A Similarity Measure for Multi-Dimensional Signals

Svetozar Valtchev, Jianhong Wu

Laboratory for Industrial and Applied Mathematics
York University
June 17, 2020

Abstract

Similarity measures are largely needed for a variety of tasks such as anomaly detection, classification and forecasting. In this paper we explore the shortcomings of current multi-dimensional correlations measures, namely the RV coefficient and the normalized RV coefficient. When a particular dimension is positively correlated between 2 samples, while another is negatively correlated, these methods can provide undesired results. We propose a new measure, the DwC (Deviation-weighted Correlation) coefficient, which overcomes these limitations. Our measure correctly incorporates positive and negative correlation on a dimension-by-dimension basis, ultimately providing a more intuitive and useful measure that generalizes to higher dimensions for the comparison of arbitrary matrices. The measure also holds some scaling properties which become useful in the presence of noise. Lastly we provide an example using accelerometer data, to classify common human activities based on maximum DwC between predefined templates and the data.

keywords: RV coefficient, Template matching, Distance measure, Similarity measure, Multivariate correlation, Dependency measure, Measures of association between matrices.

1 Introduction

With the rise of Big Data, fast and efficient data analysis has been central to many sectors. Data collection has become an industry of it's own, and with storage systems being able to handle terabytes of information effortlessly, it is common practice to store a wide variety of data. Unfortunately, the curse of dimensionality creeps up when it is time to analyse it. Time series are of particular concern as the temporal dimensionality has an associated ordering to it.

When it comes to all standard data science practices such as clustering, classification, segmentation, anomaly detection or forecasting, the concept of distance between data often creeps up. In what follows, we explore the current standard known as the RV coefficient for similarity between multidimensional datasets. We then go on to propose our own measure, the DwC coefficient, carefully constructed to overcome shortfalls currently present, and extend this to a generalized similarity measure for matrices.

1.1 Similarity Measures

Correlation is used to measure the strength of association and overall direction between data sets. When it comes to template matching, they can be thought of as an extension to distance measures in the sense that highly correlated samples are more "similar" to one another while lower correlated ones are less "similar". There are many correlation measures for bivariate samples, but when it comes to higher dimensions, less is known. We explore some of these in what follows.

As Ramsay et al. (1984) said, *"Matrices may be similar or dissimilar in a great many ways, and it is desirable in practice to capture some aspects of matrix relationships while ignoring others"*. As is often the case with distance measurements, there is no one clear cut answer. Usually the method used will depend on the task at hand.

The RV coefficient proposed by Escoufier (1969) is the most notable method for calculating association between matrices. The RV coefficient is derived to fall in the range $0 < RV(X,Y) < 1$ and $RV(X,Y) = 0$ if and only if $X'Y = 0$, ie. all dimension are orthogonal between 2 matrices X, Y, as shown in Josse and Holmes (2013). It's inspiration is to be an extension of the standard correlation coefficient ρ, as making X, Y nx1 matrices produces $RV(X,Y) = \rho^2$. This squaring however has some unfortunately effects, as it now eliminates the direction associated with the correlation given by the sign of ρ. If some dimensions have high positive association while others have high negative association, the RV coefficient will produce large overall associativity measures, which as Ramsay et al. points out, will be a decision that will need to be made on a case-by-case basis.

When it comes to template matching in images, Dawoud et al. (2011) utilize the Cross Correlation and the Sum of Absolute Difference approaches. However, this technique is not robust to rescaling or shifts in given dimension, limiting it's effectiveness for multi-variate signals in general.

Nguyen et al. (2014) also propose Multivariate Maximal Correlation Analysis (MAC), a technique which discovers correlations in the data by searching for the transformations that maximize their correlation. The results are promising, but the process is of complexity order $O(d^2 N^{\frac{2}{3}})$, with d being the number of dimensions and N the number of data points. Furthermore, the set of transformations must be pre-specified and provides ambiguity for generalized purposes.

Other common classification techniques include neural networks, where outputs can naturally be used as proxies for similarity measures between classes. However, network based approaches require retraining for any new classes added, and hence don't scale as easy as generalized template matching based on similarity measures.

2 DwC Coefficient

To overcome the shortcomings of the RV coefficient we aim to produce a new measure that will correctly account for negative correlations. A natural solution would be to collect correlations along all dimension and weight them proportionally to the size of their fluctuations.

Given 2 multivariate signals

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m,1} & x_{m,2} & \cdots & x_{m,n} \end{bmatrix},$$

$$Y = \begin{bmatrix} y_{1,1} & y_{1,2} & \cdots & y_{1,n} \\ y_{2,1} & y_{2,2} & \cdots & y_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ y_{m,1} & y_{m,2} & \cdots & y_{m,n} \end{bmatrix}$$

where n is the number of variables of each observation and m is the number of measurements, we define the shorthand notation X_i to be the ith column vector of matrix X,

$$X_i = \begin{bmatrix} x_{1,i} & x_{2,i} & \cdots & x_{m,i} \end{bmatrix}^T$$

which follows similarly for Y_i. Furthermore, σ_{X_i}, σ_{Y_i} and $Cov(X_i, Y_i)$ are the standard deviation and covariance measures for the given column(s).

We define a new correlation coefficient, the DwC coefficient as

$$DwC(X,Y) = \sum_{i=1}^{n} \gamma_i \rho_{X_i,Y_i},$$

where,

$$\gamma_i = \frac{\sigma_{X_i} + \sigma_{Y_i}}{\sum_{i=1}^{n} (\sigma_{X_i} + \sigma_{Y_i})}$$

$$\rho_{X_i,Y_i} = \frac{Cov(X_i, Y_i)}{\sigma_{X_i} \sigma_{Y_i}}$$

with $\text{Cov}(X_i, Y_i)$ being the standard linear dependance between columns X_i and Y_i. Our DwC coefficient can be rewritten in more compact form as

$$DwC = \vec{\gamma} \cdot \vec{\rho}$$

where

$$\vec{\gamma} = \begin{bmatrix} \gamma_1 & \gamma_2 & \cdots & \gamma_n \end{bmatrix}$$

$$\vec{\rho} = \begin{bmatrix} \rho_{X_1,Y_1} & \rho_{X_2,Y_2} & \cdots & \rho_{X_n,Y_n} \end{bmatrix}^T$$

Our coefficient simplifies to the inner product between 2 vectors. It's worth mentioning that the $\text{Cov}(X_i, Y_i)$ term is strictly a measure for independent identically distributed (i.i.d.) variables and only measures linear dependence. However, as we will see in section 3, this technique produces good real world results. This in effect removes the importance of the temporal ordering of the signal. It is an interesting property, which may have advantages when it comes to efficient database storage and smaller file sizes. Furthermore, notice that our matrices X and Y need not be equally spaced out in the time dimension. This is a nice property as many devices sample at rates which may oscillate depending on multiple internal factors. Financial instruments for example, also fluctuate on time intervals which are not constant. Without the need for interpolation, resampling and differentiating a given dataset into increments (ie. $S_t = X_t - X_{t-1}$) to achieve i.i.d. variables, our measure in theory should be less prone to be misused, allowing us to bypass standard preprocessing techniques.

3 Results

3.1 Synthetic Data

We begin with two simulated 3-dimensional signals X and Y, sampled at 1Hz for 40 seconds. The first dimension will be a flat signal, the second dimension a sinusoidal wave with an amplitude of 50, and the last a linear sampling in the range $[-50, 50]$, for both. All dimensions have white noise added to them ($\mu = 0, \sigma = 1$). See Figure 1 for a visualization of the two. Comparing the two signals we calculate the RV coefficient, RV coefficient with mean shifted data, RV coefficient with mean shifted data and normalized, and our DwC coefficient. Results can be seen in Table 1, under Trial 1.

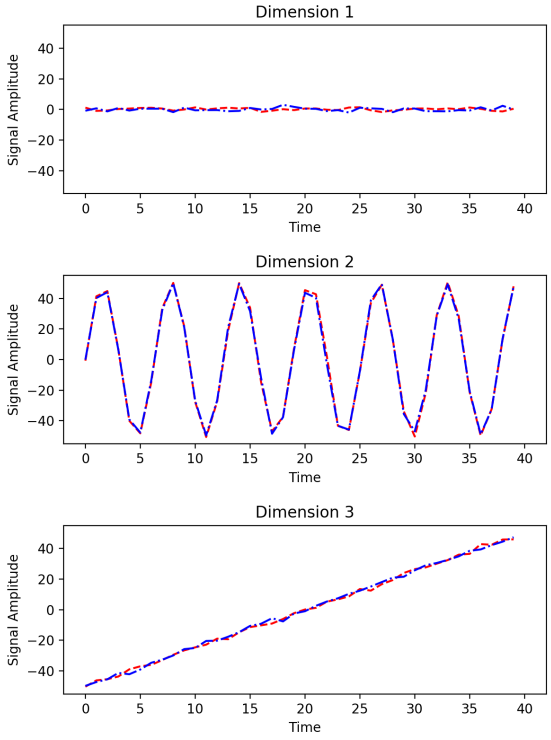

Figure 1: Positively correlated signals in 3 dimensions as seen in Trial 1. Red signal represents X while blue represents Y.

	Trial 1	Trial 2	Trial 3	Trial 4
RV	0.9979	0.3150	0.9980	0.9974
RV (MS)	0.9979	0.3152	0.9980	0.9974
RV (MS+N)	0.8680	0.3054	0.8513	0.8929
DwC	0.9819	0.4539	-0.1011	-0.9821

Table 1: Different association measure values for variations of Signals X and Y

As desired all our measures provide values close to 1, to signify high similarity between X and Y, since they only defer by noise, which does not dominate the actual signal overall. Subsequently we will change our sinusoidal wave to a cosine one in signal Y, to produce a dataset

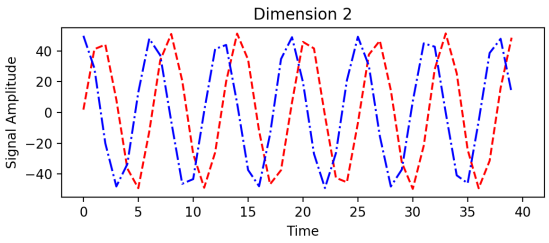

Figure 2: Dimension 2 in Trial 2. Notice the wave appears to be out of phase. Red signal represents X while blue represents Y.

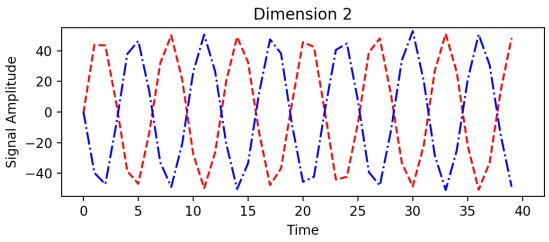

Figure 3: Dimension 2 in Trial 3. The wave is now negatively correlated between the 2 signals. Red signal represents X while blue represents Y.

which should have a relatively low similarity measure as dimension 2 between the signals will be out of phase. This is shown in Trial 2 in Table 1. Our DwC coefficient produces mid tiered values around 0.45, slightly larger than the alternatives. This is due to dimension 1 having nearly no pull on the DwC, unlike that of the RV coefficients. This justifies why our value is close to 0.5 as dimension 3 is still highly correlated while dimension 2 is not.

So far, the DwC measure doesn't outperform the rest, however it's advantage will be evident in Trial 3. We will revert signal Y's 2nd dimension back to the sinusoidal wave we began with, but this time we will reflect it in the y-axis, as to cause dimension 2 to be highly negatively correlated between X and Y. Notice how the RV based measures all show high scores close to 1, despite the fact that the signal in it's most varying dimension is negatively correlated. The DwC on the other hand, comes out at -.1011, implying a slight (but not meaningful) negative correlation. Combined with the fact that dimension 3 is still near-perfectly (positively) correlated between the signals, this is naturally the result we desire.

Lastly we reflect dimension 3 of signal Y, to produce a signal that should reflect an example of negatively correlated datasets. See Figure 4. As the RV coefficients are scaled to the range $[0, 1]$, and only measure association in the absolute sense, they produce measures close to 1 for X and Y. More precisely, in the absence of noise $Y = -X$ and so,

$$RV(X, -X) = 1 = RV(X, X)$$

Conversely,

$$DwC(X, -X) = -1$$
$$DwC(X, X) = 1$$

While the inability to detect the direction of the correlation of the RV based techniques might be warranted in some cases, it is its inability to produce intuitive results like Trial 3, that really could hurdle its use as a generalized multidimensional similarity score.

3.1.1 Analysis

Different similarity measures have strengths and weaknesses, and are largely dependant on the task at hand one aims to achieve. We explore the robustness of the DwC when it comes to translations and scaling, as well as it's sensitivity to noise in what follows.

Common time series analysis techniques rely on a transformation of the data such as calculating the increments S_t between points X_t and X_{t-1}, which clearly will be immune to translations of the signals. Our method doesn't require any special preprocessing, however it still possesses this property. Let

$$C = \vec{1}^T * \begin{bmatrix} c_1 & c_2 & \cdots & c_n \end{bmatrix}$$

,
where c_i are arbitrary constants and $\vec{1}$ being a standard row vector of ones, of size m. Then, since

$$\sigma_{X_i + C_i} = \sigma_{X_i}$$

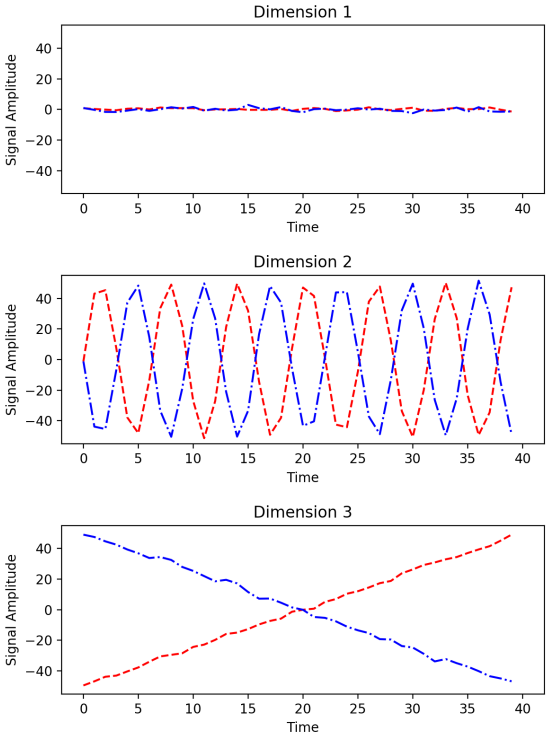

Figure 4: Negatively correlated signals in 3 dimensions from Trial 4. Red signal represents X while blue represents Y.

for a column C_i, and

$$\rho_{X_i+C_i,Y_i+C_i^*} = \frac{Cov(X_i + C_i, Y_i + C_i^*)}{\sigma_{X_i+C_i}\sigma_{Y_i+C_i^*}}$$

$$= \frac{Cov(X_i, Y_i)}{\sigma_{X_i}\sigma_{Y_i}}$$

$$= \rho_{X_i,Y_i}$$

with C* being some other matrix of constants like C, then it follows that

$$DwC(X + C, Y + C^*) = DwC(X_i, Y_i)$$

where $C^{(X)}$ and $C^{(Y)}$ are some arbitrary constant matrices such that each column consists of the same constants (eg. shift all data points X_i by $C_i^{(X)}$

Our DwC coefficient is also robust to slight changes in scale, but careful attention must be given to larger scale transformation. Suppose we scale a signal X, by scale factors $f_1, f_2, ..f_n$ in each dimension. Similar to translations, we can show again that

$$\sigma_{f_iX_i} = f_i\sigma_{X_i}$$

and

$$\rho_{f_iX_i,f_i^*Y_i} = \rho_{X_i,Y_i}$$

However, our γ_i's will be different. Let γ_i be the weights before scaling our signals and γ_i' the weights after scaling. Then

$$\gamma_i' = \frac{\sigma_{f_iX_i} + \sigma_{f_i^*Y_i}}{\sum_{i=1}^n(\sigma_{f_iX_i} + \sigma_{f_i^*Y_i})}$$

$$\frac{f_i\sigma_{X_i} + f_i^*\sigma_{Y_i}}{\sum_{i=1}^n(f_i\sigma_{X_i} + f_i^*\sigma_{Y_i})}$$

$$\neq \gamma_i$$

where f_i^* are some other scale factors similar to f_i. This has the nonlinear effect of scaling a dimension's weight γ_i, based on the scaling of that particular dimension across both signals. This is intuitively a result we might expect, as we fabricated the DwC to be precisely sensitive to larger

	$f = 0.5$	$f = 5$	$f = 50$	$f = 500$
DwC	0.7440	0.9694	0.9819	0.9958

Table 2: DwC coefficient values as we scale up dimension 2 of signal Y, in Trial 1 from Section 3.1.

	$\sigma_N = 0.1$	$\sigma_N = 1$	$\sigma_N = 10$	$\sigma_N = 100$
DwC	0.9984	0.9819	0.8200	0.1738

Table 3: DwC coefficient values as we increase the noise levels across all dimensions of both signals, X and Y.

fluctuations, and so by scaling a dimension, we also scale it's deviation.

As we can see in Table 2, scaling up a dimension with high correlation between signals, has the effect of bringing our DwC coefficient closer to 1, while scaling it down puts more emphasis on the other dimensions. In the case that they are not as strongly correlated (eg. dimension 1 being random noise), we can see that our DwC coefficient will drop. Both of these results are useful depending on the task we have at hand as we will see in Section 3.2.

Lastly, we want to look at how the noise present in each dimension, effects our DwC coefficient. More precisely, we want to see if our measure is highly sensitive to changes in noise, and if there are any shortcomings arising from it. Back in Trial 1, our noise was a simple normally distributed random variable with mean of 0 and standard deviation of 1. This was strategically chosen as dimensions 2 and 3 were of amplitudes much larger than this standard deviation. In this way, our weights γ_i, were larger for these dimensions. Furthermore, the noise in these dimensions was not large enough as to dominate the signals entirely so as to illustrate our measure well, initially. Let our noise be parameterized entire by it standard deviation σ_N. Then for large enough σ_N,

$$\rho_{X_i,Y_i} \to 0$$

since X_i, Y_i will be approximately i.i.d. random variables with respect to one another, and so $Cov(X_i, Y_i)$ will tend to 0. Numerically, this can be seen in Table 3.

It's important to note that we increased the noise in all dimensions in both signals here, unlike with scaling where we only limited our analysis to a single dimension in one signal. We do this because, usually our noise level with be

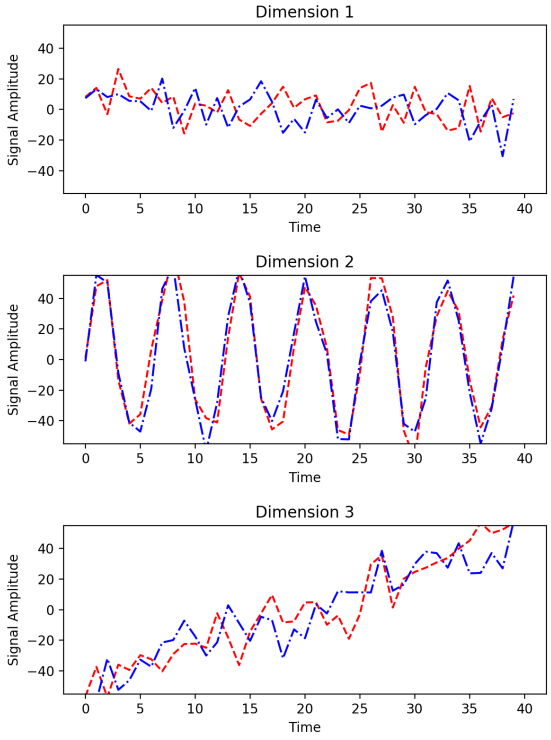

Figure 5: Our initial signals from Trial 1 after increasing the noise level tenfold. Red signal represents X while blue represents Y.

equipment dependant and not proportional to the reading. For example, a GPS system will allow accurately measure geospatial location to a level of confidence in all direction equally. A similar results should be present with dimension dependant noise levels, but a key feature of the DwC might be missed. Precisely, notice with a σ_N of 10 units, our DwC coefficient still provided a reasonable similarity between signals of 0.8200 despite the fact that dimension 1 is entirely overrun by noise and dimension 3 no longer resembles the straight line structure we initially had in Trial 1. This can be seen in Figure 5.

As we increase σ_N to a value of 100, it now is larger than the amplitudes of dimension 2 and

range of dimension 3 combined, and so as we expect, similarity between the signals is no longer present.

3.2 Real Data

We proceed to test our DwC on real world data. For this, we utilize the MHEALTH Dataset Nguyen et al. (2015); Banos et al. (2014), consisting of 10 subjects performing a routine of 12 activities. Three measurement devices were attached to each subject, collecting a wide range of body motions and vital signs. The data is already sampled at 50 Hz, and so no further preprocessing is needed. For our analysis we utilized all signals except for the magnetometer readings, as they are indicative mostly of direction which greatly reduced accuracy as expected.

Activity similarity scores can be seen in Figure 6, for a single repetition of a given activity between subjects 1 and 2. One can think of either subjects' data as the templates, and the other as samples we're trying to match to a template. Not surprisingly, our DwC coefficient is low between all the stationary activities (standing, sitting and lying down), as readings on the devices will all be relatively stable for these 3 categories. We've omitted these results in what can be seen in Figure 7. Optimal values range between 0.22 and 0.75 for matching activity classes. Despite relatively low values for some (cycling), the scores are always largest for a given activity versus all the rest. This can be useful as a template matching classifier for a known set of activities (ie. predicting what activity a subject is performing using only the information from the 3 sensors). Conversely, RV coefficient scores can be found in Figure 8. Notice how similar the RV measure finds the first 6 activities; a results that will be problematic for classification purposes. Furthermore, scores nears 0 appear using the DwC coefficient indicating dissimilarity, while the RV measures are all above 0.40 and provide no intuitive meaning.

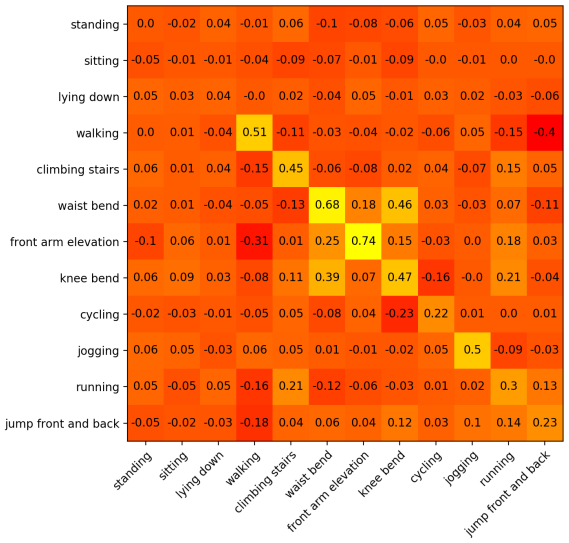

Figure 6: DwC coefficients for all 12 activities between 2 subjects.

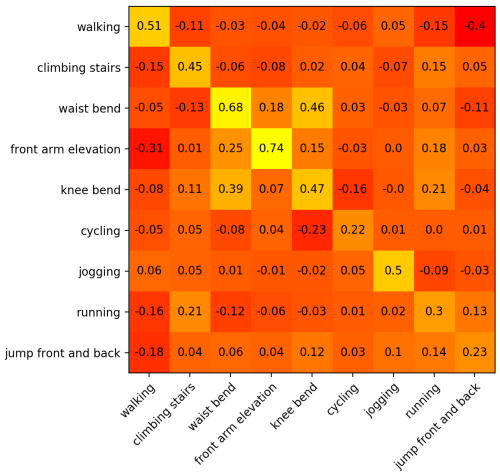

Figure 7: DwC coefficients for 9 non-stationary activities between 2 subjects.

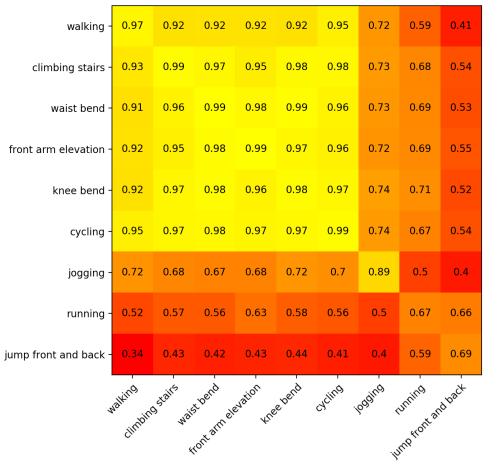

Figure 8: RV coefficient for 9 non-stationary activities between 2 subjects.

4 Discussion

Our DwC coefficient holds some useful properties when it comes to multidimensional signals, which can naturally be extended to a matrix similarity measure, using the formulation we defined in Section 2. First, is it's flexibility when it comes to dimension specific scaling and translations. This is a property that may or may not be of use to a given task. For example, given geospatial signals over some time frame, our DwC coefficient can measure similarity independent of object location. This is true since each dimensions' deviation and correlation is not shift dependant. By being robust to scaling, it also provides a measure independent of subject size and shape. For example, consider the force exerted by a professional athlete vs those of a child in the activity of running. Scaling a given dimension only has the effect of changing it's voting power on the final measure, with factors greater than 1 increasing it's weight, and factors less than 1, decreasing it.

Secondly, the measure has an intrinsic noise robustness property due to the larger deviating dimensions having larger γ_is associated with them. Any time noise levels (as measured by σ_N for example) are sufficiently smaller than the swings in the highly active dimensions, their effect will be negligible. In this manner, dimensions which may consist only of static will not sway the DwC coefficient, making it useful in a large set of applications such as template matching, sentiment analysis and forecasting.

There is no clear distance measure between matrices in general. Unlike a vector which can represent a point in n-dimensional space giving rise to a concrete idea of distance between points, matrices are representations of transformations. This is similarly true for the DwC coefficient as it may only be useful depending on the kind of data that is being analyzed and what the user would determine as "similar".

Our DwC coefficient produces similar results to the popular RV coefficient for most cases, while correctly improving on cases where inter-dimensional correlations may be of different signs (positive in some, negative in others), assigning them lower scores. It also differs in that it can produce signals of "negative association" when the coefficient is less than 0. However, such values should not be misinterpreted necessarily as being highly dissimilar. Instead values near -1 can be representative of high similarity when it comes to template matching, but under some reflection. Again, interpretation will have to be considered on a case by case basic. Nevertheless, values near zero represent low similarity.

5 Conclusion

The RV coefficient is a commonly used measure to compare associativity between matrices. Despite it's simplicity in implementation, it has a few weaknesses. Precisely, it's inability to correctly separate away negative correlation away from positive in a given dimension. Furthermore, it treats every dimension equally, which often will be undesirable in the case of dimensions being largely made up of noise as is usually the case with signal processing. We propose a new similarity measure, the DwC coefficient, that overcomes both limitations. The DwC coefficient scores closer to 0 (intuitively representing less similar signals), when some dimensions are positively correlated while others are negatively correlated. Furthermore, as is often the

case in body motion devices, larger perturbations are more indicative of the overall nature of a signal and so naturally, the DwC measure gives more weight to such signals, thereby bypassing false results caused by noise inherent in stable signals. Negative DwC values should not be confused as highly dissimilar, and low values (near zero) indicate low similarity.

Acknowledgements

This work has been supported by the Natural Sciences and Engineering Research Council of Canada, and by the Canada Research Chairs program.

References

Banos, O., García, R., Holgado-Terriza, J., Damas, M., Pomares, H., Rojas, I., Saez, A., and Villalonga, C. (2014). mhealthdroid: A novel framework for agile development of mobile health applications. volume 8868, pages 91–98.

Dawoud, N., Brahim, S., and Janier, J. (2011). Fast template matching method based optimized sum of absolute difference algorithm for face localization. *International Journal of Computer Applications*, 18:30–34.

Escoufier, Y. (1969). *Echantillonnage dans une population de variables aléatoires réelles*. Faculté des sciences-Secrétariat des mathématiques.

Josse, J. and Holmes, S. (2013). Measures of dependence between random vectors and tests of independence. literature review.

Nguyen, H. V., Müller, E., Vreeken, J., Efros, P., and Böhm, K. (2014). Multivariate maximal correlation analysis. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, page II–775–II–783. JMLR.org.

Nguyen, L. T., Zeng, M., Tague, P., and Zhang, J. (2015). Recognizing new activities with limited training data.

Ramsay, J. O., ten Berge, J. M. F., and Styan, G. P. H. (1984). Matrix correlation. *Psychometrika*, 49:403–423.