Kernel ridge regression for general noise model with its application

Shiguang Zhang\textsuperscript{a,b}, Qinghua Hu\textsuperscript{c}, Zongxia Xie\textsuperscript{c}, Jusheng Mi\textsuperscript{a,d,*}

\textsuperscript{a} College of Mathematics and Information Science, Hebei Normal University, Shijiazhuang, Hebei 050024, China
\textsuperscript{b} College of Mathematics and Computer Science, Hengshui University, Hengshui, Hebei 053000, China
\textsuperscript{c} School of Computer Science and Technology, Tianjin University, Tianjin 300072, China
\textsuperscript{d} Hebei Key Laboratory of Computational Mathematics and Applications, Shijiazhuang, Hebei 050024, China

\textbf{A R T I C L E  I N F O}

Article history:
Received 29 May 2013
Received in revised form 16 July 2014
Accepted 25 July 2014
Communicated by Christos Dimitrakakis
Available online 23 August 2014

Keywords:
Kernel ridge regression
Noise model
Loss function
Equality constraints
Short-term wind speed prediction

\textbf{A B S T R A C T}

The classical ridge regression technique makes an assumption that the noise is Gaussian. However, it is reported that the noise models in some practical applications do not satisfy Gaussian distribution, such as wind speed prediction. In this case, the classical regression techniques are not optimal. So we derive an optimal loss function and construct a new framework of kernel ridge regression technique for general noise model (N-KRR). The Augmented Lagrangian Multiplier method is introduced to solve N-KRR. We test the proposed technique on artificial data and short-term wind speed prediction. Experimental results confirm the effectiveness of the proposed model.

\textcopyright 2014 Elsevier B.V. All rights reserved.

\section{Introduction}

Ridge regression techniques are widely used in stock price prediction, marketing analysis, power consuming prediction, wind speed forecasting, etc. Although these techniques have been successfully applied in various domains, new challenges and problems are still reported in some practical applications. This topic is attracting much attention from application and research areas these years [1–3].

We first introduce some notations. Given a set of training data

\[ D_l = \{(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l)\}, \]

where \( x_i \in \mathbb{R}^d, y_i \in \mathbb{R}, i = 1, 2, \ldots, l \).

A multivariate linear model is \( f(x) = \omega^T \cdot x + b \), where \( \omega \in \mathbb{R}^d, b \in \mathbb{R} \). The task of Ridge regression is to learn the parameter vector \( \omega \) and parameter \( b \), by minimizing the objective function

\[ g_{RR} = \frac{1}{2} \cdot \omega^T \cdot \omega + C \cdot \sum_{i=1}^{l} (y_i - \omega^T \cdot x_i - b)^2. \]  

Ridge regression is a method of linear regression that implements a sum-of-squares error function together with regularization, thus controlling the bias variance trade-off [4,5]. It aims at discovering a linear structure hidden in the original data [6–9]. Meanwhile, nonlinear mappings may be estimated by kernel ridge regression [10–12], an extended version of linear ridge regression with kernel tricks. A nonlinear ridge regression model is constructed in a feature space \( H \) (the nonlinear kernel mapping \( \Phi : \mathbb{R}^d \rightarrow H \), where \( H \) is the Hilbert space), induced by the nonlinear kernel function \( K(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle, \langle \Phi(x_i), \Phi(x_j) \rangle \) is the inner product. The kernel mapping \( \Phi \) may be any positive definite Mercer kernel. So the objective function minimized in kernel ridge regression can be written as

\[ g_{KRR} = \frac{1}{2} \cdot \omega^T \cdot \omega + C \cdot \sum_{i=1}^{l} (y_i - \omega^T \cdot \Phi(x_i) - b)^2. \]  

In recent years, kernel ridge regression (KRR) is gaining popularity as a data-rich nonlinear forecasting tool [3,6,8–11], which is applicable in many different contexts [12–14], such as economic field, machine learning, and especially optical character recognition.

In 2000, Suykens et al. [15–18] proposed kernel ridge regression model with Gaussian noise (GN-KRR, also known as least squares support vector regression, LS-SVR). We know that this technique is able to find the optimal model if the errors are Gaussian. However, it was reported that the noise in some real-world applications, just like wind power forecast and direction-of-arrival estimation problem, does not satisfy Gaussian distribution, but Beta distribution, Laplace distribution, or other models. In this case, the classical regression techniques are not optimal. Despite the fact that most works on wind power forecast assume a normal distribution function to represent the probability density function,

\begin{itemize}
  \item \textbf{Keywords:} Kernel ridge regression
  \item \textbf{Noise model}
  \item \textbf{Loss function}
  \item \textbf{Equality constraints}
  \item \textbf{Short-term wind speed prediction}
\end{itemize}
(PDF) of the prediction error, in [19], a more comprehensive proposal modeling that PDF as a Beta function is justified. Fabbri et al. believed that Beta distribution function is more appropriate to fit the error than the standard normal distribution function of wind power forecasting. Bludszuweit et al. [20] showed the advantages of using Beta PDF for approximating the error distribution on wind power forecasting instead of the Gaussian PDF. The error \( e \) obeys the Beta distribution in the forecast of wind power, and the PDF of \( e \) is \( f(e)=c^{m-1}(1-c)^{n-1}, h, c \in (0, 1) \), where \( m \) and \( n \) are two parameters \((m > 1, n > 1)\), \( h \) is the normalization factor [21–23] (Laplacian distribution, Gaussian distribution and Beta distribution for different parameters are shown in Fig. 1). Zhang et al. [24] and Randazzo et al. [25] presented the estimation model under a Laplacian noise environment in the direction-of-arrival of coherent electromagnetic waves impinging estimation problem. Laplacian distribution is frequently encountered in various machine learning areas [26,27].

Based on the above analysis, we know that the error distributions do not satisfy Gaussian distribution in some application fields. We know that different loss functions should be derived for different noise models. Squared loss is fit for Gaussian distribution, and absolute loss is good for Laplacian distribution. We systematically discuss the optimal loss functions for different noise models.

It is not suitable to apply the kernel ridge regression with Gaussian noise (GN-KRR) to fit functions from data with non-Gaussian noise. In order to solve the above problems, we derive a general loss function and develop a new framework of kernel ridge regression technique for the general noise model; (3) the Augmented Lagrangian Multiplier method is applied to solve the proposed model, which guarantees the stability and validity of the solution; (4) we utilize the kernel ridge regression model for Beta noise (BN-KRR) to short-term wind speed prediction and show the effectiveness of the proposed model in practical applications.

This paper is structured as follows. In Section 2, we obtain the optimal loss function and construct a kernel ridge regression machine for the general noise model \((N-KRR)\). Section 3 gives the solution and algorithm design of \(N-KRR\). Numerical experiments are conducted on artificial data and short-term wind speed prediction in Section 4. Finally, the conclusions are drawn in Section 5.

2. Kernel ridge regression for general noise model

Kernel ridge regression can be understood as a function approximation technique. A key issue in developing a regression technique is to derive an optimal objective for the general noise model. Given a set of noisy training samples \(D_h\), we require to estimate an unknown regression function \(f(x)\). We assume that the noise is additive:

\[
y_i = f(x_i) + e_i \quad (i = 1, 2, \ldots, l),
\]

where \(e_i\) is the error, and we assume that the observations are drawn in independent and identical distributed (i.i.d.) with \(P(e)\) of standard deviation \(\sigma\) and mean \(\mu\). The objective is to estimate the function \(f(x)\) with the data set \(D_l \subseteq D_h\).

Following [29–31], the general approach is to minimize

\[
H[f] = \sum_{i=1}^{l} c(e_i) + \lambda \cdot \Phi[f]
\]

where \(c(e_i) = c(y_i - f(x_i))\) is a loss function, \(\lambda\) is a positive number and \(\Phi[f]\) is a smoothness functional.

![Fig. 1. Laplacian PDF, Gaussian PDF and Beta PDF of parameters.](image-url)
In the Bayesian approach, we regard the function $f$ as the realization of a random field with a known prior probability distribution. We are interested in maximizing the a posteriori probability of $f$ given the data $D_f$, namely $P[f|D_f]$, which can be written as

$$P[f|D_f] = P[D_f|f] \cdot P[f].$$

where $P[D_f|f]$ is the conditional probability of the data $D_f$ given the function $f$ and $P[f]$ is a priori probability of random field $f$, which is often written as $P[f] \propto \exp(-\lambda \cdot \Phi[f])$, where $\Phi[f]$ is a smoothness functional. The probability $P[D_f|f]$ is essentially a model of the noise, and if the noise is additive, as in Eq. (3) and i.i.d. with probability distribution $P(e)$, $P[D_f|f]$ can be written as

$$P[D_f|f] = \prod_{i=1}^{I} P(e_i).$$

Substituting $P[f]$ and Eq. (6) in (5), maximizing the posterior probability of $f$ given the data $D_f$ is equal to minimizing the following functional

$$H[f] = -\sum_{i=1}^{I} \log[P(y_i - f(x_i))] \cdot e^{-\lambda \cdot \Phi[f]}$$

$$= -\sum_{i=1}^{I} \log P(y_i - f(x_i)) + \lambda \cdot \Phi[f].$$

(7)

This functional is of the same form as Eq. (4) [29]. By Eqs. (4) and (7), the optimal loss function in maximum a posteriori is

$$c(e) = \frac{c(x, y, f(x))}{-\log P(y - f(x))}.$$  

(8)

We assume that the noise in Eq. (3) is Laplace, with the probability density function $P(e_i) = \frac{1}{2\sigma}|e_i|$. By Eq. (8), the loss function should be $c(e) = |e_i| (i = 1, ..., I)$.

If the noise in Eq. (3) is Gaussian, with zero mean and standard deviation $\sigma$, then by Eq. (8) the loss function corresponding to Gaussian noise is $c(e_i) = \frac{1}{2\sigma^2}e_i^2 (i = 1, ..., I)$.

And if the noise in Eq. (3) is Beta, with mean $\mu \in (0, 1)$ and standard deviation $\sigma$, then by Eq. (8), the loss function corresponding to Beta noise is $c(e_i) = (1 - m)\log(e_i) + (1 - m)\log(1 - e_i), 0 < e_i < 1 (i = 1, ..., I), m > 1, n > 1$.

The Laplacian loss function, the Gaussian loss function and the Beta loss function of parameters are shown in Fig. 2.

Given samples $D_i$, we construct a nonlinear regression function $f(x) = \omega^T \Phi(x) + b$. The uniform model of kernel ridge regression technique for the general noise model (N-KRR) can be formally defined as

$$\min \left\{ g_{\text{N-KRR}} = \frac{1}{2} \omega^T \cdot \omega + C \cdot \sum_{i=1}^{I} c(e_i) \right\}$$

(9)

with $\omega, y_i, \Phi(x_i) - b = e_i (i = 1, ..., I)$ are general convex loss functions in the sample point $(x_i, y_i) \in D_i$. $C > 0$ is a penalty parameter.

We construct a Lagrangian function from the primal objective function and the corresponding constraints by introducing a dual set of variables. Standard Lagrangian techniques are used to derive the dual problem.

**Theorem 1.** The dual problem of the primal problem (9) of N-KRR is

$$\max \left\{ g_{\text{N-KRR}} = \frac{1}{2} \sum_{i=1}^{I} \sum_{j=1}^{I} (\alpha_i \cdot \alpha_j \cdot K(x_i, x_j))$$

$$+ \sum_{i=1}^{I} (\alpha_i \cdot y_i) + C \cdot \sum_{i=1}^{I} T(e_i(\alpha_i)) \right\}$$

(10)

where $T(e_i(\alpha)) = \alpha_i^2(\alpha_i) - \alpha_i(\alpha_i) \cdot \partial c(e_i(\alpha_i)) / \partial c(e_i(\alpha_i)), \epsilon_i$ is a function of variable $\alpha_i (i = 1, ..., I), C > 0$ is constant.
Proof. We introduce the Lagrangian functional \( L(\omega, b, \alpha, e) \) as
\[
L(\omega, b, \alpha, e) = \frac{1}{2} \alpha^T \nabla^2 \omega f(\omega, b, \alpha, e) - \sum_{i=1}^l \left( c(e_i) + \sum_{j=1}^l \alpha_i (y_j - \omega^T \Phi(x_i) - b) - e_i \right).
\]
(11)

To minimize \( L(\omega, b, \alpha, e) \), we derive partial derivative \( \omega, b, \alpha, e \).

According to the KKT (Karush–Kuhn–Tucker) conditions, we have
\[
\nabla \omega = 0, \quad \nabla b = 0, \quad \nabla \alpha = 0, \quad \nabla e = 0.
\]
(12)

Thus
\[
\alpha = \sum_{i=1}^l (\alpha_i \cdot \Phi(x_i)), \quad b = \frac{1}{l} \sum_{j=1}^l \left( y_j - \sum_{i=1}^l (\alpha_i \cdot K(x_i, x_j)) - e_i(\alpha_i) \right).
\]
(13)

Substituting the extreme conditions into \( L(\omega, b, \alpha, e) \) and seeking maximum of \( \alpha \), we obtain the dual problem (10) of the primal problem (9).

Now
\[
\omega = \sum_{i=1}^l (\alpha_i \cdot \Phi(x_i)), \quad b = \frac{1}{l} \sum_{j=1}^l \left( y_j - \sum_{i=1}^l (\alpha_i \cdot K(x_i, x_j)) - e_i(\alpha_i) \right).
\]
(14)

We obtain the regression function of N-KRR as
\[
f(\omega) = \omega^T \Phi(\omega) + b = \sum_{i=1}^l (\alpha_i \cdot K(x_i, x_i)) + b.
\]
(15)

The dual problem (10) of kernel ridge regression for Gaussian noise model (GN-KRR) and kernel ridge regression for Beta noise model (BN-KRR) are described as follows:

(1) Gaussian noise model GN-KRR: References [16–18] studied the kernel ridge regression machine for Gaussian noise model. The loss function is \( c(e_i) = e_i^2 / 2 \) (1 = 1, ..., l). Thus the dual problem of GN-KRR is
\[
\begin{aligned}
\text{max} & \quad g_{D_{GN-KRR}} = -\frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l (\alpha_i \cdot \alpha_j \cdot K(x_i, x_j)) \\
& + \sum_{i=1}^l \left( \alpha_i \cdot y_i - \frac{1}{2l} \cdot \sum_{i=1}^l (\alpha_i^2) \right)
\end{aligned}
\]
(16)

(2) KRR for Beta noise BN-KRR: The loss function for Beta noise is \( c(e_i) = (1 - m) \log(e_i) + (1 - n) \log(1 - e_i), \) where \( 0 < e_i < 1 \) (i = 1, ..., l), \( m > 1, n > 1 \). On account of \( \partial c(e_i) / \partial e_i = (1 - m)e_i - (1 - n)/(1 - e_i) \), by \( \nabla \cdot \partial c(e_i) / \partial e_i - \alpha_i = 0 \), we have \( \alpha_i / C - e_i^2 / C = 2 \alpha_i / C - m - n - 1 \). Now we get
\[
e_i(\alpha_i) = \frac{2 + \alpha_i / C - m - n + \Delta^{1/2}}{2 \cdot \alpha_i / C} \\
e_i(\alpha_i) = \frac{2 + \alpha_i / C - m - n - \Delta^{1/2}}{2 \cdot \alpha_i / C},
\]
where \( \Delta = (\alpha_i / C + m - n)^2 + 4(1 + mn - m - n), \) \( m > 1, n > 1, \) As \( 0 < e_i(\alpha_i) < 1 \), we reject \( e_i(\alpha_i) \) and adopt \( e_i(\alpha_i) \) let \( e_i(\alpha_i) = e_i(\alpha_i) \). Then the dual problem of BN-KRR is
\[
\begin{aligned}
\text{max} & \quad g_{D_{BN-KRR}} = -\frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l (\alpha_i \cdot \alpha_j \cdot K(x_i, x_j)) \\
& + \sum_{i=1}^l (\alpha_i \cdot y_i - \frac{1}{2l} \cdot \sum_{i=1}^l (\alpha_i^2)) \\
& + C \cdot \sum_{i=1}^l \left( (1 - m) \cdot \log(e_i(\alpha_i)) \right) \\
& + \left( 1 - m \cdot \log(1 - e_i(\alpha_i)) \right)
\end{aligned}
\]
(17)

3. Solution based on Augmented Lagrangian Multiplier approach

Theorem 1 provides a method to solve the model of N-KRR, the dual problem (16) and (17) and the \( \nu \)-support vector regression machine for Gaussian noise model (denoted by GN-SVR) [32] are quadratic convex function. The Augmented Lagrangian Multiplier (ALM) method is a certain class of method for solving equality or inequality constrained optimization problems. For nonlinear

![Fig. 3.](image-url) The forecasting result of two models on artificial data with Gaussian noise.
programming problems with equality constraints, Powell and Hestenes [33] designed a dual method of solution, where squares of the constraint functions are added as penalties to the Lagrangian function, and a certain simple rule is used for updating the Lagrangian multipliers after each cycle (called the PH algorithm). Rockafellar [33,34] extended it for solving inequality constrained tasks (called the PHR algorithm). The basic idea is starting from the Lagrangian function of the original problem, coupled with an appropriate penalty function. The original problem is transformed into solving a series of unconstrained optimization sub-problems.

The ALM method was employed to solve Problem (10) by applying Newton’s method to a sequence of equality constrained problems. Any equality constrained minimization problem can be reduced to an equivalent unconstrained problem by eliminating the equality constraints. The gradient descent method or Newton’s method can be used to solve the problem [28,35].

In this work, we solve the kernel ridge regression machine with the ALM approach. ALM is employed to solve the dual problem (10) by applying Newton’s method to a sequence of equality constrained problems. We obtain the solution based on ALM method and the algorithm design of kernel ridge regression machine for the general noise model (N-KRR) as

(1) Let data set \( D_i = \{(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l)\} \), where \( x_i \in \mathbb{R}^k \) and \( y_i \in \mathbb{R} \), \( i = 1, \ldots, l \).

(2) Use 10-fold cross validation strategy to choose the optimal parameters \( C, m, n \) and select a kernel function \( K(\bullet, \bullet) \).

(3) Construct and solve the optimization dual problem (10). We obtain the optimal solution \( \alpha = (\alpha_1, \ldots, \alpha_l) \).

(4) Construct the function as

\[
 f(x) = \sum_{i=1}^{l} \left( \alpha_i \cdot K(x_i, x) \right) + b,
\]

where \( b = \frac{1}{l} \sum_{i=1}^{l} [y_i - \sum_{j=1}^{l} (\alpha_j \cdot K(x_i, x_j)) - e_i(\alpha_j)] \).

4. Experimental analysis

In this section, we present some experiments to test the proposed model on several regression tasks with artificial data and a real-world application of short-term wind speed prediction.

In the models of GN-KRR, GN-SVR and BN-KRR, the initial parameters of the proposed ALM approach are \( C \in [1, 201] \), \( m \in (1, 21) \). We use 10-fold cross validation strategy to search the optimal positive parameters \( C, m, n \) and the selection technique of parameters \( C, m, n \) was studied in detail in [36,37]. Many actual applications suggest that polynomial and Gaussian kernel functions tend to perform well under general smoothness assumptions, so that it should be considered especially if no additional information is used as the kernel function of N-KRR. In this work, polynomial function and Gaussian kernel function are used as the kernel functions of GN-KRR, BN-KRR and GN-SVR [17,31].

\[
 K(x_i, x_j) = ((x_i, x_j) + 1)^d, \\
 K(x_i, x_j) = e^{-\frac{||x_i, x_j||^2}{\sigma^2}},
\]

where \( d \) is a positive integer, and \( \sigma \) is positive.

We introduce three criteria, mean absolute error (MAE), mean absolute percentage error (MAPE) and the root mean square error

<table>
<thead>
<tr>
<th>Model</th>
<th>MAE (m/s)</th>
<th>RMSE (m/s)</th>
<th>MAPE (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GN-KRR</td>
<td>0.9425</td>
<td>1.1933</td>
<td>13.72</td>
</tr>
<tr>
<td>BN-KRR</td>
<td>1.4230</td>
<td>1.7576</td>
<td>29.97</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>MAE (m/s)</th>
<th>RMSE (m/s)</th>
<th>MAPE (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GN-KRR</td>
<td>1.8685</td>
<td>2.2824</td>
<td>83.43</td>
</tr>
<tr>
<td>BN-KRR</td>
<td>1.3949</td>
<td>1.7294</td>
<td>82.80</td>
</tr>
</tbody>
</table>

Fig. 4. The forecasting result of two models on artificial data with Beta noise.
(RMSE), to evaluate the performance of GN-KRR, GN-SVR and BN-KRR:

\[ MAE = \frac{1}{l} \sum_{i=1}^{l} |y_i - y'_i|, \]  

\[ MAPE = \frac{1}{l} \sum_{i=1}^{l} \frac{|y_i - y'_i|}{y_i}, \]  

\[ RMSE = \sqrt{\frac{1}{l} \sum_{i=1}^{l} (y_i - y'_i)^2}. \]  

where \( l \) is the size of the samples, \( y_i \) is the \( i \)th data and \( y'_i \) is the forecasting result of the \( i \)th data [19,20].

4.1. Artificial data sets

(1) Artificial data with Gaussian noise is \((X, Y)\), where \( X = [0.01 \times \pi : 0.01 \times \pi : 10 \times \pi] \), the relationship between \( x \) and \( y \) is as follows:

\[ y = 2 \cdot \sin(3 \cdot x + 2) + 6 \cdot \cos(2 \cdot x + 1) + \text{norm}(0, 1, 1000, 1), \]  

\( x \in X, y \in Y \). Let the number of training samples is 150 (from 1 to 150), and the number of tested samples is 150.
Fig. 3 shows the forecasting results of artificial data with Gaussian noise by GN-KRR and BN-KRR.

The indicators of MAE, MAPE and RMSE are used to assess the forecasting results of two models shown in Table 1. Experimental results on artificial data with Gaussian noise show that GN-KRR is better than BN-KRR.

(2) Artificial data with Beta noise is \((X, Y)\), where \(X = [0.01 \times \pi : 0.01 \times \pi : 10 \times \pi]\), the relationship between \(x\) and \(y\) is as follows: \(y = 2 \cdot \sin(3 \cdot x + 2) + 6 \cdot \cos(2 \cdot x + 1) + 6 \cdot \text{Beta}(2.6084, 3.0889, 1000, 1)\), \(x \in X, y \in Y\). Let the number of training samples is 150 (from 1 to 150), and the number of tested samples is 150 (from 151 to 300). Fig. 4 shows the forecasting results of artificial data with Beta noise by GN-KRR and BN-KRR.

The indicators of MAE, MAPE and RMSE are used to assess the forecasting results of two models shown in Table 2. Experimental results on artificial data with Beta noise show that BN-KRR obtains better performance than GN-KRR.

4.2. Short-term wind speed prediction

The forecasting model of N-KRR is applied in real-world data set of wind speed in Heilongjiang Province. We collect more than one year samples which record the average wind speeds in ten minutes from a wind speed farm.

Here we analyze one-month time series of wind speeds, and investigate the error distribution by using the persistence method [20]. We consider the wind speed \(x_i\) as the predicting values of \(x_{i+1}\). Then we compute the prediction error and give the statistical distribution in Fig. 5. We can see that the error \(\epsilon\) of wind speed with persistence forecast does not obey Gaussian distribution,

![Fig. 7. The boxplot of residual on wind speed prediction after 10 min in Spring.](image1)

![Fig. 8. The forecasting result of three models on wind speed data after 10 min in Summer.](image2)
Fig. 9. The boxplot of residual on wind speed prediction after 10 min in Summer.

Fig. 10. The forecasting result of three models on wind speed data after 30 min in Spring.

Fig. 11. The boxplot of residual on wind speed prediction after 30 min in Spring.
while obeys Beta distribution, and the PDF of $\epsilon$ is $f(\epsilon) = \epsilon^{2.22} \cdot (1-\epsilon)^{2.32}, \epsilon \in (0, 1)$. This result shows that the traditional modeling techniques are not optimal in wind speed prediction.

Now we test the performance of different techniques in wind speed prediction. The experiment setup is described as follows. We extract 864 samples as a training set (which includes six-day wind speeds) and the subsequent 864 samples as the tested set. We transform the original series into a multivariate task by using 12 dimensions vector $\mathbf{x}_i = (x_{i-11}, x_{i-10}, \ldots, x_{i-1}, x_i)$ as input variables to predict $x_{i+\text{step}}$, where $\text{step} = 1, 3$. That is to say, we use the above mode to predict the wind speed after 10 and 30 min of every point $x_i$. Figs. 6–13 and Tables 3–6 show the forecasting results after 10 and 30 min of every point $x_i$ given by GN-KRR, GN-SVR and BN-KRR in Spring and Summer, respectively. In each plot, we compute the predicting values and errors produced by GN-KRR, BN-KRR and GN-SVR.

1) Wind speed prediction after 10 min: Figs. 6–9 give the forecasting results after 10 min of every point $x_i$ given by GN-KRR, GN-SVR and BN-KRR in Spring and Summer. The indicators of MAE, MAPE and RMSE are used to assess the forecasting results of three models shown in Tables 3 and 4.

2) Wind speed prediction after 30 min: Figs. 10–13 illuminate the forecasting results after 30 min of every point $x_i$ given by GN-KRR, GN-SVR and BN-KRR in Spring and Summer. The indicators of MAE, MAPE and RMSE are used to assess the forecasting results of three models shown in Tables 5 and 6.

Observing from the errors, it is to derive that the errors computed with BN-KRR are a little smaller than those with GN-KRR and GN-SVR in the most cases. As the prediction horizon increases to 30-min, both the errors derived with different models rise, and the relative difference decreases. So it is not so significant in these cases. However, we can know that the Beta model is still better than the Gaussian models in terms of all the criteria of MAE, RMSE, MAPE, seen from Tables 3–6.

5. Conclusions

The traditional kernel ridge regression technique makes the assumption that the error distribution is Gaussian. However, it was reported that the noise models in some real-world applications do not satisfy Gaussian distribution. In this case, the current regression
techniques are not optimal. In this work, we derive a general loss function and develop a new kernel ridge regression technique for general noise model. According to KKT conditions, we introduce the Lagrangian functional and obtain the dual problem of the general noise model. According to KKT conditions, we introduce the dual problem of the general noise model in the future.

This work just discussed the problem of training regression models with non-Gaussian noise. In fact, there is a similar challenge for different noise models and parameter settings, Neural Comput. 9 (1997) 521–540.

Acknowledgments

This work was supported by National Program on Key Basic Research Project (973 Program) under Grant 2012CB215201, National Natural Science Foundation of China (NSFC) under Grants 61222210, 61171007, 61300121, 61105054, and the Scientific research project of Universities in Hebei Province (No. Z2010188).

References


Table 3

<table>
<thead>
<tr>
<th>Model</th>
<th>MAE (m/s)</th>
<th>RMSE (m/s)</th>
<th>MAPE (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GN-SVR</td>
<td>0.7092</td>
<td>0.9340</td>
<td>8.07</td>
</tr>
<tr>
<td>GN-KRR</td>
<td>0.7405</td>
<td>0.9736</td>
<td>8.26</td>
</tr>
<tr>
<td>BN-KRR</td>
<td>0.5736</td>
<td>0.7221</td>
<td>6.20</td>
</tr>
</tbody>
</table>

Table 4

<table>
<thead>
<tr>
<th>Model</th>
<th>MAE (m/s)</th>
<th>RMSE (m/s)</th>
<th>MAPE (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GN-SVR</td>
<td>0.7395</td>
<td>0.9552</td>
<td>14.58</td>
</tr>
<tr>
<td>GN-KRR</td>
<td>0.7055</td>
<td>0.9220</td>
<td>16.05</td>
</tr>
<tr>
<td>BN-KRR</td>
<td>0.6071</td>
<td>0.7633</td>
<td>11.28</td>
</tr>
</tbody>
</table>

Table 5

<table>
<thead>
<tr>
<th>Model</th>
<th>MAE (m/s)</th>
<th>RMSE (m/s)</th>
<th>MAPE (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GN-SVR</td>
<td>1.1516</td>
<td>1.4999</td>
<td>12.96</td>
</tr>
<tr>
<td>GN-KRR</td>
<td>1.0883</td>
<td>1.4130</td>
<td>12.40</td>
</tr>
<tr>
<td>BN-KRR</td>
<td>0.9483</td>
<td>1.2400</td>
<td>10.87</td>
</tr>
</tbody>
</table>

Table 6

<table>
<thead>
<tr>
<th>Model</th>
<th>MAE (m/s)</th>
<th>RMSE (m/s)</th>
<th>MAPE (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GN-SVR</td>
<td>0.4345</td>
<td>0.5769</td>
<td>10.21</td>
</tr>
<tr>
<td>GN-KRR</td>
<td>0.4155</td>
<td>0.5623</td>
<td>7.92</td>
</tr>
<tr>
<td>BN-KRR</td>
<td>0.2683</td>
<td>0.4137</td>
<td>4.76</td>
</tr>
</tbody>
</table>
**Shiguang Zhang** received his B.S., M.S. and Ph.D. degrees from Shaanxi Normal University, Guangxi University for Nationalities, and Hebei Normal University in 1998, 2007 and 2014, respectively. Now he is a lecturer with College of Mathematics and Computer Science, Hengshui University. His research interests are focused on machine learning and knowledge discovery for regression.

**Qinghua Hu** received his B.S., M.S. and Ph.D. degrees from Harbin Institute of Technology, Harbin, China in 1999, 2002 and 2008, respectively. He worked with Harbin Institute of Technology from 2006 to 2011. He also worked as a Postdoctoral Fellow with Department of Computing, The Hong Kong Polytechnic University from 2009 to 2011. Now he is a full professor with School of Computer Science and Technology, Tianjin University. His research interests are focused on rough sets, granular computing, data mining for classification and regression. He is a PC Co-Chairs of RSCTC 2010, CRSSC 2012 and ICMLC 2014 and serves as referee for a great number of journals and conferences. He has published more than 100 journal and conference papers in the areas of reasoning with uncertainty, pattern recognition and fault diagnosis.

**Zongxia Xie** received her B.S. from Dalian Maritime University, in 2003, and M.S. and Ph.D. from Harbin Institute of Technology, in 2005 and 2010, respectively. Now she is an associate professor with Tianjin University. Her major interests include machine learning and pattern recognition with rough sets and SVM, solar image processing and knowledge discovery. She has published more than 20 conference and journal papers on related topics.

**Jusheng Mi** received his B.S. from Hebei Normal University, in 1982, M.S. from East China Normal University, in 1992, and Ph.D. from Xi’an Jiaotong University, in 2003. Now he is a professor with Hebei Normal University. His major interests are focused on rough sets, granular computing, data mining and approximate reasoning. He has published more than 60 journal and conference papers on related topics.