SCIENCE CHINA
Technological Sciences

April 2010 Vol.53 No.4: 1088–1097
doi: 10.1007/s11431-010-0060-1

Multivariate analysis in dam monitoring data with PCA

YU Hong[1,2*], WU ZhongRu[1,2], BAO TengFei[1,2] & ZHANG Lan[1,2]

[1] *State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering, Hohai University, Nanjing 210098, China;*
[2] *National Engineering Research Center of Water Resources Efficient Utilization and Engineering Safety, Hohai University, Nanjing 210098, China*

Received July 3, 2009; accepted September 2, 2009; published online March 20, 2010

Given the limitation of traditional univariate analysis method in processing the multicollinearity of dam monitoring data, this paper reconstructs the multivariate response variables by introducing principal component analysis (PCA) method, explores the ways of determining principal components (PCs), and extracts a few PCs that have major influence on data variance. For steady observation series, a control field for the whole observation values has been established based upon PCA; for unsteady observation series that have significant tendency, a control field for the future observation values has been constructed according to PC statistical predication model. These methods have already been applied to an actual project and the results showed that data interpretation method with PCA can not only realize data reduction, lower data redundancy, and reduce noise and false alarm rate, but also be effective to data analysis, having a broad application prospect.

dam safety monitoring, multivariate response variables, principal component analysis, data reduction

Citation: Yu H, Wu Z R, Bao T F, et al. Multivariate analysis in dam monitoring data with PCA. Sci China Tech Sci, 2010, 53: 1088–1097, doi: 10.1007/s11431-010-0060-1

1 Introduction

For centuries, dams have provided mankind with essential benefits such as water supply, flood control, navigation, aquaculture, hydropower, and irrigation. They are an integral part of society's infrastructure. Dam failures are rated as one of the major "low probability, high-loss" events. A large number of dams that are 50 or more years old are of great concern, since they are generally characterized by increased risk due to structural deterioration or inadequate spillway capacity [1]. If extreme loads like catastrophic flood, earthquake, etc. are encountered, hydraulic project may fail, which will directly threaten the downstream economic society and lives.

Performance monitoring of dam plays a significant role in dam safety plan, which is mainly conducted by visual inspections and analyzing data collected from instruments. Instrumentation can timely reflect critical indicators of structural behavior, and the data interpretation has already become the important composition of hydraulic engineering safety monitoring codes in China [2]. It recent years, the lack of real time data is no longer considered as a main constraint of data interpretation, yet a large time lag between initiation of analysis and completion has emerged and restricted the work of national dam safety inspection. Its main cause is that the amount of effort put into research and innovation of data analysis method is small and out of proportion compared to the effort put in instrumentation of the dam and gathering data. So in many instances, advanced automatic data acquisition systems (ADAS) and ineffective data processing capabilities seem a paradox. With the rapid development of ADAS, a variety of methods and instruments with higher sampling frequency available to monitor the dam have emerged. But the statistical method of data is still in the stage of the univariate analysis, i.e., "one point,

*Corresponding author (email: fishred@hhu.edu.cn)

one model". It is hard to improve the data interpretation efficiency due to the universal noise in physical systems and abnormal data, which cannot be identified and processed without experts' experience.

Although the advanced data acquisition method provides detailed real time data, the direct result is that the analysis work has a geometric growth. Due to inefficient analysis, remedial treatments are potentially delayed, which means the incremental cost of repair work, even an irreparable damage. The need for effective analysis tools was recently emphasized in the 20th International Commission of Large Dams General Report by Dibiagio [3].

The principal component analysis (PCA) is now widely used for lowing redundancy and realizing reduction of data to enhance the analysis efficiency. It reduces the dimensionality of high variable space with a minimum loss of information. There are many historical examples of successful use of PCA. For example, when Stone studied US national economy, a few principal components were employed to replace the original variables which can ensure that the analysis still has more than 95% accuracy [4]. In dam monitoring field, when Behrouz et al. studied the displacements, stress and seepage in Idukki arch dam, Daniel Johnson multiple arch dam and Chute-à-Caron gravity dam, they employed PCA and the hydrostatic-season-time (HST) model to estimate parameters of response variables, and at last obtained good analysis results to summarize dam behavior [5]. Li Xuehong et al. adopted PCA to improve neural networks learning capability, and then solved the multicollinearity of dam monitoring data [6]. Liu Chengdong et al. used PCA to extract the data information to determine weights for dam multi-factors [7]. Chen Long has established a fuzzy comprehensive evaluation model of roller compacted concrete layer property based on PCA [8]. This paper utilizes PCA in the dam safety monitoring data analysis, and studies confidence ellipse for analysis and forecast.

2 False alarms, data reduction and noise elimination

False alarms, data reduction and noise elimination are the three main prevailing problems in dam safety monitoring data.

2.1 False alarms

A false alarm is a signal sent out by an instrument which is out of control and detects some distorted data beyond defensive line because of certain accidental factors, when the project is still at the safe condition in fact. From the perspective of monitoring system design, a wide variety of devices and procedures are used to monitor performance of exciting dams and factors that can influence dam operation. The following features are mostly monitored by instruments: 1) displacement, 2) pore pressure and uplift pressure, 3) water level and flow, 4) seepage flow, 5) water quality, 6) temperature, 7) crack and joint size, 8) seismic activity, 9) weather and precipitation, 10) stress and 11) strain. The ultimate result of traditional univariate analysis is a $100(1-\alpha)\%$ prediction interval for the future observation and early warning.

$$\hat{y}(x_0)-t_{\frac{\alpha}{2},n-p}\left(\sqrt{S_{y,x}^2\left(1+x_0'\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}x_0\right)}\right)\leqslant y_0 \leqslant \hat{y}(x_0)+t_{\frac{\alpha}{2},n-p}\left(\sqrt{S_{y,x}^2\left(1+x_0'\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}x_0\right)}\right), \tag{1}$$

$$S_{y,x}=\sqrt{\frac{\hat{\boldsymbol{Y}}\hat{\boldsymbol{Y}}-\boldsymbol{b}'\boldsymbol{X}'\boldsymbol{Y}}{n-p-1}}, \tag{2}$$

where n is the observation sample size, p is the predictor variables (factors) size, $\boldsymbol{b}$ is the regression coefficient vector, $\boldsymbol{x}_0$ is the predictor variables vector, $\hat{y}$ is the predicted value of observed quantity, $\boldsymbol{X}$ is the predictor variables matrix ($n\times p$), $\boldsymbol{Y}$ is the matrix of a response variable observation value, $\hat{\boldsymbol{Y}}$ is the matrix of sample estimated value with regression analysis, and $S_{y,x}$ is the estimated standard error.

A probabilistic analysis for false alarms ratio (FAR) caused by measure error can be conducted with the theory mentioned above. In present data interpretation, every response variable of dam performance can be considered as a random variable. To simplify analysis, we suppose those variables are mutually independent and have no potential correlation, and this conservative hypothesis makes FAR smaller. According to the central limit theory, standardized observations follow the normal distribution. For $\alpha = 0.05$ as bilateral quantile, the probability of observed value in the prediction interval is

$$P(A)=P\left\{y_0\left|\hat{y}(x_0)-2.575\sigma\leqslant y_0\leqslant\hat{y}(x_0)+2.575\sigma\right.\right\}=0.99. \tag{3}$$

Considering the dam failure is a kind of small probability events and abnormal data are mainly due to measure error, the univariate FAR approximately is thus equal to

$$\begin{aligned}P(B)&=P\left\{y_0\left|y_0<y_0-2.575\sigma\right.\right\}\\&\quad+P\left\{y_0\left|y_0>y_0+2.575\sigma\right.\right\}\\&=1-P(A)=1-0.99=0.01.\end{aligned} \tag{4}$$

To hydraulic projects, if the correlations among response variables are not taken into account, then the probability of no false alarm in an observation of l response variables can be estimated as

$$P(C)=(1-P(B))^l=0.99^l. \tag{5}$$

Figure 1 shows the relationship curve between FAR and

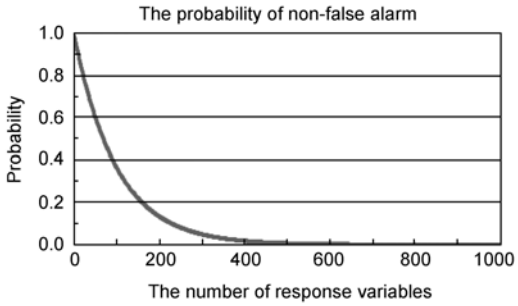

Figure 1 The probability of non-false alarm in a single observation.

the scale of monitoring response variables. Then it is recognized that FAR is already beyond 90% when the scale is up to certain extent ($l > 200$); FAR will approximately equal 1 when the scale reaches to 400. While in practice, in a medium-sized hydraulic project for instance, hundreds of instruments should be arranged to attain normal monitoring standard. As for large project like the Three Gorges Dam, more than 8000 devices have already been placed in the monitoring system design phase. Furthermore, some of those instruments can monitor several response variables, thus it makes FAR higher.

In fact, frequent false alarms, like a bottleneck, restrict the extensive application of traditional analysis method. However, this situation is not inevitable. During the dam safety monitoring practice, it was found that there is certain correlativity among many response variables, for example, the measurement value of inverse plummet at different elevations in the same dam monolith, the observations of tension wire at distinct dam monoliths and so on. Based on the correlativity, PCA can be used to implement bivariate and multivariate analyses and confidence ellipses or confidence ellipsoids can be built with PCs to test the future observation value. Thus, on the one hand PCs take the place of response variables that reduce the scale of variables, on the other hand that PCs reduce the perturbation caused by measure error to cut down FAR.

2.2 Data reduction

Because of the correlativity mentioned above, different response variables usually reflect the same feature of the dam. From this point, certain data redundancy widely exists in observations. Before ADAS technology matured, the intervals of artificial data samples were hard to keep the same steps. Consequently, it was usually necessary to use interpolation and extrapolation to create required regular data set for multivariate data processing, which is a heavy work and also hard to guarantee accuracy. Therefore, in that time multivariate analysis method was unavailable to extensive application. In addition, traditional univariate analysis method not considering the correlativity among targets and reserving high data redundancy, that limit the analysis efficiency directly. With the development of automation technology, two influences have emerged. The negative one is that the scale of arranged instruments increases and huge collected data cause the large time lag between acquisition and conclusion. The positive one is that automation technology can ensure several instruments to simultaneously work and ensure data sequences to automatically meet analysis requirement.

PCA reconstructs response variables and extracts PCs from them. In this way, information is rapidly concentrated from grand observations, and simultaneously data redundancy is eliminated. At the same time, data reduction is realized with PCA, to effectively shorten the analysis delay.

2.3 Noise elimination

To ensure the accuracy and reliability of data is the primary problem in dam safety monitoring data analysis. However, even adopting the most advanced method, measurement error is still unavoidable. Two main types of measurement errors are generally recognized: (1) systematic error and (2) random error. The former is that the measurement value is either more or less than the correct value by a fixed percentage, and this can be solved by reconfiguring the monitoring system. However, the latter is perturbation of measurement that can be on either side of the true value. Sources of random errors include uncontrolled influential factors, such as air currents, ambient temperature fluctuations, relative humidity, power source disturbances, electromagnetic interference and so on. Generally, random error is often called noise, by analogy to acoustic noise. By use of PCA, long term observed data series are analyzed and PCs or prediction factors are extracted from the original variables.

3 Principal component analysis

In PCA, under the premise of little or no information loss, the original correlated variables are transferred into the new and uncorrelated variables called the principal components. Each principal component is a linear combination of the original variables.

3.1 Basic concepts

We suppose that m response variables, each having n observations, denoted as $x_1, x_2, ..., x_m$, (for simplicity, the average of x_i is 0 and variance is 1, $1 \leqslant i \leqslant m$), construct an $n \times m$ matrix $\boldsymbol{X}$, x_{ij} is the jth observation value of x_i:

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}. \tag{6}$$

The purpose of PCA is to obtain a few uncorrelated com-

prehensive variables, which are linear combinations of m original variables (x_1, x_2, ..., x_m). Through the new variables, most original information can be well represented and explained. We may define z_1, z_2, ..., z_n as the new comprehensive variables, each of which is a linear combination of m original variables:

$$\begin{cases} z_1 = l_{11}x_1 + l_{12}x_2 + \cdots + l_{1m}x_m = \sum_{i=1}^{m} l_{1i}x_i, \\ z_2 = l_{21}x_1 + l_{22}x_2 + \cdots + l_{2m}x_m = \sum_{i=1}^{m} l_{2i}x_i, \\ \vdots \\ z_p = l_{n1}x_1 + l_{n2}x_2 + \cdots + l_{nm}x_m = \sum_{i=1}^{m} l_{ni}x_i. \end{cases} \quad (7)$$

The following conditions must be simultaneously satisfied:

(1) z_i is uncorrelated with z_j ($i \neq j$; i, j=1,2, ..., p);

(2) z_1 has the biggest variance in all linear combinations of x_1, x_2, ..., x_m; z_2 has the biggest variance in all linear combinations of x_1, x_2, ..., x_m, which are uncorrelated with z_1; … z_p has the biggest variance in all linear combinations of x_1, x_2, ..., x_m, which are uncorrelated with z_1, z_2, ..., z_{p-1}.

Then the new variables z_1, z_2, ..., z_p can be called 1st, 2nd, 3rd, …, pth principal components of original variables x_1, x_2, …, x_m respectively.

From the analysis mentioned above, it is known that the essential of PCA is to identify coefficients l_{ij} (i=1, 2, …, p, j=1, 2, …, m), which are the weights of the original variables x_j (j=1, 2, …, m) on the PCs z_i (i=1, 2, …, p). In mathematics, those coefficients are the eigenvectors of the correlation matrix of m original variables (x_1, x_2, …, x_m), corresponding to the biggest p eigenvalues respectively. Meanwhile the variance $var(z_i)$ of each comprehensive variable z_i is exactly the eigenvalue λ_i. For this reason, the contributions of PC's variances are arranged in order of decreasing eigenvalues, i.e.,

$$\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p \geqslant 0. \quad (8)$$

3.2 Procedures

3.2.1 Calculation of the correlation matrix's eigenvalues

PCA is a statistical technique applied to either the correlation matrix ($\boldsymbol{R}$) or covariance matrix ($\boldsymbol{S}$). The first step to obtain principal components is to find the solution to the eigenvalue problem:

$$(\boldsymbol{R} - \lambda \boldsymbol{I})\, p = 0, \quad (9)$$

where $\boldsymbol{I}$ is the identify matrix of order m. A number of different numerical algorithms can be used to compute the eigenvectors and eigenvalues [9]. Solving eq. (9) results in a set of eigenvalues $\lambda_j (j = 1, 2, \cdots, m)$, which can be placed as the elements of a diagonal matrix $\boldsymbol{\Lambda}$, and a corresponding set of vectors $\boldsymbol{p}_j (j = 1, 2, \cdots, m)$. Principal components are linear combinations of the original variables, where the weight on each variable is given by the eigenvectors. The percentage of original data variance explained by the first k principal components can be expressed by eq. (10) or eq. (11) as

$$\frac{\lambda_k}{\sum_{j=1}^{m} \lambda_j} \times 100\%, \quad (10)$$

or

$$\frac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{m} \lambda_j} \times 100\%. \quad (11)$$

Eq. (10) represents the percentage of original data variance explained by the k principal component, called the contribution rate. Eq. (11) represents the total percentage of original data variance explained by the first k principal components, called accumulated contribution rate.

3.2.2 Determination of PCs' number

There are four methods used to determine the number of PCs.

(1) Accumulated contribution rate: All eigenvalues are ordered by values from big to small, and the 1st principal component is structured according to the principal eigenvector. We increase the number of PCs one by one, and calculate the accumulated contribution rate with eq. (11) until it is a relatively high percentage, say 70%–90%.

(2) Average eigenvalue: Those components whose eigenvalues are greater than the average eigenvalue should be retained. Previous studies have shown that this method is fairly accurate when the number of original variables is <30 and the variables are highly correlated [10].

(3) Scree graph: Eigenvalues are plotted as a function of the number of eigenvalues. The number of components is selected where the scree graph flattens out.

(4) Signal and noise: The relationship between the principal components and secondary components (SCs) can be explained by analogy with signal and noise. The difference between adjacent components can be presented by the following equation:

$$\gamma_k = \frac{\lambda_k}{\lambda_{k+1}}, (k = 1, 2, \cdots, p-1). \quad (12)$$

When all PCs are identified, γ_k presents the ratio of the minimum PC's eigenvalue λ_k and the maximum noise's eigenvalue λ_{k+1}. Due to different features of signal and noise, the ratio γ_k, which is obviously greater than the ratios within these two sets respectively, is defined as the boundary line between PCs and SCs.

3.2.3 Construction of PCs

The original data can be reconstructed by using the first k principal components

$$\boldsymbol{Y} = \boldsymbol{Z} + \boldsymbol{E} = \boldsymbol{X}\boldsymbol{p}' + \boldsymbol{E}, \tag{13}$$

where $\boldsymbol{E}$ is the residual matrix, $\boldsymbol{Z}$ is the matrix of principal components, and $\boldsymbol{p}$ is the matrix of k principal eigenvectors $\boldsymbol{p} = \{\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_k\}$. The residual matrix contains that part of the data not explained by the PCA and most likely represents the noise in the data. The method is useful for separating signal form noise since random noise components are usually uncorrelated and are associated with low principal components.

3.3 Safety monitoring method based on principal components

If the dam is at a stable condition, then the hydrograph of principal components will not exhibit tendency variation with the time. As for this stable observation series, future observations can be used to conduct safety monitoring by establishing the control domain of whole observations. But if the dam is influenced by concrete creep, dry shrinkage, alkali-aggregate reaction and other factors, the hydrograph will show changes in trends. Therefore, it is necessary to construct control domain for future single observation to implement structural safety monitoring, which takes the model predicted value as the center. Here are two specific implementation methods.

3.3.1 Control domain of whole observation

When the dam behavior keeps stable, we can use collected data x_1, x_2, ..., x_n to extract principal components z_1, z_2, ..., z_p, and determine one or some future observations' control domains. In this condition, observations can be seen as independently identically distributed, following $N_p(u,\sigma)$.

We suppose that z_1, z_2, ..., z_p as independently identically distributed, following $N_p(u,\sigma)$. $\boldsymbol{Z}$ can be denoted as a future observation from the same distribution, and then statistics can be expressed by

$$T^2 = \frac{n}{n+1}(\boldsymbol{Z}-\overline{\boldsymbol{Z}})'\boldsymbol{S}^{-1}(\boldsymbol{Z}-\overline{\boldsymbol{Z}}), \tag{14}$$

where

$$\boldsymbol{S} = \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1p} \\ S_{21} & S_{22} & \cdots & S_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{p1} & S_{p2} & \cdots & S_{pp} \end{bmatrix} = \left\{ S_{ik} = \frac{1}{n-1}\sum_{j=1}^{n}\left(z_{ji}-\overline{z}_i\right)\left(z_{jk}-\overline{z}_k\right) \right\} \tag{15}$$

obeys the distribution of $\frac{(n-1)p}{n-p}F_{p,n-p}$, and its $100(1-\alpha)\%$ p-dimensional predicted ellipsoid is determined by all z's satisfying the following condition:

$$\frac{n}{n+1}(\boldsymbol{Z}-\overline{\boldsymbol{Z}})'\boldsymbol{S}^{-1}(\boldsymbol{Z}-\overline{\boldsymbol{Z}}) \leqslant \frac{(n-1)p}{n-p}F_\alpha(p,n-p). \tag{16}$$

It is a p-dimensional ellipsoid centered on sample average $\overline{\boldsymbol{Z}}$, whose relevant parameters can be determined as follows.

Because $\boldsymbol{S}$ is symmetric and positive definite, we can know from linear algebra that $\boldsymbol{S}$ has real eigenvalues, all of which are greater than zero. Supposing p eigenvalues are

$$\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p \geqslant 0. \tag{17}$$

And that their corresponding unit eigenvectors are $\boldsymbol{u}_1$, $\boldsymbol{u}_2$, ..., $\boldsymbol{u}_p$, which are p-dimensional column vectors and meet the following feature:

$$\boldsymbol{u}_i'\boldsymbol{u}_j = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases} \tag{18}$$

By denoting $\boldsymbol{U}$ as $(\boldsymbol{u}_1, \boldsymbol{u}_2, ..., \boldsymbol{u}_p)^{\mathrm{T}}$, $\boldsymbol{U}$ is an orthogonal matrix, that is,

$$\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} = \boldsymbol{I}. \tag{19}$$

Denote $\boldsymbol{\Lambda}$ as $\begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_p \end{bmatrix}$; then $\boldsymbol{S}^{-1} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{\mathrm{T}} = \sum_{i=1}^{p}\frac{\boldsymbol{u}_i\boldsymbol{u}_i^{\mathrm{T}}}{\lambda_i}$, and eq. (15) can be rewritten as

$$\sum_{i=1}^{p}\frac{\frac{n(n-p)}{p(n^2-1)}(\boldsymbol{Z}-\overline{\boldsymbol{Z}})'\boldsymbol{u}_i\boldsymbol{u}_i^{\mathrm{T}}(\boldsymbol{Z}-\overline{\boldsymbol{Z}})}{\lambda_i} \leqslant F_\alpha\left(p,n-p\right). \tag{20}$$

In this way, $100(1-\alpha)\%$ confidence interval of observation takes sample average $\overline{Z}$ as the center, and its half-axial lengths are $\sqrt{\frac{\lambda_1 p(n^2-1)}{n(n-p)}F_\alpha(p,n-p)}$, $\sqrt{\frac{\lambda_2 p(n^2-1)}{n(n-p)}F_\alpha(p,n-p)}$, … , $\sqrt{\frac{\lambda_p p(n^2-1)}{n(n-p)}F_\alpha(p,n-p)}$ respectively, each axial direction being eigenvector $\boldsymbol{u}_i$ corresponding to λ_i, i=1,2, ..., p.

Especially, when p equals 2, predicted ellipsoid possessing 95% probability is

$$(Z-\overline{Z})'S^{-1}(Z-\overline{Z}) \leqslant \frac{2(n^2-1)}{n(n-2)}F_{0.05}(2,n-2). \tag{21}$$

If the future observation $\boldsymbol{Z}$ falls out of the control ellipse, then it's denoted as losing control.

3.3.2 *Control domain of future single observation*

Due to concrete creep, dry shrinkage, alkali-aggregate reaction and other factors, the features of concrete change with the time. Principal components can be predicted by the hydrostatic-season-time (HST) model [11], random intervals of the predicted value can be determined, and the probability of containing observation is $1-\alpha$.

The main expressions of model (HST) are as follows:

$$D(t) = H(h) + S(\theta) + T(t), \tag{22}$$

$$H(h) = a_1 + a_2 h + a_3 h^2 + a_4 h^3 + a_5 h^4, \tag{23}$$

$$S(\theta) = a_6 \sin\theta + a_7 \cos\theta + a_8 \sin\theta\cos\theta + a_9 \sin^2\theta, \tag{24}$$

$$T(t) = c_1 t' + c_2 t'^2 + c_3 t'^3, \tag{25}$$

where $D(t)$ is the response variable, $H(h)$, $S(\theta)$, $T(t)$ are respectively reservoir level, ambient temperature and time effects components, where

$$h = \frac{H - H_{\min}}{H_{\max} - H_{\min}},\ t' = t - t_0 \ \text{ and } \ \theta = \frac{2\pi t'}{365},$$

where $H_{\min}$ and $H_{\max}$ are respectively the minimum and maximum reservoir water levels, t' is the initial date of data record in the statistic analysis collection.

Confidence interval of the predicted value, whose probability is $100(1-\alpha)\%$, can be expressed as

$$\frac{n(n-p)}{p(n-1)}(Z-\hat{Z})' S^{-1}(Z-\hat{Z}) \leqslant F_\alpha(p, n-p). \tag{26}$$

It is a p-dimensional ellipsoid centered on the predicted value $\hat{Z}$, whose relevant parameters can be determined by eqs. (17)–(19), and then eq. (26) can be rewritten as

$$\sum_{i=1}^{p} \frac{\dfrac{n(n-p)}{p(n-1)}(Z-\hat{Z})' \boldsymbol{u}_i \boldsymbol{u}_i^{\mathrm{T}} (Z-\hat{Z})}{\lambda_i} \leqslant F_\alpha(p, n-p). \tag{27}$$

In this way, $100(1-\alpha)\%$ confidence interval of observation takes the predicted value as the center, and its half-axial lengths denoted as semi-axes are respectively $\sqrt{\dfrac{\lambda_1 p(n-1)}{n(n-p)} F_\alpha(p, n-p)}$, $\sqrt{\dfrac{\lambda_2 p(n-1)}{n(n-p)} F_\alpha(p, n-p)}$, … , $\sqrt{\dfrac{\lambda_p p(n-1)}{n(n-p)} F_\alpha(p, n-p)}$, each axial direction being eigenvector $\boldsymbol{u}_i$ corresponding to λ_i, i=1,2, ..., p.

Especially, when p equals 2, the predicted ellipsoid possessing 95% probability is

$$(\boldsymbol{Z} - \hat{\boldsymbol{Z}})' \boldsymbol{S}^{-1} (\boldsymbol{Z} - \hat{\boldsymbol{Z}}) \leqslant \frac{2(n-1)}{n(n-2)} F_{2,n-2}(0.05). \tag{28}$$

If the error of predicted value itself is not taken into account, the above conclusion is accurate. But owing to the inescapability of predicted errors, it is necessary to synthetically consider the influence of half-axial length. In this paper, the author suggests that the half-axial length should be adjusted by using the following formula:

$$sa = sa + \mu S, \tag{29}$$

where μ is the reduction factor and has some connection with relevant parameters of the model, is generally 0.4–0.5. S is the regression standard deviation of the model. The reduction process should keep the eccentricity of control ellipse constant. If the future observation $\boldsymbol{Z}$ falls out of the control ellipse, then it is called as out of control.

4 Case study

In order to confirm the effectiveness of multivariable analysis, we take Chencun Hydropower Station dam as the example and carry on principal component analysis with the above method.

4.1 Introduction

Chencun Hydropower Station is situated in the upstream of Qingyi River, a tributary of Yangtze River in south of Anhui. It is a comprehensive hydro-junction, which serves mainly for power generation, and flood control, irrigation, aquaculture and navigation as well. Its main water-retaining structure is the Chencun gravity arch dam. This project was finished through three phases of construction successively. When pouring the 2nd stage concrete, layers ascended so quickly that time intervals between pouring layers were short, and the shrinkage deformation of the 2nd stage concrete was subjected to strong constraint of the 1st stage concrete. This caused the crack to appear at the top of the 1st stage concrete (near 105 height). The crack was 380 m long by more than 5 m deep, which weakened dam stiffness and had influence on dam integrity. At first, artificial joint meter was used to monitor cracks. In order to timely observe the crack opening, from August 17, 1998, automatic joint meter were installed in 5#, 8#, 18#, 26# and 28# monoliths. In this paper, the principal component analysis method was applied to process automatic collected data from August 17, 1998–July 10, 2007.

4.2 Construction of principal components

Take observations of 5 joint meters as the original values, denoted as x_1, x_2, …, x_5. After data inspection, it was found in the total 1217 groups of data, 17 groups (33 data records)

had no instrument measure records, accounting for 1.4% of the total samples. Finally, 1200 observations were obtained by excluding those abnormal data. Then we drew histograms and scatter diagrams of the original variables (shown as Figure 2), and calculated correlation coefficient matrix $\boldsymbol{R}$ (shown as Table 1).

The eigenvalues of $\boldsymbol{R}$ and their contribution rate, accumulated contribution rate and signal to noise ratio were calculated by eqs. (9)–(11), then scree graphs of eigenvalues were drawn as shown in Figure 3 and Table 2.

From Table 2, we can see only the 1st and 2nd eigenvalues are greater than 1. Combined with Figure 3, we also find 67.84% of data change volume can be interpreted by the first PC, 24.06% by the second PC, the accumulated contribution rate of these two sides has already achieved 91.9%, which can fully explain data change. In the light of signal to noise ratio (λ_k/λ_{k+1}), the ratio between 2nd and 3rd PC is the biggest. Therefore, the 3rd PC and subsequent components are defined as noise. Loadings of PCs are shown in Table 3.

After the analysis of load, it is seen that the 1st PC mainly explains the change of $x18$, $x26$, $x28$, and the 2nd PC mainly explains the changes of $x5$, $x8$. The original 5 respond variables are down to 2 through extracting PCs, which really realizes data reduction. Meanwhile, 8% data variation is seen as noise, which efficiently lessens the workload of blunder adjustment. It is illustrated that the process of extracting PC is also a process of filtering noise.

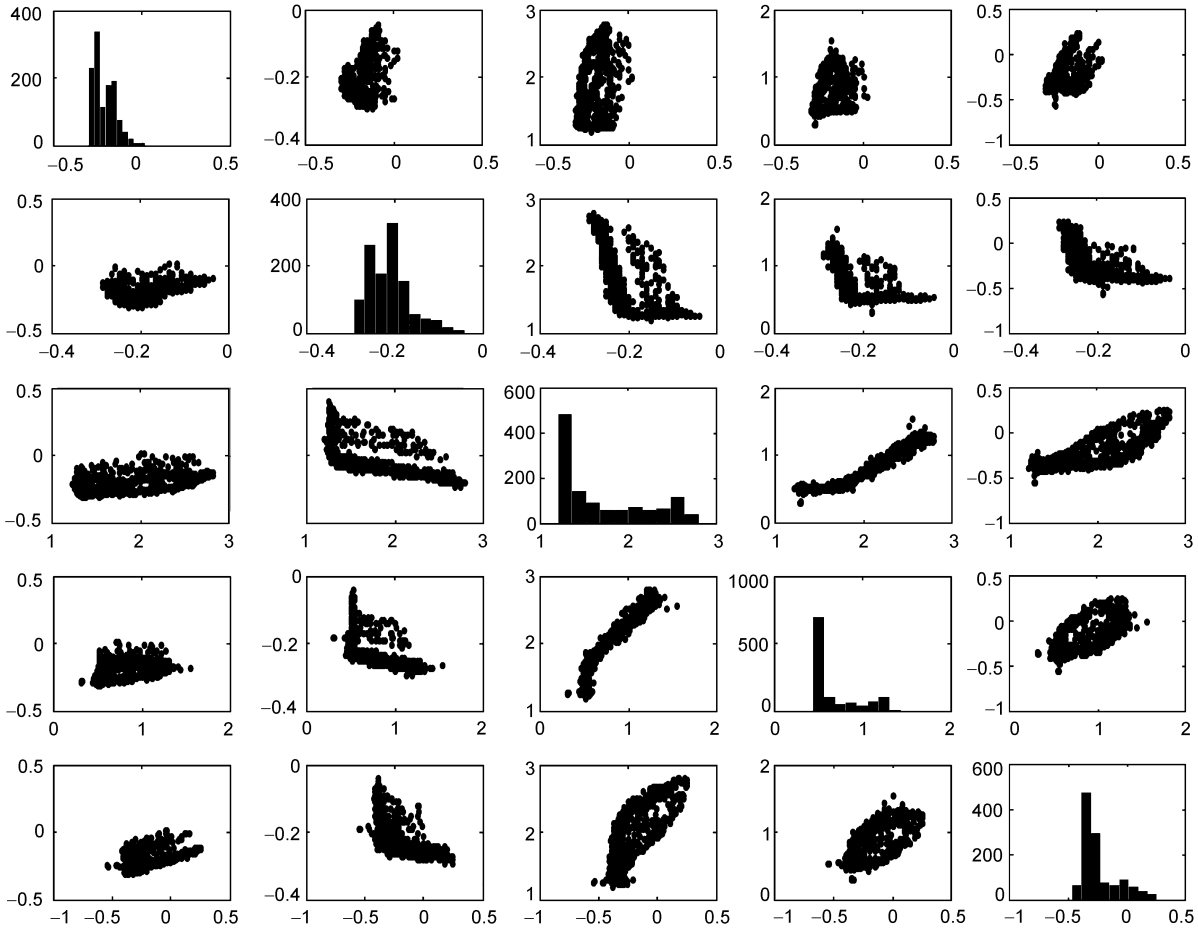

Figure 2 Scatter diagrams and histograms.

Table 1 Correlation coefficients

	x_1	x_2	x_3	x_4	x_5
x_1	1.000	0.215	0.424	0.412	0.434
x_2	0.215	1.000	–0.687	–0.604	–0.620
x_3	0.424	–0.687	1.000	0.948	0.827
x_4	0.412	–0.604	0.948	1.000	0.791
x_5	0.434	–0.620	0.827	0.791	1.000

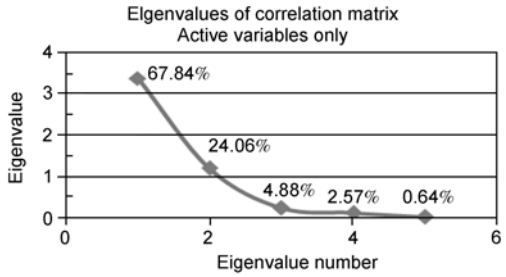

Figure 3 Scatter.

Table 2 Eigenvalues and related information of $\boldsymbol{R}$

	λ_i	Contribution rate	Accumulated contribution rate	λ_k/λ_{k+1}
1	3.392	0.678	0.678	2.819
2	1.203	0.241	0.919	4.929
3	0.244	0.049	0.968	1.900
4	0.129	0.026	0.994	4.016
5	0.032	0.006	1.000	0.006

Table 3 Loadings of PCs

Variable	Loadings	
	PC1	PC2
*x*5	–0.238	–0.801
*x*8	0.390	–0.596
*x*18	–0.530	–0.003
*x*26	–0.514	–0.036
*x*28	–0.497	–0.044

4.3 Method and index of PC monitoring

4.3.1 Construction of the control domain of whole observation

The equation of control domain can be obtained by eqs. (13)–(20) as

$$\left(Z-\begin{bmatrix}1.181\\0.271\end{bmatrix}\right)'\begin{bmatrix}0.224983 & -0.00898\\-0.00898 & 0.004166\end{bmatrix}^{-1}\left(Z-\begin{bmatrix}1.181\\0.271\end{bmatrix}\right)\leqslant 6.010. \quad (30)$$

Semi-major axis is 1.1644, its direction is $\boldsymbol{u}_1$= [–0.99918 0.040546]$^{\mathrm{T}}$; the semi-minor axis is 0.1512, its direction is $\boldsymbol{u}_2$=[–0.99918–0.040546]$^{\mathrm{T}}$. The overall scatter diagram and control domain diagram are shown in Figure 4.

4.3.2 Construction of the control domain of future observation

From the scatter diagram of whole observations and hydrographs of PC1, PC2 and environmental variables (shown in Figures 5, 6 and 7), we find there is a growing tendency of the annual peak of principal component. That is to say, the crack behavior is still unstable, which needs to establish the control domain of future observation for better monitoring. What's more, according to eqs. (22)–(25), HST prediction models of PC1 and PC2 are made, and model parameters are given in Table 4. The HST model correlation coefficient *R* of PC1 is 0.973, and its standard deviation *S* is 0.46. Correlation coefficient *R* of PC2 is 0.954, and standard deviation *S* is 0.082.

The changes of PC1 and PC2 are not synchronous as well, which means they won't simultaneously achieve the peak. Unfavorable load combination does not merely appear at the moment when principal component achieves the peak, so it is important to determine the united control domain of PC1 and PC2, which has great influences on structure safety.

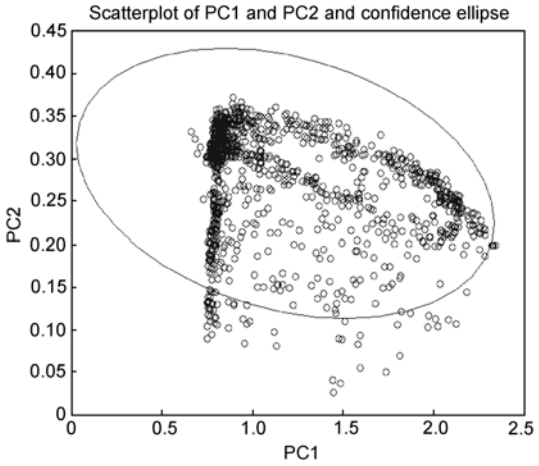

Figure 4 The control domain of whole observations.

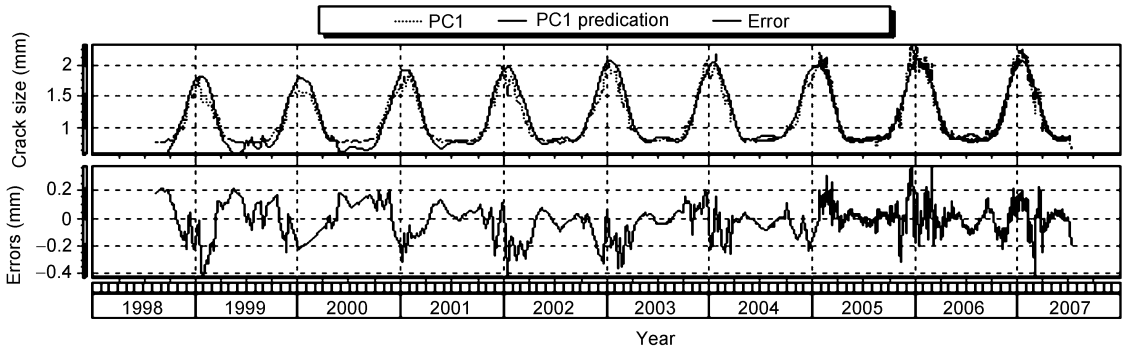

Figure 5 Fitted hydrograph of PC1 measured values.

Table 4 HST model parameters of PC1 and PC2

	Const	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	c_1	c_3
PC1	–3.142	–2.599	9.930	–14.223	6.962	0.241	0.576	0.373	–0.320	0.032	-2.206×10^{-7}
PC2	–1.462	0	0.221	–0.750	0.524	0.012	–0.040	–0.024	0.049	0.011	-5.601×10^{-8}

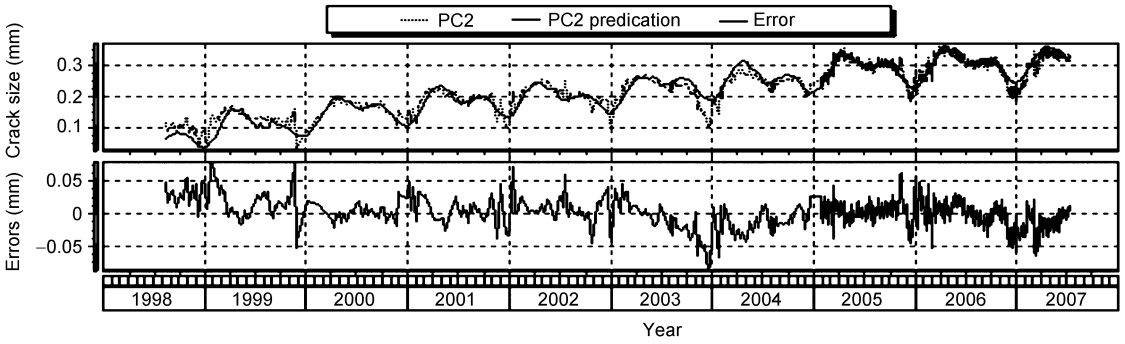

Figure 6 Fitted hydrograph of PC2 measured values.

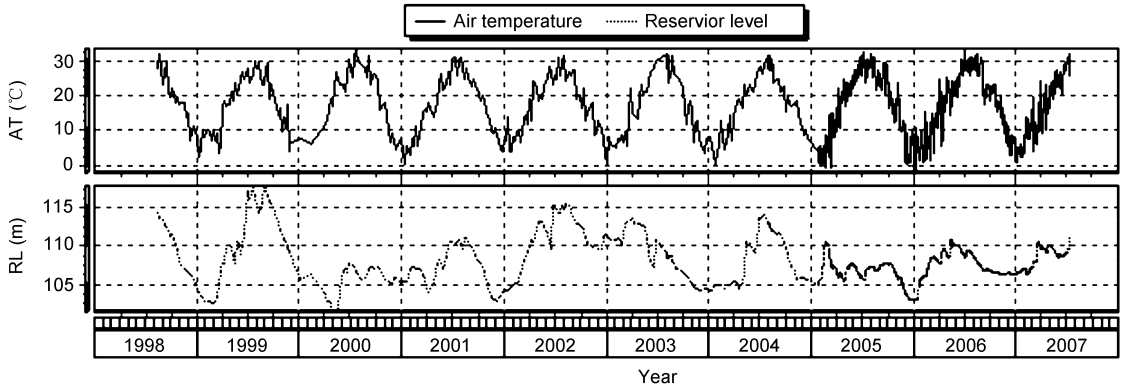

Figure 7 Hydrograph of ambient temperature and reservoir level.

The equation of control domain can be obtained according to eqs. (26)–(28) as

$$\left(\boldsymbol{Z}-\begin{bmatrix}\hat{Z}_1\\ \hat{Z}_2\end{bmatrix}\right)'\begin{bmatrix}0.224983 & -0.00898\\ -0.00898 & 0.004166\end{bmatrix}^{-1}\left(\boldsymbol{Z}-\begin{bmatrix}\hat{Z}_1\\ \hat{Z}_2\end{bmatrix}\right)\leqslant 0.005. \tag{31}$$

After taking the reduction factor of semi-major axis as 0.55, the axis reduces to 0.287 and its direction is $\boldsymbol{u}_1$= [–0.99918 0.040546]$^{\mathrm{T}}$. By taking the reduction factor of the semi- minor axis as 0.4, the axis reduces to 0.0372 and its direction is $\boldsymbol{u}_2$=[–0.99918 –0.040546]$^{\mathrm{T}}$.

Once selecting the predicted value around the PC peak in 2007, the control domain of observations could be drawn as shown in Figure 8. From the figure, we can see the control domain completely contains the observations, which illustrates the crack behavior doesn't display variation and the dam is in the normal operation under current load combination.

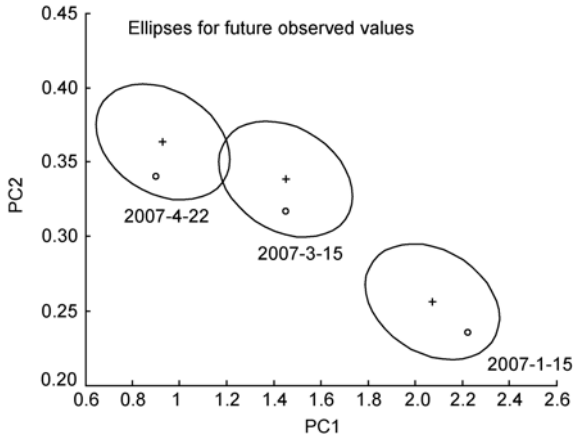

Figure 8 Confidence ellipse of future observation.

5 Conclusion

In this paper, data redundancy and measurement noise

phenomenon, which widely exists in the massive data collected from dam safety monitoring system, is fully analyzed. It is also pointed out that the traditional univariate analysis method ignores the correlativity among responsive variables, which not only causes analysis lag, but also makes frequent false alarms. That is why analysis efficiency and application are extremely restricted. Therefore, the theory and method of principal component analysis (PCA) are introduced to data analysis and monitoring, and through a specific case it is explained that the above problems can be effectively solved. The mainly conclusions of the paper are as follows.

(1) Data reduction. By use of the correlativity among respond variables, principal components are extracted, and data redundancy can be effectively reduced. The number of reconstructed PCs are generally not greater than 4–5, which means a data reduction about 60% of original size.

(2) Noise filtering. Principal component can separate more than 90% signal from noise. It may process the majority abnormal data successfully, which efficiently lessens the work of gross error adjustment.

(3) Multivariate analysis and monitoring. Multivariate values can be processed with PCA and monitored by multidimensional confidence ellipsoid, which changes the traditional mode of univariate analysis, improves data processing and structure safety monitoring. Meanwhile, false alarms are controlled at a low level, and the original monitoring system updated by PCA can be used in a large scale.

This work was supported by the National Natural Science Foundation of China (Grant Nos. 50909041, 50879024, 50809025, 50539010, 50539110), the National Supporting Program (Grant Nos. 2008BAB29B03, 2008BAB29B06), and the Natural Science Foundation of Hohai University (Grant No. 2008426811).

1 National Research Council. Safety of Existing Dams Evaluation and Improvement. Washington D.C.: National Academy Press, 1983. 4–40
2 SDJ336-89. Concrete Dam Safety Monitoring Specifications (in Chinese). Department of Energy and Water of the People's Republic of China, 1990. 1255–1256
3 Dibiagio E. Monitoring of dams and their foundations. 20th International Commission on Large Dams. Beijing: ICOLD, 2000
4 Ren R E, Wang H W. Multivariate Statistical Data Analysis-Theory, Methods, Examples (in Chinese). Beijing: National Defense Industry Press, 1997. 92
5 Behrouz A N. Multivariate Statistical Analysis of Monitoring Data for Concrete Dams. Dissertation of Doctoral Degree. Canada Montreal: McGill University, 2002. 97–102
6 Li X H, Xu H Z, Gu C S, et al. Application of principal component fuzzy neural network model to analysis of observation data of dams (in Chinese). Dam Obs Geotech Test, 2001. 14–16
7 Liu C D. Analysis Method of Multi-Factor Weightings of Dam Safety Evaluation and its Application (in Chinese). Nanjing: Hohai University, 2004. 20–22
8 Chen L. Mechanics performance with parameters changing in space & monitoring models of roller compacted concrete dam (in Chinese). Dissertation of Masteral Degree. Nanjing: Hohai University, 2006. 74–76
9 Rencher A C. Multivariate Statistical Inference and Applications. New York: John Wiley & Sons, 1998. 119–126
10 Martens H, Nase T. Multivariate Calibration. New York: John Wiley & Sons, 1989. 141–142
11 Bonelli S, Royet P. Delayed response analysis of dam monitoring data. Dams in a European Context, Swets and Zeitlinger, Lisse, 2001. 91–99