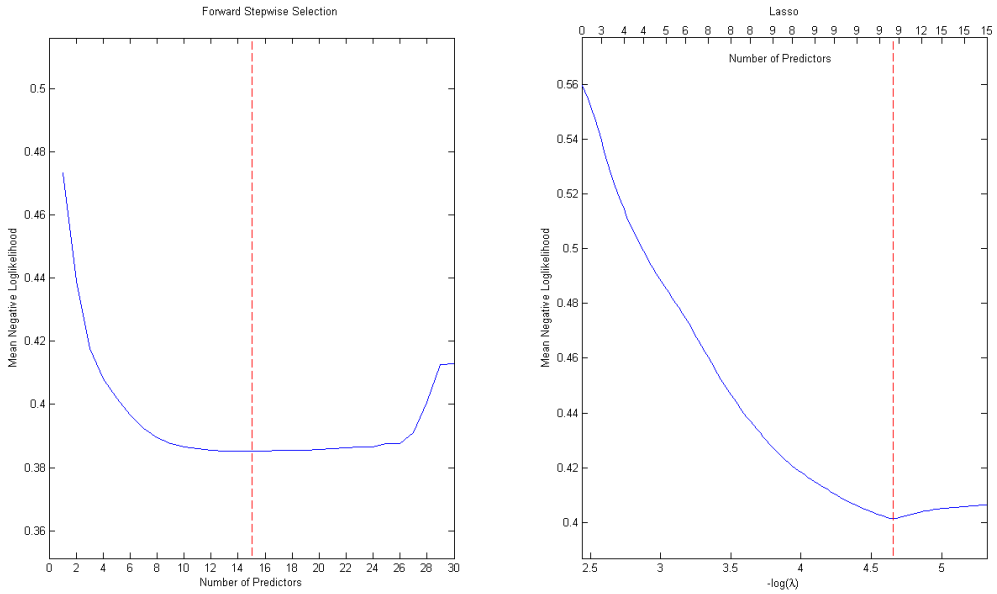

Figure 4.2: 10 Fold Cross-Validation: Forward Stepwise and Lasso

Because post-model selection statistical inference is misguided due to extra uncertainty introduced during the model selection step, we use bootstrap to deal with the issue. The standard errors of regression coefficients of the best models are estimated by bootstrapping 1000 samples and are shown in the parentheses. The regression results are in Table 4.3. We find that:

- forward stepwise and lasso identify predictors in Size, Capital Structure, Profitability, Liquidity, Market Performance and Macro-Factor, while there are no predictors selected in Debt Coverage and Business Operations. Forward stepwise selects 15 predictors as the best predicting model, while lasso identifies 9 predictors. However, 8 out the 9 predictors that lasso selects are also in forward stepwise selection result. They are: log(TA), ME/TA, SD/BD, EBIT/TA, WC/TA, SRT, SIGMA and T90RET;
- due to the shrinkage effect of lasso, all the coefficients from lasso model selection are smaller than their counter-parts in forward stepwise model selection, except for SD/BD. In forward stepwise selection, SD/BD has a coefficient of 1.062, while in lasso SD/BD has a coefficient of 1.206. The unexpected increase in the coefficient of SD/BD after shrinkage may indicate its strong effect in predicting default;
- Comparing the results of the two model selection, it seems the Lasso yields a more concise and stable model. The standard errors of the coefficients of the 9 predictors estimated by bootstrapping from lasso are uniformly smaller than the ones from forward stepwise selection, the standard errors of the other 21 predictors not selected by lasso are all very small. The coefficients of all the 9 predictors that lasso identifies have the sign matching the economical intuition. However, for the forward stepwise selection, the signs of RSIZE and GDP CH are counter-intuitive, which might be caused by multi-collinearity with other identified predictors.

Figure 4.3 and Table 4.4 shows the sequences of the predictors that are added during the model selection procedure. For forward stepwise selection, predictors that are added in the

earlier stage decrease the negative log-likelihood loss more compared with ones that are added later. For lasso, as reducing the L1 penalty λ, predictors that are added in the earlier stage have better negative log-likelihood - L1 norm penalty trade-off. For both of the two model selection procedures, predictors that are closer to the top of the selection sequences indicate stronger effects in predicting default.

Comparing the two sequences, we find that: a). WC/TA from liquidity measure are ranked 2nd and 1st in both procedures. This confirms its significance in predicting defaults; b). EBIT/TA, SRT, log(TA) and SD/BD are both ranked among the top 6 important predictors; c). The most important macro-level predictor is T90RET, whose value determines the difficulty level of firms' short-term financing; d). bExcept for TL/TA that does not appear in the forward stepwise model, all other 8 predictors selected by lasso are in the forward stepwise's first 9 predictors, which shows a very high consistency in ranking default predictors.

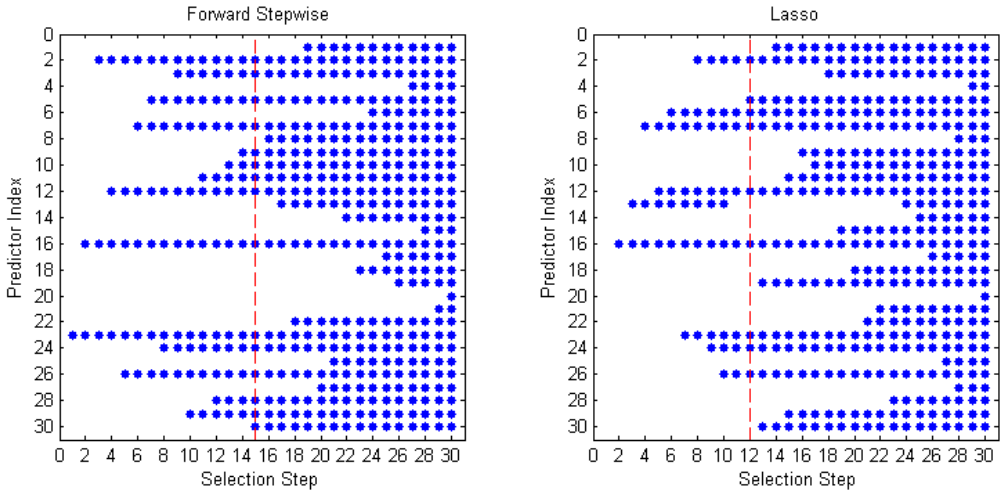

Figure 4.3: Predictor Selection Procedure

4.3.2 Prediction Results

The model selection results of forward stepwise and lasso demonstrate the relative importance of different default predictors in predicting default. It is also of great importance to compare the default prediction result based on the output of the default predicting models. In this section, we compare the 10 fold out of sample prediction results of the two best models selected by forward stepwise and lasso. From Table 4.3, forward stepwise selection achieves smaller out of sample mean negative log-likelihood than lasso. However, for the purpose of predicting defaults, a good measure of prediction accuracy should be based on actual classification result.

In our default prediction scenario, we label a default event as positive and a non-default event as negative. The output from our default prediction models are firm default probabilities. By specifying a threshold value of default probabilities, we classify each firm into

Category	Predictor	Forward Subset Selection	Lasso
Size	log(Sale)		
			(0.019)
	log(TA)	-0.442	-0.220
		(0.060)	(0.044)
	RSIZE(e-07)	0.054	
		(0.016)	(0.000)
Capital Structure	ME/BD		
			(0.000)
	ME/TA	-0.635	-0.179
		(0.147)	(0.030)
	TL/TA		0.412
			(0.138)
	SD/BD	1.062	1.206
		(0.237)	(0.193)
Growth	SALE Gth		
			(0.001)
	NI Gth	-0.002	
		(0.002)	(0.000)
	OM CH	-0.006	
		(0.007)	(0.000)
Profitability	RE/TA	-0.169	
		(0.073)	(0.006)
	EBIT/TA	-2.279	-2.276
		(0.458)	(0.386)
	NI/TA		
			(0.179)
Debt Coverage	ICR		
			(0.000)
	NI/TL		
			(0.003)
Liquidity	WC/TA	-1.253	-0.849
		(0.228)	(0.192)
	CR		
			(0.001)
	CH/TA		
			(0.001)
	QR		
			(0.000)
Business Operations	INVT/SALE		
			(0.003)
	AR/SALE		
			(0.001)
	SALE/TA		
			(0.003)
Market Performance	SRT	-1.157	-0.793
		(0.145)	(0.083)
	SIGMA	1.822	1.556
		(0.449)	(0.365)
Macro-Factor	SPRET		
			(0.000)
	T90RET	21.060	12.008
		(4.078)	(2.627)
	TSPRD		
			(0.024)
	GDP CH	21.101	
		(8.647)	(0.116)
	IPI CH	-14.251	
		(4.095)	(0.006)
	CPI CH	10.807	
		(6.113)	(0.232)
Constant		-1.455	-2.211
		(0.495)	(0.372)
Mean Negative loglikelihood		0.385	0.401
Number of Observations (firm years)		5979	5979

Table 4.3: Model Selection

Selection Sequence	Forward Stepwise	Lasso
1	SRT(Market Performance)	WC/TA(Liquidity)
2	WC/TA(Liquidity)	SD/BD(Capital Structure)
3	log(TA)(Size)	EBIT/TA(Profitability)
4	EBIT/TA(Profitability)	TL/TA(Capital Structure)
5	T90RET(Macro-Factor)	SRT(Market Performance)
6	SD/BD(Capital Structure)	log(TA)(Size)
7	ME/TA(Capital Structure)	SIGMA(Market Performance)
8	SIGMA(Market Performance)	T90RET(Macro-Factor)
9	RSIZE(Size)	ME/TA(Capital Structure)
10	IPI CH(Macro-Factor)	-
11	RE/TA(Profitability)	-
12	GDP CH(Macro-Factor)	-
13	OM CH(Growth)	-
14	NI Gth(Growth)	-
15	CPI CH(Macro-Factor)	-

Table 4.4: Selection Sequence

default (positive) or non-default(negative), and by varying the threshold, we construct different classification criterion from the default probabilities. There are four possible outcomes from the binary classifier: If the outcome from a prediction is positive (default) and the actual value is also positive (default), then it is called a true positive prediction; however if the actual value is negative (non-default) then it is said to be a false positive prediction. Conversely, a true negative has occurred when both the prediction outcome and the actual value are negative, and false negative is when the prediction outcome is negative while the actual value is positive. True positive rate is the ratio between the number of true positive instances and the total number of positive instances in the sample, and false positive rate is the ratio between the number of false positive instances and the total number of negative instances.

We introduce the Receiver Operating Characteristic (ROC) curve, a graphical plot of true positive rate versus false positive rate for a binary classifier system as its discrimination threshold varies. ROC curve provides a comprehensive and visually attractive way to summarize the accuracy of predictions. The area under the ROC curve (AUC) is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. The area under the ROC curve is closely related to the MannCWhitney U test, which tests whether positives are ranked higher than negatives. A random classifier has an area of 0.5, while an ideal one has an area of 1 where the ROC curve passes near the upper left corner of the diagram. Figure 4.4 displays the ROC curve for the forward stepwise and lasso out of sample default prediction models. The AUC under the lasso curve is 0.8812, greater than 0.8795 of AUC under forward stepwise curve, which demonstrates a more precise overall default classification results of lasso model.

The ROC curve in Figure 4.4 can be used to choose the optimal operating point (OOP), a particular threshold at which the classifier gives the best trade-off between the costs of failing to detect positives against the costs of raising false alarms. Assuming that the decision maker is indifferent between the cost of misclassifying positive and negative case (In Section 5 we will discuss and vary this assumption.), the optimal operating point is the point which lies on a 45 degree line closest to the north-west corner (0,1) of the ROC curve. The red circle points in Figure 4.4 display the OOPs of forward stepwise and lasso prediction models. The OOP of the forward stepwise model is 7.6%, and the OOP of the lasso model is 9.2%, which means that if we classify the firms with predicted default probabilities above 7.6% for the forward stepwise model

and above 9.2% for the lasso model as defaulting firms, we can achieve the best classification result. For the forward stepwise model, the false positive rate is 18.85% and the true positive rate is 81.37% at OOP, while for the lasso model, the false positive rate is 15.27% and the true positive rate is 77.64%. We can see that at the OOPs the lasso model is more conservative in making default judgements than the forward stepwise model.

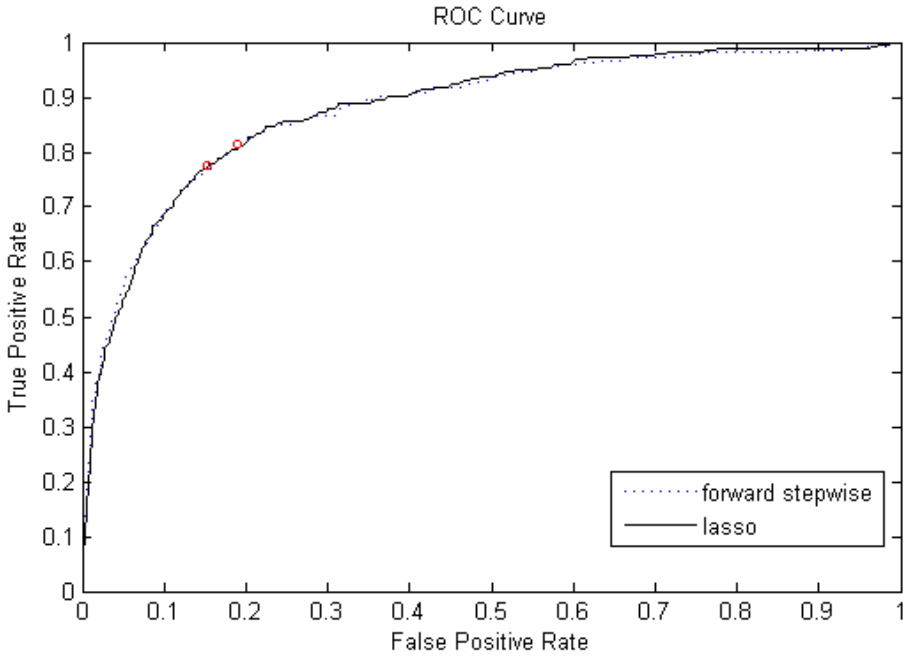

Figure 4.4: Receiver Operating Characteristic Curve

4.3.3 Default In Recessions

In section 4.3.1, we conclude with eight default predictors that are selected and ranked high by both forward stepwise and lasso model selection methods. Because defaults tend to cluster during the economic downturn and are rare events during the economic boom, it is important to investigate and to understand the predicting abilities of the predictors over economic recession periods. During the sample periods 1984-2009, according to National Bureau of Economic Research, U.S has gone through three economic recessions: the early 1990s recession caused by a combination of the debt accumulation of the 1980s, the 1990 oil price shock and the savings and loans crisis; the early 2000s recession accompanied by speculative dot-come bubble and September 11th attack; and the recent sub-prime mortgage crisis lead by national-wide over-leverage through derivatives written by financial institutions. We split our data sample into periods corresponding to the durations of the three recessions, and run the logistic regressions with the eight default predictors as the regressors. While the test conducted here is subject to the post-model selection inference issue, our goal is to identify and to compare predicting effects of default predictors over the three recession periods rather than making inferences on default predictors.

Table 4.5 displays the regression result. There are several observations from Table 4.5: First, the signs of coefficients of the eight predictors match economic intuitions with no variations over the three recession periods. This re-assures the validity of our default prediction

model. Second, all eight predictors enjoy persistent significance in predicting default over recessions, except for SRT, SIGMA in the 1990s recession and log(TA), SD/BD in the sub-prime mortgage crisis. The absolute value of coefficient and t-statistics of log(TA) both decrease over the three recessions suggesting that the size of firm is playing a less and less important role in predicting default. This is consistent what were witnessed during the recent financial crisis: giant corporations like Lehman Brothers, General Motors, Circuit City, etc, collapse together with small firms. SRT and SIGMA are not significant during the 1990s recession and the absolute coefficient values of SRT and SIGMA increase over time, because the financial market is getting more efficient today than it was in the 1990s and default risk is priced in the equity return. The fact that SD/BD is not a significant predictor for the sub-prime mortgage crisis highlights the huge impact of the sub-prime mortgage crisis when firms default due to both the maturity of short-term debt and the interest payment of long term debt.Third, judging from the coefficient values and the significance level of predictors, the 1990s recession and the sub-prime mortgage crisis share some common predictor characteristics, while the 2000s recession is different. ME/TA, EBIT/TA and T90RET explain a greater portion of default risk for the 1990s and sub-prime mortgage recessions than for the 2000s recession. This could be explained by the same over-leverage nature of the 1990s and 2007s recessions. Firms with more levered capital structures(less ME/TA) and less earning abilities are more subjected to default risk, accentuated by the more significant impact of short term interest rate. Moreover, WC/TA is more significant in the 2000s tech bubble recession than in the other two.

Recession Year	log(TA)	ME/TA	SD/BD	EBIT/TA	WC/TA	SRT	SIGMA	T90RET	Const
				Coefficient					
1989-1994	-0.923***	-0.977**	1.125*	-5.259***	-1.283**	-0.302	0.690	48.209***	1.340
1998-2003	-0.235***	-0.225*	1.275***	-1.945***	-1.280***	-1.449***	1.639***	26.536***	-2.667***
2007-2009	-0.194	-1.676**	0.581	-4.062***	-1.398**	-2.561***	3.087**	45.496***	-3.893***
				P-value					
1989-1994	0.000	0.046	0.071	0.007	0.024	0.364	0.538	0.000	0.230
1998-2003	0.000	0.093	0.000	0.000	0.000	0.000	0.003	0.000	0.000
2007-2009	0.160	0.046	0.366	0.001	0.020	0.002	0.053	0.003	0.014
				T-statistics					
1989-1994	-5.685	-1.998	1.805	-2.699	-2.252	-0.908	0.615	5.134	1.199
1998-2003	-3.698	-1.682	4.262	-3.889	-4.350	-7.075	3.020	4.293	-4.530
2007-2009	-1.406	-1.992	0.903	-3.230	-2.328	-3.089	1.936	2.995	-2.449

Table 4.5: Default in Recession

Coefficients marked ***, **, and * are significant at the 1%, 5%, and 10% significance level, respectively. There are 459 firm-year observations with 73 defaults for the period 1989-1994, 2600 firm-year observations with 270 defaults for the period 1998-2003, and 930 firm-year observations with 58 defaults for the period 2007-2009.

Chapter 5

Decision-Based Default Prediction

Why does one care about predicting corporate default? Because one needs to use the predictions to make decisions. For instance, bankers decide whether or not to issue a loan to a firm based on the firm's default probability; distress debt investors select their investments according to their default likelihood judgement; insurers choose on which firm to write a Credit Default Swap (CDS) based on firms' default probabilities; speculators on corporate default events make their bets relying on their default predictions.

For different groups of decision makers, significance of judging defaults accurately varies: for some ones the error of judging a default firm as a non-default, i.e. false negative error, is more significant than judging a non-default firm as a default, i.e. false positive error, while for others the opposite is true, and according to the characteristics of particular firms, decision makers may also put different weights, which could be proportional to the size of the firm for instance, on each default judgement. Therefore, it is very important to take into account the decision maker's loss utility when calibrating and using prediction models. However, in the default literature, decision makers' loss utility is usually ignored with only the negative log-likelihood loss function and maximum likelihood estimation method for model calibration. Maximum likelihood estimators have nice properties such as consistency and efficiency, but it is not the best choice when the underlying probabilistic model assumptions may be wrong and/or there are other objectives more than maximizing the likelihood.

In this chapter, we present a decision-based default prediction framework where we incorporate the default forecaster's loss utility into default classification and derive an optimal decision rule for this classification problem.

5.1 Optimal Decision Rules

Let random variable $Y \in \{1, -1\}$ represent default, where $Y = 1$ means default and $Y = -1$ means non-default, and random variables $X = (X^1, X^2, ..., X^p)$ taking values in $\mathcal{X}$ represent default predictor, where p is the number of default predictors in the model. Assume that the data sets $(x_i, y_i), i = 1, ..., n$, are generated independently by an unknown distribution P on $X \times Y$. Also denote the conditional distribution $P(Y = 1|X = x)$ as $p(x)$, the marginal distribution function of X as $P_X(x)$, and the marginal distribution function of Y as $P_Y(y)$. Note that here we do not differentiate firm-specific and common predictors, but express them all as X, and we suppress the firm and time subscript by assuming the firms are homogenous and default samples are taken from a distribution which is time-stationary conditioned on the default predictors.

We introduce a decision-based prediction model designed to minimize the loss (or maximize the gain) according to the decision-makers' utility-preference. Specifically, denote the decision maker's default prediction as

$$I_{\{f(x)\geq 0\}} = \begin{cases} 1, & f(x) \geq 0, \\ -1, & f(x) < 0, \end{cases}$$

where $f(x) : R^p \to R$ is a decision function adopted by the decision-maker given the value of default predictors x. Specifically, when the decision-maker observes a data point x for which $f(x) \geq 0$, he or she would consider the data point x as a default; when the decision-maker observes a data point x for which $f(x) < 0$, he or she would consider the data point x as a non-default.

The 0-1 classification loss function is defined by:

$$L_{class}(y, f(x)) = (\frac{I_{\{f(x)\geq 0\}} - y}{2})^2.$$

Note that: when the decision maker makes a right prediction, $L_{class}(y, f(x)) = 0$, otherwise, $L_{class}(y, f(x)) = 1$.

Because significance of judging defaults accurately varies for different groups of decision makers, we associate the decision maker's each prediction with a loss utility as:

$$\pi(x, y) : R^{p+1} \to R^+.$$

The loss utility servers as a weight function to penalize undesirable prediction error. Incorporating this loss utility into the 0-1 classification loss function, we write our loss function as:

$$L^{\pi}_{class}(y, f(x)) = \pi(x, y)(\frac{I_{\{f(x)\geq 0\}} - y}{2})^2.$$

Let $\mathcal{R}_{L^{\pi}_{class},P}(f) = E_P[L^{\pi}_{class}(Y, f(X))]$, where $E_P[.]$ means taking expectation under probability measure P. The default prediction goal is therefore to minimize the expect loss over all functions $f(x) : R^p \to R$:

$$\min_{f(x):R^p\to R} : \mathcal{R}_{L^{\pi}_{class},P}(f).$$

Lemma 6. $\mathcal{R}_{L^{\pi}_{class},P}(f) = \int_{f(x)<0}(\pi(x, 1) + \pi(x, -1))(p(x) - \frac{\pi(x,-1)}{\pi(x,1)+\pi(x,-1)})dP_X(x)$

Proof.

$$\begin{aligned}
\mathcal{R}_{L^{\pi}_{class},P}(f) &= E_P[L^{\pi}_{class}(Y, f(X))] \\
&= E_P[\pi(X,Y)(\frac{I_{\{f(X)\geq 0\}} - Y}{2})^2] \\
&= E_{P_X}[E_{P_Y}[\pi(X,Y)(\frac{I_{\{f(X)\geq 0\}} - Y}{2})^2 | X]] \\
&= E_{P_X}[\pi(X,1)(\frac{I_{\{f(X)\geq 0\}} - 1}{2})^2 * p(x) + \\
&\quad \pi(X,-1)(\frac{I_{\{f(X)\geq 0\}} + 1}{2})^2 * (1 - p(x))] \\
&= \int_{f(x)<0} \pi(x,1)p(x)dP_X(x) + \int_{f(x)\geq 0} \pi(x,-1)(1-p(x))dP_X(x) \\
&= \int_{\mathcal{X}} \pi(x,-1)(1-p(x))dP_X(x) + \\
&\quad \int_{f(x)<0} (\pi(x,1) + \pi(x,-1))(p(x) - \frac{\pi(x,-1)}{\pi(x,1)+\pi(x,-1)})dP_X(x).
\end{aligned}$$

□

Proposition 7. *The decision function $f^*(x)$ minimizes $\mathcal{R}_{L^{\pi}_{class},P}(.)$ if and only if $\forall x \in \mathcal{X}$:*

$$(p(x) - \frac{\pi(x,-1)}{\pi(x,1)+\pi(x,-1)})I_{\{f^*(x)\geq 0\}} \geq 0.$$

Proof. First we show the sufficient condition. Assume that a decision function $f^*(x)$ satisfies $(p(x) - \frac{\pi(x,-1)}{\pi(x,1)+\pi(x,-1)})I_{\{f^*(x)\geq 0\}} \geq 0, \forall x \in \mathcal{X}$. For any other decision function $f(x)$, according to Lemma 6, we have:

$$\begin{aligned}
\mathcal{R}_{L^{\pi}_{class},P}(f^*) - \mathcal{R}_{L^{\pi}_{class},P}(f) &= \int_{f*(x)<0} (\pi(x,1) + \pi(x,-1))(p(x) - \frac{\pi(x,-1)}{\pi(x,1)+\pi(x,-1)})dP_X(x) - \\
&\quad \int_{f(x)<0} (\pi(x,1) + \pi(x,-1))(p(x) - \frac{\pi(x,-1)}{\pi(x,1)+\pi(x,-1)})dP_X(x).
\end{aligned}$$

Denote: $A = \{x \in \mathcal{X} : f^*(x) < 0\}$, $B = \{x \in \mathcal{X} : f(x) < 0\}$ and $C = A \bigcap B$. Let $A - C$ be the set of all elements x which are members of A but not members of C and $B - C$ be the set of all elements x which are members of B but not members of C. We have:

$$\begin{aligned}
\mathcal{R}_{L^{\pi}_{class},P}(f^*) - \mathcal{R}_{L^{\pi}_{class},P}(f) &= \int_{A-C} (\pi(x,1) + \pi(x,-1))(p(x) - \frac{\pi(x,-1)}{\pi(x,1)+\pi(x,-1)})dP_X(x) - \\
&\quad \int_{B-C} (\pi(x,1) + \pi(x,-1))(p(x) - \frac{\pi(x,-1)}{\pi(x,1)+\pi(x,-1)})dP_X(x) \\
&\leq 0.
\end{aligned}$$

The last inequality holds because $\forall x \in A - C, p(x) - \frac{\pi(x,-1)}{\pi(x,1)+\pi(x,-1)} \leq 0$ and $\forall x \in B - C, p(x) - \frac{\pi(x,-1)}{\pi(x,1)+\pi(x,-1)} \geq 0$ and $\pi(x,1) + \pi(x,-1) > 0$. This concludes the sufficient condition.

Second, we show the necessary condition by contradiction. Suppose that we have an optimal decision function $f^*(x)$ and that there exist a set G such that $\forall x \in G$, $(p(x) -$

$\frac{\pi(x,-1)}{\pi(x,1)+\pi(x,-1)})I_{\{f^*(x)\geq 0\}} < 0$.

To establish contradiction, we construct another decision function $\tilde{f}(x)$ from $f^*(x)$ such that: $\tilde{f}(x) = f^*(x), \forall x \notin G$ and $(p(x) - \frac{\pi(x,-1)}{\pi(x,1)+\pi(x,-1)})I_{\{\tilde{f}(x)\geq 0\}} \geq 0, \forall x \in G$. Denote: $M = \{x \in \mathcal{X} : f^*(x) < 0\}$, $N = \{x \in \mathcal{X} : \tilde{f}(x) < 0\}$. We have:

$$\begin{aligned}
\mathcal{R}_{L^{\pi}_{class},P}(\tilde{f}) - \mathcal{R}_{L^{\pi}_{class},P}(f^*) &= \int_{\tilde{f}(x)<0} (\pi(x,1)+\pi(x,-1))(p(x) - \frac{\pi(x,-1)}{\pi(x,1)+\pi(x,-1)})dP_X(x) - \\
&\quad \int_{f^*(x)<0} (\pi(x,1)+\pi(x,-1))(p(x) - \frac{\pi(x,-1)}{\pi(x,1)+\pi(x,-1)})dP_X(x) \\
&= \int_{N\bigcap G} (\pi(x,1)+\pi(x,-1))(p(x) - \frac{\pi(x,-1)}{\pi(x,1)+\pi(x,-1)})dP_X(x) - \\
&\quad \int_{M\bigcap G} (\pi(x,1)+\pi(x,-1))(p(x) - \frac{\pi(x,-1)}{\pi(x,1)+\pi(x,-1)})dP_X(x) \\
&< 0.
\end{aligned}$$

The last inequality holds because $\forall x \in N\bigcap G, p(x) - \frac{\pi(x,-1)}{\pi(x,1)+\pi(x,-1)} \leq 0$ and $\forall x \in M\bigcap G, p(x) - \frac{\pi(x,-1)}{\pi(x,1)+\pi(x,-1)} > 0$ and $\pi(x,1)+\pi(x,-1) > 0$.

Therefore, it contradicts with the optimality of $f^*(x)$, which concludes the necessary condition.

□

Corollary 8. *For the optimal decision function $f^*(x)$,*

$$\mathcal{R}^*_{L^{\pi}_{class},P} = \mathcal{R}_{L^{\pi}_{class},P}(f^*) = \int_{\mathcal{X}} \min\{p(x)\pi(x,1), (1-p(x))\pi(x,-1)\}dP_X(x)$$

Proof.

$$\begin{aligned}
\mathcal{R}_{L^{\pi}_{class},P}(f^*) &= \int_{f^*(x)<0} \pi(x,1)p(x)dP_X(x) + \int_{f^*(x)\geq 0} \pi(x,-1)(1-p(x))dP_X(x) \\
&= \int_{\mathcal{X}} \min\{p(x)\pi(x,1), (1-p(x))\pi(x,-1)\}dP_X(x).
\end{aligned}$$

The last equation holds because:

$$\begin{aligned}
& \min\{p(x)\pi(x,1), (1-p(x))\pi(x,-1)\} = \pi(x,-1)(1-p(x)) \\
\Leftrightarrow\ & (p(x) - \frac{\pi(x,-1)}{\pi(x,1)+\pi(x,-1)}) \geq 0 \\
\Leftrightarrow\ & f^*(x) \geq 0,
\end{aligned}$$

and

$$\begin{aligned}
& \min\{p(x)\pi(x,1), (1-p(x))\pi(x,-1)\} = \pi(x,1)p(x) \\
\Leftrightarrow\ & (p(x) - \frac{\pi(x,-1)}{\pi(x,1)+\pi(x,-1)}) < 0 \\
\Leftrightarrow\ & f^*(x) < 0.
\end{aligned}$$

□

From Proposition 7, it is clear that if one has perfect knowledge of $p(x) = P(Y = 1|X = x)$, then a data point x is judged as default whenever $p(x) > \frac{\pi(x,-1)}{\pi(x,1)+\pi(x,-1)}$. However, in reality, one often knows or understands neither the underlying distribution $P(X, Y)$ that generates the data nor $p(x) = P(Y = 1|X = x)$. Two practical solutions are: a) to estimate $p(x)$ with $\hat{p}(x)$ and then make prediction according whether $\hat{p}(x) \geq \frac{\pi(x,-1)}{\pi(x,1)+\pi(x,-1)}$; b) to have an empirical estimation $\mathcal{R}_{L^{\pi}_{class},E}(f) = \frac{1}{n}\sum_{i=1}^{n} L^{\pi}_{class}(y_i, f(x_i))$ of the overall decision error $\mathcal{R}_{L^{\pi}_{class},P}(f)$ and then find an optimal decision function $f^*_E(x)$ that minimizes the empirical loss $\mathcal{R}_{L^{\pi}_{class},E}(f)$.

In section 4.2.1, we adopt the logistic regression method mostly used in the literature and assume a logistic form $p(x) = \frac{1}{1+e^{\alpha_l+<\beta_l,x>}}, \alpha_l \in R, \beta_l \in R^p$ and obtain the estimate $\hat{p}(x)$ with maximum likelihood method. Assuming that the decision maker's utility $\pi(x, 1) = \pi(x, -1)$, which means that the decision maker is indifferent to the false negative and false positive errors, we obtain the optimal operating points (OOP) on the ROC curve (figure 4.4) for lasso and forward stepwise models, which are 9.2% and 7.6% respectfully, to minimize the average out of sample classification error. However, from proposition 7, we see that if $\pi(x, 1) = \pi(x, -1)$, the optimal operating point should be 50% if our estimate $\hat{p}(x)$ is a good approximate of $p(x)$. The optimal operating points we get for both lasso and forward stepwise model are far below 50%, which shows that the logistic form estimate $\hat{p}(x)$ significantly underestimate the true default probability, therefore the logistics regression may not be appropriate for estimating accurately the default probability.

5.2 Empirical Risk Minimization

In the literature, one often relies on solution a) to obtain an estimation $\hat{p}(x)$ for a logistic functional form $p(x) = \frac{1}{1+e^{\alpha_l+<\beta_l,x>}}$, $\alpha_l \in R, \beta_l \in R^p$, and then make default prediction with $\hat{p}(x)$ by specializing a default probability threshold h for new data points. However, besides the issue of underestimating default probability shown in last section, this strategy has two other drawbacks:

- the conditional independency assumption made when writing the default likelihood as a product form may not be valid. There are on-going debates in the literature that this assumption is valid with a suitable choice of default predictors, see Lando and Nielsen (2010) [57];
- the objective of minimizing negative log-likelihood, is not consistent with the objective of minimizing default prediction/judgement errors which may depend on the decision-maker's utilities.

Here we proceed following solution b): to have an empirical estimate

$$\mathcal{R}_{L^{\pi}_{class},E}(f) = \frac{1}{n}\sum_{i=1}^{n} L^{\pi}_{class}(y_i, f(x_i)),$$

which approximates the overall decision error $\mathcal{R}_{L^{\pi}_{class},P}(f)$ and find a $f^*_E(x)$ that minimizes the empirical loss.

Note that even though the strong law of large number shows that for each $f(x)$: $\mathcal{R}_{L^{\pi}_{class},E}(f) \to \mathcal{R}_{L^{\pi}_{class},P}(f)$, solving:

$$\inf_{f:R^p \to R} \mathcal{R}_{L^{\pi}_{class},E}(f) \tag{5.1}$$

does not in general lead to an approximate minimizer of $\mathcal{R}_{L^{\pi}_{class},P}(.)$, a phenomenon called overfitting.

One common way to avoid overfitting is to choose a smaller set $\mathcal{F}$ of functions $f(x) : R^p \to R$ that is assumed to contain a reasonably good approximation of the solution of $\mathcal{R}^*_{L^{\pi}_{class},P}$. Then, instead of minimizing $\mathcal{R}_{L^{\pi}_{class},E}(.)$ over all functions, one minimizes only over $\mathcal{F}$. This empirical risk minimization approach often tends to produce the solution of the infinite-sample version: $\mathcal{R}^*_{L^{\pi}_{class},P,\mathcal{F}} = \inf_{f\in\mathcal{F}} \mathcal{R}_{L^{\pi}_{class},P}(f)$. (This equation holds if the set $\mathcal{F}$ is finite or if $\mathcal{F}$ can be approximated by a finite set of functions). Therefore, we first restrict $\mathcal{F}$ to be all functions $f(x) = g(< \beta, x > +\alpha)$, where $g(.) : R \to R$ is an invertible function so that the approximation error $\mathcal{R}^*_{L^{\pi}_{class},P,\mathcal{F}} - \mathcal{R}^*_{L^{\pi}_{class},P}$ is small. Later, we extend $\mathcal{F}$ to a richer class of functions in the reproducing kernel Hilbert spaces in section 6.3.

Note that:

$$g(< \beta, x > +\alpha) \geq 0 \Leftrightarrow < \beta, x > +\alpha \geq g^{-1}(0),$$

where $g^{-1}(x) = \inf\{y \in R : g(y) = x\}$. Therefore, it suffices to consider $\mathcal{F} = \{f(x) =< \beta, x > +\alpha\}$ for the classification purpose.

In essence, we formulate the problem toward minimizing the empirical risk: $\mathcal{R}_{L^{\pi}_{class},E,\mathcal{F}}(f) = \frac{1}{n}\sum_{i=1}^{n} L^{\pi}_{class}(y_i, f(x_i))$ over $\mathcal{F} = \{f : f(x) =< \beta, x > +\alpha, \beta \in R^p, \alpha \in R\}$.

That is:

$$\min_{\alpha,\beta} : \frac{1}{n}\sum_{i=1}^{n} \pi(x_i, y_i)(\frac{I_{\{<\beta,x_i>+\alpha\geq 0\}} - y_i}{2})^2. \tag{5.2}$$

Proposition 9. *The solution to formulation* (5.2) *is a polyhedron in* R^{p+1}.

Proof. To solve the optimization problem (5.2), we introduce a dummy decision variable $\delta_i \in \{-1, 1\}$, $i = 1, 2, ..., n$. The formulation (5.2) can be transformed to:

$$\min_{\alpha,\beta,\delta} : \frac{1}{n}\sum_{i=1}^{n} \pi(x_i, y_i)(\frac{\delta_i - y_i}{2})^2 \tag{5.3}$$

$$\text{Such that:} \delta_i I_{\{<\beta,x_i>+\alpha\geq 0\}} \geq 0, i = 1, 2, ..., n,$$

$$\delta_i \in \{-1, 1\}.$$

Note that in formulation (5.3), the objective function only involves decision variables δ_i. Therefore, the problem is reduced to finding a feasible solution of α, β satisfying:

$$< \beta, x_i > +\alpha \geq 0, \text{for all i that } \delta^*_i = 1,$$

$$< \beta, x_i > +\alpha < 0, \text{for all i that } \delta^*_i = -1,$$

where $\delta^*_i \in \{-1, 1\}$ that minimizes the objective function. Therefore, the solution α, β to formulation (5.2) is a R^{p+1} polyhedron. Note that each solution α, β of formulation (5.2) defines a separating hyperplane that minimizes the empirical classification error for the data points $x_i, i = 1, ..., n$: if $< \beta, x_i > +\alpha \geq 0$, x_i predicts as default; if $< \beta, x_i > +\alpha < 0$, x_i predicts as non-default. □

However, two questions arising from here:

- the parameters β, α that minimize the empirical classification error is not unique. How to define a truly optimal β^*, α^* that minimizes out of sample prediction error?

- the loss function $L^{\pi}_{class}(y_i, f(x_i))$ is non-differentiable and non-convex, solving the formulation (5.2) is often NP-hard and computationally expensive. How to find other loss functions that can be used to approximate $L^{\pi}_{class}(y_i, f(x_i))$ and has computational advantages?

In next chapter, we introduce the idea of "Support Vector" and "Hinge Loss function" from Support Vector Machines (SVMs) developed by Cortes and Vapnik (1995) [21] to tackle the two questions.

Chapter 6

Support Vector Machines

In this chapter, we first introduce the idea of "Support Vector" and "Hinge Loss function" from Support Vector Machines (SVMs) developed by Cortes and Vapnik (1995) [21] and then incorporate the decision maker's loss utility $\pi(x, y)$ into Support Vector Machines(SVMs), and show that minimizing the utility adjusted hinge loss is consistent with minimizing utility adjusted classification loss. Our empirical classification results of the decision-based SVMs default prediction framework demonstrates better classification accuracy and more flexibilities in achieving different decision makers' goals.

6.1 Support Vectors

Consider the problem of separating the set of training data $x_i \in R^p$ belonging to two separate classes $y_i \in \{-1, 1\}$, i=1, ..., n, with a hyperplane,

$$< \beta, x > +\alpha = 0, \tag{6.1}$$

where $\beta \in R^p, \alpha \in R$.

The set of data is said to be optimally separated by the hyperplane if it is separated without error and the distance between the closest data to the hyperplane is maximal. There is some parameter redundancy in equation (6.1), and without loss of generality, we consider a canonical hyperplane whose parameters β, α are constrained by:

$$\min_i | < \beta, x > +\alpha| = 1. \tag{6.2}$$

A separating hyperplane in the canonical form must satisfy the following constraints:

$$y_i(< \beta, x_i > +\alpha) \geq 1, i = 1, ..., n. \tag{6.3}$$

The distance $d(\beta, \alpha; x_i)$ of a data point x_i from the hyperplane (β, α) is:

$$d(\beta, \alpha, x_i) = \frac{| < \beta, x_i > +\alpha|}{||\beta||}. \tag{6.4}$$

There are many possible separating hyperplanes that can separate the data "well" and minimize the loss, but under a given appropriate metric there is only one that maximizes the distance between the hyperplanes and the nearest data point of each class, i.e., the optimal separating hyperplane. Intuitively, we would expect the decision rule implied by this hyperplane

generalize well as opposed to other hyperplanes. The optimal hyperplane is given by maximizing the margin, ρ, subject to the constraints (6.3). The margin is given by:

$$\begin{aligned}
\rho(\beta,\alpha) &= \min_{x_i:y_i=-1} d(\beta,\alpha;x_i) + \min_{x_i:y_i=1} d(\beta,\alpha;x_i) \\
&= \min_{x_i:y_i=-1} \frac{|<\beta,x_i>+\alpha|}{||\beta||} + \min_{x_i:y_i=1} \frac{|<\beta,x_i>+\alpha|}{||\beta||} \\
&= \frac{1}{||\beta||}(\min_{x_i:y_i=-1} |<\beta,x_i>+\alpha| + \min_{x_i:y_i=1} |<\beta,x_i>+\alpha|) \\
&= \frac{2}{||\beta||}, \qquad (6.5)
\end{aligned}$$

where the last equation holds because for the optimal separating hyperplane:

$$\min_{x_i:y_i=-1} |<\beta,x_i>+\alpha| = \min_{x_i:y_i=1} |<\beta,x_i>+\alpha| = \min_{x_i:y_i} |<\beta,x_i>+\alpha| = 1.$$

Hence the hyperplane that optimally separates the data is the one that solves the following optimization problem:

$$\min_{\beta,\alpha} \frac{1}{2}||\beta||^2 \qquad (6.6)$$

$$\text{such that: } y_i(<\beta,x_i>+\alpha) \geq 1, i = 1,2,...,n.$$

The optimization problem (6.6) is a quadratic programming subject to linear constraints, and its optimal solution is uniquely given by the saddle point of its Lagrange function:

$$\Phi(\beta,\alpha,\lambda) = \frac{1}{2}||\beta||^2 - \sum_{i=1}^{n} \lambda_i(y_i(<\beta,x_i>+\alpha) - 1), \qquad (6.7)$$

where $\lambda = (\lambda_1,...,\lambda_n) \geq 0 \in R^n$ is the Lagrange multiplier. The lagrange function $\Phi(\beta,\alpha,\lambda)$ has to be minimized with respect to β,α and maximized with respect to $\lambda \geq 0$. Since the original objective function 6.6 is convex, strong duality enables the primal problem to be transformed to its dual problem, given by:

$$\max_{\lambda} W(\lambda) = \max_{\lambda}(\min_{\beta,\alpha} \Phi(\beta,\alpha,\lambda)). \qquad (6.8)$$

The minimum with respect to β and α of the Lagrangian, $\Phi(\beta,\alpha,\lambda)$, is given by:

$$\frac{\partial\Phi}{\partial\alpha} = 0 \Rightarrow \sum_{i=1}^{n} \lambda_i y_i = 0,$$

$$\frac{\partial\Phi}{\partial\beta} = \mathbf{0} \Rightarrow \beta = \sum_{i=1}^{n} \lambda_i y_i x_i.$$

Hence, the dual problem is:

$$\max_{\lambda} W(\lambda) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i\lambda_j y_i y_j <x_i,x_j> + \sum_{k=1}^{n} \lambda_k. \qquad (6.9)$$

The solution to the dual problem is:

$$\lambda^* = arg\max_{\lambda} -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j y_i y_j < x_i, x_j > +\sum_{k=1}^{n}\lambda_k, \tag{6.10}$$

$$\text{such that: } \lambda_i \geq 0, i = 1, 2, ..., n,$$

$$\sum_{i=1}^{n}\lambda_i y_i = 0.$$

Solving (6.10) determines the Lagrange multipliers and the optimal separating hyperplane is given by:

$$\beta^* = \sum_{i=1}^{n}\lambda_i y_i x_i,$$

$$\alpha* = -\frac{1}{2} < \beta^*, x_r, x_l >,$$

where x_r, x_l are any data points from each class satisfying $\lambda_r, \lambda_s > 0, y_r = -1, y_l = 1$.

From Karush-Kuhn-Tucker conditions,

$$\lambda_i^*(y_i(< \beta^*, x_i > +\alpha^*) - 1) = 0, i = 1, ..., n,$$

and hence only the data points x_i^{sv} which satisfy:

$$y_i(< \beta^*, x_i^{sv} > +\alpha^*) = 1 \tag{6.11}$$

will have non-zero Lagrange multipliers. These points x_i^{SVs} satisfying (6.11) are called Support Vectors (SVs).

If the data sets $x_i, i = 1, ..., n$, are linearly separable, all the support vectors lie on the margin and the number of support vectors can be very small compared to the number of total data points. Consequently the hyperplane is determined by a small subset of the data points, the other data points can be removed from the data sets and recalculating the hyperplane would produce the same result. Therefore, support vectors can be used to summarize/compress the information contained in the data set, especially,

$$||\beta^*||^2 = \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i^*\lambda_j^* y_i y_j < x_i, x_j >= \sum_{i\in SVs}\sum_{j\in SVs}\lambda_i^*\lambda_j^* y_i y_j < x_i, x_j >,$$

and because of strong duality,

$$||\beta^*||^2 = 2W(\lambda^*) = 2 * \sum_{i=1}^{n}\lambda_i^* - \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i^*\lambda_j^* y_i y_j < x_i, x_j >,$$

we have:

$$||\beta^*||^2 = \sum_{i=1}^{n}\lambda_i^* = \sum_{i\in SVs}\lambda_i^*.$$

When the data set is not linearly separable, the optimization problem (6.6) does not have general solutions. In this case, introducing an additional loss function associated with misclassification is appropriate. Cortes and Vapnik (1995) [21] introduce non-negative slack

variables $\xi_i \geq 0, i = 1, ..., n$ and an empirical loss :

$$L(\xi) = \sum_{i=1}^{n} \xi_i,$$

where ξ_i is a continuous measure of the misclassification error. The optimization problem becomes:

$$\min_{\beta,\alpha,\xi} : \frac{1}{2}||\beta||^2 + C\sum_{i=1}^{n} \xi_i \tag{6.12}$$

$$\text{such that: } y_i(< \beta, x_i > +\alpha) \geq 1 - \xi_i, i = 1, 2, ..., n,$$

where C is a given regularization parameter which represents the trade-off between the margin $\frac{1}{2}||\beta||^2$ and classification errors ξ. The Lagrangian of the optimization problem (6.12) is:

$$\Phi(\beta, \alpha, \xi, \lambda, \mu) = \frac{1}{2}||\beta||^2 + C\sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \lambda_i(y_i(< \beta, x_i > +\alpha) - 1 + \xi_i) - \sum_{i=1}^{n} \mu_i\xi_i,$$

where $\lambda \geq 0, \mu \geq 0$ are the Lagrange multipliers. The Lagrangian has to be minimized with respect to β, α, ξ and maximized with respect to $\lambda \geq 0, \mu \geq 0$. The primal problem is a convex optimization over linear constraints, and strong duality holds. The dual problem is given by:

$$\max_{\lambda,\mu} W(\lambda, \mu) = \max_{\lambda,\mu}(\min_{\beta,\alpha,\xi} \Phi(\beta, \alpha, \xi, \lambda, \mu)).$$

The minimum with respect to β, α, ξ of the Lagrangian $\Phi(\beta, \alpha, \xi, \lambda, \mu)$ is:

$$\frac{\partial\Phi}{\partial\alpha} = 0 \Rightarrow \sum_{i=1}^{n} \lambda_i y_i = 0,$$

$$\frac{\partial\Phi}{\partial\beta} = \mathbf{0} \Rightarrow \beta = \sum_{i=1}^{n} \lambda_i y_i x_i,$$

$$\frac{\partial\Phi}{\partial\xi} = \mathbf{0} \Rightarrow \lambda_i + \mu_i = C.$$

Hence, the dual problem can be written as:

$$\max_{\lambda} W(\lambda) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i\lambda_j y_i y_j < x_i, x_j > +\sum_{k=1}^{n} \lambda_k, \tag{6.13}$$

$$\text{such that: } 0 \leq \lambda_i \leq C, i = 1, 2, ..., n,$$

$$\sum_{i=1}^{n} \lambda_i y_i = 0.$$

The solution to this dual problem is identical to the separable case except for a modification of the bounds for the Lagrange multipliers. The hyper-parameter C introduces additional capacity control over the linear classifier, and can be directly related to a regularization parameter. C must be chosen to reflect the knowledge of the noise in the data, and cross-validation is usually used to determine the value of C.

6.2 Hinge Loss and Classification Loss

In our decision-based default prediction, we modify the loss $L(\xi) = \sum_{i=1}^{n} \xi_i$ defined by Cortes and Vapnik (1995) [21] and incorporate the decision maker's utility $\pi(x, y) \geq 0$ into the loss and formulate the problem as:

$$\min_{\beta,\alpha,\xi} : \frac{1}{2}||\beta||^2 + C\sum_{i=1}^{n} \pi(x_i, y_i)\xi_i \tag{6.14}$$

$$\text{such that: } y_i(< \beta, x_i > +\alpha) \geq 1 - \xi_i, i = 1, 2, ..., n,$$

$$\xi_i \geq 0, i = 1, 2, ..., n,$$

where $\pi(x, y) : R^{p+1} \to R^+$ is given by the decision maker. And the dual problem becomes:

$$\max_{\lambda} W(\lambda) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i\lambda_j y_i y_j < x_i, x_j > +\sum_{k=1}^{n} \lambda_k, \tag{6.15}$$

$$\text{such that: } 0 \leq \lambda_i \leq C\pi(x_i, y_i), i = 1, 2, ..., n,$$

$$\sum_{i=1}^{n} \lambda_i y_i = 0.$$

Note that in formulation (6.14), $y_i(< \beta, x_i > +\alpha) \geq 1 - \xi_i$ can be written as $\xi_i \geq 1 - y_i(< \beta, x_i > +\alpha)$, and since $\xi \geq 0$, we see that the slack variables must satisfy:

$$\xi_i \geq (1 - y_i(< \beta, x_i > +\alpha)^+ = L_{hinge}(y_i, < \beta, x_i > +\alpha),$$

where $L_{hinge}(y, t) = (1 - yt)^+, y \in \{-1, +1\}, t \in R$ is called hinge loss and is convex in t. The objective function in formulation (6.14) becomes minimal in ξ_i if the above inequality is actually an equality. Therefore, minimizing formulation (6.14) is equivalent to minimizing:

$$\min_{\beta,\alpha} : \frac{1}{2}||\beta||^2 + C'\frac{1}{n}\sum_{i=1}^{n} L^{\pi}_{hinge}(y_i, < \beta, x_i > +\alpha), \tag{6.16}$$

where $C' = Cn$ and

$$L^{\pi}_{hinge}(y, f(x)) = \pi(x, y)(1 - y(f(x))^+.$$

Here we incorporate the decision maker's utility into the hinge loss. We see here that SVMs work toward minimizing the empirical loss plus a regularization penalty term. Recall that in section 5.1, our goal is to minimize the decision-based classification errors $L^{\pi}_{class}(y, f(x)) = \pi(x, y)(\frac{I_{\{f(x)\geq 0\}} - y}{2})^2$. Since $L^{\pi}_{class}(y, f(x))$ is non-convex, we introduce $L^{\pi}_{hinge}(y, f(x))$ as a surrogate loss for $L^{\pi}_{class}(y, f(x))$ and obtain a dual optimization problem that is a convex quadratic programming with linear constraints. But the question left is that: is minimizing the hinge type decision loss $L^{\pi}_{class}(y, f(x))$ consistent with minimizing the 0-1 classification error $L^{\pi}_{class}(y, f(x))$?

Zhang (2004)[82] establishes the relation that for every measurable function $f : R^p \to R$:

$$\mathcal{R}_{L_{class},P}(f) - R^*_{L_{class},P} \leq \mathcal{R}_{L_{hinge},P}(f) - R^*_{L_{hinge},P}.$$

Here we show that this relation still holds after we incorporate the decision maker's utilities $\pi(x, y)$, and therefore $L^{\pi}_{hinge}(y, f(x))$ is a reasonable surrogate for $L^{\pi}_{class}(y, f(x))$.

Proposition 10. *Given a joint distribution P on $X \times Y, X \in R^p, Y \in \{-1, 1\}$ and a positive function $\pi(x, y) : R^{p+1} \to R^+$. For any measurable function $f(x) : R^p \to R$, let $L^\pi_{class}(y, x, f(x)) = \pi(x, y)(\frac{I_{\{f(x)\geq 0\}} - y}{2})^2$ and $L^\pi_{hinge}(y, x, f(x)) = \pi(x, y)(1 - y(f(x))^+$, we have:*

$$\mathcal{R}_{L^\pi_{class},P}(f) - R^*_{L^\pi_{class},P} \leq \mathcal{R}_{L^\pi_{hinge},P}(f) - R^*_{L^\pi_{hinge},P},$$

where $R_{L,P}(f) = E_P[L(Y, f(X))]$ and $R^*_{L,P} = \inf_f E_P[L(Y, f(X))]$.

Proof. For $f(x) : R^p \to [-1, 1]$, we have:

$$\begin{aligned}
\mathcal{R}_{L^\pi_{hinge},P,[-1,1]}(f) &= E_P[\pi(x, y)(1 - yf(x))^+] \\
&= \int_{\mathcal{X}} ((1 - f(x))\pi(x, 1)p(x) + (1 + f(x))\pi(x, -1)(1 - p(x)))dP_X(x) \\
&= \int_{\mathcal{X}} ((\pi(x, 1) - \pi(x, -1))p(x) + \pi(x, -1))dP_X(x) \\
&\quad + \int_{\mathcal{X}} f(x)(\pi(x, 1) + \pi(x, -1))(\frac{\pi(x, -1)}{\pi(x, 1) + \pi(x, -1)} - p(x))dP_X(x),
\end{aligned}$$

where $p(x) = P(Y = 1|X = x)$.

Let $f^*_{class}(x) : R^p \to \{1, -1\}$ be such that: $(p(x) - \frac{\pi(x,-1)}{\pi(x,1)+\pi(x,-1)})f^*_{class}(x) \geq 0$.. Then $f^*_{class}(x)$ minimizes $\mathcal{R}_{L^\pi_{class},P}(.)$, by proposition 7. Moreover, over all functions: $f(x) : R^p \to [-1, 1]$, we see

$$\mathcal{R}^*_{L^\pi_{hinge},P,[-1,1]} = \mathcal{R}_{L^\pi_{hinge},P}(f^*_{class}(x)).$$

Note that for the hinge loss with utilities $L^\pi_{hinge}(y, x, t) = \pi(x, y)(1 - yt)^+, y \in \{-1, +1\}, t \in R$:

$$L^\pi_{hinge}(y, x, -1) \leq L^\pi_{hinge}(y, x, t), \forall t \leq -1,$$

and

$$L^\pi_{hinge}(y, x, 1) \leq L^\pi_{hinge}(y, x, t), \forall t \geq 1.$$

Therefore,

$$\mathcal{R}^*_{L^\pi_{hinge},P} = \mathcal{R}^*_{L^\pi_{hinge},P,[-1,1]} = \mathcal{R}_{L^\pi_{hinge},P}(f^*_{class}(x)),$$

and for all functions $f(x) : R^p \to R$, we have:

$$\mathcal{R}_{L^\pi_{hinge},P}(\tilde{f}) - \mathcal{R}^*_{L^\pi_{hinge},P} \leq \mathcal{R}_{L^\pi_{hinge},P}(f) - \mathcal{R}^*_{L^\pi_{hinge},P},$$

where $\tilde{f}$ is:

$$\tilde{f}(x) = \begin{cases} f(-1), & x < -1; \\ f(x), & -1 \leq x \leq 1; \\ f(1), & x > 1. \end{cases}$$

Note that for all functions $f(x) : R^p \to R$, for the classification error, we also have:

$$\mathcal{R}_{L^\pi_{class},P}(\tilde{f}) - \mathcal{R}^*_{L^\pi_{class},P} = \mathcal{R}_{L^\pi_{class},P}(f) - \mathcal{R}^*_{L^\pi_{class},P}.$$

Therefore, it suffices to proof the proposition for $f(x) : R^p \to [-1, 1]$, in which case we have:

$$\mathcal{R}_{L^\pi_{hinge},P}(f) - \mathcal{R}^*_{L^\pi_{hinge},P} = \int_{\mathcal{X}} (f(x) - f^*_{class}(x))(\pi(x, 1) + \pi(x, -1))(\frac{\pi(x, -1)}{\pi(x, 1) + \pi(x, -1)} - p(x))dP_X(x).$$

By proposition 7,

$$\mathcal{R}_{L^{\pi}_{class},P}(f) - \mathcal{R}^{*}_{L^{\pi}_{class},P} = \int_{\mathcal{X}} (I^{\{0,1\}}_{f(x)<0} - I^{\{0,1\}}_{f^{*}_{class}(x)<0})(\pi(x,1)+\pi(x,-1)) \\ (p(x) - \frac{\pi(x,-1)}{\pi(x,1)+\pi(x,-1)}) dP_X(x), \tag{6.17}$$

where $I^{\{0,1\}}_{\{.\}}$ is the usual indicator function.

Since $\pi(x,1)+\pi(x,-1) \geq 0$, the last step is to show that

$$(I^{\{0,1\}}_{f(x)<0} - I^{\{0,1\}}_{f^{*}_{class}(x)<0})(p(x) - \frac{\pi(x,-1)}{\pi(x,1)+\pi(x,-1)}) \leq (f(x) - f^{*}_{class}(x))(\frac{\pi(x,-1)}{\pi(x,1)+\pi(x,-1)} - p(x)),$$

for all $x, f(x) : R^p \to [-1,1]$.

Noting that $f^{*}_{class}(x) : R^p \to \{1,-1\}$ satisfying $(p(x) - \frac{\pi(x,-1)}{\pi(x,1)+\pi(x,-1)}) f^{*}_{class}(x) \geq 0$, the above inequality holds. □

Remark: note that if $\int_{\{X\times Y\}} \pi(x,y) dP_{\{X\times Y\}}(x,y)=1$, $\pi(x,y) = \frac{dQ_{\{X\times Y\}}(x,y)}{dP_{\{X\times Y\}}(x,y)}$ can be treated as the Radon-Nikodym derivative of measure Q with respect to measure P, where $E_P[\pi(X,Y)L(Y,f(X))] = E_Q[L(Y,f(x))]$, and therefore Proposition 10 follows naturally by the inequality established in Zhang (2004)[82].

6.3 Feature Space and Kernel Functions

Support Vector Machines can map input data $x_i \in R^p$ into a higher dimensional feature space $R^q, q \geq p$ by choosing a mapping function: $\phi : R^p \to R^q$. Support Vector Machines then work in the feature space and constructs an optimal separating hyperplane in this higher dimensional space. Since the calculation for finding the optimal separating hyperplane in the feature space involves only evaluating the inner product $< \phi(x_i), \phi(x_j) >$, it turns out that using a group of functions called "kernel", $K(.,.)$, the calculations can be performed in the input space R^p instead of the potential high (infinity) dimensional feature space where $K(x_i,x_j) =< \phi(x_i), \phi(x_j) >$.

Kernel functions that are usually used in Support Vector Machines are:

- Polynomial: $K(x_i,x_j) =< x_i, x_j >^d$;
- Gaussian Radial Basis: $K(x_i,x_j) = \exp(-\frac{||x_i-x_j||^2}{2\sigma^2})$;
- Multi-Layer Perceptron: $K(x_i,x_j) = \tanh(\rho < x_i, x_j > +\varrho)$.

Working in the feature space with kernel functions $K(.,.)$, the optimization problem (6.18) becomes:

$$\max_{\lambda} W(\lambda) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i\lambda_j y_i y_j K(x_i,x_j) + \sum_{k=1}^{n}\lambda_k, \tag{6.18}$$

$$\text{s.t: } 0 \leq \lambda_i \leq C\pi(x_i,y_i), i = 1,2,...,n,$$

$$\sum_{i=1}^{n} \lambda_i y_i = 0.$$

Solving the above optimization problem determines the Lagrange multipliers λ^*, and the optimal separating hyperplane in the feature space is given by:

$$\sum_{i \in SVs} \lambda_i^* y_i K(x_i, x) + \alpha^*,$$

where $\alpha^* = -\frac{1}{2} \sum_{i \in SVs} \lambda_i^* y_i (K(x_i, x_r) + K(x_i, x_l))$, and x_r, x_l are any data points from each class satisfying $\lambda_r^*, \lambda_s^* > 0, y_r = -1, y_l = 1$.

6.4 Empirical Results

In this section, we use the same default data set as in section 4.3 to conduct our empirical test of the decision-based SVMs default prediction model (6.14) and compare the result with results obtained from logistics regression and linear discriminant methods that are used in the literature. The empirical result in this section demonstrates that the decision-based SVMs models can achieve better prediction accuracy and are flexible in taking accounts of decision makers' utility. Model inputs are the eight predictors identified in section 4.3.1: log(TA), ME/TA, SD/BD, EBIT/TA, WC/TA, SRT, SIGMA, T90RET. Defaults are classified as positive and non-defaults are classified as negative. 10 fold cross-validation is used to estimate the penalty parameter C and the hyper-parameters of the kernel functions. All classification results are out of sample predictions obtained from 10 fold cross-validation.

Table 6.1 shows the classification results with constant unity utility associated with the classification loss. SVMs models have overall better prediction results measured by correct rate(correctly classified samples/total samples) than logistics and linear discriminant models, with SVMs with Gaussian kernel achieving the best correct rate of 0.9339. SVMs models are conservative in judging default cases and obtain lower sensitivity (correctly classified positive samples/total true positive samples) but higher Positive Predictive Rate (correctly classified positive samples/positive classified samples). This is because we weight positive and negative classification error equally here but the default and non-default groups are unbalanced with default rate being 8.07%. For decision makers like speculators for corporate default events, SVMs results may be desirable here.

In Table 6.2, we modify the decision maker's utility to assume a utility penalty to the false negative errors, i.e., errors arising from judging a default case as a non-default. Specifically, we let $\pi(x, 1) = 2(1 - \alpha), \pi(x, -1) = 2\alpha$, where $\alpha = 0.0808$ is the sample default rate. As a result, more default alarms are raised and sensitivities of SVMs models increase significantly while positive predictive rates decrease to low levels. Overall correct rates of SVMs remain above 0.8 and SVMs with Gaussian kernel have the least utility error, which is calculated as (falsely classified positive samples*$\pi(x, 1)$+falsely classified negative samples*$\pi(x, -1)$)/total samples. The classification results here could be desirable for decision makers like credit risk managers.

During the recent financial crisis, big asset firms, like Lehman Brothers, General Motors, Delta Airlines etc., defaulted and their impact is enormous. We assume a further utility penalty to the false negative errors on big asset firms. Specifically, we let $\pi(x, 1) = 2(1 - \alpha)\frac{x_1}{\mu}, \pi(x, -1) = 2\alpha$, where x_1 is the log asset value, μ is the sample average of log asset values and α is the sample default rate. Table 6.3 displays the classification result. Sensitivity II is sensitivity calculated for predicting firms with log asset values above the sample average. Sensitivity III is sensitivity calculated for predicting firms with log asset values above 90th percent sample quantile. SVMs with Gaussian kernel capture 78.91% and 79.31% of default events

for firms with above average log asset values and for firms with 90th percent quantile log asset value respectively.

In Table 6.4, in order to capture the extremely rare events of defaults for firms with huge asset, we add significantly more penalty to the false negative errors, i.e., let $\pi(x,1) = 2(1-\alpha)\frac{x_1}{\mu} * 10$ if $x_1 \geq q_{90th}$, $\pi(x,1) = 2(1-\alpha)\frac{x_1}{\mu}$ otherwise,and $\pi(x,-1) = 2\alpha$, where x_1 is the log asset value, μ is the sample average of log asset values, q_{90th} is the 90th percent quantile of sample log asset values and α is the sample default rate. Sensitivity III in Table 6.4 is increased significantly. SVMs with quadratic kernel captures 96.55% of these rare events, comparing with only 34.48% and 13.79% of logistics regression and linear discriminant analysis respectively.

The empirical test in this section only serves as an example to demonstrate the prediction result of SVMs and the idea to incorporate the decision maker's utility in predictions. More utility functions can be assumed according to the decision maker's interest. For example, one may be more interested in predicting defaults during bad economic times, therefore, we could associate an extra penalty to data with weak economic value in x and let the prediction model put more "effort" to remember (fit) those data in the hope that the prediction model can generalize better when encountering similar cases in the future. This is actually a utility-adjusted regularization idea.

Method	Correct Rate	Sensitivity	Specificity	Positive Predictive Rate	Negative Predictive Rate
SVM Linear Kernel (C=0.8)	0.9288	0.1718	0.9953	0.7615	0.9319
SVM Quadratic Kernel (C=0.005)	0.9308	0.2153	0.9936	0.7482	0.9351
SVM Polynomial Kernel (order=3, C=0.002)	0.9306	0.2422	0.9911	0.7048	0.9370
SVM Gaussian Kernel (sigma=3.1, C=7.5)	0.9339	0.2795	0.9914	0.7418	0.9400
Logistics Regression	0.7871	0.8385	0.7826	0.2531	0.9822
Linear Discriminant Analysis	0.8906	0.6066	0.9156	0.3871	0.9636

Table 6.1: Classification Result I: $\pi(x, y) = 1$

Assume that $\pi(x, y) = 1, \forall x, y$, i.e., constant unity utility associate with the classification loss. Correct Rate is calculated as correctly classified samples/total samples; Sensitivity is calculated as correctly classified positive samples/total true positive samples;Specificity is calculated as correctly classified negative samples/total true negative samples;Positive Predictive Rate is calculated as correctly classified positive samples/positive classified samples; Negative predictive value is calculated as correctly classified negative samples/negative classified samples. SVMs with Gaussian Kernel has the best correct rate, but low sensitivity comparing with logistics regression and linear discriminant analysis.

Method	Correct Rate	Sensitivity	Specificity	Positive Predictive Rate	Negative Predictive Rate	Utility Error
SVM Linear Kernel (C=0.8)	0.8289	0.7888	0.8324	0.2926	0.9782	0.0563
SVM Quadratic Kernel (C=0.005)	0.8605	0.7391	0.8712	0.3352	0.9744	0.0579
SVM Polynomial Kernel (order=3, C=0.002)	0.8792	0.6936	0.8956	0.3685	0.9708	0.0610
SVM Gaussian Kernel (sigma=3.1, C=7.5)	0.8332	0.8095	0.8353	0.3017	0.9804	0.0527
Logistics Regression	0.7871	0.8385	0.7826	0.2531	0.9822	0.0563
Linear Discriminant Analysis	0.8906	0.6066	0.9156	0.3871	0.9636	0.0710

Table 6.2: Classification Result II: $\pi(x,1) = 1.8384, \pi(x,-1) = 0.1615$

Assume that $\pi(x,1) = 1.8384, \pi(x,-1) = 0.1615, \forall x$. Correct Rate is calculated as correctly classified samples/total samples; Sensitivity is calculated as correctly classified positive samples/total true positive samples;Specificity is calculated as correctly classified negative samples/total true negative samples;Positive Predictive Rate is calculated as correctly classified positive samples/positive classified samples; Negative predictive value is calculated as correctly classified negative samples/negative classified samples. Utility Error is the utility modified classification error which is calculated as (falsely classified positive samples*$\pi(x,1)$+falsely classified negative samples*$\pi(x,-1)$)/total samples. More default alarms are raised and sensitivities of SVMs models increase significantly while positive predictive rates decrease to low levels. Overall correct rates of SVMs remain above 0.8 and SVMs with Gaussian kernel have the least utility error.

Method	Correct Rate	Sensitivity	Specificity	Positive Predictive Rate	Negative Predictive Rate	Utility Error	Sensitivity II	Sensitivity III
SVM Linear Kernel (C=0.8)	0.8312	0.7909	0.8348	0.2961	0.9785	0.0548	0.7143	0.5517
SVM Quadratic Kernel (C=0.005)	0.8650	0.7246	0.8774	0.3418	0.9732	0.0580	0.5986	0.6552
SVM Polynomial Kernel (order=3, C=0.002)	0.8791	0.7019	0.8947	0.3693	0.9715	0.0586	0.5986	0.5517
SVM Gaussian Kernel (sigma=3.1, C=7.5)	0.8314	0.8095	0.8333	0.2992	0.9803	0.0512	0.7891	0.7931
Logistics Regression	0.7871	0.8385	0.7826	0.2531	0.9822	0.0580	0.6871	0.3448
Linear Discriminant Analysis	0.8906	0.6066	0.9156	0.3871	0.9636	0.0717	0.3401	0.1379

Table 6.3: Classification Result III: Log Asset Weighted Utility

Assume more utility penalty to the false negative errors on big asset firms. Specifically, let $\pi(x,1) = 2(1-\alpha)\frac{x_1}{\mu}, \pi(x,-1) = 2\alpha$, where x_1 is the log asset value, μ is the sample average of log asset values and $\alpha = 0.0808$ is the sample default rate. Correct Rate is calculated as correctly classified samples/total samples; Sensitivity is calculated as correctly classified positive samples/total true positive samples;Specificity is calculated as correctly classified negative samples/total true negative samples;Positive Predictive Rate is calculated as correctly classified positive samples/positive classified samples; Negative predictive value is calculated as correctly classified negative samples/negative classified samples. Utility Error is the utility modified classification error which is calculated as (falsely classified positive samples*$\pi(x,1)$+falsely classified negative samples*$\pi(x,-1)$)/total samples. Sensitivity II is sensitivity calculated for predicting firms with log asset values above the sample average. Sensitivity III is sensitivity calculated for predicting firms with log asset values above 90th percent sample quantile. SVMs with Gaussian kernel capture 78.91% and 79.31% of default events for firms with above average log asset values and for firms with 90th percent quantile log asset value respectively.

Method	Correct Rate	Sensitivity	Specificity	Positive Predictive Rate	Negative Predictive Rate	Utility Error	Sensitivity II	Sensitivity III
SVM Linear Kernel (C=0.8)	0.7516	0.8054	0.7469	0.2185	0.9776	0.0695	0.9048	0.9310
SVM Quadratic Kernel (C=0.005)	0.8015	0.7867	0.8028	0.2596	0.9772	0.0624	0.7891	0.9655
SVM Polynomial Kernel (order=3, C=0.002)	0.8468	0.7391	0.8563	0.3112	0.9739	0.0849	0.6871	0.7586
SVM Gaussian Kernel (sigma=3.1, C=7.5)	0.8130	0.8219	0.8122	0.2778	0.9811	0.0685	0.8095	0.8621
Logistics Regression	0.7871	0.8385	0.7826	0.2531	0.9822	0.1295	0.6871	0.3448
Linear Discriminant Analysis	0.8906	0.6066	0.9156	0.3871	0.9636	0.1657	0.3401	0.1379

Table 6.4: Classification Result IV: Log Asset Weighted Utility - Extreme Events

we assume significantly more penalty to the false negative errors on extremely huge asset firms, i.e., let $\pi(x,1) = 2(1-\alpha)\frac{x_1}{\mu} * 10$ if $x_1 \geq q_{90th}$, $\pi(x,1) = 2(1-\alpha)\frac{x_1}{\mu}$ otherwise,and $\pi(x,-1) = 2\alpha$, where x_1 is the log asset value, μ is the sample average of log asset values, q_{90th} is the 90th percent quantile of sample log asset values and α is the sample default rate. Sensitivity III in Table 6.4 is increased significantly. SVMs with quadratic kernel captures 96.55% of these rare events, comparing with only 34.48% and 13.79% of logistics regression and linear discriminant analysis respectively. Correct Rate is calculated as correctly classified samples/total samples; Sensitivity is calculated as correctly classified positive samples/total true positive samples;Specificity is calculated as correctly classified negative samples/total true negative samples;Positive Predictive Rate is calculated as correctly classified positive samples/positive classified samples; Negative predictive value is calculated as correctly classified negative samples/negative classified samples. Utility Error is the utility modified classification error which is calculated as (falsely classified positive samples*$\pi(x,1)$+falsely classified negative samples*$\pi(x,-1)$)/total samples. Sensitivity II is sensitivity calculated for predicting firms with log asset values above the sample average. Sensitivity III is sensitivity calculated for predicting firms with log asset values above 90th percent sample quantile.

Chapter 7

Conclusion

In the literature of predicting corporate default, it is an ad-hoc process to select the predictors and different models often use different predictors. We study the predictors of U.S corporate default by forward stepwise and Lasso model selection methods. Using the U.S corporate default data from 1984-2009, we identify a set of eight default predictors that have strong effects in predicting default out of 30 candidate default predictors that have been used in the major default-predicting literature. They are: log(TA), ME/TA, SD/BD, EBIT/TA, WC/TA, SRT, SIGMA and T90RET.

By comparing the eight default predictors' predicting effects during the past three economic recessions, we find that: the predicting effects of firm size log(TA) are decreasing over the three recessions, which coincides with what we have witnessed during the recent financial crisis: giant corporations like Lehman Brothers, General Motors, etc, collapse together with small firms; market variables SRT and SIGMA are not significant during the 1990s recession and the absolute coefficient values of SRT and SIGMA are increasing over time reflecting the fact that the financial market are getting more efficient today than it was in the 1990s and default risk is priced in the equity return; ME/TA, EBIT/TA and T90RET explain a greater portion of default risk for the 1990s and sub-prime mortgage recessions than for the 2000s recession and WC/TA is more significant in the 2000s tech bubble recession than in the other two.

Having identified the set of default predictors, we move on to present a default prediction methodology where we incorporate the decision maker's loss utility $\pi(x, y) : R^{p+1} \to R^+$ into the loss function of default classification. We show that the decision function $f^*(x)$ that minimizes $\mathcal{R}_{L^{\pi}_{class},P}(.)$ if and only if: $(p(x) - \frac{\pi(x,-1)}{\pi(x,1)+\pi(x,-1)})f^*(x) \geq 0$. Comparing the optimal operating points (OOP) on the ROC curve (figure 4.4) of logistic regression results, we find that the logistic estimate $\hat{p}(x)$ significantly underestimate the true default probability and therefore may not be appropriate for the purpose of estimating accurately the default probability.

We later incorporate the decision maker's loss utility $\pi(x, y)$ into Support Vector Machines(SVMs), and show that minimizing the utility adjusted hinge loss is consistent with minimizing utility adjusted classification loss. Our empirical classification results of the decision-based SVMs default prediction framework demonstrates better classification accuracy and more flexibilities in meeting different decision makers' goals.

The idea of incorporating utilities into model calibration is analogous to regularization idea in the sense that we let the prediction model put more "effort" to remember (fit) those data which are important to us in the hope that the prediction model can generalize well when encountering similar cases in the future. This idea can be generalized to other fields beside default prediction.

Bibliography

[1] R. F. A. J. McNeil and P. Embrechts. *Quantitative Risk Management: Concepts, Techniques, and Tools*. Princeton Series in Finance, 2005.

[2] A. M. Aguilera, M. Escabias, and M. J. Valderrama. Using principal components for estimating logistic regression with high-dimensional multicollinear data. *Computational Statistics and Data Analysis*, 50:1905–1924, 2006.

[3] J. D. Akhavein, A. E. Kocagil, and M. Neugebauer. A comparative empirical study of asset correlations. *Fitch Ratings, Quantitative Financial Research Special Report*, 2005.

[4] E. Altman. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance*, 23(4):589–609, 1968.

[5] E. Altman. Predicting financial distress of companies: revisiting the z-score and zeta models. Working paper, 2000.

[6] E. Altman. Corporate distress prediction models in a turbulent economic and basel ii environment. Working paper, 2001.

[7] E. Altman, R. Haldeman, and P. Narayanan. Zeta analysis: a new model to identify bankruptcy risk of corporations. *Journal of Banking and Finance*, 6:29–54, 1977.

[8] E. I. Altman and E. Hotchkiss. *Corporate Financial Distress and Bankruptcy: Predict and Avoid Bankruptcy*. Wiley Finance, 3rd edition, 2006.

[9] S. Azizpour and K. Giesecke. Self-exciting corporate defaults: Contagion vs. frailty. Working paper, 2008.

[10] B. S. Bernanke. Nonmonetary effects of the financial crisis in the propagation of the great depression. *The American Economic Review*, 73(3):257–276, 1983.

[11] S. T. Bharath and T. Shumway. Forecasting default with the kmv-merton model. *AFA 2006 Boston Meetings Paper*, 2006.

[12] T. R. Bielecki and M. Rutkowski. *Credit Risk: Modeling, Valuation and Hedging*. Springer Series in Finance, 2004.

[13] F. Black and J. C. Cox. Valuing corporate securities: Some effects of bond indenture provisions. *Journal of Finance*, 31:351–367, 1976.

[14] F. Black and M. Scholes. The pricing of options and corporate liabilities. *Journal of Political Economy*, 81:81–98, 1973.

[15] J. Y. Campbell, J. Hilscher, and J. Szilagyi. In search of distress risk. *The Journal of Finance*, 63(6):2899–2939, 2008.

[16] U. Cetin, R. Jarrow, P. Protter, and Y. Yildirim. Modeling credit risk with partial information. *Annals of Applied Probability*, 14(3):1167–1178, 2004.

[17] S. Chava and R. A. Jarrow. Bankruptcy prediction with industry effects. *Review of Finance*, 8(4):537–569, 2004.

[18] D. G. Clayton. A model for assocation in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, (65):141–151, 1978.

[19] P. Collin-Dufresne and R. S. Goldstein. Do credit spreads reflect stationary leverage ratios? *The Journal of Finance*, 56(5):1929–1957, 2001.

[20] P. Collin-Dufresne, R. S. Goldstein, and J. Helwege. Is credit event risk priced? modeling contagion via the updating of beliefs. Working paper, 2003.

[21] C. Cortes and V. Vapnik. Support vector networks. *Machine Learning*, 20, 1995.

[22] P. Crosbie and J. Bohn. Modeling default risk. *Technical Report, Mood's KMV*, 2003.

[23] S. R. Das, D. Duffie, N. Kapadia, and L. Saita. Common failings: how corporate defaults are correlated. *The Journal of Finance*, 62(1):93–117, 2007.

[24] S. R. Das, L. Freed, G. Geng, and N. Kapadia. Correlated default risk. *The Journal of Fixed Income*, 16(2):7–32, 2006.

[25] S. R. Das and G. Geng. Correlated default processes: a criterion-based copula approach. *Journal of Investment Management*, 2(2):44–70, 2004.

[26] S. Daul, E. D. Giorgi, F. Lindskog, and A. McNeil. The grouped t-coupula with an application to credit risk. *Risk*, 16(11):73–76, 2003.

[27] M. Davis and V. Lo. Infectious defaults. *Quantitative Finance*, 1:382–387, 2001.

[28] A. C. Davison. *Statistical Models*. Cambridge, 2003.

[29] S. A. Davydenko. When do firms default? a study of the default boundary. *AFA San Francisco Meetings Paper*, 2009.

[30] D. Duffie, A. Eckner, G. Horel, and L. Saita. Frailty correlated default. *The Journal of Finance*, to appear, 2008.

[31] D. Duffie and R. Kan. A yield factor model of interest rates. *Mathematical Finance*, 6(4): 379–406, 1996.

[32] D. Duffie and D. Lando. Term structures and credit spreads with incomplete accounting information. *Econometrica*, 69:633–664, 2001.

[33] D. Duffie, L. Saita, and K. Wang. Multi-period corporate default prediction with stochastic covariates. *Journal of Financial Economics*, 83(3):635–665, 2007.

[34] D. Duffie and K. Singleton. Modeling term strucutures of defaultable bonds. *Review of financial studies*, 12:687–720, 1999.

[35] D. Duffie and K. J. Singleton. *Credit Risk: Pricing, Measurement, and Management.* Princeton Series in Finance, 2003.

[36] K. Dullmann, M. Scheicher, and C. Schmieder. Asset correlations and credit portfolio risk: an empirical analysis. *Deutsche Bundesbank Discussion Paper*, 2007.

[37] P. Embrechts. Copulas: A personal view. *Working paper*, 2009.

[38] Y. H. Eom, J. Helwege, and J. Huang. Structural models of corporate bond pricing: an empirical analysis. *Review of financial studies*, 17(2):499–544, 2004.

[39] I. Filiz, X. Guo, J. Morton, and B. Sturmfels. Graphical models for correlated defaults. *Mathematical Finance, to appear*, 2011.

[40] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1), 2010.

[41] R. Geske. The valuation of corporate liabilities as compound options. *Journal of Financial and Quantitative Analysis*, 12(4):541–552, 1977.

[42] K. Giesecke. A simple exponential model for dependent defaults. *Journal of fixed income*, 13(3):74–83, 2003.

[43] K. Giesecke and S. Weber. Cyclical correlations, credit contagion, and portfolio losses. *Journal of Banking and Finance*, 28:3009–3036, 2004.

[44] K. Giesecke and S. Weber. Credit contagion and aggregate losses. *Journal of Economic Dynamics and Control*, 30(5):741–767, 2006.

[45] E. J. Gumbel. Distributions des valeurs extremes en plusiers dimensions. *Publ. Inst. Statist. Univ. Paris*, 9:171–173, 1960.

[46] X. Guo, R. Jarrow, and H. Lin. Distressed debt prices and recovery rate estimation. *Review of Derivatives Research*, to appear, 2009.

[47] X. Guo, R. A. Jarrow, and Y. Zeng. Credit risk models with incomplete information. *Mathematics of Operations Research*, 34(2):320–332, 2009.

[48] T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer Series in Statistics, 2nd edition, 2009.

[49] J. Hull, M. Predescu, and A. White. The valuation of correlation-dependent credit derivatives using a structural model. Working paper, 2006.

[50] R. A. Jarrow and F. Yu. Counterparty risk and the pricing of defaultable securities. *Journal of Finance*, 56(5):1765–1799, 2001.

[51] I. T. Jolliffe. *Principal Component Analysis.* Springer Series in Statistics, 2nd edition, 2002.

[52] N. E. Karoui. Recent results about the largest eigenvalue of random covariance matrices and statistical application. *Acta Physica Polonica B*, 36(9), 2005.

[53] S. Kealhofer. Quantifying credit risk ii: Debt valuation. *Financial Analysts Journal*, 59(3): 78–92, 2003.

[54] S. Kealhofer. Quantifying credit risk ii: Default prediction. *Financial Analysts Journal*, 59 (1):30–44, 2003.

[55] D. Lando. Three essays on contingent claims pricing. *Ph.D. Thesis, Cornell University*, 1994.

[56] D. Lando. On cox processes and credit risky securities. *Review of Derivatives Research*, 2 (2):99–120, 1998.

[57] D. Lando and M. S. Nielsen. Correlation in corporate defaults: Contagion or conditional independence? *Journal of Financial Intermediation*, 3(19):355–372, 2010.

[58] L. H. P. Lang and R. M. Stulz. Contagion and competitive intra-industry effects of bankruptcy announcements. *Journal of Financial Economcis*, 32:45–60, 1992.

[59] A. H. Lau. A five-state financial distress prediction model. journal of accounting research. *Journal of Accounting Research*, 25(1):127–138, 1987.

[60] H. E. Leland. Predictions of default probabilities in structural models of debt. *Journal of Investment Management*, 2(2):5–20, 2004.

[61] H. E. Leland and K. Toft. Optimal capital structure, endogenous bankruptcy, and the term structure of credit spreads. *The Journal of Finance*, 51(3):987–1019, 1996.

[62] D. X. Li. On default correlation: A copula function approach. *Journal of Fixed Income*, 9 (43-54), 2000.

[63] F. A. Longstaff and E. S. Schwartz. A simple approach to valuing risky fixed and floating rate debt. *The Journal of Finance*, 50(3):789–819, 1995.

[64] D. J. Lucas. Default correlation and credit analysis. *Journal of Fixed Income*, 4(4):76–87, 1995.

[65] B. D. Marx and E. P. Smith. Principal component estimation for generalized linear regression. *Biometrika*, 77(1):23–31, 1990.

[66] R. C. Merton. On the pricing of corporate debt: The risk structure of interest rate. *Journal of Finance*, 29:449–470, 1974.

[67] F. Modigliani and M. H. Miller. The cost of capital, corporation finance and the theory of investment. *American Economic Review*, 48(3):261–297, 1958.

[68] R. B. Nelsen. *An Introduction to Copulas*. Springer, New York, 1999.

[69] J. A. Ohlson. Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, 18(1):109–131, 1980.

[70] E. Papageorgiou and R. Sircar. Multiscale intensity models and name grouping for valuation of multi-name credit derivatives. *Working paper*, 2008.

[71] K. Pearson. On lines and planes of closest fit to systems of points in space. *Phil. Mag.*, 2 (6):559–572, 1901.

[72] R.Jarrow and S.Turnbull. Pricing options on financial securities subject to default risk. *Journal of Finance*, 50:53–86, 1995.

[73] M. E. Rubinstein. *Rubinstein On Derivatives*. Risk Books, 1999.

[74] M. E. Rubinstein. Great moments in financial economics: Ii. modigliani-miller theorem. *Journal of Investment Management*, 1(2), 2003.

[75] M. E. Rubinstein. *A History of the Theory of Investments*. Wiley Finance, 2006.

[76] T. Schmidt. Copulas and dependent measurement. *Encyclopedia of Quantitative Finance*, to appear, 2008.

[77] T. Schmidt. Correlation and correlation risk. *Encyclopedia of Quantitative Finance*, to appear, 2008.

[78] T. Shumway. Forecasting bankruptcy more accurately: a simple hazard model. *Journal of Business*, 74(1):101–124, 2001.

[79] R. Tibshirani. Regression shrinkage and selection via lasso. *Journal of the Royal Statistical Society. Series B*, 58(1), 1996.

[80] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

[81] J. Zhang, F. Zhu, and J. Lee. Asset correlation, realized default correlation, and portfolio credit risk. *Technical Report, Moody's KMV*, 2008.

[82] T. Zhang. Statistical behaviour and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–134, 2004.

[83] C. Zhou. An analysis of default correlations and multiple defaults. *Review of financial studies*, 14(2):555–576, 2001.

[84] M. E. Zmijewski. Methodological issues related to the estimation of financial distress prediction models. *Journal of Accounting Research*, 24(Supplement):59–86, 1984.

Appendix A

Abbreviation	Description	Database Input
ME	Market Value of Equity	Compustat: PRCCM*CSHO
BD	Book Value of Total Debt	Compustat: DLC+DLTT
TA	Total Asset	Compustat: AT
TL	Total Liability	Compustat: LT
SD	Short Term Debt	Compustat: DLC
SALE	Sales	Compustat: SALE
NI	Net Income	Compustat: IB
RE	Retained Earnings	Compustat: RE
EBIT	Earnings Before Interest and Taxes	Compustat: EBIT
ICR	Interest Coverage Ratio	Compustat: EBIT/XINT
WC	Working Capital	Compustat: ACT-LCT
AR	Account Receivable	Compustat: RECT
INVT	Inventories	Compustat: INVT
CR	Current Ratio	Compustat: ACT/LCT
CH	Cash	Compustat: CH
QR	Quick Ratio	Compustat: (ACT-INVT)/LCT
OM CH	Change in Operating Margin	Compustat: OIBDP/SALE
SRT	Trailing One Year Stock Return	CRSP Monthly Stock
SIGMA	One Year Monthly Stock Volatility	CRSP Monthly Stock
RSIZE	Relative Market Value of Equity	ME/(NYSE, AMEX, NASDAQ Market Value)
SPRET	Trailing One Year S&P 500 Return	CRSP Monthly Stock
T90RET	90 Days Treasury Bill Rate	Global Insight
GDP CH	Gross Domestic Production	Global Insight
IPI CH	Industrial Production Index	Global Insight
CPI CH	Consumer Price Index	Global Insight
TSPRD	1 to 10 Year Treasury Spread	Global Insight

Table A.1: Abbreviation and Data Source

Source/ Variables	**Idiosyncratic**	**Common**
Altman (1968) [4]	ME/BD, WC/TA, RE/TA, EBIT/TA, Sales/TA	
Altman (1977) [7]	EBIT/TA, EVL(5-10 years), ICR, RE/TA, CR , ME/TC, TA	
Ohlson (1980) [69]	Log(TA/GNP price index), TL/TA, WC/TA, CR,$\mathbf{1}_{\{TL>TA\}}$, NI/TA, $\mathbf{1}_{\{NI<0\}}$ % Change of NI, funds by operations/TL	
Zmijewski (1984)[84]	NI/TA, TL/TA, CR	
Lau (1987) [59]	LRT, DER, WC/TL, SPT, OPES, $\mathbf{1}_{\{dividend=0\}}$, LOPA, TCEP, TWF, DVD	
Shumway (2001) [78]	NI/TA, TL/TA, RSIZE, SRT, SIGMA	
Chava and Jarrow (2004) [17]	NI/TA, TL/TA, RSIZE, SRT SIGMA	IND,NI/TA*IND,TL/TA*IND
Duffie, etc (2007, 2008) [33] [30]	DTD, SRT	T90RET, SPRET, GDP CH, IPI CH
Campbell, etc (2008) [15]	Market value version of Shumway (2001), Cash/Market TA, MB, log(TSTR), Lagged NI/Market TA, Lagged SRT	
Lando and Nielsen (2010) [57]	SRT, DTD, QR, SD/BD, log(TA)	SPRET, T90RET, IPI CH, CPI CH, TSPRD

Table A.2: Predictor by Literature

Literature	Default Data Source	Sample Period/Size
Altman (1968) [4]	-	1946-1965, 33 bankruptcies, total 66 manufacturing firms
Ohlson (1980) [69]	WSJI	1970-1976, 105 bankruptcies, 2163 firms
Shumway (2001) [78]	WSJI, CCR, Compustat, DOS	1962-1992, 300 bankruptcies, non-financial firms, 28664 firm-years
Chava and Jarrow (2004) [17]	WSJI, SDC, SEC, CCR	1962-1999, 409 bankruptcies, 72682 firm-years
Duffie, etc (2008) [30]	DRS, Bloomberg	1979-2004, 2793 exit, non-financial firms, 402434 firm-months 176 bankruptcy, 320 other default, 1047 M&A, 671 other exit
Campbell, etc (2008) [15]	WSJI, SDC, SEC, CCR	1963-2003, 1614 failures (bankruptcies, delisted, D rating), 1282853 firm months
Davydenko (2009) [29]	DRS, FISD	1997-2005, Junk firms: 213 default, 593 non-default, 7057 firm-quarters with parent-subsidiary defaults and multiple defaults within two years cleaned

Table A.3: Default Sample in Literature