scientific reports

Check for updates

OPEN

A multivariate process quality correlation diagnosis method based on grouping technique

Qing Niu[1,2]✉, Shujie Cheng[1,2] & Zeyang Qiu[1,2]

Correlation diagnosis in multivariate process quality management is an important and challenging issue. In this paper, a new diagnostic method based on quality component grouping is proposed. Firstly, three theorems describing the properties of the covariance matrix of multivariate process quality are established based on the statistical viewpoint of product quality, to prove the correlation decomposition theorem, which decomposes the correlation of all the quality components into a series of correlations of components pairs, and then by using the factor analysis method, all quality components are grouped in order to maximize the correlations in the same groups and minimize the ones between different groups. Finally, on the basis of correlations between different groups are ignored, T^2 control charts of component pairs in the same groups are established to form the diagnostic model. Theoretical analysis and practice prove that for the multivariate process quality whose the correlations between different components vary considerably, the grouping technique enables the size of the correlation diagnostic model to be drastically reduced, thus allowing the proposed method can be used as a generalized theoretical model for the correlation diagnosis.

Keywords Multivariate process quality, Correlation diagnosis, Grouping technology, Factor analysis, T^2 control chart

With the development of the modern global market, the product's quality has been one of the key factors that greatly influence the competitiveness of enterprises. In the whole formation of the product's quality, process quality is one of the most basic sessions because the product's quality will be influenced by every process' quality directly or indirectly, so process quality control is the essence of quality management in manufacturing.

The objective of managing univariate process quality was achieved by using *Shewhart's* control chart, which is a tool in the theory of statistical process control (SPC)[1–4]. But in modern manufacturing, there are many processes that involve more than one quality component. Due to the correlation of quality components, all components and their correlation must be monitored simultaneously[5,6]. The theory of monitoring the correlation shift of all the quality components using T^2 control charts was originally proposed by *Hotelling*[7]. For a p-dimensional process quality $\boldsymbol{y}=(y_1, y_2,\ldots, y_p)^T$, the T^2 statistic is defined as:

$$T^2 = (\mathbf{y}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu}) \tag{1}$$

where $\boldsymbol{\mu}$ is the mean vector, and $\boldsymbol{\Sigma}$ is the covariance matrix of $\boldsymbol{y}$. When $T^2>0$, it signifies that all the quality components in $\boldsymbol{y}$ are correlated.

The general distribution of T^2 statistics can have different forms[8–10]. Particularly, when $\boldsymbol{y}$ follows the normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the T^2 statistic follows the χ^2 distribution with p-freedom, and the proof can be found in Supplementary. Suppose α is the false probability, then the upper control limit (UCL) of the T^2 statistic is $\chi^2_\alpha(p)$, and the lower control limit (LCL) is 0. Thus, the T^2 control chart can be established to monitor the correlation shift of $\boldsymbol{y}$. T^2 control chart has the advantage of being able to fully take into account the correlation between components and gives accurate false probability under condition of component correlation, however, this control chart is unable to pinpoint the cause(s) of the correlation shift when it is out of control. Since then, on the basis of T^2 statistic, scholars have carried out a lot of research on the diagnostic methods of abnormal correlation shift between quality components, and have successively proposed diagnostic methods based on component combinations, principal component analysis, orthogonal decomposition of the T^2 statistics, and intelligent diagnostic methods.

[1]Department of Product Design, Lanzhou Jiaotong University, Lanzhou, Gansu, People's Republic of China. [2]These authors contributed equally: Qing Niu, Shujie Cheng and Zeyang Qiu. ✉email: liuqing@mail.lzjtu.cn

Diagnosis method based on component combinations

For a p-dimensional process quality $\boldsymbol{y}=(y_1, y_2,\ldots, y_p)^T$, when the T^2 control chart K, which monitors the correlation shift of all the quality components shows an abnormality, it indicates that the correlations of one or more quality component combinations must be abnormal. In order to diagnose the specific component combinations that cause the T^2 control chart K to be abnormal, a straightforward approach is to use the exhaustive method, i.e., to list all possible forms of component combinations, and for each form of combinations, to create a T^2 control chart. When the T^2 control chart K displays an abnormality, the specific component combinations that lead to the abnormality of the T^2 control chart K can be determined by analyzing the results of the T^2 control charts corresponding to all combinations of quality components one by one[11–13], this approach is referred as component combinations based diagnostic (CCBD) method in this paper. While this approach is theoretically sound and appealing, it has inherent deficiencies. For a p-dimensional process quality $\boldsymbol{y}=(y_1, y_2, \ldots, y_p)^T$, the number of T^2 control charts using this method is $N = C_p^2 + C_p^3 + \cdots + C_p^p = 2^p\text{-}p\text{-}1$, where N is an exponential function of p, and the space complexity is $O(2^p)$. When p is small, this approach has some feasibility, however, when p increases, N will increase sharply, leading to a significant expansion of the diagnostic system scale, so this method is difficult to apply in practice.

On the other hand, the defect of information redundancy in diagnostic results can not be avoided in the CCBD method. For example, in a 4-dimensional process quality $\boldsymbol{y}=(y_1, y_2, y_3, y_4)^T$, suppose the abnormal correlation shift between y_1 and y_2 is the only cause which causes the correlation of $\boldsymbol{y}$ out of control. Now in the CCBD method, besides T^2 control chart to monitor the correlation shift of (y_1, y_2) is out of control, the other combinations which contain y_1 and y_2, namely (y_1, y_2, y_3) and (y_1, y_2, y_4), their T^2 control charts are both out of control. This phenomenon that because of the correlation shift of one component combination is out of control, the correlations of other component combinations which contain the abnormal component combination are all out of control is called as the redundancy of diagnostic messages. The redundancy in diagnostic results is disturbance for process quality adjustment.

Diagnosis method based on principal component analysis

When the number of quality components to be monitored in manufacturing process is large, direct analysis of process quality data will lead to a significant increase in the computational effort of the diagnostic process. Therefore, reducing the complexity of process quality data in an appropriate means is an effective way to improve diagnostic efficiency. Because the principal component analysis (PCA) is a useful tool for dealing with high-dimensional data, scholars proposed by using the principal component analysis method[14–17], the original process quality $\boldsymbol{y}=(y_1, y_2,\ldots, y_p)^T$ is converted to p independent principal components and sorted by variance decreasing order, denoted as $\boldsymbol{z}=(z_1, z_2, \ldots, z_p)^T$. Then firstly, p *Shewhart* control charts are constructed to monitor the normality of z_i; Secondly, the first $n(n<p)$ principal components whose the cumulative sum of their variance exceeds a specified critical value are grouped as component pairs, and T^2 control charts are constructed to monitor the normality of (z_i, z_j) $(i, j \le n, i \ne j)$; At last, the normality of the rest principal components group $(z_{n+1}, z_{n+2},\ldots, z_p)$ is monitored by a T^2 control chart.

Compared with the CCBD method, the number of control charts based on PAC method is $N = p + C_n^2 + 1 = p + n(n-1)/2 + 1$, the space complexity approximately is $O(p^2)$, and the diagnostic efficiency is improved. However, n still increases rapidly while p is increasing, the scale of diagnostic system is still large. Furthermore, due to z_i generally has no engineering meaning after conversion, the cause(s) which cause the correlation shift of $\boldsymbol{y}$ out of control can only be specified by a comprehensive analysis of all the results in control charts and consulting the mapping relationship between $\boldsymbol{y}$ and $\boldsymbol{z}$, the calculation of diagnosis is increased, and the accuracy of diagnostic results is affected. Meanwhile, the redundancy of diagnostic messages also can not be avoided.

Diagnosis method based on correlation orthogonal decomposition

In 1995, *Mason*, *Young* and *Tracy*[18–20] proposed by using regression analysis method, the T^2 statistic can be decomposed into conditional and unconditional terms which have equal weight in the decomposition results and are orthogonally independent each other. Then, according to the statistical distribution of the conditional and unconditional terms, the corresponding control limits are established to diagnose the specific cause(s) when the manufacturing process is abnormal. Compared with the diagnostic methods based on principal component analysis, the conditional and unconditional terms obtained by MYT orthogonal decomposition method can be directly corresponded to the quality components or component combinations, which improves the accuracy of the diagnostic results.

As an example, in bivariate process quality $\boldsymbol{y}=(y_1, y_2)^T$, the basic idea of the MYT orthogonal decomposition method[21–24] is to decompose the T^2 statistic into the following form:

$$T^2 = T_1^2 + T_{2\cdot 1}^2 \tag{2}$$

where T_1^2, called the unconditional term, is related only to the quality component y_1 and is used to measure the contribution shift in y_1 to the T^2 statistic; and $T_{2\bullet 1}^2$, called the conditional term, whose value is related to the conditional probability $P(y_2|y_1)$ and is used to measure the contribution in the correlation between y_1 and y_2 to the T^2 statistic.

Similar to Eq. (2), the T^2 statistic can also be decomposed into another form:

$$T^2 = T_2^2 + T_{1\cdot 2}^2 \tag{3}$$

where the unconditional term T_2^2 is related only to the quality component y_2, and is used to measure the contribution shift in y_2 to the T^2 statistic; the conditional term $T_{1\bullet 2}^2$ depends on the conditional probability $P(y_1|y_2)$, and is used to measure the contribution in the correlation between y_2 and y_1 to the T^2 statistic.

Conditional probability $P(y_2|y_1) \neq P(y_1|y_2)$ when y_1 and y_2 are correlated, and hence the conditional term $T_{2\cdot 1}^2 \neq T_{1\cdot 2}^2$. For this reason, Eqs. (2) and (3) are two distinct representations of the T^2 statistic's decomposition results. In general, for a p-dimensional process quality $\boldsymbol{y} = (y_1, y_2, \ldots, y_p)^T$, the decomposition results have a total of $p(p\text{-}1)\ldots \times 2 \times 1$, and the space complexity is $O(p!)$. As the number of quality components increases, under the condition that every possible form of decomposition is analyzed, will lead to a significant increase of calculations and a serious reduction in diagnostic efficiency. At the same time, the accuracy of the diagnostic results based on this method will be affected when there are obvious correlations between different quality components.

Intelligent diagnosis methods

In addition to the traditional diagnostic methods based on mathematical model analysis, in recent years, with the development of artificial intelligence technology, intelligent diagnostic methods are applied to the field of multivariate process quality diagnosis, and the diagnostic methods based on artificial neural network (ANN)[25–28], Bayesian network[29–32], support vector machine (SVM)[33–35], etc. have been widely applied. Intelligent diagnostic methods can effectively reduce the scale of the diagnostic system and improve the diagnostic efficiency, however, these methods generally require a large amount of data to train the network's parameters, and the constructed network are generally suitable for specific applications, thus their generality will be greatly restricted. Therefore, establish a general and efficient method for multivariate process quality correlation diagnosis is a major problem to be solved in the field of quality management.

Sketch of the algorithm

In this paper, a new correlation diagnosis method based on quality component grouping is proposed. For the multivariate process quality $\boldsymbol{y} = (y_1, y_2, \ldots, y_p)^T$, three theorems describing the properties of the multivariate process quality covariance matrix are first established based on the statistical viewpoint of product quality in manufacturing processes; Then the correlation decomposition theorem is proved by drawing on the idea of decomposing the T^2 statistic in the MYT orthogonal decomposition method, which decomposes the correlation of all the quality components into the correlations of all the component pairs, to reduce the space complexity of the diagnostic system to $O(p^2)$; Next, refer to the grouping idea in the principal component analysis method, based on the correlation between different components, the quality components are grouped, so that the correlations between components in the same groups are as large as possible, and the correlations between components of different groups are as small as possible; Finally, draw on the principle of component combination diagnosis method, on the premise of ignoring the correlations between different groups, quality components in the same groups are combined as component pairs to establish the corresponding T^2 control charts, which constitutes the multivariate process quality correlation diagnostic system, thus the space complexity of the diagnostic system is reduced to approximate $O(p)$, to improve the diagnostic efficiency.

Covariance matrix properties of multivariate process quality

In the manufacturing process, factors affecting the product's quality can be attributed to 5 aspects: man, machines, materials, methods and environment (4M1E). On this basis, ISO9000 supplemented another 3 factors: the manufacturing software, auxiliary materials and utilities. Among the many factors affecting the product's quality, changes in any one of them will have an impact on the final quality of the product, so the product's quality is fluctuating in manufacturing. Tolerance theory is a direct proof of the fluctuation of the product's quality.

For the multivariate process quality $\boldsymbol{y} = (y_1, y_2, \ldots, y_p)^T$, the covariance matrix is an important parameter to describe its correlation. Combined with the fluctuation of the product's quality in the manufacturing process, this paper firstly establishes 3 theorems describing the characteristics of the covariance matrix of multivariate process quality.

Theorem 1 *In the covariance matrix $\boldsymbol{\Sigma}$ of the multivariate process quality $\boldsymbol{y} = (y_1, y_2, \ldots, y_p)^T$, all of the elements are not* 0.

Suppose the mean vector of $\boldsymbol{y}$ is $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_p)^T$. According to the definition of the covariance matrix, it is known that:

$$\begin{aligned}\boldsymbol{\Sigma} &= \begin{bmatrix} E[(y_1-\mu_1)(y_1-\mu_1)] & E[(y_1-\mu_1)(y_2-\mu_2)] & \cdots & E[(y_1-\mu_1)(y_p-\mu_p)] \\ E[(y_2-\mu_2)(y_1-\mu_1)] & E[(y_2-\mu_2)(y_2-\mu_2)] & \cdots & E[(y_2-\mu_2)(y_p-\mu_p)] \\ \vdots & \vdots & \vdots & \vdots \\ E[(y_p-\mu_p)(y_1-\mu_1)] & E[(y_p-\mu_p)(y_2-\mu_2)] & \cdots & E[(y_p-\mu_p)(y_p-\mu_p)] \end{bmatrix} \\ &= \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{21} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{p1} & \sigma_{p1} & \cdots & \sigma_{pp} \end{bmatrix}\end{aligned} \tag{4}$$

For any element $\sigma_{ij} = E[(y_i - \mu_i)(y_j - \mu_j)]$ in $\boldsymbol{\Sigma}$, the sufficient and necessary condition for it to be 0 is:

$$y_i = \mu_i \quad \text{or} \quad y_j = \mu_j \tag{5}$$

According to the properties of mathematical expectation, Eq. (5) implies that the quality component y_i or y_j is a constant in the manufacturing process. Clearly, this is in conflict with the viewpoint of the fluctuation of the product's quality, and therefore, Eq. (5) does not hold, i.e., all the elements in $\boldsymbol{\Sigma}$ are not 0.

Theorem 2 *The covariance matrix* $\boldsymbol{\Sigma}$ *of the multivariate process quality* $\boldsymbol{y} = (y_1, y_2, \ldots, y_p)^T$ *is a real symmetric positive definite matrix.*

According to Eq. (4) on the definition of the covariance matrix:

$$\sigma_{ij} = E[(y_i - \mu_i)(y_j - \mu_j)]$$
$$\sigma_{ji} = E[(y_j - \mu_j)(y_i - \mu_i)]$$

From the properties of mathematical expectation can be seen:

$$\sigma_{ij} = \sigma_{ji}$$

That is, $\boldsymbol{\Sigma}$ is a symmetric matrix.
Let p-dimensional vector $\boldsymbol{c} = (c_1, c_2, \ldots, c_p)^T \neq \mathbf{0}$.

$$\mathbf{c}^T \Sigma \mathbf{c} = (c_1, c_2, \cdots c_p)\boldsymbol{\Sigma}(c_1, c_2, \cdots c_p)^T \tag{6}$$

Bringing Eq. (4) into (6), after simplification and consolidation, we get

$$\mathbf{c}^T \Sigma \mathbf{c} = E\left[\left(\sum_{i=1}^{p} c_i(y_i - \mu_i)\right)\left(\sum_{k=1}^{p}(y_k - \mu_k)c_k\right)\right] \tag{7}$$

Let random variable $z = \sum_{i=1}^{p} c_i(y_i - \mu_i)$, bringing this into Eq. (7), we get

$$\mathbf{c}^T \Sigma \mathbf{c} = E(z^2) \geq 0$$

From the proof of Theorem 1, it is clear that according to the viewpoint of the fluctuation of the product's quality, $z \neq 0$, i.e.

$$\mathbf{c}^T \Sigma \mathbf{c} = E(z^2) > 0$$

Therefore, the covariance matrix $\boldsymbol{\Sigma}$ of the multivariate process quality $\boldsymbol{y} = (y_1, y_2, \ldots, y_p)^T$ is a real symmetric positive definite matrix.

Theorem 3 *The inverse matrix* $\boldsymbol{\Sigma}^{-1}$ *of the covariance matrix* $\boldsymbol{\Sigma}$ *of the multivariate process quality* $\boldsymbol{y} = (y_1, y_2, \ldots, y_p)^T$ *is a real symmetric positive definite matrix.*

First prove the symmetry of $\boldsymbol{\Sigma}^{-1}$. It follows from the symmetry of $\boldsymbol{\Sigma}$:

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^T$$

Inverting both ends of the above equation:

$$\boldsymbol{\Sigma}^{-1} = (\boldsymbol{\Sigma}^T)^{-1} = (\boldsymbol{\Sigma}^{-1})^T$$

The above equation shows that $\boldsymbol{\Sigma}^{-1}$ is a symmetric matrix.
Let the eigenvalues of $\boldsymbol{\Sigma}$ be $\lambda_1, \lambda_2, \ldots, \lambda_p$. By the positive definiteness of $\boldsymbol{\Sigma}$, $\lambda_i > 0$ $(i \leq p)$. According to the nature of the inverse matrix, the eigenvalues of $\boldsymbol{\Sigma}^{-1}$ are $1/\lambda_1, 1/\lambda_2, \ldots, 1/\lambda_p$, i.e., the eigenvalues of $\boldsymbol{\Sigma}^{-1}$ are all greater than 0, so $\boldsymbol{\Sigma}^{-1}$ is a positive definite matrix.

Theoretical basis for correlation grouping diagnosis

The exponential function between N and p is the main reason why applying this approach is difficult in the CCBD method. If the gradient of N with p can be lowered by proper means, the defect of diagnostic system scale expands greatly while p is increasing will be avoided to a certain extent, and thus this approach can be applied in multivariate process quality management.

Correlation decomposition

Theorem 4 In the multivariate process quality $\boldsymbol{y}=(y_1, y_2,\ldots, y_p)^T$, the sufficient and necessary condition of the correlation of all the components exists is, for any two components y_i and y_j, they are correlated.

Firstly, the sufficiency of Theorem 4 is proved. Any two components y_i and y_j in $\boldsymbol{y}$ are correlated shows that $\sigma_{ij}\neq 0$. From Theorems 2 and 3, the covariance matrix $\boldsymbol{\Sigma}$ and its inverse matrix $\boldsymbol{\Sigma}^{-1}$ are real symmetric positive definite matrix. From the definition of the T^2 statistic in Eq. (1), it is clear that for any sample data, its T^2 statistic is greater than 0, i.e., the correlation of all the components exists.

The following proves the necessity of Theorem 4 by reduction and absurdum. The existence of correlation of all the components in $\boldsymbol{y}$ implies that for any sample data, its T^2 statistic is greater than 0. From the definition of the T^2 statistic in Eq. (1), there exists an inverse matrix of the covariance matrix $\boldsymbol{\Sigma}$ of $\boldsymbol{y}$, and the rank of $\boldsymbol{\Sigma}$ is p.

$$R(\boldsymbol{\Sigma}) = p \tag{8}$$

Assume y_k and y_j in $\boldsymbol{y}$ are uncorrelated, i.e., $\sigma_{kj}=0$. By the definition of covariance, there is:

$$\sigma_{kj} = E[(y_k - \mu_k)(y_j - \mu_j)] = 0 \tag{9}$$

The sufficient and necessary condition for Eq. (9) to hold is $y_k=\mu_k$ or $y_j=\mu_j$. It may be useful to set $y_k=\mu_k$. From the definition of covariance, we know that for any component y_i ($i=1, 2, \ldots, p$), there are:

$$\sigma_{ki} = E[(y_k - \mu_k)(y_i - \mu_i)] = 0 \tag{10}$$

Equation (10) shows that in the covariance matrix $\boldsymbol{\Sigma}$ of $\boldsymbol{y}$, the kth row and kth column are both 0, i.e., $R(\boldsymbol{\Sigma})\leq p-1$. This contradicts Eq. (8), the assumption is not valid, and the necessity of Theorem 4 is proved.

Theorem 4 means that the correlation of all the quality components can be represented as correlations of component pairs, so in the correlation diagnostic system, it only needs to monitor the correlation shifts of all the component pairs. In addition, T^2 control chart to monitor the correlation shift of all the components should be added, the number of T^2 control charts is $N=C_p^2+1=p(p-1)/2+1$, N is the power function of p, the space complexity of the diagnostic system is lowered to $O(p^2)$. Compared to the CCBD method, the gradient of N with p is decreased significantly. Meanwhile, because the component pair is the minimum combination of components, the information redundancy in diagnostic results can be avoided effectively.

Grouping principle

Although the functional relation between N and p is lowered to a power function by correlation decomposition, N will still increase rapidly while p is increasing, so further proper ways should be adopted to reduce the scale of the diagnostic system on the basis of the above analysis. For this reason, this paper proposes the following grouping principle.

Theorem 5 *Let $p=p_1+p_2+\ldots+p_m$, where p and p_i ($i=1, 2, \ldots, m$) are integers greater than 0, $m>1$. In this case there is the following inequality*:

$$C_p^2 > \sum_{i=1}^{m} C_{p_i}^2 \tag{11}$$

The proof of Theorem 5 proceeds as follows:

$$\begin{aligned}
\sum_{i=1}^{m} C_{p_i}^2 &= \sum_{i=1}^{m} \frac{p_i(p_i-1)}{2} \\
&= \frac{1}{2}\Big(\sum_{i=1}^{m} p_i^2 - p\Big) \\
&< \frac{1}{2}\Big(\sum_{i=1}^{m} p_i^2 + 2\sum_{\substack{k,j=1 \\ k\neq j}}^{m} p_k p_j - p\Big) \\
&= \frac{1}{2}\Big(\Big(\sum_{i=1}^{m} p_i\Big)^2 - p\Big) \\
&= \frac{1}{2}(p^2 - p) \\
&= C_p^2
\end{aligned}$$

Theorem 5 shows that for the multivariate process quality $\boldsymbol{y}=(y_1, y_2,\ldots, y_p)^T$, if the quality components are grouped according to the degree of correlations, so that the correlations of quality components located within the same groups should be as large as possible, and the correlations of quality components located between

different groups should be as small as possible, the number of T^2 control charts in the diagnostic model can be further reduced by ignoring the correlations of the quality components located in the different groups, and the reduction of the number of T^2 control charts is $\sum_{\substack{k,j=1 \\ k\neq j}}^{m} p_k p_j$, where m is the number of quality components grouped, p_k and p_j denote the number of quality components contained in the kth and jth groups after grouping. In this case, the space complexity of the multivariate process quality correlation diagnostic model based on the grouping technique is approximated as $O(p)$.

Methodology for grouping quality components

Grouping techniques can lead to a significant reduction in the number of T^2 control charts required in the correlation diagnostic system. Typically, quality components can be grouped with reference to practical experience, but this way can not give an accurate estimate of the error before and after grouping. In order to analyze the errors quantitatively, a grouping method based on the analysis of the covariance matrix of the quality components is used here.

Before grouping, the multivariate process quality $\boldsymbol{y}=(y_1, y_2,\ldots, y_p)^T$ needs to be standardized in order to avoid differences in the observed scales from affecting the grouping results:

$$y_i^* = \frac{y_i - \mu_i}{\sigma_i} \tag{12}$$

where μ_i and σ_i are the mean and variance of y_i. In the standardization result $\mathbf{y}^* = (y_1^*, y_2^*, \cdots, y_p^*)^T$, the mean of each component is 0, and the variance is 1.

Factor analysis

Factor analysis is a method of grouping components based on the degree of correlations between different components, using the covariance matrix of a random vector as a reference. The basic model of factor analysis is as follows[36,37]:

(1) The standardized multivariate process quality $\boldsymbol{y}^*$ is an observable random vector with mean vector $E(\boldsymbol{y}^*)=\mathbf{0}$ and covariance matrix $D(\boldsymbol{y}^*)=\boldsymbol{\Sigma}^*$;
(2) The common factor vector $\boldsymbol{F}=(F_1, F_2, \ldots, F_m)^T$ $(m<p)$ is an unobservable random vector with mean vector $E(\boldsymbol{F})=\mathbf{0}$ and covariance matrix $D(\boldsymbol{F})=\boldsymbol{I}$, where $\boldsymbol{I}$ is a diagonal matrix where the main diagonal elements are 1, and the remaining elements are 0, i.e., the components in $\boldsymbol{F}$ are independent of each other;
(3) The error vector $\boldsymbol{\varepsilon}=(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p)^T$ is independent of the common factor vector $\boldsymbol{F}$ with $E(\boldsymbol{\varepsilon})=\mathbf{0}$, and the covariance matrix $\boldsymbol{D}(\boldsymbol{\varepsilon})$ is a diagonal matrix:

$$D(\boldsymbol{\varepsilon}) = \begin{pmatrix} \sigma_{\varepsilon_1}^2 & & & \\ & \sigma_{\varepsilon_2}^2 & & \\ & & \ddots & \\ & & & \sigma_{\varepsilon_p}^2 \end{pmatrix}$$

Under the above conditions, the factor analysis model can be expressed as the following equations:

$$\begin{cases} y_1^* = a_{11}F_1 + a_{12}F_2 + \cdots + a_{1m}F_m + \varepsilon_1 \\ y_2^* = a_{21}F_1 + a_{22}F_2 + \cdots + a_{2m}F_m + \varepsilon_2 \\ \vdots \\ y_p^* = a_{p1}F_1 + a_{p2}F_2 + \cdots + a_{pm}F_m + \varepsilon_m \end{cases} \tag{13}$$

Expressing the above system of equations in matrix form:

$$\mathbf{y}^* = \mathbf{AF} + \boldsymbol{\varepsilon} \tag{14}$$

where a_{ij} in the matrix $\boldsymbol{A}=(a_{ij})_{p\times p}$ is called the factor loading, and its absolute value indicates the degree of dependence between the quality component y_i^* and the common factor F_j. The matrix $\boldsymbol{A}$ formed by all the factor loadings is called the factor loading matrix.

From Eq. (14), calculate the covariance matrix of $\boldsymbol{y}^*$:

$$\boldsymbol{\Sigma}^* = D(\mathbf{y}^*) = D(\mathbf{AF}) + D(\boldsymbol{\varepsilon}) = \mathbf{A}D(\mathbf{F})\mathbf{A}^T + D(\boldsymbol{\varepsilon}) = \mathbf{A}\mathbf{A}^T + D(\boldsymbol{\varepsilon}) \tag{15}$$

On the other hand, by Theorem 2, $\boldsymbol{\Sigma}^*$ is a real symmetric positive definite matrix for which *Cholesky* decomposition is performed:

$$\boldsymbol{\Sigma}^* = \mathbf{G}\mathbf{G}^T \tag{16}$$

where $\mathbf{G} = (\sqrt{\lambda_1}\mathbf{e}_1, \sqrt{\lambda_2}\mathbf{e}_2, \cdots, \sqrt{\lambda_p}\mathbf{e}_p)$, $\lambda_i (i=1, 2, \ldots, p)$ are the eigenvalues of the covariance matrix $\boldsymbol{\Sigma}^*$ with $\lambda_1>\lambda_2>\ldots>\lambda_p$, $\boldsymbol{e}_i$ is the eigenvector corresponding to λ_i.

Comparing Eqs. (15) and (16), it can be seen that if $\boldsymbol{A}=\boldsymbol{G}$, the error vector $\boldsymbol{\varepsilon}=\mathbf{0}$ in Eq. (14), the obtained factor analysis model is accurate, but this means that after standardization, all the quality components in $\boldsymbol{y}^*$ will

be grouped into p groups, i.e., the accurate factor analysis model can only be obtained when the correlations between the quality components in $\boldsymbol{y}^*$ are completely ignored. Therefore, considering the general situation, it is necessary to retain most of the correlations between the quality components, in which case an approximation of the factor loading matrix $\boldsymbol{A}$ is constructed from the first m ($m<p$) columns of the matrix $\boldsymbol{G}$, i.e.:

$$\mathbf{A} \approx (\sqrt{\lambda_1}\mathbf{e}_1, \sqrt{\lambda_2}\mathbf{e}_2, \cdots, \sqrt{\lambda_m}\mathbf{e}_m) \tag{17}$$

Error analysis

The error vector $\boldsymbol{\varepsilon} \neq \mathbf{0}$ when building the factor analysis model from the factor loading matrix derived from Eq. (17), this implies that there must be a certain amount of information loss when grouping the quality components in $\boldsymbol{y}^*$ based on Eq. (14).

In statistics, the total amount of information contained in a random variable is generally measured by its variance. In Eq. (14), let $\boldsymbol{A}=\boldsymbol{G}$, which gives the sum of the variances of the components in $\boldsymbol{y}^*$ under the exact decomposition condition:

$$\sum_{i=1}^{p} D(y_i^*) = \sum_{i=1}^{p} \lambda_i \tag{18}$$

Equation (18) shows that under the condition of exact decomposition, the sum of the information contained in all the quality components in $\boldsymbol{y}^*$ is equal to the cumulative sum of all the eigenvalues of the covariance matrix $\boldsymbol{\Sigma}^*$ of $\boldsymbol{y}^*$.

The factor loading matrix $\boldsymbol{A}$ is then constructed according to Eq. (17), at which point it is given by Eq. (14):

$$\sum_{i=1}^{p} D(y_i^*) = \sum_{i=1}^{m} \lambda_i + \sum_{i=1}^{p} D(\varepsilon_i) \tag{19}$$

Comparing Eq. (18) with Eq. (19) shows that grouping the quality components in $\boldsymbol{y}^*$ with Eq. (14), under the condition of ignoring the correlation of the quality components between different groups, the sum of information loss is $\sum_{i=m+1}^{p} \lambda_i$. Therefore, for a specified error β, the number of quality component group m can be determined by the following inequality:

$$\eta = \frac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \geq 1-\beta \tag{20}$$

where η is the cumulative variance contribution rate of the first m eigenvalues. Empirically, when $\eta > 80\% \sim 85\%$, the number of groupings m can be determined by inequality (20). The value of η can be reasonably adjusted in combination with specific applications, but the basic principle of adjustment is that it should be conducive to the reasonable interpretation of the factor analysis model.

Correlation diagnostic algorithm based on grouping theory

The above analysis is founded on the condition that the mean μ_j and covariance matrix $\boldsymbol{\Sigma}$ of the manufacturing process are given. However, in many applications, these parameters are generally unknown. In this case, the unbiased estimator of the manufacturing process parameters can be calculated from a set of sample data $\boldsymbol{y}_i=(y_{i1}, y_{i2}, \ldots, y_{ip})^T$, ($i=1, 2,\ldots, n$) collected while the process is in stable state.

$$\mu_j = \frac{1}{n}\sum_{i=1}^{n} y_{ij} \quad (j=1,2,\ldots,p) \tag{21}$$

$$\sigma_j = \frac{1}{n-1}\sum_{i=1}^{n} (y_{ij}-\mu_j)^2 \quad (j=1,2,\ldots,p) \tag{22}$$

Then the sample data can be standardized as $\mathbf{y}_i^* = (y_{i1}^*, y_{i2}^*, \cdots, y_{ip}^*)^T$, ($i=1, 2,\ldots, n$), where

$$y_{ij}^* = \frac{y_{ij}-\mu_j}{\sigma_j} \quad (j=1,2,\ldots,p) \tag{23}$$

The covariance matrix $\boldsymbol{\Sigma}^*$ can be calculated by the standardize sample data:

$$\boldsymbol{\Sigma}^* = \frac{1}{n-1}\sum_{i=1}^{n} \mathbf{y}_i^* \mathbf{y}_i^{*T} \tag{24}$$

Based on the above analysis, after grouping the quality components in the standardized multivariate process quality $\boldsymbol{y}^*$ using factor analysis method, on the premise of ignoring the correlation of quality components between different groups, the quality components within the same groups are combined as component pairs, and the corresponding binary T^2 control charts are established to form the multivariate process quality correlation diagnostic model. The space complexity of this diagnostic model is approximated as a linear function of the quality component number p, which can lead to a significant improvement in the efficiency of the diagnosis.

The multivariate process quality correlation diagnostic model based on grouping technique can be constructed as follows:

(1) Collect sufficient quality data $\boldsymbol{y}_i (i = 1, 2, \ldots, n)$ while the manufacturing process is in stable state;
(2) Calculate the manufacturing parameters according to Eqs. (21)–(24);
(3) Calculate the eigenvalues of the covariance matrix $\boldsymbol{\Sigma}^*$ and arrange all the eigenvalues in descending order as $\lambda_1 > \lambda_2 > \cdots > \lambda_p$;
(4) Calculate the eigenvector $\boldsymbol{e}_i$ corresponding to the eigenvalue $\lambda_i (i = 1, 2, \ldots, p)$;
(5) For the given error β, calculate the number m of eigenvectors for constructing the factor loading matrix according to inequality (20), and then construct the factor loading matrix $\boldsymbol{A}$ from the first m eigenvectors according to Eq. (17);
(6) Group all the quality components according to Eqs. (13), and the grouping results are recorded as $G_1, G_2, \ldots, G_m$;
(7) For each pair of components (y_s^*, y_t^*) $(s \neq t)$ in G_k $(k = 1, 2, \ldots, m)$, build the corresponding T^2 control chart K_{st};
(8) Establish the T^2 control chart K to monitor the correlation shift of all the quality components.

In the manufacturing process, if the T^2 statistic of the new sample data exceeds the control limit in the control chart K, it indicates that the correlation shift of all the quality components is abnormal, and the cause(s) can be specified by examining the rest binary T^2 control charts in the diagnostic model.

Case study

Blades are important parts in steam turbines and aviation engines, and their machining quality directly affects the life and performance of the equipment. The contour method is a commonly used blade quality inspection technique, and its basic principle is to measure a number of cross-section contour lines of the blade along the height direction (Z-axis direction) in the way shown in Fig. 1, and then match the actual contour lines measured in different height directions with their respective theoretical contour lines by translational and rotational transformations as shown in Fig. 2, so as to decompose the blade profiling error into 4 quality components: blade contouring error before matching, blade contouring error after matching, blade positional error, and blade torsion error.

The machining process shows that the 4 quality components are correlated, so it is necessary to monitor the correlation shift during the manufacturing process and to diagnose the causes of the abnormal correlation. Here, a T^2 control chart is used to monitor the correlation shift of the 4 quality components, and the method proposed in this paper is used to diagnose the causes of the abnormal correlation.

Parameters estimation

The above 4 quality components are expressed in vector form as $\boldsymbol{y} = (y_1, y_2, y_3, y_4)^T$. 15 sample data are collected at the cross-section height $Z = 25$ mm as shown in Table 1, in order to estimate the mean vector and covariance matrix for the manufacturing process.

Experience shows that the 4 quality components to be monitored generally follow normal distribution. In order to check the normality of the sample data, set the confidence level $\alpha_t = 0.95$, and the *Shapiro–Wilk* test is done on the data of the 4 quality components in Table 1, and the results are shown in Table 2. It can be seen that

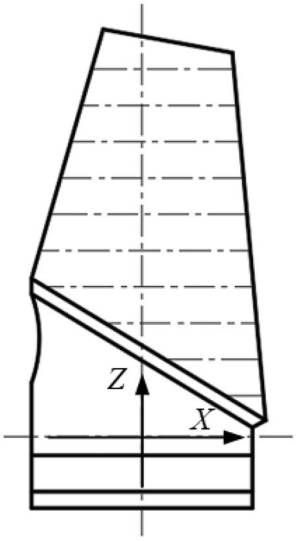

Figure 1. Contour method of blade inspection.

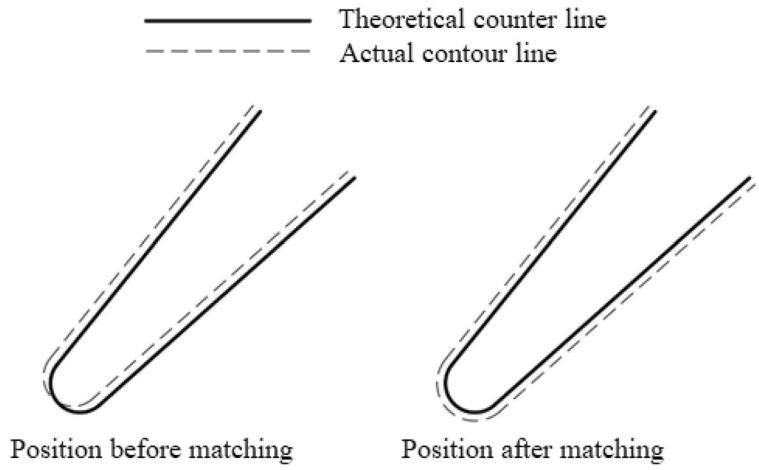

Figure 2. Theoretical and actual contour lines of the blade.

No	Blade contouring error before matching (y_1/mm)	Blade contouring error after matching (y_2/mm)	Blade positional error (y_3/mm)	Blade torsion error (y_4/')
1	0.083	0.043	0.098	2.018
2	0.086	0.045	0.099	2.216
3	0.084	0.045	0.098	2.239
4	0.077	0.036	0.093	2.108
5	0.076	0.034	0.091	2.156
6	0.081	0.042	0.095	2.332
7	0.078	0.037	0.093	2.259
8	0.087	0.046	0.101	2.184
9	0.085	0.044	0.099	2.247
10	0.089	0.051	0.105	2.318
11	0.075	0.034	0.089	2.287
12	0.088	0.047	0.104	2.047
13	0.074	0.033	0.086	2.275
14	0.082	0.041	0.098	2.066
15	0.080	0.041	0.093	2.307

Table 1. Sample data used for process parameter estimation.

Quality components	*W* statistics	Critical value *W*(15,0.05)
y_1	0.9551	0.881
y_2	0.9476	
y_3	0.9707	
y_4	0.9182	

Table 2. Results of normality test for sample data.

the *W* statistics of the four components are all greater than the critical value $W(15,0.05) = 0.881$, indicating that the sample data in Table 1 follow normal distribution.

The sample data used to estimate the process parameters must be collected while the manufacturing process is in stable state, therefore, for the sample data in Table 1, the probability of false alarm $\alpha = 0.0027$ is set for each quality component with reference to the 3σ principle of the *Shewhart* control chart. According to *Bonferroni* inequality and χ^2 distribution, take the false alarm probability $\alpha_y = 0.025$ of the correlation shift, and establish *Shewhart* control charts for the 4 quality components and T^2 control chart to monitor the correlation shift, as shown in Figs. 3, 4, 5, 6 and 7.

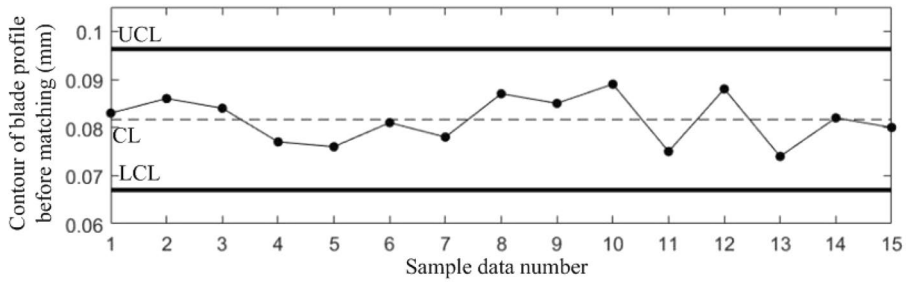

Figure 3. Shewhart control chart for sample data component y_1.

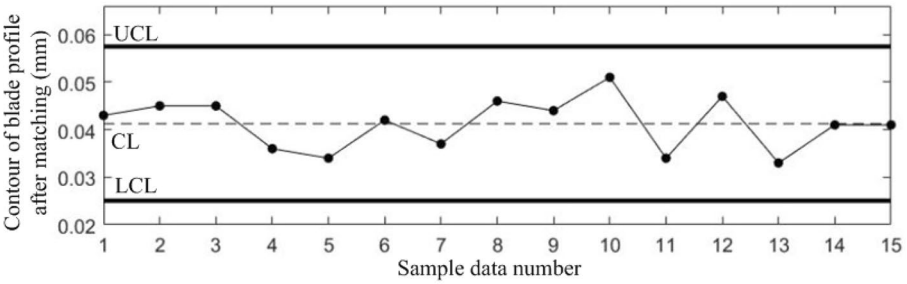

Figure 4. Shewhart control chart for sample data component y_2.

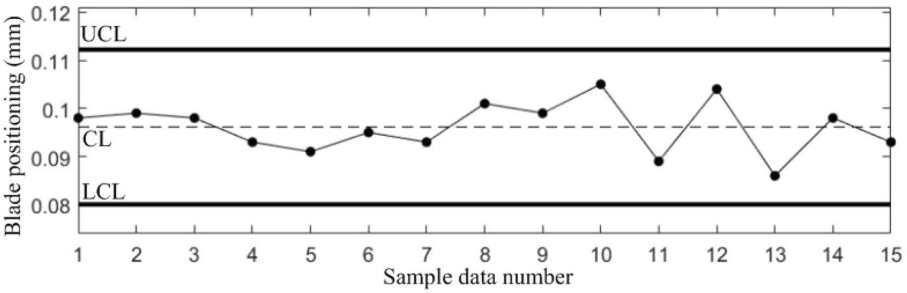

Figure 5. Shewhart control chart for sample data component y_3.

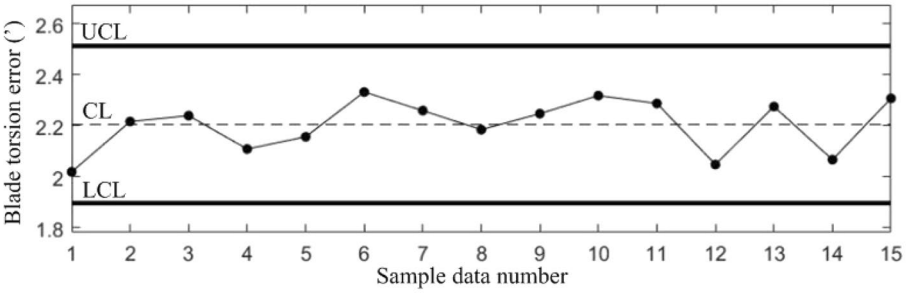

Figure 6. Shewhart control chart for sample data component y_4.

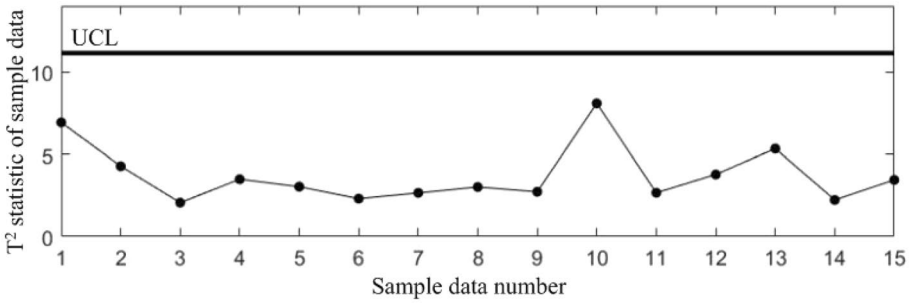

Figure 7. T^2 control chart for sample data.

Figures 3, 4, 5, 6 and 7 show that all the 5 control charts are in normal level, indicating that the sample data in Table 1 are obtained while the blade manufacturing process is in stable state, and can be used for process parameter estimation. The calculated mean vector and standard deviation are:

$$\boldsymbol{\mu} = (0.0817, 0.0413, 0.0961, 2.2039)^T$$

$$\boldsymbol{\sigma} = (0.0029, 0.0054, 0.0054, 0.1029)^T$$

The sample data in Table 1 are standardized and the results are shown in Table 3.
Calculate the covariance matrix from the data in Table 3, we get:

$$\boldsymbol{\Sigma}^* = \begin{pmatrix} 1 & 0.9814 & 0.9747 & -0.1590 \\ 0.9814 & 1 & 0.9510 & -0.0490 \\ 0.9747 & 0.9510 & 1 & -0.2730 \\ -0.1590 & -0.0490 & -0.2730 & 1 \end{pmatrix}$$

Establishment of diagnostic model

Calculate the eigenvalues and eigenvectors of the covariance matrix $\boldsymbol{\Sigma}^*$ and sort all the eigenvalues and eigenvectors in descending order of the eigenvalues, as shown in the second and third columns in Table 4. On this basis, calculate the cumulative contribution rate of the variance of the first 1 to 4 eigenvalues, as shown in the fourth column in Table 4.

As can be seen from Table 4, the first two eigenvalues of the covariance matrix, which have a cumulative contribution rate of variance of 99.11%, are already much higher than the empirical threshold of 80–85%, so let $m = 2$ to construct an approximation of the factor loading matrix $\boldsymbol{A}$ from the first two eigenvectors.

No	y^*_1	y^*_2	y^*_3	y^*_4
1	0.2733	0.3207	0.3485	−1.8061
2	0.8881	0.6908	0.5351	0.1172
3	0.4782	0.6908	0.3485	0.3406
4	−0.9564	−0.9745	−0.5849	−0.9319
5	−1.1613	−1.3445	−0.9583	−0.4656
6	−0.1366	0.1357	−0.2116	1.2440
7	−0.7514	−0.7895	−0.5849	0.5349
8	1.0930	0.8758	0.9085	−0.1936
9	0.6831	0.5057	0.5351	0.4183
10	1.5029	1.8009	1.6552	1.1080
11	−1.3633	−1.3445	−1.3316	0.8069
12	1.2979	1.0608	1.4685	−1.5244
13	−1.5712	−1.5296	−1.8917	0.6903
14	0.0683	−0.0493	0.3485	−1.3399
15	−0.3416	−0.0493	−0.5849	1.0012

Table 3. Sample data after standardization.

No	Eigenvalues	Eigenvectors	Cumulative contribution rate of variance
1	2.9772	$(-0.5751, -0.5653, -0.5747, 0.1396)^T$	74.43%
2	0.9873	$(0.0835, 0.1953, -0.0384, 0.9764)^T$	99.11%
3	0.0236	$(-0.2351, -0.5326, 0.7976, 0.1580)^T$	99.70%
4	0.0118	$(0.7791, -0.5989, -0.1793, 0.0461)^T$	100%

Table 4. Eigenvalues, eigenvectors of the covariance matrix and cumulative contribution of variance.

$$\mathbf{A} = \begin{pmatrix} -0.9923 & 0.0829 \\ -0.9754 & 0.1941 \\ -0.9916 & -0.0382 \\ 0.2409 & 0.9702 \end{pmatrix}$$

$$\begin{cases} y_1^* = -0.9923F_1 + 0.0829F_2 \\ y_2^* = -0.9754F_1 + 0.1941F_2 \\ y_3^* = -0.9916F_1 - 0.0382F_2 \\ y_4^* = 0.2409F_1 + 0.9702F_2 \end{cases}$$

It can be seen that there is a large degree of dependence between components y_1^*, y_2^*, y_3^* and factor F_1, and a smaller degree of dependence with factor F_2, so these 3 quality components are grouped together; y_4^* is only correlated with factor F_2 to a large extent, and therefore will be divided into a group alone. The final result of the grouping is $G_1 = \{y_1^*, y_2^*, y_3^*\}$, $G_2 = \{y_4^*\}$.

For group G_1, T^2 control charts K12, K13 and K23 are built to monitor the binary correlations shift of component pairs (y_1^*, y_2^*), (y_1^*, y_3^*) and (y_2^*, y_3^*); Since there is only one quality component in G_2, there is no need to create a T^2 control chart; Finally, T^2 control chart K that monitors the correlation shift of all the quality components is established, and the 4 control charts are used to form a diagnostic model of the correlation between the 4 quality components in blade processing.

Manufacturing process diagnosis

In subsequent manufacturing, 5 quality data at different moments are collected, as shown in Table 5, and the results after standardization are shown in Table 6. The T^2 statistics for the 5 data were calculated and plotted in the T^2 control chart K, as shown in Fig. 8. It can be seen that in the last three samples, the correlations of all the quality components are abnormal.

In order to diagnose the cause(s) of the abnormal control chart K, 3 control charts monitoring the binary correlation shift of the 3 component pairs shown from Figs. 9, 10 and 11 were analyzed, and the diagnostic results are shown in Table 7.

Validity analysis of diagnostic conclusions

In order to judge the accuracy of the diagnostic results in Table 7, another diagnostic model using the CCBD method is built, which contains a total of $C_4^2 + C_4^3 = 10$ T^2 control charts, as shown in Figs. 12, 13, 14, 15, 16, 17, 18, 19, 20 and 21.

No	Blade contouring error before matching (y_1/mm)	Blade contouring error after matching (y_2/mm)	Blade positional error (y_3/mm)	Blade torsion error (y_4/°)
1	0.085	0.046	0.098	2.176
2	0.077	0.037	0.092	2.148
3	0.082	0.040	0.103	2.068
4	0.085	0.042	0.095	2.312
5	0.074	0.036	0.086	2.277

Table 5. Test data collected in subsequent manufacturing.

No	y^*_1	y^*_2	y^*_3	y^*_4
1	0.6831	0.8758	0.3485	−0.2713
2	−0.9564	−0.7895	−0.7716	−0.5433
3	0.0683	−0.2344	1.2819	−1.3204
4	0.6831	0.1357	−0.2116	1.0497
5	−1.5712	−0.9745	−1.8917	0.7098

Table 6. Test data after standardization.

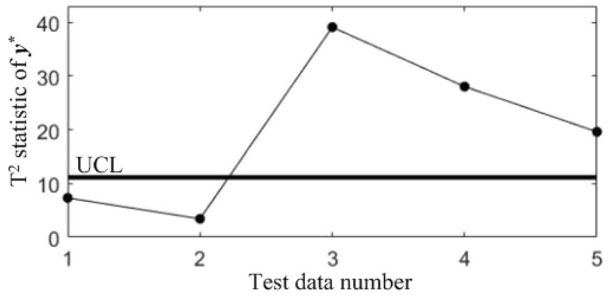

Figure 8. T^2 control chart *K* for test samples.

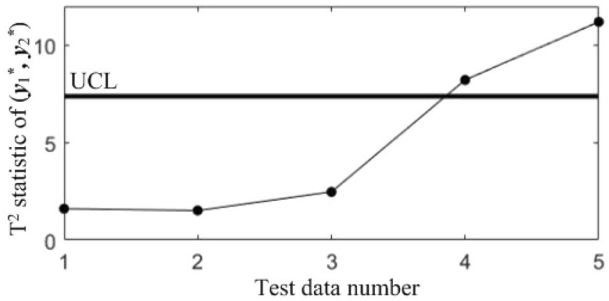

Figure 9. T^2 control chart K12 monitoring the correlation between y_1^* and y_2^*

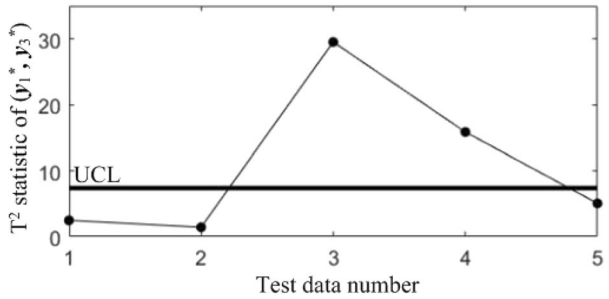

Figure 10. T^2 control chart K13 monitoring the correlation between y_1^* and y_3^*

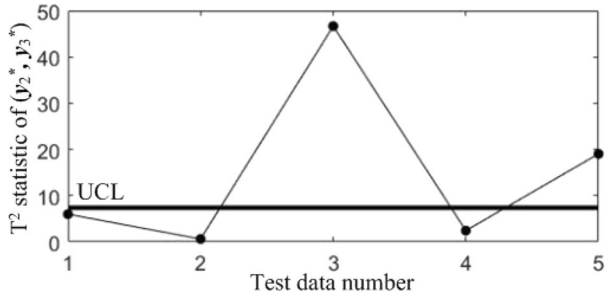

Figure 11. T^2 control chart K23 monitoring the correlation between y_2^* and y_3^*

No	Diagnostic conclusions
1	Normal
2	Normal
3	Anomalous correlations of component pairs (y^*_1, y^*_3), (y^*_2, y^*_3)
4	Anomalous correlations of component pairs (y^*_1, y^*_2), (y^*_1, y^*_3)
5	Anomalous correlations of component pairs (y^*_1, y^*_2), (y^*_2, y^*_3)

Table 7. Diagnostic results.

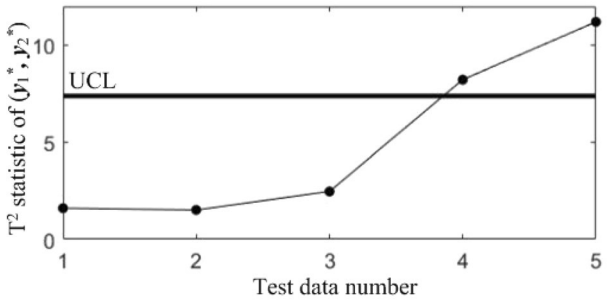

Figure 12. Diagnostic model using CCBD method: T^2 control chart KC_{12} for (y^*_1, y^*_2).

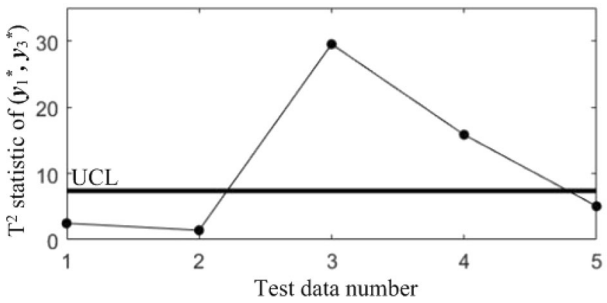

Figure 13. Diagnostic model using CCBD method: T^2 control chart KC_{13} for (y^*_1, y^*_3).

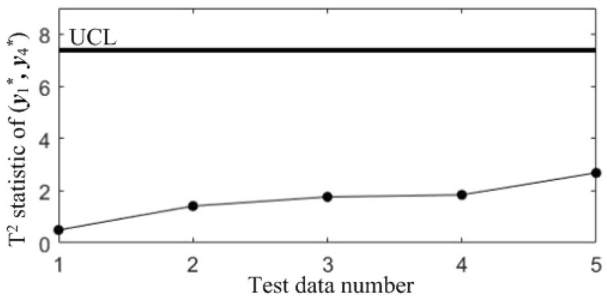

Figure 14. Diagnostic model using CCBD method: T^2 control chart KC_{14} for (y^*_1, y^*_4).

In order to compare the diagnostic conclusions derived from the two different diagnostic models, they are placed in Table 8. It can be seen that there are differences in the diagnostic conclusions of the last 3 points. Taking point 3 as an example, further analysis of the diagnostic results using the CCBD method reveals that since the correlations of component pairs (y^*_1, y^*_3), (y^*_2, y^*_3) are anomalous, the correlations of other component combinations containing (y^*_1, y^*_3) or (y^*_2, y^*_3) are bound to be in anomalous states, and thus the diagnostic results that the correlation abnormalities of component combinations (y^*_1, y^*_2, y^*_3), (y^*_1, y^*_3, y^*_4) and (y^*_2, y^*_3, y^*_4) are redundant diagnostic information. After removing the redundant diagnostic information, the diagnostic results of both

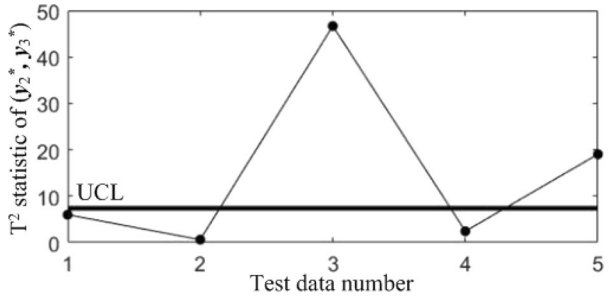

Figure 15. Diagnostic model using CCBD method: T^2 control chart KC_{23} for (y^*_2, y^*_3).

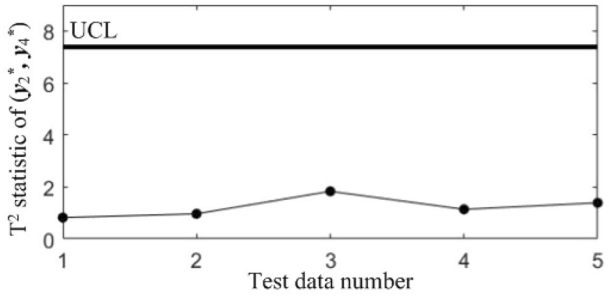

Figure 16. Diagnostic model using CCBD method: T^2 control chart KC_{24} for (y^*_2, y^*_4).

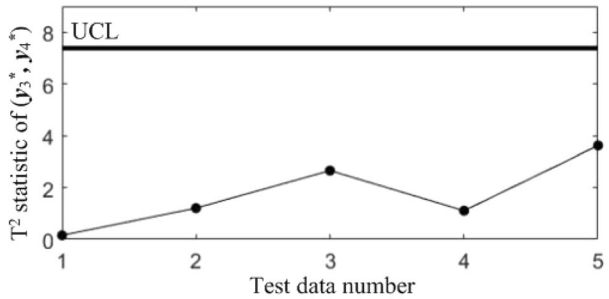

Figure 17. Diagnostic model using CCBD method: T^2 control chart KC_{34} for (y^*_3, y^*_4).

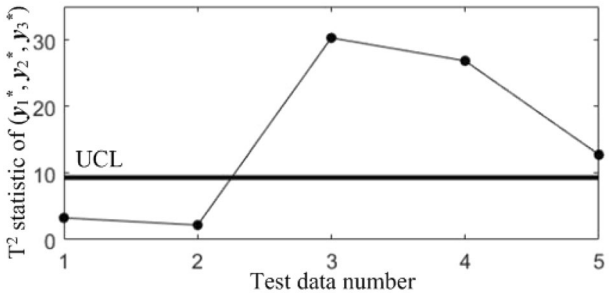

Figure 18. Diagnostic model using CCBD method: T^2 control chart KC_{123} for (y^*_1, y^*_2, y^*_3).

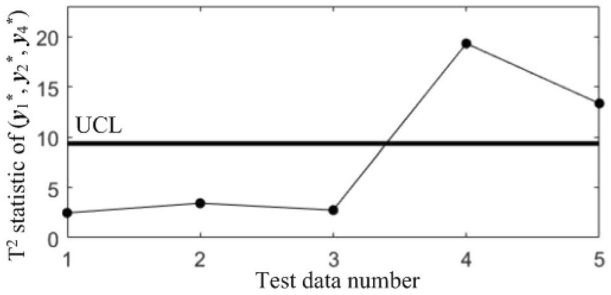

Figure 19. Diagnostic model using CCBD method: T^2 control chart KC_{124} for (y^*_1, y^*_2, y^*_4).

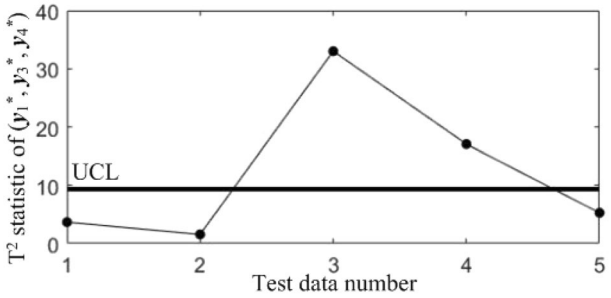

Figure 20. Diagnostic model using CCBD method: T^2 control chart KC_{134} for (y^*_1, y^*_3, y^*_4).

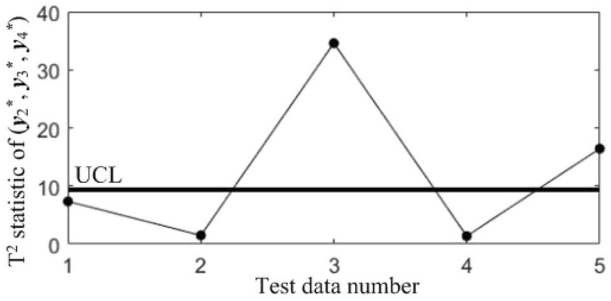

Figure 21. Diagnostic model using CCBD method: T^2 control chart KC_{234} for (y^*_2, y^*_3, y^*_4).

No	Grouping-based diagnostic model	Diagnostic model using CCBD method
1	Normal	Normal
2	Normal	Normal
3	Anomalous correlations of component pairs (y^*_1, y^*_3), (y^*_2, y^*_3)	Anomalous correlations of component combinations (y^*_1, y^*_3), (y^*_2, y^*_3), (y^*_1, y^*_2, y^*_3), (y^*_1, y^*_3, y^*_4), (y^*_2, y^*_3, y^*_4)
4	Anomalous correlations of component pairs (y^*_1, y^*_2), (y^*_1, y^*_3)	Anomalous correlations of component combinations (y^*_1, y^*_2), (y^*_1, y^*_3), (y^*_1, y^*_2, y^*_3), (y^*_1, y^*_2, y^*_4), (y^*_1, y^*_3, y^*_4)
5	Anomalous correlations of component pairs (y^*_1, y^*_2), (y^*_2, y^*_3)	Anomalous correlations of component combinations (y^*_1, y^*_2), (y^*_2, y^*_3), (y^*_1, y^*_2, y^*_3), (y^*_1, y^*_2, y^*_4), (y^*_2, y^*_3, y^*_4)

Table 8. Comparison of the diagnostic results of the two diagnostic models.

diagnostic models for the causes of the anomaly in point 3 are identical. A similar analysis of the diagnostic results for points 4 and 5 leads to the same conclusions, as shown in Table 9. Therefore, the accuracy of the diagnostic method of multivariate process quality correlation based on the grouping technique can be guaranteed.

No	Grouping-based diagnostic model	Diagnostic model using CCBD method
1	Normal	Normal
2	Normal	Normal
3	Anomalous correlations of component pairs (y^*_1, y^*_3), (y^*_2, y^*_3)	Anomalous correlations of component pairs (y^*_1, y^*_3), (y^*_2, y^*_3)
4	Anomalous correlations of component pairs (y^*_1, y^*_2), (y^*_1, y^*_3)	Anomalous correlations of component pairs (y^*_1, y^*_2), (y^*_1, y^*_3)
5	Anomalous correlations of component pairs (y^*_1, y^*_2), (y^*_2, y^*_3)	Anomalous correlations of component pairs (y^*_1, y^*_2), (y^*_2, y^*_3)

Table 9. Comparison of the two diagnostic systems after redundant diagnostic results are removed.

Discussion and conclusion

For the problem of correlation diagnosis in multivariate process quality management, this paper proposed a grouping technique based correlation diagnosis method. Compared with the present diagnostic methods, the method proposed in this paper has the following advantages:

1.1. The diagnosis is more efficient

The space complexity of the multivariate process quality correlation diagnostic method based on grouping technique is approximately $O(p)$, while the space complexity of the diagnostic algorithm based on the CCBD method, principal component analysis method and the orthogonal decomposition of the T^2 statistic are $O(2^p)$, $O(p^2)$, and $O(p!)$, respectively. Therefore, the proposed method in this paper has higher diagnostic efficiency.

2.2. The diagnostic results are more accurate

The grouping technique based multivariate process quality correlation diagnosis method takes the correlation of component pairs as the diagnostic unit. Because component pairs are the minimum combination of quality components, the disadvantage of redundant diagnostic information in diagnostic algorithms based on the CCBD method, the principal component analysis method and the orthogonal decomposition of T^2 statistics can be avoided to provide more accurate diagnostic results for manufacturing processes.

3.3. Better generality

Compared with the diagnostic methods based on artificial intelligence technology, the diagnostic method proposed in this paper is based on strict mathematical analysis as the theoretical foundation, avoids the defect of intelligent diagnostic methods in which the network structure and parameters are oriented to specific application. Therefore, the proposed method can be used as a general theoretical model for the multivariate process quality correlation diagnosis.

The multivariate process quality diagnostic model based on grouping technique has the following two issues for further discussion in its application.

(1) Judgment of the difference degree in correlations between quality components

The difference degree in correlations between quality components can be judged by the covariance matrix $\boldsymbol{\Sigma}^*$ obtained after standardizing the quality data collected in stable state. In general, if there exists at least one row of elements in $\boldsymbol{\Sigma}^*$ such that the ratio of the maximum value to the minimum value, except for the main diagonal element, is not less than 2, it can be tentatively determined that there is a large difference in the correlations between different quality components.

(2) Basis for grouping quality components

The maximum value of each row elements in the factor loading matrix $\boldsymbol{A}$ can be used as a basis for grouping the quality components. The quality component y_i^* can be assigned to group G_k represented by the common factor F_k if a_{ik} is the element with the largest absolute value in the ith row of $\boldsymbol{A}$. Experience has shown that grouping is more desirable when $a_{ik} > 0.7$. When the difference between the absolute values of the elements of a row in $\boldsymbol{A}$ is small, it indicates that the corresponding quality component has an approximately equal degree of dependence on all the common factors, and at this point, the group where the corresponding quality component is located can be rationally determined in conjunction with the actual interpretation of the factor analysis model. If the absolute values between the elements of any row in $\boldsymbol{A}$ are all approximately equal, it indicates that the degree of dependence of all quality components on all common factors is approximately equal, at this time, all quality components are located within a same group, and the diagnostic model is degraded to the diagnostic method based on the correlation decomposition. We will study this issue in depth in our later work.

Ethics declarations
The authors declare no human or animal subjects, sample or database was used in this manuscript.

Data availability
All data generated or analyzed during this study are included in this manuscript.

Received: 19 January 2024; Accepted: 12 May 2024
Published online: 08 June 2024

References
1. Fernandes, F. H., Lee, H. L. & Bourguignon, M. About Shewhart control charts to monitor the Weibull mean. *Qual. Reliab. Eng. Int.* **35**, 2343–2357 (2019).
2. Linda, L. H., Fidel, H. F. & Roberto, C. Q. Improving Shewhart control chart performance for monitoring the Weibull mean. *Qual. Reliab. Eng. Int.* **37**, 984–996 (2021).
3. Huu, D. N., Kim, P. T., Giovanni, C., Petros, E. M. & Philippe, C. On the effect of the measurement error on Shewhart and EWMA control charts. *Int. J. Adv. Manuf. Technol.* **107**, 4317–4332 (2020).
4. Malela-Majika, J. C., Shongwe, S. C., Castagliola, P. & Mutambayi, R. M. A novel single composite Shewhart-EWMA control chart for monitoring the process mean. *Qual. Reliab. Eng. Int.* **38**, 1760–1789 (2022).
5. Mjimer, I., Aoula, E. & Achouyab, E. H. Monitoring of overall equipment effectiveness by multivariate statistical process control. *Int. J. Lean Six Sig.* **13**, 847–862 (2022).
6. Yefang, S., Ijaz, Y., Yueyi, Z. & Hui, Z. Optimizing the quality control of multivariate processes under an improved Mahalanobis-Taguchi system. *Qual. Eng.* **35**, 413–429 (2023).
7. Harold, H. The generalization of student's ratio. *Ann. Math. Stat.* **2**, 360–378 (1931).
8. Mahdiyeh, E., Bahram, S. G. & Mahmoud, R. A. A new approach for monitoring healthcare performance using generalized additive profiles. *J. Stat. Comput. Simul.* **91**, 167–179 (2021).
9. Ali, Y. *et al.* A monitoring framework for health care processes using generalized additive models and auto-encoders. *Artif. Intell. Med.* **146**, 102689 (2023).
10. Mokhtar, M., Wan, Y. & Liang, C. Robust Hotelling's T^2 statistic based on M-estimator. *J. Phys. Conf. Ser.* **1988**, 012116 (2021).
11. Bahrami, H., Niaki, S. T. A. & Khedmati, M. Monitoring multivariate profiles in multistage processes. *Commu. Stat. Simul. C* **50**, 3436–3464 (2019).
12. Joshi, K. & Patil, B. Multivariate statistical process monitoring and control of machining process using principal components based Hotelling T^2 charts: a machine vision approach. *Int. J. Product. Qual. Manag.* **35**, 40–56 (2022).
13. Ershadi, M. J., Niaki, S. T. A., Azizi, A., Esfahani, A. A. & Abadi, R. E. Monitoring data quality using Hoteling multivariate control chart. *Commun. Stat. Simul. C* **52**, 1591–1606 (2023).
14. Huang, J. & Yan, X. Quality-driven principal component analysis combined with kernel least squares for multivariate statistical process monitoring. *IEEE. Trans. Control Syst. Technol.* **27**, 2688–2695 (2019).
15. Li, Q., Xiaoyun, Y., Lina, Y., Yixian, F. & Yuwei, R. Quality-related process monitoring based on improved kernel principal component regression. *IEEE Access* **9**, 132733–132745 (2021).
16. Muhammad, R., Babar, Z., Rashid, M., Nasir, A. & Muazu, A. Advanced multivariate cumulative sum control charts based on principal component method with application. *Qual. Reliab. Eng. Int.* **37**, 2760–2789 (2021).
17. Sun, C. & Hou, J. An improved principal component regression for quality-related process monitoring of industrial control systems. *IEEE Access* **5**, 21723–21730 (2017).
18. Mason, R. L., Tracy, N. D. & Young, J. C. Decomposition of T^2 for multivariate control chart interpretation. *J. Qual. Technol.* **27**, 109–119 (1995).
19. Mason, R. L., Tracy, N. D. & Young, J. C. A practical approach for interpreting multivariate T^2 control chart signals. *J. Qual. Technol.* **29**, 396–406 (1997).
20. Mason, R. L., Tracy, N. D. & Young, J. C. Improving the sensitivity of the T^2 statistic in multivariate process control. *J. Qual. Technol.* **31**, 155–165 (1999).
21. Akeem, A. A., Yahaya, A. & Asiribo, O. Hotelling's T^2 decomposition: Approach for five process characteristics in a multivariate statistical process control. *Am. J. Theor. Appl. Stat.* **4**, 432–437 (2015).
22. Huang, X. H., Xu, J. K. & Zhou, Q. Multi-scale diagnosis of spatial point interaction via decomposition of the *k* function-based T^2 statistic. *J. Qual. Technol.* **49**, 213–227 (2017).
23. Li, X. L. & Liu, S. S. Fault separation and detection algorithm based on Mason Young Tracy decomposition and Gaussian mixture models. *Int. J. Intell. Comput.* **13**, 81–101 (2020).
24. Ueda, R. M. & Souza, A. M. An effective approach to detect the source(s) of out-of-control signals in productive processes by vector error correction (VEC) residual and Hotelling's T^2 decomposition techniques. *Expert Syst. Appl.* **187**, 115979 (2022).
25. Yu, J. B., Zhang, C. Y. & Wang, S. J. Sparse one-dimensional convolutional neural network-based feature learning for fault detection and diagnosis in multivariable manufacturing processes. *Neural Comput. Appl.* **34**, 4343–4366 (2022).
26. Jiao, J. Y., Zhao, M. & Lin, J. A multivariate encoder information based convolutional neural network for intelligent fault diagnosis of planetary gearboxes. *Knowl.-Based Syst.* **160**, 237–250 (2018).
27. Samira, Z. & Moosa, A. Simultaneous fault diagnosis of wind turbine using multichannel convolutional neural networks. *ISA Trans* **108**, 230–239 (2021).
28. Xu, Q. Q., Dong, J. & Peng, K. X. A novel method of neural network model predictive control integrated process monitoring and applications to hot rolling process. *Expert Syst. Appl.* **237**, 121682 (2023).
29. Xian, X. C., Li, J. & Liu, K. B. Causation-based monitoring and diagnosis for multivariate categorical processes with ordinal information. *IEEE Trans. Autom. Sci. Eng.* **16**, 886–897 (2019).
30. Rezki, N., Kazar, O. & Mouss, L. H. A hybrid approach for complex industrial process monitoring. *J. Sci. Ind. Res. India* **76**, 608–613 (2017).
31. Wang, Y. Z., Liu, Y. & Khan, F. Semiparametric PCA and Bayesian network based process fault diagnosis technique. *Can. J. Chem. Eng.* **95**, 1800–1816 (2017).
32. Yao, W. L., Li, D. H. & Gao, L. Fault detection and diagnosis using tree-based ensemble learning methods and multivariate control charts for centrifugal chillers. *J. Build. Eng.* **51**, 104243 (2022).
33. Liang, J. P. & Zhang, K. A new hybrid fault diagnosis method for wind energy converters. *Electronics* **12**, 1263 (2023).
34. Zhang, H. Q., Wang, J. C. & Wang, M. Integration of cuckoo search and fuzzy support vector machine for intelligent diagnosis of production process quality. *J. Ind. Manag. Optim.* **18**, 195–217 (2022).
35. Tang, J. & Zhao, Q. N. Motor rolling bearing fault diagnosis based on MVMD energy entropy and GWO-SVM. *J. Vibroeng.* **25**, 1096–1107 (2023).
36. Sardarabadi, A. M. & Vanderveen, A. J. Complex factor analysis and extensions. *IEEE Trans. Signal. Process.* **66**, 954–967 (2018).

37. Forni, M., Hallin, M. & Lippi, M. Dynamic factor models with infinite-dimensional factor space: Asymptotic analysis. *J. Econom.* **199**, 74–92 (2017).

Acknowledgements
The work described in this paper was supported by the research grant from the Natural Science Foundation of Gansu Province (22JR5RA342), we hereby thank them for the financial aids.

Author contributions
Q.N. wrote the main manuscript text, S.C. prepared Figs. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, and 21, and Z.Q. prepared all the Tables. All authors reviewed the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-61954-y.

Correspondence and requests for materials should be addressed to Q.N.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.