SYMPOSIUM ON RECENT DEVELOPMENTS IN DATA ANALYSIS

J. Hamill[1], R. Van Emmerik[1], R. Miller[1], K. O'Connor[2], N. Coffey[3], D. Harrison[4]
Biomechanics Laboratory, University of Massachusetts, Amherst, MA U.S.A.[1]
Human Performance Laboratory, University of Wisconsin, Milwaukee, WI, USA[2]
Department of Statistics, University of Limerick, Limerick, Ireland[3]
Department of Physical Education and Sports Science, University of Limerick, Limerick, Ireland[4]

The purpose of this symposium is to present recent developments in biomechanical data analyses in two areas. First, current methods used in a dynamical systems approach will be described. Second, two statistical approaches, Principal Components Analysis and Functional Data Analysis, will be presented. The emphasis in this symposium will be on how to use each of these recent analysis techniques.

KEY WORDS: dynamical systems, principle components analysis, functional data analysis

INTRODUCTION:

Biomechanics has progressed significantly from both technological and analytic viewpoints. The development of new technologies has allowed biomechanists to become ever more sophisticated and ask much more complicated questions. It has also allowed biomechanists to generate large quantities of higher dimensional data. In this symposium, we will present approaches to the analysis of biomechanical data from two perspectives that are not in common use in biomechanics research. First, we will present methods of analysis that are used in a dynamical systems approach. Second, we will present two statistical methods that can be used to analyze large sets of continuous data.

DYNAMICAL SYSTEMS (Van Emmerik, Miller and Hamill):

The goal of this presentation is to present a review of current methods in the assessment of movement coordination from a dynamical systems perspective (Glass, 2001). The data analysis techniques that will be presented are essential in the assessment of stability and adaptability of movement patterns. These techniques are also important in assessing the role of movement variability in expert performance, learning/development and disease. Although traditional perspectives in biomechanics and motor control have highlighted the negative role of movement variability, dynamical and complex systems approaches have emphasized the functional role of variability in creating adaptive and stable movement patterns.

Measures of movement coordination that will be presented include relative phase and vector coding techniques (Hamill et al., 2000; Chang et al., 2008). Vector coding analysis of coordination has been primarily applied to angle-angle diagrams of lower extremity segmental or joint motions during locomotion. Relative phase techniques have traditionally been applied to bimanual coordination and lower and upper body movements during locomotion. Both continuous and discrete relative phase techniques will be discussed. Discrete relative phase (DRP) is a valuable tool to assess more complex coordination patterns containing multiple frequencies, as for example in the coordination between the respiratory and locomotor systems. Continuous relative phase (CRP) techniques are based on higher dimensional state space reconstructions (position/velocity phase plane). Issues that will be discussed in CRP analysis will include normalization of the phase plane and the use of circular statistics in the assessment of coordination patterns. Also, a comparison of the different coordination measures and their limitations will be provided.

Movement coordination is also strongly associated with the perceptual variables that may play a role in the detection of stability boundaries for upright stance or during locomotion. It is argued that a stronger emphasis on these perceptual variables is needed to assess conditions prone to postural instability. This presentation will discuss research that highlights the importance of a systematic investigation of the role of perceptual control variables such as time-to-contact (Haddad et al., 2006) as a necessary prerequisite to understand postural and gait control.

The final part will include a review of new ‘complexity’ methods to assess the structure and variability of human movement. These techniques include different measures of entropy, fractal structure and recurrence quantification analysis (RQA). Complexity analysis has been used to demonstrate changes in movement coordination and use of degrees of freedom as a function of development, disease and movement expertise.

STATISTICS:

The development of highly sophisticated data collection tools in many real-world applications (e.g. biomechanics, imaging, etc.) has resulted in the production of high dimensional data. In order to analyze these data sets, two statistical approaches have been proposed: 1) principal components analysis (Jolliffe, 2005; O’Connor and Bottum, 2009; Wrigley et al., 2006); and 2) functional data analysis (Ramsay and Dalzell, 1991; Ramsay and Silverman, 2002). In the symposium, the same data set will be analyzed to contrast the two approaches.

Principal Component Analysis (O’Connor and Hamill)*:*
In a Principal Components Analysis (PCA), the time series of each trial serves as input. PCA can be utilized to identify the dominant modes of variation within waveforms. This approach allows examinations of time series data without making a priori assumptions regarding critical dependent variables, such as maximum or minimum angles. This technique can also illuminate patterns that may not be readily obvious in examining the original waveforms. In order to facilitate data processing, each trial is time scaled to 101 data points. An *n x p* matrix is created with *n* = the number of time series and *p* = 101. The analysis is performed through an eigenvalue analysis of the covariance matrix, $S_{101 \text{ X } 101}$ which yields eigenvectors ($U_{101 \text{ X } 101}$) and eigenvalues ($L_{1 \text{ X } 101}$). The eigenvector matrix, $U_{101 \text{ X } 101}$, contains the coefficients for each of the 101 principal components (PC) that are extracted. The eigenvalue matrix, $L_{1 \text{ X } 101}$, contains the relative contribution of each PC to the total variation. An analysis is then performed to retain only those principal components that contribute modes of variation greater than an equivalently sized input matrix of randomly generated numbers. The PC scores (Z) for each of the *n* individual times series are calculated by multiplying each individual trial’s variation about the overall mean with the transpose of the eigenvector matrix:

$$Z_{\text{nx101}} = (X - (1 - x_{1\text{x}101})) \times U'$$

where $x_{1 \text{ X } 101}$ is the mean waveform of all trials. The Z scores for each retained PC can then be compared using traditional statistical techniques (e.g., gender differences).

In order to assess how well the retained PCs represent the original input data, the Q-statistic is calculated. The Q-statistic is the sum of squares of the residuals between the individual trial and the reconstructed profile based on the retained PCs. A critical Q alpha value of 0.05 is used in this analysis. Reconstructed trials with a Q statistic less than 0.05 indicate that the original data were adequately represented.

PCA also provides a unique method of investigating between-subject variability. To begin, correlations (r_{ij}) are calculated between the *i*th principal component and the *j*th time sample:

$$r_{ij} = \frac{\mathrm{U}_{\mathrm{ji}}\sqrt{\mathrm{L}_{\mathrm{i}}}}{s_j}$$

where s_j is the standard deviation at a given time in the input series. The value $\mathrm{r}_{\mathrm{ij}}^2$ is the percent explained variance across time for a given PC. Summing the variance across time provides the ability to separate the overall variation in the data into random and deterministic components. This may provide a powerful new approach to examining the nature of movement variability.

Functional Data Analysis (Coffey and Harrison):
Functional data analysis (FDA) is a statistical methodology used to analyze such data. Functional data is usually measured at a discrete number of time points but it is assumed that some underlying function $x_i(t)$ generates the observed data ($y_{i1}, \ldots\ y_{in}$) for individual i. A key idea in FDA is that the underlying function is *smooth*, i.e. pairs of adjacent values are “linked” and do not differ in value from each other by a large amount. Thus the ordering of the observed values is important. As a result, the basic idea behind FDA is to treat the entire sequence of measurements for a particular experimental unit as a single functional entity. In addition, FDA often makes use of the derivatives of the curves which greatly extends the power of FDA methods and leads to functional models as defined by differential equations (dynamical systems). It also facilitates the creation of phase plane plots which may prove highly informative about the processes generating the functions.

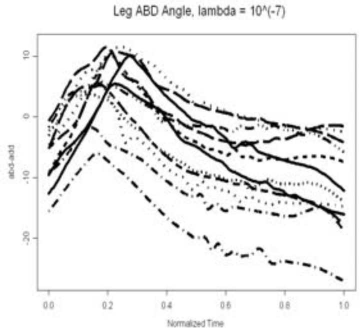
As stated above, functional data is typically measured at a finite number of time points, possibly with some measurement error. Therefore, the observed values ($y_{i1}, \ldots\ y_{in}$) can be represented as $y_{ij} = x_i(t_j) + \varepsilon_{ij}$, where $x_i(t)$ is a smooth function and ε_{ij} is measurement error. As a result, the raw data need to be converted to the smooth function $x_i(t)$. This is achieved using basis function expansions and one of several possible smoothing techniques such as regression splines, smoothing splines, etc. Regression splines estimate $x_i(t)$ as a linear combination of *K* basis functions $(\phi_1(t),\ldots,\phi_K(t))$, e.g. B-splines, Fourier splines, wavelets such that $x_i(t) = \sum_{k=1}^{K} c_{ik}\phi_k(t)$. The coefficients c_{ik} are estimated by minimizing

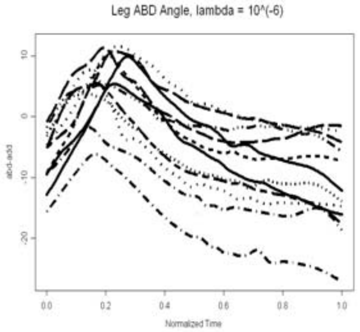
$$\sum_{i=1}^{N}\sum_{j=1}^{n}[y_{ij} - x_i(t_j)]^2 = \sum_{i=1}^{N}\sum_{j=1}^{n}[y_{ij} - \sum_{k=1}^{K} c_{ik}\phi_k(t)]^2$$

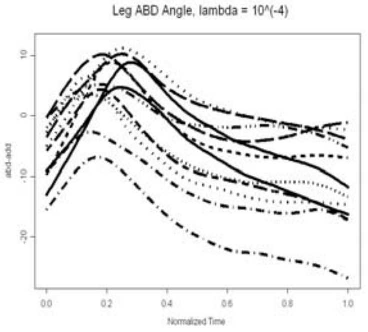
Choosing the number of basis functions is important since if *K* < *n* the data are smoothed, while if *K* = *n* the data are interpolated. This is a difficult problem and an alternative approach is to use smoothing splines which set *K* = *n* (interpolating the data) and control any over-fitting using an additional penalty term. The penalty term penalizes the curvature of the fitted function and thus curves that are more variable will be penalized more than functions that are smoother. The coefficients are now determined by minimizing

$$\sum_{i=1}^{N}\sum_{j=1}^{n}[y_{ij} - \sum_{k=1}^{K} c_{ik}\phi_k(t)]^2 + \lambda\int D^2 x_i(t)dt$$

where $D^2 x_i(t)$ is the second derivative (curvature) of $x_i(t)$ and λ is a smoothing parameter controlling the trade-off between fidelity to the data and the smoothness of the resulting fitted curve. If λ is large the fit will be smoother while if λ is small the fit will be less smooth. The figures below display the result of changing the value of λ. Smoothing splines allow the user to have more control over the amount of smoothing that is achieved and λ can be chosen using several techniques, e.g. cross-validation.







Once the raw data have been smoothed, it is possible to carry out further analyses, e.g. functional principal components analysis (FPCA), functional canonical correlation analysis, functional discriminant analysis, principal differential analysis, functional regression, etc. FPCA is an extension of multivariate principal components analysis to functional data which determines the main modes of variation in a set of *curves*. The extracted components are now functions rather than vectors, and are used to identify the characteristic features of a set of curves throughout an entire time interval. As in the multivariate case, the first few functional principal components usually account for the majority of the variation in the set of curves providing a way of looking at the variance structure which can often be more informative than a direct examination of the variance-covariance function. Functional principal component scores can also be determined for each individual and these provide a means of determining the characteristic behaviour of specific cases. They are also useful for identifying outlying observations, i.e. individuals who score very differently from the remaining individuals in a sample of curves.

REFERENCES:

Chang, R., Van Emmerik, R.E.A., & Hamill, J. (2008). Quantifying rearfoot-forefoot coordination in human walking. *Journal of Biomechanics*, 41:3101-3105

Glass, L. (2001). Synchronization and rhythmic processes in physiology. *Nature*, 410:277-284.

Haddad, J. M., Gagnon, J., Hasson, C. J., Van Emmerik, R. E. A., & Hamill, J. (2006). The use of time-to-contact measures in assessing postural stability. *Journal of Applied Biomechanics*, 22:155-161.

Hamill, J., Haddad, J.M., & McDermott, W.M. (2000). Issues in quantifying variability from a dynamical systems perspective. *Journal of Applied Biomechanics*,16:407-419.

Jolliffe, I. (2005). Principal Component Analysis. New York: John Wiley & Sons.

O’Connor, K.M., Bottum, M.C. (2009). Differences in cutting knee mechanics based on principal components analysis. *Medicine and Science in Sports and Exercise* 41:867-878.

Ramsay, J.O., Dalzell, C.J. (1991). Some tools for functional data analysis. Journal of the Royal Statistical Society 53:539-572.

Ramsay, J.O., Silverman, B.W. (2002). Applied Functional Data Analysis. New York: Springer.

Ramsay, J.O., Silverman, B.W. (2005). Functional Data Analysis. New York: Springer.

Wrigley, A.T., Albert, W.J., Deluzio, K.J., Stevenson, J.M. (2006). Principal component analysis of lifting waveforms. *Clinical Biomechanics* 21:567-578.