Renewable Energy 177 (2021) 181–192

Contents lists available at ScienceDirect

Renewable Energy

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

Very short-term probabilistic wind power prediction using sparse machine learning and nonparametric density estimation algorithms

Jiaqing Lv [a], Xiaodong Zheng [b, *], Mirosław Pawlak [a, c, d], Weike Mo [b], Marek Miśkowicz [a]

[a] AGH University of Science and Technology, Poland
[b] South China University of Technology, China
[c] University of Manitoba, Canada
[d] Information Technology Institute, University of Social Sciences, Poland

ARTICLE INFO

Article history:
Received 25 January 2020
Received in revised form
18 May 2021
Accepted 21 May 2021
Available online 28 May 2021

Keywords:
Interval forecast
Multivariate estimation
Nonparametric density estimation
Sparse modeling
Wind power prediction

ABSTRACT

In this paper, a sparse machine learning technique is applied to predict the next-hour wind power. The hourly wind power prediction values within a few future hours can be obtained by meteorological/physical methods, and such values are often broadcast and available for many wind generators. Our model takes into consideration those available forecast values, together with the real-time observations of the past hours, as well as the values in all the power generators in nearby locations. Such a model is consisted of features of high dimensions, and is solved by the sparse technique. We demonstrate our method using the realistic wind power data that belongs to the IEEE 118-bus test system named *NREL-118*. The modeling result shows that our approach leads to better prediction accuracy comparing to several other competing methods, and our results improves from the broadcast values obtained by meteorological/physical methods. Apart from that, we apply a novel nonparametric density estimation approach to give the probabilistic band of prediction, which is demonstrated by the 25% and 75% confidence interval of the prediction. The coverage rate is compared with that yielded from quantile regression.

1. Introduction

Renewable energy generation, especially wind power, are being increasingly integrated into the power systems worldwide [1]. Despite the environmental friendliness and sustainability properties, wind power challenges the economic operation, stability, and reliability of power systems with its intermittence and randomness [2,3]. Therefore, improving the accuracy of short-term (e.g., day-ahead or two-days ahead) and very short-term (i.e., up to several-hours ahead) wind power forecasts has become a paramount issue [4]. We focus in this paper on the algorithms of predicting the next-hour wind power, hence the very short-term wind power prediction.

According to Ref. [4], very short-term wind power forecasts rely much less on the numerical weather prediction (NWP) or meteorological/physical methods. For this time horizon, NWP-based methods do not obtain performance improvements with respect to persistence methods, and statistical models should outperform physical models in general [5].

For statistical learning methods, one crucial issue is to select factors that are correlated with the wind power, and avoid the "over-fitting" caused by accounting for those irrelevant elements in the learning process. In this context, lots of work has been done to select key factors. For example, in Ref. [6], appropriate factors are selected according to correlation and importance measures, and then employed to a nonparametric learning machine, i.e., a random forest predictor, to forecast the 1-h ahead wind power. The predictor is immune to irrelevant inputs. Authors in Ref. [7] use the stacked denoising auto-encoders (SDAE) with batch normalization to extract the deep features of wind speed data. In Ref. [8], after performing feature selection with the mutual information, the authors propose using an SDAE and long short term memory (LSTM) network to predict the 10-min ahead wind speed. In Ref. [9], a dilation and erosion clustering algorithm based on mathematical morphology is proposed to select the days with similar prediction information, which is shown to improve the accuracy of day-ahead wind power prediction. Regarding linear regression models, the Lasso (least absolute shrinkage and selection operator) is an

* Corresponding author.
E-mail address: z.xiaodong@mail.scut.edu.cn (X. Zheng).

https://doi.org/10.1016/j.renene.2021.05.123
0960-1481/

effective approach to feature selection. For instance, the Lasso is used in Ref. [10] to predict the long-term wind energy, which takes as input the meteorological variables, including the wind speed, temperature, and pressure. The authors in Ref. [11] use the Lasso to explore a set of different sparse structures for the vector autoregression (VAR) model. The forecast skill is improved against conventional autoregressive and vector autoregressive models, and the method is applicable to large-scale systems. Recently [12], extends the Lasso-based VAR model proposed in Ref. [11] by accounting for changes in the spatio-temporal wind power dynamics and proposing a novel coordinate descent algorithm for solving the Lasso estimator. The Lasso is adopted in Ref. [13] for solar power generation forecast. Later in Ref. [14], and with an application to the short-term prediction of solar intensity, the authors incorporate the Lasso with LSTM to capture both the linear and nonlinear relationships within the data.

Point forecasts can be regarded as the degenerated versions of probabilistic forecasts, since they yield merely the conditional expectation of the wind power for each look-ahead time [15]. Probabilistic forecasts provide more information on the random wind power, and thus they are more applicable in power system operations [16]: with the interval estimation of wind power, the voltage can be maintained within safe bounds with a desired probability [17]; by employing wind power forecasts in the form of polyhedra with certain probabilistic guarantees, the robust generation scheduling process can be carried out in a less-conservative manner [18], just to name a few.

It is shown in Ref. [19] that the probability distribution of wind power forecast error is fat-tailed. Therefore, it cannot be modeled by the Gaussian distribution; the Beta distribution cannot always perfectly models the errors as well. In Refs. [20,21], a mixture of generalized logit-normal distributions and probability masses at the bounds is developed to describe the distribution of very short-term wind power. In Ref. [22], an empirical wavelet transform is used to extract vital modes from the original data, and then a Gaussian process regression model with Student-t likelihood is adopted to forecast both the half-hour ahead and hour-ahead wind power values and distributions. Another feasible approach is the multi-distribution ensemble model. In Ref. [23], for example, an ensemble predictor built upon Gaussian, Gamma, and Laplace predictive distributions is developed for probabilistic wind power forecast. However, as has been recognized [15], the assumption that wind power and its forecast errors follow a known parametric family of distributions is quite strong. Thus, we believe that nonparametric methods, which make no assumptions on the distribution form of wind power (or errors), should be much more flexible.

Nonparametric methods have been widely investigated in recent literatures, either for distribution or interval forecasts. In order to capture the nonparametric nature of error densities observed in real-world wind power data, the epi-spline basis functions are applied in Ref. [24], and it can generate wind power scenarios that closely resemble the behavior of actual wind power observations. A nonparametric method based on the empirical dynamic modeling of wind power is proposed for very short-term probabilistic forecast by Ref. [25] recently. To quantify the uncertainty of wind speed forecast, kernel density estimation is utilized in Ref. [26] to fit the probability distribution of the Lorenz Disturbance Sequence that describes the behavior of wind speed. However, it should be noted that the density estimator employed by Ref. [26] is a univariate one. The sparse probabilistic learning method, relevance vector machine (RVM), is applied in Ref. [27] for very short-term wind power forecast. In Ref. [27], a grouping mechanism and a sampling selection method are proposed by the authors to improve the forecast efficiency and accuracy, respectively. Authors in Ref. [28] develops an ensemble of neural networks for wind power forecast, whereas RVM is adopted as a robust auxiliary predictor. In Ref. [29], a two-layer ensemble machine learning technique is proposed to predict the hour-ahead wind power, in which the first layer is designed for feature selection, and the second layer contains a blending algorithm that generates both deterministic and probabilistic forecasts. In Ref. [30], a self-adaptive evolutionary extreme learning machine is developed to directly model the prediction intervals of wind power generation with different confidence levels. In Ref. [31], an instance-based transfer learning embedded gradient boosting decision trees model is proposed to derive multiple qualtiles for wind power forecast. In Ref. [32], a multi-kernel ridge regression method is proposed to directly construct the prediction intervals for short-term forecast of wind power. To increase the accuracy, the original time series data of wind power are decomposed into a number of modes via mode-decomposition based methods. Within the broad category of predictors specifically devised for interval forecasts, quantile regression should be one of the most straightforward and efficient methods [33]. For example [26], uses quantile regression to derive the prediction interval of wind speed. A novel joint quantile model is developed in Ref. [34] to facilitate the simultaneously evaluation of uncertainties. Recently [35], proposes a novel quantile regression model that is equipped with a novel nearest neighbors quantile filter. By leveraging a modified training data set, computational superiority is gained.

For probabilistic forecasts, it is also vital to address the dimensionality issues [36]. The multivariate distribution of wind power would allow one to capture the spatio-temporal correlation, and derive a joint probability of wind power quantile for further applications like optimal power dispatch and strategic bidding. This may, however, rely heavily on a huge dataset (which maybe not available in practice), and moreover, lead to numerical difficulties [36]. To address the probabilistic forecast problem of multiple correlated wind generators, in Ref. [37], kernel density estimator is adopted to draw out the marginal power distribution of each wind generator, whereas the dependence structure between wind generators is captured by the regular vine copula. A sparse vector autoregressive process is employed in Ref. [21] to model the location parameters of multiple wind generators, from which a logit-normal distribution can be recovered for each wind generator. As such, the dimensionality issue can be well addressed. However, the scale parameter (as a measure of spread) is approximated site-by-site, and the method can only provide univariate distributions [21].

Another important fact is that the forecast errors are conditional on many other forecast regimes like the wind power generation levels [15,38,39], e.g., the magnitude and distribution of errors near the bounds may distinct a lot with those within the medium range. In Ref. [53], conditional kernel density estimation is used to draw out the wind power distribution, and by optimizing the approach towards estimation of the desired quantile, accurate wind power quantile forecasts could be achieved. It should be noting that, therein, the wind power distribution to be forecast is conditioned on the wind velocity, instead of a point forecast of wind power as in our method presented later. In Ref. [40], the authors divide the wind power dataset into multiple generation levels; based on the observation that prediction errors depend on the wind power level, then, conditional distribution of prediction errors is calculated via a kernel density estimator. Further, in Ref. [41], a diffusion-based kernel density estimator is proposed for probabilistic wind power forecast, and the cumulative distribution functions yielded are then used to calculate the lower/upper bounds for a desired confidence level.

To summarize, a skilled predictor is expected to *i*) select

correlated features from historical data, *ii*) generate probabilistic information for the operation/dispatching use, *iii*) be compatible with the nonparametric characteristic of wind power, and *iv*) address the dimensionality issue without losing too much accuracy.

The objective of this paper is to develop such a predictor in the context of very short-term wind power prediction. The contributions made in this paper include:

1. Predict the next-hour wind power based on a large number of features including: *i*) the previous real-time observations at the wind generator of modeling interest; *ii*) the previous and future power values forecast through meteorological/physical methods which have been broadcast to the public; and *iii*) the *aforementioned* data in all nearby wind generators. To the best of our knowledge, no research in the field so far has incorporated the features in Part *ii*), while only a few work has included the Part *iii*) features.
2. Use sparse machine learning algorithms to model and predict the next-hour wind power. The L_1 design enables the automatic feature selection from more than hundreds of dimensions of candidate features. The final models after learning only consist of a small number of features among the complete set of candidate features.
3. We compare our prediction accuracy with other competing approaches. These methods include: *i*) classical linear regression, and ridge regression. *ii*) our modeling approach without consideration of the candidate features obtained from the meteorological/physical methods. *iii*) the predicted values from the meteorological/physical approach. Apart from that, we also compare our approach with *iv*) support vector regression, relevance vector machine, decision trees method, time-series model, and long short-term memory method.
4. We examine the joint density of our prediction error term and the prediction value term. A nonparametric technique called the bivariate kernel density estimator is used. By formulation of the conditional density, we can obtain a confidence interval for our prediction. We would like to highlight the novel idea of conditioning the distribution of forecast error on the point forecast via bivariate density estimation, and that the examination of these two variables to build prediction intervals has not been applied in the field yet for probabilistic forecast.

2. Machine learning methods for sparse modeling

In this section, we give a brief description on the machine learning algorithms in the wind power forecast problem. In system modeling, data of the form

$$D_n = \{(\boldsymbol{X}_1, Y_1), (\boldsymbol{X}_2, Y_2), \ldots, (\boldsymbol{X}_n, Y_n)\}$$

are available, where $\boldsymbol{X}_i \in \mathbb{R}^p$, $i \in \{1, \ldots, n\}$ is a p-dimensional vector representing the input observations, and $Y_i \in \mathbb{R}^1$, $i \in \{1, \ldots, n\}$ is a 1-dimensional variable representing the output response. The aim of system modeling is to learn the mapping between them, such that for the future observations, the predicted output and the true response can be as close as possible.

In our wind power modeling problem, the output variable Y_i corresponds to the next-hour wind power value at a particular station of interest. The system input $\boldsymbol{X}_i$ includes the current and previous real-time values that have been observed, together with the current, previous, and future forecast values that have been published. It is worth noting that the input variable includes not only the features in the station of interest, but also all the wind generators nearby, and this is known as the spatial correlation of wind power. In order to model the relationship between the input and output variables, various structures can be assumed for the model. However, it is always useful to start from assuming a linear structure

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \epsilon_i, \tag{1}$$

where ϵ_i is the noise process, $\{\beta_0, \ldots, \beta_p\}$ are the weights that need to be identified. It is worth to mention that statisticians often express data in the following design matrix form:

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \qquad \boldsymbol{X} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{np} \end{bmatrix},$$

$$\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_p]^\top, \qquad \boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_n]^\top,$$

then the linear model (1) can be expressed by

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{2}$$

The most classical way to solve the linear model is through the minimization of the mean square error, i.e.,

$$\hat{\beta}_{LS} = \arg\min_{\beta} \frac{1}{n} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2, \tag{3}$$

where $\frac{1}{n}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$ is equivalent to $\frac{1}{n}\sum_{i=1}^{n}(Y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_{ij})^2$, and the solution has the following analytical form [42].

$$\hat{\beta}_{LS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{Y}). \tag{4}$$

The classical linear regression provides a simple yet straightforward solution to the linear model in (1). However, the least square approach in (4) naturally has several inherent drawbacks that make it unfeasible by modern statistical researchers. One of these drawbacks is the numerical instability, i.e., a small change in the data, such as when new observations come to be available, may result in a large variation of the estimate $\hat{\beta}$. Also, the matrix $\boldsymbol{X}^T\boldsymbol{X}$ might not be full rank, rendering a problem in the matrix inverse process.

In order to overcome the defects in the classical LS approach, the so-called ridge regression was developed in the 1970s [43]. It is to minimize the least square criterion together with a penalty term on the weights. Namely,

$$\hat{\beta}_{ridge} = \arg\min_{\beta} \left(\frac{1}{n} \left\| \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} \right\|_2^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right), \tag{5}$$

where λ is a regularization parameter that needs to be specified. It controls the amount of shrinkage in the minimization criterion. In practice, one can select λ according to some re-sampling methods, e.g., cross validation.

It is worth to mention that the ridge regression also has close-form solution as follows,

$$\hat{\beta}_{ridge} = (\boldsymbol{X}^\top\boldsymbol{X} + n\lambda\boldsymbol{I}_p)^{-1}(\boldsymbol{X}^\top\boldsymbol{Y}), \tag{6}$$

where $\boldsymbol{I}_p$ is a $p \times p$ identity matrix.

The ridge regression method processes the numerical stability in the solution, and has been widely applied in engineering and applied science. For instance, the so-called kernel ridge regression, as a variant form of the classical ridge regression has been applied

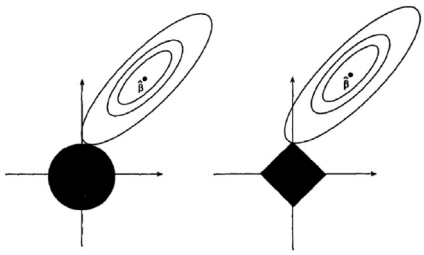

Fig. 1. The comparison between ridge regression and the Lasso.

in power engineering [32,44].

In the late 1990s, the so-called "least absolute shrinkage and selection operator" (Lasso) was developed [45]. It is also based on minimization of the least square criterion together with a penalty term. However, the penalty term is the L_1 norm of the weights, rather than the L_2 norm case as in the ridge regression case. Specifically, the Lasso has the following form,

$$\hat{\boldsymbol{\beta}}_{Lasso} = \arg\min_{\boldsymbol{\beta}} \left(\frac{1}{n} \left\| \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} \right\|_2^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right), \tag{7}$$

where λ is the regularization term that needs to be specified.

It is worth mentioning that according to the dual optimization theory, the ridge regression and the Lasso also have the following equivalent primal forms,

$$\hat{\boldsymbol{\beta}}_{ridge,primal} = \arg\min_{\boldsymbol{\beta}, \|\boldsymbol{\beta}\|_2 < s} \left(\frac{1}{n} \left\| \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} \right\|_2^2 \right), \tag{8}$$

$$\hat{\boldsymbol{\beta}}_{Lasso,primal} = \arg\min_{\boldsymbol{\beta}, \|\boldsymbol{\beta}\|_1 < s} \left(\frac{1}{n} \left\| \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} \right\|_2^2 \right), \tag{9}$$

where $\|\cdot\|_2$ and $\|\cdot\|_1$ denote the L_2 norm and the L_1 norm. Specifically, $\|\boldsymbol{\beta}\|_2 = \left(\sum_{j=1}^{p} \beta_j^2\right)^{1/2}$ and $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$. Furthermore, s is a term that can be expressed alternatively in terms of λ.

The difference between Lasso and ridge regression can be shown in the illustrative Fig. 1.

In Fig. 1, it shows an example of a bi-variate case. The sum of squares are shown. Also, the shaded area correspond to the weight constants. The left panel corresponds to the L_2 case, and the shaded area is a circular shape. In higher dimensional cases, it would be a hyper-ball. The right panel corresponds to the L_1 case since, and the shaded area is a rectangular shape. It can be seen that sometimes the optimal solution can happen at the corner points in the Lasso case, meaning part of all the features can be set to zero weight. Such scenario can almost never happen in the ridge case, meaning that ridge regression can never render a sparse solution. In high dimensional problems, it is known that the proper sparse solution can greatly enhance the model prediction accuracy due to the well known "curse of dimensionality" [46] in data science. As a consequence, it is expected that in our wind power forecast problem, the implementation of Lasso algorithm can greatly increase the model prediction performance.

Unfortunately, the solution of Lasso does not have close-form formula as the LS case (4) or the ridge case (6). In practice, after specification of regularization parameter λ, people use coordinate optimization method to calculate the solution for Lasso. This algorithm is shown in Algorithm 1, and it is also called the "shooting algorithm" for the Lasso.

Algorithm 1. Coordinate descent algorithm for computing the Lasso

Algorithm 1 Coordinate descent algorithm for computing the Lasso

1: Let $\boldsymbol{\beta}^{[0]} \in \mathbb{R}^p$ be the initial estimator. Set $s = 0$.
2: **repeat**
3: $s = s + 1$
4: **for** $j = 1, \cdots, p$ **do**
5: $\beta_j^{[m]} = \frac{\text{sign}(Z_j)(|Z_j| - \frac{\lambda}{2})_+}{\hat{\Sigma}_{jj}}$,
6: where $Z_j = \boldsymbol{X}_j^\top (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_{-j}^{[m]})/n$, $\boldsymbol{\beta}_{-j}^{[m]}$ is the same as $\boldsymbol{\beta}^{[m]}$ except the j-th component is set to zero vector, $\hat{\Sigma} = n^{-1}\boldsymbol{X}^\top\boldsymbol{X}$, and $\hat{\Sigma}_{jj}$ is the j-th diagonal component of $\hat{\Sigma}$
7: **end for**
8: **until** numerical convergence

Note that in Algorithm 1, $(c)_+$ equals to c if $c > 0$, and equals to 0 otherwise. Besides the $\beta_j, j = 1, \ldots, p$, β_0 can be estimated by taking the empirical mean of the sample $\{Y_1, \ldots, Y_n\}$. It is also worth mentioning, in order for each feature to have equal strength, the user should standardize the data so that each co-variate of $\boldsymbol{X}$ (except for the first column) should have zero mean and standard deviation. This can be done by pre-processing data when the user subtracts the mean, and then divides by the standard deviation for each feature. It is worth mentioning that it is standard practice for statistical researchers to pre-process data in such a way, before applying the machine learning/system identification methods.

In practice, the specification of the regularization parameter λ plays a critical role. It can be seen that larger λ will shrink more weights to zero. In practice, one can use cross-validation methods to select λ. Namely, on a grid of candidate λ values, estimate the model according to Algorithm 1, find the cross validation errors, and specify in the final modeling the value of λ according to the one leads to the smallest cross validation error.

3. Probabilistic forecast methods for the wind power predictor

3.1. Probabilistic approach for forming the prediction band

In the short-time wind power forecast problem, the sparse regression methods introduced in Section 2 allows a large number of features to be the candidate factors for the prediction model. Therefore it is possible to use a large number of features including the observed wind power of the previous hours, together with the broadcast wind power of the current and the past hours obtained from meteorological/physical methods. Furthermore, one can include into the model not only the observation features belong to the station of prediction interest, but also the observations belong to all the nearby stations. Namely, the model has the following expression:

$$P_t^{[l]} = g(\boldsymbol{X}^{[1]}(t), \ldots, \boldsymbol{X}_{t-1}^{[m]}(t)), \tag{10}$$

in which

$$\boldsymbol{X}^{[k]}(t) = [P_{t-1}^{[k]}, P_{t-2}^{[k]}, \ldots, P_{t-r}^{[k]}, \tilde{P}_t^{[k]}, \tilde{P}_{t-1}^{[k]}, \ldots, \tilde{P}_{t-r}^{[k]}], \tag{11}$$

where $P_t^{[k]}$ is the power output at the hour indexed by t and at the k-th station, $1 \le k \le m$, and m denotes the total number of all the stations nearby, and l indicates the index of the particular station one wants to apply the prediction model. $\tilde{P}_t^{[k]}$ is the wind power based on some kind of pilot physical forecast methods at the t-th

hour and the k-th station. Usually, everyday before at least 4 p.m., the future 24-h values of $\tilde{P}_t^{[k]}$ have already been broadcast, for example, to be used in the clearance of day-ahead energy markets or the reliability unit commitment [4]. The index r denotes only the observations up to the previous r-th hours have been considered in the model. The model in (10) shows the mechanisms of our hour-ahead prediction formulation, if the index t corresponds to the future hour to come, and $t-1$ indicates the most recent past hour.

In many research works recently published in the field of very short-term wind power forecast [6,21,32,47], only the intra-day observations are considered in the model, i.e., they made use of only $P_{t-1}^{[k]}, P_{t-2}^{[k]}, \ldots, P_{t-r}^{[k]}$ but not $\tilde{P}_t^{[k]}, \tilde{P}_{t-1}^{[k]}, \ldots, \tilde{P}_{t-r}^{[k]}$.

Let $\hat{P}_t^{[l]} = \hat{g}(\boldsymbol{X}^{[1]}(t), \ldots, \boldsymbol{X}_{t-1}^{[m]}(t))$ denotes the estimate of $P_t^{[l]}$, where $\hat{g}(\cdot)$ corresponds to the obtained model trained by data using the methodologies described in Section 2. Then let $Z_t^{[l]} = P_t^{[l]} - \hat{P}_t^{[l]}$ represents the residual. By intuition, it can be seen that $Z_t^{[l]}$ is dependent on the value of $\hat{P}_t^{[l]}$. If we obtain a prediction value $\hat{P}_t^{[l]}$, how close it is from the underlying true value can be guessed based on *conditional density* of the term $Z_t^{[l]}$ on $\hat{P}_t^{[l]}$, which can be obtained from the previously observed data. This lays the underlying mechanisms for the probabilistic forecast.

Let the conditional distribution of $Z_t^{[l]}$ on $\hat{P}_t^{[l]}$ be denoted by $f_{Z|\hat{P}}(Z_t^{[l]}|\hat{P}_t^{[l]})$. When the user use our sparse modeling method and obtain the hour-ahead prediction value, he/she can also at the same time guess how close this prediction is close to the unseen future value. For instance, if the user predicted $\hat{P}_t^{[l]}$ to be 0, then he may guess there is a good probability that the true $P_t^{[l]}$ turns out to be also 0. Besides, for instance, if the user finds $\hat{P}_t^{[l]} = p_0$, and he/she somehow has the knowledge that $f_{Z|\hat{P}}(Z_t^{[l]}|\hat{P}_t^{[l]})$ has a large density at $\hat{P}_t^{[l]} = p_0$, then he/she can speculate the prediction $\hat{P}_t^{[l]}$ at this hour should turn out to be very close to the true value. And vice versa. Therefore by studying the conditional distribution of the error term, the user can obtain a *prediction band* rather than a mere point value for the forecast.

Of course, one may argue that $Z_t^{[l]}$ depends not only on $\hat{P}_t^{[t]}$, but also a number of other factors in the entire $(2r+1)m$-dimensional vector defined in (11). However, as the problem becomes very high dimensional, the so-called "curse of dimensionality" phenomenon [46] becomes more imminent, which makes the entire problem unsolvable. On the other hand, with no doubt $\hat{P}_t^{[l]}$ should be the most informative feature among them. Therefore, we simplify the problem into a *bivariate* conditional density problem.

In order to estimate the conditional distribution $f_{Z|\hat{P}}(Z_t^{[l]}|\hat{P}_t^{[l]})$, we are to introduce the nonparametric density estimation methodology in Section 3.2.

3.2. Nonparametric density estimation

With a little of abuse of notation, here suppose $X \in \mathbb{R}^1$ is a 1-dimensional random variable, and $D_n = \{X_1, \ldots, X_n\}$ is an observed sample series of X. The density estimation refers to the recovery of the distribution function $f_X(x)$ from the observation D_n. Here we use continuous random variable as an example, and the discrete ones can be derived by similar means. In classical statistical research, X is usually assumed to follow certain class of distributions, e.g., with prior information, researchers usually assume X to follow one among Gaussian, Uniform, Laplace, Beta, Gaussian mixture model, etc. Then the entire density estimation problem comes down to the issue of estimating one or a few parameters only. For instance, if the observations are assumed to be $N(\mu, \sigma^2)$, then parametric density estimation aims to recover the parameter set (μ, σ).

However, in many cases, the random variable does not underlyingly follow any existing class of distributions. Then the aforementioned parametric estimation processes very large Bayesian error. In other words, no matter how many observations are available to the user, there exists an irreducible discrepancy between the estimated density curve and the underlying true density. This discrepancy is actually the so-called the "modeling error" referring to the difference between the best possible model within the parametric class and the true characteristics which is outside the assumed parametric space.

In order to tackle the aforementioned drawback inherent to the usual parametric density estimation approach, we need to use the so-called nonparametric estimation methodology [48]. Nonparametric estimation and nonparametric regression are modern statistics techniques that are developed in the middle and late 20th century. They have only been more frequently applied in engineering fields recently. In general, nonparametric density estimation includes a group of techniques which do not pre-assume the class of the unknown distribution, and use local methods to recover it. kernel methods, k-nearest neighbors (KNN), orthogonal series estimators, are examples of the nonparametric approach. We refer to Ref. [46] for the detailed description of these techniques.

In this paper, we use the kernel method for density estimation. For a univariate random variable X, the kernel density estimator has the following form

$$\hat{f}_{X;n}(x) = \frac{1}{nh}\sum_{i=1}^{n} \mathbb{K}\left(\frac{x - X_i}{h}\right), \tag{12}$$

where $\hat{f}_{X;n}(x)$ denotes the estimate of the unknown density $f_X(x)$ using observations of length n, $\mathbb{K}(\cdot)$ is called kernel function, and h is a scaling parameter need to be chosen. The applicable kernel function $\mathbb{K}(\cdot)$ should satisfy certain mathematical conditions [46]. One can use, e.g., the Gaussian kernel $\mathbb{K}(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$, the Epanechnikov kernel $\mathbb{K}(x) = \frac{3}{4}(1 - x^2)\mathbb{I}(|x| \le 1)$, etc. Here $\mathbb{I}(\cdot)$ is a logic indicator function, i.e., $\mathbb{I}(x) = 1$ if and only if x is true. We refer to Refs. [36,46,48] for a list of applicable kernels that can be used. In our simulation studies in Section 4, we use the aforementioned Epanechnikov kernel.

The mechanisms of the kernel desnity estimation can be intuitively understood as follows. Take the Epanechnikov kernel for example. For all the available observations $\{X_1, \ldots, X_n\}$, one aims to estimate the density at $X = x$. The kernel function defines a window $(x - h, x + h)$ where only the observations within this window will be counted. The bandwidth parameter h controls how many of the local points are taken into consideration in the estimator in (12). A large h will over-smooth the curve, while a small h will render spikes and jumps in the estimated density curve. Therefore, h should be carefully selected to reflect the balance.

Actually, the asymptotic statistics shows [48].

$$\mathbb{E}[\hat{f}_{X;n}(x) - f_X(x)]^2 = O(n^{-4/5}),$$

where $O(\cdot)$ is the "Big-O notation" in mathematics, if h is asymptotically optimally chosen as

$$h = cn^{-1/5}, \tag{13}$$

where c is a constant. Equation (13) indicates how the optimal

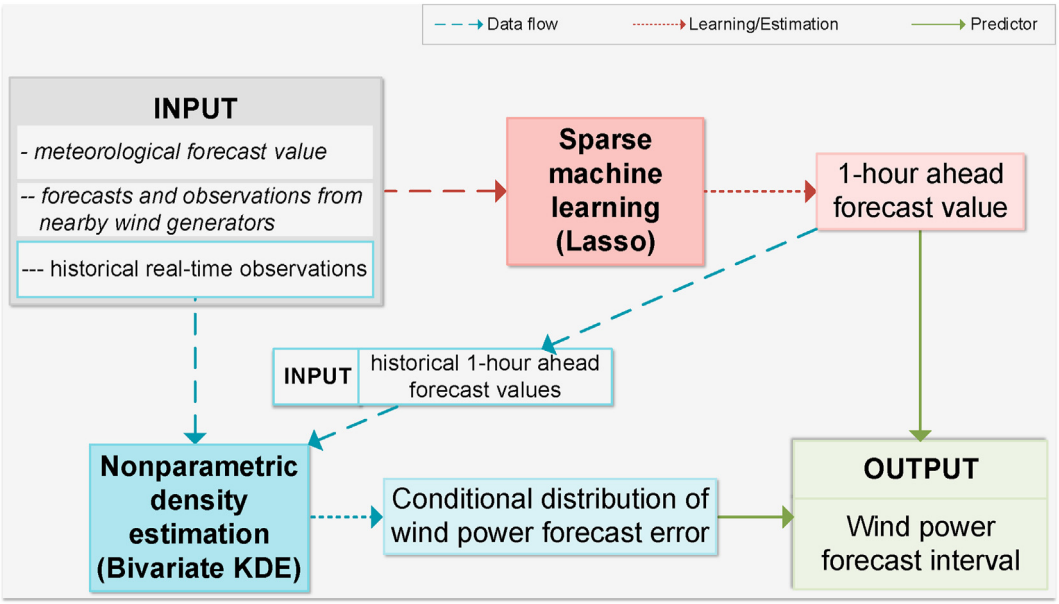

Fig. 2. Framework of the proposed probabilistic wind power prediction method.

parameter h should be selected. However, the constant c in (13) depends on the unknown density $f_X(x)$. Therefore in practice, statisticians use alternative methods that may not be optimal but usually are good enough. For instance, a rule of thumb is the so-called "Silverman's Rule" [48]. For the 1-dimensional density estimation, it takes the following form,

$$h^{1d} = 1.06\hat{\sigma}n^{-1/5},$$

where $\hat{\sigma}$ is the standard deviation of the sample $\{X_1, \ldots, X_n\}$.

The aforementioned 1-dimensional kernel density estimator (12) also has its multi-dimensional counterparts. For instance, if one wants to estimate the joint density of two random variables (X, W) at the point $(X = x, W = w)$, one may use

$$\hat{f}_{XW;n} = \frac{1}{nh_1h_2}\sum_{i=1}^{n}\mathbb{K}\left(\frac{x - X_i}{h_1}\right)\mathbb{K}\left(\frac{w - W_i}{h_2}\right), \tag{14}$$

where h_1, h_2 are the bandwidth parameters for the two dimensions individually. The Silverman's rule in this case becomes

$$h_1^{2d} = 1.06\hat{\sigma}_X n^{-1/6}, h_2^{2d} = 1.06\hat{\sigma}_W n^{-1/6},$$

where $\hat{\sigma}_X$ and $\hat{\sigma}_W$ are the sample standard deviation for X and W respectively. Accordingly the asymptotic convergence rate is

$$\mathbb{E}[\hat{f}_{XW;n}(x, w) - f_{XW}(x, w)]^2 = O(n^{-2/3}).$$

In Section 3.1, it was shown that in our probabilistic prediction problem, we need to estimate the conditional distribution. In actuality, one just needs to estimate the joint distribution of $(Z_t^{[l]}, \hat{P}_t^{[l]})$ described in Section 3.2, together with the marginal distribution of $\hat{P}_t^{[l]}$. By *dividing* the two expressions, we obtain the conditional distribution $Z_t^{[l]}|\hat{P}_t^{[l]}$.

Then, with the statistical knowledge of the error term related to each prediction $\hat{P}_t^{[l]}$, one can construct a prediction band by plotting, e.g., the 25% and 75% quantile of the error term on top of the point-value prediction. The probabilistic forecast result using this approach will be shown in Section 4.

4. Simulation studies

The data we used in our modeling studies are from a new publicly available dataset of the IEEE 118-bus system, called the *NREL-118*. This power network includes a region of wind power generators which include 17 wind generators. Time-synchronous hourly wind time series are available for over one year (366 days), together with the day-ahead forecast values obtained by meteorological methods. The locations of wind generators are chosen such that the correlation of meteorological conditions is reserved [49]. We refer the readers to Ref. [49] and https://item.bettergrids.org/handle/1001/120 for a more detailed description of these data.

The structure of the proposed very short-term probabilistic wind power predictor is revisited, and the input/output and prediction process are summarized herein. As shown in Fig. 2, the meteorological forecast values of the concerned wind generator and all nearby wind generators, together with the historical real-time observations of wind power are fed into the Lasso, which makes a point forecast of the next-hour wind power; the historical real-time observations and very short-term forecasts obtained from the Lasso are taken as input by the bivariate KDE, which produces a conditional distribution of wind power forecast error for the concerned wind generator. The combination of the point forecast and the conditional distribution then gives rise to a probabilistic predictor, which outputs the wind power forecast interval under a certain confidence level.

4.1. Forecast result for wind farm #1

We build a model to predict the next-hour power output at Wind Generator #1. The previous 72 h of all the 17 wind generators' power output are used at the input to the model, together with the current and previously broadcast meteorological forecast values across these wind generators. In this way, the total number of candidate features in the model is $p = 17 \times (72 + 73) = 2465$. The 366 days of total available data is equivalent to 8784 observations.

Table 1
Performance of various techniques to the next-hour wind power prediction problem for Wind Generator #1.The RMSE and MAE are shown as the per unit error.

	RMSE	MAE	F_0 #	F_{final} #
Lasso	0.08678	0.04977	2465	131
LR	0.11671	0.08365	2465	2465
Ridge	0.10173	0.06987	2465	2465
Lasso no Meteorological	0.09195	0.05412	1224	166
Only Meteorological	0.19522	0.13078	n/a	n/a
Time-Series Model	0.10422	0.05238	n/a	n/a
SVR	0.10652	0.07573	2465	2465
RVM	0.11116	0.06435	2465	26
Decision Tree	0.09457	0.05428	2465	2465
LSTM	0.09580	0.05083	n/a	n/a

We use the first $n = 6000$ as the training dataset, and use the remaining ones as the testing set. During our sparse modeling, we use 5-fold cross-validation to select the regularization parameter λ, and in this way we choose $\lambda = 0.00385033$.

We examine our method with another two approaches to the same sets of features, namely, the classical linear regression, and the ridge regression. Apart from that, we also compare with our modeling without the consideration of meteorological observations as input features. We also show the prediction accuracy of just using the meteorological broadcast, as a comparison. It is known in wind forecast that for very short-term forecasts like the 1-h ahead forecast, the time series model (e.g., autoregressive moving average, ARMA) usually renders good performance [4,11,50], therefore we also show the result of this approach as a comparison. Moreover, we compare our approach with some other methods recently applied in the field. These methods include multiple linear regression, ridge regression [32], support vector regression (SVR) [51], relevance vector machine (RVM) [27,28], decision tree method [52], and the long short-term memory method (LSTM) [8,14]. The prediction accuracy for these methods are shown in Table 1. We show both the root mean square error (RMSE) and the mean absolute error (MAE) for the prediction. It is worth noting that both RMSE and MAE have been transferred into per unit level.

Table 1 shows that the Lasso outperforms all the competing methods in terms of these two indexes. The RMSE of the Lasso without meteorological data is most close to that of the Lasso, while the MAE of LSTM is just slightly higher than that of the Lasso. Comparing the Lasso sparse modeling with the mere meteorological prediction approach, one can also confirm that statistical modeling really shows superiority comparing to physical/meteorological approach in the short-term modeling.

Table 1 also shows the total number of candidate features in the model, together with the number of features in the final model. We can see that our Lasso approach for the sparse modeling has eliminated 95% of the entire candidate features. By the automatic dimension reduction, our approach has achieved improved accuracy comparing to the other methods. Among the modeling methods examined in Table 1, the RVM is a sparse modeling based on the Bayesian design. This method starts from the prior distribution of parameters and makes use of empirical data to obtain the posterior distribution of parameters which allows some features to be removed from the model. Table 1 shows that the RVM reduces the dimensionality from 2465 to 26, yet does not achieve the same level of prediction accuracy as our Lasso approach.

Furthermore, if we examine the features that are selected by the Lasso method, we obtain Table 2. Described in Section 2, each dimensions of data have been standardized in the pre-processing stage of the data. Here we target to predict $P_t^{[1]} - \tilde{P}_t^{[1]}$ rather than $P_t^{[1]}$ directly. It is worth mentioning that the meaning of these notations in Table 2 follows the descriptions in Section 2. For instance, the first dominant feature $P_{t-1}^{[1]}$ represents the observations of the wind power at Generator #1 at 1-h before the prediction point; the second dominant feature $\tilde{P}_t^{[1]}$ represents the known meteorologically-foretasted values of the wind power at Generator #1 at the prediction hour; the third dominant feature $P_{t-1}^{[7]}$ represents the observations of the wind power at Generator #7 at 1-h before the prediction point, etc. It is worth mentioning that those 17 features have covered 99.7% of the total strength of all the 131 features selected by the Lasso in Table 1.

From Table 2, we can see that the wind generators at nearby locations also contribute to the accuracy of our model. The first 17 prominent features already include the observations at Generator #7, Generator #2, Generator #12, Generator #11, Generator #6, and Generator #16. These further confirms the usefulness of including a large number of candidate features in the model, as well as the reason for the success of the Lasso method in the wind power prediction problem.

Further, if we aim to forecast $P_t^{[1]}$ rather than $P_t^{[1]} - \tilde{P}_t^{[1]}$, we may not get the weight of $\tilde{P}_t^{[1]}$ as 1–0.9408. Rather, we would obtain another combination of $\tilde{P}_t^{[1]}$ and $\tilde{P}_{t-k}^{[1]}$ similar to the fashion in Table 2.

If we demonstrate weights of those features in Table 2 as well as all the ones selected by the Lasso, we obtain Fig. 3. We also show the weights selected by the ridge regression as a comparison. It is worth mentioning that among the total 2465 features in Fig. 3, the first 145 ones correspond to Generator #1, and the rest correspond to the other 16 generators. Among the first 145 features, the sequence follows $P_{t-1}^{[1]}, P_{t-2}^{[1]}, \cdots, P_{t-72}^{[1]}, \tilde{P}_t^{[1]}, \tilde{P}_t^{[1]}, \cdots, \tilde{P}_t^{[72]}$. From Fig. 3 it can be seen that the ridge regression can not eliminate any candidate features; while the Lasso can automatically set majority of features to zero weight. The desired automatic dimension-reduction property of the L_1 design thus circumvent the "curse of dimensionality" inherent in high-dimensional modeling problems, and renders high prediction accuracy in the wind power prediction problem using our framework.

Table 2
The first 17 dominant features in model solved by Lasso in forecast the next-hour wind power for Wind Generator #1.

	1	2	3	4	5	6
Weight	1.4028	−0.9408	0.2456	−0.2307	−0.1814	−0.1620
Feature	$P_{t-1}^{[1]}$	$\tilde{P}_t^{[1]}$	$P_{t-1}^{[7]}$	$P_{t-2}^{[1]}$	$\tilde{P}_{t-7}^{[1]}$	$P_{t-2}^{[2]}$
	7	8	9	10	11	12
Weight	0.1483	0.1321	0.1287	0.1278	−0.1244	−0.1043
Feature	$P_{t-1}^{[12]}$	$P_{t-1}^{[11]}$	$P_{t-1}^{[6]}$	$P_{t-1}^{[2]}$	$\tilde{P}_{t-1}^{[12]}$	$P_{t-2}^{[7]}$
	13	14	15	16	17	⋯
Weight	−0.0629	−0.0501	0.0316	0.0312	−0.0280	⋯
Feature	$P_{t-2}^{[12]}$	$\tilde{P}_t^{[12]}$	$P_{t-10}^{[11]}$	$P_{t-10}^{[16]}$	$P_{t-6}^{[7]}$	⋯

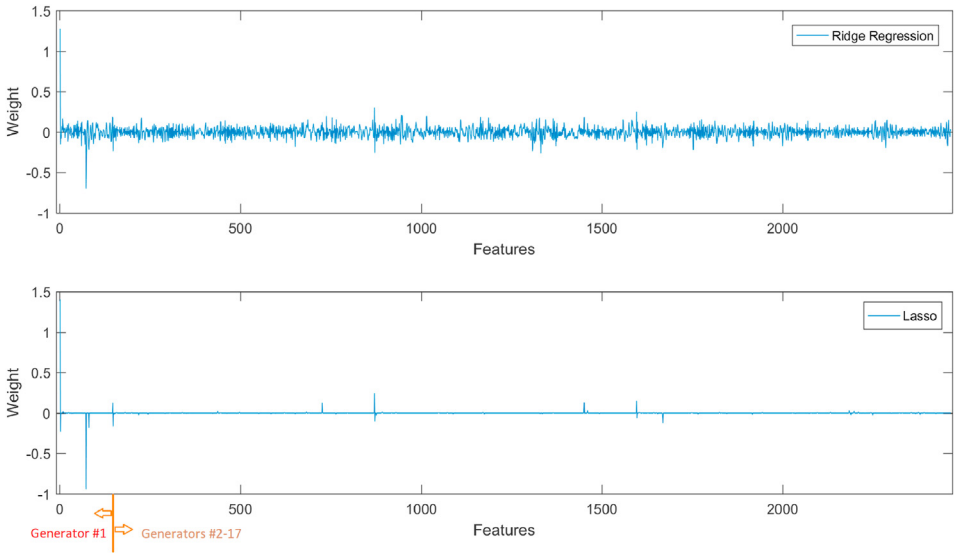

Fig. 3. The weights of the features selected by Lasso compared with the weights selected by ridge regression. The upper panel shows the ridge regression case, and the lower panel shows the Lasso case.

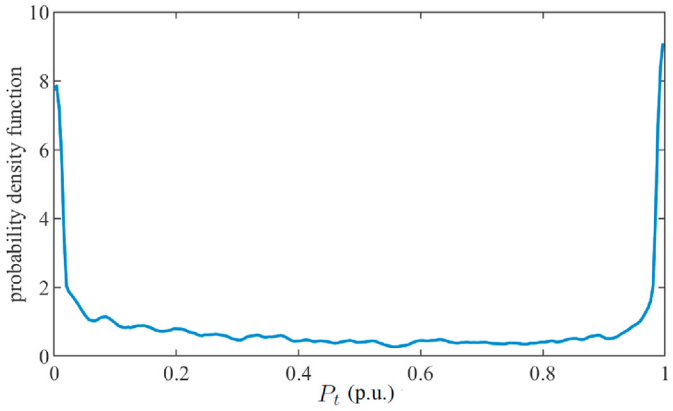

Fig. 4. The estimated density of the power observations in the Wind Farm #1. The RMSE and MAE are shown as the per unit error.

Following the discussions in Section 3.1, we thus use the 2-dimensional nonparametric density estimation technique described in Section 3.2 to examine the joint density of the prediction error term and the meteorological observation value. Similar to the notation in Section 3.1, let P_t denote the real-time wind power observation at the wind generator of our interest, let $\hat{P}_t$ be its predicted value using our sparse modeling approach, the joint density of P_t and $\hat{P}_t$ is estimated and shown in Fig. 5 (a). Alternatively, the contour line for the joint density is shown in Fig. 5 (b). It is worth noting that the univariate version of the kernel density estimation formulates the density for the power observations, which is shown in Fig. 4.

The conditional density estimator of $f_{P|\hat{P}}(P_t|\hat{P}_t)$ and its contour line equivalence are presented in Fig. 6 (a) and Fig. 6 (b).

Fig. 4 shows that the density of power observations are mostly concentrated around 0 p.u. and the maximum value 1 p.u. (the capacity of the wind generator in this case is 6.1972 MW). Fig. 6 (a) and Fig. 6 (b) illustrate that when we observe $\hat{P}_t$ to be close to 0 or 1 p.u., then we can expect the truth is closer to the estimate rather than when we observe $\hat{P}_t$ to be around 0.5 p.u., for instance. In other

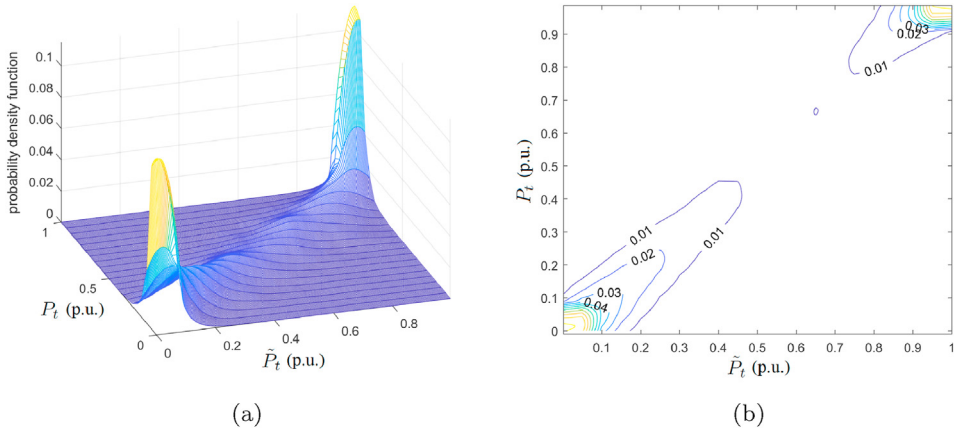

Fig. 5. The estimated (a) *conditional* density and (b) its contour line of the power observations P_t corresponding to Wind Generator #1 and its predicted value $\hat{P}_t$ based on our sparse modeling methodology.

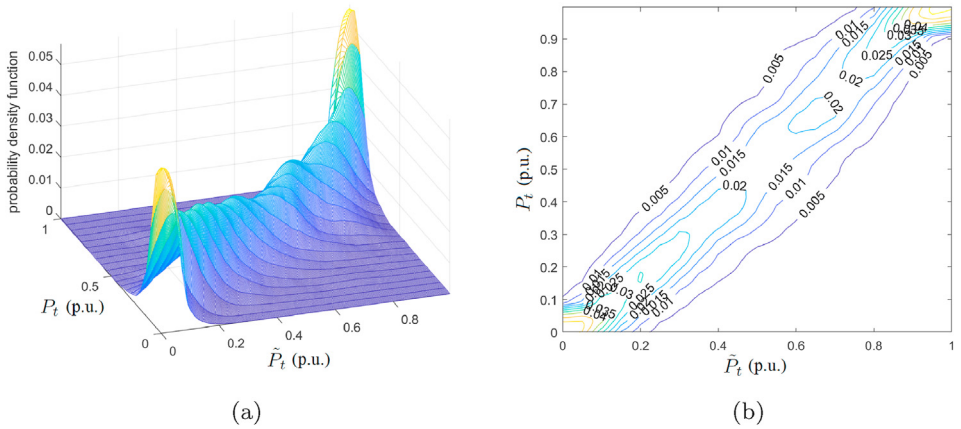

Fig. 6. The estimated (a) *joint* density and (b) its contour line of the power observations P_t corresponding to Wind Generator #1 and its predicted value $\hat{P}_t$ based on our sparse modeling methodology.

words, the former case correspond to narrower prediction band.

Alternatively, if we plot the conditional density $f_{P|\hat{P}}\left(P_t - \hat{P}_t \middle| \hat{P}_t\right)$ when different values of $\hat{P}_t$ are observed, we obtain Fig. 7. This shows the distribution of prediction error conditioned on the prediction itself. It is worth noting that Fig. 7 can be viewed to reflect the same information on different slices of the 3D plot in Fig. 6 (a).

For the first 200 values in the testing set, we plot our modeling result with the prediction value and the 25%–75% confidence band, compared with the true value. We also show the meteorological forecast values as a comparison. The result is shown in Fig. 8.

From Fig. 8, we can see that our machine learning results are relatively close to the underlying true values, while the meteorological forecast approach is significantly inferior in the 1-h prediction. The prediction band also gives an indicator of the confidence interval of our predictions.

For building the point-prediction model, it takes approximately 21.5 min in Matlab using our sparse methodology on an i7-7700HQ

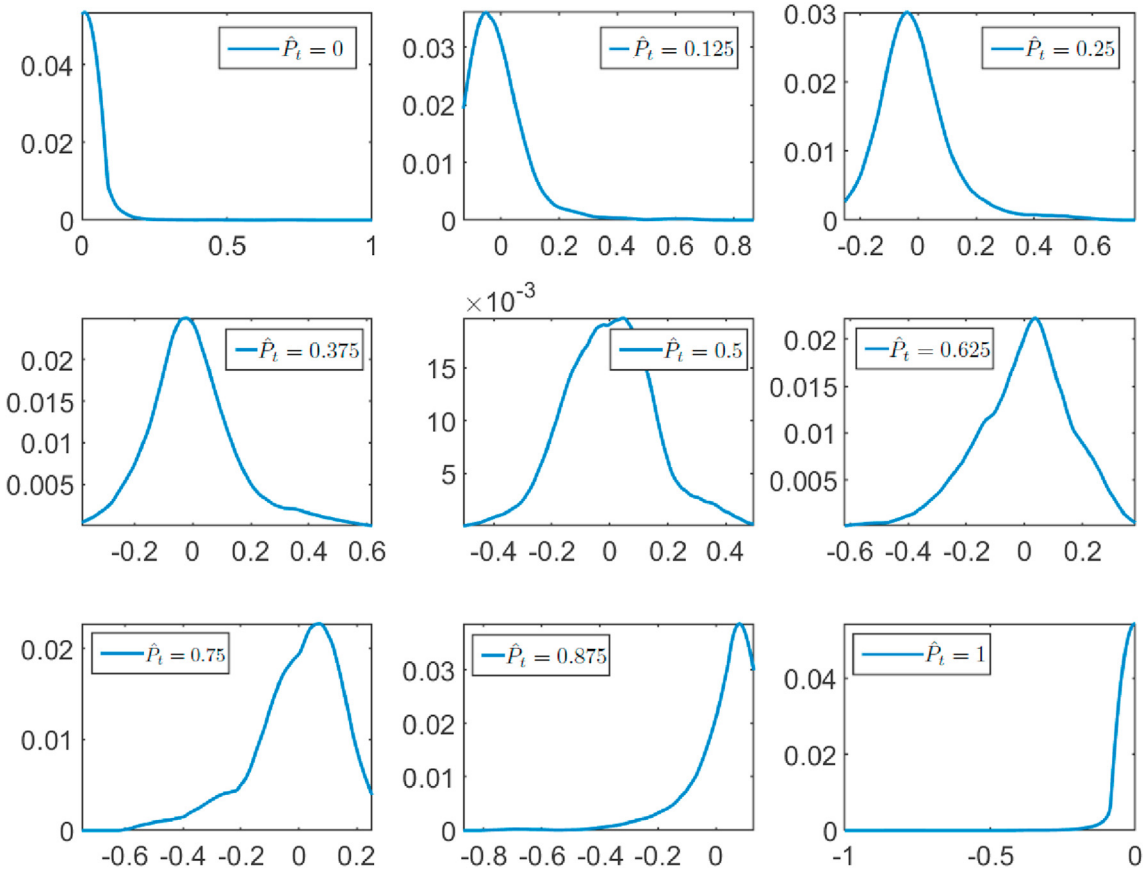

Fig. 7. The probability density of $P_t - \hat{P}_t$ corresponding to Wind Generator #1 when different values of $\hat{P}_t$ have been observed. Note that the vertical axis corresponds to the probability density, and the horizontal axis corresponds to the value of $P_t - \hat{P}_t$. The variables here are shown in the per unit level.

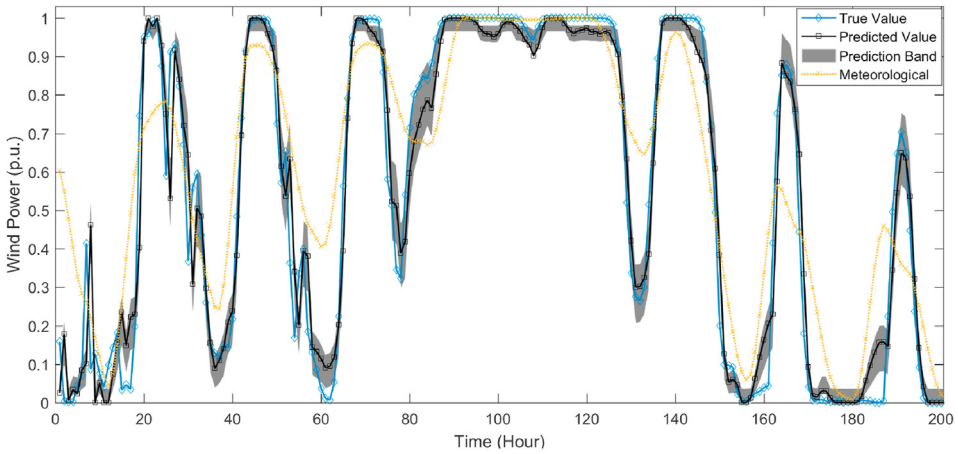

Fig. 8. The probabilistic forecast result for the first 200 data in the testing set representing Wind Generator #1 using our Lasso method equipped with the kernel density nonparametric estimation approach. The 25%–75% confidence interval of the prediction are shown.

CPU computer. For the probabilistic forecast part, it takes only 80 s. Once the model is built, it takes almost no time for make a new forecast when new observations are available. Therefore the user can make an hour-ahead forecast at the current moment immediately.

According to Ref. [16], apart from our approach, the quantile regression (robust estimate) is another way to extend existing point estimators to probabilistic estimators. However, according to the design of the Lasso, the high dimensional set-up makes the "robust Lasso" (Lasso quantile regression) very difficult in the algorithms level. Here, we also implement the quantile regression according to the algorithms in Ref. [33] using the same set of features used in our Lasso. This algorithms has already been developed as an R-language package by the same group of mathematicians/statisticians. For the first 200 values in the testing set, we also plot our probabilistic forecast using this quantile approach, and the result is shown in Fig. 9.

It takes 4.7 h for our computer to perform the quantile regression in R. It is worth to mention that R is a faster software compared to Matlab. The reason is probably that working on those large number of features is time-consuming for the majority of statistic methods, except Lasso which is designed to target the large number of features. It is also worth to mention that the RMSE and MAE for the quartile regression are 0.10032 p.u. and 0.06491 p.u. respectively. The 25%–75% band for our Lasso probabilistic modeling approach covers 57.7% of the true values in the testing set, while this number is by contrast 39.7% for the quantile regression. These further confirms the usefulness of the bivariate kernel density approach equipped with Lasso for the probabilistic forecast.

4.2. Robustness of the lasso forecast methodology

4.2.1. Forecast result for the other wind generators

In Section 4.1 we have studied the probabilistic forecast of one wind generator using our approach. To show that our sparse modeling methodology also works well for other wind generators, we firstly examine the prediction of wind power in Wind Generator #2 using a manner similar to the work in Table 1. The result is shown in Table 3.

Wind Generator #2 and Wind Generator #1 have different level of wind powers. Nonetheless Table 3 shows our sparse modeling approach demonstrate superior prediction performance.

Furthermore, if we conduct the modeling for all the 17 Wind Generators in the power network, and if we set the performance of our proposed method as unit error (for both RMSE and MSE cases), then we obtain Table 4. Table 4 confirms the superiority of our approach for all the 17 wind generators. While Table 4 only shows the average performance, it is also worth mentioning that our

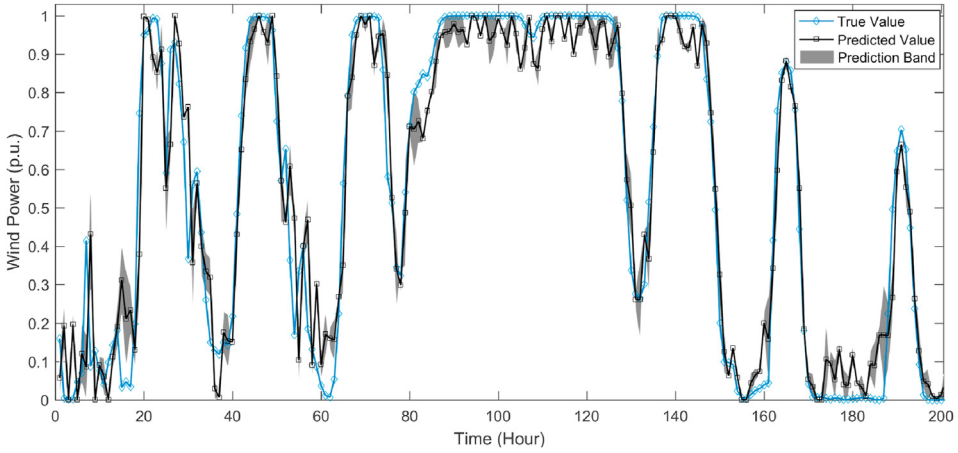

Fig. 9. The probabilistic forecast result for the first 200 data in the testing set representing Wind Generator #1 using the quantile regression as in Ref. [33]. The 25%–75% confidence interval of the prediction is shown.

sparse approach also beats all the other methods in all the individual 17 wind generator power predictions. Since the 17 wind generators are selected from several geographically dispersed wind farms [49] instead of a single wind farm, the test results across these wind generators have shown the robustness of the proposed method against geographical diversity.

Table 3
Performance of various techniques to the next-hour wind power prediction problem for Wind Generator #2. The RMSE and MAE are shown as the per unit error.

	RMSE	MAE	F_0 #	F_{final} #
Lasso	0.08810	0.04779	2465	116
LR	0.12407	0.08917	2465	2465
Ridge	0.10361	0.07205	2465	2465
Lasso no Meteorological	0.09370	0.05384	1224	101
Only Meteorological	0.17643	0.11297	n/a	n/a
Time-Series Model	0.10536	0.05037	n/a	n/a
SVR	0.11126	0.08033	2465	2465
RVM	0.10546	0.05953	2465	25
Decision Trees	0.09654	0.05403	2465	2465
LSTM	0.09985	0.05969	n/a	n/a

Table 4
Performance of various techniques to the next-hour wind power prediction problem for all the 17 wind generators on average. The RMSE and MAE are shown as the per unit error.

	RMSE	MAE
Lasso	0.08346	0.04633
LR	0.11307	0.08056
Ridge	0.09770	0.06670
Lasso no Meteorological	0.08919	0.05147
Only Meteorological	0.17832	0.11486
Time-Series Model	0.10334	0.05166
SVR	0.10185	0.07093
RVM	0.10635	0.05969
Decision Trees	0.09471	0.05470
LSTM	0.09378	0.05406

Table 5
Prediction accuracy across four trimesters. The RMSE and MAE are shown as the per unit error.

		Q1	Q2	Q3	Q4
Gen. #1	RMSE	0.10590	0.09486	0.07962	0.08098
	MAE	0.06017	0.05553	0.04901	0.04359
Gen. #2	RMSE	0.10226	0.10725	0.10099	0.07849
	MAE	0.05914	0.06786	0.06335	0.04081
⋮		⋮	⋮	⋮	⋮
All 17 Gens on average	RMSE	0.09626	0.09251	0.08002	0.07835
	MAE	0.05525	0.05839	0.04818	0.04189

Table 6
Prediction accuracy for the up to 4-h-ahead forecast using the Lasso methodology framework. The RMSE and MAE are shown as the per unit error.

		1-h ahead	2-h ahead	3-h ahead	4-h ahead
Gen. #1	RMSE	0.08678	0.10865	0.13975	0.16689
	MAE	0.04977	0.07042	0.09253	0.11162
Gen. #2	RMSE	0.08810	0.10601	0.13712	0.16625
	MAE	0.04779	0.06383	0.08480	0.10372
⋮		⋮	⋮	⋮	⋮
All 17 Gens on average	RMSE	0.08346	0.10171	0.12975	0.15445
	MAE	0.04633	0.06513	0.08596	0.10322

4.2.2. Robustness of the forecast considering the season factor

It is known that the wind power can exhibit slight variations in characteristics for different months/seasons of the year. If we train the model using the Lasso methodology described above considering the whole years data, and examine the prediction accuracy across the four quarters of the year, we obtain the results in Table 5.

From Table 5, we can see that our sparse modeling methodology works best for the third and fourth trimester of the year. There are slight increase of prediction errors for the other two trimesters.

4.2.3. Performance of the forecast for hours ahead

The aforementioned forecast framework has been focused on the 1-h ahead wind power prediction. Since the meteorological forecast usually makes available the future 4-h' values [2], our Lasso can be used to predict up to 4-h ahead for the wind power. Namely, the prediction of wind power for the 4-h ahead considers the meteorological values of the future 4 h all together with the features used in the aforementioned set-up.

The performance of forecast using the Lasso for up to the 4-h ahead prediction is shown in Table 6. From this table, we can see that the forecast accuracy decay a little bit for predicting longer number of hours in the future.

5. Concluding remarks

In this paper, we have used combined machine learning techniques for the probabilistic forecast of the very short-term wind power. We use a large number of candidate features to take into account the spatio-temporal correlation of wind power, and use the L_1 sparse modeling approach to obtain a parsimonious model. The prediction performance of such sparse modeling technique shows superiority compared to several competitive techniques. We then use the nonparametric conditional density estimation technique to build a confidence band for the sparse modeling. The resulting probabilistic forecast method shows advantages over quantile regression in terms of computational efficiency and prediction accuracy.

CRediT authorship contribution statement

Jiaqing Lv: Conceptualization, Methodology, Simulation, Validation, Writing – original draft, preparationand revision. **Xiaodong Zheng:** Conceptualization, Methodology, Data curation, Writing – original draft, Writing – review & editing. **Mirosław Pawlak:** Supervision, Writing – review & editing. **Weike Mo:** Visualization. **Marek Miśkowicz:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The work of Jiaqing Lv was supported by the Polish National Center of Science under Grant DEC-2018/31/N/ST7/03977. The work of Xiaodong Zheng and Weike Mo was supported by the National Natural Science Foundation of China under Grant 51937005. The work of Mirosław Pawlak was supported by the Polish National Center of Science under grant DEC-2017/27/B/ST7/03082. The work of Marek Miśkowicz was supported by the Polish National Center of Science under grant DEC-2018/31/B/ST7/03874.

References

[1] P. Pinson, L. Mitridati, C. Ordoudis, J. Ostergaard, Towards fully renewable energy systems: experience and trends in Denmark, CSEE journal of power and energy systems 3 (1) (2017) 26–35.

[2] H. Chen, P. Xuan, Y. Wang, K. Tan, X. Jin, Key technologies for integration of multitype renewable energy sources—research on multi-timeframe robust scheduling/dispatch, IEEE Transactions on Smart Grid 7 (1) (2015) 471–480.

[3] A. Sajadi, L. Strezoski, V. Strezoski, M. Prica, K.A. Loparo, Integration of renewable energy systems and challenges for dynamics, control, and automation of electrical power systems, Wiley Interdisciplinary Reviews: Energy Environ. 8 (1) (2019) e321.

[4] C. Monteiro, R. Bessa, V. Miranda, A. Botterud, J. Wang, G. Conzelmann, et al., Wind Power Forecasting: State-Of-The-Art 2009, Argonne National Lab.(ANL), Argonne, IL (United States), 2009. Tech. rep.

[5] C. Silva, R. Bessa, E. Pequeno, S. Jean, M. Vladimiro, Z. Zhou, A. Botterud, Dynamic Factor Graphs-A New Wind Power Forecasting Approach, Argonne National Lab.(ANL), Argonne, IL (United States), 2014. Tech. rep.

[6] A. Lahouar, J. Ben Hadj Slama, Hour-ahead wind power forecast based on random forests, Renew. Energy 109 (2017) 529–541.

[7] Y. Wu, Q. Wu, J. Zhu, Data-driven wind speed forecasting using deep feature extraction and LSTM, IET Renew. Power Gener. 13 (12) (2019) 2062–2069.

[8] X. Liu, H. Zhang, X. Kong, K.Y. Lee, Wind speed forecasting using deep neural network with feature selection, Neurocomputing 397 (2020) 393–403.

[9] Y. Hao, L. Dong, X. Liao, J. Liang, L. Wang, B. Wang, A novel clustering algorithm based on mathematical morphology for wind power generation prediction, Renew. Energy 136 (2019) 572–585.

[10] C.M. Alaíz, Barbero, J.R. Dorronsoro, Sparse methods for wind energy prediction, in: The 2012 International Joint Conference on Neural Networks, IJCNN), 2012, pp. 1–7.

[11] L. Cavalcante, R.J. Bessa, M. Reis, J. Browell, Lasso vector autoregression structures for very short-term wind power forecasting, Wind Energy 20 (4) (2017) 657–675.

[12] J.W. Messner, P. Pinson, Online adaptive lasso estimation in vector autoregressive models for high dimensional wind power forecasting, Int. J. Forecast. 35 (4) (2019) 1485–1498.

[13] N. Tang, S. Mao, Y. Wang, R.M. Nelms, Solar power generation forecasting with a lasso-based approach, IEEE Internet of Things Journal 5 (2) (2018) 1090–1099.

[14] Y. Wang, Y. Shen, S. Mao, X. Chen, H. Zou, Lasso and lstm integrated temporal model for short-term solar intensity forecasting, IEEE Internet of Things Journal 6 (2) (2019) 2933–2944.

[15] P. Pinson, G. Kariniotakis, Conditional prediction intervals of wind power generation, IEEE Trans. Power Syst. 25 (4) (2010) 1845–1856.

[16] Y. Zhang, J. Wang, X. Wang, Review on probabilistic forecasting of wind power generation, Renew. Sustain. Energy Rev. 32 (2014) 255–270.

[17] C. Zhang, H. Chen, Z. Liang, W. Mo, X. Zheng, D. Hua, Interval voltage control method for transmission systems considering interval uncertainties of renewable power generation and load demand, IET Generation, Transm. Distrib. 12 (17) (2018) 4016–4025.

[18] F. Golestaneh, P. Pinson, H.B. Gooi, Polyhedral predictive regions for power system applications, IEEE Trans. Power Syst. 34 (1) (2019) 693–704.

[19] H. Bludszuweit, J.A. Dominguez-Navarro, A. Llombart, Statistical analysis of wind power forecast error, IEEE Trans. Power Syst. 23 (3) (2008) 983–991.

[20] P. Pinson, Very-short-term probabilistic forecasting of wind power with generalized logit–normal distributions, J. Roy. Stat. Soc.: Series C (Applied Statistics) 61 (4) (2012) 555–576.

[21] J. Dowell, P. Pinson, Very-short-term probabilistic wind power forecasts by sparse vector autoregression, IEEE Transactions on Smart Grid 7 (2) (2015) 763–770.

[22] J. Hu, J. Wang, L. Xiao, A hybrid approach based on the Gaussian process with t-observation model for short-term wind speed forecasts, Renew. Energy 114 (2017) 670–685.

[23] M. Sun, C. Feng, J. Zhang, Multi-distribution ensemble probabilistic wind power forecasting, Renew. Energy 148 (2020) 135–149.

[24] A. Staid, J.-P. Watson, R.J.B. Wets, D.L. Woodruff, Generating short-term probabilistic wind power scenarios via nonparametric forecast error density estimators, Wind Energy 20 (12) (2017) 1911–1925.

[25] J. Ma, M. Yang, Y. Lin, Ultra-short-term probabilistic wind turbine power forecast based on empirical dynamic modeling, IEEE Transactions on Sustainable Energy 11 (2) (2020) 906–915.

[26] Y. Zhang, Y. Zhao, G. Pan, J. Zhang, Wind speed interval prediction based on lorenz disturbance distribution, IEEE Transactions on Sustainable Energy 11 (2) (2020) 807–816.

[27] J. Yan, Y. Liu, S. Han, M. Qiu, Wind power grouping forecasts and its uncertainty analysis using optimized relevance vector machine, Renew. Sustain. Energy Rev. 27 (2013) 613–621.

[28] F. Shahid, A. Khan, A. Zameer, J. Arshad, K. Safdar, Wind power prediction using a three stage genetic ensemble and auxiliary predictor, Appl. Soft Comput. 90 (2020) 106151.

[29] C. Feng, M. Cui, B.-M. Hodge, J. Zhang, A data-driven multi-model methodology with deep feature selection for short-term wind forecasting, Appl. Energy 190 (2017) 1245–1257.

[30] T. Mahmoud, Z.Y. Dong, J. Ma, An advanced approach for optimal wind power generation prediction intervals by using self-adaptive evolutionary extreme learning machine, Renew. Energy 126 (2018) 254–269.

[31] L. Cai, J. Gu, J. Ma, Z. Jin, Probabilistic wind power forecasting approach via instance-based transfer learning embedded gradient boosting decision trees, Energies 12 (1) (2019) 159.

[32] J. Naik, R. Bisoi, P.K. Dash, Prediction interval forecasting of wind speed and wind power using modes decomposition based low rank multi-kernel ridge regression, Renew. Energy 129 (2018) 357–383.

[33] R. Koenker, K.F. Hallock, Quantile regression, J. Econ. Perspect. 15 (4) (2001) 143–156.

[34] J. Hu, J. Tang, Y. Lin, A novel wind power probabilistic forecasting approach based on joint quantile regression and multi-objective optimization, Renew. Energy 149 (2020) 141–164.

[35] J.A.G. Ordiano, L. Groell, R. Mikut, V. Hagenmeyer, Probabilistic energy forecasting using the nearest neighbors quantile filter and quantile regression, Int. J. Forecast. 36 (2) (2020) 310–323.

[36] A. Gramacki, The curse of dimensionality, in: *In* Nonparametric Kernel Density Estimation and its Computational Aspects, Springer, 2018.

[37] Z. Wang, W. Wang, C. Liu, Z. Wang, Y. Hou, Probabilistic forecast for multiple wind farms based on regular vine copulas, IEEE Trans. Power Syst. 33 (1) (2018) 578–589.

[38] T. Gneiting, K. Larson, K. Westrick, M.G. Genton, E. Aldrich, Calibrated probabilistic forecasting at the stateline wind energy center: the regime-switching space–time method, J. Am. Stat. Assoc. 101 (475) (2006) 968–979.

[39] R.J. Bessa, V. Miranda, A. Botterud, J. Wang, E.M. Constantinescu, Time adaptive conditional kernel density estimation for wind power forecasting, IEEE Transactions on Sustainable Energy 3 (4) (2012) 660–669.

[40] B. Khorramdel, H. Khorramdel, A. Zare, N. Safari, H. Sangrody, C. Chung, A nonparametric probability distribution model for short-term wind power prediction error, in: IEEE Canadian Conference on Electrical & Computer Engineering (CCECE), IEEE, 2018, pp. 1–5, 2018.

[41] B. Khorramdel, C.Y. Chung, N. Safari, G.C.D. Price, A fuzzy adaptive probabilistic wind power prediction framework using diffusion kernel density estimators, IEEE Trans. Power Syst. 33 (6) (2018) 7109–7121.

[42] G.A. Seber, A.J. Lee, Linear Regression Analysis, John Wiley & Sons, 2012.

[43] A.E. Hoerl, R.W. Kennard, Ridge regression: applications to nonorthogonal problems, Technometrics 12 (1) (1970) 69–82.

[44] B. Jayasekara, U.D. Annakkage, Derivation of an accurate polynomial representation of the transient stability boundary, IEEE Trans. Power Syst. 21 (4) (2006) 1856–1863.

[45] R. Tibshirani, Regression Shrinkage and Selection via the Lasso, Journal of the Royal Statistical Society. Series B (Methodological), 1996, pp. 267–288.

[46] W. Greblicki, M. Pawlak, Nonparametric System Identification, vol. 1, Cambridge University Press, Cambridge, 2008.

[47] N. Safari, S.M. Mazhari, C.Y. Chung, Very short-term wind power prediction interval framework via bi-level optimization and novel convex cost function, IEEE Trans. Power Syst. 34 (2) (2019) 1289–1300.

[48] W. Härdle, et al., Smoothing Techniques: with Implementation in S, Springer Science & Business Media, 1991.

[49] I. Peña, C.B. Martinez-Anido, B. Hodge, An extended IEEE 118-bus test system with high renewable penetration, IEEE Trans. Power Syst. 33 (1) (2018) 281–289.

[50] P. Gomes, R. Castro, Wind speed and wind power forecasting using statistical models: autoregressive moving average (ARMA) and artificial neural networks (ANN), International Journal of Sustainable Energy Development 1 (1/2).

[51] Q. Hu, S. Zhang, M. Yu, Z. Xie, Short-term wind speed or power forecasting with heteroscedastic support vector regression, IEEE Transactions on Sustainable Energy 7 (1) (2015) 241–249.

[52] G.I. Nagy, G. Barta, S. Kazi, G. Borbély, G. Simon, Gefcom2014: probabilistic solar and wind power forecasting using a generalized additive tree ensemble approach, Int. J. Forecast. 32 (3) (2016) 1087–1093.

[53] J.W. Taylor, J. Jeon, Forecasting wind power quantiles using conditional 683 kernel estimation, Renew. Energy 80 (2015) 370–379.

Further reading

[1] X. Zheng, K. Qu, J. Lv, Z. Li, B. Zeng, Addressing the conditional and correlated wind power forecast errors in unit commitment by distributionally robust optimization, IEEE Trans. Sustain. Energy 12 (2) (2021) 944–954.