Solving Multicollinearity in Dam Regression Model Using TSVD

XU Chang1, DENG Chengfa2

1. Department of Municipal Engineering, Zhejiang Water Conservancy and Hydropower College, 583 Xuelin Street, Xiasha District, Hangzhou 310018, China
2. Zhejiang Institute of Hydraulic & Estuary, Hangzhou 310020, China

© Wuhan University and Springer-Verlag Berlin Heidelberg 2011

Abstract Targeting the multicollinearity problem in dam statistical model and error perturbations resulting from the monitoring process, we built a regularized regression model using Truncated Singular Value Decomposition (TSVD). An earth-rock dam in China is presented and discussed as an example. The analysis consists of three steps: multicollinearity detection, regularization parameter selection, and crack opening modeling and forecasting. Generalized Cross-Validation (GCV) function and L-curve criterion are both adopted in the regularization parameter selection. Partial Least-Squares Regression (PLSR) and stepwise regression are also included for comparison. The result indicates the TSVD can promisingly solve the multicollinearity problem of dam regression models. However, no general rules are available to make a decision when TSVD is superior to stepwise regression and PLSR due to the regularization parameter-choice problem. Both fitting accuracy and coefficients’ reasonability should be considered when evaluating the model reliability.

Keywords multicollinearity; TSVD; regularization parameter; PLSR; stepwise; dam

CLC number P207

Introduction

Statistical approaches (e.g., stepwise regression) with the advantages of convenience and no requirements for knowledge of mechanical properties are often used to describe dam deformation patterns by modeling field observables of points located on the surface and within dam structures.1,2 Unfortunately, regressors in dam regression analysis are often ill-posed due to the multicollinearity existing among independent variables (e.g., reservoir water level, temperature, and aging).3 Severe multicollinearity will lead to large variances for least-squares estimators of regression coefficients.4 In this case, other techniques (e.g., principal component regression (PCR),5 Partial Least-Squares Regression (PLSR),2 and evolutionary algorithms6,7) are proposed to circumvent the multicollinearity problem. However, PLSR or PCR with Cross-Validation (CV) criterion will ignore important information of low-variance Principal Components (PCs) in some cases.8 Evolutionary approaches require proper parameter settings, and the noise on the limited data will produce the over-fitted problem. On the other hand, dam models using evolutionary algorithms cannot separate the displacements contributed by water pressure, temperature gradient, and aging, respectively. Re-
searchers have developed numerous regularization approaches\cite{9-11} to stabilize ill-posed problems and to single out meaningful approximate solutions. In particular, Singular Value Decomposition (SVD) and Tikhonov regularization (or ridge regression) have been introduced in the field of hydraulic engineering. Wu and Cheng\cite{12} proposed a robust SVD-based statistical model to deal with noisy reservoir water level observables. Jing et al.\cite{13} discussed the practicality of ridge regression in modeling dam crack opening. Unfortunately, the regularization parameter in that paper is obtained from a ridge trace, which has a great subjective effect on the regression coefficients. As such, further studies are still necessary to put forward to discuss the reliability of regularization methods in dam deformation analysis.

In this study, we shall focus on the Truncated Singular Value Decomposition (TSVD) regularization in conjunction with two parameter-choice techniques, Generalized Cross-Validation (GCV) function and L-curve criterion, to cope with the multicollinearity problem in dam statistical models.

1 Material and methods

1.1 Data set and dam regression model

In this study, we picked up field observables (reservoir water level and crack opening) spanning 2 years of an earth-rock dam in China to establish dam statistical model. We selected first 99 pairs of the data set to build the model and last 5 pairs to examine the model predictive ability. According to the research result in the present and considering the dam characteristics (e.g., geometry, construction materials, topography, and geological condition of the foundation), we selected the following variables as the most influential factors to describe dam crack patterns:

- Hydrostatic (water) pressure factors: $H^1, H^2, H^3$,
- Thermal factors in forms of periodic functions: $T_1 = \sin(2\pi t/365 - 2\pi t_0/365)$, $T_2 = \cos(2\pi t/365 - 2\pi t_0/365)$, $T_3 = \sin(4\pi t/365 - 4\pi t_0/365)$, $T_4 = \cos(4\pi t/365 - 4\pi t_0/365)$,
- Aging factors: $\theta - \theta_0, \ln t - \ln t_0$,

where $H$ is the reservoir water level in meter above foundation, $t$ is the accumulative total number of days from starting day to observation day, and $t_0$ is the number of days from starting day to the starting day of calculating period of time, $\theta = t/100$, and $\theta_0 = t_0/100$.

Then, the observed crack opening $y$ of a point located on the dam can be modeled as:

$$y = a_0 + \sum_{i=1}^{4} a_i H^i + \sum_{j=1}^{2} b_j T_j + c_4 (\theta - \theta_0) + c_5 (\ln \theta - \ln \theta_0)$$

(1)

1.2 Truncated singular value decomposition

Set $A = [1 \ H^1 \ H^2 \ H^3 \ T_1 \ T_2 \ T_3 \ T_4 \ \theta - \theta_0, \ln \theta - \ln \theta_0]$, and $x = [a_1, a_2, a_3, b_1, b_2, b_3, c_1, c_2]^T$, then Eq. (1) can be given by the same equation:

$$Ax = y$$

(2)

Assuming $A$ has no exact zero singular values, we can obtain the least-squares solution:

$$\hat{x} = (A^TA)^{-1}A^Ty$$

(3)

Within our knowledge, the solution will not exist if $A^TA$ is close to singular, which may be caused by the multicollinearity among predictors. In this case, a small change in $y$ resulting from the field monitoring process can produce a large change in $\hat{x}$. Numerous regularization techniques are available in \cite{14} to deal with this ill-posed problem with perturbations in $y$. In this study, we shall concentrate on the TSVD regularization, a method amounts to truncating the SVD of the coefficient matrix $A$ and then solving the modified least-squares problem.\cite{15} A TSVD solution can be written as

$$x^{(k)} = \sum_{i=1}^{k} \frac{u_i^T y}{\sigma_i} v_i, 1 \leq k \leq n$$

(4)

where $u_i$ is the eigenvector of $AA^T$, $v_i$ is the eigenvector of $A^TA$, $\sigma_i$ is the singular value, and $k$ is the truncation parameter.

As shown in Eq. (4), the proper choice of $k$ is critical to the success of TSVD. However, the truncation parameter cannot be calculated accurately from the data only now but resorts to a series of trial and error (e.g., GCV and L-curve)\cite{14-17}. GCV is a statistical approach estimating the optimal value of the regularization by minimizing the functional:\cite{16}

$$G(k) = \frac{\|A \cdot x(k) - y\|_2^2}{\text{trace}(I - A \cdot A^T)}$$

(5)
where $A_i^j$ is any matrix that maps the right-hand side $y$ on to the solution $x(k)$, i.e., $x(k) = A_i^j \cdot y$.

The L-curve criterion \cite{17} consists of the analysis of the piecewise linear curve with break-points:

$$
(\log \|4x_i - y\|, \log \|Lx_i\|), \quad i = 1, \ldots, q
$$

(6)

where $q$ is the row dimension of the regularization matrix $L$. Intuitively, the best regularization parameter should lie on (or be close to) the corner of the L-curve.

1.3 Partial Least-Squares Regression (PLSR)

PLSR is a regression analysis that regresses the response variables on the extracted score vectors, which serve as new predictors (also called latent variables). \cite{18} Slightly different PLSR algorithms are available in the literature, such as NIPALS, \cite{19} PLS1 algorithms, \cite{20} and SIMPLS. \cite{21} We restrict our attention here on the PLS1 (one response variable) algorithm with the CV criterion ($\text{PRESS}/\text{RSS} \leq 0.95^2$, where $\text{PRESS}$ is the prediction error sum of squares and $\text{RSS}$ is the residual sum of squares). As in \cite{2} and \cite{22}, the PLS1 algorithm of Eq. (2) can be concluded as follows:

- Standardize both $A$ and $y$;
- $A_0 = A$, and $y_0 = y$;
- $k \leftarrow 1$;

While $\text{PRESS}/\text{RSS} \leq 0.95^2$ is not fulfilled do

- $w_k = \frac{A_i^T y_{k-1}}{\|y_{k-1}\|}$, Subject to $w_k^T w_k = 1$
- $t_i = A_{k-1} - w_k$
- $A_k = A_{k-1} - t_i p_i^T$, where $p_i = \frac{A_i^T t_i}{\|t_i\|}$
- $k \leftarrow k + 1$;

end while

- Give the final regression model $\hat{y} = T_k \hat{x}$ and recover the implied regression coefficients, where $T_k = (t_1, \ldots, t_k)$ is the sequence of derived components.

2 Comparative analysis of TSVD with PLSR and stepwise in dam crack forecasting

2.1 Multicollinearity detection

We first adopted Variance Inflation Factor ($VIF$) to detect the impact of multicollinearity among the variables. The $VIF$ can be determined by the following equation: \cite{23}

$$
VIF_i = (1 - R_i^2)^{-1}
$$

(7)

where $R_i$ is the correlation coefficient between an independent variable and the remaining variables.

Since $VIF$ of 10 and above are commonly regarded as indicating multicollinearity, \cite{24} we found that the highly correlated variables would cause a severe multicollinearity problem as evidenced by the large size of $VIF$s in Fig. 1.

![Fig. 1 VIFs of variables](image1)

2.2 Regularization parameter selection

Before building the regularized regression model with TSVD, we performed L-curve and GCV function (part of Regularization Tools \cite{14}) to determine the truncation parameter. Fig. 2 plots the performance of determining truncation parameters, and we found that the truncation parameters using L-curve and GCV are captured to 5 and 8, respectively.

![Fig. 2 Determining truncation parameter](image2)

2.3 Crack modeling and forecasting

We then built two regularized models using the calculated parameters and compared with stepwise regression and PLSR. The “penter” and “premove” values in stepwise are 0.05 and 0.10, respectively. In PLSR analysis, two PCs with $1 - \text{PRESS}/\text{RSS} = 0.0382$ were extracted to estimate regression coefficients. Table 1 presents the estimated coefficients; Table 2 lists comparative summary including coefficient of determination ($R$), $\text{RSS}$, and residual standard deviation ($S$); and Fig. 3 plots the residual curve of TSVD, stepwise,
and PLSR, respectively. We should note herein that the inclusion of stepwise and PLSR is not the main objective of the paper but rather to supplement the identification of TSVD and add confidence in our approach.

Based on the preceding analyses, we establish an empirical (statistical) perdition model as in [2]. We picked up five continuous sample points to test the predictive ability of the models.

Table 1  Regression coefficients

| Factors | Coefficients |
|---------|--------------|
|         | TSVD         | TSVD          | Stepwise | PLSR |
|         | (L-curve)    | (GCV)         |          |      |
| $H^1$   | $-0.0082$    | $0.5884$      | $0$      | $-0.0158$ |
| $H^2$   | $-0.0001$    | $0.0005$      | $0$      | $-0.0003$ |
| $H^3$   | $-1.63 \times 10^{-6}$ | $-0.0002$ | $0$      | $-4.81 \times 10^{-6}$ |
| $T_1$   | $-0.2062$    | $-0.5388$     | $-0.2618$| $0.0083$ |
| $T_2$   | $0.2626$     | $0.2610$      | $0.1872$ | $-0.0020$ |
| $T_3$   | $-0.0070$    | $-0.0683$     | $0$      | $0.0308$ |
| $T_4$   | $-0.3212$    | $-0.2386$     | $-0.3944$| $-0.2205$ |
| $\theta-\theta_0$ | $-0.1566$ | $0.3522$      | $0$      | $-0.1209$ |
| $\ln \theta/\ln\theta_0$ | $-0.2247$ | $-1.8322$     | $-0.5670$| $-0.2237$ |
| const   | $1.0222$     | $-7.5594$     | $1.4210$ | $1.2591$ |

Table 2  Model summary

| Model       | $R$  | RSS (mm$^2$) | $S$ (mm) |
|-------------|------|--------------|----------|
| TSVD (L-curve) | $0.801$ | $27.16$     | $0.527$  |
| TSVD (GCV)  | $0.858$ | $19.99$     | $0.452$  |
| Stepwise    | $0.804$ | $26.80$     | $0.523$  |
| PLSR        | $0.775$ | $30.22$     | $0.555$  |

Table 1 manifests that the coefficients of TSVD model are sensitive to the truncation parameter. Tables 1 and 2 and Fig. 3 confirm that the multicollinearity among variables may not significantly affect the fitting accuracy of regression models but affect the estimation of regression coefficient. Fig. 4 indicates that the severe multicollinearity can lead to a regression model, which fits the training data reasonably well but generalizes poorly to new data. In this case, regression coefficients may be wrongly estimated even with high fitting and predictive accuracy.

Fig. 4  Predicted and observed crack openings

In addition, the PLSR with more PCs included may have better fitting accuracy but will lead to different coefficients obtained, making dam regression models difficult to interpret. In fact, the true value of regression coefficient is mostly unknown in real applications. We therefore cannot decide which model is the most effective and reliable if only the fitting and forecasting accuracy are considered. In light of these concerns, we tentatively recommend using both the fitting accuracy and reasonability of regression coefficients to evaluate the model reliability. Within our knowledge, hydrostatic pressure and thermal gradient generally reflect the elastic displacement under load, and aging reflects the non-elastic behavior of dams. We believe the stepwise model herein may be unreasonable, because most of the water level, temperature, and linear components of aging are excluded. On the other side, we found that cracks contributed by water pressure in TSVD model with GCV function are too large (not provided herein), which is not consistent with the deformation trends of the dam. Both considering the predictive accuracy and coefficient’s reasonability, we believe our TSVD model with L-curve criterion behaves relatively better than PLSR.

3 Conclusion

This paper extends the method of TSVD in dam deformation analysis. We found that TSVD is a promising means to solve the multicollinearity problem in dam regression models. Due to the regularization parameter-choice problem and without prior knowledge of regression coefficients, we have no general rules to make a decision when the TSVD
analysis is superior to PLSR in dam deformation analysis. We tentatively put forward to use both the fitting accuracy and the coefficients’ reasonability to evaluate the model reliability. There are still remaining problems that require further studies.

References

[1] Wu Zhongru (1990) Theory of safety monitoring of hydraulic constructions and its applications [M]. Nanjing: Hohai University Press (in Chinese)

[2] Deng Nianwu, Wang Jianguo, Anna Szostak Chrzanowski (2008) Dam deformation analysis using the partial least squares method [C]. The 13th FIG Int. Symp. on Deformation Measurements and Analysis & 4th IAG Symp on Geodesy for Geotechnical and Structural Engineering, Lisbon

[3] Yang Jie (2006) Analysis methods and applications research on empty uncertainty problems in large dam monitoring and control [D]. Nanjing: Hohai University (in Chinese)

[4] Roman Rosipal, Leonard J Trejo (2001) Kernel partial least squares regression in reproducing Kernel Hilbert Space [J]. Journal of Machine Learning Research, 2: 97-123

[5] Li Xuehong, Xu Hongzhong, Gu Chongshi, et al. (2001) Application of principal component fuzzy neural network model to analysis of observation data of dams [J]. Dam Observation and Geotechnical Tests, 25(5): 14-16 (in Chinese)

[7] Su Huazhi, Wen Zhiping, Wu Zhongrui (2009) An early-warning model of dam safety based on SVM theory [J]. Journal of Basic Science and Engineering, 17(1): 40-48

[8] Ian T Jolliffe (1982) A Note on the use of principal components in regression [J]. Journal of the Royal Statistical Society, Series C (Applied Statistics), 31: 300-303

[9] Tychonoff A N, Arsenin V Y (1977) Solution of ill-posed problems [M]// Gani J (Eds.). Perspectives in Probability and Statistics: Papers in Honour of M S Bartlett on the occasion of his sixty-fifth birthday. London: Academic Press

[10] Hofmann B (1986) Regularization for applied inverse and ill-posed problems [M]. Leipzig: Teubner

[11] Neumaier A (1998) Solving ill-conditioned and singular linear systems: A tutorial on regularization [J]. SIAM Review, 40: 636-666

[12] Wu Xiongwei, Cheng Weiping (2007) SVD algorithm-based dam monitoring data regression model [J]. Hydropower Automation and Dam Monitoring, 31(3): 53-33 (in Chinese)

[13] Jing Ji, Bao Tengfei, Gu Yanchang, et al. (2007) Analysis of ridge regression for the multicollinearity of monitoring data [J]. Hydropower Automation and Dam Monitoring, 31(3): 56-59 (in Chinese)

[14] Hansen P C (1994) Regularization tools: A Matlab package for analysis and solution of discrete ill-posed problems [J]. Numerical Algorithms, 6(1): 1-35

[15] Wu Limin (2003) A parameter choice method for Tikhonov regularization [J]. Electronic Transactions on Numerical Analysis, 16: 107-128

[16] Elden L (1984) A Note on the computation of the generalized cross-validation function for ill-conditioned least squares problems [J]. BIT, 24: 467-472

[17] Hansen P C (1992) Analysis of discrete ill-posed problems by means of the L-curve [J]. SIAM Rev, 34(4): 561-580

[18] Roman Rosipal, Nicole Krämer (2006) Overview and recent advances in partial least squares [C]. SLSFS 2005, LNCS, Bohinj, Slovenia

[19] Wold H (1975) Soft modelling by latent variables: the nonlinear iterative partial least squares approach [M]// Gani J (Eds.). Perspectives in Probability and Statistics: Papers in Honour of M S Bartlett on the occasion of his sixty-fifth birthday. London: Academic Press

[20] Wold S, Martens H, Wold H (1983) The multivariate calibration problem in chemistry solved by the PLS method [J]. Lecture Notes in Mathematics, 973: 286-293

[21] S de Jong (1993) SIMPLS: An alternative approach to partial least squares regression [J]. Chemometrics and Intelligent Laboratory Systems, 18: 251-263

[22] Kondylis A (2006) PLS methods in regression model assessment and inference [D]. Switzerland: University of Neuchâtel

[23] Fox J, Monette G (1992) Generalized collinearity diagnostics [J]. Journal of the American Statistical Association, 87: 178-183

[24] Robert M O’Brien (2007) A caution regarding rules of thumb for variance inflation factors [J]. Quality and Quantity, 41(5): 673-690

[25] Wu Daowen, Han Dajian (1996) Effect of independent variable correlation on regression analysis of dam monitoring data [J]. Dam Observation and Geotechnical Tests, 20(3): 23-25 (in Chinese)