Multidimensional Wind Power Correlation Analysis and Modeling Based on Pair Copula

Peng Wei, Jiankun Liu, Qian Zhou and Xinyao Zhu

ABSTRACT

The introduction of large-scale wind farms makes the analysis of correlation more complicated. It is significant to describe the randomness and correlation of multiple wind farms and to analyze the influence of wind power on the operation of power system accurately. The correlation of two elements can be described accurately with current methods, such as Copula. But they are not accurate enough for higher-dimensional model. Hence, a high-dimensional wind power correlation model based on canonical-vine Pair Copula theory and a corresponding sampling method are proposed. Pair Copula can describe the different correlation structure of wind power, which can reflect the complex multidimensional correlation better. Moreover, the modeling procedure is simple, flexible and applicable. The modeling and analysis of the wind farm output samples in Australia confirm the superiority of the proposed method. Finally, an example of the IEEE 118 bus system is used to illustrate the characterization of the wind power correlation, which can analyze the operating characteristics of the power system more accurately with wind power access.

Keywords: Pair Copula; Multi-dimensional correlation; wind; probabilistic load flow.

INTRODUCTION

By the end of 2013, the installed capacity of wind power in China was 91.4 GW [1], and a large number of wind farms were connected to the power system. However, due to the inherent randomness and volatility of natural factors (wind speed), wind power output shows strong intermittency. In addition, with the increase of the wind farms, the correlation between wind farm outputs is becoming increasingly complex [2]. Therefore, it is necessary to consider the probability characteristics of the wind power output when analyzing the wind power system with multiple wind farms [3].

In order to solve the problem of wind power correlation, a third-order polynomial normal transformation method is proposed in the probabilistic load flow calculation in [4]. A ninth-order polynomial normal transformation is proposed in [5]. But these two methods are based on linear correlation coefficient modeling, not suitable for describing nonlinear relations. The global correlation is described by the rank correlation coefficient in [6], but it can’t reflect all the relevant properties, such as tail correlation.
Copula model can be used to describe the nonlinearity, asymmetry and tail correlation of two random variables, which is widely used in power system, such as probabilistic load flow [7], optimal load flow [8], risk assessment [9] and assessment on available transfer capability [10],[11]. The Nataf transformation proposed in [12] is actually a special case of the Copula function, but the computation is more complicated. A two-dimensional hybrid Copula function is proposed to further improve modeling accuracy in [13]. For higher-dimensional random variables, only a few Copula functions are available [14], and random variables can only be modeled using Copula functions of the same type. However, with the increase in access to wind farms, there are complex spatial position correlations among the large number of wind farms, and the correlations between them are often different. Different types of two-dimensional Copula functions should be used to describe their relative characteristics. It is suggested in [15] that the wind power can be divided into two groups, which can describe the correlation of each group well, but the criterion is not clear. Therefore, the existing copula function is difficult to accurately model multidimensional wind power [16].

Aiming at the above problems, this paper presents a multidimensional correlation model of wind power based on C-Pair Copula and a quasi-Monte Carlo (QMC) sampling method [17]. This paper analyzes the characteristics of multiple wind farm output samples in Australia and applies them to the probabilistic load flow calculation of IEEE 118 bus system, which proves the validity and superiority of the proposed method.

INTRODUCTION TO COPULA THEORY

Copula function is used to model the edge distribution and dependency structure of random variables separately, so as to construct the joint probability distribution. Let F be the joint probability distribution function of random variable \( X = [x_1, x_2, \ldots, x_n] \) with edge distribution \( F_1, F_2, \ldots, F_n \). Sklar theorem states that there exists a Copula probability distribution function \( C(\cdot) \) for any \( X \in \mathbb{R}^n \):

\[
F(x_1, x_2, \ldots, x_n) = C(F(x_1), F(x_2), \ldots, F(x_n))
\]  

Commonly used Copula functions include: Gumbel-Copula, Clayton-Copula, Frank-Copula, Normal-Copula and t-Copula. These five functions can be used to model various binary correlation characteristics.

PAIR COPULA MULTIDIMENSIONAL CORRELATION MODELING

In order to solve the problem of complicated wind power correlation structure in multi-wind farms, this paper proposes to use Pair Copula to model the correlation of wind power output.

Pair Copula Structure

Pair Copula was first proposed by K. Aas and so on [18]. C vine is a construction in Pair Copula. The following C rattan Pair Copula abbreviated as Pair Copula. N-dimensional Pair Copula has \( n-1 \) layers, denoted as \( T_i(i = 1, 2, \ldots, n-1) \), each layer has a root node connected with the rest of the nodes, by building a multi-level combination of probability distribution. Here, each node is a binary Copula function. Let \( u_i = F_i(x_i) \), n-dimensional Pair copula structure is shown in Figure 1.
In Figure 1, $c_{1,1}$ is the abbreviation of $c_{1,1}(u_{1i},u_{1i})$ $(i = 1, 2, \ldots, n-1)$ in layer $T_1$. In the remaining layers, $c_{j,j+1,\ldots,j+1}(F(u_j | u_1, u_2, \ldots, u_{j-1}), F(u_{j+1} | u_1, u_2, \ldots, u_{j+1}))$ $(j = 2, 3, \ldots, n-1, i = 1, 2, \ldots, n - j)$. It should be noted here that $c(\cdot)$ denotes the Copula probability density function and the relationship between $c(\cdot)$ and $C(\cdot)$ is shown in (2).

\[
c(u_1, u_2, \ldots, u_n) = \frac{\partial C(u_1, u_2, \ldots, u_n)}{\partial u_1 \partial u_2 \cdots \partial u_n}
\]

(2)

The joint probability density $f(x_1, x_2, \ldots, x_n)$ is

\[
f(x_1, x_2, \ldots, x_n) = c(F_1(x_1), F_2(x_2), \ldots, F_n(x_n)) \prod_{i=1}^{n} f_i(x_i) = c(u_1, u_2, \ldots, u_n) \prod_{i=1}^{n} f_i(x_i)
\]

(3)

In (3), $f_i(x_i)$ is the probability density function of $x_i$.

According to Pair Copula theory:

\[
c(u_1, u_2, \ldots, u_n) = (\prod_{i=1}^{n-1} c_{1,1}(u_{1i}, u_{1i})) \prod_{j=1}^{n} \prod_{i=1}^{n-j} c_{j,j+1,\ldots,j+1}(F(u_j | u_1, u_2, \ldots, u_{j-1}), F(u_{j+1} | u_1, u_2, \ldots, u_{j+1}))
\]

(4)

The conditional distribution in (4) has the following properties:

\[
F(u_{j+1} | u_j, u_2, \ldots, u_{j-1}) = \frac{\partial C_{j+1,2}(u_{j+1}, u_{j+1})}{\partial u_{j+1}} \quad i = 0, 1, \ldots, n - 2
\]

(5)

\[
F(u_{j+1} | u_j, u_2, \ldots, u_{j-1}) = \frac{1}{\partial F(u_{j+1} | u_j, u_2, \ldots, u_{j-2})} \cdot \partial C_{j+1, j+2, \ldots, j+2} \quad i = 0, 1, \ldots, n - j, j = 3, 4, \ldots, n
\]

(6)

Pair Copula combines the random variables into two corresponding Copula functions, which can introduce many kinds of copula functions and improve the fitting precision of the probability model.

**Goodness of Fit Test**

In order to judge the probability model of the original sample data fitting effect, the goodness of fit test is required. The Euclidean distance [8] and $K(z)$ [18] are used for this test.

Euclidean distance is the sum of the square of the difference between the actual and empirical values of the empirical distribution. Let $(x_{1i}, x_{2i}, \ldots, x_{mi}) (i = 1, 2, \ldots, m)$ be the historical data sample of the n-dimensional stochastic quantity X. m is the total number of historical data samples. The Euclidean distance of the Copula or Pair Copula function C is fitted by (7).
\[ d^2 = \sum_{i=1}^{m} |C_m(F_i(x_u), F_2(x_u), \ldots, F_n(x_u)) - C(F_i(x_u), F_2(x_u), \ldots, F_n(x_u))|^2 \]

\[ C_m(\cdot) \text{ is the actual probability value at the sample point and can be obtained from the historical data.} \]

The smaller the Euclidean distance, the better the probabilistic model is. In addition, in order to directly reflect the goodness of fit, the \( K(z) \) test is used as shown in (8). The closer the empirical distribution is to the \( K(z) \) of Copula, the better the fitting is.

\[ K(z) = \Pr(C(F_i(x_u), F_2(x_u), \ldots, F_n(x_u)) \leq z) \] (8)

**Modeling Process**

Based on the structure of Pair Copula and the test of fitting degree, the concrete steps of determining the optimal Pair Copula function are given as follows.

**Step 1:** Obtain the historical data of wind power \( X \) to get the corresponding probability distribution function and probability density function. Let \( u_i = F_i(x_i) \), obtain the sample data of \( U \) composed of \( u_i \).

**Step 2:** Based on the \( U \)-sample data, the Copula function is used to fit the first-level Copula sequence \( \{c_{1,2}, c_{1,3}, \ldots, c_{1,n}\} \), and the logarithm maximum likelihood estimation is used to get the corresponding parameters. According to Euclidean distance and \( K(z) \) test of (7), the optimal Copula sequence is selected.

**Step 3:** The conditional distribution shown in (5) is taken as a new random variable, and its samples are obtained by using the samples of \( U \). Based on the sample of the conditional distribution in (5), similar to Step 2, the second layer of Copula sequence \( \{c_{2,3,1}, c_{2,4,1}, \ldots, c_{2,n,1}\} \) is established.

**Step 4:** Using the samples of the previous conditional distribution and (6) to get the sample of the conditional distribution of this layer, similar to the above steps, we get the optimal Copula sequence up to the last layer and complete the modeling.

**Pair Copula Sampling Method**

The steps of Pair Copula sampling method are as follows.

**Step 1:** Use QMC [19] to generate random numbers with independent uniform distribution \( z_j, (j = 1, 2, \ldots, n) \). QMC generates pseudo-random numbers, which are superior to Monte Carlo and Latin hypercube in terms of convergence rate and sampling efficiency.

**Step 2:** For each sampling point of \( z_j, (j = 1, 2, \ldots, n) \), the corresponding random variable \( u \) samples are obtained step by step by iteratively solving equation (9) using (5) and (6).

\[
\begin{align*}
\hat{z}_i &= u_i \\
\hat{z}_j &= F(u_j | u_1, u_2, \ldots, u_{j-1}) \quad j = 2, 3, \ldots, n
\end{align*}
\] (9)

**Step 3:** Obtain the corresponding wind power \( X \) sampling value based on \( u_i = F_i(x_i) \). Further, the deterministic load flow calculation is performed for each \( X \) sampling point, and the results of the probabilistic load flow calculation can be obtained.
CASE STUDY

Probabilistic Model Validation

An example is given to illustrate the accuracy of the Pair Copula. Taking 13 wind farms in Australia as examples (wind power stations 1, 2, ..., 13, the corresponding power $P_1, P_2, ..., P_{13}$), the output data of the wind power plants are selected from July 11, 2009 to July 31, 2010 by the hour [20], according to their respective maximum power standard. Since the number of samples is enough, we use the empirical distribution to obtain the edge probability distribution.

In order to investigate the limitations of correlation coefficient, t-Copula and Normal-Copula, two-dimensional wind power is taken as an example to analyze. Through the analysis, it is found that the use of linear or rank correlation coefficient to describe the correlation is incomplete and inaccurate. In addition, the wind power has a variety of correlation structure, so it’s difficult to accurately describe its related properties using a single t-Copula or Normal-Copula.

Then, the wind power of all the wind farms is modeled and the results of the model are fitted to some wind power. There are 1287 kinds and 715 kinds of wind power in 5 and 9 dimensions respectively. The test results are shown in Figure 2. The Euclidean distance and the modeling time of the model are shown in Table I, and the $K(z)$ test is shown in Figure 3.

Figure 3 and Table I show that the Pair Copula Euclidean distance is the smallest and the overall fitting effect is the best. Figure 2 shows that this Pair Copula model is also optimal for some wind power, with a Euclidean distance of about 0.5. Therefore, for the wind power with complex correlation characteristics, the overall and local fitting effects of Pair Copula can be optimized, which can provide the theoretical basis for the calculation and analysis of the operating characteristics of power system. It can be seen from Figure 2 that the Euclidean distance between the wind power of some wind farms is as high as 4 for t-Copula and Normal-Copula, which shows that the local correlation of wind farm is not good.

Table I. Euclidean Distance and Estimation Time of Copula Functions of All Wind Power Outputs.

<table>
<thead>
<tr>
<th>Copula function</th>
<th>Euclidean distance</th>
<th>Estimation time/s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal-Copula</td>
<td>1.189</td>
<td>0.032</td>
</tr>
<tr>
<td>t-Copula</td>
<td>0.787</td>
<td>17791.000</td>
</tr>
<tr>
<td>Pair Copula</td>
<td>0.391</td>
<td>540.000</td>
</tr>
</tbody>
</table>

Figure 2. Euclidean Distance of Copula Functions of 5 and 9 Dimensional Wind Power Outputs.
In addition, although the t-Copula Euclidean distance is smaller than Normal-Copula in Table I, and the overall fitting effect is better than Normal-Copula, it’s not practical with respect to its complex model expression and the long parameter fitting time (about 17791s).

Examples of Pair Copula Applications

Use the IEEE 118-bus system [21] to verify the application of the proposed model in probabilistic load flow calculation. Specific simulation model and parameters are as follows.

1) Wind farms 1, 2, ..., 9 replace the original generators at nodes 12, 25, 26, 49, 54, 59, 61, 103 and 111 respectively. The wind farms 10, 11, 12, 13 are connected to nodes 1, 4, 6 and 8, respectively, with a rated capacity of 100 MW. All the wind farms are used constant pressure control.

2) The node load obeys the independent normal distribution, the expected value is the value at the static equilibrium point, the standard deviation is 5% of the expected value, and the power factor maintains the value at the equilibrium point.

The relative error of the expected value and standard deviation of the output variables is given by (10) and (11), using the probabilistic load flow results calculated from wind power history data as reference values.

\[
\varepsilon_\mu^\gamma = \left| \frac{\mu_{em} - \mu_c}{\mu_{em}}\right| \times 100\% \\
\varepsilon_\sigma^\gamma = \left| \frac{\sigma_{em} - \sigma_c}{\sigma_{em}}\right| \times 100\% 
\]

\(\mu_{em}\) and \(\sigma_{em}\) are the expected value and standard deviation calculated from the historical data. \(\mu_c\) and \(\sigma_c\) are the expected value and standard deviation, respectively, calculated from the probability model. The superscript \(\gamma\) is the output variable type, which includes the line reactive power \(Q\) and active power \(P\), node voltage phase angle \(\theta\) and amplitude \(V\), the system total active loss \(P_{Loss}\) and so on.

Since the number of output random variables may be more than one, the average errors \(\bar{\varepsilon}_\mu\) and \(\bar{\varepsilon}_\sigma\) and the maximum values \(\max(\varepsilon_\mu^\gamma)\), \(\max(\varepsilon_\sigma^\gamma)\) of the relative error of each output random variable are used to measure the global error. As the t-Copula fitting time is too long, we don’t consider it there. The results of the error index calculation are shown in Table II and Table III, and the independent variables in the table represent the probability models of wind power independence.
From Table II and Table III, it can be seen that the errors of the three models are close to each other, so the wind power correlation has little effect on the expected value. However, there are significant differences in the accuracy of the standard deviation between the three models. The standard deviation error is greatest when correlation is not taken into account. Normal-Copula has greatly improved the accuracy of calculation because of the consideration of wind power correlation. Pair Copula error is only half of Normal-Copula, and the precision is the highest.

Furthermore, the influence of correlation on system performance is investigated by taking the active power $P_{40-41}$ between node 40 and node 41 as an example. The probability density curve of $P_{40-41}$ is shown in Figure 4.

<table>
<thead>
<tr>
<th>Error index</th>
<th>Average relative error/%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Normal-Copula</td>
</tr>
<tr>
<td>$\bar{\varepsilon}_P$</td>
<td>0.29000</td>
</tr>
<tr>
<td>$\bar{\varepsilon}_Q$</td>
<td>1.77400</td>
</tr>
<tr>
<td>$\bar{\varepsilon}_V$</td>
<td>0.33400</td>
</tr>
<tr>
<td>$\bar{\varepsilon}_\theta$</td>
<td>1.65600</td>
</tr>
<tr>
<td>$\bar{\varepsilon}_\mu$</td>
<td>0.00074</td>
</tr>
<tr>
<td>$\bar{\varepsilon}_\sigma$</td>
<td>1.21100</td>
</tr>
<tr>
<td>$\bar{\varepsilon}_\theta$</td>
<td>1.04100</td>
</tr>
<tr>
<td>$\bar{\varepsilon}_{Loss}$</td>
<td>2.74200</td>
</tr>
<tr>
<td>$\bar{\varepsilon}_{Loss}$</td>
<td>0.39400</td>
</tr>
<tr>
<td>$\bar{\varepsilon}_{Loss}$</td>
<td>3.15100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Error index</th>
<th>Average relative error/%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Normal-Copula</td>
</tr>
<tr>
<td>max($\varepsilon_P$)</td>
<td>2.949</td>
</tr>
<tr>
<td>max($\varepsilon_Q$)</td>
<td>5.033</td>
</tr>
<tr>
<td>max($\varepsilon_V$)</td>
<td>3.621</td>
</tr>
<tr>
<td>max($\varepsilon_\theta$)</td>
<td>3.856</td>
</tr>
<tr>
<td>max($\varepsilon_\mu$)</td>
<td>0.017</td>
</tr>
<tr>
<td>max($\varepsilon_\sigma$)</td>
<td>4.881</td>
</tr>
<tr>
<td>max($\varepsilon_\theta$)</td>
<td>4.573</td>
</tr>
<tr>
<td>max($\varepsilon_{Loss}$)</td>
<td>3.095</td>
</tr>
</tbody>
</table>
As can be seen from Figure 4 (a), the Pair Copula is closer to the actual result than the Normal-Copula, consistent with the above conclusion. In Figure 4 (b), the probability density curve of $P_{40-41}$ tends to the normal distribution when the correlation is neglected, but the actual distribution is asymmetric and obviously non-normal distribution. This is due to the computational error of the higher order moments of the random quantities when the correlation is not considered, and the larger standard deviation errors in Tables II and III reflect this phenomenon. When the correlation is not considered, the high-value section of the line's active power and the high and low voltage portions of the node voltage are underestimated, i.e., the possible risk of overrun is underestimated.

**CONCLUSION**

In this paper, C-Pair Copula is used to model the correlation of multi-dimensional wind power. The theoretical and numerical examples show that: 1) Pair Copula possesses the advantages of existing Copula theory in two-stochastic modeling, such as the edge distribution is unrestricted, and so on. 2) Pair Copula uses the corresponding binary Copula function to describe the different correlation characteristics of the stochastic quantities, and uses the hierarchical structure to construct the multidimensional wind power correlation structure model. The modeling procedure is simple, flexible and applicable. 3) Compared with the existing Copula function, Pair Copula has high precision of fitting the whole and local correlation, and can obtain the more accurate probability flow calculation results.

**REFERENCES**


