독립성분분석을 이용한 다변량 시계열 모의

Multivariate Time Series Simulation With Component Analysis

이태삼[1], 호세 살라스[2], 주하 카바넨[3], 노재경[4]

Taesam Lee, Jose D. Salas, Juha Karvanen and Jaekyoung Noh

Abstract

In hydrology, it is a difficult task to deal with multivariate time series such as modeling streamflows of an entire complex river system. Normal distribution based model such as MARMA (Multivariate Autorgressive Moving average) has been a major approach for modeling the multivariate time series. There are some limitations for the normal based models. One of them might be the unfavorable data-transformation forcing that the data follow the normal distribution. Furthermore, the high dimension multivariate model requires the very large parameter matrix. As an alternative, one might be decomposing the multivariate data into independent components and modeling it individually. In 1985, Lins used Principal Component Analysis (PCA). The five scores, the decomposed data from the original data, were taken and were formulated individually. The one of the five scores were modeled with AR-2 while the others are modeled with AR-1 model. From the time series analysis using the scores of the five components, he noted "principal component time series might provide a relatively simple and meaningful alternative to conventional large MARMA models". This study is inspired from the researcher's quote to develop a multivariate simulation model.

The multivariate simulation model is suggested here using Principal Component Analysis (PCA) and Independent Component Analysis (ICA). Three modeling step is applied for simulation. (1) PCA is used to decompose the correlated multivariate data into the uncorrelated data while ICA decomposes the data into independent components. Here, the autocorrelation structure of the decomposed data is still dominant, which is inherited from the data of the original domain. (2) Each component is resampled by block bootstrapping or K-nearest neighbor. (3) The resampled components bring back to original domain. From using the suggested approach one might expect that a) the simulated data are different with the historical data, b) no data transformation is required (in case of ICA), c) a complex system can be decomposed into independent component and modeled individually. The model with PCA and ICA are compared with the various statistics such as the basic statistics (mean, standard deviation, skewness, autocorrelation), and reservoir-related statistics, kernel density estimate.

key words: principal component analysis, independent component analysis, multivariate simulation

[1] 정회원 · Department of Civil and Environmental Engineering, Colorado State University, CO., USA · E-mail : tae3lee@gmail.com

[2] 비회원 · Department of Civil and Environmental Engineering, Colorado State University, CO., USA · E-mail : jsalas@engr.colostate.edu

[3] 비회원 · Department of Health Promotion and Chronic Disease Prevention, National Public Health Institute, Mannerheimintie 166, FIN-00300 Helsinki, Finland · E-mail : juha.karvanen@ktl.fi

[4] 정회원 · 충남대학교 지역환경토목과 교수 · E-mail : jknoh@cnu.ac.kr

1. Introduction

Stochastic simulation of streamflows has been employed for the evaluation of alternative designs and operation rules, analysis of significant drought and flood events. The intricacy of multivariate modeling is not limited only on the hydrologic fields. Statistician describes it as 'the curse of dimension' referred from Bellman (1957) as the exponential growth of hypervolume as a function of dimensionality. Therefore, we suggest the multivariate simulation model using the decomposition of the multivariate data into independent variables and modeling the variables with the various univariate model KNN. In the suggested model, the multivariate variables are decomposed into the other variables that are independent on each variable. The Principal and Independent component analysis are applied for decomposition of the variables. From the decomposed data, the univariate parametric and nonparametric models are applied for simulation. And the individually simulated decomposed data is back-transformed into the original domain.

2. Multivariate Decomposition Analysis and time series simulation modeling

The multisite data can be decomposed into independent or uncorrelated component with ICA or PCA, respectively. The composition analysis might be adopted in the multi-site streamflow modeling. A possible approach is to decompose the multivariate data into independent sources. Each independent sources are modeled using time series model. And the independent sources are generated from the fitted model and back-transformed into the original domain. The applied generation procedure is illustrated in Fig. 1. Karvanen (2003) tested the generation of correlated non-gaussian random variables with generalized lambda distribution for the sources. For individual time series modeling for each component, k-nearest neighbor resampling (Lall and Sharma, 1996) is applied. Notice that the estimated sources from ICA are not normally distributed. Traditional normal based time series model such as ARMA(p,q) model might also be employed with the appropriate transformation for ICA.

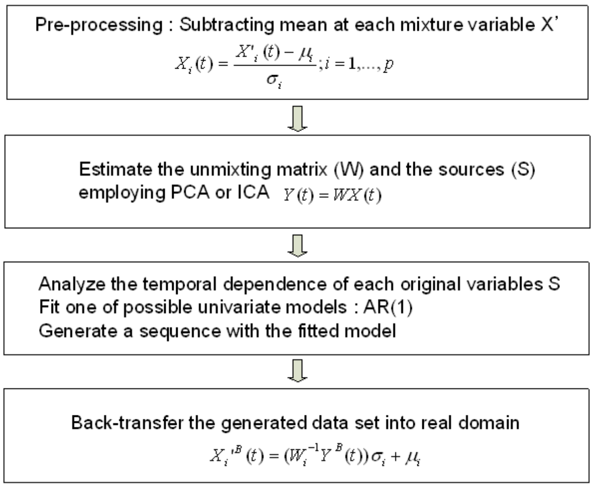
Pre-processing : Subtracting mean at each mixture variable X'
$X_i(t) = \frac{X'_i(t) - \mu_i}{\sigma_i}; i = 1,\ldots,p$
Estimate the unmixting matrix (W) and the sources (S)
employing PCA or ICA $Y(t) = WX(t)$
Analyze the temporal dependence of each original variables S
Fit one of possible univariate models : AR(1)
Generate a sequence with the fitted model
Back-transfer the generated data set into real domain
$X_i'^B(t) = (W_i^{-1}Y^B(t))\sigma_i + \mu_i$

Fig. 1. Flowchart for Decomposition and Bootstrapping or AR(1) modeling

3. Modification of Time series modeling of PCA and ICA

Time independency and Modified Principal Component Analysis (PCA)

The PCA and ICA algorithms linearly transform the mixture variables into uncorrelated or independent variables, respectively. However, in their algorithm the timely crossed dependence structure are not handled. This might lead to remain the time-lagged cross dependence in the uncorrelated or independent components. The dependent structure in different levels is presented in Fig. 2 in case of two variables presented as

$$\mathbf{Y}_t = \begin{bmatrix} Y_t^1 \\ Y_t^2 \end{bmatrix} = \begin{bmatrix} w_{11}X_t^1 + w_{12}X_t^2 \\ w_{21}X_t^1 + w_{22}X_t^2 \end{bmatrix} = \mathbf{W}\cdot\mathbf{X}_t$$

The full line implies the full strength of dependency, the dash line no dependency, and the dashed-two dots line is the loss of dependency. In Fig. 2, the dependent structure of the multi-site streamflow data (original variables X) presented in (a). From the PCA or ICA algorithm, the decomposed variables (Y) will have the structure in (b). If each variable is modeled individually represented by (c), the final dependent structure might be resulted in (d). The underestimation of the serial dependency in the original variables is a significant consideration.

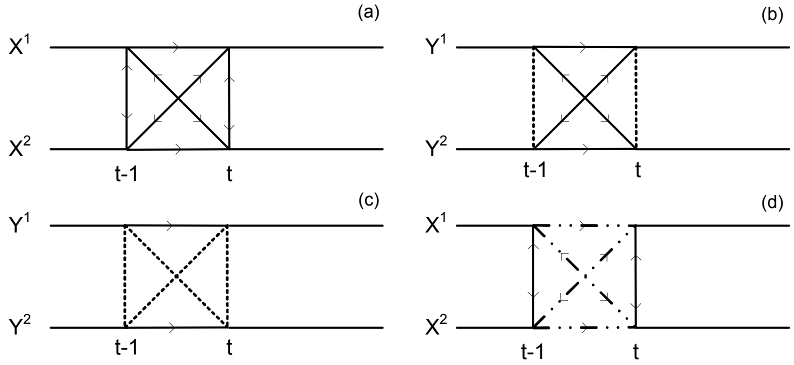

Fig. 2. Dependent structure of the mixture variables, X, and the decomposed variabes, S in different cases : dependent structure of (a) original data, (b) decomposed data, (c) individual modeling, (d) back-transformed data where the full line implies that full strength of dependency, the dash line no dependence, and the segment line underestimation of dependency

When S=2, $\zeta_h^{ij}(z) = E[z_{t-h}^i z_t^j]$, and $\{\psi_{ij}\}_{i,j\in\{1,2..m\}} = \Psi = \mathbf{W}^{-1}$ as an example, lag-1 auto-covariance of X of the first variable is

$$\zeta_1^{11}(x) = \psi_{11}\psi_{11}\zeta_1^{11}(y) + \psi_{11}\psi_{12}\zeta_1^{21}(y) + \psi_{12}\psi_{11}\zeta_1^{12}(y) + \psi_{12}\psi_{12}\zeta_1^{22}(y) \quad (1)$$

By assuming that $\zeta_h^{ij}(y) = 0$,$i \neq j$

$$\zeta_1^{11}(x) = \psi_{11}\psi_{11}\zeta_1^{11}(y) + \psi_{12}\psi_{12}\zeta_1^{22}(y) \quad (2)$$

$$\zeta_1^{22}(x) = \psi_{21}\psi_{21}\zeta_1^{11}(y) + \psi_{22}\psi_{22}\zeta_1^{22}(y) \quad (3)$$

In matrix form,

$$\mathbf{\Lambda}(x) = \mathbf{\Psi}^{(2)}\mathbf{\Lambda}(y) \tag{4}$$

where $\mathbf{\Lambda}(z) = [\zeta_1^{11}(z), \zeta_1^{22}(z)]'$ and $\mathbf{\Psi}^{(2)} = \{(\psi_{ij})^2\}_{i,j\in\{1,2\}}$. $\mathbf{\Lambda}(y)$ is solved easily by

$$\mathbf{\Lambda}(y) = \left[\mathbf{\Psi}^{(2)}\right]^{-1}\mathbf{\Lambda}(x) \tag{5}$$

The decomposed variables with PCA can be modeled utilizing the estimate of Eq.(5). A parametric model should be used for this purpose. In this study, the decomposed data are modeled with the simple time series model AR(1) (Salas et al., 1980) formulated as

$$y_t = \mu_y + \phi_1(y_{t-1} - \mu_y) + \varepsilon_t \tag{6}$$

where $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$ and $\phi_1 = \mathrm{cov}[Y_t Y_{t-1}]/\mathrm{var}[Y_t] = E[Y_t Y_{t-1}]/E[YY_t]$, $\sigma_\varepsilon^2 = \sigma_y^2(1-\phi^2)$ and $E[Y_t] =$ μ_y =0 as mentioned already. From the Cauchy-Schwarz inequality theorem (Grimmett and Stirzaker, 2001) such that,

$$\left|E[Y_t Y_{t-1}]\right| \le \left(E\left[Y_t^2\right]E\left[Y_{t-1}^2\right]\right)^{1/2} \quad \text{or} \quad \left|\mathrm{cov}[Y_t Y_{t-1}]\right| \le \mathrm{var}(Y_t) \tag{7}$$

This implies that the auto-covariance of higher order components can be neglected in case that the variance of the components is significantly smaller than the low order components. For example, it can be describe as $\zeta_1^{11}(x) = \psi_{11}\psi_{11}\zeta_1^{11}(y)$ from neglecting the autocovariance, $\zeta_1^{22}(y)$. Rather than using the parameter estimation , optimization approach can be employed such that

$$\mathbf{\Lambda}(y) = \min\left\{\sum_{i=1}^{V}\left|\sum_{j=1}^{V}\psi_{ij}^2\zeta_1^{jj}(y) - \zeta_1^{ii}(x)\right|\right\} \tag{8}$$

where, $\zeta_1^{ll}(y) = 0$ where *l* is the selected orders from the highest. This minimization problem is solved with the taxi cab method (Powell, 1998).

4. Data Description and Application

The Colorado River System consists of 29 selected gaging stations that characterize the river flow (Fig. 2). The historical gaged data has been naturalized over the 29 stations through 2003. Part of the data has been extended by Lee and Salas (2006) back to 1906. Nine Colorado River streamflow sites are selected for the research area. All selected sites are in Green river basin. Sites 8, 16, and 20 are chosen as the representative sites. The applied models is (1) PCA with KNN - PCA (2) PCA with AR(1) - M_PCA (3) ICA with KNN - ICA_ML and (4) ICA with TAR(1) - M_ICA_ML. The Skewness and lag-1 correlation are illustrated at Fig. 3. The PCA and M_PCA does not preserve the observed skewness property while ICA_ML and M_ICA_ML reproduces this statistics pretty well. And lag-1 serial correlation statistics are well preserved in all models except ICA_ML which has slight underestimation. More detailed statistics are tested, not shown here. Therefore, we can conclude that ICA decomposition multivariate modeling have the ability to reproduce the higher order statistics such as skewness and the time-lagged serial correlation also can be modeled with appropriate modification.

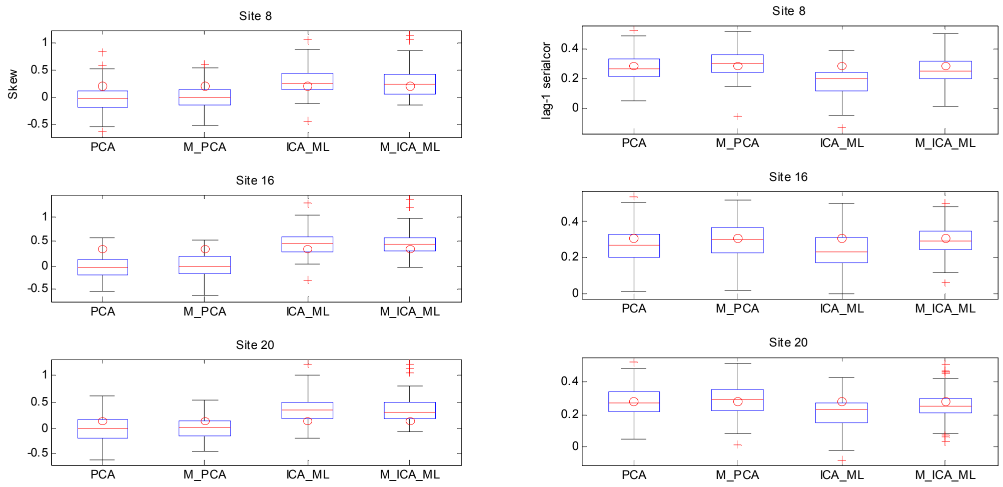

Fig. 3. Skewness (Left) and Lag-1 serial correlations (Right) of the historical data (circle) and 100 set of generated data (boxplot) with PCA, M_PCA, ICM_ML, and Modified ICA_ML

5. Conclusions

From the modified model, brief conclusions can be made. PCA or ICA with individual component models can be reasonable alternatives for multivariate model. The modification for the suggested model to improve the preservation of the historical statistics is successful. Further improvements is required to account for longer temporal dependence.

References

1. Bellman, R.E.(1957). Dynamic Programming, Princeton University Press, Princeton, NJ.
2. Grimmett, G.R. and Stirzaker, D.R.(2001). Probability and Random Processes, Oxford press.
3. Karvanen J., Koivunen V.(2004). Independent component analysis via optimum combining of kurtosis and skewnes-based cirteria, J. of the Franklin Institute, V.341, pp.401-418.
4. Karvanen, J.(2003). Generation of correlated non-Gaussian random variables from independent components, Proceedings of Fourth International Symposium on Independent Component Analysis and Blind Signal Separation, ICA2003, pp.769-774, Nara, Japan.
5. Lall, U., and Sharma, A.(1996). A nearest neighbor bootstrap for resampling hydrologic time series, *Water Resources Research*, 32(3), 679-693.
6. Lee, T., and Salas, J.D.(2006). Record Extension of Monthly Flows for the Colorado River System, Bureau of Reclamation, U.S. Department of Interior.