Expert Systems With Applications 58 (2016) 210–220

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Nonparametric machine learning models for predicting the credit default swaps: An empirical study

Youngdoo Son[a], Hyeongmin Byun[b], Jaewook Lee[c,*]

[a] Research Institute of Engineering and Entrepreneurship, Seoul National University, Republic of Korea
[b] Department of Industrial Engineering, Seoul National University, Republic of Korea
[c] Department of Industrial Engineering, Seoul National University, Republic of Korea

ARTICLE INFO

Article history:
Received 18 December 2014
Revised 28 March 2016
Accepted 29 March 2016
Available online 4 April 2016

Keywords:
Financial forecasting
Nonparametric models
Credit default swap
Empirical analysis

ABSTRACT

Credit default swap which reflects the credit risk of a firm is one of the most frequently traded credit derivatives. In this paper, we conduct a comprehensive study to verify the predictive performance of nonparametric machine learning models and two conventional parametric models on the daily credit default swap spreads of different maturities and different rating groups, from AA to C. The whole period of data set used in this study runs from January 2001 to February 2014, which includes the global financial crisis period when the credit risk of firms were very high. Through experiments, it is shown that most nonparametric models used in this study outperformed the parametric benchmark models in terms of prediction accuracy as well as the practical hedging measures irrespective of the different credit ratings of the firms and the different maturities of their spreads. Especially, artificial neural networks showed better performance than the other parametric and nonparametric models.

1. Introduction

Credit market is one of the most important financial markets and has received wide attention especially from the credit crisis in 2008. A default probability is one of the typical measures to represent the credit risk of a firm or nation but it is difficult to determine since many firms and nations, or obligors, are linked by various contracts and obligations and thus a credit-related event, including default, of one obligor may affects many other obligors. The default probability is usually measured by the credit derivatives traded in the credit market because their prices are not highly affected by the other factors than credit risk unlike defaultable bond prices. For example, Bühler and Trapp (2009) showed that the 95% of the credit default swap (CDS) spread stems from the credit risk while only 4% from the liquidity. The mispricing of these derivatives can lead to misunderstanding of default probability; thus, accurate pricing for credit derivatives from the credit crisis period has become an important consideration.

During the last two decades, many researches have been made to price credit derivatives and their models can be categorized into two classes of models. One class of models, called *structural models*, assume that a certain stochastic process for the fundamental value of the firm and defines an event of default as the fundamental value hits a predetermined barrier (Black & Cox, 1976; Finger, Finkelstein, Lardy & Pan, 2002; Merton, 1974). the other class of models, called *reduced-form models* or *intensity-based models*, assume that the default is driven by an exogenous factors and an event of default follows a Poisson process with a stochastic intensity (Cox, Ingersoll, & Ross, 1985; Jarrow & Turnbull, 1995; Vasicek, 1977). There have also existed a large number of studies that compared those models by the predicted credit derivative prices (Bakshi, Madan, & Zhang, 2006; Duffee, 1999; Eom, Helwege, & Huang, 2004; Gündüz & Uhrig-Homburg, 2011; Jones, Mason, & Rosenfeld, 1984; Lyden & Saraniti, 2001; Ogden, 1987) but there has not been a robust conclusion that a certain model overwhelms the others for pricing and predicting credit derivatives traded in the real market.

On the other hand, nonparametric learning models have extensively been used to predict financial time series in recent years due to their flexibility which fits the models to the data well. Most of those results have been focused on the stock (Chen, Shih, & Wu, 2006; Liao & Chou, 2013; Son, Noh, & Lee, 2012; Ticknor, 2013) and its derivative markets (Han & Lee, 2008; Hutchinson, Lo, & Poggio, 1994; Park, Kim, & Lee, 2014; Park & Lee, 2012; Yang & Lee, 2011) and achieved accurate prediction results. However, relatively a few studies have been conducted for the other markets including the fixed-income market (Cao & Tay, 2003; Kim & Noh, 1997) and

* Corresponding author. Tel.: +8228807176.
E-mail addresses: hand02@snu.ac.kr (Y. Son), ragnarok@snu.ac.kr (H. Byun), jaewook@snu.ac.kr (J. Lee).

http://dx.doi.org/10.1016/j.eswa.2016.03.049
0957-4174/

the foreign exchange market (Bhattacharyya, Pictet, & Zumbach, 2002; Osuna, Freund, & Girosi, 1997). For the credit market, most of studies using learning methods have concentrated on credit rating analysis. Lee (2007) and Kim and Ahn (2012) used support vector machines to classify the rating of firm and Huang, Chen, Hsu, Chen, and Wu (2004) classified the rating of corporate bonds using both support vector machines and artificial neural networks. For credit derivatives pricing, Gündüz and Uhrig-Homburg (2011) applied the support vector regression (Drucker, Burges, Kaufman, Smola, & Vapnik, 1997) to predicting one-dimensional output corresponding five-year maturity CDS spread of one firm using the spreads of other firms at the same moment called a cross-sectional design or using the past value of spreads of the same firm called a time series design and compared it with those of the Merton model (Merton, 1974) and the constant intensity model (Jarrow & Turnbull, 1995). However, only one specific spread was used in prediction although considering and predicting the spreads of other liquid maturities at the same time are practically important and no advanced state-of-the-art machine learning models other than support vector regression were used for comparison.

To our knowledge, no empirical studies have been made that prices and predict the CDS spreads with diverse maturities simultaneously using several nonparametric learning models and even the earlier studies on other financial markets such as stocks or options were made to determine and predict one-dimensional outputs for its price values. In this paper we aim to conduct a comprehensive study that compares the predictive power of several nonparametric models using the real market CDS spread data of diverse maturities from January 2001 to February 2014 as well as those of different credit ratings. Since the spreads with several different maturities are predicted simultaneously, the employed models can capture the effect of the past spreads with different maturities on the current spreads. For our experiment, we applied four well-known state-of arts nonparametric learning regression models (*support vector regression* (Drucker et al., 1997), *artificial neural networks, Bayesian neural networks*, and *Gaussian processes* (Cressie, 1993; Rasmussen, 1996) to prediction of six dimensional outputs consisting of CDS spreads with six different maturities, 1, 2, 3, 5, 7, and 10 years. Also to verify the relative predictive performance of nonparametric learning models, we have applied two benchmark parametric models, called constant intensity model (Jarrow & Turnbull, 1995) and Cox-Ingersoll-Ross (CIR) intensity model (Cox et al., 1985).

The organization of this paper is as follows. In the next section, we review some literatures related to this study. Then, we give a brief description of nonparametric machine learning models and two benchmark parametric models, constant intensity model and CIR model, used to price the credit derivatives in Section 3. We describe the data used for this study and explain the design of the experiment and the performance measure and present the experimental results with some discussions on the results in Section 4 and 5. Finally we conclude this paper and provide directions for future work in Section 6.

2. Related work

Since the credit market is one of the largest markets among several financial markets, there have been a large number of literatures that related to the analysis of the CDS spreads. In this section, we reviewed some milestone literatures related to the parametric models of the CDS spreads and some recent studies related to the analysis of the CDS spreads from the different point of view.

There have been two main streams of the parametric models about credit derivatives, the *structural models* and the *reduced-form models*. In structural models, the asset value of the firm is assumed to follow a certain stochastic process, usually a geometric Brownian motion, and the even of default of the firm is defined as when the firm's asset value hits a predetermined barrier or becomes below it. Merton (1974) assumed that the asset value of the firm follows the geometric Brownian motion and the default occurs if the value of the firm is below the liability at the maturity. Therefore the default can only occur at the maturity and the formula for European option in Black and Scholes (1973) can be applied to find the theoretical value of the CDS contract. In contrast with Merton (1974), the first passage model (Black & Cox, 1976) defined the event of default of a firm as when the asset value of the firm hits the predetermined barrier. Thus the default can occur any time before maturity in this model. CreditGrades model (Finger et al., 2002) is one of the recent and widely used structural models, which imposed the randomness on the default barrier. The main framework of CreditGrades model is similar to the first passage model, but the barrier follows a lognormal distribution to reflect the uncertainty in the recovery. In reduced-form models, another main stream of credit derivative models, assumes that the default is caused by the exogenous factors which are independent of market information and there is no information about the arrival time of default. Thus, the occurence of default is usually assumed to follow the poisson process in the reduced-form models. Jarrow and Turnbull (1995) firstly proposed the reduced-form model with the constant intensity of the poisson process but the inhomogeneous poisson process like Vasicek model (Vasicek, 1977), CIR model (Cox et al., 1985), and even the intensities including jumps (Schoutens & Cariboni, 2009) have also been employed to analyze the credit derivatives. However, both of these structural and reduced-form models explained above are parametric models that assume a certain form of the solution. Although, these parametric models have an advantage in finding the financial implication like the explanation of stylized facts, they do not concentrate on fitting the data itself compared to the machine learning models. The more detailed description of these parametric models for credit derivatives is well summarized in Lando (2004) and Schoutens and Cariboni (2009).

CDS has also been extensively studied and used in recent literatures. Brigo and El-Bachir (2010) proposed the exact pricing model of defaultable swaptions under the assumption of the jump diffusion process and used the CDS spreads for calibration. Cont and Kan (2011) proposed the multivariate time series model for the CDS spreads of three different firms and found that the proposed model showed better performance than the intensity models with jumps and random walk models for predicting loss quantiles of CDS portfolios. Jarrow, Li, and Ye (2011) found the statistical arbitrage opportunities based on reduced-form models. Bianchi and Fabozzi (2015) compared several parametric models, Brownian motion, three Lèvy processes, and a Sato process, for the prediction of CDS prices. Gündüz and Uhrig-Homburg (2011) employed the support vector regression, Merton model, and the constant intensity model, which are a nonparametric machine learning model, a structural model, and a reduced-form model, respectively, to predict the CDS spreads and compared the results. However, they only used the CDS spreads with one specific maturity and no state-of-the-art machine learning models other than the support vector regression were not employed. Most of the literatures above, except Gündüz and Uhrig-Homburg (2011), also focused on or used the parametric models to find the CDS spreads. However, since these parametric models assume a special form of the model with predefined parameters, they cannot use the information which is not predefined in the model compared to the machine learning models which are versatile in the kinds and number of input variables. For example, Galil, Shapir, Amiram, and Ben-Zion (2014) found that the three variables, stock return, the change in the stock return volatility, and the change in the median CDS spreads in the rating class, were effective to the change of the CDS spreads and

Bijlsma, Lukkezen, and Marinova (2014) measured the too-big-to-fail funding advantages from the CDS spreads of small banks. The machine learning models easily allow these variables as input variables without a large change of the model while a new model should be constructed to deal with new inputs for the parametric models. Mayordomo, Peña, and Schwartz (2014) compared six major data provider of the CDS spread data and found that there existed significant deviations among them. However, the framework using nonparametric machine learning models employed in this study can be applied to the CDS spreads data from any data provider without any change.

3. Preliminaries

In this section, we briefly describe some preliminaries of this study. First, two benchmark parametric models, *constant intensity model* and *Cox-Ingersell-Ross model*, are described and then four nonparametric machine learning models employed in this study are presented. For more detailed descriptions can be found in Lando (2004) and Schoutens and Cariboni (2009), for parametric models, and Bishop (2006), for nonparametric models.

3.1. Parametric Constant intensity model

Constant intensity model (Jarrow & Turnbull, 1995) is one of the widely used reduced-form models for credit derivative pricing. Reduced-form models generally assume that the default happens stochastically and independently from market information thus they do not assume any fundamental values unlike the structural models which assume a fundamental asset value process of a firm. The process of default in the reduced-form models is usually described as a Cox process (Cox, 1955), which is a Poisson process with a stochastic intensity. The intensity, or hazard rate, of default process is defined as

$$\lambda_t = \lim_{\delta t \to 0} \frac{P(t < \tau < t + \delta t)}{P(\tau > t)\delta t} \tag{1}$$

where $P(\tau \in B)$ denotes the probability that the default time τ is included in a set of time B. Once the intensity process is defined, the default probability density $P(\tau \in [t, t + dt])$ can be easily calculated as

$$P(\tau \in [t, t + dt]) = \mathbb{E}[\lambda_t e^{-\int_0^t \lambda_s ds}]dt. \tag{2}$$

In constant intensity model, the intensity process is defined as a constant, i.e. $\lambda_t = \lambda$. Thus, the Cox process for the default becomes a Poisson process with intensity λ and so the default time follows the exponential distribution with the parameter λ. Thus the survival probability in this model simply becomes $P(\tau > t) = e^{-\lambda t}$ and the premium leg and protection leg become as follows (Duffie & Singleton, 2003):

$$V_{\text{premium}} = F\hat{s}\sum_{i=1}^{N} e^{-(r(i)+\lambda)T(i)}(T(i) - T(i-1)) \tag{3}$$

$$V_{\text{protection}} = F(1 - R)\sum_{i=1}^{N} e^{r(i)T(i)}(e^{-\lambda T(i-1)} - e^{-\lambda T(i)}) \tag{4}$$

where F is the notional value of the contract, $r(i)$ is the interest rate at time $T(i)$, R is the recovery rate, and $\hat{s}$ is the theoretical fair spread. Equating two values above, we obtain the fair CDS spread in this model:

$$\hat{s} = \frac{(1 - R)\sum_{i=1}^{N} e^{r(i)T(i)}(e^{-\lambda T(i-1)} - e^{-\lambda T(i)})}{\sum_{i=1}^{N} e^{-(r(i)+\lambda)T(i)}(T(i) - T(i-1))}. \tag{5}$$

and simply $\hat{s} = \frac{(1-R)(e^{\lambda \Delta t} - 1)}{\Delta t}$ when the time interval $T(i+1) - T(i) = \Delta t$ for all $i = 1, \ldots, N$.

3.2. CIR intensity model

In CIR model, the intensity process of the Cox process, λ_t has the dynamics of the form

$$d\lambda_t = \kappa(\theta - \lambda_t)dt + \sigma\sqrt{\lambda_t}dW_t \tag{6}$$

where W_t is the standard Brownian motion, $\kappa > 0$ is the parameter describing mean-reverting rate of the process, θ is a long-term mean of process, and σ is the volatility which adjusted by $\sqrt{\lambda_t}$. Since CIR intensity model is a special case of an affine jump diffusion process, which has been extensively studied for the . For example, it is known that if the intensity follows an affine jump diffusion process, the corresponding default probability has the form as

$$P(t) = \exp A(t) + B(t)\lambda_t \tag{7}$$

where $A(t)$ and $B(t)$ can be obtained by Riccati equations (Brigo & Mercurio, 2007). Then the theoretical value of CDS spreads can be obtained by equating the premium leg and protection leg of the contract,

$$V_{\text{premium}} = F\hat{s}\sum_{i=1}^{N} e^{-\{r(i)T(i)+A(T(i))+B(T(i))\lambda_{T(i)}\}}(T(i) - T(i-1)) \tag{8}$$

$$V_{\text{protection}} = F(1 - R)\sum_{i=1}^{N} e^{r(i)T(i)} \cdot [e^{-\{A(T(i-1))+B(T(i-1))\lambda_{T(i-1)}\}} - e^{-\{A(T(i))+B(T(i))\lambda_{T(i)}\}}] \tag{9}$$

where F is the notional value of the contract, $r(i)$ is the interest rate at time $T(i)$, R is the recovery rate, and $\hat{s}$ is the theoretical fair spread.

3.3. Artificial neural networks

Artificial neural networks (Rosenblatt, 1962) are extensively used highly nonlinear nonparametric model which can be used for the regression task. Mimicking a human brain, an artificial neural network model is composed of layers which also consist of nodes, conducting a role as neurons in the brain. There are usually three types of layers in an artificial neural network: input layer, output layer, and hidden layer. The input layer is the first layer and it has nodes that propagates the value of input variables to the next layer. The output layer is the last layer has nodes that make the overall outputs. The hidden layers are located at between the input layer and the output layer and they have nodes that makes the nonlinear output from inputs propagated from the previous layer. The nonlinear output $f(\mathbf{x})$ of each node has a form as follows:

$$f(\mathbf{x}) = g(\sum_i w_i x_i) \tag{10}$$

where $\mathbf{x}$ is the input vector from the previous layer and g is an activation function. The sigmoid functions, S-shaped functions such as hyperbolic tangent function, logistic function, and probit function, are commonly used for the activation function. Providing the activation function and transform functions are given or chosen by the user, training of artificial neural network means the optimization of weights w_i's in (10). The widely used optimization algorithm is *back-propagation algorithm* (Rumelhart, Hinton, & Williams, 1986).

In back-propagation algorithm, the weights are chosen backwardly from the output layer to the first hidden layer to minimize the loss, squared sum of errors in usual. The detailed derivation of the backpropagation algorithm for multilayer artificial neural networks is as follows.

Assume that there are $N+1$ layers, index 0 for the input layer and index N for the output layer. Let w_{ij}^n be a weight from ith node of $n-1$th layer to jth node of nth layer, o_j^n be the output value of the jth node in nth layer, and x_j^n be the input of jth node in nth layer. In other words,

$$o_j^n = g(x_j^n) \tag{11}$$

$$x_j^n = \sum_i w_{ij}^n o_i^{n-1} \tag{12}$$

where g is the activation function. The goal of the backpropagation algorithm is to find the partial derivative $\frac{\partial E_p}{\partial w_{ij}^n}$ where $E_p = \frac{1}{2}\sum_k (y_k - o_k^N)^2$ is the training error for the single input point indexed by p. If $n = N$, the output layer, the computation of the partial derivative can be conducted as follows:

$$\begin{aligned}\frac{\partial E_p}{\partial w_{ij}^N} &= \frac{\partial E_p}{\partial o_j^N}\frac{\partial o_j^N}{\partial x_j^N}\frac{\partial x_j^N}{\partial w_{ij}^N} \\ &= -(y_j - o_j^N) g'(x_j^N) o_i^{N-1}.\end{aligned} \tag{13}$$

However, if $n < N$, the computation of the partial derivative is not trivial. Consider

$$\frac{\partial E_p}{\partial o_j^n} = \sum_l \frac{\partial E_p}{\partial o_l^{n+1}}\frac{\partial o_l^{n+1}}{\partial x_l^{n+1}}\frac{\partial x_l^{n+1}}{\partial o_j^n}. \tag{14}$$

It can be observed that the partial derivative $\frac{\partial E_p}{\partial o_j^n}$ can be computed if $\frac{\partial E_p}{\partial o_l^{n+1}}$ is known for every l since the other partial derivatives can be easily calculated from (11) and (12). Then, we can calculate

$$\frac{\partial E_p}{\partial w_{ij}^n} = \frac{\partial E_p}{\partial o_j^n}\frac{\partial o_j^n}{\partial x_j^n}\frac{\partial x_j^n}{\partial w_{ij}^n}$$

from (14). Finally, given the learning rate η, the backpropagaion algorithm updates the weights as

$$w_{ij}^n \leftarrow w_{ij}^n + \eta \frac{\partial E_p}{\partial w_{ij}^n} \tag{15}$$

until one of the stopping criteria satisfies.

The artificial neural networks show good performances after training the weights but it is difficult in this model to determine the relationship between inputs and outputs, or the meaning of the model parameters. It is a common characteristic of many machine learning methods.

3.4. Bayesian neural networks

Bayesian neural network is a variant of neural network training algorithm, which was originally proposed in MacKay (1992). Similar to other algorithms with Bayesian nature, this algorithm assumes a Gaussian-type prior over weights on the networks, or equivalently regularizes the error function via sum of squares of weights.

Suppose we have n input-output pairs $\{x_i, y_i\}$ with $y_i = f(x_i) + \epsilon_i$, where ϵ_i are i.i.d Gaussian errors. In Bayesian neural network setting, the objective function is represented as $F(\mathbf{w}, \alpha, \beta; \{x_i, y_i\}) = \beta E_D + \alpha E_W$, where $E_D = \sum_{i=1}^n (y_i - a_i)^2$ is the sum of squares of errors from network output a_i and target y_i corresponding to x_i, and E_W is the sum of squares of the network weights. The relative importance between regularization and fitting the data is determined by adjusting relative size between α and β, which can be done by maximizing the posterior distribution $P(\alpha, \beta | \{x_i, y_i\})$. Thus the training of Bayesian neural networks implies to find the weights, $\mathbf{w}$, and the relative importance parameters, α and β, simultaneously.

Dan Foresee and Hagan (1997) proposed an algorithm which iteratively optimizes the weight and parameters, using Gauss-Newton approximation to the Hessian of the objective function F. After initializing α, β and weights $\mathbf{w}$, the algorithm to find appropriate α, β, and $\mathbf{w}$ fitting the data runs as follows:

1. Take one step of the Levenberg-Marquardt algorithm to minimize the objective $F(\mathbf{w}) = \beta E_D + \alpha E_W$ and find new weights $\mathbf{w}$.
2. Compute the effective number of parameters $\gamma = N - 2\alpha \mathrm{tr}(\mathbf{H})^{-1}$, making use of the Gauss-Newton approximation to the Hessian: $\mathbf{H} = \nabla^2 F(\mathbf{w}) \approx 2\beta J^T J + 2\alpha I_N$ where J is the Jacobian matrix of the training set errors.
3. Compute new estimates for the objective function parameters: $\alpha = \frac{\gamma}{2E_W(\mathbf{w})}$ and $\beta = \frac{n-\gamma}{2E_D(\mathbf{w})}$.
4. Return to (1) and repeat the procedure until the stopping criterion is satisfied.

3.5. Support vector regression

Support vector regression (Drucker et al., 1997) is a kernel regression that minimizes the ϵ-insensitive loss function

$$\mathcal{L}(y_1, y_2) = \max\{\epsilon, |y_1 - y_2|\} - \epsilon \tag{16}$$

with some $\epsilon > 0$. This loss function is zero if $|y_1 - y_2| < \epsilon$ and $|y_1 - y_2| - \epsilon$ otherwise. Defining the regression function $f(\mathbf{x}, \mathbf{w}) = <\mathbf{w}, \boldsymbol{\phi}(\mathbf{x})> + b$ with basis functions $\boldsymbol{\phi}$ and penalize with the errors larger than ϵ, the support vector regression problem, which minimizes $\|\mathbf{w}\|^2$ to reduce the complexity of model, becomes

$$\min_{\mathbf{w}, b} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^n (\xi_i^+ + \xi_i^-) \tag{17}$$

with the constraints

$$\begin{aligned} y_i - f(\mathbf{x}_i, \mathbf{w}) &\le \epsilon + \xi_i^+ \\ f(\mathbf{x}_i, \mathbf{w}) - y_i &\le \epsilon + \xi_i^- \\ \xi_i^+, \xi_i^- &\ge 0 \end{aligned} \tag{18}$$

for all $i = 1, \ldots, n$. Using Karush-Kuhn-Tucker conditions, we can get the following dual problem

$$\begin{aligned}\max_{\boldsymbol{\alpha}^+, \boldsymbol{\alpha}^-} &-\frac{1}{2}\sum_{i=1}^n \sum_{j=1}^n (\alpha_i^+ - \alpha_i^-)(\alpha_j^+ - \alpha_j^-) k(\mathbf{x}_i, \mathbf{x}_j) \\ &-\epsilon \sum_{i=1}^n (\alpha_i^+ - \alpha_i^-) + \sum_{i=1}^n (\alpha_i^+ - \alpha_i^-) y_i \end{aligned} \tag{19}$$

with the constraints $0 \le \alpha_i^+, \alpha_i^- \le C$ where the kernel function, $k(\mathbf{x}_i, \mathbf{x}_j) = <\boldsymbol{\phi}(\mathbf{x}_i), \boldsymbol{\phi}(\mathbf{x}_j)>$ and $\mathbf{w} = \sum_{i=1}^n (\alpha_i^+ - \alpha_i^-)\boldsymbol{\phi}(\mathbf{x}_i)$. The training of the support vector regression implies to solve this dual problem and it can be solved by a quadratic programming solver including LIBSVM (Chang & Lin, 2001). Then the predictive value for the new input $\mathbf{x}^*$ becomes

$$y^* = \sum_{i=1}^n (\alpha_i^+ - \alpha_i^-) k(\mathbf{x}_i, \mathbf{x}^*) + b \tag{20}$$

where $b = y_k - \sum_{i=1}^n (\alpha_i^+ - \alpha_i^-) k(\mathbf{x}_i, \mathbf{x}_k)$ for any $k = 1, \ldots, n$. For more details, see Drucker et al. (1997) and Bishop (2006).

For multiple output version, Pérez-Cruz, Camps-Valls, Soria-Olivas, Pérez-Ruixo, Figueiras-Vidal and Artes-Rodriguez (2002) proposed that the multiple output support vector regression where the ϵ-insensitive zone has a shape of hypersphere instead of a shape of hypercubic. This hypersphere ϵ-insensitive zone showed better performance than the same volume of the hypercubic ϵ-insensitive zone (Sánchez-Fernández, de Prado-Cumplido, Arenas-Garcia, & Pérez-Cruz, 2004; Tuia, Verrelst, Alonso, Pérez-Cruz, & Camps-Valls, 2011).

3.6. Gaussian process regression

Gaussian process regression (Cressie, 1993; Rasmussen, 1996) is a collection of random variables such that any finite combination of them follows the Gaussian distribution. A Gaussian process $f(\mathbf{x})$ can be completely determined by its mean function $m(\mathbf{x})$ and covariance function $k(\mathbf{x}, \mathbf{x}')$ as

$$m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})]$$
$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))].$$

Assume that the set of data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ is given with the noisy output y_i where the variance of the output noise is denoted by σ^2. Then, the covariance of the output vector $\mathbf{y} = (y_1, \dots, y_n)$ is given as

$$\mathrm{cov}(\mathbf{y}) = \mathbf{K}$$

where $\mathbf{K}$ is an $N \times N$ matrix whose i,jth entry is $k(\mathbf{x}_i, \mathbf{x}_j)$. Then, with a new input $\mathbf{x}^*$ and define the mean function $m(\mathbf{x}) = 0$,

$$\begin{bmatrix} \mathbf{y} \\ f^* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K} + \sigma^2\mathbf{I} & \mathbf{k}_*^T \\ \mathbf{k}_* & k_{**} \end{bmatrix}\right) \tag{21}$$

where $f^* = f(\mathbf{x}^*)$, $k_{**} = k(\mathbf{x}^*, \mathbf{x}^*)$, and $\mathbf{k}_* = (k(\mathbf{x}_1, \mathbf{x}^*), \dots, k(\mathbf{x}_n, \mathbf{x}^*))^T$. Then the distribution for predictive output f^* can be easily calculated by using the conditional distribution for normal distribution as follows:

$$P(f^*|\mathcal{D}) = \mathcal{N}(\mathbf{k}_*^T(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{y}, k_{**} - \mathbf{k}_*^T(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{k}_*). \tag{22}$$

As shown in (22), the Gaussian process regression gives the variance of the predictive output as well as the mean value, thus it belongs to the class of Bayesian regression.

Training of Gaussian process refers to optimizing the hyperparameters in the kernel function k and the output noise σ^2 by maximizing the log-likelihood function

$$\log P(\mathbf{y}|\mathcal{D}) = -\frac{1}{2}\mathbf{y}^T(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{y} - \frac{1}{2}\log\det(\mathbf{K}+\sigma^2\mathbf{I}) - \frac{N}{2}\log 2\pi. \tag{23}$$

The optimization can be done by usual techniques such as gradient descent and quasi-Newton method. The log-likelihood function in (23) varies with the exploited kernel functions and the most widely used kernel function is a squared exponential kernel function which has the form $k(\mathbf{x}, \mathbf{x}') = C\exp(-\gamma\|\mathbf{x} - \mathbf{x}'\|^2)$ where C and γ are hyperparameters. For more details, see Williams and Rasmussen (2006) and Bishop (2006).

4. Design of experiments

4.1. Data description

We used daily CDS contract data obtained from *MARKIT* database. The whole period of data set used is from January 2001 to February 2014. First, we eliminated data by the type of currency and region. The only US dollar denominated and Northern American contracts were used for the experiments. Then we selected five representative firms for each implied rating, AA, A, BBB, BB, B, and C. The implied rating is graded based on the five year CDS spread of the firm. Also, any two firms having the same rating are not included in the same industrial sector. The names and tickers of the selected firms are shown in Table 1. For the CDS spreads we have used six different maturities, 1, 2, 3, 5, 7, and 10 years, since they are very liquidly traded credit derivatives among several credit derivatives.

The statistics of the selected data are summarized in Table 2 and 3. The spreads are represented as a percentage. Noticeably, data of most rating groups have very high standard deviations and maxima. This is a typical feature of the positively skewed data set and this coincides the highly positive skewness values as shown in the tables. This positive skewnesses seem to be originated from the high CDS spreads of the global financial crisis period.

Table 1
Names and tickers of the selected firms.

Implied rating	Ticker	Name
AA	AMGN	Amgen, Inc.
	CSX	CSX Corporation
	IBM	International Business Machines Corp.
	JPM	JPMorgan Chase & Co.
	T	AT&T, Inc.
A	CAT	Caterpillar Inc.
	IP	International Paper Co.
	AIG	American International Group, Inc.
	CU	First Trust ISE Global Copper Index Fund
	CNP	CenterPoint Energy, Inc.
BBB	CA	CA, Inc.
	F	Ford Motor Co.
	MAY	Meadow Bay Gold Corp.
	PH	Parker-Hannifin Corp.
	N	NetSuite Inc.
BB	KBH	KB Home
	HRB	H&R Block, Inc.
	SLMA	SLM Corp.
	DPL	Diplomat Pharmacy, Inc.
	EP	El Paso LLC
B	AMKR	Amkor Technology, Inc.
	INTEL	Intelsat Ltd.
	PLCOAL	Peabody Energy Corp.
	THC	Tenet Healthcare Corp.
	FST	Forest Oil Corp.
C	BOW	Bowater Inc.
	DYN	Dynegy Inc.
	AMR	Armour Group PLC
	VC	Visteon Corp.
	JCP	J C Penney Company Inc.

Fig. 1 shows the term structures of the mean and median spreads for each rating group. In the graph of mean spreads, we can see that the spreads decrease as the implied rating improves when the maturity is fixed except for the group of BBB rating. The reason for this exception is that the mean value is sensitive to some large values. Especially, Ford Motor Company in the group of BBB has very large maximum spreads values because the automotive industry was one of the industries that had a very severe situation in the financial crisis period. The graph of median spreads shows more typical structures than that of mean spreads. The spreads decrease as the implied rating improves except for the B and C grades when the maturity is greater than or equal to five years. In addition, the spreads increase when the maturity increases in the graph of median spreads while there are decreasing term structure graphs for some implied rating groups in the graph of mean spreads.

4.2. Experimental procedures

We predicted the CDS spreads for six different maturities using the CDS spreads of past 14 days and those past values are employed as the input variables without any manipulations. In Frühwirth and Sögner (2006), the constant intensity model is analyzed with real data from German market and it was observed that using 5 to 25 preceding days for the analysis resulted in good performances. Especially, using 14 preceding days gave one of the best results. Gündüz and Uhrig-Homburg (2011) also used the spreads of 14 past days to predict the future spread values with Merton model, constant intensity model, and support vector regression. Therefore, following these previous literatures, we also decided to use the spreads of 14 preceding days as input variables.

Table 2
Basic statistics of the selected data set of the AA to BBB rating groups. The spreads are represented as a percentage (100 bp).

Rating	Statistics	Maturity					
		1Y	2Y	3Y	5Y	7Y	10Y
AA	Mean	0.3705	0.4424	0.5077	0.6418	0.7211	0.8081
	Std.dev.	0.6938	0.6909	0.6532	0.6446	0.6292	0.6173
	Min	0.0140	0.0223	0.0305	0.0574	0.0825	0.1120
	Median	0.1572	0.2316	0.3035	0.4410	0.5470	0.6544
	Max	8.4001	7.8672	7.7486	7.3746	7.0579	6.8820
	Skew	6.2316	5.7746	4.9098	4.1485	3.9504	3.7062
A	Mean	1.1168	1.1935	1.2925	1.4633	1.5205	1.5799
	Std.dev.	3.4282	2.9976	2.8149	2.4867	2.2196	2.0047
	Min	0.0110	0.0200	0.0381	0.0677	0.0900	0.1023
	Median	0.3021	0.4304	0.5712	0.8129	0.9392	1.0560
	Max	67.7230	53.0722	45.6853	38.1119	33.2684	29.1860
	Skew	8.9471	7.6979	6.9729	6.2172	5.7942	5.4023
BBB	Mean	3.2994	3.5232	3.6397	3.7565	3.7107	3.6576
	Std.dev.	12.0486	11.1026	10.2291	9.2389	8.5481	7.9098
	Min	0.0447	0.0752	0.1000	0.1800	0.2117	0.2516
	Median	0.6347	0.7921	0.8928	1.1566	1.2384	1.2909
	Max	134.238	129.014	123.176	117.764	114.133	108.237
	Skew	7.4621	7.1876	7.0362	7.0393	7.1650	7.2775

Table 3
Basic statistics of the selected data set of the BB to C rating groups. The spreads are represented as a percentage (100 bp).

Rating	Statistics	Maturity					
		1Y	2Y	3Y	5Y	7Y	10Y
BB	Mean	2.1716	2.4243	2.6317	2.9900	3.0781	3.1377
	Std.dev.	3.8042	3.4495	3.1998	2.9326	2.7774	2.6447
	Min	0.0369	0.0658	0.0951	0.1730	0.2100	0.2773
	Median	0.7279	1.1917	1.5603	2.1566	2.4054	2.6094
	Max	58.6189	37.7039	34.7189	29.4581	26.3604	22.7197
	Skew	4.4169	3.3698	2.8427	2.1328	1.8674	1.6544
B	Mean	2.6014	3.3860	4.0913	5.0095	5.1440	5.2297
	Std.dev.	2.2829	2.6063	2.8358	2.9178	2.7293	2.5829
	Min	0.0950	0.3750	0.4500	0.7519	0.7537	0.8345
	Median	1.8617	2.7500	3.4869	4.6540	5.0072	5.2866
	Max	16.0751	18.0306	17.9076	18.3501	16.7750	15.2661
	Skew	1.5652	1.5299	1.3985	1.0524	0.8066	0.6073
C	Mean	10.95254	10.9582	10.8783	10.7015	10.2037	9.8036
	Std.dev.	33.4680	29.4195	25.8971	22.8022	20.5104	18.7787
	Min	0.0489	0.0857	0.1906	0.3719	0.5518	0.7164
	Median	2.8764	3.4114	3.9314	4.6025	4.6884	4.8191
	Max	393.177	429.800	402.680	397.940	267.801	239.643
	Skew	7.3444	7.3754	6.9731	6.9465	6.6013	6.5255

Since there are six values of CDS spreads, one for each maturity, the total dimension of input variables is 84 and the total dimension of target variables is 6. This structure of data object is displayed in Fig. 2. Then we divided the constructed data set into non-overlapping one-month subperiods. Since the period of the whole data set includes the contracts from January 2001 to February 2014, 158 subperiods were constructed in total for each firm.

For the prediction using nonparametric models, we used the roll-over strategy as follows. First, we constructed 157 sub-data sets, each of which contains two consecutive subperiods. Then, for each sub-data set, the data instances of the preceding subperiod were used to train the model and the model was applied to the following subperiod to find the prediction error. This roll-over strategy is briefly summarized in Fig. 3. Since the similar roll-over strategies have been used in several literatures that predicted financial variables by using machines learning models, more detailed description can be found in those literatures including Son et al. (2012) and Park et al. (2014). This roll-over strategy was also used to calibrate the parameters for the benchmark models. The parameters of the benchmark models were calibrated to show the smallest relative RMSE value in the prediction of the preceding subperiod of a sub-data set and applied to the following subperiod to measure the prediction performance. The obtained parameters can be different for each sub-data set.

The parameters and other settings of nonparametric models were basically the same with those of the parametric benchmark models explained above. For artificial neural networks and Bayesian neural networks, the sigmoid function was used as an activation function except for the last layer for which the linear function is used as an activation function. For support vector regression, the radial basis kernel, $K(\mathbf{x}, \mathbf{y}) = \exp(-\gamma \|\mathbf{x} - \mathbf{y}\|^2)$, was used for the basis kernel function. Gaussian process used constant mean and covariance functions with the Gaussian likelihood function for hyperparameters. The specific values of parameters and other variables were selected from the preceding, or train, subperiod, with an appropriate validation task, and applied to the following, or test, subperiod.

For measuring the performance, we took the average value of relative root-mean-squared error (RMSE) of the test subperiods of all sub-data sets. The relative RMSE for each sub-data set is

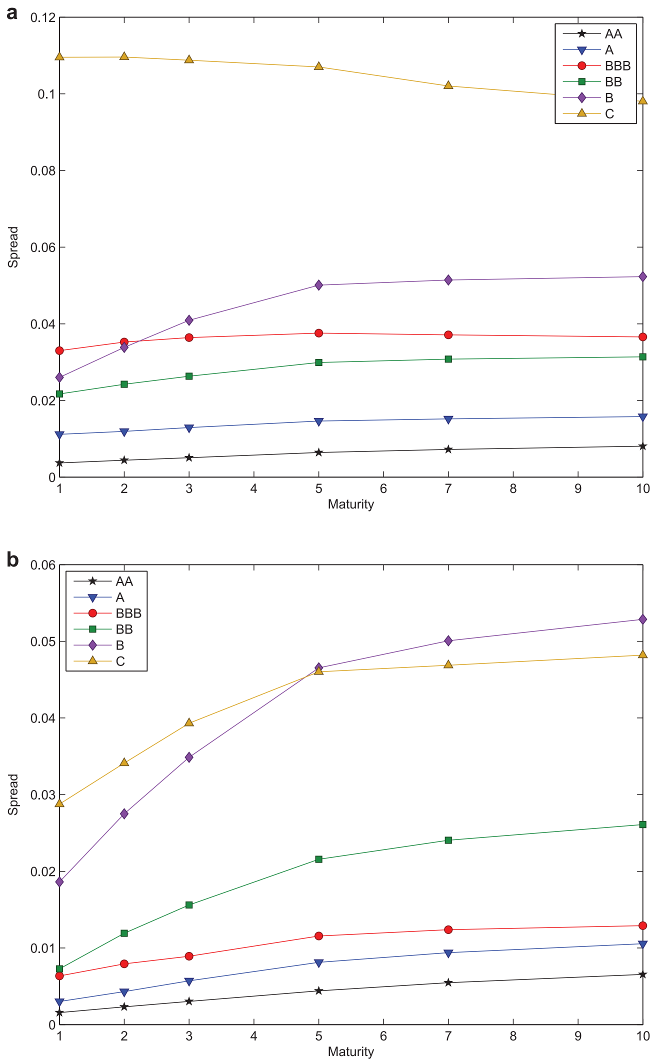

Fig. 1. The term structure of mean and median spreads for each rating group. (a) mean spreads (b) median spreads

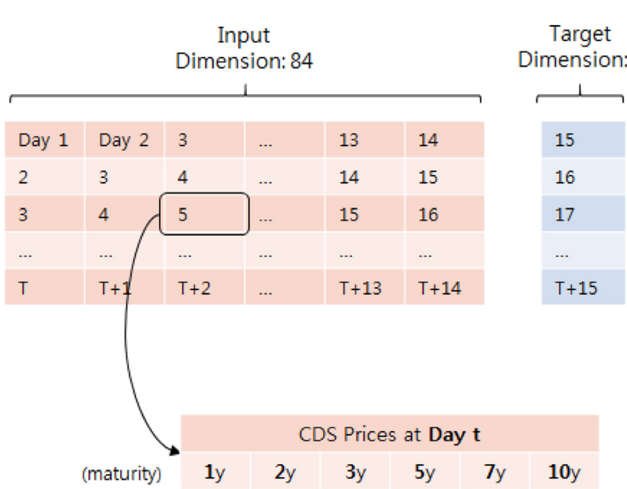

Fig. 2. The structure of data set. Each instance has 84 input variables and 6 target variables.

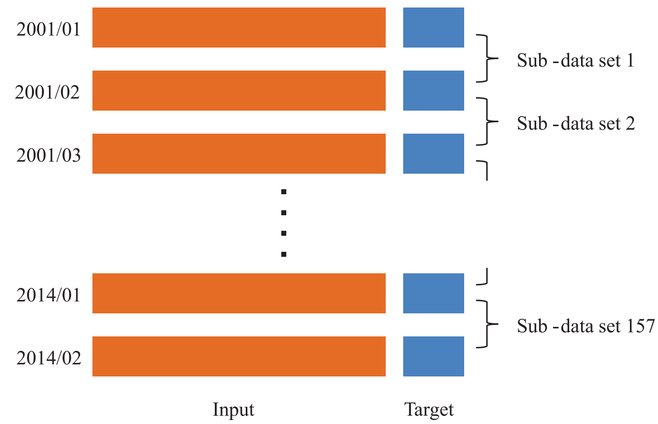

Fig. 3. Construction of the sub-data sets by using roll-over strategy.

computed as

$$\text{Rel. RMSE} = \left[\frac{1}{|D|} \sum_{t \in D} \left\{ \frac{1}{|M|} \sum_{m \in M} \left(\frac{s(m,t) - \hat{s}(m,t)}{s(m,t)} \right)^2 \right\} \right]^{\frac{1}{2}} \tag{24}$$

where D is the set of dates that the test subperiod of the sub-data set includes, M is the set of maturities, $s(m, t)$ is the actual spread value for the maturity m and date t, and $\hat{s}(m,t)$ is the predicted value. The weights for the different maturities and dates are set to be all the same.

5. Experimental results

In this section, we present the experimental results for which four nonparametric models and two parametric benchmark models are applied to prediction for the CDS spreads with six different maturities. Table 4 shows the predicted relative RMSE, averaged over the whole period, of employed models for each firm. The parameters of each model were chosen to show show the smallest relative RMSE in the validation error of the training data for each sub-data set and thus the parameters can be different for each sub-data set. The 10-fold cross-validation method was used as the validation method. The best result for each firm is bold-faced. ANN, BNN, SVR, and GP in the tables and figures refer to artificial neural networks, Bayesian neural networks, support vector regression, and Gaussian process regression respectively. Also, Constant and CIR refer to constant intensity model and Cox-Ingersoll-Ross model, respectively. It is observed that the nonparametric machine learning models perform better than the benchmark parametric models for most cases. Especially, the artificial neural network model showed the best and robust performances for all cases.

For more analysis, we averaged the prediction results for each implied rating group and this results are shown in Table 5 and Fig. 4. For the averaged results, artificial neural network model showed the best performances since it performed the best in all cases and Gaussian processes showed the better result than two parametric benchmark models in average. Support vector regression showed good performance compared to the parametric benchmark models, especially to the constant intensity model and this result contradicts the previous research of Gündüz and Uhrig-Homburg (2011) that only considered one value CDS spread with 5 year maturity. Especially, CIR model did not perform well for lower rating groups, BB, B, and C, while the results of the other models did not vary much for different implied rating groups, according to Fig. 4.

Additionally, to compare the predictive performance of the models in financial crisis, we applied the prediction models to the global financial crisis period from 2007 to 2009. The results are

Table 4
Averaged relative RMSE of predicted CDS spreads for each firm. The **boldface** represents the best result.

Implied rating	Ticker	ANN	BNN	SVR	GP	Constant	CIR
AA	AMGN	**0.1178**	0.1406	0.1658	0.1903	0.4983	0.1846
	CSX	**0.1065**	0.1258	0.1532	0.1793	0.4835	0.1842
	IBM	**0.1094**	0.1287	0.1471	0.1723	0.4841	0.1617
	JPM	**0.1161**	0.1346	0.1706	0.2019	0.4769	0.1971
	T	**0.1075**	0.1273	0.1393	0.1761	0.4604	0.1902
A	CAT	**0.1126**	0.1321	0.1534	0.1867	0.4783	0.1871
	IP	**0.0924**	0.1143	0.1312	0.1637	0.4585	0.1877
	AIG	**0.1301**	0.1523	0.1723	0.2829	0.4746	0.2223
	CU	**0.0867**	0.1047	0.1198	0.1945	0.3519	0.2526
	CNP	**0.0839**	0.0997	0.1195	0.1460	0.4837	0.1830
BBB	CA	**0.1181**	0.1404	0.1450	0.2071	0.3414	0.2104
	F	**0.1020**	0.1227	0.1378	0.1967	0.4027	0.2787
	MAY	**0.1020**	0.1272	0.1455	0.1652	0.3836	0.1514
	PH	**0.0883**	0.0071	0.1240	0.1355	0.4577	0.4445
	N	**0.1113**	0.1494	0.1653	0.1871	0.5392	0.2033
BB	KBH	**0.1096**	0.1290	0.1441	0.2013	0.3894	0.2096
	HRB	**0.1377**	0.1546	0.1714	0.2242	0.4188	0.4958
	SLMA	**0.0981**	0.1210	0.1336	0.1907	0.4588	0.2095
	DPL	**0.0951**	0.1132	0.1384	0.1721	0.4913	0.5283
	EP	**0.1080**	0.1316	0.1401	0.2001	0.3856	0.2294
B	AMKR	**0.1085**	0.1431	0.1477	0.2016	0.3691	0.4968
	INTEL	**0.1014**	0.1251	0.1398	0.1904	0.4648	0.5450
	PLCOAL	**0.0928**	0.1098	0.1349	0.1347	0.5154	0.1677
	THC	**0.0829**	0.1035	0.1063	0.1490	0.3986	0.2801
	FST	**0.0965**	0.1162	0.1513	0.1983	0.4529	0.2345
C	BOW	**0.0986**	0.1145	0.1283	0.1782	0.3391	0.2283
	DYN	**0.1282**	0.1539	0.1539	0.2469	0.3489	0.3388
	AMR	**0.1095**	0.1354	0.1357	0.2062	0.2771	0.3408
	VC	**0.1176**	0.1453	0.1600	0.2230	0.3153	0.3028
	JCP	**0.1123**	0.1334	0.1489	0.1968	0.4098	0.2421

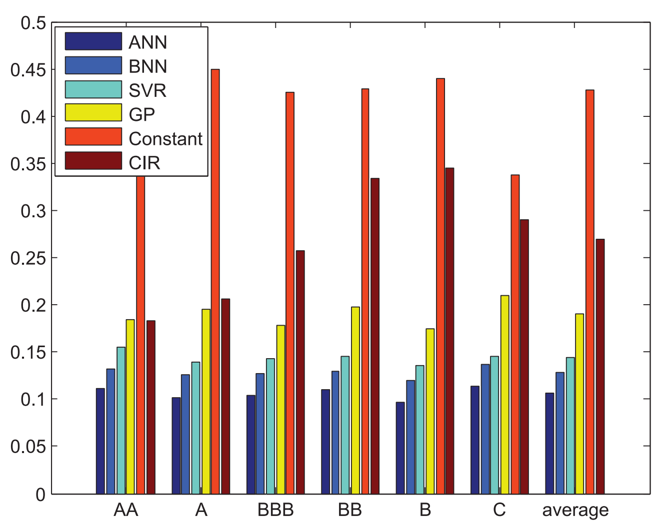

Fig. 4. Averaged relative RMSE of predicted CDS spreads for each implied rating group.

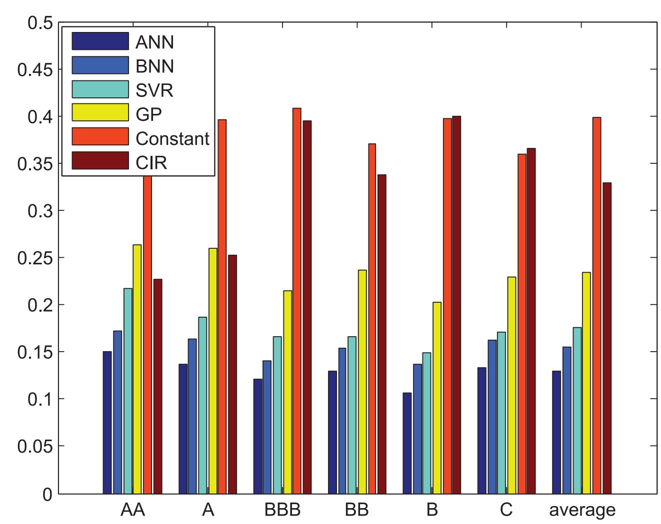

Fig. 5. Averaged relative RMSE of predicted CDS spreads for the global financial crisis period.

Table 5
Averaged relative RMSE of predicted CDS spreads for each implied rating group. The **boldface** represents the best result.

Implied rating	ANN	BNN	SVR	GP	Constant	CIR
AA	**0.1114**	0.1314	0.1552	0.1840	0.4807	0.1836
A	**0.1011**	0.1260	0.1393	0.1948	0.4494	0.2065
BBB	**0.1044**	0.1274	0.1435	0.1783	0.4249	0.2577
BB	**0.1097**	0.1299	0.1455	0.1977	0.4288	0.3345
B	**0.0964**	0.1195	0.1360	0.1748	0.4402	0.3448
C	**0.1133**	0.1365	0.1454	0.2102	0.3381	0.2905
Average	**0.1061**	0.1285	0.1442	0.1900	0.4270	0.2696

shown in Fig. 5. It can be easily observed that the relative RMSE of this period is higher than the average of the whole period for all models except the constant intensity model. Especially, the results of CIR model for the rating group from BBB to C did not perform well. In the crisis period, CDS spreads of most firms increased sharply thus even the high grade firms might have large spread values than the low grade firms of the moderate period. Therefore, CIR which showed bad performances for low grade firms in the whole period did not forecast well even for BBB grade firms in the crisis period. Therefore, the nonparametric machine learning models can be useful compared to the parametric benchmark models in predicting the credit risks of low-rated firms or in making forecasts during a crisis when the credit risks of most firms are large.

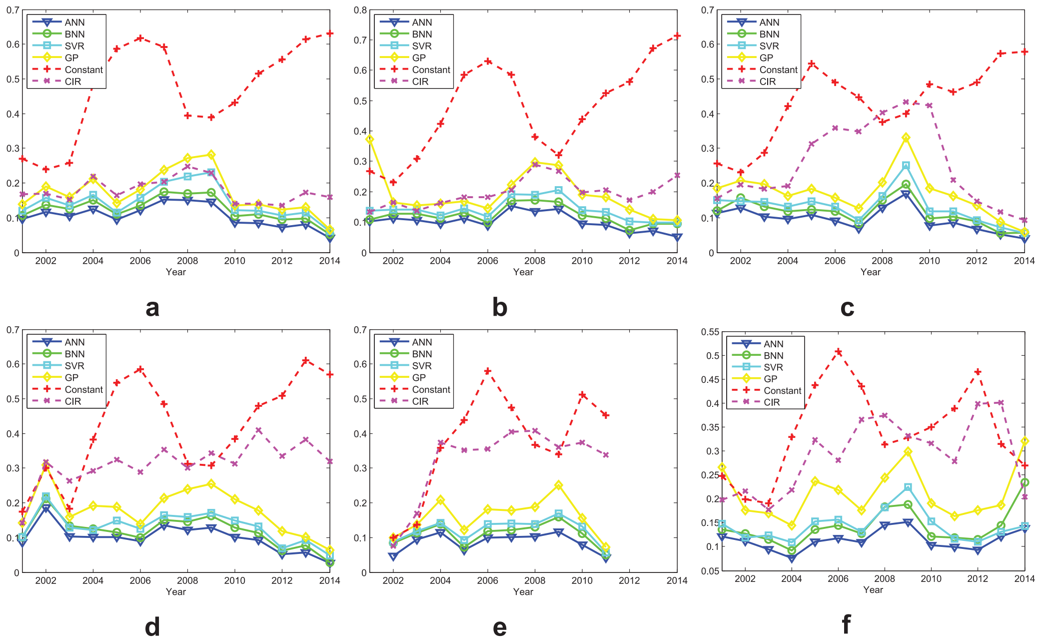

Fig. 6. Time series plot of annually averaged relative RMSE of each rating group of (a) AA, (b) A, (c) BBB, (d) BB, (e) B, and (f) C.

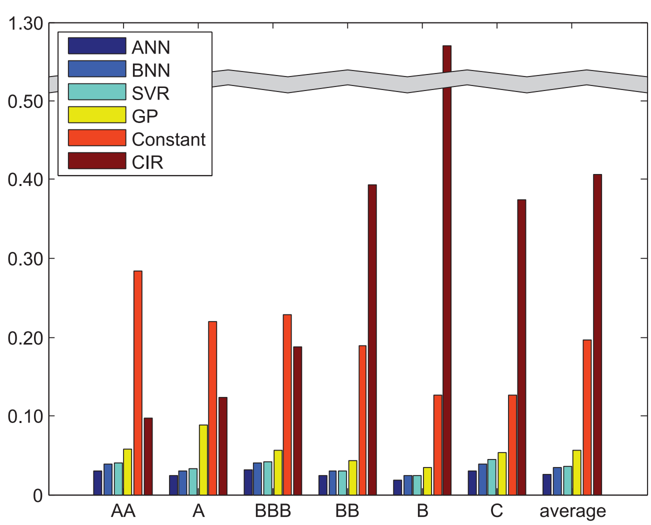

Fig. 7. Averaged relative RMSE of the ratio between 10 year maturity spread and 5 year maturity spread.

Fig. 6 showed the time series plot of annually averaged relative RMSE of each rating group. It can also be observed that the machine learning models except Gaussian processes usually performed better than the two parametric benchmarks. In addition, the relative RMSEs in the crisis period, 2007 to 2009, are usually larger than the other years for all prediction models except the constant intensity model and this feature appears clearly for the nonparametric models in all rating groups but does not occur clearly for CIR model in low rating groups.

Finally, we measured the performance of the model in the relative RMSE of the ratio between 10 year maturity spread and 5 year one instead of the absolute value of the credit spreads themselves and the results are shown in Fig. 7. This measure is meaningful because the CDS contract is hedged by the CDS contract with different maturities. It can be noticed that the performance differences between the machine learning models and the parametric benchmarks are larger than those measured in the usual RMSE of the credit spread values. Therefore, the machine learning nonparametric models can be applied to the prediction of CDS spreads for the practical hedging purpose as well.

6. Conclusion and future work

In this study, we conducted an empirical study on the predictive performance of nonparametric machine learning models for the CDS spread prediction. This study has several features. First, four state-of-arts nonparametric machine learning models including artificial neural networks, Bayesian neural networks, support vector regression, and Gaussian process regression, have been compared to verify their performances to predict CDS spreads with two benchmark parametric models whereas the previous studies were usually focused only on parametric models or additional one nonparametric model. Second, the data set used in this study is very extensive to range daily contract data from January 2001 to February 2014 including the global financial crisis period. Third, CDS spreads of various rating of firms, from AA to C, were used for the prediction whereas most of the earlier studies focused on AA or A ratings. Finally, the prediction for the CDS spreads with different maturities were conducted simultaneously not just the spread of the one maturity since it is very important to analyze the close relationship among the spreads of different maturities at a specific time for any firms for market users.

As a result of this study, nonparametric machine learning models possibly except the Gaussian processes outperformed than the benchmark parametric models for every maturity and implied rating group. Especially, artificial neural networks showed the best performances in every case, whereas Gaussian processes performed worse than the CIR benchmark model in some cases. This characteristic was also shown in the analysis on the global financial crisis period for high rated firms as well as for low rated firms where both of the parametric models such as the CIR and the constant intensity models showed worse performance than the compared machine learning models. Especially, in the global financial crisis period, the relative RMSEs were higher than those in the moderate period for most of the models used in this study except the constant intensity model which did not perform well in the moderate period, neither. The parametric and the machine learning models were also compared in the relative RMSEs of the ratio between 10 year maturity CDS spreads and 5 year one to find which models can be better for the practical hedging purpose where the machine learning models showed better performance than the benchmarks in this measure as well. This result shows the impact and significance of the machine learning methods on expert and intelligent systems that help the market practitioners to analyze and predict the price of CDS contracts and hedge a CDS contract with different maturities.

This study still has possibilities to be improved on some points. Technically, the sophisticated input variables traditionally exploited for the time series prediction like moving averages can also be used for the CDS prediction in this study due to the versatility of the machine learning models in managing the number and kinds of input variables. The other developed parametric models can also be used to the same data set to compare the predictive powers with the nonparametric models. As a firm specific view, in addition, conducting the prediction of CDS spreads when a special event on the credit risk of a certain firm occurs using nonparametric models can be one of extensions of this study. In this study, thirty firms including five firms from each rating group were used since there were not many firms with enough records of the CDS spreads in a longer period as of 15 years. For a shorter period, like one year, where a large number of firms are available, our analysis using nonparametric machine learning models may also be easily extended, which needs to be further investigated.

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MEST) (No. 2011-0017657).

References

Bakshi, G., Madan, D., & Zhang, F. X. (2006). Investigating the role of systematic and firm-specific factors in default risk: Lessons from empirically evaluating credit risk models. *The Journal of Business, 79*(4), 1955–1987.

Bhattacharyya, S., Pictet, O. V., & Zumbach, G. (2002). Knowledge-intensive genetic discovery in foreign exchange markets. *Evolutionary Computation, IEEE Transactions on, 6*(2), 169–181.

Bianchi, M. L., & Fabozzi, F. J. (2015). Investigating the performance of non-gaussian stochastic intensity models in the calibration of credit default swap spreads. *Computational Economics, 46*(2), 243–273.

Bijlsma, M. J., Lukkezen, J., & Marinova, K. H. (2014). Measuring too-big-to-fail funding advantages from small banks CDS spreads. Available at SSRN. TILEC Discussion Paper No. 2014-012.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer.

Black, F., & Cox, J. C. (1976). Valuing coporate securities: Some effects of bond indenture provision. *The Journal of Finance, 31*(2), 351–367.

Black, F., & Scholes, M. (1973). The pricing of options and corporate liabilities. *The Journal of Political Economy, 81*(3), 637–654.

Brigo, D., & El-Bachir, N. (2010). An exact formula for default swaptions' pricing in the SSRJD stochastic intensity model. *Mathematical Finance, 20*(2), 365–382.

Brigo, D., & Mercurio, F. (2007). *Interest rate models - Theory and practice: With smile, inflation and credit* (2nd ed.). Berlin, New York: Springer.

Bühler, W., & Trapp, M. (2009). Time-varying credit risk and liquidity premia in bond and CDS markets. CFR working paper 09–13.

Cao, L. J., & Tay, F. E. H. (2003). Support vector machine with adaptive parameters in financial time series forecasting. *Neural Networks, IEEE Transactions on, 14*(6), 1506–1518.

Chang, C. C., & Lin, C. J. (2001). LIBSVM: A library for support vector machines. Software available at https://www.csie.ntu.edu.tw/~cjlin/libsvm/.

Chen, W. H., Shih, J. Y., & Wu, S. (2006). Comparison of support-vector machines and back propagation neural networks in forecasting the six major asian stock markets. *International Journal of Electronic Finance, 1*(1), 49–67.

Cont, R., & Kan, Y. H. G. (2011). Statistical modeling of credit default swap portfolios. Available at SSRN 1771862. Working paper.

Cox, D. R. (1955). Some statistical methods connected with series of events. *Journal of the Royal Statistical Society. Series B (Methodological)*, 129–164.

Cox, J. C., Ingersoll, J. E., Jr, & Ross, S. A. (1985). A theory of the term structure of interest rates. *Econometrica: Journal of the Econometric Society*, 385–407.

Cressie, N. (1993). Statistics for spatial data. In *Wiley series in probability and statistics*. Princeton, New Jersey: John Wiley & Sons.

Dan Foresee, F., & Hagan, M. T. (June 1997). Gauss-newton approximation to bayesian learning. In *Neural networks, 1997., international conference on: 3* (pp. 1930–1935). IEEE.

Drucker, H., Burges, C. J., Kaufman, L., Smola, A., & Vapnik, V. (1997). Support vector regression machines. *Advances in neural information processing systems, 9*, 155–161.

Duffee, G. R. (1999). Estimating the price of default risk. *Review of Financial Studies, 12*(1), 197–226.

Duffie, D., & Singleton, K. J. (2003). *Credit risk: pricing, measurement, and management*. Princeton, New Jersey: Princeton University Press.

Eom, Y. H., Helwege, J., & Huang, J. Z. (2004). Structural models of corporate bond pricing: An empirical analysis. *Review of Financial studies, 17*(2), 499–544.

Finger, C., Finkelstein, V., Lardy, J. P., Pan, G., Ta, T., & Tierney, J. (2002). Creditgrades technical document. *RiskMetrics Group*, 1–51.

Frühwirth, M., & Sögner, L. (2006). The jarrow/turnbull default risk model?evidence from the german market. *The European Journal of Finance, 12*(2), 107–135.

Galil, K., Shapir, O. M., Amiram, D., & Ben-Zion, U. (2014). The determinants of CDS spreads. *Journal of Banking & Finance, 41*, 271–282.

Gündüz, Y., & Uhrig-Homburg, M. (2011). Predicting credit default swap prices with financial and pure data-driven approaches. *Quantitative Finance, 11*(12), 1709–1727.

Han, G. S., & Lee, J. (2008). Prediction of pricing and hedging errors for equity linked warrants with gaussian process models. *Expert Systems with Applications, 35*(1), 515–523.

Huang, Z., Chen, H., Hsu, C. J., Chen, W. H., & Wu, S. (2004). Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decision support systems, 37*(4), 543–558.

Hutchinson, J. M., Lo, A. W., & Poggio, T. (1994). A nonparametric approach to pricing and hedging derivative securities via learning networks. *The Journal of Finance, 49*(3), 851–889.

Jarrow, R., Li, H., & Ye, X. (2011). *Exploring statistical arbitrage opportunities in the term structure of CDS spreads*. RMI Working Paper.

Jarrow, R. A., & Turnbull, S. M. (1995). Pricing derivatives on financial securities subject to credit risk. *The Journal of Finance, 50*(1), 53–85.

Jones, E. P., Mason, S. P., & Rosenfeld, E. (1984). Contingent claims analysis of corporate capital structures: An empirical investigation. *The Journal of Finance, 39*(3), 611–625.

Kim, K. J., & Ahn, H. (2012). A corporate credit rating model using multi-class support vector machines with an ordinal pairwise partitioning approach. *Computers & Operations Research, 39*(8), 1800–1811.

Kim, S. H., & Noh, H. J. (1997). Predictability of interest rates using data mining tools: a comparative analysis of korea and the US. *Expert Systems with Applications, 13*(2), 85–95.

Lando, D. (2004). *Credit risk modeling: Theory and applications*. Princeton, New Jersey: Princeton University Press.

Lee, Y. C. (2007). Application of support vector machines to corporate credit rating prediction. *Expert Systems with Applications, 33*(1), 67–74.

Liao, S. H., & Chou, S. Y. (2013). Data mining investigation of co-movements on the taiwan and china stock markets for future investment portfolio. *Expert Systems with Applications, 40*(5), 1542–1554.

Lyden, S., & Saraniti, D. (2001). An empirical examination of the classical theory of corporate security valuation. Available at SSRN 271719. Working paper.

MacKay, D. J. C. (1992). Bayesian interpolation. *Neural Computation, 4*, 415–447.

Mayordomo, S., Peña, J. I., & Schwartz, E. S. (2014). Are all credit default swap databases equal? *European Financial Management, 20*(4), 677–713.

Merton, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. *The Journal of Finance, 29*(2), 449–470.

Ogden, J. P. (1987). Determinants of the ratings and yields on corporate bonds: Tests of the contingent claims model. *Journal of Financial Research, 10*(4), 329–340.

Osuna, E., Freund, R., & Girosi, F. (September 1997). An improved training algorithm for support vector machines. In *Neural networks for signal processing [1997]VII. proceedings of the 1997 IEEE workshop* (pp. 276–285). IEEE.

Park, H., Kim, N., & Lee, J. (2014). Parametric models and non-parametric machine learning models for predicting option prices: Empirical comparison study over KOSPI 200 index options. *Expert Systems with Applications, 41*(11), 5227–5237.

Park, H., & Lee, J. (2012). Forecasting nonnegative option price distributions using bayesian kernel methods. *Expert Systems with Applications, 39*(18), 13243–13252.

Pérez-Cruz, F., Camps-Valls, G., Soria-Olivas, E., Pérez-Ruixo, J. J., Figueiras-Vidal, A. R., & Artes-Rodriguez, A. (2002). Multi-dimensional function approximation and regression estimation. In *Artificial neural networks-ICANN 2002* (pp. 757–762). Berlin Heidelberg: Springer.

Rasmussen, C. E. (1996). *Evaluation of Gaussian processes and other methods for non–linear regression (Doctoral dissertation)*. University of Toronto.

Rosenblatt, F. (1962). *Principles of Neurodynamics: Perceptrons and the theory of brain mechanisms*. Washington DC: Spartan.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In *Parallel distributed processing: Explorations in the microstructure of cognition: vol. 1* (pp. 318–362). Cambridge, Massachusetts: MIT Press.

Sánchez-Fernández, M., de Prado-Cumplido, M., Arenas-Garcia, J., & Pérez-Cruz, F. (2004). SVM multiregression for nonlinear channel estimation in multiple-input multiple-output systems. *Signal Processing, IEEE Transactions on, 52*(8), 2298–2307.

Schoutens, W., & Cariboni, J. (2009). *Lèvy processes in credit risk*. Chichester, United Kingdom: Wiley.

Son, Y., Noh, D. J., & Lee, J. (2012). Forecasting trends of high-frequency KOSPI200 index data using learning classifiers. *Expert Systems with Applications, 39*(14), 11607–11615.

Ticknor, J. L. (2013). A bayesian regularized artificial neural network for stock market forecasting. *Expert Systems with Applications, 40*(14), 5501–5506.

Tuia, D., Verrelst, J., Alonso, L., Pérez-Cruz, F., & Camps-Valls, G. (2011). Multioutput support vector regression for remote sensing biophysical parameter estimation. *Geoscience and Remote Sensing Letters, IEEE, 8*(4), 804–808.

Vasicek, O. (1977). An equilibrium characterization of the term structure. *Journal of financial economics, 5*(2), 177–188.

Williams, C. K. I., & Rasmussen, C. E. (2006). *Gaussian processes for machine learning*. MIT Press.

Yang, S. H., & Lee, J. (2011). Predicting a distribution of implied volatilities for option pricing. *Expert Systems with Applications, 38*(3), 1702–1708.