CLUSTERING OF MIXED DATA: SIMULTANEOUS CONSIDERATION OF KINEMATIC AND FUNCTIONAL OUTCOME MEASURES

Kim Parker[1], Tom Chau[1,2], Tim Daniels[2,3], Rhys Thomas[3]
[1]Bloorview MacMillan Children's Centre, [2]University of Toronto, [3]St. Michael's Hospital

ABSTRACT

Kinematic and functional outcome measures obtained from 28 healthy adults (controls) and 27 adults with ankle arthrodesis (patients) were incorporated into a mixed data clustering model. Selected Fourier coefficients were used to parametrize the kinematic data. Principal component analysis (PCA) reduced an original data set of 165 variables to five principal components representing over 75% of the original data variance. A fuzzy clustering algorithm separated the data into 2 clusters, with the dominating principal component (PC) primarily consisting of three functional outcome measures representative of function and pain. Kinematic Fourier coefficients had little effect on the clustering of the data. Approximately half of the patients were clustered with the controls in a high functioning group while the remaining patients were clustered in a low functioning group.

BACKGROUND

In quantitative gait analysis, many researchers have used various analytical techniques, including fuzzy clustering and neural networks, to automatically group subjects. A thorough review is provided by Chau (2001) [1] To date, subject classification has been based largely on parameterizations of temporal data such as electromyography, kinematics, and kinetics. Few techniques can readily incorporate mixed data consisting of both temporal data and static functional outcome measures, such as clinical and health questionnaire scores. Recently, Su et al. (2001) used a fuzzy cluster paradigm on sagittal plane Euler angles for the hindfoot and forefoot from a population of controls and patients with ankle arthrodesis [2]. They considered each time step in the time normalized data series as an independent variable. Clustering revealed different gait patterns that were clinically interpretable. Due to the high dimensionality of the data in this approach, an enormous data set is required to justify clustering.

There is a need to reduce the dimensionality of the data when dealing with smaller sample sized data sets, especially when considering all the lower limb joints in all the three planes of movement. Through the application of multivariate statistical techniques such as PCA, one can reduce the dimensionality of the data. The cyclic nature of kinematic data lends itself to Fourier series representation. Anywhere from 5 to 13 harmonics have been used to describe kinematic waveforms [3, 4]. Experimental methods and recursive algorithms have been used to determine the number of harmonics required to accurately represent the kinematic series [5, 6]. Through the use of both Fourier series representation and PCA, one can reduce the data's dimensionality to one that is manageable with conventional clustering techniques and that will hopefully provide additional insight into pathological gait

METHODOLOGY

Kinematic and functional outcome measures were obtained from 28 healthy adults and 27 adults with unilateral ankle arthrodesis. Euler angles were obtained for the pelvis, hip, knee, hindfoot and forefoot in the sagittal, frontal and transverse planes for the affected side of the patient population and a randomly chosen side for the control population. Kinematics of the foot were obtained using an adapted two segment foot model to describe the motion of the forefoot relative to the hindfoot and the hindfoot relative to the shank [7]. The motion of the distal segment relative to the proximal segment was obtained using Euler angles and the neutral position to correct the dynamic joint angles. A rotation sequence from the sagittal plane to the coronal and then transverse planes was used to define the forefoot and hindfoot relative angles to allow comparisons to previously published data [7]. The remainder of segment rotations were defined using an existing mathematical Euler sequence of coronal, followed by transverse and then sagittal plane rotations [8]. A six-camera Vicon system monitored the three-dimensional body segment

kinematics at a sampling rate of 60 Hz. Coordinate data were passed through a low-pass 6 Hz digital Butterworth filter prior to Euler-angle calculations. Cadence and stride lengths were calculated based on the posterior heel marker trajectories. Five cycles of gait were time normalized and averaged for each subject to obtain individual mean and standard deviation representative data.

Functional outcome measures included AOFAS ankle-hindfoot scale, MODEMS (includes SF-36), and Ankle Osteoarthritis Scale pain and disability components [9,10,11]. In some instances, data were missing from the outcome measures, in which case the entire data point was excluded from the analysis. In total, data from five patients were excluded. Missing information in the control population was replaced by the mean of the control population.

The kinematic data were represented using Fourier series. The required number of harmonics was determined by considering the mean and standard deviation of the kinematic waveform for each subject, for each joint and movement plane. The minimum number of harmonics required to reconstruct 96% of the mean waveform within one standard deviation was found. The first and last 2% of the curve were ignored due to discontinuities in the data. The highest number of harmonics required across all subjects was chosen as the representative number of coefficients. The data set consisted of the required number of Fourier coefficients for the pelvis, hip, knee, hindfoot and forefoot in the sagittal, frontal and transverse planes, mean cadence normalized by height, and mean stride length normalized by height. The functional outcome data included the SF-36 physical composite score, AOFAS total score, and the first two components of the AOS score measuring pain and disability. PCA was applied to this 165-dimensional data across 48 subjects using a singular value decomposition approach [12, 13]. The number of PCs retained for further analysis was chosen such that 75% of the original data variance was captured. Subsequent to PCA, cluster analysis was applied to the new space defined by the retained PCs. A fuzzy k-means methodology [14] was applied using a weighting exponent value of 2 and a minimal value of improvement, 10^{-5}, for the objective function similar to Su et al. (2001) [2]. The number of clusters were selected by minimizing a measure of overall cluster compactness relative to cluster centre separation [15].

RESULTS

The required number of harmonics for Fourier series representation is presented in Table 1. The original data set consisted of 165 variables, 159 of them Fourier coefficients. Examination of the data set revealed positive and negative correlation coefficients with magnitudes greater than 0.7 both within Fourier coefficients and outcome measures but not between the measures. Five PCs were retained to describe over 75% of the variance in the original data. Based on the three highest absolute factor loadings, the first PC (PC1) uncovered an inverse relationship between the AOFAS total score and outcome measures of pain and disability. PC1 can be referred to as the subjective function. The second PC (PC2) revealed an inverse relationship between mean hip flexion and mean pelvis back tilt. Principal components three to five were dominated by the mean Fourier coefficient term for the hip, knee, pelvis and forefoot in varying planes of motion.

Table 1. Number of Fourier harmonic terms required for individual kinematic waveform representation.

JOINT/BODY	SAGITTAL PLANE Flexion/extension	TRANSVERSE PLANE Internal/external rotation	FRONTAL PLANE Abduction/Adduction
PELVIS	4	5	5
HIP	4	5	5
KNEE	5	6	5
HINDFOOT	5	6	4
FOREFOOT	4	4	5

Using the aforementioned cluster validity algorithm [15], a minimum cluster number of 2 was found for the five PCs. Figure 1 shows the cluster centres and projected data for control and patient populations along PC1 and PC2, which capture 59% of the original data. Cluster analysis revealed one group of 11 patients exhibiting significant differences in subjective function values indicating lower function and greater pain. The second cluster encompassed all the controls and a subset of 10 patients.

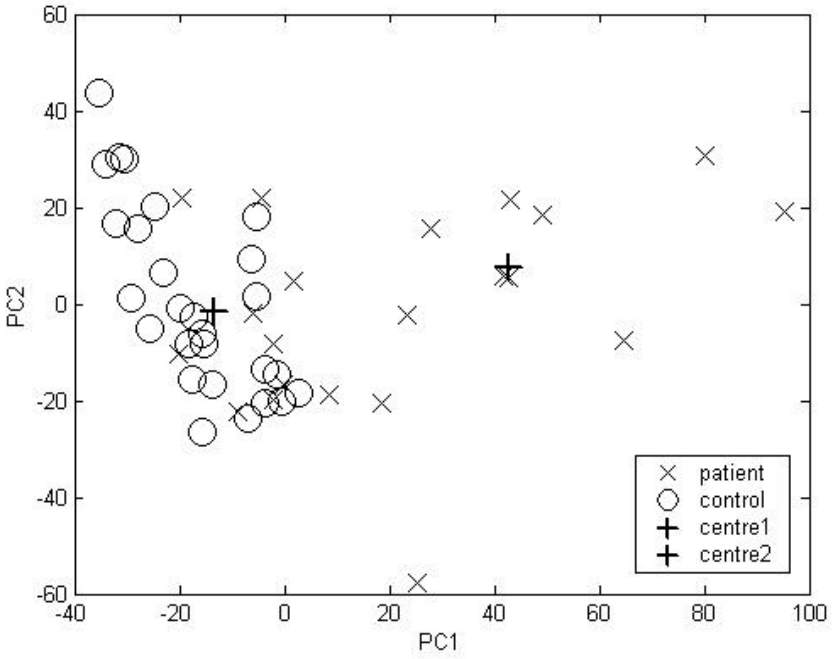

Figure 2. Patient and control population in the space defined by PC1 (function) and PC2 (pelvis and hip) along with the identified cluster centres.

DISCUSSION

PCA revealed that the majority of the variance in this data can be explained by three of the functional outcome measures and the mean joint rotations. The absence of the other Fourier harmonic terms may be due to their decreased relative contribution to the original signal. As well, variability in marker placement and individual static standing posture can lead to an increased variability in the mean rotation values. The subjective functional outcome measures were the dominating factors separating the population into a group of patients who are in pain and low functioning compared to a group of patients and controls with little pain and high function values. Although this is informative, it is surprising that some patients and controls are included in the same cluster as measurable differences were anticipated between these two groups. Perhaps the measured differences between some patients and controls were small or the chosen parameterization occluded some of the differences.

CONCLUSION

We have demonstrated the clustering of mixed data through the application of Fourier analysis and PCA. Approximately half of the patients were clustered with the controls in a high functioning group while the remaining patients were clustered in a low functioning group with pain and function outcome measures being the dominant factors. Future work will focus on other ways of incorporating kinematic waveforms such that their relative contribution to the mixed data model is strengthened.

REFERENCES

1. Chau T. A review of analytical techniques for gait data. Part1: fuzzy, statistical and fractal methods. Gait Posture 2001;13:49-66.

2. Su FC, Wu WL, Cheng YM, Chou YL. Fuzzy clustering of gait patterns of patients after ankle arthrodesis based on kinematic parameters. Med EngPhys. 2001;23:83-90.
3. Witra RW, Golbranson FL. A technique for the display of joint movement deviations. Bulletin of Prosthetics Research 1980; 17: 73-9.
4. Sutherland DH, Kaufman KR, Campbell K, Ambrosini D, Wyatt M. Clinical use of prediction regions for motion analysis. Dev Med Child Neur.1996;38:773-81.
5. Sutherland DH, Olshen R, Cooper L, Woo SLY. The development of mature gait. J Bone Joint Surg Am. 1980;62-A:336-53.
6. Schneider E, Chao EY. Fourier analysis of ground reaction forces in normals and patients with knee joint disease. J Biomech 1983; 16:591-601.
7. Wu WL, Su FC, Cheng YM, Huang PJ, Chou YL, Chou CK. Gait analysis after ankle arthrodesis. Gait Posture 2000;11:54-61
8. Apkarian J., Naumann, S., and Cairns, B. A three-dimensional kinematic and dynamic model fo the lower limb. J Biomech. 1989; 22(2): 143-55
9. Kitaoka HB, Alexander IJ, Adelaar RS, Nunley JA, Myerson MS, Sanders M. Clinical rating systems for the ankle-hindfoot, midfoot, hallux and lesser toes. Foot Ankle Int. 1994; 15:349-53
10. Saltzman CL, Mueller C, Zwlor-Maron K, Hoffman RD. A primer on lower extremity outcome measurement instruments. Iowa Orthop J. 1998; 18:101-11.
11. Domsic RT, Saltzman CL. Ankle osteoarthritis scale. Foot Ankle Int. 1998;19:466-71.
12. Ripley BD. Pattern Recognition and Neural Networks. Cambridge, Great Britain: Cambridge University Press 1996
13. Manly BFJ. Multivariate Statistical Methods: A Primer. London, UK: Chapman & Hall 1994
14. Bezdek JC. Cluster validity with fuzzy sets. J Cybernet 1974; 3:58-73.
15. Xie XL, Beni G. A validity measure for fuzzy clustering. IEEE Trans. Pattern Anal. Machine Intell. 1991;13:841-47.