Pointwise Elimination Method on the Abnormal Data and Robust LS Polynomial Fitting of Wind Power

Dong Wang¹, Jingxin Guo², Xiaoyu Zhang²,*, Jianping Fei¹ & Haidong Qian¹

¹Zhoushan Power Supply Company, Zhejiang 315000, P.R. China
²Department of Mathematics, Beijing Forestry University, Beijing 100083, P.R. China

ABSTRACT: In this paper, we establish the 5th order polynomial robust LS curve-fitting model, research the pointwise elimination method on the abnormal data of wind power, and develop an improved stopping-rule combining with the method of T test. Under the nonlinear model, we prove that when the stopping-parameter is \[ w = \frac{r^2}{1-\alpha (m-1)} \] , we can maximally eliminate the abnormal data and reserve the good data, where \( \alpha \) is the confidence level and \( m \) is the number of observation data. Finally, numerical experiments show simulation results and illustrate the effectiveness of our method.

1 INTRODUCTION

The technology of wind power has been developed relatively mature, however, the characteristics of wind such as the intermittent and volatility (Wang Hongtao, 2011) always lead to the intermittent and volatility of power generation, which increase the difficulty of dispatching and requirement for network preparation dosage (Wang Lijie, 2009). Therefore, developing wind-power research actively to improve the accuracy of forecasting has a realistic significance on dispatching and reducing the operation cost of the system (Yang Xiuyuan, 2005). Some improved forecasting models have been obtained (Feng Lei, 2013) (Wang Caixia, 2010). However, less models have been considered the effects of abnormal observation data to the accuracy of model. In order to improve the accuracy of the forecast, it is necessary to eliminate the abnormal data at first. In 1997, Wang (Wang Zhengmin, 1997) developed the one-by-one method of outliers rejection on first linear order model.

In this paper, basing on the fact that the observation data usually shows the characteristic of nonlinear relationship, first of all, we choose the 5th order polynomial robust LS curve fitting model. Secondly, we research the pointwise elimination method on the abnormal data of wind
power, and obtain a criterion of stopping-rule combining with the method of T test. Under choosing an appropriate stopping-parameter, we can maximally eliminate the abnormal data and reserve the good data. Finally, we provide numerical experiments to show the effectiveness of our method.

2 POINTWISE ELIMINATION METHOD

To establish the model based on the observation data of wind power, we take advantage of the LS curve-fitting method (4), which is well known effectiveness.

Let fitting function be \( y = p(x) \), the observation data \( y_i, i = 1, \ldots, m \) and residual error \( \delta_i = \delta_1, \ldots, \delta_m \) \( \delta_i = p(x_i) - y_i \). The key point of LS curve-fitting method finds the least residual sum of squares is \( \min \sum_{i=1}^{m} \delta_i^2 \).

As the observation data usually shows the characteristic of high order polynomial relationship, we adopt the LS curve-fitting function as follows: \( p(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n \).

Generally, there will inevitably be some abnormal observation data among statistical data. And they would cause negative effect on establishing the forecasting model on the observation data of wind power. Therefore, in order to improve the accuracy of forecasting, it is important to detect and eliminate the abnormal observation data at first. Considering the polynomial model:

\[
Y_i = X_i \beta + e_i
\]

(2.1)

where \( X_i = (1, x_i, x_i^2, \ldots, x_i^n) \), \( X = X_{m \times d} = (x_1, x_2, \ldots, x_m)^T \), \( x_i \) is the time at the point \( i \), \( Y = (Y_1, Y_2, \ldots, Y_m)^T \), \( Y_i \) is the observation data at the time \( x_i \), \( \beta = (\beta_0, \beta_1, \ldots, \beta_n)^T \), \( \beta \) is the \( n + 1 \) dimensional parameter vector, \( e = (e_1, e_2, \ldots, e_m)^T \), \( e_i \) is the random error \( e_i \sim N(0, \sigma^2) \), \( i = 1, 2, \ldots, m \). Let
\[
H = (h_{ij})_{mn} = X(X^T X)^{-1} X^T
\]
\[
\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_m)^T = (I_m - H)Y
\]
\[
\zeta = (\zeta_1, \zeta_2, \ldots, \zeta_m)^T, \zeta_i = (1 - h_i)^{-1}\sigma_i^2
\]
\[
h_i = h_{ij} = \sum_{j=1}^{m} h_{ij}^2, 1 \leq i, j \leq m
\]

(2.2)

**Lemma 1 (Tong Li, 2001)** If \( h < \frac{1}{4} \), \( k > 1 \), we have \( E\zeta_i < E\zeta_j \), \( i \in M, j \in N \).

If \( Y_i \) is the abnormal data, \( h \) is much less than \( 1/4 \).

Lemma 1 shows that, the greater value of \( \zeta_i \), the greater probability of \( i \in N \). We can identify the abnormal data by comparing the value of \( \zeta_i \). The pointwise elimination method is to find out the maximum of \( \zeta_i \) one by one, and to eliminate the corresponding data \( Y_i \). However, if there isn’t a stopping-rule, the good data will be eliminated as well. Under the first order polynomial model, the stopping-rule was developed in [6] as follows:

**Lemma 2 (Wang Zhengmin, 1997)** For first order polynomial model (2.1), let \( \omega = \frac{7.29}{K-L} \), when \( \zeta_i = \max_i \zeta_i > \omega\|\sigma\|^2 \), it concludes that the observation data \( Y_i \) is abnormal.

Since \( E\|\sigma\|^2 = (m-p)\delta^2 \), \( \zeta_i\delta^2 \sim \chi^2(1) \), according to the high probability (about 0.95), we have \( \zeta_i < 7.29\delta^2 = \frac{7.29}{K-L}\|\sigma\|^2 \).

3 AN IMPROVED STOPPING-RULE

In this part, we introduce a test statistic \( T_i \), which satisfies the T test. Furthermore, the improved stopping-rule of high order polynomial model is developed by combining the T test and the criterion of stopping-parameter is given.

3.1 Hypothesis Test

Through LS curve-fitting, we could obtain the high order polynomial model, let \( \hat{y}_i \) be the fitted value at the point \( x_i \). If \( y_i \) is normal, we would have \( E(y_i - \hat{y}_i) = 0 \), let \( \sigma_i = y_i - \hat{y}_i \).
σ_i ∼ N(0, δ_i). In general, the error σ_i satisfies the same gauss distribution, which
means δ_j = δ_i = δ, i, j = 1, 2,..., m. Thus, if E(σ_i) = 0, y_i is correct.

Consider test statistic T_i as

\[ T_i = \frac{\sum_{j \neq i} (\hat{y}_j - y_j)}{\sqrt{\frac{\sum_{j \neq i} (\hat{y}_j - y_j)^2}{(m-1)}}}. \]

Since \( \frac{\hat{y}_j - y_j}{\delta_j} \sim N(0,1), \) and

\[ \sum_{j \neq i} (\hat{y}_j - y_j)^2 \sim \chi^2(m-1), \] where m is the number of observation data.

Therefore, we have test statistic \( T_i \sim t(m-1). \) According to T hypothesis test, if

\[ P\{ |T_i| > t_{\text{critical}} \} \geq \alpha, \] where \( \alpha \) is the confidence level, and \( t_{\text{critical}} = t_{1-\alpha}, \) we can conclude that the observation data \( y_i \) is abnormal.

3.2 The improved stopping-rule

Under the high order polynomial model, the effect of using the stopping-parameter given by
lemma 2 to eliminate abnormal data is non-ideal. Thus, we consider the high order polynomial model.

**Lemma 3  (Wang Zhengmin, 1997) Under the assumption of model (2.1)**

\[ P\{ \max_{1 \leq i \leq m} \xi_i \geq w \| \sigma \|^2 \} \geq \alpha \]

(3.1)

For a given stopping-parameter w, from (3.1), the probability of abnormal data \( Y_i \) is misjudged as normal data is less than \( \alpha. \)

**Theorem 1** Under the assumption of model (2.1), let \( n \geq 2, \) when

\[ P\{ \max_{1 \leq i \leq m} \xi_i \geq w \| \sigma \|^2 \} \geq \alpha \]

where \( m \) is the number of observation data, \( Y_i \) is abnormal observation data, we have the concrete form of stopping-parameter \( w = \frac{t_{1-\alpha}(m-1)}{(m-1)}. \)

**Proof:** From (2.2), we have

\[ \sigma_i^2 = (1-h_i)\xi_i^2, \ 1 \leq i \leq m, \]

(3.2)
the LS curve-fitting function: \( y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \beta_2 x_i^2 + \cdots + \beta_n x_i^n \), and \( \sigma_i = y_i - \hat{y}_i \sim N(0, \delta^2) \), we have
\[
\frac{y_i - \hat{y}_i}{\delta} \sim N(0,1) \tag{3.3}
\]
From (3.3), there exists the test statistic \( T_i \)
\[
T_i = \frac{\sigma_i}{\sqrt{\sum_{j \neq i} \sigma_j^2 / (m-1)}} \sim t(m-1) \tag{3.4}
\]
where \( \sum_{j \neq i} \sigma_j^2 = \|\sigma\|^2 \).

Combined with (3.2) and (3.4), we obtain
\[
T_i = \sqrt{\frac{(1-h_i)}{\|\sigma\|^2 / (m-1)}} \sim t(m-1) \tag{3.5}
\]
We could simplify (3.5) based on lemma 2.1 as
\[
T_i' = \sqrt{\frac{1}{\|\sigma\|^2 / (m-1)}} \sim t(m-1) \tag{3.6}
\]
According to hypothesis test, the critical of \( T_i' \) is \( t_{1-\alpha}(m-1) \).

By (3.6) and lemma 3.1, it follows \( Y_i \) is the abnormal data when
\[
T_i = \sqrt{\frac{\hat{\epsilon}_i}{\|\epsilon\|^2 / (m-1)}} \sim t_{1-\alpha}(m-1) \tag{3.7}
\]
Rewrite (3.7) as \( \hat{\epsilon}_i = t_{1-\alpha}(m-1) \|\epsilon\|^2 / (m-1) \), we have \( \omega\|\epsilon\|^2 = t_{1-\alpha}(m-1) \|\epsilon\|^2 \), which implies \( w = t_{1-\alpha}(m-1) / (m-1) \).

This completes the proof of Theorem 1. Proceeding as in the proof of Theorem 1, we can also obtain the following criterion.
Criterion 1: For high polynomial model (2.1), denote \( \sigma = (I - H)e \),
\[
\zeta_i = (1-h_i)^{-1}\sigma_i^2, 1 \leq i \leq m. \]
If \( \max_{1 \leq i \leq m} \zeta_i \geq w_2^2 \), and the stopping-parameter \( w = \frac{1-\alpha}{(m-1)} \),
then \( Y_i \) is abnormal data, where \( \alpha \) is the confidence level.

4 NUMERICAL EXAMPLES

Example 4.1: In the following Table, there are wind power data of different moment which is divided into two groups, provided by some wind power company of China in December 2008.

<table>
<thead>
<tr>
<th>moment</th>
<th>wind power</th>
</tr>
</thead>
<tbody>
<tr>
<td>0:00-0:55</td>
<td>68 66 66 67 69 70 69 70 72 74 79 81</td>
</tr>
<tr>
<td>1:00-1:55</td>
<td>82 83 84 84 80 79 77 79 80 80 85 86</td>
</tr>
</tbody>
</table>

Figure 4.1. The fitting curves before and after elimination.

First of all, we get 5th order polynomial model by LS curve-fitting:

\[
p(x) = a_0 + a_1x + a_2x^2 + a_3x^3 + a_4x^4 + a_5x^5.
\]

Then, by means of pointwise elimination method and the theorem 1, we obtain the excluding critical condition is \( \zeta_i = \max_{1 \leq i \leq m} \zeta_i \geq \sqrt{\frac{2}{m}} \), where \( m = 12 \), \( w = \frac{t_{0.95}^2(11)}{11} = 0.169 \).
Finally, by using the theorem 1, we can get the estimated results which were showed as following table 4.2. The last column is the residual sum of squares. The first line is the true value of $a$, the second line is the estimated result on all data. And the other lines are the results after excluding the outliers $(Y_6, Y_{10}, Y_{10}, Y_{4})$, and $(Y_{22}, Y_{22}, Y_{16}, Y_{22}, Y_{16}, Y_{20})$.

Table 4.2. The table of wind power eliminating results.

<table>
<thead>
<tr>
<th></th>
<th>$a_0$</th>
<th>$a_1$</th>
<th>$a_2$</th>
<th>$a_3$</th>
<th>$a_4$</th>
<th>$|T|$</th>
</tr>
</thead>
<tbody>
<tr>
<td>raw data</td>
<td>-9381.60</td>
<td>13686.41</td>
<td>-6979.60</td>
<td>1510.21</td>
<td>-114.27</td>
<td>68.19</td>
</tr>
<tr>
<td>6.00</td>
<td>-9719.19</td>
<td>12596.45</td>
<td>-6352.57</td>
<td>1366.96</td>
<td>-104.37</td>
<td>68.13</td>
</tr>
<tr>
<td>6.10</td>
<td>-9468.27</td>
<td>13067.12</td>
<td>-6590.85</td>
<td>1389.49</td>
<td>-104.55</td>
<td>68.12</td>
</tr>
<tr>
<td>6.10,4</td>
<td>-10304.66</td>
<td>14488.25</td>
<td>-7089.30</td>
<td>1466.00</td>
<td>-105.84</td>
<td>68.08</td>
</tr>
</tbody>
</table>

It showed clearly in the table 3.4 and figure 3.1:

1) After eliminating outliers, we can find that the residual error of LS curve-fitting model will have a substantial decline. Meanwhile, the estimated results can get a great improvement.

2) By means of pointwise elimination method and theorem 1 which is combined with T test, let $w = 0.1681$, we can identify the abnormal observation data accurately and protest the good ones greatly.

3) The 5th order polynomial model we obtained on the good data fit well to the real tendency which means that the model has a good reference value for wind power protection.

REFERENCES


