Modeling Multivariate Time Series with Uneven Spacing on Multiple Time Scales to Estimate Large Scale Time Trend

T. Zeleny[1,*] and D. Marx[1]

[1] *University of Nebraska-Lincoln, Department of Statistics; tzeleny1@huskers.unl.edu, david.marx@unl.edu*
Corresponding author

Abstract. *In certain situations, observations may be made on a multivariate time series on a given temporal scale. However, there may be an underlying, unobserved time series on a larger temporal scale that is of greater interest. Often times, identifying the behavior of the data over the course of the larger scale is a key objective. Because this large scale trend is not being directly observed, describing the trends of the data can be more difficult. To further complicate matters, the observed data on the smaller time scale can be unevenly spaced from one larger scale time point to the next. This means it may be more appropriate to view the observations as coming from multiple, shorter multivariate time series occurring at each large scale time point as opposed to a single, long multivariate time series. We discuss the process of modeling each of these smaller scale multivariate time series and obtain estimates of the corresponding parameters. We then introduce a method to use these parameter estimates to estimate the unobserved values of the larger scale multivariate time series at each large scale time point. A model is then fit to the estimated unobserved values of the larger scale multivariate time series in order to estimate the parameters that describe the large scale temporal trends of the data.*

Keywords. *Multivariate Time Series; Multiple Time Scales; Multivariate Multiple Regression; Multivariate Delta Method*

1 Motivating Example

The motivation for this model comes from data collected by the University of Nebraska football team on players lifting weights. Each weight lifting station is equipped with motion capture technology that is able to measure the velocity of the bar as the athletes perform each lift.

For each rep that is completed, both the average power over the course of the rep and the peak power over the course of that rep are recorded. This results in a bivariate time series of data for each player, with average power and peak power being the two variables.

The data in this example occurs on two different time scales: the rep-to-rep scale within a single day and the day-to-day scale within a training session. For example, an athlete may complete between 15 to 30 reps for a given exercise in a given day. The winter training session lasts approximately six weeks with athletes working out two to three times per week, resulting in 12 to 18 days of data collection. Of course the observations are made on the smaller rep-to-rep scale; however, the coaches may be more interested in the trends of the athletes' lifting performances on the larger day-to-day scale over the course of the training period.

Further, when looking at the data on the rep-to-rep scale over the course of the entire training period, there will be wide variations in spacing. The time spacing between consecutive reps within the same day will likely be only seconds apart, while the time spacing between consecutive reps across days will be many hours apart. Because of this, when considering the data on the rep-to-rep scale it may be more appropriate to treat data collected on separate days as separate, shorter bivariate time series that may exhibit different behavior as opposed to a single, longer bivariate time series. By doing this, the performance of an athlete on a given day should be more accurately modeled.

Once the rep-to-rep bivariate time series have been evaluated for each day, a measure of an athlete's performance for each day can be obtained for both variables. This day-to-day performance for the two variables represents the bivariate time series on the larger day-to-day scale. The trends present in this large scale time series are the ones that are of interest here. The goal is to develop a model that is able to make these evaluations on both the small and large scales and estimate these large scale trends.

2 Vector Autoregression

One of the most common, successful, and flexible models for analyzing multivariate time series is the vector autoregressive (VAR) model. Suppose $\mathbf{x}_t = (x_{1t}, x_{2t}, ..., x_{nt})'$ is an (n x 1) vector of time series variables, then a general p-lag VAR model (known as VAR(p)) can be written as

$$\boldsymbol{x}_t = \boldsymbol{\Phi}_1 \boldsymbol{x}_{t-1} + \boldsymbol{\Phi}_2 \boldsymbol{x}_{t-2} + ... + \boldsymbol{\Phi}_p \boldsymbol{x}_{t-p} + \boldsymbol{\varepsilon}_t \tag{1}$$

where $\boldsymbol{\Phi}_j$ (j = 1 to p) are (n x n) matrices of coefficients and $\boldsymbol{\varepsilon}_t$ (t = 1 to T) is an (n x 1) vector of error terms such that E($\boldsymbol{\varepsilon}_t$) = $\mathbf{0}$, E($\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t$') = Σ and E($\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_m$') = $\mathbf{0}$ for all $t \neq m$. Another common assumption in addition to the mean and covariance matrix of the error terms is that they also follow a multivariate normal distribution i.e. $\boldsymbol{\varepsilon}_t \sim$ N($\mathbf{0}$, Σ).

The model outlined above is a purely stochastic process with zero mean. In this case, it is appropriate to also include a deterministic term that describes the time trends of the data. After adding the deterministic term the model can be written as

$$\boldsymbol{y}_t = \boldsymbol{\mu}_t + \boldsymbol{x}_t \tag{2}$$

where the deterministic term $\boldsymbol{\mu}_t$ is a linear time trend i.e. $\boldsymbol{\mu}_t = \boldsymbol{\mu}_0 + \boldsymbol{\mu}_1 t$. In the context of the motivating weight lifting example, it is this deterministic term on the larger time scale that describes the time trend that is of interest. However, it is yt that is actually the vector of observed variables. Thus $\boldsymbol{\mu}_t$, which is of greatest interest, is unobserved. The previous equation can be written as

$$\boldsymbol{y}_t = \boldsymbol{\theta}_0 + \boldsymbol{\theta}_1 t + \boldsymbol{\Phi}_1 \boldsymbol{y}_{t-1} + ... + \boldsymbol{\Phi}_p \boldsymbol{y}_{t-p} + \boldsymbol{\varepsilon}_t \tag{3}$$

where $\boldsymbol{\theta}_0 = (\mathbf{I}_n - \sum_{j=1}^{p} \boldsymbol{\Phi}_j)\boldsymbol{\mu}_0 + (\sum_{j=1}^{p} j\boldsymbol{\Phi}_j)\boldsymbol{\mu}_1$ and $\boldsymbol{\theta}_1 = (\mathbf{I}_n - \sum_{j=1}^{p} \boldsymbol{\Phi}_j)\boldsymbol{\mu}_1$. By using the model given here, the parameters will be on the scale of the observed data and will be more easily estimable. Once estimates have been found for $\boldsymbol{\theta}_t$ and each $\boldsymbol{\Phi}_j$ (j = 1 to p), the estimates for $\boldsymbol{\mu}_t$, which are of greatest interest, can be found by using

$$\boldsymbol{\mu}_0 = (\mathbf{I}_n - \sum_{j=1}^{p} \boldsymbol{\Phi}_j)^{-1}\boldsymbol{\theta}_0 - (\mathbf{I}_n - \sum_{j=1}^{p} \boldsymbol{\Phi}_j)^{-1}(\sum_{j=1}^{p} j\boldsymbol{\Phi}_j)(\mathbf{I}_n - \sum_{j=1}^{p} \boldsymbol{\Phi}_j)^{-1}\boldsymbol{\theta}_1 \quad (4)$$

$$\boldsymbol{\mu}_1 = (\mathbf{I}_n - \sum_{j=1}^{p} \boldsymbol{\Phi}_j)^{-1}\boldsymbol{\theta}_1 \quad (5)$$

3 Estimation

By rewriting the above equations using matrices, the estimates for the parameters $\boldsymbol{\Phi}_j$, $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ can be found through the use of ordinary least squares, which under the assumptions of the model will also be the maximum likelihood estimates. The distribution of these estimates will be approximately normal with mean and covariance that can be determined through the methods of multivariate linear regression.

The above equations can then be used to find estimates for $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$. By using the multivariate delta method, it can be shown that these estimates will also be approximately normal with estimable mean and covariance which will allow for statistical inference to be run on the parameters $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$.

4 Multiple Time Scales

By fitting a VAR model to the observations on the smaller time scale at each large scale time point (which in the case of the weight lifting would be the equivalent of a day) the behavior of the data at each large scale time point can be estimated. A VAR model can then be fit to these estimates on the large time scale in order to estimate the behavior, and the trend, of the data on the large scale. This can be done by using the methods of vector autoregression, least squares estimation, and the multivariate delta method as outlined previously.

Simulations are run in order to demonstrate the ability of these methods, as well as to approximate the type I error rate and power of doing hypothesis testing for the large scale time trends through the use of these methods.

References

[1] Geyer, C. (2001). *Convergence Concepts Continued* [Course Notes]. Retrieved from http://www.stat.umn.edu/

[2] Ahamada, I. (2008). *Multivariate Time Series* [Course Notes]. Retrieved from http://eurequa.univ-paris1.fr

[3] Mills, S. (2014). *Multivariate Linear Regression Models* [Course Notes]. Retrieved from http://people.math.carleton.ca

[4] Levy, D. (2011). *Numerical Differentiation* [Course Notes]. Retrieved from http://www2.math.umd.edu/

[5] Luetkepohl, H. (2011). *Vector Autoregressive Models*. European University Institute, Florence.

[6] Zivot, E., & Wang, J. (2006). *Modeling Financial Time Series with S-Plus*. Retrieved from http://faculty.washington.edu/ezivot/econ584/notes