Measuring Abnormality in High Dimensional Spaces: Applications in Biomechanical Gait Analysis

Michael Marks\*, Improvement Path Systems, Rochester, MI

Trevor Kingsbury, Naval Medical Center San Diego, San Diego, CA

Richard Bryant, Improvement Path Systems, Rochester, MI

John David Collins, University of Delaware, Newark, DE

Marilynn Wyatt, Naval Medical Center San Diego, San Diego, CA

Using high-dimensional datasets to measure a subject’s overall level of abnormality as compared to a reference population is often needed in outcomes research. Utilizing applications in instrumented gait analysis, this article demonstrates how using data that is inherently non-independent to measure overall abnormality can cause biased results. The authors then put forth a methodology to address this bias to accurately measure overall abnormality in high dimensional spaces. This methodology is in line with others found in the literature: the general Mahalanobis distance and the gait normalcy index proposed by Schutte et al. (2000). However, it differs in two major ways: it can be applied to datasets in which the number of observations is less than the number of features/variables, and it can be abstracted to any number of domains or dimensions. After applying the proposed methodology to the original data, the researcher is left with a set of uncorrelated variables (i.e. principal components) with which overall abnormality can be measured without bias. To best the meet the needs of various researchers, the authors discuss different considerations when deciding the appropriate number of these principal components to keep and the aggregate distance measure to utilize (i.e. Euclidean, Mean Absolute Deviation, etc.). This methodology is made freely available via an R package.

# Introduction

Instrumented gait analysis has been widely used for a variety of pediatric and adult pathologies as a means of either quantifying a functional movement deficit or evaluating improvements due to rehabilitation treatments1. A typical gait data collection yields thousands of unique measurement dimensions (i.e. variables, features, etc.) that quantify joint position and force production. Measuring a patient’s overall level of abnormality as compared to a reference population across these many dimensions is required to empower both clinical decision making and outcomes research.

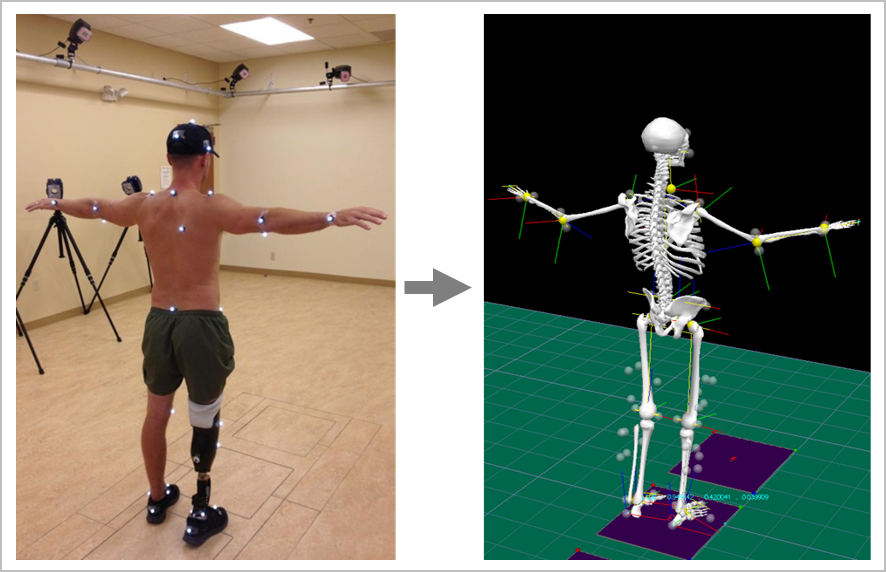


Figure 1: An example of a patient preparing for instrumented gait analysis

Measuring a patient’s overall level of abnormality is typically done by comparing a patient’s gait data to a reference population data-set of able-bodied controls. When this comparison is done in a single measurement dimension, simple distance measures (e.g. Euclidean, Manhattan, etc.) provide unbiased results. However, typically collected gait data consists of thousands of non-independent measurement dimensions. If this dependency structure is not properly addressed, using standard distance measures to define overall abnormality can produce biased results.

The Mahalanobis distance measure2 attempts to address this bias, but it cannot be calculated when the number of observations () is less the number of measurement dimensions ()3,4. This is problematic since for many high-dimensional data-sets, especially those found in instrumented gait analysis. Schutte et al.1 also attempted to address this bias for their gait normalcy index; however, this measure was made up of only 16 univariate parameters that provide an incomplete picture of overall gait5.

Attempting to address this incompleteness, two new gait normalcy indexes were created by Schwartz and Rozumalski5 and Baker et al.6 that measured normalcy across 459 dimensions during a gait cycle. However, these two methodologies use standard distance measures and fail to address the dependency structure of the underlying data. Despite this, the results from these methodologies have become standard gait summary measures reported in various populations7,8,9,10,11,12,13.

Having an overall gait abnormality measure that takes into account all the dimensions of human gait, while still accounting for the dependent nature of all those dimensions, would ensure proper conclusions are drawn in clinical decision making and outcomes research. Therefore, the purpose of this study is to: 1) Demonstrate the inherent non-independent nature of data produced in instrumented gait analysis, 2) illustrate how this dependency structure can bias measures of overall abnormality, and 3) put forth a methodology to accurately measure overall abnormality in high dimensional spaces.

# Results

## The Inherent Non-Independent Nature of Gait Data

Human gait is a complex movement that consists of both open and closed kinematic chain movements. While motion is typically analyzed at the joint level, movement of one joint can result in changes at other joints. Thus, gait elements are non-independent and the data representing it will not be either. To demonstrate this non-independence, gait data were collected for 32 able-bodied males and assembled into a 32 x 459 matrix (See Methods section for collection methodology). These are the same 459 dimensions (9 kinematic joint angles x 51 points each) used by Schwartz and Rozumalski5 and Baker et al.6 for their overall gait abnormality measures (Figure 2).

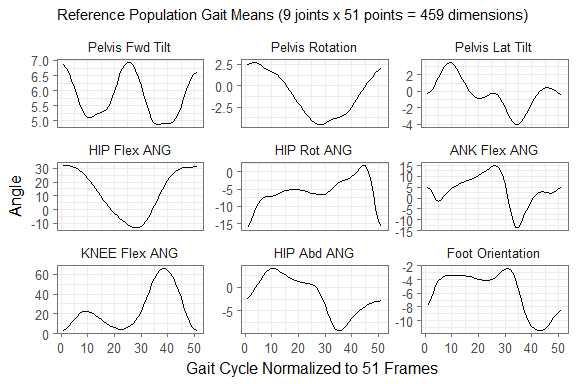


Figure 2: Reference population (n=32) means for nine joint angles across the gait cycle normalized to 51 points. These 459 total dimensions are the same as those used by Schwartz and Rozumalski5 and Baker et al.6 for their overall gait abnormality measures. This data was aggregated into a 32 x 459 matrix for the multicollinearity analysis.

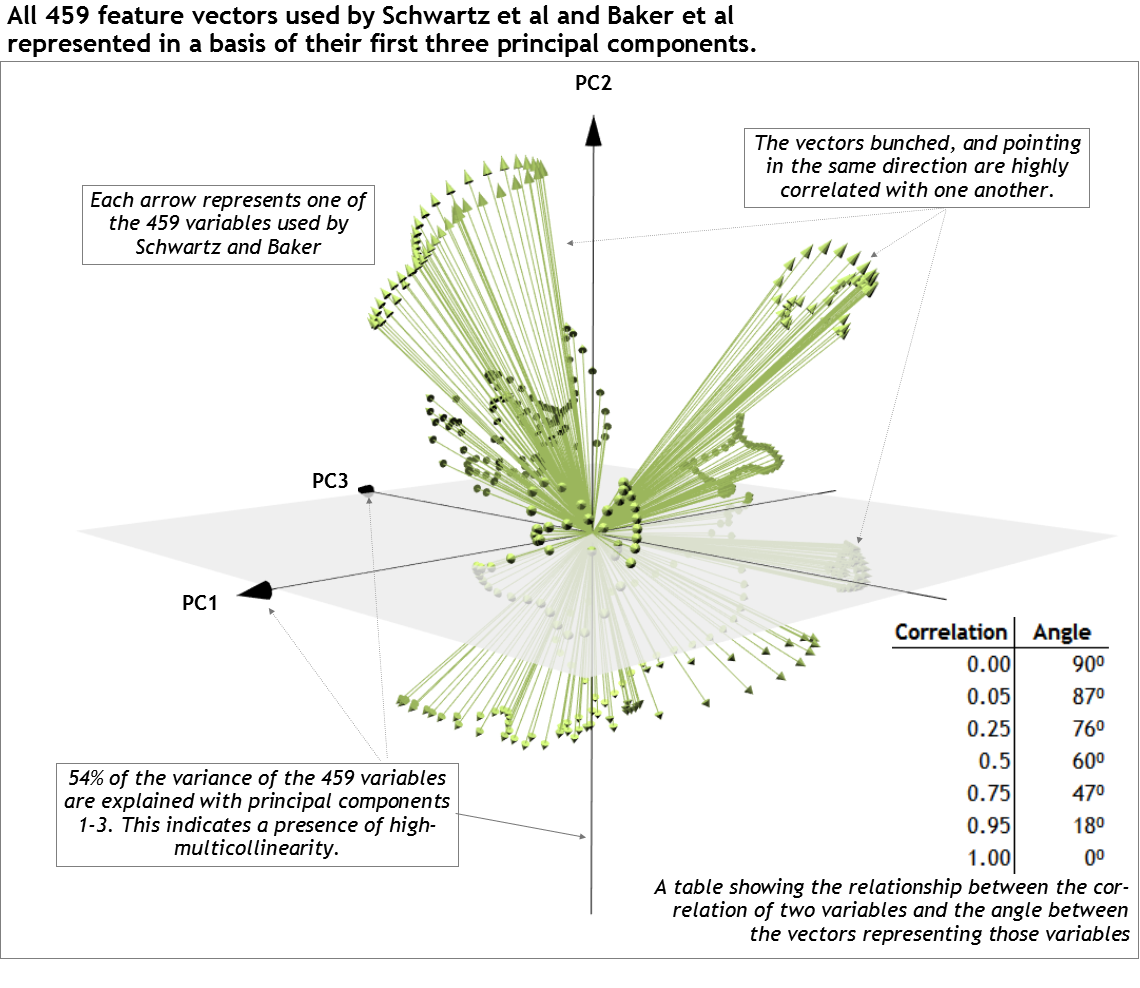


Figure 3: All 459 feature vectors of the reference gait data matrix projected onto their first three eigenvectors (principal components). The non-random distribution of the vectors in this space indicate a strong dependency among the gait features.

By projecting all 459 feature vectors of this reference matrix onto their first three eigenvectors (principal components), the non-independent nature of the gait data can be seen visually (Figure 3)14. There is a clear, non-random pattern to the vectors, demonstrating the dependent nature of the underlying variables they represent.

## The Non-Independence Bias Problem

This dependency structure can bias measures of overall abnormality if not properly addressed. This bias is most easily demonstrated by visually exploring overall abnormality measurements of a subject against a reference population calculated with non-independent variables in two-dimensional space.

The four graphs in Figure 4 show the values of two correlated variables ( and , ) for a reference population (blue dots, ) and two different subjects (orange), and . On the left, the variables are in their scaled and centered form (). We will refer to this basis as the standard basis . On the right, and are transformed into a basis of their orthonormal principal component vectors utilizing the methodology laid out in the methods section. We will refer to this basis as the principal component basis . Combinations of and , and and , that are within two standard deviations of the mean of the reference population will be inside the blue ellipse. Since the data have been centered and scaled (), the mean reference subject lies at the origin for both graphs.

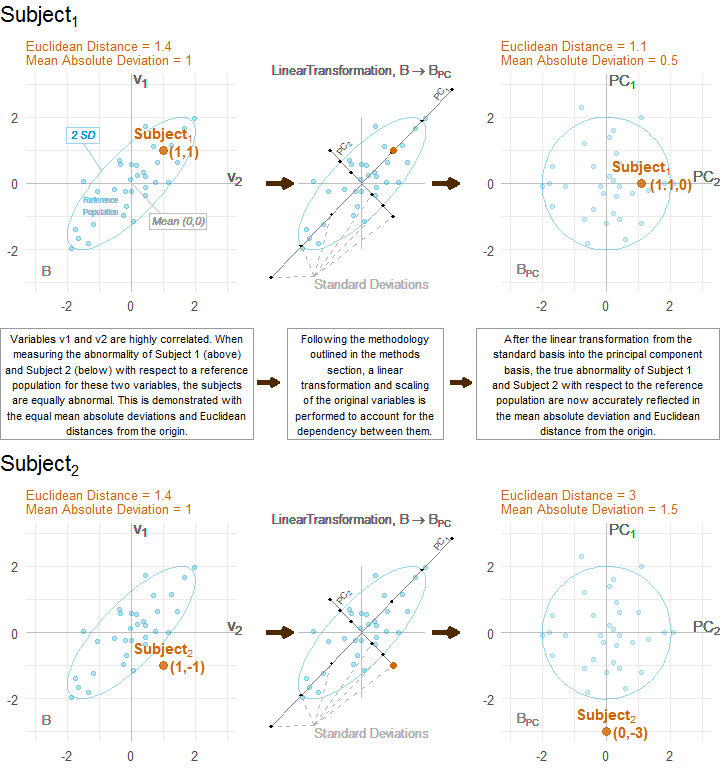


Figure 4: A two-dimensional example of the non-independence bias problem when measuring overall abnormality with two correlated variables (r=0.83) and how that bias is addressed with the proposed methodology.

The Euclidean distance between the subject and the origin represents the level of abnormality. When comparing the Euclidean distances of and , the issues with measuring normalcy in the standard basis, , can be seen. In (Figure 4), both and are units from the origin, indicating equal abnormality. However, due to the strong positive correlation between and , ’s *combination* of the two variable is more abnormal than ’s. This is demonstrated by comparing locations in relation to the blue ellipse representing two standard deviations from the origin. This interesting example shows that not addressing multicollinearity can lead to biased results and how the proposed methodology negates this bias.

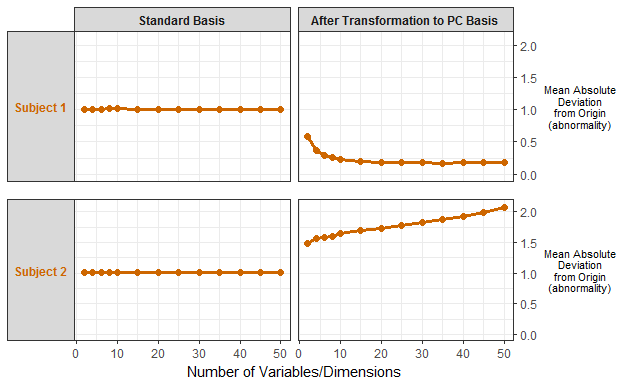


Figure 5: The results of a simulated extension of the example from Figure 4 into higher dimensional spaces. For each simulation iteration: a reference population () was generated with all variables having and correlated at r = 0.75; two subjects were then generated in which Subject 1 had all values equal 1 and Subject 2 had a random distribution of values equal 1 and values equal -1; the mean absolute deviation was then taken between the simulated reference population and subjects in both the standard basis, and the principal component basis (following the methodology outlined and utilizing all principal components). This procedure was done 500 times and averaged for each value of p. The results show that the bias when measuring overall abnormality can actually increase as the number of dimensions increase.

This example can be extended into higher dimensions as well. Figure 5 shows a simulated extension of the Figure 4 example into higher dimensional spaces (Methods - Simulated Example of Bias in Higher Dimensional Spaces). This simulated example shows that the bias seen in Figure 4 actually gets larger as the number of dimensions increase. When using a Euclidean distance or mean absolute deviation (MAD) in the standard basis, Subject 1 (all values of 1) and Subject 2 (half values of 1 and half values of -1) share the same level of normalcy in any number of dimensions. This is despite a reference population whose features are all positively correlated (r=0.75). This strong positive correlation makes Subject 2’s values more and more abnormal as the number of dimensions increase. This expected level of abnormality is reflected when using a mean absolute deviation in the principal component (PC) space, thus demonstrating the effectiveness of this method.

As seen by the examples in Figure 4 and Figure 5, not addressing the non-independent nature of the underlying data in a measure of overall abnormality can lead to biased results. This bias only gets worse as the correlation between the underlying variables increases (Figure 6 - Appendix 1). However, as seen by these results, utilizing the methodology laid out to transform the original variables into a basis of their orthonormal principal component vectors negates this bias.

# Discussion

The purpose of this study was to demonstrate the inherent non-independent nature of data produced in instrumented gait analysis, illustrate how this non-independence can bias measures of overall abnormality, and put forth a methodology to produce a new coordinate system with which to accurately measure overall abnormality in high dimensional spaces. In addition to the new coordinate system, the number of principal components to retain and the proper distance measure to utilize are important discussion topics. The two applications depend on the specific clinical or research question but have several validated options available.

##### Determining the Number of Principal Components

Selecting the right number of principal components is an extremely well researched topic (Jackson 199315, Peres-Neto et al.16, and Ferre 199517 provide good surveys). The extensive research done on this problem speaks to its importance: if the number of axes is not correctly estimated, one can introduce noise (overestimation) or loss of information (underestimation) in the analysis18.

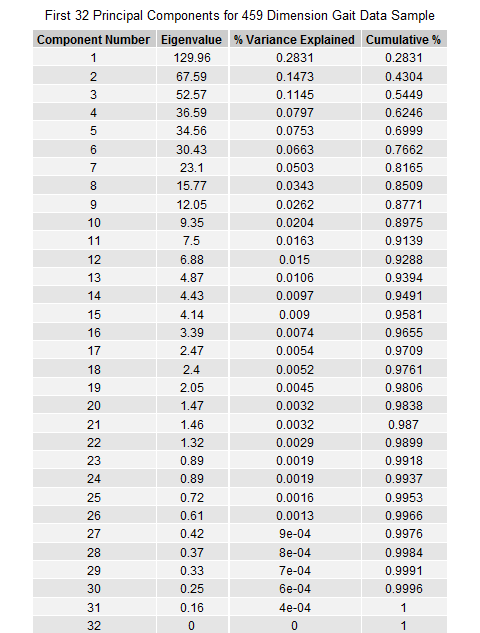


Table 1: A breakdown of the first 32 Principal Components for the 459 Dimension Gait Data Sample

In the case of this study’s gait data (and many other high dimensional datasets), there are fewer observations () than measurement dimensions (). If is smaller than , the data cannot occupy more than an -dimensional subspace of the -dimensional space. Therefore, projection into a lower-dimensional space does not necessarily lose information. If the data really are a lower-dimensional manifold in the high dimensional space, then a projection can preserve its structure exactly19. This is the case with our example gait data-set as illustrated by the first 32 principal components accounting for 100% of the total variation in the data (Table 1).

One could choose to retain all these principal components for measuring overall abnormality. This is what was done by Schutte et al.1 for their gait normalcy index; however, the data used in that study only had 16 dimensions. Keeping all the principal components removes all risk that information relevant to a subject’s abnormality is lost. However, including principal components with small eigenvalues may introduce unnecessary noise (e.g. measurement error) that could bias results. Furthermore, utilizing fewer principal components may be beneficial when making clinical interpretations since one could determine what data are represented in each PC and determine the most meaningful data that is being evaluated.

Deciding how many of those principal components (in this case 32) to keep is more of a subjective art than a perfect science. Figure 8 shows a scree plot with the results of some common methodologies used to determine the appropriate number of principal components when applied to the sample gait data. Percent total variance explained (% TVE) is a common, yet relatively arbitrary, method; three different cutoff values are presented (90%, 95%, and 99%). Additionally, the results of the broken stick20, Kaiser-Guttman21,22, and parallel23 methods are included as well24.

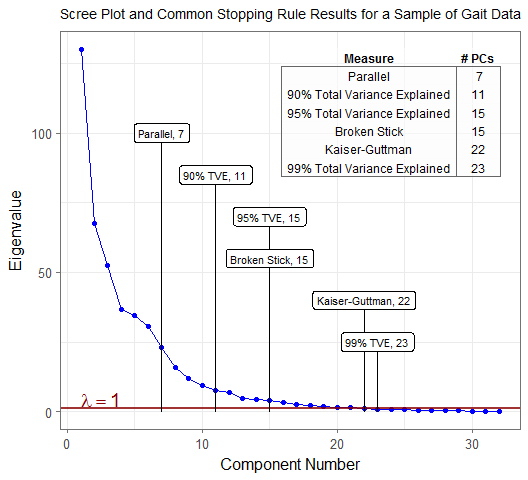


Figure 8: A scree plot with the results of some common methodologies used to determine the appropriate number of principal components when applied to the sample gait data

For the purposes of measuring abnormality in high dimensional gait analysis, it is likely that removing principal components with eigenvalues 1 (Kaiser-Guttman) strikes a good balance between the risks of unnecessary noise and information loss. An eigenvalue greater than one indicates that its corresponding principal component contains more information than any single original variable15. If an eigenvalue equals one for a particular principal component, that variability could be associated exclusively with one single original variable that is orthogonal to the rest of the data. While this scenario is highly unlikely, one would probably not want to exclude a variable which exhibits these properties for the purposes of abnormality measurement. Keeping all principal components with an eigenvalue 1 would ensure this does not happen.

An additional consideration is that for very high dimensional datasets with p much larger than n, the eigenvalues from the data are not consistent estimators of the population eigenvalues25. Therefore, caution should be exercised before drawing conclusions from the eigenvalues or TVEs from the sample. A possible approach in this circumstance is to use bootstrapping. This allows one to create a confidence interval either of the true eigenvalue corresponding to each principal component, or the percentage of variance explained by the first k PCs, through resampling the data. This confidence interval can be used to test, at a specified confidence level, that an eigenvalue < 1 or TVE is 90%. Therefore, the first PC that satisfies this test can be used as the cutoff point for PCs to retain.

For purely clinical applications, interest lies in utilizing a summary measure to enable specific interventions. To do this, some knowledge of why a subject is abnormal is necessary. Assessing the loading significance of the original variables on the resultant principal components26 can enable a clinician to see where a patient is most abnormal. This task becomes more complex with more principal components, so using fewer PCs would enable some clinical interpretability. As a result, even though a lower amount of variance would be explained, the clinical meaning would be improved with fewer principal components.

##### Selecting a Distance Measure

Measuring distances and the properties of different distance measures in high-dimensional spaces is a well-studied topic for the purposes of outlier detection in computer science27,28,29,30,31,32. However, computer scientists are typically more concerned with how points relate to one another in terms of distance, instead of normalcy which would simply be distance from average (which would be the origin if the variables are standardized). Further research is required to truly understand the implications of measuring distance from the origin in these higher dimensional spaces.

Established measures of overall gait abnormality use various distance measures in their methodologies. Schutte et al.1 uses the square of a Euclidean distance. Schwartz and Rozumalski5 use the natural logarithm of a Euclidean distance as a raw score, and utilize a scaled version () of the raw score for clinical interpretation purposes. Baker et al.6 utilize a simple root mean squared (RMS) difference.

Ultimately, the metric used to measure distance should be chosen based on how it will be used. If it is for statistical tests, one would want a measure whose output has good mathematical properties (e.g. normally distributed). This was the rational for Schwartz and Rozumalski5 using the natural logarithm of a Euclidean distance. If it is for interpretation purposes, one would want something the user of the data will understand (The rationale for Schwartz and Rozumalski’s scaled version5).

Any number of distance measures (e.g. Euclidean, MAD, RMS, Manhattan, etc.) all have their merit; however, in an ideal world, usability testing would be done to establish a unit of measurement that is most valuable to its users. For example, if the users are familiar with the idea of a standard deviation, a mean absolute deviation (MAD) of the standardized principal components could be used. This would improve clinical interpretability and likely enhance utility in a multidisciplinary setting.

# Conclusion

Biomechanical gait analysis is a powerful tool for collecting large amounts of outcomes data. Aggregating all this data into a single measure of abnormality greatly aids in clinical decision making and outcomes research. As a result, previous attempts at creating such measures have become widely used in many academic and clinical circles5,6. These measures have been a big step forward for the gait community; however, as this study has demonstrated, these measures have bias because they fail to address the inherent dependent nature of gait data.

This study has given the researcher a methodology to address these dependency issues when creating overall abnormality measures. These methods are not exclusive to measuring abnormality in gait and could be applied to other high-dimensional, multicollinear data-sets. Given each researcher/clinician has their own needs when using or creating outcome measures, different considerations can be utilized that will affect the resultant measure. Application of these methods and considerations will empower researchers to create useful measures of overall abnormality in high dimensional spaces.

# Methods

## Subject Data Collection

Thrity-two able-bodied males, with no history of orthopedic injuries or surgeries that would affect gait, were studied while walking at their self-selected speed. The average age of these subjects was 30 ± 6 years, and their average BMI was 27 ± 2 kg/m2. All subjects were studied using a 34-marker modified Helen Hayes marker set and data were collected using a 12-camera Motion Analysis Eagle system (Motion Analysis Corporation, Santa Rosa, CA). This marker set allowed for nine joint angles (pelvic and hip angles in all three planes, knee flex/extension, ankle dorsi/plantarflexion, and foot progression) to be calculated at 2% increments throughout the entire gait cycle of 100%, giving 51 data points per joint angle. The analysis of the volunteer’s left side only resulted in 459 total measurement dimensions for each subject with which to measure normalcy (9 angles x 51 points each = 459 dimensions) (Figure 2). To mirror the data used by Schwartz and Rozumalski5 and Baker et al.6, the exported data on these subjects were aggregated into a 32 x 459 matrix for the multicollinearity analysis.

## Methodology to Address the Non-Independence Bias Problem

To address the multicollinearity bias problem to accurately measure normalcy in dimensions:

Let be the matrix representing the reference population.

The rows of the matrix represent subjects in the reference population; the columns represent features of the subjects in the reference population.

Let be the vector of those same features for the subject with whom we want to compare to the reference population.

To get these two objects into the principal component basis, , we must first scale and center both based on and where:

and the th element of is:

And where:

where the th element of is:

To get the scaled and centered version of (We will call this ), is subtracted from all rows of , all rows are then divided by . The same process is done on to convert it to :

To determine the change of basis matrix, let represent the change of basis matrix from basis to basis . which is simply composed of the eigenvectors of the covariance matrix of . Since the covariance matrix is symmetric, its eigenvectors are orthogonal, thus using a matrix of these eigenvectors as a change of basis matrix results in a rotation of the original data. A change of location (or translation) by scaling and centering followed by a rotation does not alter the intrinsic statistical properties of the data33. The overall goal of this transformation is to create a new set of independent variables with which to measure the distinct properties of the reference population and how a subject differs from that population.

We will refer to the results of the projection of and onto the eigenvectors of as, and respectively.

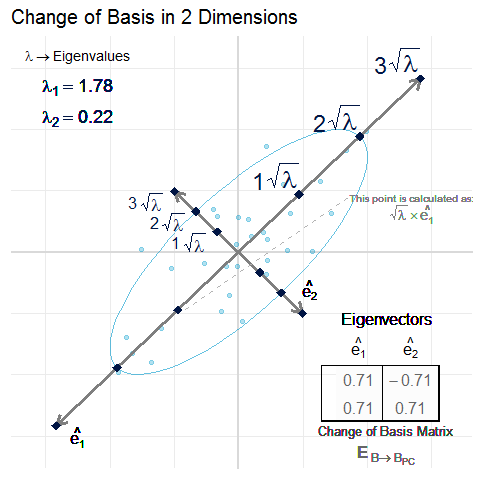


Figure 7: A two-dimensional example of the scaling that is done to ensure that each of the new independent variables has equal variance. This is done by dividing each new variable by the square root of its corresponding eigenvalue. The square root of the eigenvalue is the standard deviation of its corresponding eigenvector.

The square root of the eigenvalue is the standard deviation of its corresponding eigenvector, so each new independent variable is divided by the square root of its corresponding eigenvalue to ensure equal variance (Figure 7). According to Schutte et al.1, scaling the new variables this way accounts for the magnitude of variation inherent in certain variables. In other words, if one of the original variables (or some linear combination of the variables) has a large variation within the reference population, then a large deviation from the average value of that variable will not count excessively against the ‘normalcy’ of a subject. It could be argued that the eigenvectors associated with small eigenvalues represent variable combinations that may be small random fluctuations and should not be magnified through division by their eigenvalue1. This is a valid concern which could be addressed by removing principal components with small eigenvalues from the analysis altogether. This is a common practice and its merits are considered in the discussion section.

The re-scaled versions of and will be referred to as and . It is with respect to these variables that normalcy can be accurately measured.

These new variables are a different representation of the original points defining a subject in space. The original points have not changed their location in space, but only the construct in which they are expressed has changed.

Since and have been scaled and centered, the mean for each variable lies at the origin (i.e. ) of the dimensional space in which they occupy. The distance from the origin can therefore be thought of as a level of abnormality. The further away a subject is from the origin, the more abnormal they are.

## Simulated Example of Bias in Higher Dimensional Spaces

Let be the matrix representing the covariance matrix used to generate a reference population () from a multivariate normal distribution34 with .

For each simulation iteration, a reference population, , is generated from . A subject is then generated as either or with a random distribution of values equal to 1 and values equal to -1.

The Euclidean distance and mean absolute deviation are taken between and , and and . These distance measures are taken in both the standard basis, and the principal component basis (following the methodology outlined and utilizing all principal components).

The random sampling of , , and and the subsequent abnormality calculations (Euclidean and MAD in standard and PC basis) were done 500 times for each value of (2, 4, 6, 8, 10, 15, 20, 25, 30, 35, 40, 45, 50). Results were averaged for each value of and reported as seen in Figure 5.

## Data Availability

This article was written with RMarkdown35,36. All source data and code to reproduce the entire manuscript are organized into an R Project35,37 and are freely available at [*Insert URL here*]

# Appendix

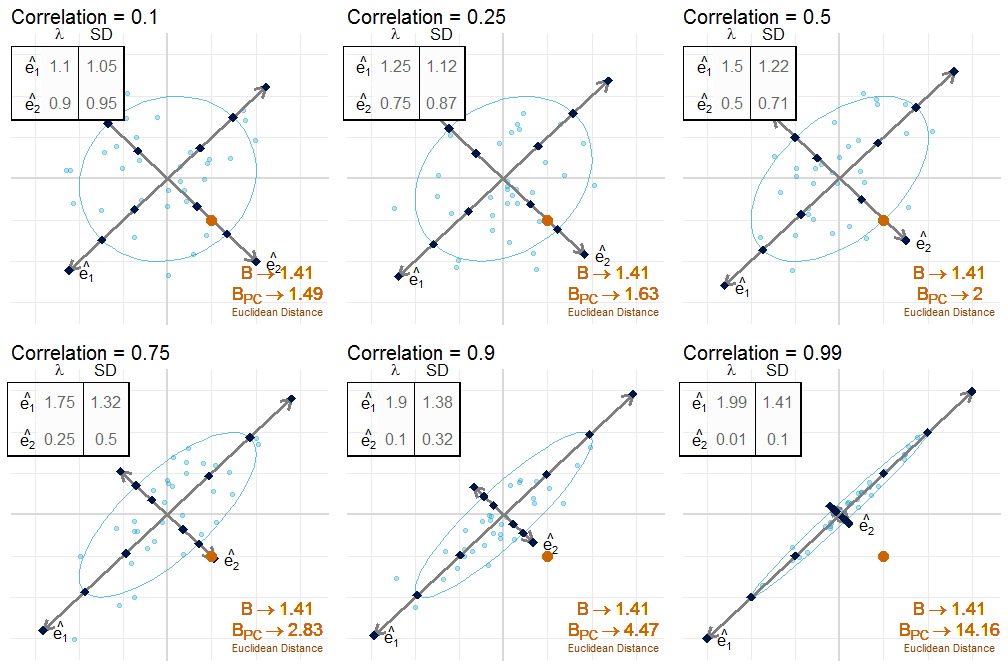


Figure 6: A visual example of the relationship between correlation, eigenvalues and overall abnormality measurement. As correlation between the two variables increases, the difference in overall abnormality (i.e. Euclidean distance from the origin) between the standard basis and the principal component basis for the subject (orange) grows. This shows that the bias introduced when measuring overall abnormality in the standard basis increases as correlation increases.

# References

1. Schutte, L. *et al.* An index for quantifying deviations from normal gait. *Gait & posture* **11,** 25–31 (2000).

2. Mahalanobis, P. C. On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India, 1936* 49–55 (1936).

3. Filzmoser, P., Maronna, R. & Werner, M. Outlier identification in high dimensions. *Computational Statistics & Data Analysis* **52,** 1694–1711 (2008).

4. De Maesschalck, R., Jouan-Rimbaud, D. & Massart, D. L. The mahalanobis distance. *Chemometrics and intelligent laboratory systems* **50,** 1–18 (2000).

5. Schwartz, M. H. & Rozumalski, A. The gait deviation index: A new comprehensive index of gait pathology. *Gait & posture* **28,** 351–357 (2008).

6. Baker, R. *et al.* The gait profile score and movement analysis profile. *Gait & posture* **30,** 265–269 (2009).

7. McGinley, J. L. *et al.* Single-event multilevel surgery for children with cerebral palsy: A systematic review. *Developmental Medicine & Child Neurology* **54,** 117–128 (2012).

8. Thomason, P. *et al.* Single-event multilevel surgery in children with spastic diplegia: A pilot randomized controlled trial. *JBJS* **93,** 451–460 (2011).

9. Molloy, M., McDowell, B., Kerr, C. & Cosgrove, A. Further evidence of validity of the gait deviation index. *Gait & posture* **31,** 479–482 (2010).

10. Cimolin, V., Galli, M., Vimercati, S. L. & Albertini, G. Use of the gait deviation index for the assessment of gastrocnemius fascia lengthening in children with cerebral palsy. *Research in developmental disabilities* **32,** 377–381 (2011).

11. Esbjörnsson, A. *et al.* Quantifying gait deviations in individuals with rheumatoid arthritis using the gait deviation index. *Scandinavian journal of rheumatology* **43,** 124–131 (2014).

12. Celletti, C. *et al.* Use of the gait profile score for the evaluation of patients with joint hypermobility syndrome/ehlers–Danlos syndrome hypermobility type. *Research in developmental disabilities* **34,** 4280–4285 (2013).

13. Pau, M. *et al.* Novel characterization of gait impairments in people with multiple sclerosis by means of the gait profile score. *Journal of the neurological sciences* **345,** 159–163 (2014).

14. Weiner, J. *Pca3d: Three dimensional pca plots*. (2017).

15. Jackson, D. A. Stopping rules in principal components analysis: A comparison of heuristical and statistical approaches. *Ecology* **74,** 2204–2214 (1993).

16. Peres-Neto, P. R., Jackson, D. A. & Somers, K. M. How many principal components? Stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics & Data Analysis* **49,** 974–997 (2005).

17. Ferré, L. Selection of components in principal component analysis: A comparison of methods. *Computational Statistics & Data Analysis* **19,** 669–682 (1995).

18. Dray, S. On the number of principal components: A test of dimensionality based on measurements of similarity between matrices. *Computational Statistics & Data Analysis* **52,** 2228–2237 (2008).

19. Skillicorn, D. B. *Understanding high dimensional spaces*. (Springer, 2012).

20. Frontier, S. Decrease of eigenvalues in principal component analysis-comparison with broken stick model. *Journal of Experimental Marine Biology and Ecology* **25,** 67–75 (1976).

21. Guttman, L. Some necessary conditions for common-factor analysis. *Psychometrika* **19,** 149–161 (1954).

22. Kaiser, H. F. The application of electronic computers to factor analysis. *Educational and psychological measurement* **20,** 141–151 (1960).

23. Horn, J. L. A rationale and test for the number of factors in factor analysis. *Psychometrika* **30,** 179–185 (1965).

24. Raiche, G. *An r package for parallel analysis and non graphical solutions to the cattell scree test*. (2010).

25. Johnstone, I. M. & Lu, A. Y. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association* **104,** 682–693 (2009).

26. Peres-Neto, P. R., Jackson, D. A. & Somers, K. M. Giving meaningful interpretation to ordination axes: Assessing loading significance in principal component analysis. *Ecology* **84,** 2347–2363 (2003).

27. Zimek, A., Schubert, E. & Kriegel, H.-P. A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining* **5,** 363–387 (2012).

28. Verleysen, M. & François, D. The curse of dimensionality in data mining and time series prediction. in *Computational intelligence and bioinspired systems: 8th international work-conference on artificial neural networks, iwann 2005, vilanova i la geltrú, barcelona, spain, june 8-10, 2005. proceedings* (eds. Cabestany, J., Prieto, A. & Sandoval, F.) 758–770 (Springer Berlin Heidelberg, 2005). doi:[10.1007/11494669\_93](https://doi.org/10.1007/11494669_93)

29. Aggarwal, C. C. & Yu, P. S. Outlier detection for high dimensional data. in *ACM sigmod record* **30,** 37–46 (ACM, 2001).

30. Muller, E., Assent, I., Steinhausen, U. & Seidl, T. OutRank: Ranking outliers in high dimensional data. in *Data engineering workshop, 2008. icdew 2008. ieee 24th international conference on* 600–603 (IEEE, 2008).

31. Shyu, M.-L., Chen, S.-C., Sarinnapakorn, K. & Chang, L. Principal component-based anomaly detection scheme. in *Foundations and novel approaches in data mining* 311–329 (Springer, 2006).

32. Aggarwal, C. C., Hinneburg, A. & Keim, D. A. On the surprising behavior of distance metrics in high dimensional space. in *International conference on database theory* 420–434 (Springer, 2001).

33. Fieller, N. *Basics of matrix algebra for statistics with r*. (CRC Press, 2016).

34. Venables, W. N. & Ripley, B. D. *Modern applied statistics with s*. (Springer, 2002).

35. R Core Team. *R: A language and environment for statistical computing*. (R Foundation for Statistical Computing, 2013).

36. Allaire, J. *et al.* *Rmarkdown: Dynamic documents for r*. (2017).

37. RStudio Team. *RStudio: Integrated development environment for r*. (RStudio, Inc., 2015).