Modelling and Prediction for Functional Relationships between Time-series

X. Lu

***Abstract*—Much research in the sciences involves modelling and prediction of the functional relationship between simultaneous time series. In this paper we present a new method to infer functional relationship between simultaneous time series based on the modified version of Singular Value Decomposition method. We firstly extract the dominant relationships, perform pattern analyses on the dominant relationships, and construct the model equations of functional relationship between the time series through the time before the forecast period of interest. We then conduct predictions on the future values of one time series from another time series based on the model equations. Several regression schemes are proposed to serve as a prediction basis. The proposed model is applied to predict the real simultaneous health outcome time series. The simulation, the predictions and the real data show good agreement.**

***Index Terms*—Time series, functional relationship, mathematical modelling, prediction.**

I. Introduction

Much research in the sciences involves modelling the relationship between simultaneous time series. For example, the current research on health and medical fields links to various aspects of inferring functional relationships between outcomes and predictors in order to examine weather the relationships are dynamic across the investigated factors [1]-[5]. Inferences vary in type and degree depending on the purpose for it and the context in which it is performed. Very often, it deals with prediction of one time series with the knowledge of another time series and its past values. One of the key advantages of studying functional relationships is that it can provide a deep quantitative and qualitative understanding of how system parameters act and the mechanisms behind work. It gives new insights into the mechanisms of the data based system.

However, methods of modelling such functional relationship are at present inadequate. A common approach is based on correlation analysis which implies only statistical dependence and is rather crude [6]. Correlation is often used as a simple and naive measure for the statistical dependence between simultaneous time series.

In this paper, we propose a new approach to model the functional relationship between simultaneous time series. The modified version of Singular Value Decomposition (SVD) method is adopted to capture dominant relationships. Then prediction of the future values of one time series from another time series is based on a regression model such as time series and other statistical analysis. Several regression schemes are proposed to serve as a prediction basis. The proposed model is tested using the obtained simultaneous time series data regarding the computer-related workload and health outcomes and achieves satisfactory results.

Manuscript received March 23rd, 2009.

Xiaoshu Lu is with the Finnish Institute of Occupational Health, Topeliuksenkatu 41 a A, FIN-00250 Helsinki, FINLAND (corresponding author; phone: 358-30-4742505; fax: 358-30-4742008; e-mail: xiaoshu@cc.hut.fi).

II. Model Development

A. The Model

The model is an extension of our previous model [7]. Here we explain only general ideas. We apply a modified version of SVD [8] to extract the dominant relationships between the simultaneous time series. Firstly, we generate two matrices from the simultaneous time series. The matrix construction depends on the analysis purposes that come with the data and other conditions. We then apply SVD to the matrices to capture the dominant right singular vectors and regress the right singular vectors. The left singular vectors are analysed to construct simulation model equation that can be used for further predictions. The analysis phase takes early time series points. The retrospective analysis provides an indication of how closely the simulation model matches actual data. Future projections are made based on both the analysis phase and the constructed simulation model. The proposed methodology is outlined below:

Step 1: Generating two sample $m \times n$ matrices A_x and A_y from the simultaneous time series;

Step 2: Applying SVD to A_x and A_y to capture the dominant right singular vectors v_{xi}, v_{yj}, and the left singular vectors u_{xi}, u_{yj};

Step 3: Regressing the dominant right singular vectors v_{xi} and v_{yj};

Step 4: Analysing and constructing the model equations between the left singular vectors u_{xi} and u_{yj};

Step 5: Modelling functional association by combining the above regression equations and making predictions.

B. Model Predictions

Depending on the study goals, the appropriate method is not apparent. We suggest four primary methods used for the prediction:

(i). Linear model

By linear model, we mean the classical linear model. The most important assumption of the model is that the

observations of the dependent variables are uncorrelated. Under this assumption, the maximum likelihood parameter estimates can be obtained with well-known least square method. The application of the model will lead to a loss of power if there is correlation among the observations. Such complication can be handled with other modelling approaches for example: marginal and random effect models.

(ii). Marginal model

An extension of the linear model is the marginal model which incorporates correlations among the observations. The marginal variance depends on the marginal means and the variance-covariance structure which specifies the correlation structure of the observations. The parameters are commonly estimated using Generalised Estimating Equations (GEE) [9] rather than a likelihood based method.

(iii). Random effect model

The fundamental difference between marginal and random effect models is that the latter allows the regression of dependent variables on the independent variables differ among the subjects by introducing random-effects parameters. The random parameters, or random variables, can vary, for example, under repeated sampling with more flexible specification of the covariance matrix [10]-[11]. The parameters are usually estimated using likelihood based methods. Very often, it is difficult to justify a particular distribution for the random effects. Maximum likelihood estimation based on the marginal distribution of the observations integrates out the random effects [12].

(iv). Time series model

Much of the literature assumes that the mapping from one time series to another can be adequately approximated over the range of interest by the impulse response function. The Box-Jenkins approach [13] has been widely used to derive the impulse response function. The Box-Jenkins approach starts with fitting an autoregressive integrated moving average (ARIMA) model to the differenced times series for prewhitening the series. The cross-covariance function of the filtered time series is then applied to obtain a good estimate of the impulse response function. For a more detailed calculation explanation, refer to Box & Jenkins [13] where more estimation methods can be found for example the commonly used state-space and the Kalman-filter models.

III. APPLICATION OF THE MODEL TO HEALTH OUTCOME TIME SERIES DATA

To apply and validate the model, a measured health outcome time series dataset is employed. The dataset is simultaneous time series based on the daily records of human computer-related work exposure and the correspondent health outcomes expressed as discomfort ratings at different body sites including eyes, head, neck, and many others. We firstly identify the functional association between computer-related work exposure series and the discomfort ratings. Based on the results, prediction of the future discomfort ratings is made. The measured discomfort ratings are plotted against the simulated and predicted ones for assessing the accuracy of the model.

A. Data

The study population consists of office staff in Finland. They did office work for at least four hours a day and had reported a moderate amount of musculoskeletal symptoms. The data collection procedure was carried out in two-week periods before the intervention, and at the 2-month and 10-month follow-up. The computer-related work exposure was measured with the software (Work-Pace^{TM}, Niche Software Limited, New Zealand) which continuously monitored the staff' keyboard and mouse entries with an accuracy of ten milliseconds. Data were then summed up presenting computer-related workload as a daily base. Simultaneously with the recordings of computer use the subjects were asked to fill in a questionnaire-diary three times a day: in the morning, at noon and in the evening. The questionnaire-diary contained a body map diagram and questions about the existence of musculoskeletal discomfort in different body regions. Each item was assessed using 5-point rating scale from "5-feel good" to "1-feel very uncomfortable". Data were averaged to indicate the daily health outcomes expressed as discomfort ratings at different body sites. A detailed description of the data collection procedures can be found in [14]. For illustrative clarity, not any deeper reason, only the discomfort ratings of eyes, head and neck are selected as health outcomes to demonstrate the model equations and the analysis results.

B. Model Equations

The model equations were constructed from the simultaneous time-series of both computer-related workload and discomfort ratings from week one to three. The model was applied to forecast health outcomes for week four. Previous studies have shown that both the computer-related workload and discomfort ratings had only one dominant pattern and contributed over 90% variability [7]. Therefore i, j =1 from Step 1 to Step 4 which are omitted in the following equations. The model equations are

$$\boldsymbol{v}_x = \alpha_{workload\_1} + \alpha_{workload\_2}\, t \tag{1}$$

$$\boldsymbol{v}_y = (\alpha_1 + \frac{\alpha_2 - \alpha_1}{1 + 10^{t-\alpha_3}}) \tag{2}$$

and their closed-form is

$$\boldsymbol{v}_y = (\alpha_1 + \frac{\alpha_2 - \alpha_1}{1 + 10^{\frac{v_x - \alpha_{workload\_1}}{\alpha_{workload\_2}} - \alpha_3}}) \tag{3}$$

where α_1, α_2, and α_3 are body site-related parameters such as eyes, head and neck.

The next step, Step 3, is to analyse and construct the model equations between the left singular vectors u_x and u_y which are two time series denoted here as $\{X_t\}$ and $\{Y_t\}$ for

convenience. Very often, a time series model, ARIMA model, is used for forecasting purpose [13]. The model is superior to many common time series and multivariate regression models in that it accounts for the correlation between the error residual and the lagged values. For example, the 'weekly differences' in this study are likely to be autocorrelated since the measurements were made from a single subject. Therefore ARIMA model should be generally applied to model $\{X_t\}$ and $\{Y_t\}$.

However, due to the short measurement time for small size of the data in this study, ARIMA model cannot be constructed. Therefore, a relatively simple forecasting method, a generalised linear model, that relies exponential smoothing method is adopted. Moreover, considering the intervention effect on health outcomes, more weight to recent observations and less weight to observations further in past are assigned in the forecasting procedure [15]. The forecasting is performed in the following way:

$$Y_{t+1} = h_{t+1}\, X_{t+1} \tag{4}$$

where

$$h_{t+1} = c_0\, h_t + c_1\, h_{t-1} + \ldots \text{ and } c_i = 0.8\,(1-0.8)^i \tag{5}$$

Fig 1 to Fig. 7 show the results. Because of the small size of the data which do not allow statistical evaluation of agreement, the model performance is studied through a rigorous direct comparison between simulation and measurements only. Simulation is made for two-subject observations.

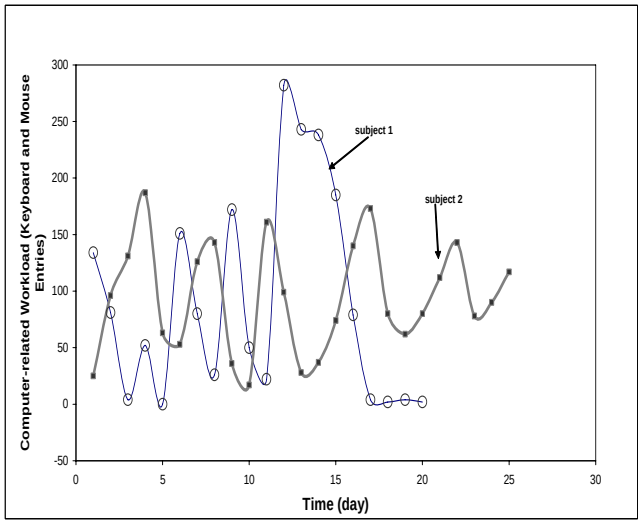

Fig. 1. Computer-related workloads for two subjects evaluated with the numbers of keyboard and mouse entries.

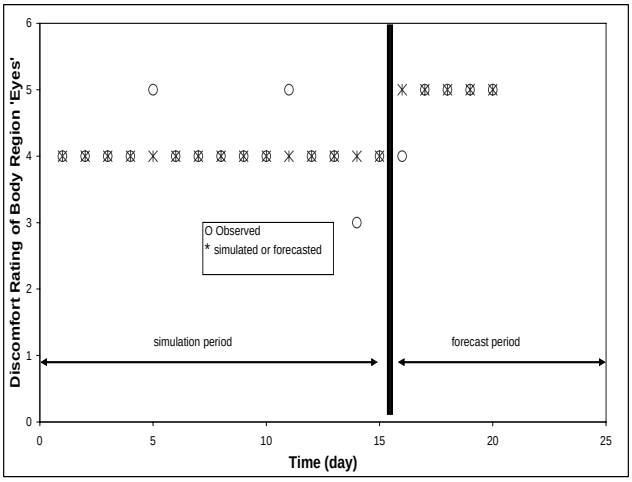

Fig. 2. Eye-discomfort rating for subject 1; 5, feel good; 1, feel very uncomfortable.

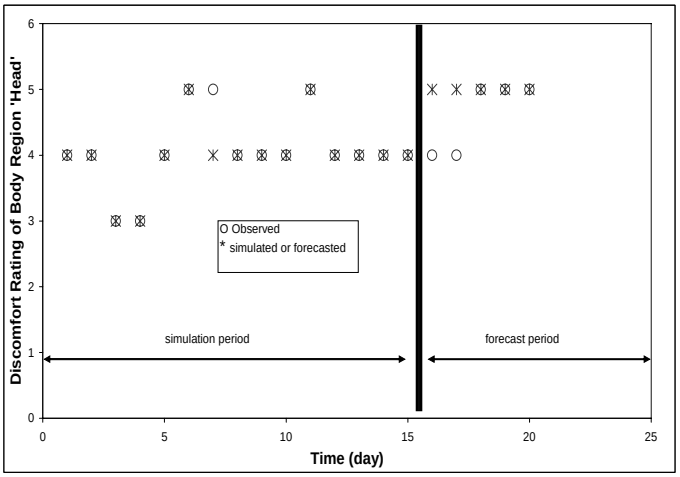

Fig. 3. Head-discomfort rating for subject 1; 5, feel good; 1, feel very uncomfortable.

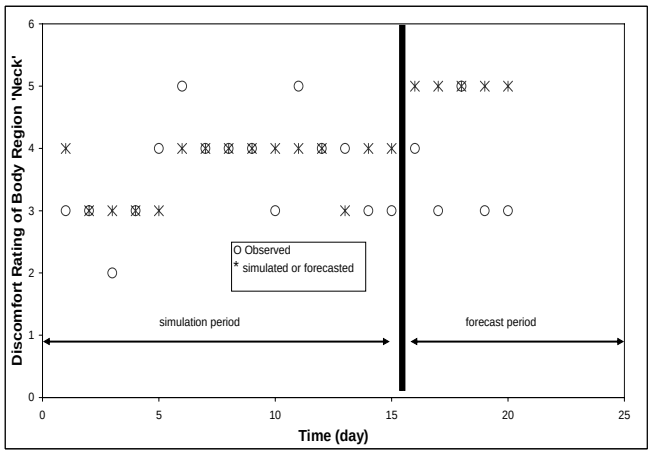

Fig. 4. Neck-discomfort rating for subject 1; 5, feel good; 1, feel very uncomfortable.

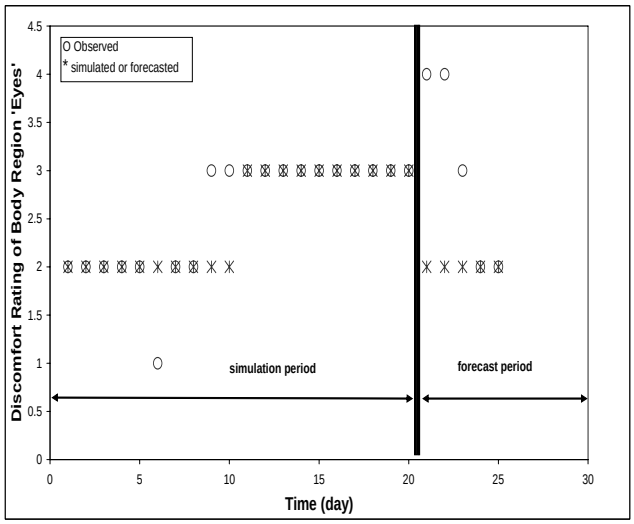

Fig. 5. Eye-discomfort rating for subject 2; 5, feel good; 1, feel very uncomfortable.

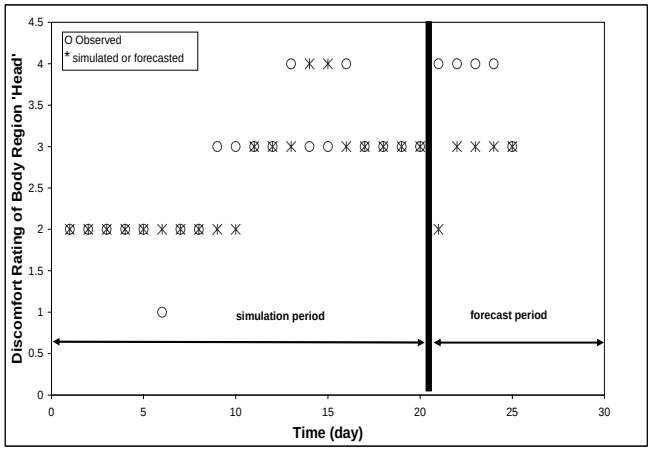

Fig. 6. Head-discomfort rating for subject 2; 5, feel good; 1, feel very uncomfortable.

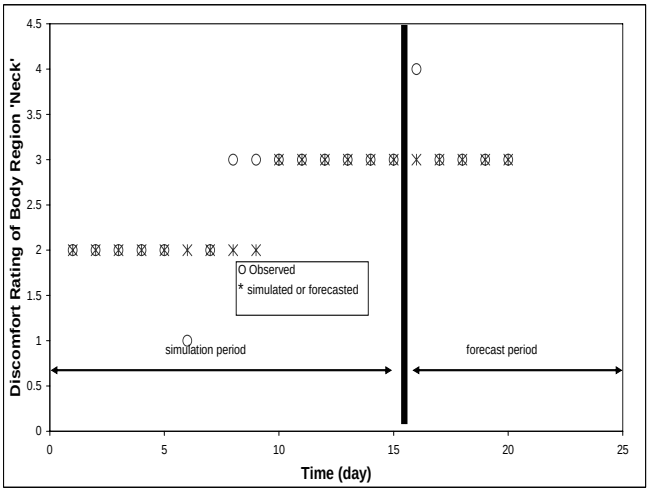

Fig. 7. Neck-discomfort rating for subject 2; 5, feel good; 1, feel very uncomfortable.

Note that even though individual workloads varied several orders of magnitude depending on many unknown factors (Fig.1), the simulation and forecasting are satisfactory.

IV. CONCLUSIONS

Functional relationship between simultaneous time series can assist in characterization of mechanisms of the study system and is studied in various applications. Most often, statistical dependence, for example cross-correlation alone, is used to examine the strength of the relation between the simultaneous time series even if there is a strict functional relationship between the time series. Functional association is more desirable goal than just simplicity of the correlation as it allows for prediction to a certain extent. This paper provides a procedure to find and construct the closed form of functional relationship. SVD procedure offers a way to carry it out. By using SVD, the dominant relationships between two time series can be captured. The method provides a simple and robust data-driven procedure to handle various noisy time series depending on the data structures and study purposes. In addition, computation algorithms are relatively simple which are easily computed by computers with available commercial software. The functional relationship can be used to explore complex interplay among the mechanical and physical factors which govern the system and to predict the future values of one time series based on the other time series.

The dataset of measured computer-related workload and health outcomes was used to test the proposed model with promising results even though the data suffer from a number of limitations such as collection of time series of the data is short.

ACKNOWLEDGMENT

I am indebted to Dr. Esa-Pekka Takala and Risto Toivonen for providing discussion and the measurement data.

REFERENCES

[1] T.W. Beck, T.J. Housh, J.T. Cramer, J.P. Weir, G.O. Johnson, J.W. Coburn, M.H. Malek, and M. Mielke. (2005). Mechanomyographic amplitude and frequency responses during dynamic muscle actions: a comprehensive review. *BioMedical Engineering OnLine*. 4 (1). pp.67. (DOI: 10.1186/1475-925X-4-67).

[2] D.L. Rhatigan, A.E. Street, and D.K. Axsom. (2006). A critical review of theories to explain violent relationship termination: Implications for research and intervention. *Clinical Psychology Review*. 26 (3). pp. 321-345.

[3] C.A. Stratakis. (2006). Cortisol and growth hormone: clinical implications of a complex, dynamic relationship. *Pediatr Endocrinol Rev*. 3 (Suppl 2). pp.333-338.

[4] S.L. Whitney, G.F. Marchetti, and A.I. Schade (2006). The Relationship between falls history and computerized dynamic posturography in persons with balance and vestibular disorders. *Archives of Physical Medicine and Rehabilitation*. 87 (3), pp. 402-407.

[5] R.D. Conger, and M.B. Donnellan. (2007). An interactionist perspective on the socioeconomic context of human development. Annual Review of Psychology. 58. pp.175-199.

[6] J.M. Sonnergaard. (2006). On the misinterpretation of the correlation coefficient in pharmaceutical sciences. *International Journal of Pharmaceutics* 321. pp.12-17.

[7] X. Lu, and E-P. Takala,. (2008). A novel mathematical approach for modelling and characterising time-dependent musculoskeletal outcomes for office staff. *Statistics in Medicine*. 27. pp.4549-4568.

[8] G. Golub, and C.F. Van Loan, *Matrix Computations*. Maryland: Third ed, 2715 North Charles Street, Baltimore, 1996.

[9] K.Y. Liang, and S.L. Zeger. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*. 73. pp.13–22.

[10] C.R. Henderson, "Statistical Method in Animal Improvement: Historical Overview", In *Advances in Statistical Methods for Genetic Improvement of Livestock*, D. Gianola, and K. Hammond, Ed, New York: Springer-Verlag, 1990, pp. 2-14.

[11] S.R. Searle, G. Casella, and C.E. McCulloch, *Variance Components*. New York: John Wiley & Sons, Inc., 1992.

[12] J.C. Pinheiro, and D.M. Bates, *Mixed Effects Models in S and S-Plus*. Springer-Verlag: New York, 2000.

[13] G.E.P. Box, and G.M. Jenkins, 1970 *Time Series Analysis: Forecasting and Control*. San Francisco, CA: Holden-Day, 1970.

[14] R. Ketola, R. Toivonen, M. Häkkänen, R. Luukkonen, E-P. Takala, and E. Viikari-Juntura. (2002). Effects of ergonomics intervention in work with video display units. *Scand. J. Work. Environ. Health*. 28. pp.18-24.

[15] C. Chatfield, *The Analysis of Time Series An Introduction*. Chapman & Hall: London, 1996.