Gait & Posture 51 (2017) 153–158

Contents lists available at ScienceDirect

Gait & Posture

journal homepage: www.elsevier.com/locate/gaitpost

Full length article

Correcting waveform bias using principal component analysis: Applications in multicentre motion analysis studies

Allison L. Clouthier[a,*], Eric R. Bohm[b], John F. Rudan[c], Barbara L. Shay[d], Michael J. Rainbow[a], Kevin J. Deluzio[a]

[a] Department of Mechanical and Materials Engineering, Queen's University, 130 Stuart St., Kingston, ON, Canada
[b] Department of Surgery, University of Manitoba, 820 Sherbrook St., Winnipeg, Manitoba, Canada
[c] Department of Surgery, Queen's University, 76 Stuart St., Kingston, ON, Canada
[d] Department of Physical Therapy, University of Manitoba, 771 McDermot Ave., Winnipeg, Manitoba, Canada

ARTICLE INFO

Article history:
Received 26 May 2016
Received in revised form 6 September 2016
Accepted 11 October 2016

Keywords:
Principal component analysis
Motion analysis
Gait
Waveform data
Inter-laboratory

ABSTRACT

Multicentre studies are rare in three dimensional motion analyses due to challenges associated with combining waveform data from different centres. Principal component analysis (PCA) is a statistical technique that can be used to quantify variability in waveform data and identify group differences. A correction technique based on PCA is proposed that can be used in post processing to remove nuisance variation introduced by the differences between centres. Using this technique, the waveform bias that exists between the two datasets is corrected such that the means agree. No information is lost in the individual datasets, but the overall variability in the combined data is reduced. The correction is demonstrated on gait kinematics with synthesized crosstalk and on gait data from knee arthroplasty patients collected in two centres. The induced crosstalk was successfully removed from the knee joint angle data. In the second example, the removal of the nuisance variation due to the multicentre data collection allowed significant differences in implant type to be identified. This PCA-based technique can be used to correct for differences between waveform datasets in post processing and has the potential to enable multicentre motion analysis studies.

1. Introduction

The kinematic and kinetic waveform data obtained in three dimensional motion analysis can provide insight into research questions regarding human mobility. Thus far, motion analysis has generally been limited to single centre studies. However, multicentre studies would allow sample sizes to be much larger, which may provide insight into the functional outcomes of infrequently performed procedures as well as the pathomechanics of rare conditions. Such studies have the potential to accelerate research in some areas while providing additional power to other studies through increased sample sizes.

Multicentre motion analysis studies are not typically performed because of the so-called laboratory effect. Discrepancies arise among data collected on a homogeneous cohort in different centres due to differences in laboratory environment, software, hardware, protocol, and data analysis techniques. This nuisance variation introduced by the laboratory effect often results in a bias between the waveforms of two centres. Inconsistencies in locating anatomical landmarks and marker placement contribute to a large portion of this variation [1] and, therefore, rigorous standardization of protocol has been suggested to mitigate the laboratory effect [2]. However, even when the same protocol is used in the same laboratory, inter-assessor reliability is not ideal [3,4]. Synthesis of pre-existing data collected using different methods is also precluded.

Many studies have attempted to quantify the reliability of gait data between sessions, assessors, protocols, and laboratories. Many of these have used correlation coefficients to quantify reliability [3,5,6]; however, such statistics are difficult to interpret clinically and it is unclear for which values the data may be deemed 'reliable enough' [7]. Additionally, some coefficients, such as the coefficient of multiple correlation (CMC), are influenced by the range of motion of the joint. Other methods have been implemented which address most of these concerns such as the

* Corresponding author at: 130 Stuart Street, McLaughlin Hall, Queen's University, Kingston, ON, K7L 3N6, Canada.
E-mail addresses: allison.clouthier@queensu.ca (A.L. Clouthier), ebohm@cjrg.ca (E.R. Bohm), rudanj@kgh.kari.net (J.F. Rudan), barbara.shay@umanitoba.ca (B.L. Shay), michael.rainbow@queensu.ca (M.J. Rainbow), kevin.deluzio@queensu.ca (K.J. Deluzio).

http://dx.doi.org/10.1016/j.gaitpost.2016.10.008
0966-6362/

standard deviation [2], mean absolute difference [8], or functional limits of agreement [9]. While these may quantify reliability in a more meaningful way, they do not model what exactly the differences are, and therefore, may not be easily extended to correct the differences that exist. A method has been proposed to assess the reliability of marker placement using a generalized Procrustes analysis. Briefly, the method measures how well the markers are placed on a participant compared to a large database of previous marker placements performed by an expert [10]. Although this could be useful as a training tool, it requires that the same marker set be used and also cannot be used to correct for errors in post-processing.

Principal component analysis (PCA) is a statistical technique that has been widely used in motion analysis [11–13]. One benefit of PCA is that the entire gait waveform is analyzed so temporal information is considered, as opposed to analyzing selected discrete points. This has made PCA a valuable tool in motion analysis and it has been used to identify and quantify waveform features that differ between groups and conditions as a result of pathologies [14,15], surgeries [16,17], and other factors affecting human motion [18,19]. Similarly, PCA could easily be applied to the issue of inter-laboratory motion analyses to identify whether differences exist in data collected from different centres. However, since this statistical technique also produces features that quantify waveform bias, it is possible to extend PCA to calculate a correction for the identified differences. If the populations of two centres represent a homogeneous cohort, then this method can be used to correct the waveform bias that occurs as a result of the laboratory effect. Therefore, the purpose of this study is to define and evaluate a PCA-based method that corrects differences between data collected in different centres on a homogeneous cohort.

2. Methods

2.1. Correction technique

Principal component analysis (PCA) is used to identify and remove differences between datasets. With PCA, features are extracted from a set of waveform data and every individual waveform receives a score for each feature. Following this, scores for different groups can be compared to identify waveform features for which the groups differ significantly. Because the entire waveform is considered, these features include both timing and magnitude information.

PCA is first applied to the combined waveform data from the two centres to statistically quantify the existing differences [20]. The p waveforms $\boldsymbol{X} = \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$ can be converted into p uncorrelated principal components (PCs) $\boldsymbol{z} = z_1, z_2, \ldots, z_p$ through an orthogonal transformation:

$$\boldsymbol{z} = \boldsymbol{U}^t(\boldsymbol{X} - \overline{\boldsymbol{x}})$$

where $\boldsymbol{U} = \boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_p$ are the eigenvectors of the covariance matrix of $(\boldsymbol{X} - \overline{\boldsymbol{x}})$, or loading vectors, and $\overline{\boldsymbol{x}}$ is the mean waveform. The loading vectors, $\boldsymbol{u}_i$, form an orthogonal basis for the waveform dataset and represent features contained in the dataset. The PC scores contained in $\boldsymbol{z}_i$ measure the degree to which the shape of an individual waveform corresponds to the feature represented by the loading vector for that PC, $\boldsymbol{u}_i$. It is possible to identify principal components for which the two data sets differ by performing a t-test on the PC scores. If the data originated from a homogeneous population, no differences should exist in PC scores between the two groups. Any differences that do exist can be termed laboratory effect. This effect may result from any number of sources, but if it is not related to the variable of interest, it is essentially a nuisance factor. It is then possible to correct for this waveform bias in the data collected in centre B, for example, such that centres A and B will have the same mean PC scores on all principal components. This is accomplished by adding the difference in mean PC scores between centres multiplied by the loading vector to each waveform from centre B for each principal component where the two centres differ significantly:

$$\boldsymbol{x}_{corr} = \boldsymbol{x} + \sum_{i \in S} (\overline{z}_{Ai} - \overline{z}_{Bi})\boldsymbol{u}_i$$

where $\boldsymbol{x}$ is an individual waveform, $\overline{z}_{Ai}$ and $\overline{z}_{Bi}$ are the mean PC scores for the ith principal component for centres A and B, and $\boldsymbol{u}_i$ is the loading vector for the ith principal component. The summation is carried out over the principal components deemed to differ significantly between the two centres (S). The summation term creates a single correction waveform that is added to each individual waveform from centre B. Since the majority of the variation is explained by the first few principal components [11], it is sufficient to consider only these first few PCs for inclusion in the set S. In this case, a 90% trace criterion was used where only the first k PCs are retained that together explain at least 90% of the variance in the data [21].

For the applications demonstrated here, PCA and the correction technique were applied separately to each plane of the joint angles and moments. PCA and the correction procedure were performed in MATLAB (The Mathworks Inc., Natick, MA). Here, the mean PC scores from each centre were used to calculate the correction; however, if the dataset contains outliers that may influence the mean, it may be advisable to use a trimmed mean or other variant when calculating $\overline{z}_{Ai}$ and $\overline{z}_{Bi}$

2.2. Application on simulated crosstalk error

To ensure that the correction is reasonable, the technique was first applied to a dataset with a known induced error. Knee kinematics were calculated for 24 healthy participants (15 female, age 39.5 ± 16.4 years, weight 71.3 ± 10 kg) with no history of mobility impairment or injury walking overground at the Human Mobility Research Laboratory in Kingston, ON. Each participant gave informed written consent and the study was approved by the Research Board for Health Sciences at Queen's University. Segment coordinate systems were created according to the model used by the Istituto Ortopedico Rizzoli [22]. Crosstalk was then induced by rotating the thigh coordinate system about its long axis by −5, 5, and 10°, thus creating three additional sets of knee angles for each participant. This simulates the error induced by differences in marker placement and landmark misidentification [23]. The resulting biases were corrected using the proposed technique. Principal components for which the datasets differed significantly ($p < 0.05$) were identified using Student's t-tests on the PC scores. No knowledge of the source of the induced error was used in the correction process. The calculated correction was compared with the induced error for each participant to ensure that the correction was reasonable. The induced error was calculated based on the angle of rotation about the long axis for each participant [24]. The changes in data variability were calculated using the trace of the covariance matrix, which is equivalent to the sum of the variance at each time point in the waveform data.

2.3. Application to multicentre data

Since the intended purpose of this technique is to enable multicentre studies, it was also demonstrated using a separate dataset of knee kinematics and kinetics collected in two different centres on similar populations. In both laboratories, participants had received total knee arthroplasty to treat severe knee osteoarthritis. All participants were randomly assigned to receive