Modeling and Forecasting Corporate Default Counts Using Hidden Markov Model

Lu Li and Jie Cheng

***Abstract*—In this paper, a Hidden Markov Model is employed to fit global, U.S. and European annual corporate default counts. The Expectation-Maximization algorithm is applied to calibrate all parameters while the standard errors of the estimated parameters are conducted by Monte Carlo method. Parametric bootstraps are used to compute the nonlinear forecasts. The empirical results show that the Hidden Markov model is useful in distinguishing the periods of expansion from the periods of recession (relative to the points identified by the NBER). Moreover, it obtains relatively satisfactory forecasts especially in capturing the state switching while incorporating more original observations.**

***Index Terms*—Corporate default counts, expectation-maximization algorithm, hidden Markov model, parametric bootstrap.**

I. Introduction

The issue regarding estimating potential risk levels and forecasting default events of financial assets has increasingly became the interest of many financial, economic, and mathematical researchers in contemporary society. Previously, due to the achievement of Moody [1], the Binomial Expansion Technique (BET) was created to estimate the expected loss of collateralized bond and loan obligations (CBOs and CLOs). However, it is ideal that there exists a pure binomial distribution with independent defaults. With the introduction of diversity score which is used to distinguish a smaller portfolio of independent and homogenous financial assets, it is easier to assume that all these financial assets (bonds and loans) have the same default probability and default independently, resulting in the binominal distribution regarding the quantity of observed default events in single time stage. More importantly, as mentioned by Düllmann [2], some shortages of BET method can be optimized to some extent by the model created by Davis and Lo [3]. In particular, it was related to infectious defaults which increase the default risk of other financial assets. There are two types of risk (normal risk and enhanced risk, respectively), and the latter risk level is enhanced by multiplying infectious factor k. In this case, the similar approach named Hidden Markov Model will be used to detect risk periods in the economy, and related parameters are estimated by Expectation-Maximization algorithm (EM algorithm). More importantly, what this paper pays more attention to concerns corporate default counts forecast, which is different from emphasizing on estimation process and detection of expansion and recession periods in previous researches. The forecast process is achieved by the parametric bootstrap approach according to Tsay [4], which is used to perform the nonlinear forecasts.

The content of this paper is divided into six aspects. Detailed methods or approaches utilized in this article will be included and explained in Section II and the simulation part Section III is to test the effectiveness of parameter estimation. Then imperial analysis in Section IV incorporates some small related aspects, data introduction, for example. Section V sketches the final conclusion, and the further improvement for this paper is offered in Section VI.

Manucsript received November 11, 2013; revised January 7, 2014.

Lu Li and Jie Cheng are with the Xi'an Jiaotong-Liverpool University, China (e-mail: lu.li1002@student.xjtlu.edu.cn).

II. Methodology

A. Model Introduction and Description

In this paper, a two-state discrete HMM is used, and two hidden states are normal risk state and enhanced risk state, respectively, denoting as 1 and 2. According to BET published by Moody's Investors Service [1], in this case, the default counts N in state 1 and 2 follow different binomial distributions with the parameters P_1 and P_2, representing the observed default probabilities in each state.

$$P_1(N)=\binom{n}{N}P_1^N(1-P_1)^{n-N} \tag{1}$$

$$P_2(N)=\binom{n}{N}P_2^N(1-P_2)^{n-N} \tag{2}$$

where n denotes total number of surviving financial assets (bonds or loans) in the market. More specially, the parameter P_2 is obtained by multiplying P_1 with one factor k $(k \geq 1)$, which describes enhanced effect in state 2.

Moreover, besides the number of states s (2 states in this paper) and the observations per state, the parameters of an HMM also include initial state distribution $\pi = P[q_1 = S_i]$ which means the probability regarding the initial observation occurrence in state i, an observation symbol probability distribution $B = \{b_j(m)\}$ to represent the probability of observing m events on state j (two binominal distributions here), and the state transition matrix $A = \{a_{ij}\}$ which describes the transition probability from state i to state j [5]. More specifically, in our approach, the parameters to describe the constant transition matrix are demonstrated as follows:

$$A=\begin{pmatrix} a_{11} & 1-a_{12} \\ 1-a_{22} & a_{22} \end{pmatrix} \tag{3}$$

where a_{11}, a_{22} represent the probability of retaining in state 1 and 2 respectively. Hence, complete parameters utilized in our two-state HMM are summarized as $\lambda\ (A, B, \pi)$.

B. Parameter Estimation

Given the real observation sequence $O = O_1 O_2 \cdots \cdots O_T$, the challenges are to estimate HMM parameters (A, B, k) and maximize $P(O|\lambda)$. In 1989, Rabiner recommended one efficient method named EM algorithm to cope with problems, which is utilized to calculate the maximum likelihood value when there is unobserved variables [5]. In forward-backward procedure, there are two separate variables containing forward variable $\alpha_t(i)$ and backward variable $\beta_t(i)$. In detail, $P(O_1 O_2 \cdots O_t, q_t = S_i|\lambda)$ can be defined as $\alpha_t(i)$ which represents given λ, the probability of the partial observation sequence when reaching state S_i at time *t*. Similarly, backward variable $\beta_t(i) = P(O_{t+1} O_{t+2} \cdots O_T, q_t = S_i$ denotes the probability of the rest observation sequence from *t*+1 to the final after arriving in state S_i at time t with given model parameters λ. In order to compute the probability of arriving state S_i at time *t* ($\gamma_t(i)$) and transition probability a_{ij} from state S_i at time t to state S_j at time *t*+1 ($\xi_t(ij)$), they can be defined in the following form:

$$\gamma_t(i) = \frac{a_t(i)\beta_t(i)}{\sum_{t=1}^{s} a_t(i)\beta_t(i)} \tag{4}$$

$$\xi_t(ij) = \frac{a_t(i)a_{ij}b_j(o_{t+1})\ \beta_{t+1}(i)}{\sum_{i=1}^{s}\sum_{j=1}^{s} a_t(i)a_{ij}b_j(o_{t+1})\ \beta_{t+1}(i)} \tag{5}$$

Detailed information can refer to Rabiner [5].

C. Forecasts Based on HMM (Parametric Bootstrap)

Unlike the non-parametric bootstrap, the parametric bootstrap is used to draw samples from a distribution formed from a sample set by a model [6]. In this case, nonlinear forecasts are calculated by the parametric bootstrap. Referring to Tsay [4], the values of $x_{T+1}, x_{T+2} \cdots x_{T+l}$ are computed by drawing new realizations from specified distribution of the model if estimated parameters are given, where T and l ($l > 0$) represent the forecast origin and the forecast horizon, respectively. Additionally, by the model, the original observations, the forecast of $x_{T+1}, x_{T+2} \cdots x_{T+l-1}$ and repeating the procedure (*M* times), *M* realizations of x_{T+l} can be obtained, and then the forecast of x_{T+l} is regarded as the average of *M* values drawn before.

In this paper, the forecasting process works in the following steps.

1) Start from forecast origin *T* and the record before *T* is the first data set to forecast.
2) Perform one-step ahead forecast. In detail, the start point is to calculate the expected smoothed probabilities in *T*+1 for each risk level ($\gamma_{T+1}(1)$ and $\gamma_{T+1}(2)$) given estimated parameters generated from all available data before *T*. After that, it is essential to randomly draw the default counts from two risk levels and compute the expectation of default count in *T*+1 by multiplying the $\gamma_{T+1}(1)$ and $\gamma_{T+1}(2)$ respectively. By reduplicative M realizations (1000 in this trial), the forecasting default count in $T+1$ is the sample average of 1000 expected default counts calculated before. The general forecast process is calculated below, assuming $X_{T+l}^{j}(1)\ and\ X_{T+l}^{j}(2)$ ($j = 1 \cdots M,\ l > 0$) are *j* realization for $T + l$ drew from state 1 and 2 at j times:

$$\gamma_{T+l}(1) = \gamma_{T+l-1}(1) * a_{11} + \gamma_{T+l}(2) * a_{21} \tag{6}$$

$$\gamma_{T+l}(2) = \gamma_{T+l-1}(1) * a_{12} + \gamma_{T+l}(2) * a_{22} \tag{7}$$

$$X_{T+l}^{j} = X_{T+l}^{j}(1) *\ \gamma_{T+l}(1) + X_{T+l}^{j}(2) *\ \gamma_{T+l}(2)\ \ j = 1 \cdots M \tag{8}$$

$$X_{T+l} = \frac{1}{M}\sum_{j=1}^{M} X_{T+l}^{j} \qquad j = 1 \cdots M \tag{9}$$

3) Incorporate one more real default count each time according to the order. The next is to re-estimate related parameters for each data set and repeat forecasting process until all the data are utilized.

Data length in varied regions for forecast is different. The forecast origin for global default counts is *T*=81 covering the period 1920-2000. Due to the limited data collected about U.S. and Europe, their forecast origins are quite shorter than globe's (*T*=29 from1981 to 2009 and *T*=24 from 1986 to2009, respectively).

D. Covariance Matrix

The standard errors of the estimated parameters (a_{11}, a_{22}, P_1, k) are computed by Monte Carlo Method, which is implemented in the Matlab. In particular, the square root of the values on the diagonal of the covariance matrix concerned is the standard errors of the estimated parameters mentioned before.

The initial step is to generate an observation sequence by prior estimated EM estimators, and repeat this process t times. Next, for each generated sequence, we need to re-estimate parameters (a_{11}, a_{22}, P_1, k). Finally, the covariance matrix is computed below:

$$\mathrm{C} = \frac{1}{t-1}\sum_{i=1}^{t}\left(\theta_i - \hat{\theta}\right)' \cdot \left(\theta_i - \hat{\theta}\right) \tag{10}$$

$$\hat{\theta} = \frac{1}{t}\sum_{i=1}^{t}\theta \tag{11}$$

where θ is a vector containing four estimated parameters for each generation.

TABLE I: SIMULATION RESULTS

Parameters	True parameter	Initial parameter(1st)	Estimated parameter	Initial parameter(2nd)	Estimated parameter
a_{11}	0.9	0.7	0.9057	0.8	0.9207
			(0.045583)		(0.04145)
a_{22}	0.9	0.75	0.9132	0.25	0.9157
			(0.045846)		(0.0527)
P_1	0.007	0.002	0.007	0.0015	0.0067
			(0.000385)		(0.00037)
k	4	2	3.7935	2	4.1991
			(0.254197)		(0.2839)

III. SIMULATION RESULTS

In order to testify the effectiveness of parameter estimation, referring to Davis [7], Zhu and Cheng [8], the similar method will be used in this paper. The first step is to simulate 100 default counts observations within one *n*=1000 bond portfolio. Meanwhile, two sets of initial parameters are chose

and applied in the EM algorithm. The detailed results are demonstrated in the Table I below, which satisfyingly agrees with the true parameters and supports the effectiveness of parameter estimation. (The brackets here represent the standard errors regarding corresponding estimated parameters)

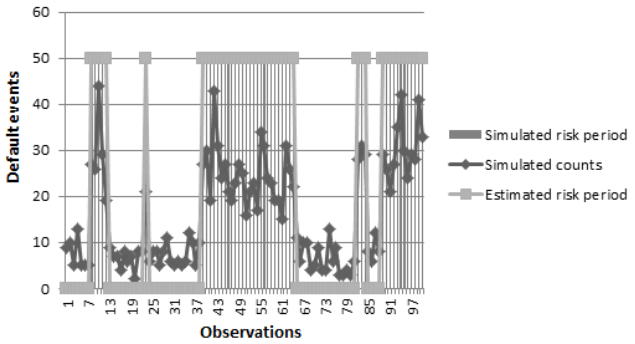

Fig. 1. Simulated default counts, simulated and estimated risk periods for 1st set of initial parameters. The solid bar demonstrates risk level in state 2. Hence, the algorithm is satisfying to detect enhanced risk periods.

IV. Imperial Analysis

A. Data Description

The data sources used in this paper consist of Moody's and Standard & Poor's annual default studies.

The global data are extracted from Exhibit 16 and Exhibit 30 of Moody's annual default study [9] which include the number of annual global cooperate issuer default events and annual issuer-weighted corporate default rates from 1920 to 2012. Here we just use actual global default counts from 1920 to 2000 to perform estimation, and the rest data started from 2001 will be applied to forecast. In particular, it should be noticed that all the default counts in Moody's report only cover Moody's all-rated cooperate issuers.

As for Europe default counts, it covers the period 1986-2012 in this paper. European default rates are collected from Exhibit 17 of Moody's European Corporate Default and Recovery Rates [10], and corresponding default counts come from Moody's annual default study (Excel data), Exhibit 18 [11].

The United States default counts derive from The Standard & Poor's annual U.S. corporate default studies Table I, covering U.S. default counts from 1981 to 2012 [12]. Meanwhile, these tables also offer corresponding annual default rate.

Additionally, U.S. and European default counts ranging from 1981 to 2009 are used to perform the estimation process and the forecast point begins in 2010, due to the short data length provided by Moody's and Standard & Poor's annual default studies.

Data regarding U.S. business cycle are collected from the National Bureau of Economic Research (NBER) [13], which is plotted in Fig. 3 roughly.

B. Default Definition

The statistical data collected from Moody's and Standard & Poor's annual default studies implements different default definitions (the difference is described in the annual U.S. Corporate Default Study And Rating transitions [12] and Moody's Rating Symbols and Definitions [14]). Moreover, the definition of issuer-weighted default rate is explained in the appendix of Latin American Corporate Default and Recovery Rates [15]

C. Estimation Results

The global, U.S. and European estimation results are demonstrated in the Table II with standard errors within the brackets.

TABLE II: Estimated Results of Globe U.S., and Europe

Parameters	Globe	U.S.	Europe
a_{11}	0.9412 (0.03335)	0.85 (0.092572)	0.766 (0.11869)
a_{22}	0.75 (0.21775)	0.75 (0.151191)	0.506 (0.23195)
P_1	0.0001 (0.000031)	0.003 (0.0000138)	0.0001(0.00026)
k	7.9418 (0.84115)	4.613 (0.4026)	19.0185(5.8858)

As can be seen from the Table II, the estimation results are quite reliable and stable. Fig. 2- Fig. 4 below demonstrate real observations in globe, U.S. and Europe, along with estimated state switching process. It is clear that the risk states estimated by our model are effective to detect enhanced risk periods, especially in U.S., which capture the real business cycle roughly.

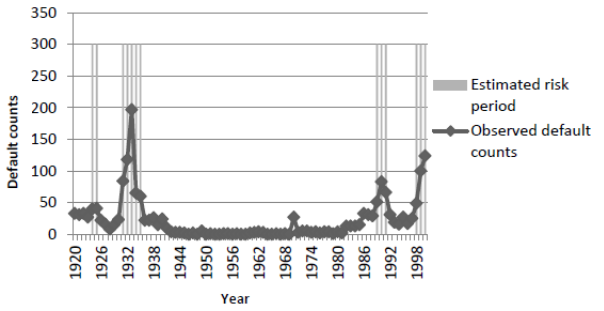

Fig. 2. Real global default counts with estimated risk periods.

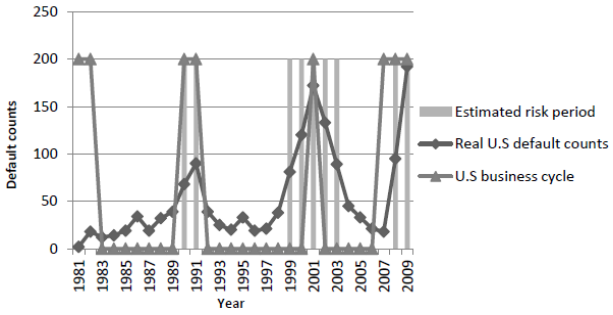

Fig. 3. Real U.S. default counts and estimated risk periods along with U.S. business cycle.

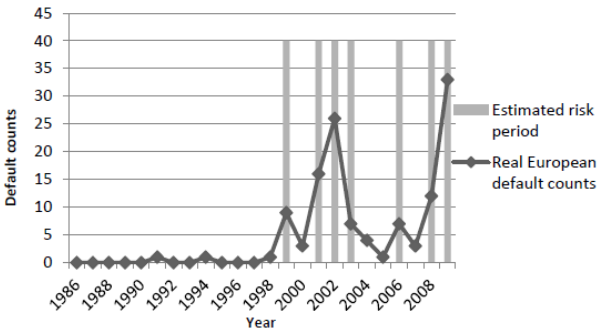

Fig. 4. Real European default counts with estimated risk periods.

D. Forecast results and Analysis

As mentioned in the previous section, the parametric bootstrap will be utilized to predict the global default counts from 2001 to 2012, U.S. and European default counts from 2010 to 2012.

The Fig. 5 below sketches the comparison of observed global default counts with their forecasts from 2001 to 2012. Moreover, the Table III contains two sets of smoothed probabilities in state 1 from 2001 to 2012, one is obtained from applying all available real global default counts from 1920 to 2012, and the other is from rolling estimation process while incorporating a new observation. The detailed data record regarding the comparison between the observed default counts and corresponding forecasts is included in the Table III as well.

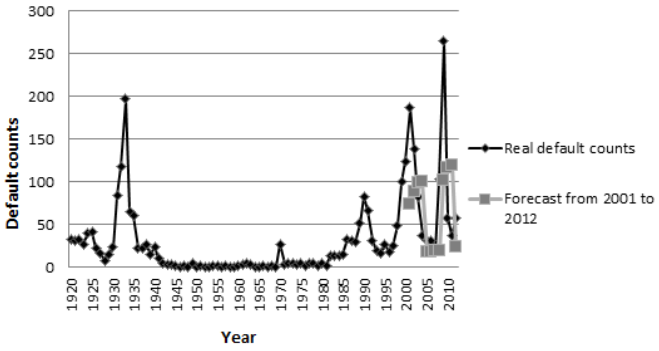

Fig. 5. Global default counts and forecasts.

TABLE III: GLOBAL SMOOTHED PROBABILITIES AND FORECAST RESULTS

	Smoothed probability (1920-2012)	Smoothed probability (Rolling process)	Real default counts	Forecast
2001	$2.3538 * 10^{-134}$	0.25003	187	74
2002	$3.0258 * 10^{-89}$	0.1871	138	89
2003	$4.9212 * 10^{-36}$	0.16714	82	98
2004	1	0.1549	37	100
2005	1	0.95668	31	18
2006	1	0.95769	31	18
2007	1	0.9583	18	19
2008	$2.2888 * 10^{-55}$	0.95888	103	19
2009	$3.6769 * 10^{-206}$	0.21444	265	102
2010	$6.1136 * 10^{-14}$	0.20293	58	116
2011	0.9997	0.1875	37	119
2012	$7.0126 * 10^{-13}$	0.94595	58	24

While comparing the real global annual default counts with corresponding forecasts from 2001 to 2012, it is interesting to find that there exist large differences between them. Actually, it is reasonable that about 75% of probability that the real default counts in 2001 remains at enhanced risk state, which results from the estimated parameter $a_{22} = 0.75$ computed by the data from 1920 to 2000. Our forecast smoothed probabilities are the weighted mean of being in two states, however, the real case is that observations only occur in one state, which results in such a large difference between the real and predicting case. Obviously, the results can be remedied by incorporating more original data. It is clear that the smoothed probabilities over time obtained by one-step ahead forecast and the parametric bootstrap approach approximately follow the tendency or fluctuation of real global default counts record, which reflects risk state switching.

As for U.S., The similar Fig. 6 and Table IV reflect its default counts, corresponding forecasts and two sets of smoothed probabilities in normal state ranging from 2010 to 2012. It should be noticed that two sets of smoothed probabilities consist of outcomes computed by whole annual U.S. default counts from 1981 to 2012 and one-step ahead forecast process at each rolling step since 2009.

More satisfying results can be obtained from U.S. default observations from 1981 to 2012. In detail, the transition probability a_{22} calculated by U.S. data covering the period 1981 - 2009 is 0.75 and the observation in 2009 is estimated at state 2, which lead to 75% of probability that the real default count in 2010 still remains at state 2. More importantly, the forecast smoothed probabilities are the weighted mean of being in two states, however, the real observations only occur in one state, which results in such a large difference between the observed default counts and forecasts. However, with more data, our results will be modified effectively. Obviously, the smoothed probabilities computed by rolling re-estimates process approximately catches the switch between normal and enhanced risk states.

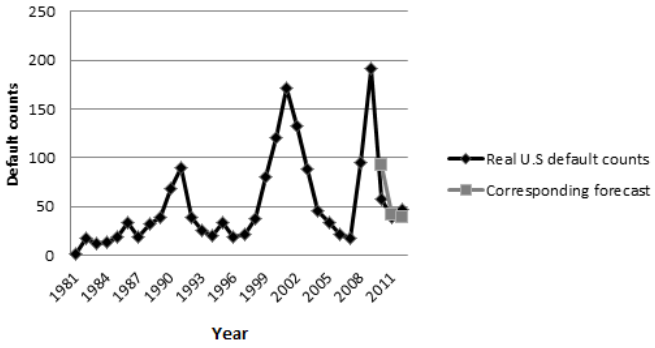

Fig. 6. U.S. default counts and forecasts.

TABLE IV: U.S. SMOOTHED PROBABILITIES AND FORECAST RESULTS

	Smoothed probability (1981-2012)	Smoothed probability (Rolling process)	Real default counts	Forecast
2010	0.9946	0.25	58	92
2011	1	0.81978	39	41
2012	1	0.85708	47	39

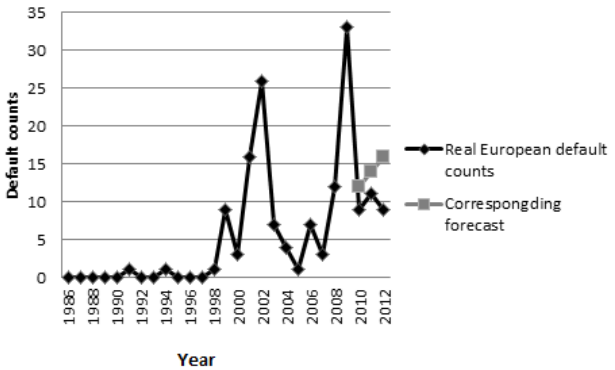

Fig. 7. European default counts and forecast.

TABLE V: EUROPEAN SMOOTHED PROBABILITIES AND FORECAST RESULTS

	Smoothed probability (1986-2012)	Smoothed probability (rolling process)	Real default counts	Forecast
2010	$2.34002 * 10^{-7}$	0.49398	9	12
2011	$7.3127 * 10^{-10}$	0.41639	11	14
2012	$7.2718 * 10^{-7}$	0.35583	9	16

The European outcomes are included in the Fig. 7 and Table V.

Although there are large differences between real European default counts and forecasts, what our model estimate can generally catch the state switching, and show that there is a higher probability remaining in recession periods (enhanced risk state) since 2010.

V. CONCLUSION

In this paper, a two-state Hidden Markov Model is used to fit real data in the macro-economy (Globe) and two geographic regions (U.S. and Europe). Related parameters are obtained by EM algorithm. In our case, the U.S. estimated enhanced risk periods capture and follow most recession periods in its real business cycle, which is beneficial for proving the effectiveness of our model. For forecast perspective, the parametric bootstrap is to apply in rolling one-step ahead forecast. Though there are relatively large differences between the real observations and their forecasts, generally, the approach used in this article captures real state switching effectively and forecasts accurately while incorporating more original data.

VI. FURTHER WORKS

Although the model applied in this article can detect enhanced risk periods and capture risk switching, it still needs further works to improve. In detail, firstly, we do not consider cross regional default counts correlation by assuming that they are independent. However, in reality, the global default events in a certain year may trigger the more default events in other geographic regions. Next, the data length is one of our major shortages as well, especially the European default counts only covering from 1986 to 2012, which may cause inaccurate estimation and forecast. Moreover, somehow Markov Chain includes autocorrelation function (ACF), where the next observation depends only on the current observation. There is one alternative approach named integer-valued autoregressive model (INAR) introduced by McKenzie, Alzaid and Al-osh [16], which may improve our method, because we find the significant ACF at lag 1, 2, and 3 while applying global default counts from 1920-2000.

ACKNOWLEDGMENT

Firstly, Lu Li would like to express thanks to Xi'an Jiaotong-Liverpool University (XJTLU) and XJTLU Summer Undergraduate Research Fellowship for funding and assisting in our project. Then it is very grateful to the Department of Mathematical Sciences in XJTLU for offering this precious opportunity to do this research.

REFERENCES

[1] Moody's Investors Service, "The binomial expansion method applied to CBO/ CLO analysis," New York: Moody's, 1996.
[2] K. Düllmann. (2006). Measuring business sector concentration by an infection model. Germany: Deutsche Bundesbank. [Online]. Available: http://www.fbv.kit.edu/symposium/10th/papers/Duellmann%20-%20Measuring%20Business%20Sector%20Concentration%20by%20an%20Infection%20Model.pdf
[3] M. Davis and V. Lo, "Modeling default correlation in bond portfolios," in *Mastering Risk, Applications*, London: Pearson Education, vol. 2, pp. 141-151, 2001.
[4] R. S. Tsay, *Analysis of Financial Time Series*, 2nd ed. Canada: John Wiley & Sons, INC, pp. 192, 2005.
[5] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, pp. 257-286, 1989.
[6] P. V. Heijden, H. Hart, and J. Dessens, "A parametric bootstrap procedure to perform statistical tests in a LCA of anti-social behavior," *Journal of Educational and Behavioral Statistics*, pp. 196-208, 1997.
[7] G. Giampieri, M. Davis, and M. Crowder, "A hidden Markov model of default interaction," *Quantitative Finance*, vol. 5, pp. 27-34, 2005.
[8] Y. J. Zhu and J. Cheng, "Using hidden Markova model to detect macro-economic risk level," *Review of Integrative Business and Economics*, vol. 2, no. 1, pp. 239-249, 2013.
[9] Moody's Investors Service, "Annual default study: corporate default and recovery rates, 1920-2012," New York: Moody's, 2013.
[10] Moody's Investors Service, "European corporate default and recovery rates, 1985 – 2012," New York: Moody's, 2013.
[11] Moody's Investors Service. (2013). Annual default study: corporate default and recovery rates, 1920-2012 - Excel data. [Online]. Available: https://www.moodys.com/researchandratings/research-type/default-ratings-analytics/default-studies/003009000/4294965103/4294966848/0/0/-/0/-/-/-/-/-/-/-/en/global/pdf/-/rra
[12] Standard& Poor's. (2013). 2012 Annual U.S. corporate default study and rating transitions. [Online]. Available: https://www.globalcreditportal.com/ratingsdirect/renderArticle.do?articleId=1098627&SctArtId=145785&from=CM&nsl_code=LIME
[13] National Bureau of Economic Research. (2012). U.S. business cycle expansion and contractions. [Online]. Available: http://www.nber.org/cycles/cyclesmain.html
[14] Moody's Investors Service, "Moody's rating symbols and definition," New York: Moody's, 2013
[15] Moody's Investors Service, "Latin American corporate default and recovery rates, 1990–July 2013," New York: Moody's, 2013
[16] M. Kachour and J. F. Yao, "First-order round integer-valued autoregressive (RINAR (1)) process," *Journal of Time Series Analysis*, vol. 30, no. 4, pp. 417-488, 2008

Lu Li was born on August 28, 1991, in Chengdu, Sichuan Province, China. She is one undergraduate student in Year 4 majoring in Financial Mathematics in Xi'an Jiaotong-Liverpool University (XJTLU) in Suzhou, Jiangsu Province, China. During the summer in 2013, she participated in the 2nd annual Summer Undergraduate Research Fellowship (SURF) funded by XJTLU. This SURF program encourages undergraduates to work on a research project under supervision, thus providing them with a valuable range of practical skills which will set them in good stead not only for the remainder of their studies at XJTLU, but also for future research projects at the postgraduate level and in their professional careers. Currently, she is involved in her final year project regarding financial time series with Dr. Jie Cheng from Mathematical Sciences Department in XJTLU.