Chapter 8
Using Public Information to Predict Corporate Default Risk

C.N. Peng and J.L. Lin

Abstract Corporate defaults are often affected by many factors that are roughly divided into the two types: internal factors and external factors. Internal factors can be measured precisely with firm-specific financial statistics while external factors contain qualitative data, like related news. There are large amount of timely information from news which affects the default probability of corporates. Efficient extraction information contained in the news is the main focus of this study and we propose to use empirical Bayes and Bayesian Networks to achieve this goal. First, we retrieve both macroeconomic and firm-specific news published by major newspapers in Taiwan. Then, word segmentation is applied, keywords are extracted and then the news variables are computed. Instead of adding the news variables to the logistic regression model, we convert them into prior distribution for the parameters in the corporate default model. Finally, we compute the posterior distribution of the model parameters to predict the corporate default. The estimation is performed using the integrated nested Laplace approximations which, to our belief, is better than the traditional Markov Chain Monte Carlo for our model. Empirical analysis using Taiwanese data finds that news has a significant impact on the corporate default rate prediction. Adding the news variable does improve the forecast precision and prove its usefulness.

C.N. Peng (✉) · J.L. Lin
Commerce Development Research Institute, Taiwan, Republic of China
e-mail: Kenal.peng@cdri.org.tw

J.L. Lin
e-mail: jlin@gms.ndhu.edu.tw

W.K. Härdle et al. (eds.), *Applied Quantitative Finance*, Statistics and Computing,
DOI 10.1007/978-3-662-54486-0_8

8.1 Introduction

Due to the rapid development of internet, we can get instant global economic news on all the financial media around the clock. There are basically two kinds of news based on its frequency and involved entity. One is regularly published government economic data and forecast, and the other is occasional occurrence of corporate litigation, financial earning information, personnel changes or industry dynamics. News such as Taiwan HTC's infringement cases sued by the US Apple, US Apple's announcement of its unexpected decrease of sales, or the talk of Morris Chang, TSMC's chairman, will have direct or indirect impacts on business, industry and the overall economic environment. Extracting and interpreting these financial news to forecast corporate default rates have been an important issue. However, since news is mostly qualitative, and is often released irregularly, it is difficult to quantify such information as variables to be included in the econometric models. In practice, credit rating agencies such as S and P and Moody's and other credit rating agencies have taken into account non-quantitative factors to adjust their credit rating results obtained from the statistical models.

Financial information can also be classified as qualitative and quantitative types. News about European debt crisis is qualitative data while credit rating or economic growth rate is quantitative data. Both types of data have significant impacts on corporate earnings and should be included in the corporate default prediction models. For the quantitative data, one can directly feed them into statistical models for empirical analysis. As for extracting information from qualitative data, it would be much more convenient to perform the task using the Bayesian models, which combine prior distribution and likelihood function into posterior distribution. Qualitative information is for prior distribution as data is for the likelihood function. In other words, textual news can be coerced into priori distribution. Yet, there is still one obstacle for this implementation. In traditional Bayesian models, priori distribution is formulated for the model parameters in likelihood functions or in regression models. While we could easily make a statistical inference from the news about its impact on default rate, its implication for models parameters is unclear. For example, the Euro debt crisis will not only increase the potential default probability of the bank, but also slowdown economic growth. Such information is difficult to be converted into priori distribution of model parameters. Therefore, the main purpose of this paper is to quantify financial news and embed it in a Bayesian framework to forecast corporate default rates. The computation and simulation are performed using the Integrated Nested Laplace Approximations (INLA) which is believed to be more efficient than the popular Markov Chain Monte Carlo (MCMC) for our model. It is worth mentioning that our model could be further developed as a real-time and dynamic default prediction model that is very useful for credit risk management.

8.2 Literature Review

Credit rating reflects the soundness of the enterprise and related literatures are voluminous. We shall first examine the influences of credit rating by some major credit rating agencies, followed by evaluating these credit ratings. Then, we discuss papers on modeling corporate default probability and introduce information theory and its application. Finally, we review models containing quantitative and qualitative variables.

Brooksaff et al. (2004) used the Standard & Poor and Fitch's credit ratings to assess their impacts on the global stock market. The empirical analysis confirmed significant effects, especially when the credit rating are downward graded. Yet, it is not the case for newly developing countries. Ferreira and Gama (2007) also found a spillover effect on the stock markets of other countries when a country's rating is downward graded. Kim and Wu (2008) discover some impacts on credit markets when credit rating agencies release long and short term ratings. Orth (2013) applied Bayesian simulation approach to adjust the rating of sovereign debt securities and corporate debt securities. There exist under-estimation of risk for Standard & Poor's credit rating, especially when the rating is downward graded. Literatures on modeling corporate default probability are voluminous and can be roughly divided into two categories: structural model and reduced-form model. Merton's model as in Black and Scholes (1973) and Merton (1974) is the representative structural model. Credit rating agency, Moody, further revised it as Merton-KMV model. In this model, when the market value of a corporate's assets is lower than its liabilities, the company will soon reach default. It uses European option pricing to calculate the default probability. This model has been called firm-value based model. Vasicek (1977) and Shimko (1993) use stochastic interest rates to evaluate the Bond prices. Longstaff and Schwartz (1995) and Hui et al. (2003) relax part of the assumptions and modify Merton model. However, in addition to the internal factors from within the corporate, there are many external factors that could cause corporate default. The changing external environment has gradually made structural model less popular. Reduced-form model, also known as intensity model, mainly explores the linkage among corporate default and the explanatory variables. It was first proposed by Jarrow and Turnbull (1995) and a great deal of related models were developed, including multiple regression analysis (West 1970), multivariate discriminant analysis and Z-score model (Altman 1968), logistic model (Ohlson 1980), Probit model (Zmijewski 1984), order probability model (Gentry et al. 1985; Blume et al. 1998; Guttler and Wahrenburg 2007), fixed proportional hazards model (Cox 1972; Lane et al. 1986; Bharath and Shumway 2008), discrete-time hazard model (Shumway 2001; Chava and Jarrow 2004), credit rating transition matrix (Lando and Skodeberg 2002) and dynamic default intensity model (Duffie et al. 2007). It is worth noting that (Duffie et al. 2007) and its extended models belong to the application of survival models, which use macroeconomic, industry, firm-specific and other variables to estimate the default intensity.

Information arrives in many forms but all affect the corporates performance. While information about corporate earnings and other general information are released on

quarterly or monthly based, the daily stock market is often strongly influenced by the news of the day so that the daily close price reflect daily market information rather than corporate real operating conditions. Brown et al. (1988), Braun et al. (1995), Pandher and Currie (2013), Coval and Shumway (2001) and others interpret this phenomenon from different angles. Tetlock (2007) studied medias (Wall Street Journal) impact on investors and found significant impacts of negative news on stock trading volume. Tetlock et al. (2008) show that negative wording will affect corporate revenue and can be used as an important predictor for the stock returns and the corporate revenue. Antweiler and Frank (2004) studied the impact of the web news on stock market. Yet, it is rather difficult to evaluate the composite impacts of news from different sources as their basic characteristics might be different from each other in a fundamental way.

For Bayesian credit risk literature, Czado (1994) derived Bayesian inference of binary regression models with parametric link; Gössl (2005), and McNeil and Wendin (2007) used Bayesian inference method to revise portfolio credit risk calculation; Kiefer (2008, 2009, 2010), Jacobs and Kiefer (2010, 2011), Gössl (2005) and McNeil and Wendin (2007) included outside experts opinions via Bayesian framework to compute the posterior density of underlying parameters in credit risk models. Orth (2013) studied the evaluation of sovereign and corporate credit risk, and calculated credit rating transition matrix. Lock and Gelman (2010) transforms the poll results into a priori distribution and then combine it with the general regression model to predict the US presidential election results. Ben-Gal (2007) and Fernandez and Salmeron (2008) show that Bayesian network model could be represented by directed acyclic graph, which describes the relationship between two or more nodes, and the node strength was expressed by probability. Yet, this approach requires clear definitions of all nodes with real data that limited its applicability. Among few related researches, Alexander (2000) used Bayesian belief networks (BBNs) to design work insurance policy. Pourret et al. (2008a), mentioned that Denmark's largest financial services company (Nykredit) applied BBNs to predict the default probability of large corporates. It is worth noting that Bayesian network model is mainly applied in computational biology and bioinformatics gene regulatory networks, gene expression analysis, document classification, information retrieval, decision support systems and so on.

Furthermore, both Back et al. (2001) and Kloptchenko et al. (2004) combined firm-specific variables with news processed using text mining methods to evaluate the impact of the news on the corporation. However, this approach is limited to specific event and is difficult to generalize to general cases. Only few studies combine quantitative and qualitative data into a single model to predict corporate default rates and Lu et al. (2012) is one exception. He retrieved keywords from news, classified these keywords into crisis and non-crisis categories, use chi-square test to screen proper keywords and then assign weights to construct Intensity of Default-Corpus (ITDC) which latter is fed into a logistic regression model for corporate default probability prediction. The empirical results showed that the closer to the crisis point the better estimation of default probability.

8.3 Econometric Models

We shall discuss our econometric models in two parts. The conventional corporate default model is first introduced and then the news variables are added.

8.3.1 Logistic Models for Default Rate

Among existing models, we select Shumway (2001)'s model as our base model because it is a dynamic discrete-time hazard model. Let T denote the time of default and the firm starts at $t = 1$. Then, the survival probability at Δt, is

$$\varphi(t|x) \equiv p(t \in [t, t + \triangle t|T \geq t, x]) \tag{8.1}$$

$$= \frac{1}{1 + e^{-\theta_1 g(t) - \theta_2 X}} \tag{8.2}$$

The multi-period logistic model for empirical analysis. Equation (8.2) now becomes

$$\lambda(t|x) \equiv ln(\varphi(t|x)) = \theta_1 g(t) + \theta_2 X \tag{8.3}$$

where $g(t) = ln(t)$ is a function of t, θ_1, θ_2 are estimated parameters, and x could be firm-specific earnings or macroeconomic variables. By plugging-in estimated parameters into the model, we get the strength of default, the higher the value the higher the default probability. Note that model defined in (8.3) will be reduced to standard logistic model if the term $g(t)(= ln(t))$ is removed.

8.3.2 Default Models Including News Information

In Bayesian models, past data can be used to specify priori distribution (Robbins (1985), Brandel (2004)). Assume that $p(x|\theta)$ is the likelihood function of x, and θ is the unknown parameter of interest. Let $g(\theta|\eta)$ be the prior distribution of θ, where η is called hyper-parameters vector. Brandel (2004) applied Bayes theory and obtained posterior distribution as

$$p(\theta|x, \eta) = \frac{p(x|\theta)g(\theta|\eta)}{m(x|\eta)} = \frac{p(x|\theta)g(\theta|\eta)}{\int p(x|\theta)g(\theta|\eta)d\theta}$$

where $m(x|\eta) = \int p(x|\theta)g(\theta|\eta)d\theta$ is the marginal distribution of x. Then the expectation of posterior density is

$$E[\theta|x] = \frac{\int \theta p(x|\theta)g(\theta|\eta)d\theta}{\int p(x|\theta)g(\theta|\eta)d\theta} \tag{8.4}$$

In (8.4), the estimation result will be affected by the hyper-parameters vector, η. Estimation is straightforward if η is known but η is usually unknown in practice. In turn, Marginal Maximum Likelihood Estimation (MMLE) can be applied and the resulting marginal distribution $m(x|\eta)$ of x is then used to estimate η. This process is called empirical Bayes method.

Obviously, for default probability model, the dependent variable is 0 (event does not occur) or 1 (event occurs), the estimated default probability is within (0,1] and the explanatory variables are macroeconomic or firm-specific financial variables. This explains why Kleinman (1973) Wilhelmsen et al. (2009), Kiefer (2009, 2010), and Jacobs and Kiefer (2010, 2011) all choose Beta-Binomial model. Assume that variable Y_{it} represents the default status of i-th corporate at time t. $Y_{it} = 1$ when it defaults or $Y_{it} = 0$ when it does not default. Y_{it} has Bernoulli(π_i) distribution, where π_i is default probability of corporate i. Assume the default status of corporate i is independent over time. Let X_i be the default status up to time n_i, we have the following formula

$$X_i = \sum_{t=1}^{n_i} Y_{it} \sim B(n_i, \pi_i)$$

where X_i has binomial distribution and variable X_i will vary with π_i. The maximum likelihood function of corporate i with default probability at time n_i is

$$p(X_i = x_i|\pi_i) = C_{x_i}^{n_i} \pi_i^{x_i} (1 - \pi_i)^{n_i - x_i}$$

Through dynamic default probability model, we can solve for π_i. Assume π_i has $Beta(r, s)$ distribution and is re-parameterized as $Beta_{rep}(\mu, M)$ where

$$\mu = \frac{\gamma}{\gamma + s}, M = \gamma + s$$

Put (8.5) into $Beta(\mu, M)$, the joint probability density function is

$$g(\Pi = \pi|\mu, M) = \frac{\Gamma(M)}{\Gamma(M\mu)\Gamma(M(1-\mu))} \pi^{M\mu - 1}(1 - \pi)^{M(1-\mu)-1}$$

Thus the marginal probability function is

$$m(X = x|\mu, M) = C_x^n \frac{\Gamma(M)}{\Gamma(M\mu)\Gamma(M(1-\mu))} \frac{\Gamma(x + M\mu)\Gamma(n - x + M(1-\mu))}{\Gamma(n + M)}$$

Finally, the posterior distribution is $Beta(r_{EB}, s_{EB})$ where

$$\gamma_{EB} = x + M\mu, S_{EB} = n - x + M(1 - \mu)$$

In the estimation process, the relationship of hyper-parameters requires simulation estimation. While there exist a great of simulation estimation methods, Markov Chain Monte Carlo (MCMC) or EM-algorithm are commonly used. This paper adopts more efficient Integrated Nested Laplace Approximations (INLA). Wilhelmsen et al. (2009) compared the difference between MCMC and INLA, and found that the efficiency and accuracy of INLA are better than that of MCMC. See Rue et al. (2009) for details.

8.3.3 Bayesian Network Model

Ben-Gal (2007) pointed out that the main structure of Bayesian network model is non-circulate probability graphical model where there exist sequential causal relationships among various events. In this paper, we shall estimate $\lambda(t|x)$ in the discrete-time hazard model. As it is affected not only by firm-specific variables at time t, but also by the news information at time $t - 1$. Hence, we specify the default probability function as

$$f(Y_t|X_{i,t-1}), i = 1, 2, ..., n$$

where $X_{i,t-1}$ is the news information factor at time $t - 1$. News information will be retrieved, quantified and its probability distribution will be simulated. Finally, using Bayesian network method, we can get revised default probability as

$$f(Y_t, X_{i,t-1}), i = 1, 2, ..., n$$

From above, assume there are two news X_1 and X_2 then

$$f(Y, X_1, X_2) = f(Y|X_1, X_2)f(X_2|X_1)f(X_1)$$

where $f(Y|X_1, X_2)$ is the default probability from the corporate default model, and $f(X_2|X_1)$ is mutual impact between news events. This, in principle, can be used to estimate the impact of sequent news events on default probability but it is difficult to implement in practice. Thus, we follow Fernandez and Salmeron (2008) and Rijmen (2008) and apply regression analysis. Since Bayesian network is non-circulate directed, each news event can be treated as an explanatory variable, and the dependent variable is the corporate default variable. Under multiple news events, we need to consider whether they are related with each other. Its mathematical formula is

$$f(Y|X) = \alpha X + \varepsilon$$

were Y represent default of the corporate, X is news event and ϵ is random error. Obviously, we have

$$f(Y|X) = \frac{f(Y,X)}{f(X)} = f(X|Y)\frac{f(Y)}{f(X)}$$

and

$$f(X|Y) \propto f(Y)f(X|Y)$$

It is called Naive Bayes (NB) when each news event is independent and Tree Augmented Naive Bayes (TAN) when news are dependent. Rijmen (2008) adopts logistic regression in Bayesian Network model where the weight of each segmented word is estimated with the logistic regression model. Wilhelmsen et al. (2009) assumed the prior distribution of logistic regression coefficients is

$$\beta_j \sim \pi(\beta_j|\theta_j), \quad j = 0, 1, ..., M$$

where $\pi(\cdot|\theta)$ denotes all possible distributional function, and θ_j is a scalar or vector parameter. In this paper, θ_j is assumed as a scalar from news information, we obtain posterior distribution as

$$\pi(\beta, \theta|y) = \frac{\pi(\beta, \theta, y)}{\pi(y)} \propto \pi(y|\beta, \theta)\pi(\beta|\theta)\pi(\theta) \tag{8.5}$$

$$= \prod_i \pi(y_i|\beta, \theta)\pi(\beta|\theta)\pi(\theta) \tag{8.6}$$

Solved by INLA, we obtain $\pi(\beta, \theta|y)$ where news information is included.

Rue et al. (2009) derive the test for parameters. Let $y = (y_1, y_2, \cdots, y_n)$ be the observed variable, its probability function be $\pi(\beta|\theta)$, and the model for unknown parameter β be $\pi(\beta|\theta)$, and θ is hyper-parameter. $\pi(\theta)$ is distribution function of hyper-parameter, and through Bayesian theory we get marginal posterior distribution as

$$\pi(\beta_i|y) = \int_\theta \pi(\beta_i|\theta, y)\pi(\theta|y)d\theta \tag{8.7}$$

$$\pi(\theta_j|y) = \int \pi(\theta|y)d\theta_{-j} \tag{8.8}$$

Through INLA, we get the approximation of marginal posterior distribution as

$$\tilde{\pi}(\beta_i|y) = \int_\theta \tilde{\pi}(\beta_i|\theta, y)\tilde{\pi}(\theta|y)d\theta \tag{8.9}$$

$$\tilde{\pi}(\theta_j|y) = \int \tilde{\pi}(\theta|y)d\theta_{-j} \tag{8.10}$$

where $\int \pi(\theta|y)d\theta_{-j}$ represents integration over all but the j-th parameter. In other words, to obtain the estimated value of the parameters, we have to integrate over all hyper-parameters. As the parameter vector θ is multi-dimensional, we must use the Laplace estimate. In order to improve accuracy, latent Gaussian models are applied. To obtain the estimation of $\tilde{\pi}(\beta_i|y)$, we need to get an approximation of $\tilde{\pi}(\beta_i|\theta, y)$ and $\tilde{\pi}(\theta|y)$, which are assumed as Gaussian distribution. We use Kullback–Leibler Divergence (KLD) test which is defined as below:

$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} ln(\frac{p(x)}{q(x)})p(x)dx$$
$$D_{KL}(P||Q) = \sum_{x} ln(\frac{p(x)}{q(x)})p(x)$$

where P, Q are respective two cumulative probability distribution for continuous and discrete random variables. Let Q be normal distribution, when $D_KL(P||Q) \approx 0$, P is also normally distributed.

For model selection, we shall use two methods: in-sample Receiver Operating Characteristic Curve (ROC) and out-of-sample forecasting error (Lin and Tsay 2007). Altman and Bland (1994) proposed ROC as a method of diagnostic test, which is widely used by biometrics. Within a 2×2 table, P denotes positive and N negative.

Diag	P	N	Total
Truth			
P	TP	FN	
N	FP	TN	
Total			Nobs

The True Positive(TP) and True Negative (TN) are the cells for the right diagnostics. Let Nobs denote total number of samples, then accuracy ratio, sensitivity and specificity are respectively defined as (TP+TN)/Nobs and TP/(TP+FN). ROC is based upon sensitivity and specificity, and can be used for model comparison.

As for out of sample forecasting error, we can calculate its Root Mean Square of Error (RMSE)

$$RMSE_t = \sqrt{\sum_{i=t+1}^{T} \frac{(\hat{y}_i - y_i)^2}{n_i}}$$

To summarize, these news frequencies are used to obtain the prior distribution of the regression parameters β in Shumways model.

8.4 Extracting News Information

Chinese characters can be divided into three types: classical, vernacular and other dialects. Their usages and the structures are all different from each other. Vernacular is currently used, which might vary due to the geographical environment and social backgrounds, but in general follows certain syntax. Tsay (2008) pointed out that a sentence is constituted by two basic components, subject and predicate. Subject is the major part of the sentence, either the perpetrators of the action, or the objects being interpreted, clarified or depicted. The predicate is the statement to clarify the subject. In this paper, news from various media also follow a set of rules. For example, editorial manual of Central News Agency depicts the main structure and term usage. We further classify economic news in Taiwan into two categories. One is economic news containing economic data, business cycle indicators, or economic policy announcement released by the government official or agencies, which does not make judgment of any corporate. The other one is public talks or comments on specific corporate. In addition to Taiwan's local news, foreign financial news also has a considerable impact. We must distinguish their impacts.

8.4.1 News Keywords

Keyword is set in accordance with the commonly used terms and categorized by subject, verb and adjective. Six main structures of the subject are set including raw materials, European debt crisis, people and institutions, economic data release, as well as business and policy agreements. Within each main structure, at least eight keywords are selected, which can be different words with same meaning. The predicate is mainly verbs, such as recovery, recess, rise, fall, up, down, strength and the like. Default keywords defined by Taiwan Economic Journal (TEJ) are also included. There are 10 categories: bankruptcies, restructuring, bounced checks, bail out, take over, CPAs doubt on continuous operation, net worth is negative, unlist, tight budget, negative worth, and shut down. Finally, these keywords are classified as positive, neutral and negative.

8.4.2 Keyword Conversion

Segmented keywords from all news items (documents) are compiled into the document-term matrix where columns are news items, and rows are keywords. For each cell, 0 and 1 indicate if there is such keyword. For each keyword, summing over all news items during any specific quarter will produce frequency of keywords. This process is repeated separately for positive and negative keywords and their ratios are then computed.

8.5 Empirical Analysis and Results

This paper uses quarterly firm-specific data of all listed companies in Taiwan from 2000 to 2012. The data is taken from TEJ, excluding incomplete data entries, financial firms and news media corporations. There are 908 corporates where 805 are still listed at the end of the sample period and 103 are unlisted. As for news, there are mainly two sources: newspaper and networks news. Yet, as the latter is only available for one month after posting, we only use newspapers news. The major four newspapers in Taiwan are China Times, United Daily News, Free News and Apple Daily. The data is collected daily from the first quarter of 2008 to the fourth quarter of 2012, amounting to about 270,000 news items.

8.5.1 Empirical Models

This empirical analysis is illustrated in two parts. First, we follow previous research in selecting firm-specific quantitative variables under the constraint that the resulting ROC curve is above 90%. Second, as for news variables, we employ empirical Bayes and Bayesian networks to convert as quantitative variables and then feed them into the base default model as is introduced previously. We compare the performance of the following six models:

1. Model I: Earnings model
   This is the conventional default model only based upon firm-specific financial variables and $ln(t)$. Standard logistic regression estimation will suffice.
2. Model II: Earnings-macroeconomic model
   In additional to firm-specific financial variables and $ln(t)$, macroeconomic variables are also included in the model for default prediction. Again, the model is estimated using standard logistic regression.
3. Model III: Bayesian earnings model
   Earnings models are formulated under Bayesian framework and is used to predict corporate defaults. Empirical Bayes is used for model estimation. To be specific, default variable is first regressed against firm-specific financial variables and the estimation results are then converted into prior distribution of the associated parameters using INLA algorithm. Finally, the posterior distribution are derived with prior and likelihood function.
4. Model IV: Bayesian earnings-macroeconomic model
   Both firm-specific financial variables and macroeconomic variables are included in the model under Bayesian framework. Estimation procedure is the same as Bayesian earnings model except that macroeconomic variables are added.
5. Model V: Bayesian news-earnings model
   News variables are added to the Bayesian earning model via empirical Bayes method and INLA. To be specific, firm-specific news are classified as good news, and bad news and their relative frequencies to all news are computed.

For macroeconomic news, only those containing the five most and least frequent keywords, such as *price, monetary policy* are counted. These news are further classified as good news or bad news. Next, regress the firm default variable against $ln(t)$, firm-specific good news and firm-specific bad news for each firm. Then, for each ten macroeconomic key variables, regress firm default variable against macroeconomic good news and bad news for each firm. Summing the predicted probability distribution obtained from five most frequent keywords and firm-specific regressions give rise to model 5(L). Similarly, summing the predicted probability distribution obtained from five least frequent keywords and firm-specific regressions gives rise to model 5(S). It is worth noting that the idea of Bayesian network model is used in this step. Now, we could combine news effects with Shumway's model with firm-specific variable using INLA.

6. Model VI: Bayesian news-earning-macroeconomic model
   News variables are added to the Bayesian earnings-macroeconomic model via empirical Bayes method and INLA. Computation procedure is the same as Bayesian news-earnings model except for the added macroeconomic variables.

8.5.2 Variable Selection

In the discrete-time hazard model, explanatory variables must be included to predict corporate default probability. Altman (1968), Ohlson (1980) and Zmijewski (1984) used three to nine financial ratio variables. Shumway (2001) included two financial ratios and three market-driven variables. Chava and Jarrow (2004) added industrial variables to those in Altman (1968) and Zmijewski (1984). Lee and Yeh (2004) focused on the relationship between corporate governance and financial distress. Duffie et al. (2007) added macroeconomic variables to the dynamic intensity model. Campbell et al. (2008) added two firm-specific financial ratios and stock return to the list of variables compiled by Shumway. Standard & Poor consider eighteen variables on liquidity, terms of profitability, capital structure, cash flow and ability to repay interest etc. in corporate's credit rating.

After taking all these literatures into consideration, we select seven variables: assets-liabilities ratio, quick ratio, ratio of retained earnings to total assets, earnings per share, operating expense ratio, unemployment rate, and TAIEX (Taiwan Stock Exchange Capitalization Weighted Stock Index) return. The definitions of the selected variables are reported in Table 8.1. In additional to the variable definition and type of variables, their expected sings are also listed. Table 8.2 summarizes basic statistics of the variables. Except for unemployment rate and the stock market return, extremely large skewness and kurtosis of firm-specific financial variables indicate obvious departure from normal distribution assumption. Table 8.3 reports the parameter estimates for Model I and II. As can be seen from the table, except for the ratio of retained earnings to total assets, all variables are significant and their signs are consistent with prediction from finance theory. The Bayesian estimates for Model III

Table 8.1 Variable definitions

Category	Name	Variable definition	Sign
Financial structure	Asset-liability ratio	Total asset/total liability	Negative
Solvency	Quick ratio	(Liquid asset-inventory)/liquid liability	Negative
Profitability	Ratio of retained earnings to total assets	Retained earning/total assets	Negative
	Earning per share	Earning/number of shares	Negative
Operating capacity	Operating expense ratio	Operating expense/net revenue	Positive
Macro variables	Unemployment rate		Positive
	Stock market return		Negative

Table 8.2 Summary statistics for explanatory variables

Variable	Mean	Std	Median	Skewness	Kurtosis
Assets-liabilities ratio	3.52	5.11	2.61	25.81	1083.35
Quick ratio	1.65	4.08	1.06	26.86	1139.92
Retained earnings to total assets ratio	0.06	0.66	0.10	−49.58	3273.97
Earnings per share	1.15	3.56	0.66	54.42	5886.05
Operating expense ratio	0.26	6.49	0.10	110.16	14210.83
Unemployment rate	4.48	0.72	4.32	0.19	2.77
Stock market return	3.80	26.47	6.89	0.13	3.22

Table 8.3 Parameter estimates for Model I and II. Signif. codes: $p < 0.001$ ****; $p < 0.01$ ***; $p < 0.05$ **; $p < 0.1$ *

	Model I		Model II	
	Est.	t-stat	Est.	t-stat
Intercept	−0.9814	−3.306****	−2.8829	−6.067****
Time trend	−0.2667	−4.151****	−0.4406	−5.145****
Assets-liabilities ratio	−1.0066	−6.734****	−1.0574	−6.427****
Quick ratio	−3.1559	−11.742****	−3.0579	−10.585****
Retained earnings to total assets ratio	0.0776	1.495	0.0767	1.470
Earnings per share	−0.1998	−9.201****	−0.1974	−8.292****
Operating expense ratio	0.0085	3.188***	0.0080	2.457**
Unemployment rate			0.5361	5.726****
Stock market return			−0.0041	−1.668*

Table 8.4 Estimation results using empirical Bayes method

	Model III					Model IV				
	Mean	Std	2.50%	97.50%	KLD	Mean	Std	2.50%	97.50%	KLD
Intercept	−0.923	0.414	−1.729	−0.103	7.07E-14	−2.847	0.618	−4.074	−1.645	1.08E-13
Time trend	−0.264	0.090	−0.437	−0.085	1.23E-12	−0.436	0.111	−0.655	−0.217	0.00E+00
Asset-lib rat	−1.037	0.209	−1.464	−0.644	1.17E-11	−1.087	0.214	−1.523	−0.682	8.67E-12
Quick ratio	−3.169	0.375	−3.932	−2.459	8.80E-12	−3.066	0.376	−3.830	−2.353	8.13E-12
Rtn earnings/Total asset	0.081	0.072	−0.037	0.245	2.67E-10	0.080	0.068	−0.031	0.234	2.39E-10
Earn per share	−0.202	0.030	−0.260	−0.142	1.55E-13	−0.199	0.031	−0.259	−0.138	1.79E-13
Oper. exp. rat	0.009	0.004	0.002	0.017	9.15E-11	0.008	0.004	0.001	0.018	2.37E-10
Unemp rat						0.542	0.122	0.303	0.781	4.67E-15
Stock mkt rtn						−0.004	0.003	−0.010	0.002	1.50E-14

Table 8.5 Estimation results of logistic model with news variables Signif. codes: p < 0.001****; p < 0.01***; p < 0.05**; p < 0.1*

	Pooled news		Category news	
	est.	t-stat	est.	t-stat
Intercept	2.19453	0.464	0.8633	0.164
Time trend	−2.18977	−1.711	−1.9608	−1.387
Pooled news	−0.01627	−1.247		
Positive news			−1.688	−2.234**
Negative news			1.6849	2.627***

and IV are summarized in Table 8.4. In additional to mean, standard deviation, 2.50 and 97.5% quantiles, we also compute Kullback-Leibler Divergence (KLD) statistics which measures divergence from normal distribution. KLD values of all parameters are very small, indicating little divergence of the posterior distribution from normal distribution. Furthermore, we also find that except for the ratio of retained earnings to total assets and TAIEX return, the 95% confidence interval of all parameters do not include 0.

8.5.3 Adding News Variables

For the purpose of comparison, we perform a logistic regression of corporate default indicator directly against news variables and put the results in Table 8.5. On the left panel of the table all news are pooled together while on the right panel positive and negative news are separated. As is expected, pooled news variable is not significant while negative news has stronger effect than positive news on corporate default rate though both estimates are significant. Similar findings were found in Lu et al. (2012).

Now we turn to models V and VI where news variables are added to Shumway's model on the Bayesian framework. Empirical results are reported in Table 8.6. A detailed comparison of estimation results, we make the following observations. First, estimation results of Shumway model without news variables are similar whether it is estimated within classical logistic model or empirical Bayesian model. Second, the results of Model V and VI are similar to those of models III and IV that except for the ratio of retained earnings to total assets and TAIEX return, the 95% confidence interval of all parameters do not include 0 and all KLDs are close to 0. Third, adding news variables to the Bayesian model would change the parameter estimates a great deal. For example, the impacts on quick ratio double in Models V and VI. Fourth, as is in Fig. 8.1 where RMSE for out-of-sample forecast over time are graphed, model II with macroeconomic variables consistently outperform the base model I with only firm-specific variable. Fifth, as is shown in Fig. 8.2, ROC curves of all six models are all above 90%, but the difference is small among models.

Table 8.6 Estimation results using INLA with news variables

	Model V					Model VI				
	Mean	Std	2.50%	97.50%	KLD	Mean	Std	2.50%	97.50%	KLD
Intercept	5.370	4.966	−4.330	15.157	7.56E-15	−3.244	7.574	−19.120	10.672	2.56E-11
Time trend	−1.751	1.323	−4.373	0.822	7.71E-14	−0.256	1.803	−3.648	3.444	1.29E-11
Asset-lib rat	−1.810	0.478	−2.810	−0.930	1.94E-11	−1.831	0.476	−2.831	−0.957	2.43E-11
Quick ratio	−1.577	0.545	−2.709	−0.565	2.44E-11	−1.504	0.542	−2.636	−0.503	2.20E-11
Rtn earnings to total asset rat	0.312	0.239	−0.079	0.851	1.97E-10	0.305	0.240	−0.085	0.847	2.47E-10
Earn per share	−0.149	0.047	−0.238	−0.055	4.28E-12	−0.152	0.046	−0.238	−0.058	1.46E-11
Oper. exp. ratio	0.011	0.006	0.002	0.024	2.05E-09	0.011	0.006	0.001	0.025	2.75E-09
Unemp rat					0.673	0.291	0.126	1.272	1.38E-11	
Stock mkt rtn					−0.006	0.006	−0.019	0.006	9.48E-12	

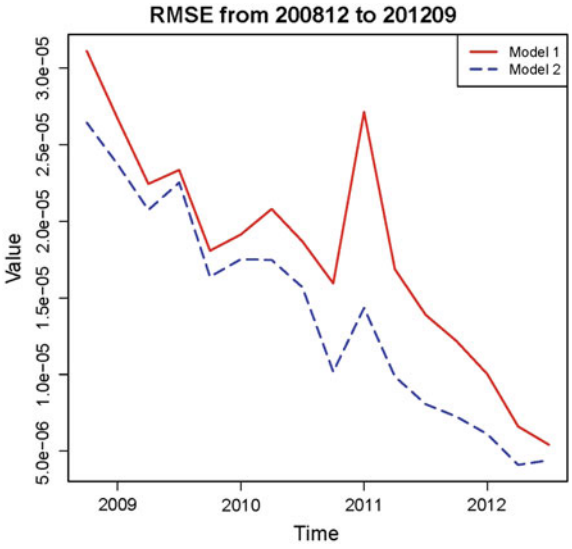

Fig. 8.1 RMSEs for Model I and II

Figure 8.3 is the time series graph of average corporate default rate where annual default rates are put in the left panel whereas quarterly default rates are put in the right panel. The upper panel are based on ratio of number of unlisting stocks to total stocks while the bottom panel is computed following the definition of default in TEJ. As is obvious from the figures, the peaks and troughs of default rate defined by TEJ leads those defined by unlisting.

Figure 8.4 displays the average corporate default intensity of all six models which is the simple average of each corporate's default intensity in each model respectively. Comparing the resulting intensity figures of paired models will highlight their differences. Models I, III and V do not contain macroeconomic variables and are put in left panel of the figure while Model II, IV and VI include macroeconomic variables and are put in the right panel. From the figure, we make the following findings. First, the estimated default intensity of empirical Bayesian model (model III/IV) are smaller than those from Shumway model (model I/II). Both estimates differ from each other by a big margin from 2002 to 2008 when the subprime mortgage crisis broke out. Yet both estimates converge after 2008 crisis. The patterns are similar for both paired models with and without macroeconomic variables. Second, as news variables are collected from Jan 1, 2008 to Dec 31, 2012, comparing estimation results of two subperiods with and without news variable would reveal the impacts of new variables. Considering that each keyword might have different impact on corporates default probability, we add one more step. We first perform a logistic regression again each macroeconomic keyword, compute the squared root of residual sum of squares, RSS, and then sort them in ascending order. Next, we select the keywords with the 5 largest

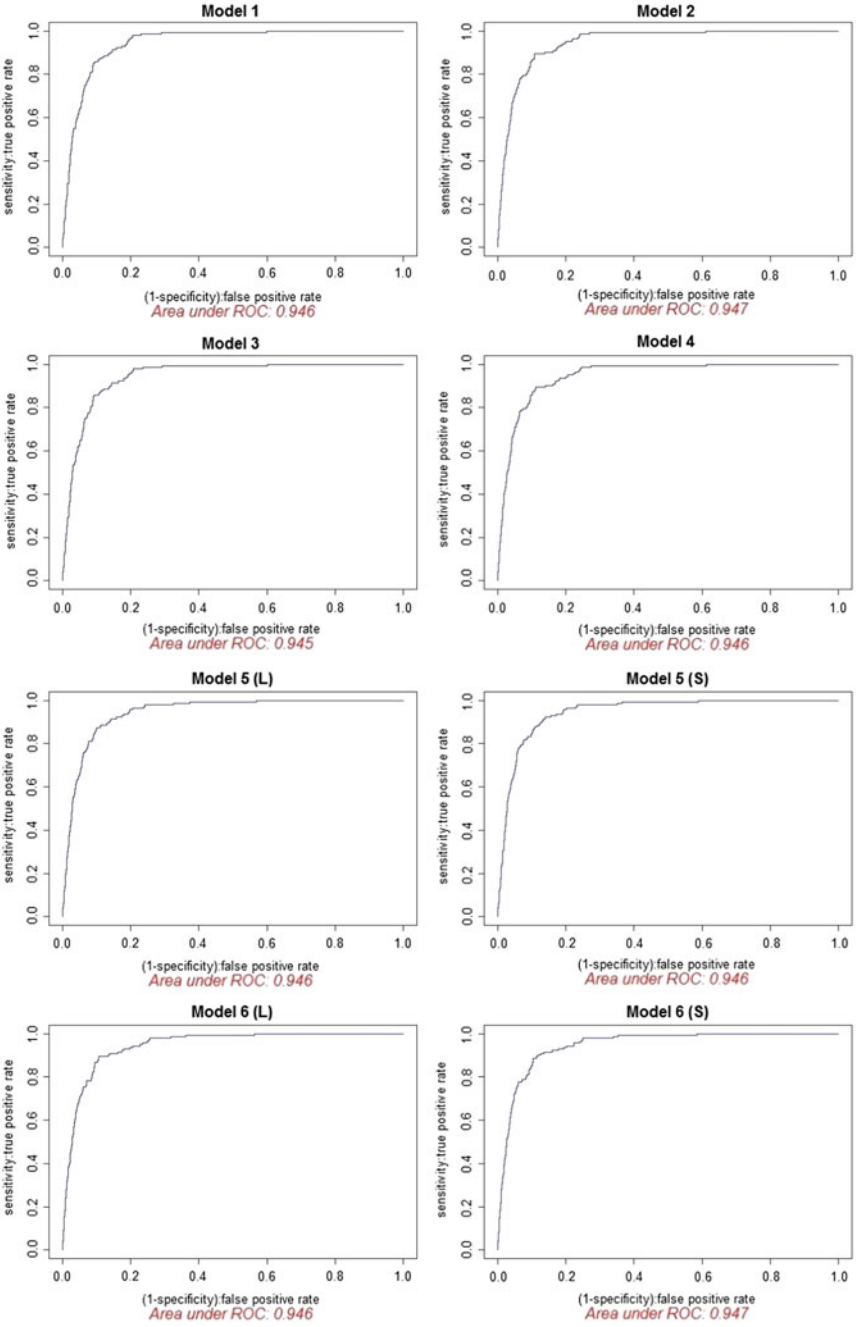

Fig. 8.2 ROC curves for all six models

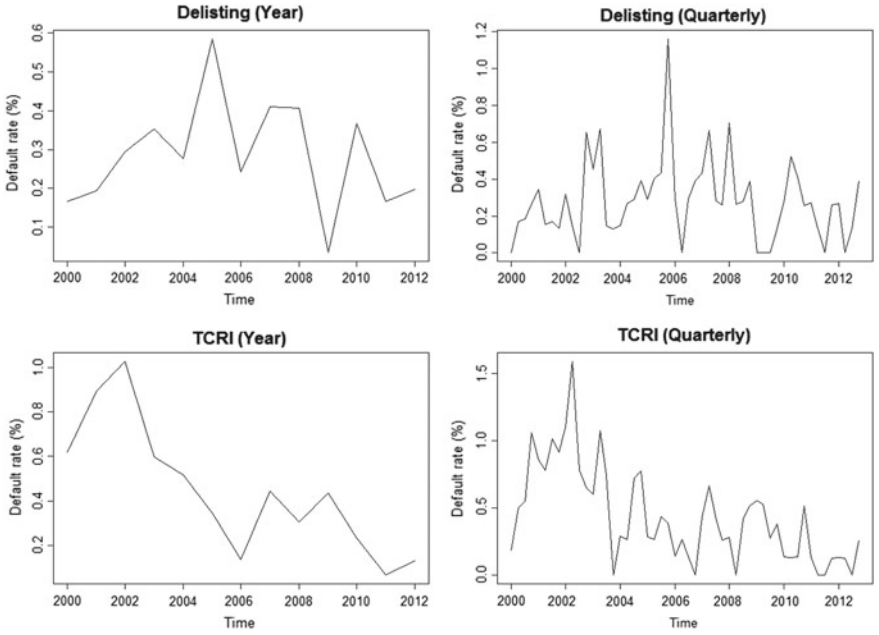

Fig. 8.3 Time series plot of average default rate

RSSs (denoted as L-keywords) and keywords with 5 smallest RSSs (denoted as S-keywords). These L- and S-keywords are then respectively combined with keywords for each corporate, fed into the Bayesian models and estimated using INLA the algorithm. The results are put in the middle and bottom panels of Fig. 8.4. From the figure, we observe that without macroeconomic variables, adding S-keywords produces a sharp increase of corporate default intensity in early 2008 while the impact of S-keyword are much smaller. The situation is reversed when macroeconomic variables are included in the model where L-keywords has a stronger impact on default intensity than S-keywords. It deserves further investigation to explain this phenomenon. Finally, the ROC curves for all six models are reported in Fig. 8.2. They are all above 90% but adding news variables does not significantly increase the ROC curve.

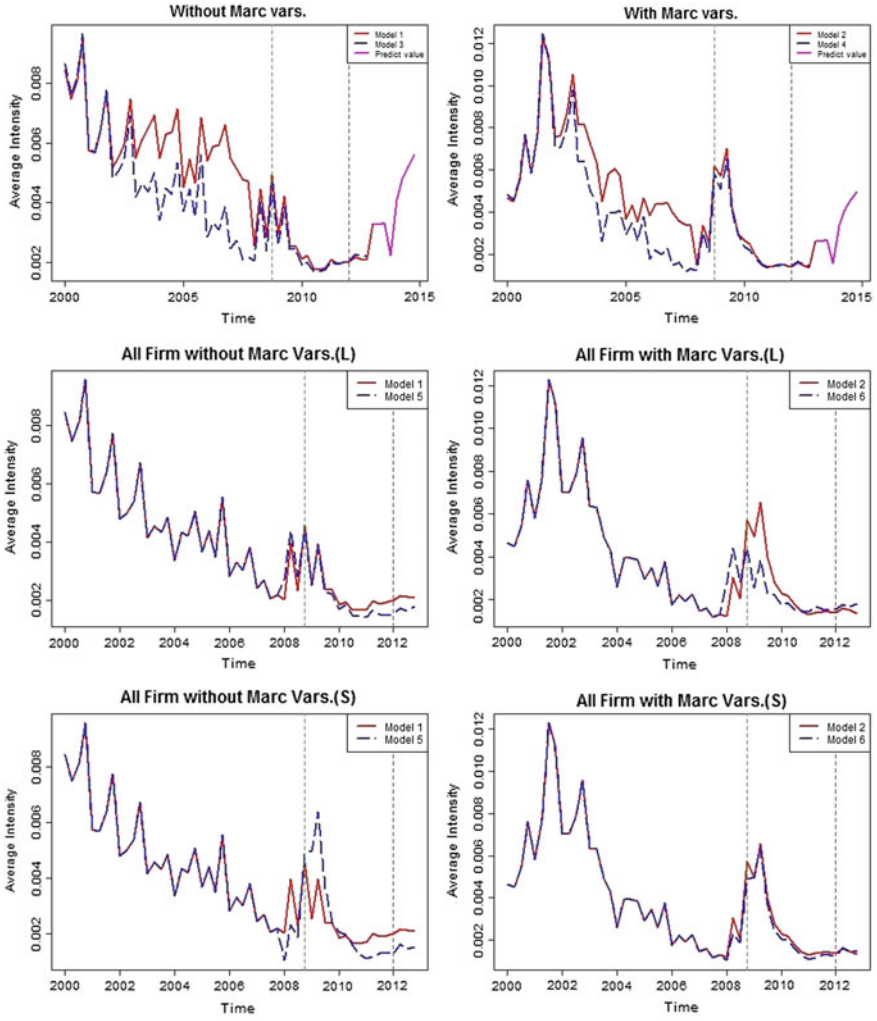

Fig. 8.4 Time series plot of default intensity of all six models

8.6 Conclusions

While corporates' financial reports are released on a quarterly basis, daily economic or financial news could provide timely and useful information about the corporate default probability. This paper provides a framework to extract information from text-based news to improve corporate default prediction. Instead of converting news as a new variable in a standard logistic regression model, we employ the complicated INLA method to transform news into prior information of corporate default and then estimate its impact within a Bayesian model. The conversion is completed using the

INLA. Empirical analysis confirms usefulness of the proposed method though there are rooms for improvement. For example, each keyword might have different weight and the timing of the news within each quarter might be important. These issues deserve further investigation in the future.

References

Alexander, C. (2000). Bayesian Methods for Measuring Operational Risks. *Discussion papers in Finance*.

Altman, D. G, & Bland, J M. (1994). Diagnostic tests. 1: Sensitivity and specificity. *BMJ: British Medical Journal*, *308*(6943), 1552

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

Antweiler, W., & Frank, M. Z. (2004). Is all that talk just noise? the information content of internet stock message boards. *The Journal of Finance*, *59*, 1259–1294.

Ben-Gal, I. (2007). Bayesian networks. In F. Ruggeri, R. Kenett, & F. Faltin (Eds.), *Encyclopedia of statistics in quality and reliability*. New York: Wiley.

Back, B., Toivonen, J., Vanharanta, H., & Visa, A. (2001). Comparing numerical data and text information from annual reports using self-organizing maps. *International Journal of Accounting Information Systems*, *2*, 249–269.

Bharath, S. T., & Shumway, T. (2008). Forecasting default with the merton distance to default model. *Review of Financial Studies*, *21*, 1339–1369.

Black, F., & Scholes, M. (1973). The pricing of options and corporate liabilities. *The Journal of Political Economy*, *81*, 637–654.

Blume, M. E., Lim, F., & MacKinlay, A. C. (1998). The declining credit quality of U.S. corporate debt: Myth or reality? *Journal of Finance*, *53*, 1389–1414.

Brandel, J. (2004). Empirical Bayes methods for missing data analysis. *Department of Mathematics Uppsala University, Project Report*.

Braun, P. A., Nelson, D. B., & Sunier, A. M. (1995). Good news, bad news, volatility, and betas. *The Journal of Finance*, *50*, 1575–1603.

Brooks, R., Faff, R. W., Hillier, D., & Hillier, J. (2004). The national market impact of sovereign rating changes. *Journal of Banking and Finance*, *28*, 233–250.

Brown, K. C., Harlow, W. V., & Tinic, S. M. (1988). Risk aversion, uncertain information, and market efficiency. *Journal of Financial Economics*, *22*, 355–385.

Campbell, J. Y., Hilscher, J., & Szilagyi, J. (2008). In search of distress risk. *Journal of Finance*, *63*, 2899–2939.

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

Coval, J. D., & Shumway, T. (2001). Is sound just noise? *The Journal of Finance*, *56*, 1887–1910.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society Series B*, *34*, 187–220.

Czado, C. (1994). Bayesian inference of binary regression models with parametric link. *Journal of Statistical Planning and Inference*, *41*, 121–140.

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

Fernández, A., & Salmerón, A. (2008). Extension of Bayesian network classifiers to regression problems. In H. Geffner, R. Prada, I. M. Alexandre, & N. David (Eds.), *Advances in artificial intelligence - IBERAMIA 2008* (Vol. 5290, pp. 83–92), Lecture notes in artificial intelligence. Berlin: Springer.

Ferreira, M. A., & Gama, P. M. (2007). Does sovereign debt ratings news spill over to international stock markets? *Journal of Banking and Finance*, *31*, 3162–3182.
Gentry, J. A., Newbold, P., & Whitford, D. T. (1985). Predicting bankruptcy: If cash flow's not the bottom line, what is? *Financial Analyst's Journal*, *41*, 47–56.
Gössl, C. (2005). Predictions based on certain uncertainties-a Bayesian credit portfolio approach. *Working Paper*, Hypo Vereinsbank AG.
Güttler, A., & Wahrenburg, M. (2007). The adjustment of credit ratings in advance of defaults. *Journal of Banking and Finance*, *31*, 751–767.
Hui, C. H., Lo, C. F., & Tsang, S. W. (2003). Pricing corporate bonds with dynamic default barriers. *Journal of Risk*, *5*(3), 17–37.
Jacobs, M, Jr., & Kiefer, N. M. (2010). *The bayesian approach to default risk: A guide* (2nd ed.). Ithaca: Center for Analytical Economics.
Jacobs Jr, M., & Kiefer, N. M. (2011). The bayesian approach to default risk analysis and the prediction of default rates. *Discussion paper*.
Jarrow, R. A., & Turnbull, S. M. (1995). Pricing derivatives on financial securities subject to credit risk. *The Journal of Finance*, *50*, 53–85.
Kiefer, N. M. (2008). Default estimation for low-default portfolios. *Journal of Empirical Finance*, *16*, 164–173.
Kiefer, N. M. (2009). Correlated defaults, temporal correlation, expert information and predictability of default rates. *CAE Working Paper*.
Kiefer, N. M. (2010). Default estimation and expert information. *Journal of Business and Economic Statistics*, *28*, 320–328.
Kim, S. J., & Wu, E. (2008). Sovereign credit ratings, capital flows and financial sector development in emerging markets. *Emerging Markets Review*, *9*, 17–39.
Kleinman, J. C. (1973). Proportions with extraneous variance: Single and independent sample. *Journal of the American Statistical Association*, *68*, 46–54.
Kloptchenko, A., Eklund, T., Karlsson, J., Back, B., Vanharanta, H., & Visa, A. (2004). Combining data and text mining techniques for analysing financial reports. *Intelligent systems in accounting, finance and management*, *12*, 29–41.
Lando, D., & Skodeberg, T. (2002). Analyzing ratings transitions and rating drift with continuous observations. *Journal of Banking and Finance*, *26*, 423–444.
Lane, W. R., Looney, S. W., & Wansley, J. W. (1986). An application of the cox proportional hazards model to bank failure. *Journal of Banking and Finance*, *10*, 511–531.
Lee, T. S., & Yeh, Y. H. (2004). Corporate governance and financial distress: Evidence from Taiwan. *Corporate Governance*, *12*(3), 378–388.
Lin, J. L., & Tsay, R. S. (2007). Comparisons of forecasting methods with many predictors. Working paper, Department of Finance, National DongHwa University.
Lock, K., & Gelman, A. (2010). Bayesian combination of state polls and election forecasts. *Political Analysis*, *18*, 337–348.
Longstaff, F. A., & Schwartz, E. S. (1995). A Simple Approach to Valuing Risky Fixed and Floating Rate Debt. *The Journal of Finance*, *50*(3), 789–819.
Lu, Y. C., Wei, Y. C., Chang, T. Y., & Liao, W. J. (2012). Does soft information from news improve the forecasting performance of habitually used Taiwan corporate credit risk index? *Taiwan Banking and Finance Quarterly*, *13*(4), 27–53. (in Chinese).
McNeil, A. J., & Wendin, J. P. (2007). Bayesian inference for generalized linear mixed models of portfolio credit risk. *Journal of Empirical Finance*, *14*, 131–149.
Merton, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. *The Journal of Finance*, *29*, 449–470.
Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, *18*, 109–31.
Orth, W. (2013). Default probability estimation in small samples-with an application to sovereign bonds. *Quantitative Finance*, *13*, 1891–1902.

Pandher, G., & Currie, R. (2013). CEO compensation: A resource advantage and stakeholder-bargaining perspective. *Strategic Management Journal*, *34*, 22–41.

Pourret, O., Naim, P., & Marcot, B. (Eds.). (2008a). *Bayesian networks: a practical guide to applications* (Vol. 73). New Jersey: Wiley.

Rijmen, F. (2008). Bayesian networks with a logistic regression model for the conditional probabilities. *International Journal of Approximate Reasoning*, *48*, 659–666.

Robbins, H. (1985). An empirical bayes approach to statistics. In Samuel Kotz & Norman L. Johnson (Eds.), *Breakthroughs in Statistics* (pp. 388–394). New York: Springer.

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the royal statistical society: Series B*, *71*, 319–392.

Shimko, D. C. (1993). Bounds of probability. *Risk Magazine*, *6*(4), 33–37.

Shumway, T. (2001). Forecasting bankruptcy more accurately: A simple hazard model. *The Journal of Business*, *74*, 101–124.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, *62*, 1139–1168.

Tetlock, P. C., Saar-Tsechansky, M., & Macskassy, S. (2008). More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance*, *63*, 1437–1467.

Tsay, Z. Y. (2008). *Chinese grammer*. Taipei: Wanjuan. (in Chinese).

Vasicek, O. (1977). An equilibrium characterization of the term structure. *Journal of Financial Economics.*, *5*, 177–188.

West, R. (1970). An alternative approach to predicting corporate bond ratings. *Journal of Accounting Research*, *7*, 118–127.

Wilhelmsen, M., Dimakos, X. K., Husebo, T., & Fiskaaen, M. (2009). Bayesian modelling of credit risk using integrated nested laplace approximations. Working paper. Oslo, Norway: Norwegian Computing Center.

Zmijewski, M. E. (1984). Methodological issues related to the estimation of financial distress prediction models. *Journal of Accounting Research*, *22*, 59–82.