A random version of principal component analysis in data clustering

Luigi Leonardo Palese[a,*]

[a] *University of Bari "Aldo Moro", Department of Basic Medical Sciences, Neurosciences and Sense Organs (SMBNOS), Bari, 70124, Italy*

Abstract

Principal component analysis (PCA) is a widespread technique for data analysis that relies on the covariance/correlation matrix of the analyzed data. However, to properly work with high-dimensional data sets, PCA poses severe mathematical constraints on the minimum number of different replicates, or samples, that must be included in the analysis. Generally, improper sampling is due to a small number of data respect to the number of the degrees of freedom that characterize the ensemble. In the field of life sciences it is often important to have an algorithm that can accept poorly dimensioned data sets, including degenerated ones. Here a new random projection algorithm is proposed, in which a random symmetric matrix surrogates the covariance/correlation matrix of PCA, while maintaining the data clustering capacity. We demonstrate that what is important for clustering efficiency of PCA is not the exact form of the covariance/correlation matrix, but simply its symmetry.

Keywords: Principal Component Analysis, Random Projection, Dimensionality Reduction, Data Clustering, Protein Structure, Structural Bioinformatics

1. Introduction

Science today is surrounded by large amounts of data. These are produced by techniques and instruments able to measure a huge number of variables on a large number of samples, or are deposited in an increasing number of online databases that grow exponentially Gross (2011); Berger et al. (2013). Also modern numerical simulations can produce very large and high-dimensional outputs Dror et al. (2012). The challenge of the growing size of data concerns all fields, but the one in which we have seen the most spectacular growth is

[☆] Published in final form as https://doi.org/10.1016/j.compbiolchem.2018.01.009

[☆☆] Preprint available at https://arxiv.org/pdf/1610.08664.pdf

[*] Corresponding author

Email address: luigileonardo.palese@uniba.it (Luigi Leonardo Palese)

probably that of life sciences, where the advancement of genomics, proteomics and other high-throughput technologies has produced an overwhelming amount of data, more and more often freely available to all researchers. Beside the large number of samples, these data are big also because they are high-dimensional: this means that each sample, or instance, of a typical data set contains a large number of degrees of freedom. Such high-dimensionality makes visualization and exploration of samples and data sets very difficult. To overcome these limitations, a series of techniques have been developed that help researchers in visualization, exploration and mining of large data Van Der Maaten et al. (2009); Hassanien et al. (2013).

Among the various algorithms that reduce the dimensionality of data, while retaining the important information, one of the most successful is principal component analysis (PCA) Ringnér (2008). PCA nowadays allows a huge number of tasks, including the phylogenetic classification of the proteins encoded in complete genomes Tatusov et al. (2001), or to obtain insights into protein functional dynamics Yang et al. (2009); Bossis and Palese (2013); Palese (2015b,a, 2016). PCA has been reinvented several times, but it has been developed in its modern form by Pearson and Hotelling Pearson (1901); Hotelling (1933); Bro and Smilde (2014). How PCA works will be briefly recalled below, but here it is important to note that, in its classical implementation, PCA relies on the covariance (or also correlation) matrix of the analysed data. This is a point often overlooked by end-users, but it should be stressed that the number of samples needed to accurately estimate the covariance/correlation matrix of a system containing n degrees of freedom should be (much) larger than n. Otherwise the covariance/correlation matrix will be full of spurious correlations, or even rank deficient from a mathematical point of view if the number of samples is less than n. However here we will show that what is important for the functioning of the method in data clustering, and the related ability to reduce the dimensionality, it is not a particular covariance/correlation matrix, but rather the symmetry that characterizes this type of matrices. The algorithm which will be described can be of general application as will be demonstrated by the analysis of some classic data sets, but our attention will focus particularly on a set of crystallographic structures of the same protein. This data set, being characterized by a low number of samples with respect to the degrees of freedom that describe the system, requires special precautions to be properly analyzed.

2. Theory

Dimensionality reduction consists in the application of mathematical and statistical techniques that reduce the number of variables necessary to the system description. These techniques generally use linear transformations in determining the intrinsic dimensionality of the manifold in which the data set is located and in extracting its principal directions. Among these techniques we can mention linear discrimination analysis, canonical correlation analysis, discrete cosine transform, random projection (RP) and finally PCA, which is certainly the most widely used.

2.1. The PCA algorithm

PCA is a statistical procedure in which a transformation maps a set of observations of (possibly) correlated variables into a set of values of linearly uncorrelated new variables called *principal components*. The first principal component has the largest variance; each of the subsequent components has the restriction of being orthogonal with respect to the previous one. In general, few principal components are needed to account for the majority of variance of the original data set. From a mathematical point of view, PCA is an orthogonal linear transformation. In practice there are different implementations of the PCA; here we will focus on the PCA implementation that is based on the eigenvector decomposition of the correlation matrix Van Der Maaten et al. (2009); Ringnér (2008); Bro and Smilde (2014); Bossis and Palese (2013); Palese (2015b,a); Shlens (2014); Raschka (2015).

We assume that our data are arranged in a matrix such that each row represents a sample (observation or instance), and each column represents a degree of freedom. After the centroid subtraction, the covariance matrix of the data set is obtained as

$$C_{ij} = \langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle$$

where $\langle \ldots \rangle$ represents the average over all the samples in the data set. The correlation matrix is calculated from this matrix as

$$P_{ij} = \frac{C_{ij}}{\sqrt{C_{ii} C_{jj}}}$$

and this square symmetric matrix is diagonalised as

$$R^T P R = \Lambda$$

using standard numerical routines (see the Methods section), where R is an orthonormal transformation matrix (whose column vectors are the eigenvectors of P), the superscript ^{T} means transposition and Λ is a diagonal matrix whose elements are the eigenvalues. After sorting the columns of the eigenvector matrix R and eigenvalue diagonal matrix Λ in order of decreasing eigenvalues, the empirical matrix is projected onto the eigenvectors to give the principal components.

It is interesting to note that the power of PCA in data analysis is not only related to the noise reduction when used as a preparatory step before the application of more dedicated data clustering algorithms. In fact, this noise reduction property alone is not adequate to explain the PCA effectiveness: it was demonstrated that the principal components are the continuous solutions of the class membership indicators in k-means clustering. This means that the dimensionality reduction operated by PCA implies the data clustering according to the k-means objective functions Ding and He (2004).

2.2. The RCA algorithm

In dimensionality reduction and unsupervised data clustering, it should be considered that what really we are interested in is not the identification of the axes that describe the greatest variance of the data (axes which do not have a particular *a priori* meaning), but instead an orthogonal linear transformation of data that could be useful in exploratory data analysis. We can relax the request that the correlation-covariance matrix (the true or the approximated one) is needed for such transformation: it is possible that what is important in PCA as clustering tool may not be the use of a *particular* matrix, but instead of a matrix belonging to a particular *symmetry class*. The bases for such a hypothesis are rooted in the fact that good models for the covariance matrices for the protein configurations obtained from molecular dynamics Palese (2015b,a, 2016) are a class of symmetric random matrices Edelman and Wang (2013). Moreover, the fact that in the Pearson original view Pearson (1901); Bro and Smilde (2014) of PCA which is important is the subspace and not the axes as such, furnish us a further justification.

Among the techniques for reducing the dimensionality of the data sets we previously mentioned the RP. This is a set of simple and efficient techniques for dimensionality reduction which is being increasingly used in recent yearsXie et al. (2016); Geppert et al. (2015); Tasoulis et al. (2014); Varmuza et al. (2011); Palmer et al. (2015). The core idea behind this class of algorithms comes from the Johnson-Lindenstrauss' Lemma Johnson and Lindenstrauss (1984):

Johnson-Lindenstrauss' Lemma: *given* $\forall \epsilon > 0$*, positive integer* n *and* k*, such that* $k \geq k_0 = O(\epsilon^{-2} \ln n)$*. For every set* S *of* n *points in* R^d *there is a linear map* $f : R^d \longrightarrow R^k$ *such that*
$\forall (x_i, x_j) \in S$, $(1-\epsilon)||x_i - x_j||^2 \leq ||f(x_i) - f(x_j)||^2 \leq (1+\epsilon)||x_i - x_j||^2$.

From the above Lemma, we can state that the distance between any two points in a vectorial space of sufficiently high dimension is ϵ-preserved when they are projected in a suitable lower-dimensional space. Given samples x_i in R^d we can project them in R^k by a random projection matrix $W^{k \times d}$ ($k \ll d$) and preserving the distances. From this seminal result, a series of works have shown that RP is a promising class of unsupervised learning algorithms Papadimitriou et al. (1998); Kaski (1998); Achlioptas (2001); Bingham and Mannila (2001). Interestingly, it has been demonstrated that RP can make spherical also highly eccentric clusters Dasgupta (2000). A drawback of RP is that it is highly unstable: even if some algorithms can overcome (at least partially) these difficulties Fern and Brodley (2003); Xie et al. (2016), different projection may lead to different clustering of high dimensional data.

Here we suggest a new RP algorithm that we will call random component analysis (RCA) because of the similarity with the PCA. The central idea for this RP variant, beside the above mentioned Lemma, derives from the empirical observation of the structures and symmetries of the correlation matrices obtained from molecular dynamics experiments Bossis and Palese (2013); Palese (2013, 2015b,a, 2016), and particularly their relation to a class of random matrices Palese (2015b,a, 2016). So, the RCA algorithm is conceived to be performed exactly as the PCA, except for the fact that the square symmetric correlation matrix is replaced by a random symmetric one. This random symmetric matrix

M is defined as

$$M = \frac{G + G^T}{2}$$

where G is a normal distributed random square matrix, so that M belongs to the Gaussian Orthogonal Ensemble Edelman and Wang (2013); Palese (2015b,a, 2016). Thus, the proposed algorithm could be described as a version of classical PCA with relaxed constraints respect to the matrix to be used in calculating the new orthonormal reference system, where only the matrix symmetry is preserved. Obviously, this immediately relaxes also the constraint of the need to have a sufficiently larger number of samples with respect to the degrees of freedom of the system. Although this is not a problem in many areas, as for example in the molecular dynamics data analysis Yang et al. (2009); Bossis and Palese (2013); Palese (2013, 2015b,a, 2016), this could be the case in other applications.

3. Methods

3.1. Well dimensioned data sets

In order to test the performance of the proposed RCA algorithm on well dimensioned data sets (i.e. those ones with a large number of instances or samples respect to the degrees of freedom), three classical and well-known data sets have been used. The first one is the Iris data set, which is perhaps the best known database in the pattern recognition literature Fisher (1936); Anderson (1936). This data set consists of 50 samples from each of three species of *Iris setosa*, *Iris virginica* and *Iris versicolor*. The features reported in the data set are the length and the width of the sepals and petals. Fisher developed a linear discriminant model to distinguish the species from each other on the basis of these characteristics.

We analysed also two chemiometric data sets, both containing a series of chemical features of wine. The Wine data set Forina et al. (1994); Aeberhard et al. (1992), reports the results of a chemical analysis of wines obtained in the same region in Italy but derived from three different cultivars. For for each of the 178 samples in the data set, 13 attributes are reported. The second chemiometric data set (the Wine-quality data set) is related to white variants of a Portuguese wine (4898 samples and 11 attributes; this database contains also the red variant, but we have not considered this part of the data set in our analysis) Cortez et al. (2009). These chemiometric data sets require an additional standardization step before use Raschka (2015). This can be described by the equation

$$x_i{}^{std} = \frac{x_i - \langle x_i \rangle}{\sigma_x}$$

which is part of the standard pre-processing tools in machine learning software. In this work the function implemented in the scikit-learn software package has been usedPedregosa et al. (2011).

All the above mentioned data sets were obtained from the the UCI (University of California at Irvine, School of Information and Computer Science) Machine Learning Repository Lichman (2013).

3.2. A not well dimensioned set of data: the albumin data set

A good example of not well dimensioned data set (i.e. with the number of available samples much lower than the number of degrees of freedom that are necessary for a proper description of the system) can be assembled using an ensemble of crystallographic structures of related proteins. In order to build up a suitably large data set of protein structures we searched in the Protein Data Bank (PDB) Berman et al. (2000) for the albumin entries, with the constraints of specie (human), single protein type in the structure, and resolution of 3.30 Å or better. We will call it the human serum albumin (HSA) data set. The choice fell on this protein simply because it is well represented in the PDB, as well as for the fact that, despite being a monomeric protein, it shows two different conformations (see Results).

After the initial screening, because some N- and C-terminal residues are often not present in the deposited structure, and in order to include the largest possible number of structures as complete as possible, the ones starting after the SER 5 and ending before ALA 569 were excluded from the database. Finally, the structures containing a number of α-carbon atoms different of 565 were also excluded. The final data set contained 58 HSA structures Sugio et al. (1999); Bhattacharya et al. (2000a,b); Petitpas et al. (2001b,a, 2003); Wardell et al. (2002); Zunszain et al. (2003); He and Carter (1992); Ghuman et al. (2005); Yang et al. (2007); Ryan et al. (2011); Zhu et al. (2008); Guo et al. (2009); Hein et al. (2010); Buttar et al. (2010); He et al. (2011); Sivertsen et al. (2014); Wang et al. (2013a,b); Zhang et al. (2015); Bijelic et al. (2016) which are reported in the Supplementary Table 1.

A pdb file of the protein moiety for each of these structures was written in VMD Humphrey et al. (1996) (from SER 5 to ALA 569); these structures were aligned using MultiSeq Roberts et al. (2006) and the pdb files were updated to the new coordinates. The same software was used to calculate the distance trees (RMSD and Q$_h$ style) O'Donoghue and Luthey-Schulten (2005); Russell and Barton (1992). The clusters obtained by these analyses are reported in the Supplementary Table 1.

To obtain the data set in a matrix form, the updated pdb files were loaded in VMD and the α-carbon atom coordinates were extracted and written in a text file such that each row described a structure, by a Tcl (www.tcl.tk) script. Final editing of the raw text file was performed by vim scripting (www.vim.org), so as to obtain the data matrix in a readable file format by the numerical analysis software.

3.3. Numerical implementation

The PCA and RCA algorithms were implemented in the Python language (www.python.org) in an IPython notebook Pérez and Granger (2007). The NumPy numerical software library Van Der Walt et al. (2011) was used, which is part of the Scipy Oliphant (2007) software package. The Pandas McKinney (2010) and Matplotlib Hunter (2007) packages were used to import the Iris and the two chemiometric data sets and to obtain the all graphical outputs, respectively (both packages were obtained from Scipy; www.scipy.org). The implementation of these algorithms is reported in Python format as Supplementary data. Note that two versions of the RCA algorithm are reported: the first one requires the data set and the dimension of the dummy correlation matrix as arguments, while the second requires as arguments the data set and the random matrix that will be used for the calculation of the orthogonal projection system. These files are easily customizable; as it is provided, the software requires seconds or less for the analysis of the proposed data sets (the HSA data set described above, the Iris and the two chemiometric data sets) on an Intel Core i7 machine or a Xeon equipped workstation, both running Ubuntu 14.04 LTS. Very large data sets (as in the case of molecular dynamics outputs; not shown) could require up to (also several) minutes to be analyzed. Since the RCA algorithm performs a random projection, multiple runs of it must be carried out. This because, in a small percentage of cases the algorithm does not get a (two-dimensional) projection that separates the samples in different clusters, although they may be detected (see the Results section). This is the only, and expected, drawback of the implementation of the RCA algorithm here described, which, however, is common to all methods that implement random projection.

4. Results

4.1. Comparing the clustering power of PCA and RCA

The RCA algorithm has been developed bearing in mind the need to obtain an efficient dimensionality reduction and unsupervised clustering of data sets not properly dimensioned. This was the main reason for the introduction of a random symmetric matrix as surrogate of the correlation matrix, which is employed in the classical PCA algorithm. However, it must first be demonstrated, at least, the non-inferiority of this algorithm in the exploratory data analysis of data sets where the performance of the PCA is perfectly known. For such purpose three data sets, retrieved from the UCI Machine Learning Repository, were analyzed with both algorithms. These data sets are not particularly challenging, but they are universally used as a test of machine learning algorithms, particularly the famous Iris data set. These represent different situations, namely a case in which two clusters are certainly present in the data, and two situations in which only one wide cluster can be identified. In one of these last two sets of data it is evident the presence of outliers. The results of PCA and RCA on the Iris data set are reported in Figure 1. As it can be appreciated by looking at the figure, both algorithms easily differentiate the *Iris setosa* cluster from the other two species, whereas the *Iris virginica* and *Iris versicolor* can be only partially discriminated by all the algorithms of this class, since they partially overlap in low-dimensional projections. In the full set of RCA runs, carried out on the Iris data set, similar clustering results have been obtained. The algorithm (almost) always discriminates two clusters in the two-dimensional projections, the composition of which is identical to that obtained by the PCA. Using this

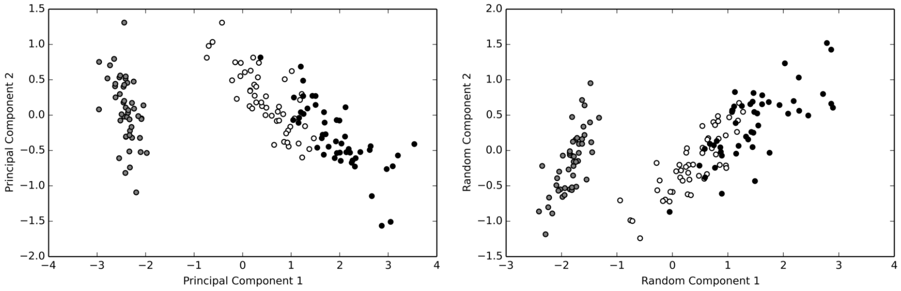

Figure 1: **The Iris data set.** Principal component analysis (left) and random component analysis (right) of the Iris data set are reported. This data set contains 150 entries, 50 for each of the species *Iris virginica* (black), *Iris setosa* (gray) and *Iris versicolor* (white).

particularly simple data set, the RCA algorithm rarely fails in the identification of the two clusters (a rough estimate of the non-recognition of clusters is about 5% of the test performed). It can be stated that, using the Iris data set, RCA is at least not inferior to PCA in clustering purposes, and that the results are reproducible.

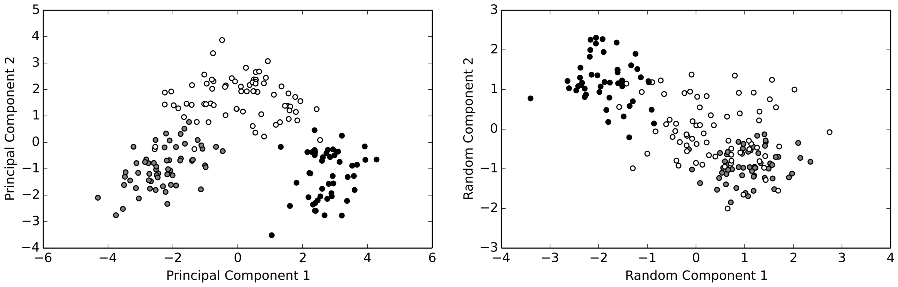

Figure 2: **The Wine data set.** Principal component analysis (left) and random component analysis (right) of the Wine data set are reported. This data set contains 178 entries belonging to three different cultivars, which are reported as black, gray and white circles.

The two chemiometric data sets are a bit more challenging for the linear algorithms. PCA is not able to separate the three cultivars present in the Wine data set as distinct clusters. The Figure 2, left panel, shows that all of them overlap (note that the markers in the figures are externally imposed, and not determined by the classification algorithms). Then the PCA algorithm predicts the existence of a single cluster, even if we can appreciate a preferential localization for different types of sample. Similarly, the RCA algorithm invariably detects a single cluster, with a partial overlap, but with preferential localization, of the cultivars (see Figure 2, right panel). These observations suggest that the RCA algorithm does not exceed the PCA algorithm in the clustering performance.

The analysis of the Wine-quality data set points out another interesting feature shared between the two methods. The projection of this data set onto the first two principal components reveals a single large cluster of data points and two entries that are far away from all other. Figure 3, right panel, allows to visually appreciate the presence of these two outliers, which are highlighted in the Figure. In fact, PCA is also a method employed in the detection of this type of "anomalous" data in large multivariate data sets. Interestingly, these outliers are also detected by the random projection operated by the RCA algorithm. As can be appreciated by inspecting the right panel of Figure 3, these entries are considerably distant from the bulk also when the data set is projected onto the random orthogonal reference system by RCA.

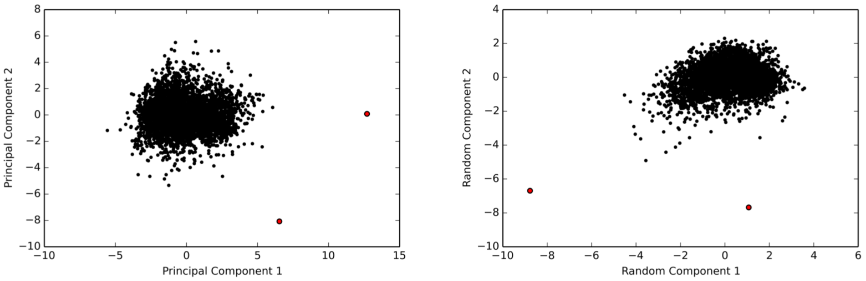

Figure 3: **The Wine-quality data set.** Principal component analysis (left) and random component analysis (right) of the Wine-quality data set are reported. This data set contains 4898 entries which are reported as black dots. Two outliers in PCA are highlighted by red circles in the left panel; the same points are highlighted by red circles (color online) also in the right panel.

These data collectively suggest that RCA has a performance in dimensionality reduction, and cluster detection, comparable to classical PCA. When the entries in a data set can be separated in different clusters by PCA also RCA can do this task. This is true also in the case of single data points or outliers (see Figure 3). If the data cannot be separated in clusters, RCA returns a single cluster, exactly as PCA. These facts, from one hand, tell us that the RCA algorithm is not better than PCA in the unsupervised classification of data. But, from the other, this assures us that it does not introduce any artefactual separations in data.

4.2. The HSA data set

To perform a structural analysis similar to the PCA in a protein structure data set containing a low number of samples respect to the degrees of freedom that describe the protein, we choose to analyze the HSA available structures in the PDB. HSA, Fanali et al. (2012) the most abundant protein in plasma, is a monomeric multi-domain molecule. HSA is a non-glycosylated, all-α protein chain of 65 kDa, with a globular heart-shaped conformation consisting of three homologous domains (I-III). Each domain is composed by two subdomains (A and B). It is an important transport protein with different binding sites able to accommodate a number of chemically different ligands. HSA represents the main carrier for fatty acids (there are seven binding sites for fatty acids, labeled as FA1 to FA7), and it is a depot and carrier for exogenous compounds (mainly, but not exclusively at the Sudlow's sites I and II), thus affecting the pharmacokinetics of many drugs. Among the available structures, we selected 58 structure for the analysis (see the Methods section for the selection criteria). After structural alignment, the α-carbon atom Cartesian coordinates were extracted and arranged in a data matrix (see Methods) which is a coarse-grained representation of the HSA structures. This data matrix was composed of 58 rows and 1695 columns (since 565 α-carbon atoms were finally included in the analysis). This is clearly a degenerated data set, as it is impossible to obtain the true correlation matrix of a multivariate system with 1695 degree of freedom by using only 58 samples. If we calculate the correlation matrix, this will be, at best, only a rank deficient approximation of the true one in which a large number of false correlations must be expected. While it is true that, using a careful error handling (and silencing) program, or also using algorithms that estimate the principal components without ever computing the covariance matrix, it is generally possible to calculate the first principal componentsRoweis (1998); Halko et al. (2011), the classical PCA is not calculable on this data set.

We applied to the albumin data set the RCA algorithm by using, as a dummy covariance-correlation matrix, a square symmetric random matrix of dimension 1695 × 1695. The results of this analysis are reported in Figure 4. As can be easily appreciated by inspecting the figure, RCA leads to two well defined clusters of structures, and what is more interesting is that one cluster contains all and only the HSA molecules with bound fatty acid, the other one only structures without fatty acid. These cluster are reproducible (not shown) and

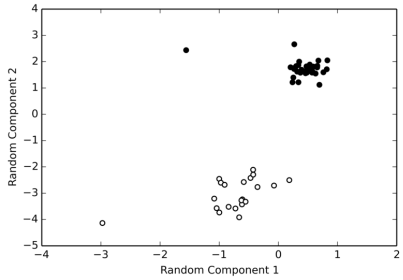

Figure 4: **Random component analysis of the HSA structures.** The Figure reports a random component analysis on the HSA structures contained in the data set described in the text. The HSA structures with bound fatty acids are reported as solid (black) circles, whereas the structures without bound fatty acids are reported as void (white) circles. The algorithm clearly permits to differentiate two clusters of structures in the data set, and the discriminant is the presence of absence, respectively, of bound fatty acids. Two similar cluster of structures have been obtained in all the random component analysis calculations carried out on the HSA data set (snot shown).

are similar to those obtained by different protocolsO'Donoghue and Luthey-Schulten (2005); Russell and Barton (1992) (see Methods and Supplementary Table 1). It is worth noting that a large number of structural and functional works on HSA lead to the conclusion that two structures, possibly related to the presence of fatty acids, are discernible for this protein Fanali et al. (2012); Ascenzi and Fasano (2010). Our RCA analysis permits to go further, as it clearly demonstrates that the only discriminant for such structural switch in the whole data set is the presence or absence of bound fatty acid.

5. Discussion

We developed the RCA method mainly in order to calculate an unsupervised clustering of not well dimensioned data sets, such as those constituted by crystallographic structures. Proteins are structurally and dynamically complex objects Frauenfelder (2002); Palese (2013). Their structure can be studied by molecular dynamics, which is actually at a level of accuracy that permits to predict experimentally observables Bossis and Palese (2011); Dror et al. (2012). In the analysis of molecular dynamics trajectories PCA is of widespread use, as the high-dimensional large number of different molecular conformations that constitute the output of a molecular dynamics experiment is an ideal data set for PCA Kitao and Go (1999); Yang et al. (2009); Bossis and Palese (2013); Palese (2013, 2015b,a, 2016). On the other hand, the number of protein structures reported in the PDB Berman et al. (2000) is collectively large, but there are few structures of a single protein. Although it is possible to find dozens or even hundreds of versions of a single protein in the PDB, the number of available structures is incomparably smaller than the number of degree of freedom of a typical protein. So while PCA can be used in the analysis of the thousands of conformations obtained from molecular dynamics simulations, in its classical implementation PCA can not be used in the analysis of the experimental structures as the low number of different conformations reported in the PDB does not allow an accurate calculation of the covariance matrix. However, in this work it has been shown that a RP-based algorithm can perform in a comparable way respect to the classical PCA algorithm.

The reported data collectively suggest that the proposed RCA algorithm has a performance in dimensionality reduction, and cluster detection, comparable to classical PCA. When the entries in a data set can be separated in different clusters by PCA, as in the case of the *Iris setosa* cluster respect to the *Iris virginica* and *Iris versicolor* one, also RCA can do this task. This is true also in the case of single data points or outliers (see Figure 3). If the data cannot be separated in clusters, as in the case of the species *Iris virginica* and *Iris versicolor* or the bulk entries in both the chemiometric data sets, RCA returns a single cluster, exactly as PCA. These facts show that the RCA algorithm does not outperform PCA in the unsupervised classification of data, and that it does not introduce any artefactual separations in data. But on the other hand the RP algorithm proposed in this communication is easy to implement, conceptually simple and numerically robust. Its performance in dimensionality reduction and unsupervised clustering of large multivariate data sets is, at least, comparable to that of PCA. It is another example of useful application of random matrix theory, Palese (2015b,a, 2016); Edelman and Wang (2013) whose pervasiveness is even more evident in a large number of fields. This work demonstrates that what is important for clustering efficiency of PCA is not the exact form of the covariance/correlation matrix, but instead simply its symmetry, as in our RCA algorithm. The fact that good and informative clustering can be achieved by random projection is nowadays an emerging concept that, beside practical applications, could have far reaching implications also from a conceptual point of view. Finally, this work suggests that an excessive confidence on correlations (which are often spurious) and on large covariance should be avoided, if a simple random matrix could well surrogate them in cluster generation.

Supplementary Information

Supplementary Table and code are available online

References

Achlioptas, D., 2001. Database-friendly random projections, in: Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, ACM. pp. 274–281. doi:10.1145/375551.375608.

Aeberhard, S., Coomans, D., De Vel, O., 1992. Comparison of classifiers in high dimensional settings. Dept. Math. Statist., James Cook Univ., North Queensland, Australia, Tech. Rep .

Anderson, E., 1936. The species problem in Iris. Ann. Mo. Bot. Gard. 23, 457–509.

Ascenzi, P., Fasano, M., 2010. Allostery in a monomeric protein: the case of human serum albumin. Biophys. Chem. 148, 16–22. doi:10.1016/j.bpc.2010.03.001.

Berger, B., Peng, J., Singh, M., 2013. Computational solutions for omics data. Nat. Rev. Genet. 14, 333–346. doi:10.1038/nrg3433.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., 2000. The protein data bank. Nucl. Acids Res. 28, 235–242. doi:10.1093/nar/28.1.235.

Bhattacharya, A.A., Curry, S., Franks, N.P., 2000a. Binding of the general anesthetics propofol and halothane to human serum albumin high resolution crystal structures. J. Biol. Chem. 275, 38731–38738. doi:10.1074/jbc.M005460200.

Bhattacharya, A.A., Grüne, T., Curry, S., 2000b. Crystallographic analysis reveals common modes of binding of medium and long-chain fatty acids to human serum albumin. J. Mol. Biol. 303, 721–732. doi:10.1006/jmbi.2000.4158.

Bijelic, A., Theiner, S., Keppler, B.K., Rompel, A., 2016. X-ray structure analysis of indazolium trans-[tetrachlorobis (1H-indazole) ruthenate (III)](KP1019) bound to human serum albumin reveals two ruthenium binding sites and provides insights into the drug binding mechanism. J. Med. Chem. doi:10.1021/acs.jmedchem.6b00600.

Bingham, E., Mannila, H., 2001. Random projection in dimensionality reduction: applications to image and text data, in: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, ACM. pp. 245–250. doi:10.1145/502512.502546.

Bossis, F., Palese, L.L., 2011. Molecular dynamics in cytochrome c oxidase Mössbauer spectra deconvolution. Biochem. Biophys. Res. Commun. 404, 438–442. doi:10.1016/j.bbrc.2010.11.140.

Bossis, F., Palese, L.L., 2013. Amyloid beta (1–42) in aqueous environments: effects of ionic strength and E22Q (Dutch) mutation. Biochim. Biophys. Acta 1834, 2486–2493. doi:10.1016/j.bbapap.2013.08.010.

Bro, R., Smilde, A.K., 2014. Principal component analysis. Anal. Methods 6, 2812–2831. doi:10.1039/C3AY41907J.

Buttar, D., Colclough, N., Gerhardt, S., MacFaul, P.A., Phillips, S.D., Plowright, A., Whittamore, P., Tam, K., Maskos, K., Steinbacher, S., Steuber, H., 2010. A combined spectroscopic and crystallographic approach to probing drug–human serum albumin interactions. Bioorg. Med. Chem. 18, 7486–7496. doi:10.1016/j.bmc.2010.08.052.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., Reis, J., 2009. Modeling wine preferences by data mining from physicochemical properties. Decis. Support Syst. 47, 547–553. doi:10.1016/j.dss.2009.05.016.

Dasgupta, S., 2000. Experiments with random projection, in: Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence, Morgan Kaufmann Publishers Inc.. pp. 143–151.

Ding, C., He, X., 2004. K-means clustering via principal component analysis, in: Proceedings of the twenty-first international conference on Machine learning, ACM. p. 29. doi:10.1145/1015330.1015408.

Dror, R.O., Dirks, R.M., Grossman, J., Xu, H., Shaw, D.E., 2012. Biomolecular simulation: a computational microscope for molecular biology. Annu. Rev. Biophys. 41, 429–452. doi:10.1146/annurev-biophys-042910-155245.

Edelman, A., Wang, Y., 2013. Random matrix theory and its innovative applications, in: Advances in Applied Mathematics, Modeling, and Computational Science. Springer, pp. 91–116. doi:10.1007/978-1-4614-5389-5_5.

Fanali, G., di Masi, A., Trezza, V., Marino, M., Fasano, M., Ascenzi, P., 2012. Human serum albumin: from bench to bedside. Mol. Aspects Med. 33, 209–290. doi:10.1016/j.mam.2011.12.002.

Fern, X.Z., Brodley, C.E., 2003. Random projection for high dimensional data clustering: A cluster ensemble approach, in: ICML, pp. 186–193.

Fisher, R.A., 1936. The use of multiple measurements in taxonomic problems. Ann. Eugen. 7, 179–188.

Forina, M., Lanteri, S., Armanino, C., Leardi, R., Drava, G., 1994. Parvus: An extendable package of programs for data explorative analysis, classification and regression analysis, version 1.

Frauenfelder, H., 2002. Proteins: paradigms of complexity. Proc. Natl. Acad. Sci. U.S.A. 99, 2479–2480. doi:10.1073/pnas.012579999.

Geppert, L.N., Ickstadt, K., Munteanu, A., Quedenfeld, J., Sohler, C., 2015. Random projections for bayesian regression. Stat. Comput. , 1–23doi:10.1007/s11222-015-9608-z.

Ghuman, J., Zunszain, P.A., Petitpas, I., Bhattacharya, A.A., Otagiri, M., Curry, S., 2005. Structural basis of the drug-binding specificity of human serum albumin. J. Mol. Biol. 353, 38–52. doi:10.1016/j.jmb.2005.07.075.

Gross, M., 2011. Riding the wave of biological data. Curr. Biol. 21, R204–R206. doi:10.1016/j.cub.2011.03.009.

Guo, S., Shi, X., Yang, F., Chen, L., Meehan, E.J., Bian, C., Huang, M., 2009. Structural basis of transport of lysophospholipids by human serum albumin. Biochem. J. 423, 23–30. doi:10.1042/BJ20090913.

Halko, N., Martinsson, P.G., Tropp, J.A., 2011. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. SIAM review 53, 217–288. doi:10.1137/090771806.

Hassanien, A.E., Al-Shammari, E.T., Ghali, N.I., 2013. Computational intelligence techniques in bioinformatics. Comput. Biol. Chem. 47, 37–47. doi:10.1016/j.compbiolchem.2013.04.007.

He, X.M., Carter, D.C., 1992. Atomic structure and chemistry of human serum albumin. Nature 358, 209–215. doi:10.1038/358209a0.

He, Y., Ning, T., Xie, T., Qiu, Q., Zhang, L., Sun, Y., Jiang, D., Fu, K., Yin, F., Zhang, W., Sheng, L., Wang, H., Jianjun, L., Lin, Q., Sun, Y., Li, H., Zhu, Y., Yang, D., 2011. Large-scale production of functional human serum albumin from transgenic rice seeds. Proc. Natl. Acad. Sci. U. S. A. 108, 19078–19083. doi:10.1073/pnas.1109736108.

Hein, K.L., Kragh-Hansen, U., Morth, J.P., Jeppesen, M.D., Otzen, D., Møller, J.V., Nissen, P., 2010. Crystallographic analysis reveals a unique lidocaine binding site on human serum albumin. J. Struct. Biol. 171, 353–360. doi:10.1016/j.jsb.2010.03.014.

Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. J. Educ. Psychol. 24, 417.

Humphrey, W., Dalke, A., Schulten, K., 1996. VMD: visual molecular dynamics. J. Mol. Graphics 14, 33–38. doi:10.1016/0263-7855(96)00018-5.

Hunter, J.D., 2007. Matplotlib: A 2D graphics environment. Comput. Sci. Eng. 9, 90–95. doi:10.1109/MCSE.2007.55.

Johnson, W.B., Lindenstrauss, J., 1984. Extensions of Lipschitz mappings into a Hilbert space. Cont. Math. 26, 189–206.

Kaski, S., 1998. Dimensionality reduction by random mapping: Fast similarity computation for clustering, in: Neural Networks Proceedings, 1998. IEEE World Congress on Computational Intelligence. The 1998 IEEE International Joint Conference on, IEEE. pp. 413–418. doi:10.1109/IJCNN.1998.682302.

Kitao, A., Go, N., 1999. Investigating protein dynamics in collective coordinate space. Curr. Opin. Struct. Biol. 9, 164–169. doi:10.1016/S0959-440X(99)80023-2.

Lichman, M., 2013. UCI machine learning repository. URL: http://archive.ics.uci.edu/ml.

McKinney, W., 2010. Data structures for statistical computing in Python. volume 445. pp. 51–56.

O'Donoghue, P., Luthey-Schulten, Z., 2005. Evolutionary profiles derived from the QR factorization of multiple structural alignments gives an economy of information. J. Mol. Biol. 346, 875–894. doi:10.1016/j.jmb.2004.11.053.

Oliphant, T.E., 2007. Python for scientific computing. Comput. Sci. Eng. 9, 10–20. doi:10.1109/MCSE.2007.58.

Palese, L.L., 2013. Protein dynamics: complex by itself. Complexity 18, 48–56. doi:10.1002/cplx.21434.

Palese, L.L., 2015a. Correlation analysis of Trp-cage dynamics in folded and unfolded states. J. Phys. Chem. B 119, 15568–15573. doi:10.1021/acs.jpcb.5b09678.

Palese, L.L., 2015b. Random Matrix Theory in molecular dynamics analysis. Biophys. Chem. 196, 1–9. doi:10.1016/j.bpc.2014.08.007.

Palese, L.L., 2016. Protein states as symmetry transitions in the correlation matrices. J. Phys. Chem. B 120, 11428–11435. doi:10.1021/acs.jpcb.6b09216.

Palmer, A.D., Bunch, J., Styles, I.B., 2015. The use of random projections for the analysis of mass spectrometry imaging data. J. Am. Soc. Mass Spectrom. 26, 315–322. doi:10.1007/s13361-014-1024-7.

Papadimitriou, C.H., Tamaki, H., Raghavan, P., Vempala, S., 1998. Latent semantic indexing: A probabilistic analysis, in: Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems, ACM. pp. 159–168. doi:10.1145/275487.275505.

Pearson, K., 1901. LIII. On lines and planes of closest fit to systems of points in space. Philos. Mag. 2, 559–572.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Pérez, F., Granger, B.E., 2007. IPython: a system for interactive scientific computing. Comput. Sci. Eng. 9, 21–29. doi:10.1109/MCSE.2007.53.

Petitpas, I., Bhattacharya, A.A., Twine, S., East, M., Curry, S., 2001a. Crystal structure analysis of warfarin binding to human serum albumin anatomy of drug site I. J. Biol. Chem. 276, 22804–22809. doi:10.1074/jbc.M100575200.

Petitpas, I., Grüne, T., Bhattacharya, A.A., Curry, S., 2001b. Crystal structures of human serum albumin complexed with monounsaturated and polyunsaturated fatty acids. J. Mol. Biol. 314, 955–960. doi:10.1006/jmbi.2000.5208.

Petitpas, I., Petersen, C.E., Ha, C.E., Bhattacharya, A.A., Zunszain, P.A., Ghuman, J., Bhagavan, N.V., Curry, S., 2003. Structural basis of albumin–thyroxine interactions and familial dysalbuminemic hyperthyroxinemia. Proc. Natl. Acad. Sci. U. S. A. 100, 6440–6445. doi:10.1073/pnas.1137188100.

Raschka, S., 2015. Python Machine Learning. Packt Publishing, Birmingham, UK.

Ringnér, M., 2008. What is principal component analysis? Nat. Biotechnol. 26, 303–304. doi:10.1038/nbt0308-303.

Roberts, E., Eargle, J., Wright, D., Luthey-Schulten, Z., 2006. Multiseq: unifying sequence and structure data for evolutionary analysis. BMC Bioinformatics 7, 1. doi:10.1186/1471-2105-7-382.

Roweis, S., 1998. EM algorithms for PCA and SPCA. Adv. Neural Inf. Process. Syst. , 626–632.

Russell, R.B., Barton, G.J., 1992. Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. Proteins 14, 309–323. doi:10.1002/prot.340140216.

Ryan, A.J., Ghuman, J., Zunszain, P.A., Chung, C.w., Curry, S., 2011. Structural basis of binding of fluorescent, site-specific dansylated amino acids to human serum albumin. J. Struct. Biol. 174, 84 – 91. doi:http://dx.doi.org/10.1016/j.jsb.2010.10.004.

Shlens, J., 2014. A tutorial on principal component analysis. arXiv preprint arXiv:1404.1100 .

Sivertsen, A., Isaksson, J., Leiros, H.K.S., Svenson, J., Svendsen, J.S., Brandsdal, B.O., 2014. Synthetic cationic antimicrobial peptides bind with their hydrophobic parts to drug site II of human serum albumin. BMC Struct. Biol. 14, 1. doi:10.1186/1472-6807-14-4.

Sugio, S., Kashima, A., Mochizuki, S., Noda, M., Kobayashi, K., 1999. Crystal structure of human serum albumin at 2.5 Å resolution. Protein Eng. 12, 439–446. doi:10.1093/protein/12.6.439.

Tasoulis, S., Cheng, L., Välimäki, N., Croucher, N.J., Harris, S.R., Hanage, W.P., Roos, T., Corander, J., 2014. Random projection based clustering for population genomics, in: Big Data (Big Data), 2014 IEEE International Conference on, IEEE. pp. 675–682. doi:10.1109/BigData.2014.7004291.

Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D., Koonin, E.V., 2001. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucl. Acids Res. 29, 22–28. doi:10.1093/nar/29.1.22.

Van Der Maaten, L., Postma, E., Van den Herik, J., 2009. Dimensionality reduction: a comparative review. J. Mach. Learn. Res. 10, 66–71.

Van Der Walt, S., Colbert, S.C., Varoquaux, G., 2011. The NumPy array: a structure for efficient numerical computation. Comput. Sci. Eng. 13, 22–30. doi:10.1109/MCSE.2011.37.

Varmuza, K., Engrand, C., Filzmoser, P., Hilchenbach, M., Kissel, J., Krüger, H., Silén, J., Trieloff, M., 2011. Random projection for dimensionality reductionapplied to time-of-flight secondary ion mass spectrometry data. Anal. Chim. Acta 705, 48–55. doi:10.1016/j.aca.2011.03.031.

Wang, Y., Yu, H., Shi, X., Luo, Z., Lin, D., Huang, M., 2013a. Structural mechanism of ring-opening reaction of glucose by human serum albumin. J. Biol. Chem. 288, 15980–15987. doi:10.1074/jbc.M113.467027.

Wang, Z.m., Ho, J.X., Ruble, J.R., Rose, J., Rüker, F., Ellenburg, M., Murphy, R., Click, J., Soistman, E., Wilkerson, L., Carter, D.C., 2013b. Structural studies of several clinically important oncology drugs in complex with human serum albumin. Biochim. Biophys. Acta 1830, 5356–5374. doi:10.1016/j.bbagen.2013.06.032.

Wardell, M., Wang, Z., Ho, J.X., Robert, J., Ruker, F., Ruble, J., Carter, D.C., 2002. The atomic structure of human methemalbumin at 1.9 Å. Biochem. Biophys. Res. Commun. 291, 813–819. doi:10.1006/bbrc.2002.6540.

Xie, H., Li, J., Zhang, Q., Wang, Y., 2016. Comparison among dimensionality reduction techniques based on random projection for cancer classification. Comput. Biol. Chem. 65, 165–172. doi:10.1016/j.compbiolchem.2016.09.010.

Yang, F., Bian, C., Zhu, L., Zhao, G., Huang, Z., Huang, M., 2007. Effect of human serum albumin on drug metabolism: structural evidence of esterase activity of human serum albumin. J. Struct. Biol. 157, 348–355. doi:10.1016/j.jsb.2006.08.015.

Yang, L.W., Eyal, E., Bahar, I., Kitao, A., 2009. Principal component analysis of native ensembles of biomolecular structures (PCA_NEST): insights into functional dynamics. Bioinformatics 25, 606–614. doi:10.1093/bioinformatics/btp023.

Zhang, Y., Lee, P., Liang, S., Zhou, Z., Wu, X., Yang, F., Liang, H., 2015. Structural basis of non-steroidal anti-inflammatory drug diclofenac binding to human serum albumin. Chem. Biol. Drug Des. 86, 1178–1184. doi:10.1111/cbdd.12583.

Zhu, L., Yang, F., Chen, L., Meehan, E.J., Huang, M., 2008. A new drug binding subsite on human serum albumin and drug–drug interaction studied by X-ray crystallography. J. Struct. Biol. 162, 40–49. doi:10.1016/j.jsb.2007.12.004.

Zunszain, P.A., Ghuman, J., Komatsu, T., Tsuchida, E., Curry, S., 2003. Crystal structural analysis of human serum albumin complexed with hemin and fatty acid. BMC Struct. Biol. 3, 6. doi:10.1186/1472-6807-3-6.