Article

Kernel Ridge Regression Model Based on Beta-Noise and Its Application in Short-Term Wind Speed Forecasting

Shiguang Zhang [1], Ting Zhou [2], Lin Sun [1,*] and Chao Liu [1]

[1] College of Computer and Information Engineering, Henan Normal University, Xinxiang 453007, China;
     sg201106@163.com (S.Z.); l1040062050@163.com (C.L.)
[2] The State-Owned Assets Management Office, Henan Normal University, Xinxiang 453007, China;
     zhouting7606@163.com
*   Correspondence: linsunok@gmail.com

check for
updates

Abstract: The Kernel ridge regression (*KRR*) model aims to find the hidden nonlinear structure in raw data. It makes an assumption that the noise in data satisfies the Gaussian model. However, it was pointed out that the noise in wind speed/power forecasting obeys the Beta distribution. The classic regression techniques are not applicable to this case. Hence, we derive the empirical risk loss about the Beta distribution and propose a technique of the kernel ridge regression model based on the Beta-noise (*BN-KRR*). The numerical experiments are carried out on real-world data. The results indicate that the proposed technique obtains good performance on short-term wind speed forecasting.

Keywords: ridge regression model; kernel function; Beta-noise empirical risk loss; wind speed forecasting

1. Introduction

Linear regression (*LR*) is an approach to using the least squares method to model the relationship between a scalar dependent variable and one or more explanatory variables. It also refers to the plane points that are fitted with a straight line or the points in a high dimension space that are fitted with a hyperplane. This method is very sensitive to predictors being in a configuration of near-collinearity. Ridge regression (*RR*) is a variant of linear regression whose goal is to circumvent the problem of predictors collinearity. The ridge regression model is a powerful technique of machine learning which was introduced by Hoerl [1] and Hastie et al. [2], and it is a method from classical statistics that implements a regularized form of least squares regression [3]. Ridge regression is an alternative method for learning function based on a regularized extension of least squares techniques [4].

Given the data-set

$$D_N = (x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N) \tag{1}$$

where $x_i \in X = R^n$, $y_i \in R$, $i = 1, \ldots, N$ is the data-set. A multiple *LR* is $f(x) = \omega^T \cdot x + b$. R represents real number set, R^n is n dimensional Euclidean space, N is the number of sample points, superscript T denotes the matrix transpose. *LR* and *RR* determine the parameter vector $\omega \in R^n$ by minimizing the objective functions, respectively:

$$g_{LR} = \sum_{i=1}^{N} (y_i - \omega^T \cdot x - b)^2, \tag{2}$$

$$g_{RR} = \frac{1}{2} \cdot \omega^T \cdot \omega + C \cdot \sum_{i=1}^{N} (y_i - \omega^T \cdot x - b)^2. \tag{3}$$

The objective function used in ridge regression implements a form of Tikhonov [5] regularization of a sum-of-squares error metric, which is a regularization parameter controlling the bias-variance trade-off [6]. This corresponds to penalized maximum likelihood estimation of ϖ, assuming the targets have been corrupted by independent identical probability distribution (i.i.d.) samples from a Gaussian noise process with zeros mean and variance σ^2, i.e., $y_i = f(x_i) + \xi_i, \xi \in N(0, \sigma^2), i = 1, \ldots, N$.

The *KRR* model based on Gaussian-noise characteristic is derived by Saunders et al [7]. *RR* [1,3,5] aims to find the hidden nonlinear structure in the raw data, while nonlinear mapping is approximated by means of *KRR* based on kerneltechniques [7–11]. Therefore a linear *RR* model is constructed in a feature-space H ($\Phi : R^n \longrightarrow H$), induced by a nonlinear kernel function defining the inner product $K(x_i, x_j) = (\Phi(x_i) \cdot \Phi(x_j))$ $(i, j = 1, \ldots, N)$. The kernel function $\Phi : R^n \to H$ may be any positive definite Mercer kernel. Therefore, the objective function of *KRR* based on Gaussian-noise (*GN-KRR*) minimization can be written as

$$g_{GN\text{-}KRR} = \frac{1}{2} \cdot \varpi^T \cdot \varpi + C \cdot \sum_{i=1}^{N} (y_i - \varpi^T \cdot \Phi(x) - b)^2. \tag{4}$$

Suppose the noise is Gaussian, the $GN - KRR$ model may meet the requirements. However, the noise in wind speed and wind power forecast does not obey the Gaussian distribution, but the Beta distribution. The classic regression techniques are not applicable to above case. The uncertainty of wind power predictions was investigated in [12]. The statistics of the wind power forecasting error were not Gaussian. The work in [13] also found that the output of wind turbine systems is limited between zero and the maximum power and the error statistics do not follow a normal distribution. It also proved that using the Beta-function is justifiable for wind power prediction about chi-squared tests. In [14], the standard deviation of the data set was a function of the normalized predicted power $p = p_{pred}/p_{inst}$, where p_{pred} is the predicted power and p_{inst} is the wind power installed capacity. Fabbri [14] pointed out standardized production power p be within the interval $[0, 1]$ and Beta-function are more suitable than standard normal distribution. Literature [15] exhibited the advantages of using Beta-probability distribution function (pdf) instead of Gaussian pdf for approximating the forecasting error. Based on the above literature [12–16], this work plans to study the error of Beta-distribution between the predicted values x_p and the measured values x_m in the wind speed forecasting, and pdf of ε_i is $f(\varepsilon_i) = \varepsilon_i^{u-1} \cdot (1 - \varepsilon_i)^{v-1} \cdot h, \varepsilon_i \in (0, 1), i = 1, 2, \cdots, N$, plotted below in Figure 1. Where u, v are parameters, h is normalization-factor, and the parameters u, v may be determined by the given values of mean and standard deviation [17].
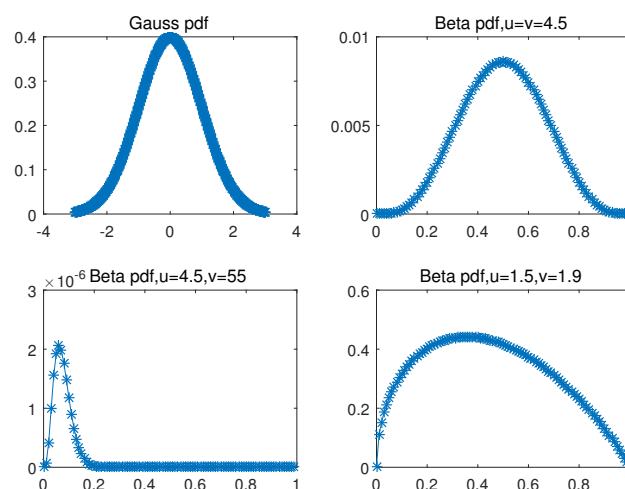

Figure 1. Beta pdf and Gauss pdf.

It is not suitable to apply the *KRR* techniques based on Gaussian-noise model (*GN-KRR*) to fit functions from data-set with Beta-noise. In order to solve the above problem, this work focuses on the utilization of optimization theory and Beta-noise loss function and derives a method of *KRR* based on Beta-noise characteristic (*BN-KRR*). It also introduces a forecasting technique that can deal with high-dimensionality and nonlinearity simultaneously.

This paper is organized as follows. In Section 2, we will derive the Beta-noise empirical risk loss by the Bayesian principle. Section 3 describes the proposed *KRR* model based on the Beta-noise. Section 4 gives the solution and algorithm design of *KRR* of the Beta-noise characteristic based on Genetic Algorithm. The numerical experiments are carried out on *BN-KRR* to short-term wind speed and wind power prediction in Section 5, respectively. Finally, the conclusions and future work are given in Section 6.

2. Bayesian Principle to Beta-Noise Empirical Risk Loss

Learning to fit data with noise is an important problem in many real-world data mining applications. Given a training set D_N of (1) with noise is additive

$$y_i = f(x_i) + \xi_i, i = 1, \cdots, N \tag{5}$$

where ξ_i is random i.i.d. $P(\xi_i)$ with standard deviation σ and mean μ.

The objective is to find regressor f minimizing the expected risk [18,19] $R[f] = \int l(x, y, f(x))dP(x, y)$ based on the empirical data D_l, where $l(x, y)$ is a empirical risk loss (determining how we will penalize estimation errors). Since we do not know the distribution $P(x, y)$, it can only use data-set D_N to estimate a regressor f and minimize $R[f]$. A possible approximation consists of replacing the integration by the empirical estimate to get the empirical risk $R_{emp}[f] = \frac{1}{N} \cdot \sum_{i=1}^{N} l(x, y_i, f(x_i))$. In general, we should add a capacity control term in *RR* and *KRR*, which leads to the regularized risk functional [18,20]

$$R_{reg}[f] = R_{emp}[f] + \frac{\lambda}{2} \cdot \|\varpi\|^2, \tag{6}$$

where $\lambda > 0$ is a regularization constant, $R_{emp}[f]$ is the empirical risk. It is well known that $l(\xi_i) = \frac{1}{2} \cdot \xi_i^2$ is the empirical risk loss of Gaussian-noise characteristics for *LR* (2), *RR* (3), and *KRR* (4). However, what is the empirical risk loss about Beta-noise of *KRR* model? The Beta-noise empirical risk loss by the use of Bayesian principle is given as follows.

The regressor $f(x)$ is unknown, the objection is to estimate the regressor $f(x)$ from $g \in D_N$. According to the literature [20–22], the optimal empirical risk loss from maximum likelihood be

$$l(x, y, f(x)) = -logp(y - f(x)). \tag{7}$$

The maximum likelihood estimation be

$$X_f = \{(x_1, f(x_1)), (x_2, f(x_2)), \ldots, (x_N, f(x_N))\}, \tag{8}$$

$$p(X_f|X) = \prod_{i=1}^{N} p(f(x_i)|(x_i, y_i)) = \prod_{i=1}^{N} p(y_i - f(x_i)). \tag{9}$$

Maximizing $p(X_f|X)$ is equivalent to minimizing $-log(p(X_f|X))$. Using Equation (7), we have

$$l(x, y, f(x)) = -logp(X_f|X). \tag{10}$$

Suppose noise in Equation (5) adheres to Beta distribution with mean $\mu \in (0,1)$ and variance σ^2, thus we can get $u = (1-\mu) \cdot \mu^2/\sigma^2 - \mu$, $v = (1-\mu)/\mu \cdot u$ [13,14], where $h = \Gamma(u+v)/\Gamma(u) \cdot \Gamma(v)$ is the normalization-factor. By Equation (10), the Beta-noise empirical risk loss is

$$l(\xi) = l(y - f(x)) = (1-u)log(\xi) + (1-v)log(1-\xi). \tag{11}$$

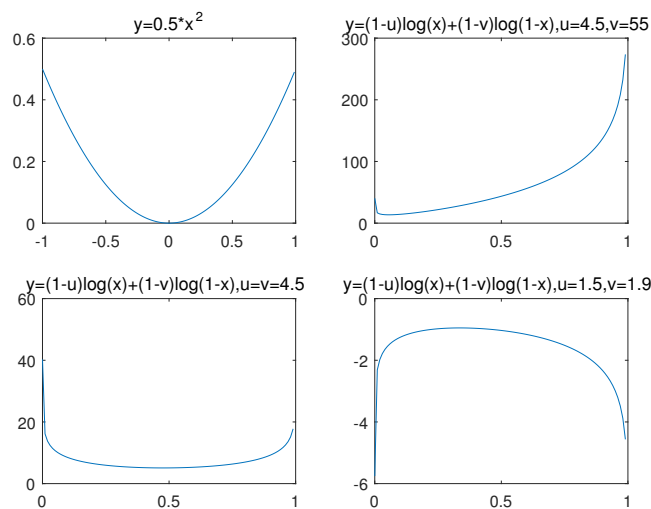Empirical risk loss of Gauss-noise and Beta-noise with different parameters is shown in Figure 2.

Figure 2. Empirical risk loss of Gauss-noise and Beta-noise.

3. *KRR* Model Based on Beta-Noise

It is not appropriate to apply the *KRR* model based on Gaussian-noise characteristic (*GN-KRR*) to deal with tasks with the Beta-noise distribution. Consequently, we use Beta-noise loss function and maximum likelihood method to estimate the optimal loss function. Now, we derive the optimal empirical risk loss about Beta-noise distribution, and propose a new technique of the *KRR* model based on Beta-noise characteristic (*BN-KRR*).

First, considering constructing *LR* regressor $f(x) = \omega^T \cdot x_i + b$, where $x_i = (1, x_{i1}, \ldots, x_{in})^T$ ($i = 1, 2, \cdots, N$), $\omega = (\omega_0, \omega_1, \cdots, \omega_n)^T$. We use kernel techniques and construct the kernel function $K(\bullet, \bullet)$ where $K(x_i, x_j) = (\Phi(x_i) \cdot \Phi(x_j))$, $\Phi : R^n \to H$, H is Hilbert space, and $(x_i \cdot x_j)$ ($i, j = 1, 2, \cdots, N$) is inner product of H. Then we extend kernel techniques to the ridge regression model based on the Beta-noise characteristic.

Let the set of inputs be $\{(x_i, y_i), i = 1, \cdots, N\}$, where i represents the indicator for the i-th sample in D_l. For the general Beta-noise characteristic, it is Formula (11) that the Beta-noise loss function $c(\xi_i)$ in the sample point $\{(x_i, y_i)\}$ of D_N. Owing to the fact that ridge regression and *KRR* techniques with Gaussian-noise characteristic ($GN - KRR$) are not suitable to Beta-noise distribution in time series problems, the Formula (11) is selected as Beta empirical risk loss to overcome the shortage of $GN-KRR$. The primal problem of *KRR* model with the Beta-noise (Denoted by $BN-KRR$) can be described as follows ($C > 0$)

$$min\{g_{P_{BN\text{-}KRR}} = \frac{1}{2}\omega^T \cdot \omega + C \cdot \sum_{i=1}^{N} c(\xi_i)\}$$

$$P_{BN\text{-}KRR} : s.t. \quad y_i - \omega^T \cdot \Phi(x_i) - b = \xi_i. \tag{12}$$

where $c(\xi_i) = (1-u)log(\xi_i) + (1-v)log(1-\xi_i)$, $\xi_i = y_i - f(x_i)i = 1, \cdots, N$.

Theorem 1. *Model BN-KRR's Solution to original Problem (12) about ω exists and is unique.*

Proof. The existence of solutions is trivial. The uniqueness of solutions is shown below. If we have solutions $\overline{\omega}, \widetilde{\omega}$, then Problem (12) exist solutions $(\overline{\omega}, \overline{b}, \overline{\xi})$ and $(\widetilde{\omega}, \widetilde{b}, \widetilde{\xi})$. Define (ω, b, ξ) as follows

$$\omega = \frac{1}{2}(\overline{\omega} + \widetilde{\omega}), b = \frac{1}{2}(\overline{b} + \widetilde{b}), \xi = \frac{1}{2}(\overline{\xi} + \widetilde{\xi}). \tag{13}$$

We have

$$y_i - \omega^T \cdot \Phi(x_i) - \xi_i = y_i - \frac{1}{2}(\overline{\omega} + \widetilde{\omega})^T \cdot \Phi(x_i) - \frac{1}{2}(\overline{b} + \widetilde{b}) - \frac{1}{2}(\overline{\xi} + \widetilde{\xi})$$

$$= \frac{1}{2}(y_i - \overline{\omega}^T \cdot \Phi(x_i) - \overline{b} - \overline{\xi}_i) + \frac{1}{2}(y_i - \widetilde{\omega}^T \cdot \Phi(x_i) - \widetilde{b} - \widetilde{\xi}_i) = 0,$$

where $\frac{1}{2}(\overline{\xi} + \widetilde{\xi}) \geq 0 (i = 1, \cdots, l)$, (ω, b, ξ) is a feasible solution of Problem (12). Further,

$$\frac{1}{2}\|\omega\|^2 + C \sum_{i=1}^{N} c(\xi_i) \geq \frac{1}{2}\|\overline{\omega}\|^2 + C \sum c(\widetilde{\xi}_i), \tag{14}$$

$$\frac{1}{2}\|\omega\|^2 + C \sum_{i=1}^{N} c(\xi_i) \geq \frac{1}{2}\|\widetilde{\omega}\|^2 + C \sum c(\widetilde{\xi}_i). \tag{15}$$

By Inequalities (14) and (15), we get $2\|\omega\|^2 \geq \|\overline{\omega}\|^2 + \|\widetilde{\omega}\|^2$. $\omega = \overline{\omega} + \widetilde{\omega}$ is substituted into the above inequality

$$\|\overline{\omega} + \widetilde{\omega}\|^2 \geq 2(\|\overline{\omega}\|^2 + \|\widetilde{\omega}\|^2), \tag{16}$$

as $\|\overline{\omega} + \widetilde{\omega}\| \leq \|\overline{\omega}\| + \|\widetilde{\omega}\|$, then

$$(\|\overline{\omega}\| + \|\widetilde{\omega}\|)^2 \geq 2(\|\overline{\omega}\|^2 + \|\widetilde{\omega}\|^2). \tag{17}$$

In addition, $2\|\overline{\omega}\| \cdot \|\widetilde{\omega}\| \leq \|\overline{\omega}\|^2 + \|\widetilde{\omega}\|^2$, by (17), get

$$2\|\overline{\omega}\| \cdot \|\widetilde{\omega}\| = \|\overline{\omega}\|^2 + \|\widetilde{\omega}\|^2, \|\overline{\omega}\| = \|\widetilde{\omega}\|, \|\overline{\omega} + \widetilde{\omega}\| = \|\overline{\omega}\| + \|\widetilde{\omega}\|.$$

For $\widetilde{\omega} = m \cdot \overline{\omega}$, thus $m = 1$ or $m = -1$. Since $m = -1$, then $\overline{\omega} + \widetilde{\omega} = 0$. By $\|\overline{\omega} + \widetilde{\omega}\| = \|\overline{\omega}\| + \|\widetilde{\omega}\|$, get $\|\overline{\omega}\| = \|\widetilde{\omega}\| = 0$. Namely, $\overline{\omega} = \widetilde{\omega} = 0$. For $m = 1$, thus $\overline{\omega} = \widetilde{\omega}$.

In conclusion, Solution to Problem (12) exists and is unique. $\square$

Theorem 2. *Model BN-KRR's dual Problem of primal Problem (12) is*

$$\begin{aligned} Max\{g_{D_{BN\text{-}KRR}} &= -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha_i \cdot \alpha_j \cdot K(x_i, x_j)) + \sum_{i=1}^{N}(\alpha_i \cdot y_i) \\ &+ C\sum_{i=1}^{N}((1-u)\log(\xi_i(\alpha_i)) + (1-v)\log(1-\xi_i(\alpha_i)))\} \\ D_{BN\text{-}KRR} &: s.t. \quad \sum_{i=1}^{l}(\alpha_i) = 0 \end{aligned} \tag{18}$$

where $\xi_i(\alpha_i) = \frac{2 + \alpha_i/C - u - v - \Delta^{\frac{1}{2}}}{2 \cdot \alpha_i/C}$, $\Delta = (\alpha_i/C + u - v)^2 + 4 \cdot (1 + u \cdot v - u - v)$ $(i = 1, \cdots, N)$ and $C > 0$ is constant.

Proof. The introduction of Lagrange functional $L(\omega, b, \alpha, \xi)$ is

$$L(\omega, b, \alpha, \xi) = \frac{1}{2}\omega^T \cdot \omega + C \cdot \sum_{i=1}^{N}(c(\xi_i)) + \sum_{i=1}^{N}\alpha_i(y_i - \omega^T \cdot \Phi(x_i) - b - \xi_i).$$

For the sake of the minimum $L(\omega, b, \alpha, \xi)$, seek partial derivative to ω, b, α, ξ respectively. From Karush-Kuhn-Tucker (KKT) conditions, obtain

$$\nabla_\omega(L) = 0, \nabla_b(L) = 0, \nabla_\xi(L) = 0.$$

So $\omega = \sum_{i=1}^N (\alpha_i \cdot \Phi(x_i)), \sum_{i=1}^N (\alpha_i) = 0, C \cdot \frac{\partial(c(\xi_i))}{\partial \xi_i} - \alpha_i = 0.$

Substituting the extreme conditions into $L(\omega, b, \alpha, \xi)$ and finding the Maximum of α, thus derive the dual Problem (18) of Problem (12). □

On account of $\frac{\partial(c(\xi_i))}{\partial(\xi_i)} = \frac{1-\alpha}{\xi_i} - \frac{1-\beta}{1-\xi_i}$, by $C \cdot \frac{\partial(c(\xi_i))}{\partial(\xi_i)} - \alpha_i = 0$, we have $\alpha_i/C \cdot \xi_i^2 - (2 + \alpha_i/C - \alpha - \beta) \cdot \xi_i + 1 - \alpha = 0.$

Now get

$$\xi_{i1}(\alpha_i) = \frac{2 + \alpha_i/C - \alpha - \beta + \Delta^{\frac{1}{2}}}{2 \cdot \alpha_i/C}, \xi_{i2}(\alpha_i) = \frac{2 + \alpha_i/C - \alpha - \beta - \Delta^{\frac{1}{2}}}{2 \cdot \alpha_i/C}.$$

where $\Delta = (\alpha_i/C + \alpha - \beta)^2 + 4 \cdot (1 + \alpha \cdot \beta - \alpha - \beta)$. Because of $0 < \xi_i(\alpha_i) < 1$, we reject $\xi_{i1}(\alpha_i)$ and let $\xi_i(\alpha_i) = \xi_{i2}(\alpha_i)$.

We have $\omega = \sum_{i=1}^N (\alpha_i \cdot \Phi(x_i)), b = y_i - \sum_{i=1}^N (\alpha_j \cdot K(x_j, x_i)) - \xi_i(\alpha_i)$, gain the decision-making function of *BN-KRR* is

$$f(x) = \omega^T \cdot \Phi(x) + b = \sum_{i=1}^N (\alpha_i \cdot K(x_i, x)) + b.$$

Note: The *KRR* of the Gaussian-noise characteristic ($GN - KRR$) was discussed in [9–11]. The Gaussian empirical risk loss in the sample point $(x_i, y_i) \in D_N$ is $c(\xi_i) = \frac{1}{2}\xi_i^2$, thus the dual Problem of model *KRR* based on Gaussian-noise characteristic ($GN - KRR$) is

$$max\{g_{D_{GN\text{-}KRR}} = -\frac{1}{2}\sum_{i=1}^N \sum_{j=1}^N (\alpha_i \cdot \alpha_j \cdot K(x_i, x_j)) + \sum_{i=1}^N (\alpha_i \cdot y_i) - \frac{1}{2C} \cdot \sum_{i=1}^N (\alpha_i^2)\} \tag{19}$$

$$D_{GN\text{-}KRR} : s.t. \quad \sum_{i=1}^N \alpha_i = 0.$$

The dual Problem of model *RR* based on the Gaussian-noise characteristic ($GN\text{-}RR$) is

$$max\{g_{D_{GN\text{-}RR}} = -\frac{1}{2}\sum_{i=1}^N \sum_{j=1}^N (\alpha_i \cdot \alpha_j \cdot (x_i, x_j)) + \sum_{i=1}^N (\alpha_i \cdot y_i) - \frac{1}{2C} \cdot \sum_{i=1}^N (\alpha_i^2)\} \tag{20}$$

$$D_{GN\text{-}RR} : s.t. \quad \sum_{i=1}^N \alpha_i = 0.$$

4. Solution Based on Genetic Algorithm

We get the Solution and algorithm design of model *KRR* based on Beta-noise characteristic (*BN-KRR*) as follows.

(1) Let training samples $D_N = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$, where $x_i \in X = R^n$, $y_i \in R$ $(i = 1, \dots, N)$.
(2) Select the appropriate positive C, u, v and the suitable kernel $K(\bullet, \bullet)$.
(3) Solve optimization Problem (18), gain optimal Solution $\alpha = (\alpha_1, \cdots, \alpha_N)$.
(4) Construct the decision-making function

$$f(x) = \omega^T \cdot \Phi(x) + b = \sum_{i=1}^N (\alpha_i \cdot K(x_i, x)) + b,$$

and $b = y_i - \sum_{i=1}^N (\alpha_j \cdot K(x_j, x_i)) - \xi_i(\alpha_i)$.

The confirmation of unknown parameters of model $BN - KRR$ is a complicated process and the appropriate parameter combination of the models can enhance the regression accuracy of the kernel ridge regression based on Beta-noise. Genetic Algorithm (GA) [23–25] is a search heuristic that mimics the process of natural evolution, this heuristic is routinely used to generate useful solutions to optimize and search problems. In GA, the evolution usually starts from a population of randomly generated individuals and happens in generations. In each generation, the fitness of every individual in the population is evaluated, multiple individuals are stochastically selected from the current population and modified to form a new population. The new population is then used in the next iteration of the algorithm. Commonly, the algorithm terminates when either a maximum number of generations has been produced, or a satisfactory fitness level has been reached for the population. If the algorithm has terminated due to a maximum number of generations, a satisfactory solution may or may not have been reached.

GA is considered as one of the modern optimization algorithms to solve the combinatorial optimization problem and is used to determine the parameters of model *BN-KRR*. Based on the survival and reproduction of the fitness, GA is continually applied to get new and better solutions without any pre-assumptions, such as continuity and unimodality [26–28]. The proposed model *BN-KRR* has been implemented in Matlab 7.8 programming language. The experiments are made on the 8.0 GHz Core (TM) i7-4790 CPU personal computer with 3.60 GB memory under Microsoft Windows XP Professional. The initial parameters of GA are $Max - cgen = 100$, $C \in [1, 201]$, $u, v \in (0, \infty)$. Many practical applications display that polynomial and Gaussian kernels perform well under general smooth assumptions [29]. This work, polynomial, and Gaussian kernels can be used as the kernel for models v-*SVR*, *GN-KRR*, and *BN-KRR*:

$$K(x_i, x_j) = ((x_i \cdot x_j) + 1)^d, K(x_i, x_j) = \exp(\frac{-\|x_i - x_j\|^2}{\sigma^2}),$$

where d is positive integer, and let $d = 1, 2,$ or 3. σ is positive, and take $\sigma = 0.2$.

As we all know, no prediction model forecasts perfectly. There are also certain criteria, such as mean absolute error (MAE), the root mean square error (RMSE), mean absolute percentage error (MAPE), and standard error of prediction (SEP) are used to evaluate the predictive performance of models v-*SVR*, *GN-KRR*, and *BN-KRR*. The four criteria are defined as follows:

$$MAE = \frac{1}{N} \cdot \sum_{i=1}^{N} |x_{p,i} - x_{m,i}|, \tag{21}$$

$$MAPE = \frac{1}{N} \cdot \sum_{i=1}^{N} \frac{|x_{p,i} - x_{m,i}|}{x_{m,i}}, \tag{22}$$

$$RMSE = \frac{1}{N} \cdot \sqrt{\sum_{i=1}^{N} (x_{p,i} - x_{m,i})^2}, \tag{23}$$

$$SEP = \frac{\sqrt{\sum_{i=1}^{N} (x_{p,i} - x_{m,i})^2}}{\sum_{i=1}^{N} x_{m,i}}, \tag{24}$$

where l is the size of the selected samples, m, i is the measured result of data-point x_i, and p, i is the predictive result of data-point x_i ($i = 1, 2, \cdots, N$) [14–16].

## 5. Short-Term Wind Speed and Wind Power Forecasting with Real Data-Set

The model *BN-KRR* is applied to the multi-factors actual data-set for wind speed sequence prediction from Jilin Province. The wind speed data contain more than a year of samples which are collected in intervals of ten minutes, and the number of wind speed data is 62,466. Each column

attribute is mean, variance, minimum, and maximum, respectively. The short-term wind speed forecast is studied as follows.

Suppose the training sample number is 2160 (from 1 to 2160 for 15 days), and the number of test samples is 720 (from 2161 to 2880 for 5 days). The input vector is $\overrightarrow{x_i} = (x_i, x_{i+1}, x_{i+2}, \cdots, x_{i+11})$, the output value is $x_{i+11+step}$, and $step = 1, 3$. Namely, the pattern above is used to forecast the wind speed each interval of 10 and 30 min at each Point x_{i+11}, respectively [30,31].

1. Forecast wind speed at point x_{i+11} each interval of 10 min

The short-term wind speed sequence forecast results at point x_{i+11} each interval of 10 min given by *GN-KRR* [7,8,32], *v-SVR* [33,34], and *BN-KRR* are illuminated with Figure 3. In *GN-KRR*, parameter $C = 151$. In *v-SVR*, parameter $C = 151, v = 0.54$. In *BN-KRR*, parameters $C = 181, u = 3.6084, v = 3.0889$.
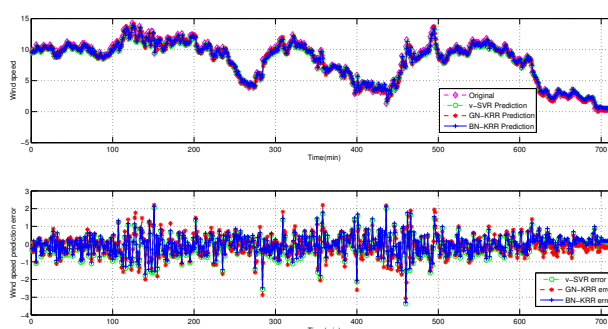
Figure 3. The forecasting results of *v-SVR*, *GN-KRR*, *BN-KRR* (step = 1).

MAE, MAPE, RMSE, and SEP indicators are used to evaluate the prediction results of the three models at point x_{i+11} each interval of 10 min shown in Table 1.
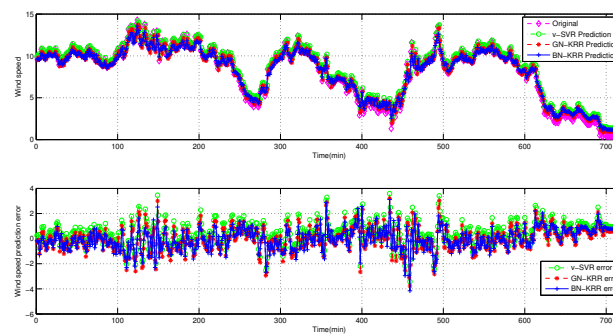
Table 1. Error statistic of three models (step = 1).

Model	MAE	RMSE	MAPE(%)	SEP(%)
v-SVR	0.4280	0.5833	7.02	7.02
GN-KRR	0.4219	0.5768	7.94	7.06
BN-KRR	0.3668	0.4233	6.84	5.23

2. Forecast wind speed at point x_{i+11} each interval of 30 min

The short-term wind power sequence forecast results at point x_{i+11} each interval of 30 min given by *GN-KRR*, *v-SVR*, and *BN-KRR* are illuminated with Figure 4. In $GN - KRR$, parameter $C = 151$. In *v-SVR*, parameter $C = 151, v = 0.54$. In *BN-KRR*, parameters $C = 181, u = 3.6084, v = 3.0889$.

Figure 4. The forecasting results of v-*SVR*, *GN-KRR*, *BN-KRR* (step = 3).

MAE, MAPE, RMSE, and SEP indicators are used to evaluate the prediction results of the three models at point x_{i+11} each interval of 30 min shown in Table 2.

Table 2. Error statistic of three models (step = 3).

Model	MAE	RMSE	MAPE(%)	SEP(%)
v-SVR	0.7979	1.0116	23.36	12.53
GN-KRR	0.7109	0.9226	17.17	11.43
BN-KRR	0.6640	0.8417	18.82	10.43

The results of wind speed forecasting experiments indicate that *BN-KRR* has better performance than *GN-KRR* and v-*SVR* in 10-min and 30-min short-term wind speed forecasting.

We have predicted the short-term wind speed from the Jilin Province wind farm, so we can calculate the wind power according to the Formula (25):

$$P_M = \begin{cases} 0, & v < v_{cut-in}, \quad \text{or} \quad v > v_{cut-out} \\ \frac{v - v_{cut-in}}{v_r - v_{cut-in}}, & v_{cut-in} \leq v < v_r \\ P_r, & v_r \leq v < v_{cut-in}. \end{cases} \tag{25}$$

where v_{cut-in} and $v_{cut-out}$ represent cut-in wind speed and cut-out wind speed of wind turbine, respectively. v_r and P_r represent rated wind speed and rated power of wind turbine, respectively. The predictive wind speed is substituted into the Formula (25), we can obtain the predicted wind power.

6. Conclusions and Future Work

In this work, we propose a new version of kernel ridge regression model based on the Beta-noise (*BN-KRR*) to predict the uncertainty system of Beta-noise. Novel results have been obtained by the use of the model *BN-KRR*, which takes the Bayesian principle to Beta-noise empirical risk loss and improves the prediction accuracy. The numerical experiments are carried out on real-world data (the short-term wind speed). Comparing the model *BN-KRR* and models *GN-KRR* and v-*SVR* by criteria *MAE*, *MAPE*, *RMSE*, and *SEP* verifies the validity and feasibility of our proposed model $BN - KRR$. Further, the forecasting results indicate that the proposed technique can obtain good performance on short-term wind speed forecasting.

In practical regression problems, data uncertainty is inevitable. The observed data are usually described in linguistic levels or ambiguous metrics, like the weather forecast, the forecast results of dry and wet, or sunny and cloudy, and so on. We should consider developing fuzzy kernel ridge regression algorithms with different noise models.

We verify the validity and feasibility of the model.

References

1. Hoerl, A.E. Application of ridge analysis to regression problems. *Chem. Eng. Prog.* **1962**, *58*, 54–59.
2. Hastie, T.; Tibshirani, R.; Friedman, J.H. *The Elements of Statistical Learning*; Springer: Berlin/Heidelberg, Germany, 2001.
3. Hoerl, A.E.; Kennard, R.W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **1970**, *12*, 55–67. [CrossRef]
4. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer: New York, NY, USA, 2009.
5. Tikhonov, A.A.; Arsenin, V.Y. *Solutions of Ill-Posed Problems*; Wiley: New York, NY, USA, 1977.
6. Geman, S.; Bienenstock, E.; Doursat, R. Neural networks and the bias/variance dilemma. *Neural Comput.* **1992**, *4*, 1–58. [CrossRef]
7. Saunders, C.; Gammerman, A.; Vovk, V. Ridge Regression Learning Algorithm in Dual Variables. In Proceedings of the 15th International Conference on Machine Learning, Madison, WI, USA, 24–27 July 1998; Morgan Kaufmann: San Francisco, CA, USA, 1998; pp. 515–521.
8. Suykens, J.A.K.; Lukas, L.; Vandewalle, J. Sparse Approximation using Least-squares Support Vector Machines. *IEEE Int. Symp. Circuits Syst. Geneva* **2000**, *2*, 757–760.
9. Gavin, C.C.; Nicola, L.C.T. Reduced rank kernel ridge regression. *Neural Process. Lett.* **2002**, *16*, 293–302.
10. Zhang, Z.H.; Dai, G.; Xu, C.F. Regularized Discriminant Analysis, Ridge Regression and Beyond. *J. Mach. Learn. Res.* **2010**, *11*, 2199–2228.
11. Orsenigo, C.; Vercellis, C. Kernel ridge regression for out-of-sample mapping in supervised manifold learning. *Expert Syst. Appl.* **2012**, *39*, 7757–7762. [CrossRef]
12. Lange, M. On the uncertainty of wind power predictions-Analysis of the forecast accuracy and statistical distribution of errors. *J. Sol. Energy Eng.* **2005**, *127*, 177–184. [CrossRef]
13. Bofinger, S.; Luig, A.; Beyer, H.G. Qualification of wind power forecasts. In Proceedings of the 2002 Global Wind Power Conference, Paris, France, 2 April 2002.
14. Fabbri, A.; Roman, T.G.S.; Abbad, J.R.; Quezada, V.H.M. Assessment of the cost associated with wind generation prediction errors in a liberalized electricity market. *IEEE Trans. Power Syst.* **2005**, *20*, 1440–1446. [CrossRef]
15. Bludszuweit, H.; Antonio, J.; Llombart, A. Statistical analysis of wind power forecast error. *IEEE Trans. Power Syst.* **2008**, *23*, 983–991. [CrossRef]
16. Madhiarasan, M.; Deepa, S.N. A novel criterion to select hidden neuron numbers in improved back propagation networks for wind speed forecasting. *Appl. Intell.* **2016**, *44*, 878–893. [CrossRef]
17. Canavos, G.C. *Applied Probability and Statistical Methods*; Little, Brown and Company: Toronto, ON, Canada, 1984.
18. Vapnik, V.N. *The Nature of Statistical Learning Theory*; Springer: New York, NY, USA, 1995.
19. Smola, A.; Scholkopf, B. A tutorial on support vector regression. *Stat. Comput.* **2004**, *14*, 199–222. [CrossRef]
20. Wei, C.; Keerthi, S.S.; Chong, J.O. Bayesian support vector regression using a unified cost function. *IEEE Trans. Neural Netw.* **2004**, *15*, 29–44.
21. Girosi, F. *Models of Noise and Robust Estimates*; A.I. Memo No. 1287; Massachusetts Institute of Technology, Artificial Intelligence Laboratory: Cambridge, MA, USA, 1991.
22. Pontil, M.; Mukherjee, S.; Girosi, F. On the Noise Model of Support Vector Machines Regression. In Proceedings of the 11th International Conference on Algorithmic Learning Theory, Sydney, NSW, Australia, 11–13 December 2000; pp. 316–324.

23. Goldberg, D.E.; Holland, J.H. Genetic algorithm and machine learning. *Mach. Learn.* **1988**, *3*, 95–99. [CrossRef]

24. Eren, B.; Vedide Rezan, U.; Ufuk, Y.; Egrioglu, E. A modified genetic algorithm for forecasting fuzzy time series. *Appl. Intell.* **2014**, *41*, 453–463.

25. Wei, H.; Tang, X.S.; Liu, H. A genetic algorithm(GA)-based method for the combinatorial optimization in contour formation. *Appl. Intell.* **2015**, *43*, 112–131. [CrossRef]

26. Zojaji, Z. Semantic schema theory for genetic programming. *Appl. Intell.* **2016**, *44*, 67–87. [CrossRef]

27. Shi, K.S.; Li, L.M. High performance genetic algorithm based text clustering using parts of speech and outlier elimination. *Appl. Intell.* **2013**, *38*, 511–519. [CrossRef]

28. Trivedi, A.; Srinivasan, D.; Biswas, S.; Reindl, T. A genetic algorithm—Differential evolution based hybrid framework: Case study on unit commitment scheduling problem. *Inf. Sci.* **2016**, *354*, 275–300. [CrossRef]

29. Wu, Q.; Law, R. Fuzzy support vector regression machine with penalizing Gaussian noises on triangular fuzzy number space. *Expert Syst. Appl.* **2010**, *37*, 7788–7795. [CrossRef]

30. Gajowniczek, K.; Zabkowski, T. Simulation study on clustering approaches for short-term electricity forecasting. *Complexity* **2018**, *2018*, 3683969. [CrossRef]

31. Massidda, L.; Marrocu, M. Smart meter forecasting from one minute to one year horizons. *Energies* **2018**, *11*, 3520. [CrossRef]

32. Petković, D.; Shamshirband, S.; Saboohi, H.; Ang, T.F.; Anuar, N.B.; Pavlović, N.D. Support vector regression methodology for prediction of input displacement of adaptive compliant robotic gripper. *Appl. Intell.* **2014**, *41*, 887–896.

33. Chang, C.C.; Lin, C.J. Training v-Support Vector Regression: Theory and Algorithms. *Neural Comput.* **2002**, *14*, 1959–1977. [CrossRef]

34. Chalimourda, A.; Schölkopf, B.; Smola, A.J. Experimentally optimal v in support vector regression for different noise models and parameter settings. *Neural Netw.* **2004**, *17*, 127–141. [CrossRef]