Modeling Expected Loss

S. Chava*, C. Stefanescu**, and S. M. Turnbull***

October 29, 2006

[1]

Abstract

This paper develops a methodology for modeling and estimating expected loss over arbitrary horizons. We jointly model the probability of default and the recovery rate given default. Different model specifications are estimated using an extensive default and recovery data set that contains the majority of defaults between 1980–2004 of AMEX, NYSE and NASDAQ listed companies. We undertake extensive out-of-sample performance tests for both the default prediction models and recovery rate given default models. Under the joint model specification, we find that the probability of default and the recovery rate given default are negatively correlated out-of-sample, and that the magnitude of the correlation varies with the credit cycle. We also compare the accuracy of the out-of-sample one year ahead default predictions using quarterly and annual data.

[1]*Texas A&M, Business School,**London Business School and ***Bauer College of Business

1 Introduction

This paper provides a new methodology for modeling and estimating expected loss, defined as the product of the probability of default, loss given default, and exposure at default. Under the advanced internal ratings based approach in the New Basel Capital Accord (Basel II), banks are allowed to develop their own estimates of these three parameters so that they reflect the nature of their portfolios. The estimates are subject to supervisory review to ensure that they are "reasonable". The Accord is not explicit as to all the steps that banks must take to show that their models are reasonable in order to gain regulatory approval, instead it requires banks to present enough evidence about the properties and performance of their models to satisfy regulators. However, regulators themselves are unsure about how to assess whether the models that an institution uses are reasonable.[2] In this paper we develop a methodology for modelling and estimating expected loss over arbitrary horizons in the presence of unobservable heterogeneity, and we demonstrate how to assess the performance of the different components of this methodology.

This paper makes four contributions. First, we provide a methodology for the joint modeling of the probability of default and of the recovery rate given default. Our framework models the evolution of the state variables explaining the probability of default and the loss given default, and can therefore be used for estimating the expected loss over arbitrary horizons. This is important, since banks typically examine their risk exposure over multi–year horizons, while the majority of extant methodologies are static in nature[3] and generate predictions over a given horizon, usually one year.[4]

Second, we address explicitly the issue of estimating the effects of unobserved heterogeneity and incomplete information. Investors usually have only incomplete information about the true state of a firm. There are individual variations among firms that affect the probability of default but that are not directly observable, such as differences in managerial styles, in the skill sets of workers, and in firm culture. Even differences in such areas as production skills, resource usage, cost control, and risk management are only partially revealed in accounting statements. While the quarterly reports of a public firm must satisfy generally accepted accounting principles, they are generated based on a large set of unidentified assumptions. Investors do not have complete knowledge about these implicit assumptions, or the reliability of the firm's reporting system, or the integrity of its auditors,[5] yet they must

[2]See the recent report issue by the BIS (May 2005) on the validation of the internal rating systems.

[3]However, see Janosi, Jarrow and Yildirim (2002) and Duffie, Saita and Wang (2005) who estimate the stochastic processes describing the covariates.

[4]This coincides with the one year Bank of International Settlement (BIS) regulatory horizon.

[5]The Public Company Accounting Oversight Board found significant deficiencies in a quarter of the audit

base their projections on the reports. The uncertainty surrounding these projections will, in general, depend on the state of the economy, the state of the particular sector in which the company operates, and the unique characteristics of the firm.

Third, we estimate the parameters of different extant models using an extensive default and recovery data set, containing the majority of defaults of companies listed on the AMEX, NYSE and NASDAQ between 1980–2004. Many start-up firms are listed on NASDAQ and tend to have negative cash flows over the initial few years; failure rates also tend to be higher among start-ups than among established firms. This increases the heterogeneity of the sample universe. However, most extant academic studies on default prediction limit their analyses to data from firms listed on AMEX and NYSE. We investigate whether there is a NASDAQ effect on model performance, and we find that both the in-sample and the out-of-sample performance of default prediction models are indeed better for the more homogeneous AMEX and NYSE data set, while the relative ranking of models remains unchanged.

Fourth, we assess the out-of-sample performance of eight different default models and six recovery models over the period 1996–2004, in order to address the concerns of regulators about the reasonableness of model specification. For default prediction, we investigate another facet of the effects of data heterogeneity on model performance, by separately considering the set non–financial firms and the set of manufacturing firms. Our results confirm the superiority of the Shumway (2001) default prediction model. Over the same period, we find that several different models for the recovery rate given default have similar out-of-sample performance. Finally, we compare the one year out-of-sample default model performance using yearly and quarterly data. With quarterly data, one year represents four periods and it is then necessary to also estimate the parameters describing the stochastic processes for the covariates. Due to the added covariate estimation error, we find that a history of at least five years of quarterly data is needed for similar levels of out-of-sample performance as that of one-period models estimated with annual data. This is an issue for young firms that typically trade on NASDAQ.

The Basel II framework recognizes that changes in the probability of default and the loss given default are generally related for most asset classes,[6] and it requires financial institutions to recognize this dependence.[7] We demonstrate that our joint model specification implies that out-of-sample, the probability of default and the recovery rate given default are

engagements undertaken by one of the top four accounting firms in 2004. See the *Financial Times*, September 30, 2005, p1.

[6]The dependence between the probability of default and the loss given default (one minus the recovery rate), implies that for a portfolio of loans the distribution describing the loss can vary substantially from that estimated employing the foundation IRB approach with its assumed loss given default.

[7]See the Basel Committee on Banking Supervision (2005) guideline.

negatively correlated. Furthermore, the magnitude of the correlation varies with the credit cycle.

This paper is related to several different strands of previous research. There is a large and growing literature devoted to the modeling of the probability of default — see Shumway (2001), Chava and Jarrow (2004), Campbell, Hilscher and Szilagyi (2004) and Duffie, Saita and Wang (2005). An extensive survey of methodologies is given in Altman and Hotchkiss (2005). There is also an emerging literature addressing the modeling of the determinants of the recovery rate given default. A survey of empirical evidence regarding the properties of recovery rates is given in Schuermann (2004) — see also Acharya, Bharath and Srinivasan (2003). Several recent studies model the dependence between the probability of default and the recovery rate given default, by assuming there is a common latent factor affecting both (Frye, 2000; Pykhtin, 2003; Dullmann and Trapp, 2004). However, to the best of our knowledge, there are no empirical studies that attempt to explicitly model the covariates affecting the probability of default, the recovery rate given default, and their dependence.

The paper is structured as follows. In Section 2 we develop our modeling methodology, and in Section 3 we describe the data set used in this study. The empirical results for the estimation of the probability of default and the recovery rate over a one-period horizon are given in Section 4. In Section 5 we examine the implications for the modeling of the expected loss over arbitrary horizons. Section 6 concludes the paper with a summary of our findings.

2 The Default and Recovery Models

In this section we first describe the specification of default models with unobservable heterogeneity and develop the estimation methodology. Next, we discuss several specifications of recovery rate models.

2.1 The Default Models

The sample data contains firms grouped in G groups or industries. Let n_i be the number of firms in the ith group, and $n = \sum_{i=1}^{G} n_i$ be the total number of firms in the sample. During the observation period $[0, T]$, any particular firm may experience a default, may leave the sample before time T for reasons other than default (for example a merger, an acquisition, or a liquidation), or may survive in the sample until time T. A firm's lifetime is said to be censored if either default does not occur by the end of the observation period, or if the

firm leaves the sample because of a non–default event. Let T_{ij} denote the observed (possibly censored) lifetime of the jth firm in the ith group, and let N_{ij} be the censoring indicator, where $N_{ij} = 1$ if T_{ij} is a default time and $N_{ij} = 0$ if T_{ij} is a censoring time. The total number of failures in group i is given by $N_{i.} = \sum_{j=1}^{n_i} N_{ij}$. For every $s = 1, \ldots, d$, let $\delta_{ij}(s) = 1$ if the jth firm in the ith group is in the sample at time t_s, and zero otherwise. For example, if the firm is in the sample at the beginning of the observation period and censoring only occurs at time T, then $\delta_{ij}(s) = 1$, for $s = 1, ..., d$.

Let $X_{ij}(t)$ be a $1 \times K$ vector of covariates at time t. The vector $X_{ij}(t)$ usually includes a constant component representing an intercept term, and it is composed of both firm-specific variables and macroeconomic variables. Information about the firm-specific variables terminates at time T_{ij}, and information about the macroeconomic variables is available at all times. We observe the covariates at discrete time intervals $0 < t_1 < t_2 \cdots < t_d \leq T$, and assume that $X_{ij}(t)$ is constant during the period between two consecutive observations.

Let $\lambda_{ij}(t)$ be the default intensity function (the hazard function) for the jth firm in the ith group. In order to model the correlation between defaults of firms in the same group, we assume that the unobservable heterogeneity can be represented by a latent non–negative random variable Y_i, common to all firms in the same industry, which we shall refer to as frailty[8] and which represents the effects of the unobservable measurement errors and missing variables.[9] The shared frailty Y_i acts multiplicatively on the intensity functions $\lambda_{ij}(t)$, so

[8] An introduction to frailty models is given in Kiefer (1988), Klein and Moeschberger (1997, chapter 13), and Hougaard (2000, chapter 7). There is a large biostatistical and demographic literature on frailty modelling, but to–date there have been only a small number of applications in the credit risk area. Gagliardini and Gourieroux (2003) and Schönbucher (2003a) introduce the notion of unobservable heterogeneity or frailty to model information driven contagion.

[9] Let $X^T(t)$ represent the true value of the vector of covariates and $X(t)$ be the observed covariates, where we assume that

$$X_k^T(t) = X_k(t) + e_k(t).$$

Here $e_k(t)$ is the measurement error of the kth covariate for the firm. Hence $X^T(t)\beta = X(t)\beta + y(t)$, where $y(t)$ represents the effects of the measurement errors and β is a vector of parameters giving the dependence of the default intensity on the covariate vector. We shall assume that the baseline default intensity is $\lambda_0(t) = \exp(X(t)\beta)$.

If there are missing variables, let $m(t)$ denote the vector of missing variables and β^M the corresponding vector of parameters. The intensity is now given by

$$\lambda(t) = \exp(X(t)\beta + m(t)\beta^M + y(t)),$$

which we can rewrite as

$$\lambda(t) = Y(t)\exp(X(t)\beta).$$

that the hazard rates are specified by

$$\lambda_{ij}(t) = Y_i \exp(X_{ij}(t)\beta), \tag{1}$$

where β denotes the $K \times 1$ vector of regression parameters. Conditional on the unobserved Y_i, the lifetimes of firms in the ith group are independent. When the unknown Y_i is integrated out, the lifetimes become dependent; the dependence is induced by the common value of Y_i.

The shared frailty model specified by (1) is a natural approach for modeling dependence and taking into account unobservable heterogeneity. The model can be easily extended to the case where the frailties are time-varying, multivariate rather than univariate, or obligor specific rather than shared by all obligors in the same sector. Such extensions allow modeling of more flexible patterns of default dependence. For example, the shared frailty model (1) implies positive correlation of defaults within an industry; in practice, however, some degree of negative correlation may be conceivable, for example due to competition. The multivariate lognormal frailty model (Stefanescu and Turnbull, 2006) can accommodate negative default dependence as well.

The frailty has an assumed prior distribution which is updated as the default information set evolves over time. For example, if no firms within a particular sector default, this might help to increase confidence in the credit worthiness of the firms in this sector. Conversely, if there is a failure in a particular sector or the aggregate number of defaults in the economy increases, this might adversely affect the assessment of credit worthiness. There is a range of choices for the distribution of the frailties — the most popular is the gamma distribution $G(r, \alpha)$, partly due to mathematical convenience.[10] With gamma frailties, the scale parameter needs to be restricted for identifiability reasons, and the standard restriction is $r = \alpha$ as this implies a mean of one for Y_i. We complete the specification of model (1) by assuming that the sector frailties Y_i are independent and identically distributed with a gamma distribution $G(1/\theta, 1/\theta)$, with $\theta > 0$. The unconditional frailty means are thus equal to one, while the conditional means vary across sectors.

We next show how to estimate the parameters of the default model in a maximum likelihood framework. Let γ denote the vector of the parameters to be estimated for the stochastic processes $\{X_{ij}(t)\}$, and L_X denote the likelihood function of the covariates.[11] The likelihood of the sample is a product of the survival likelihood conditional on the frailties,

[10]The gamma density function of Y_i is given by $f(y_i) = \alpha^r y_i^{r-1} \exp(-\alpha y_i) \cdot \frac{1}{\Gamma(r)}$, where $\Gamma(r)$ is the gamma function. The expected value is $E[Y_i] = r/\alpha$ and the variance $var(Y_i) = r/\alpha^2$.The parameter α is referred to as the scale parameter and r as the shape parameter.

[11]The covariate likelihood may correspond, for example, to an autoregressive time series process.

the likelihood of the frailties, and the likelihood of the covariates:

$$L = L(\theta, \beta \,|\, \{N\}, \{T\}, \{Y\}, \{X\}) \cdot L_Y(\theta) \cdot L_X(\gamma),$$

where the likelihood function for the frailties is given by

$$L_Y(\theta) = \prod_{i=1}^{G} f(y_i) = \prod_{i=1}^{G} \frac{1}{\theta^{1/\theta}\Gamma(1/\theta)} y_i^{1/\theta - 1} \cdot \exp(-y_i/\theta).$$

The parameters to be estimated are the regression coefficients β, the frailty variance θ, and the parameters γ for the covariate processes $\{X_{ij}\}$. The maximization program separates, implying that γ is estimated separately from β and θ. In general, the estimation of γ is the standard numerical procedure of fitting a multivariate time series process to the covariate vectors $\{X(t)\}$. In this section we shall focus on the estimation of β and θ.

Let $L(\theta, \beta)$ denote the likelihood conditional on the data $\{T_{ij}, X_{ij}, N_{ij}\}$ and including the frailties. This is given by

$$L(\theta, \beta) = L_Y(\theta) \prod_{i=1}^{G} \prod_{j=1}^{n_i} L(\theta, \beta \,|\, T_{ij}, N_{ij}, X_{ij}, y_i),$$

where

$$L(\theta, \beta \,|\, T_{ij}, N_{ij}, X_{ij}, y_i) = [y_i \exp(X_{ij}(T_{ij})\beta)]^{N_{ij}} \cdot \exp\left(- \int_0^{T_{ij}} \lambda_{ij}(t) dt \right),$$

and the integrated hazard is given by

$$\int_0^{T_{ij}} \lambda_{ij}(t) dt = Y_i \sum_{s=1}^{d} \delta_{ij}(s) \exp(X_{ij}(t_s)\beta) \equiv Y_i \Lambda_{ij}.$$

The log–likelihood function is therefore

$$\log L(\theta, \beta) = \log L_Y(\theta) + \sum_{i=1}^{G} \sum_{j=1}^{n_i} \log L(\theta, \beta \,|\, T_{ij}, N_{ij}, X_{ij}, y_i),$$

where

$$\log L(\theta, \beta \,|\, T_{ij}, N_{ij}, X_{ij}, y_i) = N_{ij}[\log(y_i) + X_{ij}(T_{ij})\beta] - y_i \Lambda_{ij}, \tag{2}$$

and

$$\log L_Y(\theta) = \sum_{i=1}^{G} [(1/\theta - 1)\log(y_i) - y_i/\theta - \log\Gamma(1/\theta) - (1/\theta)\log(\theta)]. \tag{3}$$

From (2) and (3) it follows that the log–likelihood function is given by

$$\begin{aligned} \log L(\theta, \beta) &= \sum_{i=1}^{G} [(1/\theta - 1 + N_{i\cdot})\log(y_i) - y_i/\theta] \\ &\quad -G[\log\Gamma(1/\theta) + (1/\theta)\log(\theta)] \\ &\quad + \sum_{i=1}^{G}\sum_{j=1}^{n_i} N_{ij} X_{ij}(T_{ij})\beta - y_i \Lambda_{ij}. \end{aligned} \tag{4}$$

In order to maximize the likelihood, we use the Expectation–Maximization (EM) algorithm (Dempster, Laird and Rubin, 1977), which is the classic tool for obtaining maximum likelihood estimates from incomplete or missing data. The complete data for model (1) consists of the realized values of the frailties $Y_1, \ldots, Y_G$ and the uncensored lifetimes. The observed but incomplete data consists in the observed lifetimes $\{T_{ij}\}$ and the censoring indicators $\{N_{ij}\}$. The EM algorithm starts with some initial estimates; for the β coefficients these can be computed by ignoring the frailty terms, and the initial estimate for the frailty variance θ can be set equal to one. Then the algorithm iterates between two steps: the expectation (E) step computes expected values of the sufficient statistics for the complete data, conditional on the observed data and current values of the parameters. In the maximization (M) step, new estimates of the unknown parameters are obtained by numerically maximizing the likelihood computed with the expected values of the sufficient statistics from the previous E–step. These two steps are repeated until convergence is achieved, and it can be shown that, under mild conditions, the EM algorithm converges to the maximum likelihood estimates.

Conditional on the observed data $\{T_{ij}, N_{ij}, X_{ij}\}$ and on the current values of parameters θ and β, the frailty Y_i has a gamma distribution $G(A_i, C_i)$ with scale parameter $C_i = 1/\theta + \sum_{j=1}^{n_{ij}} \Lambda_{ij}$ and shape parameter $A_i = N_{i\cdot} + 1/\theta$. The conditional means are therefore

$$\begin{aligned} E[Y_i] &= A_i/C_i \\ E[\log(Y_i)] &= \psi(A_i) - \log(C_i), \end{aligned} \tag{5}$$

where $\psi(\cdot)$ is the digamma function. From (4) and (5) it follows that the expected value of

the log–likelihood function which is maximized in the M step is given by

$$\begin{aligned} E[\log L(\theta,\beta)] &= \sum_{i=1}^{G}(1/\theta - 1 + N_{i\cdot})[\psi(A_i) - \log(C_i)] - [A_i/C_i]/\theta \\ &\quad -G[\log\Gamma(1/\theta) + (1/\theta)\log(\theta)] \\ &\quad + \sum_{i=1}^{G}\sum_{j=1}^{n_i} N_{ij}X_{ij}(T_{ij})\beta - [A_i/C_i]\Lambda_{ij}. \end{aligned}$$

After convergence of the EM algorithm, the standard errors of the estimates of θ and β can be computed from the inverse of the observed information matrix. Using these estimates, we can also calculate the expected value of the frailty for each group.

This methodology can be easily extended to the case of competing risks.[12] Firms may exit the sample for reasons other than default, such as a merger or an acquisition. These non–default events are all competing risks that may cause censoring of a firm's lifetime. For the jth firm in the ith group, let M_{ij} denote the indicator if the firm exits the sample for reasons other than default, and let $\alpha_{ij}(t)$ be the intensity function for the competing risks which will also depend on the firm's covariates $X_{ij}(t)$).

With multiple causes for exit, we consider a bivariate frailty model whereby the frailty for the ith group is given by $Y_i = (Y_{i1}, Y_{i2})$. The hazard rates are specified by

$$\lambda_{ij}(t) = Y_{1i}\exp(X_{ij}(t)\beta_1), \tag{6}$$

and

$$\alpha_{ij}(t) = Y_{2i}\exp(X_{ij}(t)\beta_2), \tag{7}$$

where Y_{1i} is the frailty associated with default and Y_{2i} is the frailty associated with exit for reasons other than default for the ith group. We assume that Y_{1i} and Y_{2i} are independent, with gamma distributions $G(1/\theta_1, 1/\theta_1)$ and $G(1/\theta_2, 1/\theta_2)$ respectively. Let $\theta = (\theta_1, \theta_2)$ denote the vector of parameters for the frailty distributions, and $\beta = (\beta_1, \beta_2)$ denote the vector of covariate coefficients. The likelihood conditional on the data $\{T_{ij}, X_{ij}, N_{ij}, M_{ij}\}$ and including the frailties is given by

$$L(\theta,\beta) = L_Y(\theta)\prod_{i=1}^{G}\prod_{j=1}^{n_i} L(\theta,\beta \,|\, T_{ij}, N_{ij}, M_{ij}, X_{ij}, y_i),$$

[12] An introduction to competing risk models is given in Crowder (2001). See also Hougaard (2000), Lawless (2003), and Duffie et al. (2005).

where

$$L_Y(\theta) = L_{Y_1}(\theta_1) \cdot L_{Y_2}(\theta_2),$$

and

$$\begin{aligned} L(\theta, \beta \,|\, T_{ij}, N_{ij}, M_{ij}, X_{ij}, y_i) &= [y_{i1} \exp(X_{ij}(T_{ij})\beta_1)]^{N_{ij}} \exp(-y_{i1}\Lambda_{ij1}) \\ &\quad \times [y_{i2} \exp(X_{ij}(T_{ij})\beta_2)]^{M_{ij}} \exp(-y_{i2}\Lambda_{ij2}). \end{aligned} \tag{8}$$

Here Λ_{ij1} and Λ_{ij2} are the integrated hazards for default and competing risks respectively.

The likelihood function in expression (8) is separable. Maximum likelihood estimates of θ and β can be computed using an extension of the EM algorithm as outlined previously.

2.2 The Recovery Rate Models

Let $R_{ij}(t)$ be the recovery rate of the jth firm in the ith sector at time t. We assume that the recovery rate depends on the set of covariates $X_{ij}(t)$ through a function of the linear form $U_{ij}(t) = X_{ij}(t)\beta_r$, where β_r is a vector of regression coefficients. Recovery rates are non–negative and usually less than one.[13]

Several different approaches have been used in the literature to model the dependence of recovery rate on covariates. Acharya et al. (2003) and Varma and Cantor (2005) assume that

$$R_{ij}(t) = X_{ij}(t)\beta_r,$$

implying that $U_{ij}(t) \equiv R_{ij}(t)$ and the recovery rates are normally distributed and unconstrained. Schönbucher (2003b) models the recovery rate through a logit specification

$$R_{ij}(t) = \frac{1}{1 + \exp(X_{ij}(t)\beta_r)},$$

implying that $U_{ij}(t) \equiv \log(L_{ij}(t)/R_{ij}(t))$, where $L_{ij}(t) = 1 - R_{ij}(t)$ is the loss given default.

Andersen and Sidenius (2005) use the probit transformation:

$$R_{ij}(t) = \Phi(X_{ij}(t)\beta_r),$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. Then

[13] It is possible for recovery rates to be greater than one, especially if bond prices within one month of default are used. In our data set, four recovery rates were greater than one and these were eliminated for the empirical estimation.

$U_{ij}(t) \equiv \Phi^{-1}(R_{ij}(t))$.

With missing variables or measurement errors, we can write

$$U_{ij}(t) = X_{ij}(t)\beta_r + e_{ij},$$

where e_{ij} represents the effects of the missing variables.

3 Data Description

In this section we first describe the data sources and then discuss the covariates used at different stages of the analysis.

3.1 Data Sources

3.1.1 Bankruptcy, Defaults and Recovery Data

Bankruptcy is defined as the event that a company makes either a Chapter 7 or a Chapter 11 filing.[14] The initial source of bankruptcy data is the data set from Chava and Jarrow (2004). This database consists of all bankruptcy filings as reported in the Wall Street Journal Index (1962–1980), the SDC Database (Reorganizations module 1980–2002), SEC filings (1978–2002), CCH Capital Changes Reporter and New Generation Research. As such, the bankruptcy data includes most of the bankruptcy filings between 1962–2004 of publicly traded companies on either the NYSE, AMEX or NASDAQ stock exchanges. To our knowledge, this is the most comprehensive bankruptcy database available. In this paper we focus on bankruptcies during the 1980–2004 time period.[15]

Data on defaults and recovery is taken from the *Moody's Default Risk Service Database* for the period 1980–2004. In addition to the recovery data, this database has detailed issue level information. We supplement this information with the Mergent's Fixed Income Securities Database, when necessary. See Varma and Cantor (2005) and Covitz and Han (2004) for more details on Moody's DRS.

[14] A Chapter 11 filing does not necessarily imply that a company will file for Chapter 7 (liquidation).

[15] The Bankruptcy Reform Act of 1978 took effect on October 1, 1979, and substantially revamped bankruptcy practices. A strong business reorganization chapter was created, Chapter 11. This replaced the old Chapters X, XI and XII that were created by the 1898 Act and amended by the Chandler Act. In general, the Reform Act of 1978 made it easier for both businesses and individuals to file a bankruptcy and to reorganize.

Default in this paper refers to Moody's definition of default. A default is said to occur if

- There is a missed or delayed disbursement of interest and/or principal, including delayed payments made within a grace period.
- The company files for bankruptcy, administration, legal receivership, or other legal blocks to the timely payment of interest or principal.
- A distressed exchange occurs when:
  - the issuer offers bondholders a new security or a package of securities that represent a diminished financial obligation (such as preferred or common stock, or debt with a lower coupon or par amount, lower seniority, or longer maturity), or
  - the exchange has apparent purpose of helping the borrower avoid default.

3.1.2 Other Forms of Exit Data

We use the information in the CRSP de-listing files to determine other forms of exit. At any given point of time, a firm can be active, become acquired or merged into another firm, go bankrupt, or get liquidated and de-listed for performance or other reasons. We use the de-listing codes in the CRSP files to determine the nature of exit. Specifically, we focus on exit through merger and acquisitions and construct a separate database for exits other than default.

3.1.3 Corporate Data

The firm level balance sheet data is taken from quarterly COMPUSTAT (active and research) files for the period 1980–2004, and market data is taken from CRSP. Both the accounting and market data are lagged by one quarter, so that they are observable by the market at the beginning of each quarter. This is an attempt to ensure that at the time of estimation we use only the accounting and market data that is available to market participants at that time.

3.2 Covariates

3.2.1 Firm Level Factors

The following firm level variables are used in the default prediction models:

- *rsize* denotes the relative size of the firm and is defined as the logarithm of each firm's equity value divided by the total NYSE/AMEX market equity value. This variable is statistically significant in Shumway (2001), Chava and Jarrow (2004), Campbell et al. (2004), and Beaver, McNichols and Rhie (2005). The effect of this variable is expected to be negative, because the firm size might proxy for differential firm power with respect to the ability to negotiate with creditors. The larger the firm, the greater its ability to negotiate and the lower the probability of failure.

- *exret* denotes excess return and is defined as the return on the firm minus the value-weighted CRSP NYSE/AMEX index return. The monthly returns are cumulated to obtain the quarterly return.[16] This variable is statistically significant in Shumway (2001), Campbell et al. (2004), and Chava and Jarrow (2004). The effect of this variable is expected to be negative: the larger the excess rate of return, the lower the probability of default.

- *nita* represents the ratio of Net Income to Total Assets of the firm.[17] In Shumway (2001) this variable is not statistically significant, while Campbell et al. (2004), Beaver et al. (2005), and Chava and Jarrow (2004) find it significant. The effect of this variable is expected to be negative: the larger the ratio, the lower the probability of default.

- *tlta* represents the ratio of Total Liabilities to the Total Assets of the firm, and it is statistically significant in Shumway (2001), Campbell et al. (2004), and Beaver et al. (2005). The effect of this variable is expected to be positive: the larger the leverage ratio, the greater the probability of default.

- *retl* denotes the firm's trailing one year stock return. A negative relation is expected: the greater the return in the previous year, the stronger the firm and the lower the probability of default.

- *sigma* represents the standard deviation of daily stock returns of the previous quarter. This variable is significant in Shumway (2001), Campbell et al. (2004), and Chava and Jarrow (2004). The effect of this variable is expected to be positive: the larger the standard deviation, the greater the probability of default.

- *dd* represents the Distance to Default and is constructed similarly to Bharath and Shumway (2005). Note that this variable combines firm specific information and mar-

[16] Shumway (2001) uses an indicator function if the stock's cumulative excess returns have been in the lowest 5% of all the NYSE/AMEX stock returns during the last three years. This indicator is not statistically significant.

[17] We measure Total Assets as the book value of liabilities plus the market value of equity.

ket information. Hillegeist et al. (2004) examine the sensitivity of the default point to different measures of the liabilities due, and find that their results are relatively insensitive to the specification. The effect of this variable is expected to be negative: the greater the distance to default, the lower the probability of default. This variable is statistically significant in Campbell et al. (2004), Duffie et al. (2005), and Hillegeist et al. (2004).

3.2.2 Macro Economic Factors

Many different macroeconomic variables have been used in previous studies. Hillegeist et al. (2004) use the previous year economy wide default rate to calibrate the baseline hazard rate and find it to be an important variable. Duffie and Wang (2003) investigate personal income growth and find a negative and statistically significant relation. Duffie et al. (2005) consider the firm's trailing one year stock return, the trailing one year return on the S&P index, and the three month Treasury bill rate. In the presence of these covariates, other covariates are found to be statistically insignificant. Campbell et al. (2004) experiment with different NBER indicators of recession, though none improves the fit of their model. They also use the slope of the Treasury curve and the corporate bond spread, and find that several interaction terms between leverage, the Treasury slope and the credit spread are significant.

In this study we investigate the effects of six macroeconomic variables:

- *termspread*, computed as the difference of the ten year Treasury yield and the one year Treasury yield.
- *creditspread*, computed as the difference between AAA and BAA yields. The greater the spread, the higher is the aggregate credit risk of the economy. One would expect the greater the credit spread, the greater will be the probability of default for an individual obligor.
- *growth in real GDP* (Δgdp) — a negative relation is expected: the greater the growth in the economy, the lower the probability of default.
- *growth in personal income* (Δpi) — a negative relation is expected: the greater the growth in personal income, the stronger the economy and the lower the probability of default.

  The information on the last two variables is taken from the Federal Reserve's website.

- *the three month Treasury yield (tbsm3)* — a positive relation is expected: the greater the Treasury bill rate, the higher the probability of default.
- *the S&P 500 index trailing one year return (spretl)* — a negative relation is expected: the greater the return in the previous year, the stronger the economy and the lower the probability of default.

To avoid any outlier effects, all variables are winsorized at the 1% and 99% of the cross–sectional distributions.

3.3 Additional Recovery Covariates

The following additional covariates are used in the recovery rate models.

- *seniority* — we identify five classes of seniority: junior, subordinated, senior subordinated, senior unsecured and senior secured. The choice of these five classes was dictated by the availability of data.
- *log(issuesize)* — logarithm of the initial amount issued. Larger issues may earn higher recoveries than smaller issues, as a larger stakeholder may be able to exert greater bargaining power in the bankruptcy proceedings.
- *log(matoutstand)* — logarithm of time to maturity
- *couprate* — the coupon rate on the bonds at the time of default. Acharya et al. (2003) argue that if a bond is issued at a discount or premium, then the coupon on the bond will affect the accelerated amount payable to bondholders in bankruptcy,[18] as will the remaining maturity of the issue.
- *log(ta)* – logarithm of the size of the firm as measured by the total assets
- *mtb* denotes the market to book ratio of the firm. The variable is a proxy for the firm's growth prospects, and thus it should have a positive effect on recoveries.
- *ebitdasales* denotes the ratio of earnings before interest, tax, depreciation and amortization to the total sales of the firm, and it is a measure of the firm's profitability. Acharya et al. (2003) find a statistically significant effect of this variable. A positive relation is expected: the higher the ratio, the greater should be the recovery.

[18] A common clause in bond indentures is that the accelerated amount payable to bondholders in bankruptcy equals the remaining promised cash flows discounted at the original issue yield.

- *tanta* denotes the ratio of property plant and equipment to the total assets of the firm, and it is a measure of the firm's tangible assets. Acharya et al. (2003) find the effect of this variable to be statistically insignificant.

4 Empirical Results

In this section we discuss the results of estimating the default and recovery models described in Section 2, using annual data.

4.1 Default Prediction Results

We consider eight different models for the probability of default, spanning the array of current models that have been used by academics. The first model, M1, contains the same covariates as in Shumway (2001), except for firm age which Shumway found to be statistically insignificant. The second model, M2, is the reduced form model considered by Chava and Jarrow (2004). The third model, M3, is obtained by replacing the volatility variable *sigma* in M2 with a distance-to-default variable. The fourth model, M4, is obtained by adding two macroeconomic variables, the Treasury spread and the credit spread to model M1. The fifth model, M5, is obtained by adding the two macroeconomic variables to model M3, which includes the distance-to-default. The sixth model, M6, is a private firm model that does not utilize any equity market based variables. The seventh model, M7, is obtained from model M4 by adding two more macroeconomic variables: the change in real gross domestic product and the change in personal income. The eighth model, M8, uses the same covariates as in Duffie et al. (2005).

The analyses are done on two subsets of the data, the first consisting of all non–financial firms and the second consisting of only manufacturing firms.[19] In order to set a benchmark, we first estimate the parameters of all models without frailty. We then re-estimate the parameters of the models with frailty and assess their performance out-of-sample.

4.1.1 No frailty

Table 2 reports the results for the case when there is no frailty. Panel A presents the results for the sample of non-financial firms. For model M1 (Shumway, 2001), all covariates are

[19] A non–financial firm has a SIC code less than 6000 or greater than 7000. A manufacturing firm has a SIC code between 2000 and 4000.

significant and have the correct sign, except for net income to total assets.[20] All coefficients in model M2 (Chava and Jarrow, 2004) and model M3 have the expected sign and are significant, except for relative size of the firm in M3. The credit spread coefficient is significant and has the expected sign in model M4, and it is insignificant in model M5. The term spread coefficient is significant in model M5, although it has a positive sign. The addition of the credit spread and term spread variables in models M4 and M5 has only a small effect on the magnitudes of the other coefficients. For the private firm model M6, the credit spread covariate is insignificant, while the other coefficients are significant and have the expected sign. For model M7, the Treasury spread remains insignificant, while the change in real gross domestic product and the change in personal income are significant. Finally, for model M8 (Duffie et al, 2005) all covariates are statistically significant, except the three month Treasury bill rate. The coefficient for the one year trailing S&P 500 index is positive, consistent with findings in Duffie et al. (2005) who argue that this may be due to the correlation between the individual stock returns and the S&P 500 index, or perhaps to the trailing nature of the returns and the business cycle dynamics.

In Panel B we report the results of the analysis for the sample of manufacturing firms. The results are broadly similar; the net income to total assets covariate becomes insignificant in models M1 and M4, and the change in real gross domestic product is now insignificant in model M7.

As a measure of in–sample fit, the log-likelihood function is highest for models M1 and M4, suggesting that these are the two best fitting models for both the non-financials and the manufacturing samples. The difference in fit between M1 and M4 is not statistically significant.

4.1.2 One frailty per sector

For the second part of the analysis we assume that there is one frailty specific to each sector. We allocate firms to different sectors, first on the basis of the 4-digit SIC industry codes from COMPUSTAT,[21] then on the basis of the Fama–French sector classification.[22] The two classifications gave very similar results, the only substantial difference being a higher estimated frailty variance when firms are classified based on 4-digit SIC codes. This is to be expected, as the frailty variance is a measure of within sector homogeneity, and sectors are more ho-

[20]The coefficient of this variable is negative in Shumway(2001), though not statistically significant — see Table 6B.

[21]See Kahle and Walkling (1996) for the merits of using the Compustat versus CRSP SIC codes.

[22]There are 442 sectors defined by the 4-digit SIC industry codes, and 48 sectors defined by the Fama–French sector classification.

mogeneous when they are defined according to a more refined classification.[23] Consequently we only report the coefficients obtained from the classification using COMPUSTAT 4-digit industry codes.

Table 3, Panel A presents the results for the non-financial sample. The frailty variance is statistically significant for all models. In comparison to the results for the no-frailty models in Panel A from Table 2, the signs of all coefficients generally remain unchanged, and there are only small changes in the magnitudes of the coefficients. The net income to total assets covariate is now statistically insignificant except in the private firm model M6, while the term spread coefficient varies in sign and in statistical significance across models. The results for the sample of manufacturing firms reported in Panel B from Table 3 lead to similar insights.

For all models, there is a significant improvement in the log-likelihood function over the corresponding model without frailty, in both the non-financials and the manufacturing samples. A χ^2 test confirms that, for the same sets of covariates, a frailty default model provides a statistically significant improvement in fit over a default model without frailty.

4.1.3 Competing Risks

Table 4 reports the estimation results from fitting the competing risks models described in (6)–(7). This is the first study to examine the impact of exit due to reasons other than default on a range of default prediction models. In comparison with Table 3, the estimated frailty variance has decreased by about a factor of two, but still remains highly significant. The coefficients for exit due to default change little from those reported in Table 3.

For exit due to other reasons, the coefficients for the excess return, the net income to total assets, the change in real gross domestic product, the lagged stock return, and the lagged S&P return are all positive and statistically significant. The coefficients for the relative size, credit spread, and term spread are all negative and statistically significant in most models. The coefficients for sigma, total liabilities to total assets, and the three month Treasury yield are generally not significant. The results for model M8 are broadly consistent with the findings reported in Duffie et al. (2005).

[23] As the classification of the sectors becomes more refined, implying that firms within a sector become less heterogeneous, the dispersion across sectors increases. Hence the variance θ of the sector specific frailty Y increases.

4.2 Out-of-Sample Performance

We investigate the out-of-sample forecasting performance of the different default models, using a one year horizon which is suggested by regulatory requirements. We estimate the model coefficients using data between 1980–1995, then we compute the one-year probability of default for each firm at the beginning of each year in which the firm is alive between 1996–2004. The probabilities of default for every year are then ranked by descending order and grouped into deciles. We record the yearly number of actual defaults in each decile, and we compute the aggregated percentage of defaults in each decile over the 1996–2004 period for all models. The top two decile percentages are reported in Table 5, Panel A.

For the sample of non-financial firms, there is little difference in the out-of-sample performance for the models with and without frailty. Models M1 and M4 have the best performance without frailty, correctly identifying 77 percent of the defaulting firms in the first two deciles. This compares well with the results in Chava and Jarrow (2004), who correctly identify 79 percent[24] of the defaulting firms. With frailty, there is a minor deterioration in the out-of-sample performance of models M1 and M4. The best performance comes from model M7, which correctly identifies 79.8 percent of the defaulting firms.

For the more homogeneous sample of manufacturing firms, the out-of-sample performance of all models is significantly better. Models with frailty provide better or equal classification than models without frailty in all cases, although the differences tend to be small. Models M1 and M4 have again the best out-of-sample performance, correctly identifying 84 percent of the defaulting firms.

The NASDAQ Effect

Most academic studies, including Shumway (2001) and Beaver et al. (2005), restrict their empirical analysis to AMEX and NYSE traded firms. Many start-up firms, however, are listed on NASDAQ. These firms tend to have negative cash flows over the initial few years, and failure rates tend to be higher among start-ups than among established firms.[25] These considerations will, in general, affect the estimation of the coefficients and the performance of default prediction models when the sample data includes firms listed on NASDAQ.

This issue is examined in Chava and Jarrow (2004). First, they consider only firms with price data from AMEX or NYSE during the period 1962–1999, including 404 bankruptcies.

[24]Chava and Jarrow (2004) use data on 1,197 defaulting firms over the period 1962 to 1999. For out-of-sample testing, they estimate the coefficients over the period 1962 to 1990 and test the out-of-sample performance over the period 1991 to 1999.

[25]This was the especially the case in the 1990s and the crash of the dot com epoch in the early 2000s.

They test the out-of-sample performance of several models over the period 1991–1999, and find that the Shumway model identifies 86.40 percent of bankruptcies in the top two deciles. Next, they include in the analysis firms with price data from NASDAQ, and the size of their bankruptcy sample increases to 1,066. They find that the model with best performance in out-of-sample testing is now a public firm model with industry dummy variables, and this model identifies only 79.12 percent of bankrupt firms in the top two deciles.

To investigate the NASDAQ effect in our sample, we first ran the analysis on the unrestricted sample including firms listed on AMEX, NYSE and NASDAQ. The summary forecasting results are reported in Table 5, Panel A, and we discussed them earlier. Next, we restricted our analysis to firms trading only on AMEX and NYSE, and we report the forecasting results in Table 5, Panel B. Comparing Panel B with Panel A, the out-of-sample performance improves for all models, especially for models M1, M4 and M8. For non-financials, without frailty models M1 and M4 again give the best performance, identifying 84.5 percent of the defaulting firms in the top two deciles. This compares favorably with Shumway (2001) who identifies 87.5 percent of the bankrupt firms,[26] and with Beaver et al. (2005) who identify 88.1 percent of the bankrupt firms in the top two deciles.[27] With frailty, models M1 and M4 identify 83.2 percent of the defaulting firms in the top two deciles, while the best performance with 84.5 percent is given by model M7. For manufacturing, without frailty models M1 and M4 correctly identify 86.3 percent of the defaulting firms, and with frailty they still give the best performance, identifying 89.2 percent of the defaults in the top two deciles.

The numerical differences between the results in Panels A and B are of the order of 7%, which is substantial. The results illustrate, once again, the basic point that model performance depends crucially on the characteristics of the data set.[28] For regulators, the important message to be gleaned from these results is that while a bank can demonstrate the performance of a model on a particular data set, there is no guarantee that it will have similar performance on another data set.

[26]Shumway (2001) uses a sample of 239 bankruptcies over the period 1962 to 1992 for estimating the default model and tests the out-of-sample performance of his model over the period 1984 to 1992.

[27]Beaver et al. (2005) use a sample of 544 bankrupt firms covering the period 1962 to 2002. Using a model similar to Shumway (2001), they estimate the coefficients over the period 1962 to 1993 and test the out-of-sample performance over the period 1994 to 2002.

[28]This point is also demonstrated in Beaver et al. (2005). They divide randomly their data set for the period 1962 to 1993 into two subsamples, then they estimate the model coefficients using data from one subsample and test the model performance on correctly identifying the bankruptcies in the other subsample. A similar exercise is then undertaken using data for the period 1994 to 2002. The percentage of bankrupt firms correctly identified in the top two deciles varies between 76.90 and 85.60 percent.

4.3 Recovery Rate Results

Our analysis follows that of Acharya et al. (2003) and considers contract specific, firm specific, and macroeconomic variables, all of which have been described in Section 3.2.

The empirical results are shown in Table 6. We test the linear, logit and probit specifications described in Section 2.2. The estimation results for the logit model are similar to the ones for the probit model, and the out-of-sample performance of the probit model is slightly superior to the performance of the logit model. Consequently we only report the linear model results in Panel A and the probit model results in Panel B. For all models, the dummy variables representing seniority are positive and are statistically and economically important. The ordering is as expected for the dummy variables representing the senior secured, senior unsecured, and senior subordinated recovery rates. However, the coefficient for senior subordinated is less than the coefficient for subordinated recovery rate. This is not suprising, as the mean senior subordinated recovery rate in the sample is 28.49 cents per dollar (72 observations), compared to the mean subordinated recovery rate of 29.24 cents per dollar (145 observations). In Acharya et al. (2005), the senior secured coefficient is less than the senior unsecured coefficient and both are statistically significant, while the coefficients for the other classes of seniority are not statistically significant. We find that the coefficient of the coupon variable is positive and statistically significant, while in Acharya et al. (2005) it is not significant. The coefficients for the logarithm of issue size and for the maturity outstanding are insignificant, similar to Acharya et al. (2005). We find that all of the firm specific variables are insignificant. Among the macroeconomic variables, the three month Treasury bill yield is significant, while the lagged S&P return has the expected positive sign and is borderline significant in several models.

In order to assess the out-of-sample performance of the recovery rate models, we choose the initial sample period 1980–1995, then generate rolling one year ahead forecasts for the period 1996–2004. The average root mean square error for each of the models is reported in the last row of Table 6. We do not know of any extant studies that examine the out-of-sample performance of recovery rates models, and consequently we have no benchmark. The out-of-sample RMSE values are very similar across the six recovery rate models, and also between the linear and probit specifications.[29]

[29]However, in the simulations used to estimate the expected loss over arbitrary horizons — discussed in Section 5.1 — the linear model often led to negative values for estimated future recovery rates, because in the linear formulation the recovery rate dependent variable is unconstrained.

5 Expected Loss

In this section we outline the methodology for computing the expected loss over arbitrary horizons, using the estimated parameters from the default and recovery models. The probability of default and the loss given default for firm j both depend on the set of covariates $\{X_j(t)\}$. Let I_{jt} denote an indicator function that equals 1 if firm j defaults in period t conditional on survival up to period t, and 0 if default does not occur in period t. The loss in period t is

$$\widehat{L}_{jt} = \begin{cases} L_{jt}; & I_{jt} = 1 \\ 0; & I_{jt} = 0 \end{cases}$$

The expected loss over a one period horizon is

$$E_t[\widehat{L}_{jt+1}] = E_t[I_{jt} \cdot L_{jt}] = p_{jt+1}(X_j(t)) \cdot E_t[L_{jt+1}(X_j(t))],$$

where $p_{jt+1}(X_j(t))$ represents the probability of default over the next period $t+1$ conditional on survival up to $t+1$.

The expected loss in the second period is

$$E_t[\widehat{L}_{jt+2}] = E_t\{[1 - p_{jt+1}(X_j(t))] \cdot p_{jt+2}(X_j(t+1)) \cdot L_{jt+2}(X_j(t+1))\},$$

where the expectation is taken over the stochastic processes of the covariates $\{X_j(t+1)\}$. To compute the expected loss we simulate the paths of the covariates, calculate the term

$$[1 - p_{jt+1}(X_j(t))] \cdot p_{jt+2}(X_j(t+1)) \cdot L_{jt+2}(X_j(t+1)),$$

and estimate its mean value.

5.1 Default and Recovery Correlation

We have no direct benchmark against which to calibrate our expected loss model. Consequently, we examine some of its properties. Empirical evidence shows that the frequency of default and the loss given default are negatively correlated, therefore we examine the correlation between the probability of default and the expected loss given default implied by our model.

Similar to the methodology for out-of-sample testing, we estimate the parameters for the probability of default model M4 and for the recovery rate Model 3 (under the probit

specification) over the period 1980–1995, separately for manufacturing and for non-financial firms. For every year t during the out-of-sample period 1996–2004, we estimate the one-year conditional probability of default and the expected recovery rate[30] during year t for every firm that is alive at the beginning of year t. We then compute the cross-sectional correlation in year t between the estimated probability of default and the expected recovery rate, denoted by $\rho(t) = corr(p(t), R(t))$. Figure 1 gives a time series plot of the cross correlations $\rho(t)$ and shows that for all nine years the correlations are negative and significant, as expected. The magnitude of the correlation varies with the credit cycle. The correlation decreases as the economy slides into recession during the late 1990s, and increases as the economy improves during the new millennium.

5.2 Estimation Results with Quarterly Data

We estimate the default models M1–M8 with frailty using quarterly data, for purposes of comparison with the similar analyses using annual data. The results are given in Table 7, Part A for the non-financial sample and Part B for the manufacturing sample. They are broadly similar with the corresponding results for annual data reported in Table 4. The only differences are that the coefficient of net income to total assets is now significant in models M1 and M4, and the coefficient of the three month Treasury yield becomes significant in model M8, both for the non-financials and the manufacturing samples.

5.3 Multiperiod Out-of-Sample Performance

The methodology that we developed in Section 2 can be used for estimating the expected loss over arbitrary horizons, rather than just over one period. This is particularly important for banks, since they typically examine their risk exposure over multi–year horizons. For modelling expected loss over arbitrary horizons it is also necessary to estimate the stochastic

[30] For the linear case the expected recovery is

$$E_t[R(t+1)] = X(t)\beta.$$

For the probit case we have

$$E_t[R(t+1)] = E_t[\Phi(X(t)\beta + \sigma_e e(t))],$$

where $e(t)$ is the error term with zero mean and unit variance, σ_e is the standard deviation of the error term, and $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. Evaluating the above expression gives

$$E_t[R(t+1)] = \Phi(X(t)\beta\varpi),$$

where $\varpi \equiv 1/(1+\sigma_e^2)^{1/2}$.

processes describing the evolution of the covariates, in addition to estimating the parameters for default and recovery models.

We choose a one-year horizon and we compare the out-of-sample performance of default model M4 estimated with quarterly data with the performance of model M4 estimated with annual data. Note that in the case of quarterly data, the one-year horizon is in fact a four period horizon over which we need to simulate the evolution of the covariate processes. In the case of annual data, the one-year horizon is a one period horizon over which the last observed covariate values will be used for predicting default probabilities; therefore it is not necessary to simulate the evolution of the covariate processes in this case.

We estimate the coefficients in model M4 using both quarterly and annual data over the initial period 1980–1995. For the case of quarterly data, we assume that the seven state variables in model M4 follow a multivariate AR(1) process. Note that for the economy level covariates the entire period 1980–1995 is available for estimation, while for firm level covariates we only have available the history during the lifetime of the firm. The obligors with short lifetimes did not have sufficient data to allow the estimation of the multivariate time series process. Therefore, we first restricted the sample to those firms with at least eight quarters of data. To investigate the impact of available history on model performance, we next repeated the analysis on the sample restricted to those firms with at least twenty quarters of data; this substantially altered the size of the sample.[31] For every year between 1996–2004 and each obligor who is in the sample at the beginning of that year, we jointly estimate the coefficients of the multivariate time series process for the underlying covariates.[32] The obligor's probability of default over the next one year horizon is then estimated using simulation, and the procedure for assessing the out–of–sample performance of the default model becomes similar to that described in Section 4.2.

The results are shown in Table 8, where Panel A reports the results for non-financials and Panel B for industrials. Two salient features are noteworthy. First is the NASDAQ effect: the model performance improves substantially when we remove the NASDAQ traded firms and thus decrease the heterogeneity in the sample. This is not unexpected, given the discussion of the NASDAQ effect in Section 4.2. Second is the history length effect: the model performance improves again substantially when using at least twenty quarters of data for the estimation of the parameters of the state variables process. This is also not surprising,

[31] In the case of a filter of 8 quarters, the restricted manufacturing sample contains 5,218 firms and 427 defaults. The restricted non-financials sample contains 11,143 firms and 1,058 defaults.

In the case of a filter of 20 quarters, the restricted manufacturing sample contains 3,619 firms and 272 defaults. The restricted non-financials sample contains 7,206 firms and 615 defaults.

[32] Estimation was performed using the software described in Neumaier and Schneider (2001).

given that model performance depends on the ability to accurately describe the evolution of the different correlated covariates. To estimate even a simple AR(1) process with only eight observations is problematic. Unfortunately, defaults often occur for young firms, so that the corresponding time series data are limited. This implies that a model that relies only on market driven covariates, such as M8, has perhaps an advantage over firm specific type models such as M4. However, the loan portfolios of many banks consist of a majority of private firms, and hence model M8 is not applicable.

6 Summary

This paper develops a methodology for modeling and estimating expected loss over arbitrary horizons in the presence of unobservable heterogeneity in firm characteristics. In this framework we model jointly the probability of default and the recovery rate given default. Unobservable heterogeneity, representing the effects of measurement errors and missing variables, is modeled as a non-negative latent random variable that acts multiplicatively on the default intensity function. Since firms in the same industry share the same latent random variable, this specification induces within–industry correlation of default intensities. We estimate the parameters of the different models using an extensive default and recovery data set, containing the majority of defaults of companies listed on the AMEX, NYSE and NASDAQ exchanges between 1980–2004. We also allow for exit due to reasons other than default, and examine how this affects parameter estimation for the default prediction models.

We undertake extensive out-of-sample performance tests for both default prediction models and recovery rate given default models. This is the first study to test the out-of-sample performance of models for the recovery rate given default. Our joint model specification implies that out-of-sample, the probability of default and the recovery rate given default are negatively correlated, and the magnitude of the correlation varies with the credit cycle. Using our methodology, it is straightforward to estimate the expected loss over arbitrary horizons instead of simply over a one-period horizon. We compare the out-of-sample default predictions using quarterly and annual data, and find that the performance over multi-periods is sensitive to the estimation accuracy for the parameters of the covariate processes. This has practical implications for the choice of models, given data availability.

Under the advanced internal ratings based approach of the new Basel Capital Accord (Basel II), banks can use their own estimates of the probability of default and loss given default, subject to regulatory approval. The Accord is not explicit as to all the steps that

banks must take to gain regulatory approval, instead it requires banks to present enough evidence about the properties and performance of their models to satisfy regulators. This paper provides banks with a framework that directly addresses the regulatory concerns about demonstrating the performance and robustness of models for expected loss.

References

[1] Acharya, V. V., S. T. Bharath, and A. Srinivasan (2003), Understanding the recovery rates on defaulted securities, Working paper, London Business School.

[2] Altman, E. I., and E. Hotchkiss (2005), *Corporate Financial Distress and Bankruptcy*, Third Edition, Wiley, New York.

[3] Altman, E. I., B. Brady, A. Resti, and A. Sironi (2005), The link between default and recovery rates: theory, empirical evidence and implications, *Journal of Business* 78, 2203–2227.

[4] Andersen, L., and J. Sidenius (2005), Extensions to the Gaussian copula: random recovery and random factor loadings, *Journal of Credit Risk* 1, 29–70.

[5] Basel Committee on Banking Supervision (2005a), *Guidance on Paragraph 468 of the Framework Document.*

[6] Basel Committee on Banking Supervision (2005b), *Studies on the Validation of the Internal Rating Systems*, Working paper N0. 14.

[7] Beaver, B., M. McNichols, and J. Rhie (2005), Have financial statements become less informative?, Working paper, Stanford University.

[8] Bharath, S. T., and T. Shumway (2005), Forecasting default with the KMV Merton model, Working paper, University of Michigan.

[9] Campbell, J. Y., J. Hilscher, and J. Szilagyi (2004), In search of distress risk, Working paper, Harvard University.

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

[11] Covitz, D., and S. Han (2004), An empirical analysis of bond recovery rates: exploring a structural view of default, Working paper, The Federal Reserve Board of Washington.

[12] Crowder, M. (2001), *Classical Competing Risks*, Chapman & Hall, New York.

[13] Dempster, A. P., N. M. Laird, and D. B. Rubin (1977), Maximum likelihood from incomplete data via the EM algorithm (with discussion), *Journal of the Royal Statistical Society B* 39, 1–38.

[14] Duffie, D., and K. Wang (2003), Multi–period corporate failure prediction with stochastic covariates, Working paper, Stanford University.

[15] Duffie, D., L. Saita, and K. Wang (2005), Multi–period corporate failure prediction with stochastic covariates, forthcoming *Journal of Financial Economics*.

[16] Dullmann, K., and M. Trapp (2004), Systematic risk in recovery rates — an empirical analysis of U.S. corporate credit exposure, Working paper, Deutsche Bundesbank, Frankfurt, Germany.

[17] Frye, J. (2000), Depressing recoveries, *Risk Magazine* 13, 108–111.

[18] Gagliardini, P., and C. Gourieroux (2003), Spread term structure and default correlation, Working paper, University of Toronto.

[19] Hillegeist, S. A., E. K. Keating, and D. P. Cram (2004), Assessing the probability of default, *Review of Accounting Studies* 9, 5–34.

[20] Hougaard, P. (2000), *Analysis of Multivariate Survival Data*, Springer, New York.

[21] Janosi, T., R. Jarrow, and Y. Yildirim, (2002), Estimating the expected losses and liquidity discounts implicit in debt prices, *Journal of Risk* 5, 1–38.

[22] Kahle, K., and R. A. Walkling (1996), The impact of industry classification on financial research, *Journal of Financial and Quantitative Analysis* 31, 309–335.

[23] Kiefer, N. M. (1988), Economic duration data and hazard functions, *Journal of Economic Literature* Vol. XXV1, 646–679.

[24] Klein, J. P., and M. Moeschberger, (1997), *Survival Analysis*, Springer, New York.

[25] Lawless, J. L., (2003), *Statistical Models and Methods for Lifetime Data*, John Wiley & Sons, New Jersey.

[26] Neumaier, A., and T. Schneider (2001), Estimation of parameters and eigenmodes of multivariate autoregressive models, *ACM Transactions on Mathematical Software*, 27, 27–57.

[27] Pykhtin, M. (2003), Unexpected recovery risk, *Risk* **16**, 74–78.

[28] Schönbucher, P. J. (2003a), Information driven default contagion, Working paper, ETH, Zurich.

[29] Schönbucher, P. J. (2003b), *Credit Derivatives Pricing Models*, John Wiley & Sons Ltd. New Jersey.

[30] Schuermann, T. (2004), What do we know about loss given default. In *Credit Risk Models and Management*, Risk Books, London.

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

[32] Stefanescu, C., and B. W. Turnbull (2006), Multivariate frailty models for exchangeable survival data with covariates, forthcoming in *Technometrics*.

[33] Varma, P., and R. Cantor (2005), Determinants of recovery rates on defaulted bonds and loans for North American corporate issuers: 1983–2003, *Journal of Fixed Income* 14, 29–44.

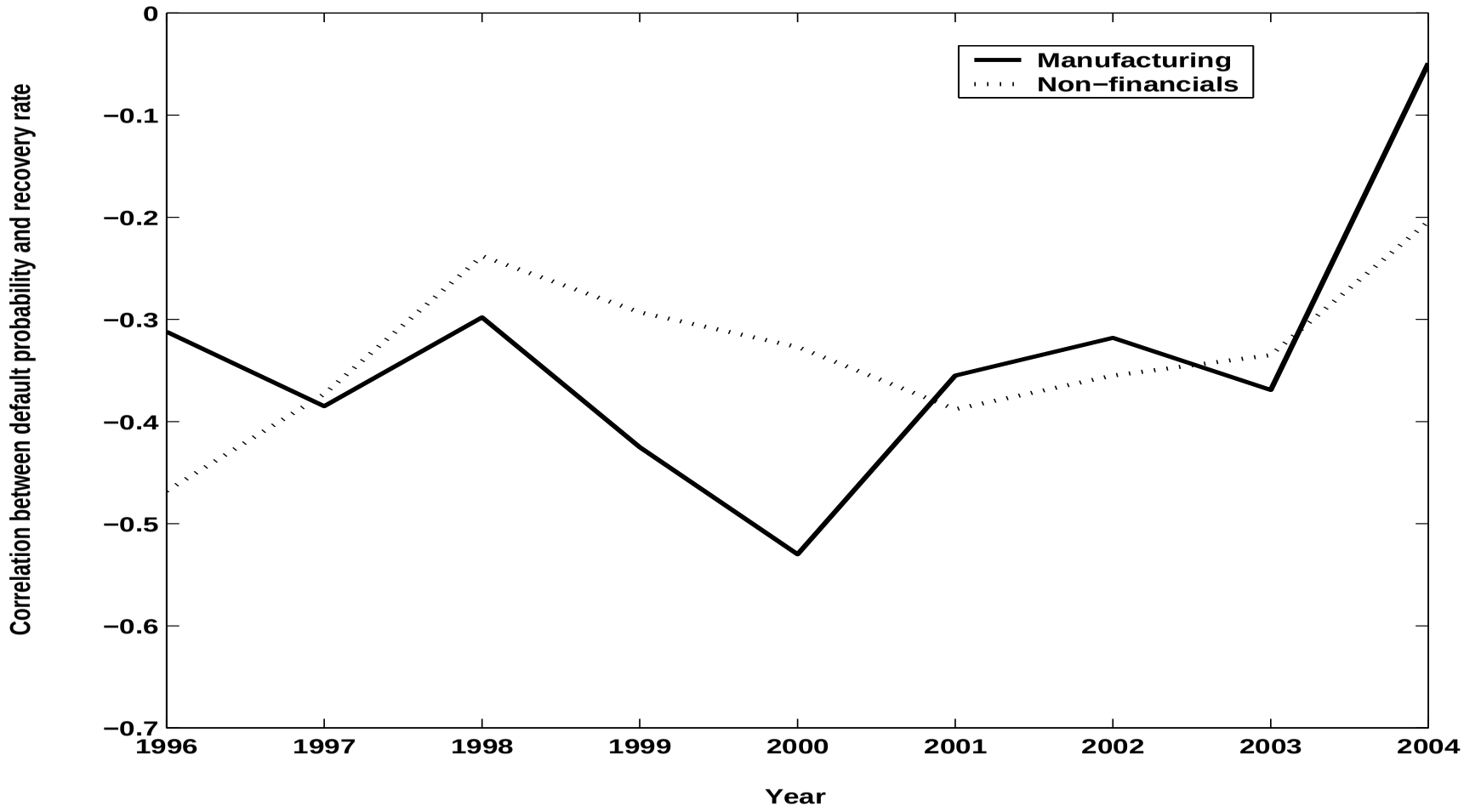

Figure 1: Correlation between estimated default probabilities and recovery rates for every year between 1996–2004 for manufacturing and non-financial firms.

Table 1: **Descriptive Statistics**

The following table presents the descriptive statistics for the variables used in the default and recovery estimation. The sample is restricted to *non-financial* firms with information in CRSP and COMPUSTAT during 1980-2004. *exret* denotes excess return and is defined as the return on the firm minus the value-weighted CRSP NYSE/AMEX index return. *rsize* denotes the relative size of the firm and is defined as the logarithm of each firm's equity value divided by the total NYSE/AMEX market equity value. *sigma* represents the standard deviation of daily stock returns of the previous quarter. *nita* represents the ratio of Net Income to Total Assets of the firm. *tlta* represents the ratio of Total Liabilities to the Total Assets of the firm. *dd* represents the Distance to Default, constructed similarly to Bharath and Shumway (2005). *retl* denotes the firm's trailing one year stock return. *termspread* is the difference of the ten year Treasury yield and the one year Treasury yield, *creditspread* is the difference between AAA and BAA yields. Δgdp is the growth in real GDP. Δpi is the growth in personal income. *tbsm3* is the three month Treasury yield. *spretl* is the trailing one year S&P 500 index return. *rec* is the recovery on the bond (bond price within a month after default as given by Moody's DRS). *couprate* is the coupon rate on the bonds. *log(issuesize)* is the logarithm of the initial amount of the bond issued. *log(matoutstand)* is the logarithm of time to maturity. *ebitdasales* denotes the earnings before interest, tax, depreciation and amortization to the total sales of the firm. *tanta* denotes the ratio of property plant and equipment to the total assets of the firm. *mtb* is the market to book ratio of the firm.

	mean	25^{th} pctl	50^{th} pctl	75^{th} pctl
Firm level covariates				
exret	0.0248	-0.3134	-0.0706	0.1941
rsize	-10.9699	-12.3407	-11.1137	-9.7547
sigma	0.5876	0.3089	0.4794	0.7378
nita	-0.0372	-0.0241	0.0272	0.0710
tlta	0.5492	0.3423	0.5466	0.7298
dd	6.7301	-0.4435	5.3237	10.4771
retl	0.1842	-0.2193	0.0722	0.3769
Macroeconomic covariates				
termspread	1.1161	0.2800	0.8700	2.1600
creditspread	1.0865	0.6900	0.9800	1.2800
Δgdp	3.4287	1.9000	3.1000	5.4000
Δpi	6.1205	3.8000	6.6000	8.1000
*tbsm*3	5.7974	4.3900	5.2000	7.6300
spretl	0.1160	-0.0154	0.1462	0.2633
Recovery covariates				
rec	0.3434	0.1440	0.2800	0.5237
couprate	9.7648	7.8750	9.8750	12.0000
log(issuesize)	4.8642	4.3175	4.8283	5.4380
log(matoutstand)	8.2765	8.0548	8.2041	8.3859
ebitdasales	-0.0326	0.0164	0.0739	0.1778
tanta	0.4100	0.2238	0.3805	0.5737
mtb	1.2755	0.9684	1.1307	1.4004

Table 2: **Default Estimation: No Frailty**

Panel A: Non-financials. Annual Data.

The following table presents the estimates for the default prediction model with no frailty and exponential hazards. Standard errors are given in parantheses. The sample is restricted to *non-financial* firms with information in CRSP and COMPUSTAT during 1980-2004. Variable definitions are given in Table 1. The models M1 to M8 differ in the specification of the covariates. The last row gives the out-of-sample forecasting accuracy for each model. Data from 1980–1995 is used to compute the parameter estimates for 1996, then rolling 1-year ahead estimates are generated for 1997–2004. Every year, firms are ranked into deciles according to their estimated probability of default, and the aggregate percentages of defaults in the top two deciles are presented for each model.

	M1	M2	M3	M4	M5	M6	M7	M8
intercept	-7.746	-6.641	-4.367	-7.930	-4.312	-5.563	0.518	-3.781
	(0.190)	(0.182)	(0.204)	(0.208)	(0.233)	(0.095)	(0.178)	(0.088)
exret	-2.738	-2.775	-2.088	-2.734	-2.064		-0.688	
	(0.094)	(0.093)	(0.089)	(0.094)	(0.090)		(0.061)	
sigma	0.199	0.215		0.205			0.374	
	(0.024)	(0.022)		(0.024)			(0.011)	
rsize	-0.106	- 0.117	-0.022	-0.108	-0.013		0.382	
	(0.016)	(0.016)	(0.016)	(0.016)	(0.017)		(0.012)	
nita	0.231			0.221		-0.730	4.311	
	(0.081)			(0.081)		(0.070)	(0.219)	
tlta	2.252			2.247		2.224	2.553	
	(0.088)			(0.088)		(0.090)	(0.087)	
*tbsm*3								-0.014
								(0.010)
creditspread				0.159	0.007	-0.010	-0.354	
				(0.063)	(0.064)	(0.061)	(0.085)	
Δgdp							0.052	
							(0.012)	
Δpi							-0.135	
							(0.007)	
termspread				-0.022	0.088	-0.083	-0.018	
				(0.023)	(0.022)	(0.021)	(0.025)	
retl								-1.752
								(0.078)
spretl								0.888
								(0.162)
dd			-0.183		-0.191		-0.391	-0.196
			(0.009)		(0.010)		(0.007)	(0.010)
Log L	-6,284	-6,604	-6,398	-6,280	-6,390	-7,162	-7,500	-6,414
Out–of–sample performance								
Top 2 deciles (%)	76.97	71.10	74.82	77.25	74.67	51.50	74.68	75.11

Table 2: **Default Estimation: No Frailty**

Panel B: Manufacturing. Annual Data.

The following table presents the estimates for the default prediction model with no frailty and exponential hazards. Standard errors are given in parantheses. The sample is restricted to *manufacturing* firms with information in CRSP and COMPUSTAT during 1980-2004. Variable definitions are given in Table 1. The models M1 to M8 differ in the specification of the covariates. The last row gives the out-of-sample forecasting accuracy for each model. Data from 1980–1995 is used to compute the parameter estimates for 1996, then rolling 1-year ahead estimates are generated for 1997–2004. Every year, firms are ranked into deciles according to their estimated probability of default, and the aggregate percentages of defaults in the top two deciles are presented for each model.

	M1	M2	M3	M4	M5	M6	M7	M8
intercept	-8.711	-7.803	-5.820	-9.084	-6.141	-6.118	2.836	-4.249
	(0.318)	(0.315)	(0.354)	(0.347)	(0.401)	(0.153)	(0.274)	(0.145)
exret	-2.528	-2.688	-2.231	-2.529	-2.251		-1.624	
	(0.153)	(0.153)	(0.148)	(0.154)	(0.151)		(0.123)	
sigma	0.284	0.332		0.313			0.911	
	(0.064)	(0.052)		(0.064)			(0.035)	
rsize	-0.163	- 0.189	-0.114	-0.166	-0.118		0.609	
	(0.027)	(0.028)	(0.028)	(0.027)	(0.028)		(0.019)	
nita	0.196			0.188		-0.737	0.712	
	(0.124)			(0.125)		(0.107)	(0.147)	
tlta	2.390			2.380		2.545	2.635	
	(0.140)			(0.141)		(0.143)	(0.127)	
*tbsm*3								0.009
								(0.015)
creditspread				0.319	0.200	0.087	-1.346	
				(0.100)	(0.102)	(0.096)	(0.156)	
Δgdp							-0.031	
							(0.020)	
Δpi							-0.097	
							(0.012)	
termspread				-0.033	0.056	-0.086	-0.059	
				(0.036)	(0.035)	(0.033)	(0.040)	
retl								-2.011
								(0.133)
spretl								0.666
								(0.265)
dd			-0.160		-0.160		-0.284	-0.181
			(0.015)		(0.015)		(0.013)	(0.015)
Log L	-2,477	-2,623	-2,567	-2,472	-2,563	-2,809	-2,924	-2,582
Out–of–sample performance								
Top 2 deciles (%)	83.54	72.01	74.89	83.95	74.48	65.02	72.43	74.90

Table 3: **Default Estimation: One Frailty per Sector**

Panel A: Non-financials. Annual Data.

The following table presents the estimates for the default prediction model with one frailty per sector and exponential hazards. Standard errors are given in parantheses. The sample is restricted to *non-financial* firms with information in CRSP and COMPUSTAT during 1980-2004. Variable definitions are given in Table 1. The models M1 to M8 differ in the specification of the covariates. The last row gives the out-of-sample forecasting accuracy for each model. Data from 1980–1995 is used to compute the parameter estimates for 1996, then rolling 1-year ahead estimates are generated for 1997–2004. Firms are ranked into deciles according to their estimated probability of default, and the aggregate percentages of defaults in the top two deciles are presented for each model.

	M1	M2	M3	M4	M5	M6	M7	M8
frailty variance	0.256	0.338	0.288	0.254	0.294	0.397	0.223	0.290
	(0.043)	(0.048)	(0.043)	(0.042)	(0.044)	(0.056)	(0.040)	(0.044)
intercept	-7.582	-6.629	-4.498	-7.751	-4.348	-5.411	-5.338	-3.602
	(0.200)	(0.189)	(0.213)	(0.217)	(0.244)	(0.105)	(0.271)	(0.096)
exret	-2.671	-2.778	-2.137	-2.670	-2.103		-1.967	
	(0.095)	(0.094)	(0.092)	(0.095)	(0.092)		(0.094)	
sigma	0.216	0.255		0.221			0.164	
	(0.026)	(0.024)		(0.025)			(0.036)	
rsize	-0.103	-0.124	-0.038	-0.106	-0.025		0.012	
	(0.017)	(0.016)	(0.017)	(0.017)	(0.017)		(0.018)	
nita	0.033			0.027		-0.858	0.016	
	(0.085)			(0.086)		(0.077)	(0.086)	
tlta	2.092			2.088		2.109	1.902	
	(0.095)			(0.095)		(0.096)	(0.096)	
*tbsm*3								-0.028
								(0.010)
creditspread				0.140	-0.035	-0.035	0.179	
				(0.064)	(0.066)	(0.061)	(0.080)	
Δgdp							0.050	
							(0.012)	
Δpi							-0.038	
							(0.009)	
termspread				-0.015	0.104	-0.078	0.013	
				(0.023)	(0.035)	(0.021)	(0.026)	
retl								-1.775
								(0.080)
spretl								0.820
								(0.165)
dd			-0.175		-0.186		-0.158	-0.193
			(0.010)		(0.010)		(0.010)	(0.010)
Log L	-4,644	-4,611	-4,581	-4,647	-4,544	-5,112	-4,572	-4,582
Out–of–sample performance								
Top 2 deciles (%)	75.68	71.38	74.72	75.68	74.68	53.07	79.79	75.11

Table 3: **Default Estimation: One Frailty per Sector**

Panel B: Manufacturing. Annual Data.

The following table presents the estimates for the default prediction model with one frailty per sector and exponential hazards. Standard errors are given in parantheses. The sample is restricted to *manufacturing* firms with information in CRSP and COMPUSTAT during 1980-2004. Variable definitions are given in Table 1. The models M1 to M8 differ in the specification of the covariates. The last row gives the out-of-sample forecasting accuracy for each model. Data from 1980–1995 is used to compute the parameter estimates for 1996, then rolling 1-year ahead estimates are generated for 1997–2004. Firms are ranked into deciles according to their estimated probability of default, and the aggregate percentages of defaults in the top two deciles are presented for each model.

	M1	M2	M3	M4	M5	M6	M7	M8
frailty variance	0.238	0.323	0.297	0.230	0.305	0.335	0.232	0.295
	(0.063)	(0.070)	(0.067)	(0.062)	(0.068)	(0.077)	(0.062)	(0.067)
intercept	-8.640	-7.803	-5.914	-8.990	-6.113	-5.973	-7.194	-3.672
	(0.331)	(0.321)	(0.363)	(0.358)	(0.411)	(0.165)	(0.448)	(0.093)
exret	-2.509	-2.740	-2.292	-2.515	-2.296		-2.070	
	(0.154)	(0.156)	(0.152)	(0.155)	(0.154)		(0.154)	
sigma	0.266	0.319		0.290			0.253	
	(0.064)	(0.051)		(0.063)			(0.084)	
rsize	-0.168	- 0.199	-0.127	-0.171	-0.125		-0.076	
	(0.028)	(0.028)	(0.028)	(0.028)	(0.029)		(0.031)	
nita	-0.003			0.000		-0.909	-0.060	
	(0.132)			(0.132)		(0.119)	(0.133)	
tlta	2.259			2.254		2.437	2.081	
	(0.149)			(0.150)		(0.153)	(0.151)	
tbsm3								-0.008
								(0.016)
creditspread				0.284	0.145	0.035	0.270	
				(0.100)	(0.103)	(0.096)	(0.125)	
Δgdp							0.029	
							(0.019)	
Δpi							-0.005	
							(0.014)	
termspread				-0.018	0.083	-0.075	0.035	
				(0.036)	(0.035)	(0.034)	(0.041)	
retl								-2.043
								(0.135)
spretl								0.592
								(0.267)
dd			-0.151		-0.156		-0.162	-0.178
			(0.015)		(0.015)		(0.010)	(0.015)
Log L	-2,141	-2,188	-2,157	-2,143	-2,148	-2,407	-2,098	-2,171
Out–of–sample performance								
Top 2 deciles (%)	83.54	76.13	79.01	84.37	79.02	66.67	75.18	79.01

Table 4: **Default Estimation: One Frailty per Sector, Competing Risks.**

Panel A: Non-financials. Annual Data.

The following table presents the estimates for the default prediction model with one frailty per sector, exponential hazards and competing risks. Standard errors are given in parantheses. The sample is restricted to *non-financial* firms with information in CRSP and COMPUSTAT during 1980-2004. Variable definitions are given in Table 1. The models M1 to M8 differ in the specification of the covariates. The last row gives the out-of-sample forecasting accuracy for each model. Data from 1980–1995 is used to compute the parameter estimates for 1996, then rolling 1-year ahead estimates are generated for 1997–2004. Firms are ranked into deciles according to their probability of default, and the aggregate percentages of defaults in the top two deciles are presented for each model.

	M1		M2		M3		M4	
	Default	Merger	Default	Merger	Default	Merger	Default	Merger
frailty variance	0.133		0.165		0.149		0.132	
	(0.017)		(0.018)		(0.017)		(0.017)	
intercept	-7.611	-3.603	-6.611	-3.459	-4.457	-3.666	-7.779	-3.025
	(0.197)	(0.104)	(0.186)	(0.095)	(0.210)	(0.109)	(0.215)	(0.113)
exret	-2.694	0.037	-2.797	0.039	-2.145	0.035	-2.692	0.038
	(0.095)	(0.008)	(0.094)	(0.008)	(0.091)	(0.008)	(0.095)	(0.007)
sigma	0.214	-0.058	0.252	-0.136			0.219	-0.109
	(0.026)	(0.044)	(0.024)	(0.044)			(0.025)	(0.046)
rsize	-0.103	-0.039	-0.120	-0.032	-0.033	-0.035	-0.106	-0.037
	(0.017)	(0.009)	(0.016)	(0.009)	(0.016)	(0.009)	(0.017)	(0.009)
nita	0.068	0.441					0.062	0.380
	(0.085)	(0.079)					(0.085)	(0.079)
tlta	2.116	0.047					2.112	0.046
	(0.093)	(0.067)					(0.093)	(0.067)
creditspread							0.141	-0.372
							(0.063)	(0.037)
termspread							-0.017	-0.126
							(0.023)	(0.012)
dd					-0.176	0.012		
					(0.009)	(0.002)		
Log L	-17,344		-16,738		-16,937		-17,267	
Out–of–sample performance								
Top 2 deciles (%)	77.68		72.82		75.68		77.68	

Table 4: **Default Estimation: One Frailty per Sector, Competing Risks. Panel A: Non-financials. Annual Data (contd.).**

The following table presents the estimates for the default prediction model with one frailty per sector, exponential hazards and competing risks. Standard errors are given in parantheses. The sample is restricted to *non-financial* firms with information in CRSP and COMPUSTAT during 1980-2004. Variable definitions are given in Table 1. The models M1 to M8 differ in the specification of the covariates. The last row gives the out-of-sample forecasting accuracy for each model. Data from 1980–1995 is used to compute the parameter estimates for 1996, then rolling 1-year ahead estimates are generated for 1997–2004. Firms are ranked into deciles according to their probability of default, and the aggregate percentages of defaults in the top two deciles are presented for each model.

	M5		M6		M7		M8	
	Default	Merger	Default	Merger	Default	Merger	Default	Merger
frailty variance	0.151		0.171		0.126		0.149	
	(0.017)		(0.020)		(0.016)		(0.017)	
intercept	-4.315	-3.231	-5.444	-2.677	-5.359	-3.568	-3.636	-3.267
	(0.241)	(0.119)	(0.101)	(0.060)	(0.268)	(0.136)	(0.093)	(0.047)
exret	-2.111	0.036			-1.980	0.035		
	(0.092)	(0.008)			(0.094)	(0.008)		
sigma					0.162	-0.070		
					(0.036)	(0.048)		
rsize	-0.021	- 0.038			0.012	-0.050		
	(0.017)	(0.009)			(0.018)	(0.009)		
nita			-0.845	0.368	0.049	0.394		
			(0.076)	(0.073)	(0.085)	(0.079)		
tlta			2.121	0.034	1.924	0.098		
			(0.095)	(0.067)	(0.094)	(0.067)		
tbsm3							-0.026	0.009
							(0.010)	(0.005)
creditspread	-0.033	-0.339	-0.032	-0.373	0.180	-0.076		
	(0.066)	(0.037)	(0.061)	(0.037)	(0.080)	(0.044)		
Δgdp					0.050	0.089		
					(0.012)	(0.008)		
Δpi					-0.037	-0.047		
					(0.009)	(0.005)		
termspread	0.101	-0.150	-0.080	-1.124	0.011	-0.198		
	(0.022)	(0.012)	(0.021)	(0.012)	(0.026)	(0.015)		
retl							-1.783	0.026
							(0.080)	(0.007)
spretl							0.831	0.633
							(0.165)	(0.098)
dd	-0.187	0.018			-0.159	0.013	-0.194	0.006
	(0.010)	(0.002)			(0.010)	(0.003)	(0.009)	(0.002)
Log L	-16,745		-17,136		-17,167		-16,924	
Out–of–sample performance								
Top 2 deciles (%)	75.68		54.65		80.55		76.54	

Table 4: **Default Estimation: One Frailty per Sector, Competing Risks.**

Panel B: Manufacturing. Annual Data.

The following table presents the estimates for the default prediction model with one frailty per sector, exponential hazards and competing risks. Standard errors are given in parantheses. The sample is restricted to *manufacturing* firms with information in CRSP and COMPUSTAT during 1980-2004. Variable definitions are given in Table 1. The models M1 to M8 differ in the specification of the covariates. The last row gives the out-of-sample forecasting accuracy for each model. Data from 1980–1995 is used to compute the parameter estimates for 1996, then rolling 1-year ahead estimates are generated for 1997–2004. Firms are ranked into deciles according to their probability of default, and the aggregate percentages of defaults in the top two deciles are presented for each model.

	M1		M2		M3		M4	
	Default	Merger	Default	Merger	Default	Merger	Default	Merger
frailty variance	0.073		0.102		0.094		0.071	
	(0.019)		(0.021)		(0.021)		(0.018)	
intercept	-8.665	-3.805	-7.776	-3.611	-5.857	-3.793	-9.027	-3.276
	(0.325)	(0.147)	(0.317)	(0.134)	(0.358)	(0.157)	(0.352)	(0.161)
exret	-2.521	0.092	-2.736	0.091	-2.280	0.083	-2.525	0.104
	(0.154)	(0.022)	(0.155)	(0.021)	(0.151)	(0.021)	(0.154)	(0.021)
sigma	0.273	-0.016	0.324	-0.127			0.300	-0.059
	(0.064)	(0.069)	(0.051)	(0.070)			(0.063)	(0.071)
rsize	-0.165	-0.045	-0.193	-0.037	-0.120	-0.039	-0.168	-0.042
	(0.028)	(0.013)	(0.028)	(0.013)	(0.028)	(0.012)	(0.028)	(0.013)
nita	0.081	0.609					0.080	0.559
	(0.129)	(0.121)					(0.130)	(0.120)
tlta	2.314	0.088					2.308	0.082
	(0.145)	(0.101)					(0.146)	(0.101)
creditspread							0.301	-0.312
							(0.100)	(0.053)
termspread							-0.026	-0.136
							(0.036)	(0.017)
dd					-0.154	0.010		
					(0.015)	(0.004)		
Log L	-10,456		-10,453		-10,426		-10,410	
Out–of–sample performance								
Top 2 deciles (%)	81.90		74.48		74.07		81.89	

Table 4: **Default Estimation: One Frailty per Sector, Competing Risks Model. Panel B: Manufacturing. Annual Data (contd.).**

The following table presents the estimates for the default prediction model with one frailty per sector, exponential hazards and competing risks. Standard errors are given in parantheses. The sample is restricted to *manufacturing* firms with information in CRSP and COMPUSTAT during 1980-2004. Variable definitions are given in Table 1. The models M1 to M8 differ in the specification of the covariates. The last row gives the out-of-sample forecasting accuracy for each model. Data from 1980–1995 is used to compute the parameter estimates for 1996, then rolling 1-year ahead estimates are generated for 1997–2004. Firms are ranked into deciles according to their probability of default, and the aggregate percentages of defaults in the top two deciles are presented for each model.

	M5		M6		M7		M8	
	Default	Merger	Default	Merger	Default	Merger	Default	Merger
frailty variance	0.096		0.093		0.073		0.092	
	(0.021)		(0.021)		(0.019)		(0.020)	
intercept	-6.097	-3.426	-6.024	-2.833	-7.256	-3.896	-4.120	-3.414
	(0.405)	(0.173)	(0.159)	(0.083)	(0.441)	(0.198)	(0.149)	(0.066)
exret	-2.290	0.093			-2.081	0.096		
	(0.153)	(0.021)			(0.154)	(0.021)		
sigma					0.255	-0.028		
					(0.084)	(0.074)		
rsize	-0.120	- 0.043			-0.076	-0.052		
	(0.029)	(0.013)			(0.030)	(0.014)		
nita			-0.858	0.489	0.026	0.574		
			(0.114)	(0.110)	(0.130)	(0.120)		
tlta			2.471	0.069	2.142	0.143		
			(0.149)	(0.101)	(0.147)	(0.102)		
tbsm3							-0.002	0.000
							(0.016)	(0.007)
creditspread	0.162	-0.281	0.055	-0.320	0.287	0.024		
	(0.102)	(0.052)	(0.096)	(0.052)	(0.125)	(0.063)		
Δgdp					0.029	0.105		
					(0.019)	(0.011)		
Δpi					-0.005	-0.049		
					(0.014)	(0.007)		
termspread	0.072	-0.161	-0.081	-1.132	0.028	-0.205		
	(0.035)	(0.018)	(0.034)	(0.017)	(0.041)	(0.022)		
retl							-2.038	0.065
							(0.134)	(0.018)
spretl							0.622	0.700
							(0.267)	(0.150)
dd	-0.158	-3.426			-0.118	0.012	-0.179	0.004
	(0.015)	(0.004)			(0.015)	(0.004)	(0.015)	(0.003)
Log L	-10,362		-10,664		-10,307		-10,439	
Out–of–sample performance								
Top 2 deciles (%)	74.07		65.44		83.95		74.48	

Table 5: **Forecasting Accuracy: Summary**

Panel A: AMEX, NYSE and NASDAQ data

This table summarizes the out-of-sample forecasting accuracy for the various models presented earlier. Data from 1980–1995 is used to compute the parameter estimates for 1996, then rolling 1-year ahead estimates are generated for 1997–2004. Every year, firms are ranked into deciles according to their estimated probability of default, and the aggregate percentages of defaults in the top two deciles are presented for each model.

Model	Non-Financials*		Manufacturing**	
	No Frailty	Frailty	No Frailty	Frailty
M1	76.97	75.68	83.54	83.54
M2	71.10	71.38	72.01	76.13
M3	74.82	74.72	74.89	79.01
M4	77.25	75.68	83.95	84.37
M5	74.67	74.68	74.48	79.02
M6	51.50	53.07	65.02	66.67
M7	74.68	79.79	72.43	75.18
M8	75.11	75.11	74.90	79.01

* For non-financial firms there are 1,430 defaults for the whole period 1980–2004, and 699 defaults for the out–of–sample period 1996–2004.

** For manufacturing firms there are 541 defaults for the whole period 1980–2004, and 243 defaults for the out–of–sample period 1996–2004.

Table 5: **Forecasting Accuracy: Summary**

Panel B: AMEX and NYSE data

This table summarizes the out-of-sample forecasting accuracy for the various models presented earlier. Data from 1980–1995 is used to compute the parameter estimates for 1996, then rolling 1-year ahead estimates are generated for 1997–2004. Every year, firms are ranked into deciles according to their estimated probability of default, and the aggregate percentages of defaults in the top two deciles are presented for each model.

Model	Non-Financials*		Manufacturing**	
	No Frailty	Frailty	No Frailty	Frailty
M1	84.51	83.19	86.28	89.22
M2	83.19	82.30	81.38	79.42
M3	81.86	82.30	78.44	80.40
M4	84.51	83.19	86.28	89.22
M5	82.30	82.30	78.44	80.40
M6	53.54	58.85	66.67	72.55
M7	76.55	84.52	78.43	87.26
M8	82.74	84.07	82.36	82.35

* For non–financial firms there are 489 defaults for the whole period 1980–2004, and 226 defaults for the out–of–sample period 1996–2004.

** For manufacturing firms there are 213 defaults for the whole period 1980–2004, and 100 defaults for the out–of–sample period 1996–2004.

Table 6: **Determinants of Recovery Rate**

Panel A: Linear Models

This table reports coefficient estimates from the linear regression relating the recovery rate to the bond, firm and macroeconomic variables during 1980–2004. The dependent variable is the recovery rate. Other variable definitions are given in Table 1. Robust standard errors adjusted for firm level clustering are given in parantheses, and the adjusted R^2 and the number of observations N are also reported. For out–of–sample forecasting the initial sample period is 1980–1995; rolling one year ahead forecasts and the RMSE are calculated for the period 1996–2004, and the mean RMSE is presented.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
couprate	0.0182	0.0182	0.0272	0.0222	0.0240	0.0133
	(0.0055)	(0.0056)	(0.0068)	(0.0063)	(0.0062)	(0.0078)
log(issuesize)	-0.0332	-0.0274	-0.0070	-0.0033	-0.0079	0.0105
	(0.0185)	(0.0189)	(0.0195)	(0.0205)	(0.0194)	(0.0235)
log(matoutstand)	-0.0000	-0.0012	0.0087	-0.0248	0.0155	-0.0145
	(0.0303)	(0.0301)	(0.0290)	(0.0360)	(0.0293)	(0.0253)
subord	0.1849	0.1919	0.2332	0.2133	0.2136	0.1364
	(0.0433)	(0.0421)	(0.0472)	(0.0478)	(0.0456)	(0.0887)
sensub	0.1610	0.1636	0.1958	0.1666	0.1718	0.1392
	(0.0649)	(0.0636)	(0.0650)	(0.0665)	(0.0647)	(0.0861)
sensec	0.3161	0.3245	0.3289	0.3481	0.3150	0.3055
	(0.0746)	(0.0724)	(0.1307)	(0.1089)	(0.1320)	(0.1218)
senuns	0.2558	0.2635	0.2974	0.2957	0.2957	0.2508
	(0.0580)	(0.0567)	(0.0531)	(0.0531)	(0.0489)	(0.0705)
tbsm3	-0.0319	-0.0301	-0.0306	-0.0319	-0.0350	-0.0302
	(0.0133)	(0.0131)	(0.0139)	(0.0146)	(0.0131)	(0.0132)
ebitdasales		0.0450				0.0114
		(0.0229)				(0.0208)
tanta						0.1419
						(0.0819)
log(ta)						-0.0071
						(0.0172)
mtb						-0.0223
						(0.0276)
creditspread						0.1892
						(0.0875)
Δgdp						-0.0086
						(0.0113)
Δpi						0.0090
						(0.0133)
dd			0.0158		0.0178	0.0132
			(0.0096)		(0.0121)	(0.0106)
retl				0.0423	-0.0244	0.0125
				(0.0466)	(0.0593)	(0.0510)
spretl				0.2119	0.2565	0.2867
				(0.1375)	(0.1250)	(0.1408)
intercept	0.2625	0.2293	-0.1127	-0.1475	-0.1283	0.0123
	(0.3074)	(0.3059)	(0.2854)	(0.3376)	(0.2977)	(0.3083)
Adj R^2	0.122	0.131	0.185	0.174	0.211	0.256
N	444	444	337	334	334	332
Out–of–sample RMSE	0.2544	0.2530	0.2496	0.2770	0.2666	0.2674

Table 6: **Determinants of Recovery Rate**

Panel B: Probit Models

This table reports coefficient estimates from the probit regression relating the recovery rate to the bond, firm and macroeconomic variables during 1980–2004. The dependent variable is probit of the recovery rate. Other variable definitions are given in Table 1. Robust standard errors adjusted for firm level clustering are given in parantheses, and the adjusted R^2 and the number of observations N are also reported. For out–of–sample forecasting the initial sample period is 1980–1995; rolling one year ahead forecasts and the RMSE are calculated for the period 1996–2004, and the mean RMSE is presented.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
couprate	0.0674	0.0671	0.0941	0.0776	0.0831	0.0489
	(0.0177)	(0.0180)	(0.0214)	(0.0204)	(0.0199)	(0.0250)
log(issuesize)	-0.1337	-0.1106	-0.0600	-0.0477	-0.0621	0.0152
	(0.0634)	(0.0627)	(0.0694)	(0.0714)	(0.0677)	(0.0728)
log(matoutstand)	0.0241	0.0195	0.0739	0.1265	0.0977	-0.0004
	(0.0982)	(0.0977)	(0.0962)	(0.1144)	(0.0963)	(0.0861)
subord	0.5370	0.5648	0.6996	0.6287	0.6295	0.3788
	(0.1582)	(0.1531)	(0.1513)	(0.1550)	(0.1477)	(0.2779)
sensub	0.4309	0.4412	0.5592	0.4572	0.4737	0.3319
	(0.2242)	(0.2194)	(0.2154)	(0.2219)	(0.2160)	(0.2897)
sensec	0.9726	1.0060	1.0226	1.0783	0.9761	0.9118
	(0.2403)	(0.2309)	(0.4254)	(0.3561)	(0.4282)	(0.3890)
senuns	0.8128	0.8431	0.9771	0.9652	0.9655	0.8081
	(0.2012)	(0.1943)	(0.1723)	(0.1740)	(0.1609)	(0.2172)
tbsm3	-0.1067	-0.0995	-0.1039	-0.1096	-0.1191	-0.0932
	(0.0409)	(0.0397)	(0.0430)	(0.0453)	(0.0407)	(0.0432)
ebitdasales		0.1781				0.0779
		(0.1027)				(0.1043)
tanta						0.5303
						(0.2569)
log(ta)						-0.0345
						(0.0548)
mtb						-0.1002
						(0.0930)
creditspread						0.5640
						(0.2921)
Δgdp						-0.0144
						(0.0363)
Δpi						0.0234
						(0.0411)
dd			0.0501		0.0549	0.0403
			(0.0305)		(0.0383)	(0.0344)
retl				0.1449	-0.0609	0.0763
				(0.1432)	(0.1823)	(0.1580)
spretl				0.7560	0.8963	0.8664
				(0.4436)	(0.4093)	(0.4610)
intercept	-0.8731	-1.0045	-2.2061	-2.3092	-2.2499	-1.7773
	(1.0170)	(1.0014)	(0.9652)	(1.1017)	(0.9963)	(1.0351)
Adj R^2	0.141	0.154	0.198	0.197	0.227	0.272
N	444	444	337	334	334	332
Out–of–sample RMSE	0.2581	0.2553	0.2528	0.2830	0.2713	0.2685

Table 7: **Default Estimation: One Frailty per Sector**

Panel A: Non-financials. Quarterly Data.

The following table presents the estimates for the default prediction model with one frailty per sector and exponential hazards. Standard errors are given in parantheses. The sample is restricted to *non-financial* firms with information in CRSP and COMPUSTAT during 1980-2004. Variable definitions are given in Table 1. The models M1 to M8 differ in the specification of the covariates.

	M1	M2	M3	M4	M5	M6	M7	M8
frailty variance	0.260	0.394	0.302	0.252	0.289	0.367	0.205	0.270
	(0.044)	(0.054)	(0.045)	(0.043)	(0.044)	(0.054)	(0.038)	(0.041)
intercept	-11.778	-10.493	-4.964	-12.082	-5.724	-8.381	-8.240	-4.859
	(0.215)	(0.203)	(0.279)	(0.234)	(0.295)	(0.122)	(0.316)	(0.107)
exret	-2.235	-2.841	-1.314	-2.225	-1.308		-1.303	
	(0.096)	(0.101)	(0.103)	(0.097)	(0.103)		(0.097)	
sigma	0.209	0.282		0.215			0.190	
	(0.015)	(0.014)		(0.015)			(0.033)	
rsize	-0.279	-0.364	-0.087	-0.286	-0.105		-0.085	
	(0.017)	(0.016)	(0.019)	(0.017)	(0.019)		(0.020)	
nita	-1.736			-1.718		-4.254	-1.569	
	(0.223)			(0.223)		(0.211)	(0.222)	
tlta	3.649			3.651		4.165	3.142	
	(0.110)			(0.111)		(0.115)	(0.112)	
*tbsm*3								0.051
								(0.012)
creditspread				0.281	0.565	0.035	0.351	
				(0.071)	(0.070)	(0.068)	(0.075)	
Δgdp							0.001	
							(0.011)	
Δpi							0.014	
							(0.008)	
termspread				-0.061	-0.015	-0.079	-0.053	
				(0.026)	(0.026)	(0.026)	(0.026)	
retl								-2.004
								(0.104)
spretl								0.757
								(0.156)
dd			-0.396		-0.400		-0.247	-0.342
			(0.015)		(0.015)		(0.022)	(0.014)
Log L	-5,597	-5,973	-5,749	-5,615	-5,761	-6,020	-5,506	-5,629

Table 7: **Default Estimation: One Frailty per Sector**

Panel B: Manufacturing. Quarterly Data.

The following table presents the estimates for the default prediction model with one frailty per sector and exponential hazards. Standard errors are given in parantheses. The sample is restricted to *manufacturing* firms with information in CRSP and COMPUSTAT during 1980-2004. Variable definitions are given in Table 1. The models M1 to M8 differ in the specification of the covariates.

	M1	M2	M3	M4	M5	M6	M7	M8
frailty variance	0.251	0.365	0.319	0.234	0.292	0.309	0.206	0.280
	(0.066)	(0.077)	(0.071)	(0.063)	(0.067)	(0.073)	(0.060)	(0.065)
intercept	-12.704	-11.461	-6.123	-13.225	-7.182	-9.150	-9.993	-4.946
	(0.352)	(0.338)	(0.469)	(0.379)	(0.493)	(0.196)	(0.519)	(0.162)
exret	-1.959	-2.711	-1.376	-1.962	-1.373		-1.278	
	(0.156)	(0.166)	(0.169)	(0.157)	(0.169)		(0.155)	
sigma	0.174	0.261		0.180			0.195	
	(0.025)	(0.022)		(0.025)			(0.059)	
rsize	-0.320	-0.426	-0.155	-0.330	-0.180		-0.154	
	(0.027)	(0.027)	(0.032)	(0.028)	(0.032)		(0.033)	
nita	-1.624			-1.592		-3.855	-1.442	
	(0.357)			(0.358)		(0.342)	(0.356)	
tlta	4.091			4.097		4.686	3.649	
	(0.178)			(0.180)		(0.185)	(0.182)	
*tbsm*3								0.061
								(0.018)
creditspread				0.456	0.757	0.214	0.519	
				(0.105)	(0.106)	(0.102)	(0.114)	
Δgdp							0.004	
							(0.018)	
Δpi							0.018	
							(0.013)	
termspread				-0.076	-0.019	-0.088	-0.061	
				(0.042)	(0.040)	(0.041)	(0.042)	
retl								-1.810
								(0.157)
spretl								0.711
								(0.254)
dd			-0.351		-0.359		-0.195	-0.346
			(0.023)		(0.023)		(0.022)	(0.021)
Log L	-2,554	-2,872	-2,735	-2,554	-2,727	-2,787	-2,493	-2,683

Table 8: **Multiperiod Forecasting Accuracy: Summary**

This table summarizes the out-of-sample forecasting accuracy for the default model M4 with frailty. Data from 1980–1995 is used to compute the parameter estimates for 1996, then rolling one-year ahead estimates are generated for 1997–2004. Every year, firms are ranked into deciles according to their estimated probability of default, and the aggregate percentages of defaults in the top two deciles are presented.

Panel A: Non–financial firms

Sample data	Quarterly data		Annual data
	At least 8 quarters	At least 20 quarters	
AMEX and NYSE	72.43	80.00	83.19
AMES, NYSE and NASDAQ	64.48	71.34	75.68

Panel B: Manufacturing firms

Sample data	Quarterly data		Annual data
	At least 8 quarters	At least 20 quarters	
AMEX and NYSE	85.56	92.40	89.22
AMES, NYSE and NASDAQ	76.10	83.33	84.37