Curve and surface reconstruction based on MTLS algorithm combined with k-means clustering

Tianqi Gu¹, Hongxin Lin¹, Dawei Tang³, Shuwen Lin¹ and Tianzhi Luo²,*

¹School of Mechanical Engineering and Automation, Fuzhou University, Fuzhou 350108, China
²CAS Key Laboratory of Mechanical Behaviour and Design of Materials, Department of Modern Mechanics, University of Science and Technology of China, Hefei 230022, China
³Centre for Precision Technologies, University of Huddersfield, Huddersfield HD1 3DH, UK

Abstract

Curve and surface reconstruction methods play an important role in many research and engineering fields. It is an imperative procedure to carry out surface reconstruction from measurement data in reverse engineering, which is complicated with the presence of outliers. To achieve better accuracy and robustness of reconstruction, an improved moving total least squares (MTLS) algorithm based on k-means clustering called a KMTLS method is proposed in this article. Based on MTLS, KMTLS adjusts the weight of discrete points within the support domain by adopting a two-step fitting procedure. Firstly, an ordinary least squares (OLS) method is adopted to obtain the pre-fitting result and calculate the residuals as the input of k-means clustering. In k-means clustering, abnormal nodes are classified into one cluster and a weight function based on clustering information is introduced to deal with these nodes. Secondly, based on the compact weight function in MTLS and the weight obtained in the pre-fitting procedure, a weighted total least squares method is conducted to determine the final estimated value. The process of detecting outliers is

Corresponding author: tzluo@ustc.edu.cn
automatic without setting threshold artificially. The simulation and experiment show that KMTLS has great robustness and accuracy.

**Keywords:** Surface reconstruction; Moving least squares; K-means clustering; Outliers.

1. Introduction

As a classical parametric regression method, OLS is a powerful technique that has extensive applications in many research fields such as point set registration [1], economic analysis [2], curve fitting [3,4]. However, for a function with a complex feature, it requires a higher-order basis to depict its shape and the Runge phenomenon easily occurs on this occasion [5]. More importantly, OLS is a global regression method so that the local characteristic of the profile may be ignored, which makes OLS not feasible to fit the function of complex shape. To overcome this disadvantage of OLS, Shepard [6] proposed a pointwise non-parametric regression method, i.e., moving least squares (MLS), to fit arbitrary two-dimensional curve and Lancaster [7] promoted this method to fit the three-dimensional surface. Based on OLS, they introduced a weight function with compact support. Compared to OLS, MLS only considers those nodes within support domain and the assigned weights are positive related to the distance to the centre of support domain, which makes MLS able to fit the local shape of the curve with low order basis [8]. Since MLS was proposed, it has been widely used in the surface reconstruction and data approximation field [9, 10]. Moreover, owing to its approximation way, MLS has become an imperative method of constructing the shape function in the meshless method [11-14].

In the actual measurement system, the measurement errors are composed of systematic errors, random errors, and outliers (or gross errors) [15]. Even though systematic errors can be eliminated to some extent after calibration while random errors and outliers are inevitable in measurement data. Therefore, the reconstruction performance of measurement data depends on the capability to deal with these two errors. Random errors not only occur in the dependent variable but exist in the independent variable. In OLS, the optimization target is to minimise the residual
errors between the fitting curve and discrete points, which implies that only errors of
the dependent variable are considered in OLS. Therefore, the OLS-based method, i.e.,
MLS, cannot deal with the error of the independent variable in theory. To solve this
problem, Scitovski [16] proposed MTLS based on total least squares (TLS) to take the
errors of all variables into account. In TLS, the fitting result is obtained by minimising
the orthogonal distance between the fitting curve and discrete points. Similar to MLS,
by introducing a compact weight function to TLS, MTLS can also reflect the local
feature of complex function.

Even though MTLS is a more rational method for approximating measurement
data, the outlier is still challenging for MTLS. The fitting accuracy of MTLS will be
greatly influenced when the measurement data contain outliers. The reason is that
when the estimated point is near the outliers, the weight of outliers is large, leading to
the estimated value deviating from the ideal curve [17]. In fact, the outlier not only
has an adverse effect on surface generation, but also has a huge impact on industrial
monitoring filed in which the controlled variable highly depends on the monitoring
data [18, 19]. Hence, the investigation of the solution to the outlier detection and
rectification method is of great importance. Abundant of scholars have investigated
and proposed many outlier detection methods based on statistic [20,21], machine
learning [22], heuristic algorithm [23] and so on. In the surface reconstruction field,
one possible solution to this problem is to directly delete the outliers by setting a
threshold value [24, 25]. In this type of solution, the performance on handling outlier
is highly associated with the threshold value itself so that the determination of a
reasonable threshold value is difficult for fitting different measurement data.
Moreover, an artificially determined threshold may induce additional negative effect.
An alternative solution is to assign a small weight to the abnormal points to weaken
their influence [26]. The key issue of this solution is to define the weight of outliers
without influencing the weight of normal points.

Currently, there are few reports on improving MTLS methods to simultaneously
suppress random errors and outliers in measurement data. In this paper, an MTLS
based reconstruction method combined with k-means clustering is proposed called a
KMTLS method. KMTLS adopts a two-step fitting procedure to obtain the estimated value. The first fitting procedure called the pre-fitting procedure uses an OLS method to fit the discrete points within the support domain. According to the residuals calculated in the pre-fitting procedure, the k-means clustering method is adopted to classify the discrete points into two clusters and a weight function based on clustering result is used to adjust the weight of outliers. One of the advantages of KMTLS is that no manual threshold values are introduced so that the method can automatically deal with the outlier based on data itself. The performance tests show that by adjusting the weight of points based on the result of k-means clustering, the outliers as well as random errors can be suppressed. Arrangements for this article are shown as follows.

In section 2, the basic theory of MLS and MTLS is introduced concisely. In section 3, a brief introduction to the standard k-means clustering algorithm is given and KMTLS is proposed. In section 4, some fitting cases are used to test and compare the performances of KMTLS, MLS and MTLS method. In section 5, the performance of KMTLS is discussed in detail. In section 6, the whole work in this paper is concluded.

2. Introduction to the fundamental theory

2.1 MLS method

In this section, a concise introduction to MLS is given. Take the function \( u(x) \) defined in the domain \( \Omega \) into consideration. The independent variable \( x \) can be further expanded as \( \{x\} \) and \( \{x, y\} \) in space \( \mathbb{R}^1 \) and \( \mathbb{R}^2 \), respectively. In MLS, supposed that estimated point locates at \( x \), and support domain \( \Omega_x \) is defined by \( \Omega_x = \{y : \|y - x\| < r\} \) where \( r \) is the radius of support domain. Then, MLS gives the estimated value \( \hat{u}(x) \) by conducting weighted least squares (WLS) method within the support domain \( \Omega_x \) with the expression

\[
\hat{u}(x) = \sum_{i=1}^{m} p_i(x) \alpha_i(z) \bigg|_{z=x} = p^T(x) \alpha(z) \bigg|_{z=x}
\]

where \( p(x) \) is a basis vector with a dimension of \( m \). In 1D, \( p^T(x) = [1 \ x] \) with linear basis and \([1 \ x \ x^2]\) with quadratic basis. \( \alpha(z) \big|_{z=x} \) is a coefficient vector depending on \( x \), which can be determined by minimising the error function [27]
\[
J(\alpha(z)|_{z=x}) = \sum_{y=x_1}^{x_n} w(y-x,r)[u(y) - p^T(y)\alpha(z)]_{z=x}^2
= (P^T_{\Omega, x} \alpha(z)|_{z=x} - u_{\Omega, x})^T W(x) (P^T_{\Omega, x} \alpha(z)|_{z=x} - u_{\Omega, x})
\]

with
\[
\begin{align*}
\vec{u}_{\Omega, x} &= (u(x_1) \ u(x_2) \ldots \ u(x_n))^T \\
\vec{p}_{\Omega, x} &= (p^T(x_1) \ p^T(x_2) \ldots \ p^T(x_n))^T \\
W(x) &= \text{diag}(w(x_1-x,r) \ w(x_2-x,r) \ldots \ w(x_n-x,r))
\end{align*}
\]

where \(x_1, x_2, \ldots, x_n\) are the discrete points in \(\Omega_x\) and \(w(y-x,r)\) is a compact weight function in which the weight of point within support domain is positively related to the distance to the centre of \(\Omega_x\). The widely used exponential weight function is adopted in this paper with the expression
\[
w(y-x,r) = \begin{cases} 
    e^{-s \theta^s} - e^{-\theta}, & \text{if } s \leq 1 \\
    1 - e^{-\theta}, & \text{if } s > 1
\end{cases}
\]

where \(s = ||y-x||/r\) and \(\theta\) is a parameter related to the convergence rate of the weight value. In this paper, \(\theta\) is set to 9 for all cases. Fig.1 shows the exponential weight functions at different estimated points. In Fig.1, it can be found that the weight of point \(x_i\) changes with the estimated point moving from \(x_{i-1}\) to \(x_{i+1}\). In MLS, the weights of the discrete points in different support domains are independent with each other.

![Fig. 1. The distribution of weight at different estimated points](image)

The solution of the coefficient vector in Eq. (2) can be obtained by
\[
\frac{\partial J}{\partial \alpha} = M^{-1}(x) \alpha(z)|_{z=x} - P^T_{\Omega, x} W(x) u_{\Omega, x} = 0
\]

Thus, we get
\[ a(z)_{\varepsilon} = \mathbf{M}^{-1}(\mathbf{X})\mathbf{P}_{\Omega}^T W(x)\mathbf{u}_{\Omega}, \quad (6) \]

with

\[ \mathbf{M}(\mathbf{x}) = \mathbf{P}_{\Omega}^T W(x)\mathbf{P}_{\Omega}, \quad (7) \]

where \( \mathbf{M}(\mathbf{x}) \) is the moment matrix with the dimension \( m \times m \), and the dimension of the matrix \( \mathbf{P}_{\Omega} \) is \( n_\Omega \times m \). By substituting Eq. (6) to Eq. (1), the estimated value \( \hat{u}(\mathbf{x}) \) can be obtained

\[ \hat{u}(\mathbf{x}) = \mathbf{p}^T(\mathbf{x})\mathbf{M}^{-1}(\mathbf{x})\mathbf{P}_{\Omega}^T W(x)\mathbf{u}_{\Omega}, \quad (8) \]

In the application of MLS for data approximation, the moment matrix \( \mathbf{M}(\mathbf{x}) \) sometimes will be ill-conditioned, which makes the inversion of \( \mathbf{M}(\mathbf{x}) \) further decrease the calculation precision and numerical stability. Consequently, MLS sometimes will give incorrect result [28, 29]. Therefore, to improve the robustness of MLS, it is usually needed to scale the monomials in the basis by transferring the global variable to the local variable or adopt the orthogonal basis. If the estimated point is at \( \mathbf{x}_I \), the basis vector \( \mathbf{p}(\mathbf{y}) \) in Eq. (2) is transferred to \( \mathbf{p}((\mathbf{y}-\mathbf{x}_I)/r) \) called shifted polynomial basis [30, 31]. With this transformation, the estimated value \( \hat{u}(\mathbf{x}_I) \) can be obtained

\[ \hat{u}(\mathbf{x}_I) = \mathbf{p}^T(0)\mathbf{M}^{-1}(\mathbf{x}_I)\mathbf{P}_{\Omega}^T W(x)\mathbf{u}_{\Omega}, \quad (9) \]

with

\[ \mathbf{P}_{\Omega} = (\mathbf{p}^T \frac{\mathbf{x}_1 - \mathbf{x}_I}{r} \quad \mathbf{p}^T \frac{\mathbf{x}_2 - \mathbf{x}_I}{r} \quad \ldots \quad \mathbf{p}^T \frac{\mathbf{x}_{n_\Omega} - \mathbf{x}_I}{r})^T \quad (10) \]

and \( \mathbf{p}^T(0)=[1 \ 0] \) with linear basis and \( \mathbf{p}^T(0)=[1 \ 0 \ 0] \) with quadratic basis in 1D case.

### 2.2 MTLS method

According to Eq. (2), it can be found that MLS only minimises the residual errors between the fitting curve and discrete points without considering the error of independent variable \( \mathbf{x} \). Similar to the discussion of MLS in the last section, the function \( u(\mathbf{x}) \) and a given point \( \mathbf{x} \) are also considered in this section. To consider the error of independent variable, MTLS begins with the Errors-in-Variable (EIV) model which defines the relationship between the dependent variable and independent variable as [32]
\[(P_{\Omega} + E_x)^T \beta(z) \bigr|_{z=x} = u_{\Omega} + e_u\]  

(11)

where \(E_x\) is the error matrix of \(P_{\Omega}\) related to \(x\) and \(e_u\) is the error vector of \(u_{\Omega}\). In MTLS, the coefficient vector \(\beta(z)\bigr|_{z=x}\) is determined by minimising the error function

\[K(\beta(z)\bigr|_{z=x}) = e_s^T W(x) e_s + e_u^T W(x) e_u, \text{ with } e_s = \text{vec}(E_x)\]  

(12)

Compared with Eq. (2), it can be found that in MLS method, only the second term in the right of equation is considered in the error function. To determine the coefficient \(\beta(z)\bigr|_{z=x}\), the SVD-based solution is usually adopted due to its efficiency and simplicity.

In MTLS, the SVD of the augmented matrix is

\[C := W(x)[P_{\Omega}, u_{\Omega}] = \begin{bmatrix} U_{n,x} \Sigma_{n,x} \end{bmatrix} V_{(m+1)x(n+1)^T}\]  

(13)

where \(U^T U = I_n\), \(V^T V = I_{m+1}\), \(\Sigma = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_{m+1}, 0_{(m+1)\times(n+1)})^T\), and \(\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{n+1}\) are the singular values of augmented matrix \(C\). Let \(\Sigma' := \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_{m+1})\) and the partition of matrix \(V\) and \(\Sigma'\) are respectively

\[V' := \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}_{m} \quad \Sigma' := \begin{bmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{bmatrix}_{m}\]  

(14)

A solution exists when \(V_{22}\) is non-singular and the solution is unique if \(\sigma_n \neq \sigma_{n+1}\).

In this case, the solution of Eq. (11) can be obtained by

\[\beta(z)\bigr|_{z=x} = -V_{12} V_{22}^{-1}\]  

(15)

Through substituting the coefficient vector \(\beta(z)\bigr|_{z=x}\) into Eq. (1), MTLS gives the estimated value \(\hat{u}(x)\) at \(x\). Similar to the treatment of basis in MLS, through the transformation of the basis vector mentioned in the last section, the robustness of MTLS in fitting the data with large random errors can be improved. Hence, in this paper, all MLS, MTLS and KMTLS adopt the shifted polynomial basis to fit curve and surface.

3. Proposed KMTLS

3.1 K-means clustering method

In this section, a brief introduction to k-means clustering is presented. K-means clustering is a classical and effective clustering method, and based on k-means, many novel clustering methods are further developed such as H-K-means (Hierarchical
K-means) [33], ATPSO (Accelerated Two-Stage Particle Swarm Optimization) [34] and so on. In k-means clustering, suppose that a set of discrete points \( X = \{x_1, x_2, \ldots, x_n\} \) in space \( \mathbb{R}^d \), k-means divides the point set \( X \) into \( k \) clusters \( C = \{C_1, C_2, \ldots, C_k\} \) subject to \( U_{i,j}, C_i = X \) and \( C_i \cap C_{i'} = \emptyset \), \( i \neq i' \) by minimizing the squared distance between discrete points and their nearest centroids [35]

\[
\sum_{i=1}^{k} \sum_{j \in C_i} \| x_j - \mu_i \|^2
\]  

(16)

where \( \| \cdot \| \) denotes the Euclidian norm and \( \mu_i \) is the centroids of the cluster \( C_i \) calculated by the average of \( C_i \). To obtain the global minima of Eq. (16) is an NP-hard problem [36]. Lloyd firstly proposed the solution of k-means clustering in an iterative scheme to calculate the local minima with great simplicity and flexibility [37]. After that, almost all scholars adopted the generalized Lloyd’s algorithm as the solution of k-means clustering [38]. The process of generalized Lloyd’s algorithm is adopted in this paper, which can be expressed as follow [39]

1. **Step 1**: Input discrete point set \( X \) and give the number of cluster \( k \).
2. **Step 2**: Randomly select \( k \) points as the initial centroids of \( k \) clusters.
3. **Step 3**: Calculate the distance \( d_i = \| x_j - \mu_i \| \) between point \( x_j \) and all centroids \( \mu_i \) \((1 \leq i \leq k)\) respectively and assign point \( x_j \) to the nearest cluster.
4. **Step 4**: Recalculate the centroid of all clusters.
5. **Step 5**: Repeat Step 3 – Step 4 until all the centroids are stable.

It is noted that the clustering result obtained through the iterative scheme is the local minima because the clustering result can be greatly influenced by the initial position of the centroids. Furthermore, the parameter \( k \) is needed to be defined manually, which is one of the major problems of this clustering method [40]. Another problem is that the clustering result is also sensitive to outliers [41]. Therefore, by combining the k-means clustering with MTLS, the above-mentioned problems should be considered properly, which will be discussed in the next section.

### 3.2 KMTLS

As stated in section 2, MTLS is more suitable for dealing with measurement data when random errors exist in all variables. However, MTLS is also sensitive to outliers.
Hence, an improved MTLS method which aims to enhance the reconstruction robustness for dealing with outliers is proposed.

Fig. 2. The principle of KMTLS

Fig. 2 shows the principle of KMTLS. In KMTLS, to obtain the estimated value at an estimated point $x^{(e)}_i$, a two-step fitting procedure is adopted. The first step is a pre-fitting procedure. An OLS method is adopted within the support domain to obtain the local coefficients and calculate the residual $e_{i,j}$ of each point. Then, k-means clustering is adopted to classify the discrete points in the support domain into two clusters according to their residuals obtained in the pre-fitting procedure. In k-means clustering, the parameter $k$ is set to 2 and two points are randomly selected as the two initial centroids of cluster 1 and 2. Points in cluster 1 (with lower centroid) denotes that these points are well distributed in the support domain and their weights will not be influenced subsequently. While the points in cluster 2 (with higher centroid) denotes that these points are distributed abnormally, and their weights should be weakened accordingly. The weight of points in cluster 2 is defined by:

$$w(e_{i,j}) = 1 - \frac{l_{i,j}}{L}$$

(17)

where $L$ is the difference between the maximum residual and the centroid of the first cluster, and $l_{i,j}$ is the difference of the residual of $x'_{i,j}$ and the centroid of the first cluster. In this way, the weight of the point with maximum residual is set to 0 and the
weights of other points in the second cluster are reduced according to Eq. (17). In the second fitting procedure, a weighted total least squares (WTLS) method is conducted in which the weight is defined by the product of compact weight function and the weight obtained by the pre-fitting result.

The flow chart of KMTLS in a 1D situation is presented in Fig. 3.

**Fig. 3.** The flow chart of KMTLS

4. Case validation

To validate the capability of KMTLS for handling outliers and random errors, simulation and experimental measurement data are taken as the input data and processed by these three methods (MLS, MTLS, KMTLS). In Cases 1 and 2, the simulation data are generated by adding random errors and outliers to the ideal data.

4.1. Case 1

The 1D function

\[
y = \sin(0.2\pi x) + 0.25\sin(0.6\pi x), \quad x \in [0, 10]
\]

is taken as a benchmark for MLS, MTLS and KMTLS in 1D case. To generate the simulation data, a set of points \{((x_1, y(x_1)), (x_2, y(x_2)), \ldots, (x_n, y(x_n))\} with uniform space is taken as the original discrete points. Random errors are introduced by
\( (x'_i = x_i + \epsilon_i, \quad y'_i = y(x_i) + \delta_i) \), where \( \epsilon_i \sim N(0, \sigma_x^2) \) and \( \delta_i \sim N(0, \sigma_y^2) \). Subsequently, some outliers are further added to the discrete points \( (x'_i, y'_i) \) to obtain the input data. In this case, three points are selected to generate the outlier \( E_1, E_2, E_3 \) respectively.

Set the number of equispaced estimated points \( n' \) to 201 and the radius of the support domain \( r \) to \( (\max(x_i) - \min(x_i)) \times 5/100 \). The sum of differences \( s \) between the fitting value \( \hat{y}_i \) and theoretical value \( y_i \) expressed as

\[
s = \sum_{i=1}^{n'} |y_i - \hat{y}_i|
\]

is used to evaluate the fitting property of three methods. Table 1 gives the simulation results. For each row of Table 1, simulations are conducted 1000 times under each combination of \( \sigma_x \) and \( \sigma_y \). In each simulation run, MLS, MTLS and KMTLS are applied to fit the simulation data, respectively. Take \( s' \) as the average of \( s \) values in 1000 simulation runs for each method as shown in Table 1.

<table>
<thead>
<tr>
<th>( \sigma_x )</th>
<th>( \sigma_y )</th>
<th>MLS</th>
<th>MTLS</th>
<th>KMTLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1E-06</td>
<td>1E-03</td>
<td>1.879194</td>
<td>1.582363</td>
<td>0.375409</td>
</tr>
<tr>
<td>1E-05</td>
<td>1E-03</td>
<td>1.879900</td>
<td>1.583016</td>
<td>0.376054</td>
</tr>
<tr>
<td>1E-04</td>
<td>1E-03</td>
<td>1.879613</td>
<td>1.582803</td>
<td>0.375713</td>
</tr>
<tr>
<td>1E-03</td>
<td>1E-03</td>
<td>1.879439</td>
<td>1.582985</td>
<td>0.376240</td>
</tr>
<tr>
<td>1E-03</td>
<td>1E-04</td>
<td>1.878659</td>
<td>1.580337</td>
<td>0.362461</td>
</tr>
<tr>
<td>1E-03</td>
<td>1E-05</td>
<td>1.878561</td>
<td>1.580202</td>
<td>0.361966</td>
</tr>
<tr>
<td>1E-03</td>
<td>1E-06</td>
<td>1.878495</td>
<td>1.580135</td>
<td>0.362195</td>
</tr>
</tbody>
</table>

Fig.4 presents the fitting results in a simulation run by the three methods under the condition of \( \sigma_x = \sigma_y = 10^{-3} \). From the comparison results, it can be found that the fitting curves obtained by MLS and MTLS distort significantly near the outliers. In comparison, the fitting curve of KMTLS is barely influenced by the outliers. As shown in Table 1 and Fig.4, KMTLS shows great robustness for dealing with the outliers.
4.2. Case 2

The 2D function [42]

\[ z = \frac{3}{4} e^{-\left(\frac{1}{4}(x-y)^2+(y-y)^2\right)} + \frac{3}{4} e^{-\left(\frac{1}{4}(x+y)^2+(y+y)^2\right)} + \frac{1}{2} e^{-\left(\frac{1}{4}(x+y)^2+(y-y)^2\right)} - \frac{1}{5} e^{-\left(\frac{1}{4}(x+y)^2+(y+y)^2\right)} \]  

(20)

is taken as a benchmark for three methods in 2D case, where \((x, y) \in \Omega = [0, 1] \times [0, 1]\).

Set the total number of estimated points to \(n' \cdot m' = 41 \cdot 41 = 1681\), where \(n', m'\) are the number of equispaced estimated points on \(x\) and \(y\) axis respectively, and let \(r = (\max(x_i) + \max(y_j)) \times 0.01\) in Case 2. Similar with Case 1, the \(s\) value can be defined as

\[ s = \sum_{i=1}^{n'} \sum_{j=1}^{m'} |z_{i,j} - \hat{z}_{i,j}| \]  

(21)

where \(z_{i,j}\) and \(\hat{z}_{i,j}\) are theoretical value and fitting value, respectively. Fig.5 shows the fitting results by three methods. When the outliers exist in the measurement data, MLS and MTLS give poor estimation near the outliers, while KMTLS can suppress all the outliers in this case and gives a smooth approximation result. Table 2 show the \(s'\) values obtained by three methods under the existence of outliers. The same conclusion can be made that KMTLS has outstanding performance on suppressing the outlier.

<table>
<thead>
<tr>
<th>(\sigma_x, \sigma_y)</th>
<th>(\sigma_z)</th>
<th>(s')</th>
<th>MLS</th>
<th>MTLS</th>
<th>KMTLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>(10^{-3})</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Fitting results and error distributions in Case 2

Fig. 5. Fitting results and error distributions in Case 2
To further investigate the performance of KMTLS in dealing with the random errors, only random errors are added to Cases 1 and 2, respectively. Table 3 and 4 show the comparison results of $s'$ values without outliers. In this situation, it can be found that KMTLS also shows the best performance on suppressing random errors, which may be because that KMTLS suppresses the point with a large degree of deviation in the fitting procedure even though outliers are not considered.

**Table 3.** The $s'$ values in Case 1 without outliers

<table>
<thead>
<tr>
<th>$\sigma_x$</th>
<th>$\sigma_y$</th>
<th>MLS</th>
<th>MTLS</th>
<th>KMTLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1E-06</td>
<td>1E-03</td>
<td>0.816006</td>
<td>0.415274</td>
<td>0.367198</td>
</tr>
<tr>
<td>1E-05</td>
<td>1E-03</td>
<td>0.815587</td>
<td>0.414818</td>
<td>0.366544</td>
</tr>
<tr>
<td>1E-04</td>
<td>1E-03</td>
<td>0.816003</td>
<td>0.415433</td>
<td>0.367148</td>
</tr>
<tr>
<td>1E-03</td>
<td>1E-03</td>
<td>0.817417</td>
<td>0.417512</td>
<td>0.369343</td>
</tr>
<tr>
<td>1E-03</td>
<td>1E-04</td>
<td>0.814956</td>
<td>0.412476</td>
<td>0.348679</td>
</tr>
<tr>
<td>1E-03</td>
<td>1E-05</td>
<td>0.815198</td>
<td>0.413029</td>
<td>0.349096</td>
</tr>
<tr>
<td>1E-03</td>
<td>1E-06</td>
<td>0.814786</td>
<td>0.412639</td>
<td>0.348982</td>
</tr>
</tbody>
</table>

**Table 4.** The $s'$ values in Case 2 without outliers

<table>
<thead>
<tr>
<th>$\sigma_x$</th>
<th>$\sigma_y$</th>
<th>MLS</th>
<th>MTLS</th>
<th>KMTLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1E-06</td>
<td>1E-03</td>
<td>1.805159</td>
<td>2.92329</td>
<td>2.48402</td>
</tr>
<tr>
<td>1E-05</td>
<td>1E-03</td>
<td>1.811909</td>
<td>2.96089</td>
<td>2.48731</td>
</tr>
<tr>
<td>1E-04</td>
<td>1E-03</td>
<td>1.813532</td>
<td>3.10099</td>
<td>2.62700</td>
</tr>
<tr>
<td>1E-03</td>
<td>1E-03</td>
<td>2.087255</td>
<td>2.016168</td>
<td>1.859258</td>
</tr>
<tr>
<td>1E-03</td>
<td>1E-04</td>
<td>1.841913</td>
<td>1.496398</td>
<td>1.407924</td>
</tr>
<tr>
<td>1E-03</td>
<td>1E-05</td>
<td>1.836036</td>
<td>1.482572</td>
<td>1.399788</td>
</tr>
<tr>
<td>1E-03</td>
<td>1E-06</td>
<td>1.837691</td>
<td>1.483085</td>
<td>1.397583</td>
</tr>
</tbody>
</table>

**4.3. Case 3**

KMTLS shows outstanding fitting accuracy compare to MLS and MTLS according to simulation results. To further verify the performance of KMTLS on fitting actual measurement data, a coordinate measuring machine (CMM) is utilised to measure the outer shape of a cylindrical part with a radius of 40.1840 mm calibrated by the Taylor Hobson PGI 1240 profilometer.
The CMM measurement platform in Fig. 6 mainly consists of three-axis mobile rails and a laser displacement sensor. In this measurement platform, SILVERA 080 series precision rail is adopted as the movement part controlled by Parker 1505 controller with a repetitive positioning error of about 15 μm. LK-G150 laser displacement sensor from KEYENCE with a repetitive error of around 5 μm is adopted and clamped on the Z-rail. The measurement data of the part profile is obtained by LK-G150 and the scanning movement of X-rail, respectively. In the measurement procedure, the cylindrical part is mounted on the workbench.

The principle of LK-G150, which mainly adopts the triangulation to measure the object [43], is shown in Fig. 7. When the object moves from position $a$ to $b$, the relative displacement $m$ can be calculated through the similarity between $\triangle M'P'N'$ and $\triangle MPN$ with the expression
\[ m = \frac{(OM - f)M'N'\sin(\gamma)}{f \sin(\theta) - M'N'(1 - \frac{f}{OM})\sin(\theta + \gamma)} \quad (22) \]

where \( f \) is the focal distance of the receiving lens, \( \gamma \) is the angle between CCD and receiving laser beam from position \( a \), and \( \theta \) is the angle between transmitting beam and the receiving laser beam from position \( a \).

![Fig. 8. The fitting results of measurement data](image)

In this case, the measurement data with a length of 42mm and the measurement points number of 681 is adopted. Let \( r = (\max(x_i) - \min(x_i)) \times 2.5/100 \). Firstly, three methods are adopted to fit the experimental data, respectively. Then, each fitting result is taken as the input data for parameter regression based on a simulated annealing algorithm to obtain a regression radius of the cylindrical part. Fig.8 shows the fitting results by three methods and Table 5 gives the regression radii of different fitting results. In Table 5, the radius obtained by KMTLS is the closest to calibrated radius of the part, which validates the performance of KMTLS in handling measurement data.

<table>
<thead>
<tr>
<th>( \sigma_x )</th>
<th>( \sigma_y )</th>
<th>Regression radius</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLS</td>
<td>MTLS</td>
<td>KMTLS</td>
</tr>
<tr>
<td>0.015</td>
<td>0.005</td>
<td>40.110147</td>
</tr>
</tbody>
</table>

5. Discussion
In Cases 1 and 2, the outliers are set to be far away from each other so that there is always only a single outlier in each support domain. KMTLS has shown good approximation properties in such a situation. However, in fitting the actual measurement data, the situation may be more complex, that is, there are multiple outliers in one support domain with different amplitudes (Fig.9). Under this situation, in the fitting procedure of KMTLS, the weight of the furthest outlier is set to 0 according to Eq. (17) while part of outliers are classified into the first cluster, leading to the deviation of the estimated value as shown in Fig.9.

![Fig. 9. The fitting result by KMTLS with multiple outliers](image)

To extend the proposed KMTLS to solve this problem, an iterative form of KMTLS is investigated in this section. In the iterative KMTLS, a weighted residual $e'_{i,j,h}$ is introduced as the input of k-means clustering at each iteration defined by

$$e'_{i,j,h} = \begin{cases} w^{(h-1)}(e'_{i,j,h-1}) \cdot e_{i,j} , & h > 1 \\ e_{i,j} , & h = 1 \end{cases} (23)$$

where $w^{(h-1)}(e'_{i,j,h-1})$ is the weight calculated by clustering the weighted residual $e'_{i,j,h-1}$ obtained in the pre-fitting result of $(h-1)$th iteration, and $e_{i,j}$ is the residual defined in Fig.2. In the pre-fitting procedure of $h$th iteration, a WLS method is adopted to fit the discrete points with the weight $w_{i,j,h}$. The estimated value at $h$th iteration can be obtained through WTLS with the weight $w_{i,j,h}$.
\[
\begin{align*}
    w_{i,j,h} &= \begin{cases} 
    w(x'_{i,j} - x_i, r) \cdot w^{(h-1)}(e'_{i,j,h-1}), & x'_{i,j} \in D_{i,1,h} \\
    w(x'_{i,j} - x_i, r) \cdot w^{(h-1)}(e'_{i,j,h-1}) \cdot w^{(h)}(e'_{i,j,h}), & x'_{i,j} \in D_{i,2,h} 
    \end{cases}
\end{align*}
\]

where \(D_{i,1,h}\) and \(D_{i,2,h}\) are the cluster 1 and cluster 2 in \(h^{th}\) iteration at estimated point \(x^{(e)}_i\), respectively. In this way, multiple outliers with different amplitudes can be suppressed step by step. In Fig.10(a), the furthest outlier is classified into the first cluster owing to its weighted residual is 0 through the first iteration step. Then in the second step, the secondly furthest outlier is classified into the second cluster. And the pre-fitting result is closer to the ideal curve. In Fig.10(d), through several iteration steps, all outliers are suppressed and the difference between estimated value and ideal value is much smaller than that in the previous iteration.

**Fig. 10.** The fitting procedure of KMTLS with iterative form

Fig.11 shows the relationship between the iteration number and the sum of
weighted residual $e'_{i,j,h}$ in the OLS pre-fitting procedure. It can be found that the residual drops rapidly in the first three steps, meaning that the outliers are suppressed in this stage according to Fig.10(a)-(c). After the third iteration step, the degree of reduction on the residual error decreases significantly because the random errors are mainly suppressed in this stage. If the iteration procedure is conducted continually (iteration number >7 in Fig.11), the residual will be close to 0 because the weights of almost all points in the support domain are weakened, which will make the fitting procedure unstable.

![Fig. 11. Relationship between iteration and the sum of weighted residual](image)

To deal with this problem, the iteration procedure should be stopped after all the outliers are processed. Therefore, the ratio $L_1/L_2$ based on the clustering result is proposed as the criterion for quitting the iteration procedure. As shown in Fig.12, $L_1$ is the difference between origin and centroid 1, $L_2$ is the difference between centroid 1 and centroid 2, and $e'_{i,j}$ is the weighted residual mentioned above. When the residual difference between the two steps is lower than $L_1/L_2$, the iteration will be stopped. In the first three iteration steps, because of the existence of outliers, the distance between two clusters is much larger than the ratio $L_1/L_2$ so that the iteration step will continuously be conducted. In the stage of processing random errors, the residual difference between two iteration steps is much smaller than that in the stage of processing outliers, which makes the iteration be terminated after the outliers and large random errors are processed.
Fig. 12. The definition of $L_1$ and $L_2$

Take Case 1 as an example and add many more outliers than the previous situation. The maximum iteration number is set to 10. The radius of support domain is the same as Case 1 and the number of discrete points increases from 200 to 500. Fig.13 shows the fitting result of Case 1 with multiple outliers. Compared with the fitting results by three methods in Fig.9, iterative KMTLS shows great robustness for handling multiple outliers.

Fig. 13. The fitting result by iterative KMTLS with multiple outliers

6. Conclusions

In this paper, a novel MTLS-based reconstruction method combined with k-means clustering (KMTLS) is proposed to enhance the robustness and accuracy of MTLS in handling the measurement data with outliers. The proposed KMTLS adopts a two-step fitting procedure to automatically detect and weaken the influence of outlier without
setting a threshold value. The pre-fitting procedure uses the OLS method to evaluate
the distribution of discrete points within the support domain. By conducting k-means
clustering method on the residuals obtained in the pre-fitting procedure, outliers or the
points with large random errors will be classified into one cluster (cluster 2) while the
other normal points will be classified into another cluster (cluster 1). A weight
function defined by the clustering result is proposed to adjust the weight of the points
in cluster 2. Through the above procedures, WTLS is used to calculate the estimated
value. Through simulation and experiment measurement data validation, KMTLS
shows higher fitting accuracy even though there are no obvious outliers in the data.
Moreover, when multiple outliers exist in the support domain, an iterative KMTLS
shows great robustness by processing the outliers step by step.

Acknowledgements

This work was supported by the National Natural Science Foundation of China
(Grant No. 51605094 and 11572316), the Thousand Young Talents Program of China,
and the Fundamental Research Funds for the Central Universities (Grant No.
WK2090050042).

References

https://doi.org/10.1109/JAS.2019.1911579.
correlated coordinates to a straight line, Meas. Sci. Technol. 22 (2011) 035101.
https://doi.org/10.1088/0957-0233/22/3/035101.
approximation on an equispaced grid and Mock–Chebyshev subset interpolation, Appl. Math.


X. Jiang, P. Liu, Z. Li, Data reconciliation and gross error detection for operational data in power plants, Energy 75 (2014) 14-23. https://doi.org/10.1016/j.energy.2014.03.024.


