ELSEVIER

The application of principal component analysis to drug discovery and biomedical data

Alessandro Giuliani

Environment and Health Department, Istituto Superiore di Sanità, Roma, Italy

There is a neat distinction between general purpose statistical techniques and quantitative models developed for specific problems. Principal Component Analysis (PCA) blurs this distinction: while being a general purpose statistical technique, it implies a peculiar style of reasoning. PCA is a 'hypothesis generating' tool creating a statistical mechanics frame for biological systems modeling without the need for strong a priori theoretical assumptions. This makes PCA of utmost importance for approaching drug discovery by a systemic perspective overcoming too narrow reductionist approaches.

Introduction

Principal Component Analysis (PCA) is the by far most widespread multidimensional data analysis technique [1]. Its application range goes from theoretical physics to meteorology, psychology, biology, chemistry, and engineering. Traversing these fields of inquiry, PCA changed its name: Factor Analysis, Singular Value Decomposition (SVD), Singular Spectrum Analysis (SSA), Karhunen–Loeve transformation, Essential Dynamics are some of the names the same technique (with only relatively minor modifications) took along a longer than a century (the first theoretical paper on the argument dates back to 1873) history [1–8].

The applications of PCA range across all the main themes of pharmacology and biomedical sciences as well, going from Quantitative Structure Activity Relationships [5,6], to data mining [5] and different 'omics' approaches [10–12]. The 'data deluge' consequent to both the diffusion of high-throughput techniques and the development of big data bases will further increase the relevance of principal component analysis in pharmacological and biomedical research.

The reader can easily find many excellent works explaining the different facets of PCA (see for e.g.: [1,6–9]). Here I will focus on the fact PCA is not only a 'smart data analysis technique', but implies a somewhat unusual systemic view in doing science.

In his seminal 1901 paper [3], Karl Pearson synthetically defined the main goal of PCA: 'In many physical, statistical and biological investigations it is desirable to represent a system of points in plane, three or higher dimensioned space by the 'best fitting' straight line or plane'. The need to collapse multidimensional information scattered over different (and sometimes heterogeneous) descriptors into a lower number of relevant dimensions is one of the main pillars of scientific knowledge.

Pearson continues: 'In nearly all the cases dealt with in the textbooks of least squares, the variables on the right of our equations are treated as independent, those on the left as dependent variables'. This implies that the minimization of the sum of squared distances only deals with the dependent (y) variable. The variance along independent (x) variable, being the consequence of the choice of the scientist (e.g. dose, time of observation.) is supposed to be strictly controlled and thus does not enter in least squares computation.

The novelty of PCA lies in a different look at reality, again Pearson: 'In many cases of physics and biology, however, the 'independent' variable is subject to just as much deviation or error as the 'dependent' variable, we do not, for example, know x accurately and then proceed to find y, but **both x and y are found by experiment or observation**'.

This new attitude is the core of the peculiar 'best fitting' procedure set forth by Pearson. Fig. 1 reports on the left the original plot of Karl Pearson and on the right the classical regression scheme.

In PCA (left panel) the distances to minimize are perpendicular to the model of the data (the straight-line correspondent to the first principal component of x, y space), while in the classical regression model (right panel) the distances are perpendicular

E-mail address: alessandro.giuliani@iss.it.

1359-6446/© 2017 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.drudis.2017.01.005

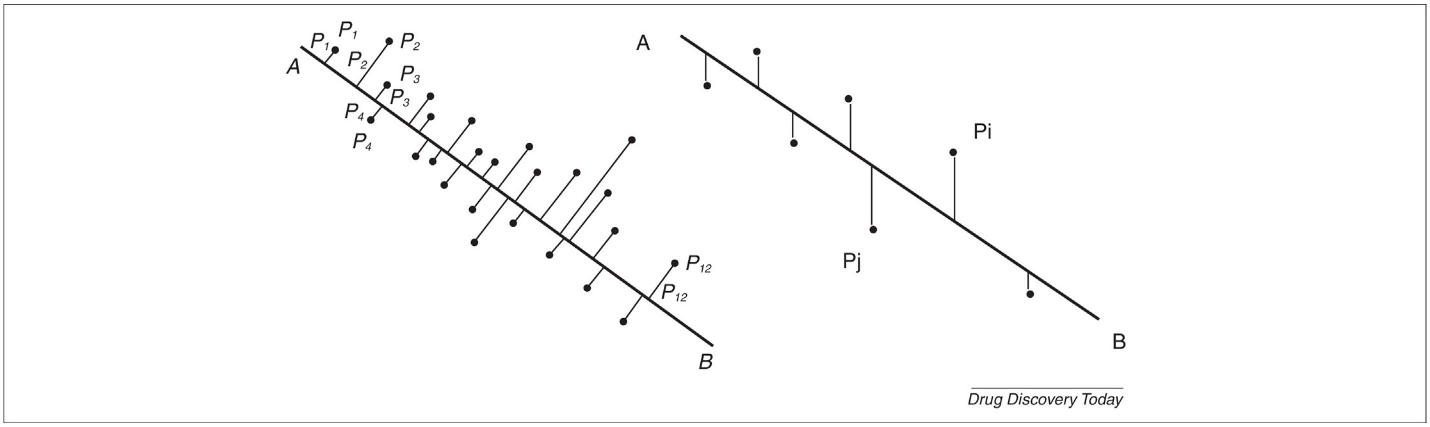

FIGURE 1

Two different least-squares strategies. The figure reports a graphical comparison of the least squares optimization proposed by Karl Pearson (left panel, original Pearson drawing from [3]) and the usual least squares paradigm adopted in linear regression (right panel). The neat separation between an independent *X* variable whose variance is decided by the experimentalist and a dependent *Y* variable correspondent to the experimental outcome, makes the fit only dependent on the scattering of the vector points parallel to *Y* axis (linear regression, right panel). In the left panel (Principal Component Analysis) both *X* and *Y* are affected by error, consequently the distances of the experimental points from the line (component) are computed on the bi-dimensional *X*, *Y* space (orthogonal to the component). The independent-dependent distinction is abolished.

to *x* axis, because the only uncertainty taken into account refers to *y*. This apparently minor geometrical detail encompasses a sort of revolution in the style of doing science. Principal components are linear combinations of the *n* variables defining the studied system according to the formula:

$$\mathrm{PC} = ax1 + bx2 + cx3 + \ldots + kxn.$$

where $X1 - Xn$ are the experimental/observational variables defining the statistical units, the *a*, *b*, *c*,. . .,*k* coefficients are estimated by least squares optimization. Principal components are both the 'best summary', in a least square sense, of the information present in the *n*-dimensional data cloud, and the directions along which the between variables correlation is maximal.

Pearson did not limit the number of components to one (line) but extends it to two (plane) or more. More in general, when in presence of *n* initial variables, it is ever possible to rotate the original *n*-dimensional space into another reference set spanned by *n* mutually independent axes (components), preserving the entire initial information. When a relevant among variables correlation structure does exist, a '*good fit*', i.e. a reliable reconstruction of the original information present in the data set, can be obtained by *p* components with $p < n$. This reduction of dimensionality corresponds to the '*desirable representation*' Karl Pearson refers to in the initial phrase of his 1901 paper. These 'desirable representations' take with them an implicit model of the system at hand [1–9]. To make this model explicit we need to give a name to the extracted components. This can be done by considering the degree of participation of the original variables to the PCs (hypothesis generating phase) and by subsequent hypothesis driven research on expected modulation of component scores by suitable chosen external agents (hypothesis testing phase).

In the following I will introduce the PCA-based approach to complex systems modeling. Chapter 'Catching the sense of experimental results: a case study on animal behavior' reports an exemplar application of PCA in order to illustrate the effective use of the technique in research practice. Chapter 'Enlarging the view' enlarges the view to different fields of biological investigation. Chapter 'Collective parameters from PCA: biological statistical mechanics' deals with the possible role of PCA in the foundation of a sort of biological statistical mechanics, with special reference to network pharmacology.

Catching the sense of experimental results: a case study on animal behavior

The need to de-convolve the hidden independent factors modulating a given set of observed variables, is particularly cogent in behavioral research. An example is given by one of the most popular behavioral test to estimate learning and memory in animals: the Morris Water Maze (MWM). In the reported case, MWM was applied to the comparison between female rats exposed to an enriched environment (EE) and housed in standard conditions (SC) (13).

The MWM paradigm works like this: The test subject is put into a circular pool provided with an invisible platform, located in a fixed position, 2 cm below the surface of the water. Rats undergo different consecutive training sessions in which they swim until they reach the escape platform and climb on it. Progressively the animal learns the position of the platform, the process is described by five variables: (1) the time each animal spent to reach the platform (*latency*), (2) percentage of time spent in quadrant containing the platform (*target*), (3) distance travelled (*distance*), (4) swimming *speed* (distance/latency), (5) time spent in the target quadrant 24 h after the removal of platform (*probe*).

The MWM descriptors are a blend of two 'hidden variables' we can grossly define as motility (i.e. efficiency of swim) and learning. In order to unambiguously evaluate the nature of the effect exerted by treatment on the experimental animals, we need to disentangle the two motor and learning components.

To this aim a PCA was computed on the data set having as statistical units 31 rats (16 SC, 15 EE) defined by the above described five variables. This gave rise to the distribution of explained variance on of the eigenvalues of the between descriptors correlation matrix reported in Table 1.

TABLE 1

Eigenvalue distribution.

	Eigenvalue	Difference	Proportion	Cumulative
1	***2.328***	***0.819***	***0.4657***	***0.4657***
2	***1.509***	***0.892***	***0.3020***	***0.7676***
3	***0.618***	***0.083***	***0.1236***	***0.8913***
4	0.535	0.526	0.1070	0.9983
5	0.008	0.002	1.0000	

The table reports the Eigenvalues of the correlation matrix relative to the five MWM original variables. The eigenvalues are in descending order, the second column reports the subsequent eigenvalues differences, while third column corresponds to the proportion of variance explained by each eigenvalue. The last column is the cumulative variance explained by the different solutions at increasing number of components. The three component solution (bold italic values) was further analyzed in terms of loading pattern.

The initial data set has five dimensions (MWM descriptors), thus a five component solution, correspondent to a rotation of the system of points, allows for explaining the 100% of initial information. The existence of a correlation structure among the MWM descriptors provokes an asymmetric explained variance distribution on the components: the first three components jointly explain the by far major part of total variance (89%), being PC1 responsible for 47%, PC2 for 30% and PC3 for 12% of total variance respectively.

There are many empirical methods to select the number of components to retain, the most straightforward one being the visual 'scree' test: the number of 'signal' components to be retained corresponds to the reaching of a plateau of eigenvalue at increasing number of components [14]. A frequently adopted 'rule-of-thumb' is to retain (in the case of PCA on correlation matrix, like this one) all the components having an eigenvalue greater than one. This rule descends from the fact the use of correlation matrix implies each variable has a unit standard deviation (being the correlation correspondent to the covariance of standardized variables) thus a component having an eigenvalue lesser than one has an explanatory power lower than what expected by a single descriptor. The adoption of this criterion generates a two component (77% of total variance explained), as *bona fide* signal [1,14].

Beside any empirical 'recipe', the last word on component selection should reside in the analyst judgement that, inspecting the 'component loadings', correspondent to the Pearson correlation between original variables and components (Table 2), tries to 'give a name' to the components on the basis of the loading pattern.

TABLE 2

Component loadings.

	PC1	PC2	PC3
latency	**0.917**	0.	0.099
distance	**0.758**	**0.644**	0.029
speed	−0.283	**0.826**	−0.172
target	−**0.656**	0.418	−0.386
Probe	−**0.633**	0.390	**0.655**

Loading pattern of the five MWM on the three main principal components. The loadings correspond to the Pearson correlation coefficients between variables and components. The loadings of the variables most relevant for the component interpretation are bolded.

Looking at Table 2, it is immediate to note how PC1 has mainly to do with *latency* (loading = 0.92) and, on a much lesser extent, with *distance, target* and *probe* variables. The opposite signs of *target* and *probe* with respect to *latency* and *distance* loadings, are consistent with the interpretation of PC1 as 'learning component'. Rats with high PC1 scores are those that spend a lot of time to reach the platform (high values of *latency*, the pivot variable of PC1), have an erratic motion in the pool (high *distance*) not easily reaching the correct quadrant (low values of *target*) while in the same time performing poorly at probe test (low values of *probe*). That is to say that PC1 is an inverse measure of learning: low values of PC1 point to 'good learners' while high values to 'poor learners' [13].

The variable most correlated with PC2 is *speed* (loading = 0.83), *distance* has a relevant (0.64) loading on PC2, while the other variables show low correlation with PC2. This prompts us to define PC2 as a 'motility' component. Rats with high values of PC2 are those having an elevated motility (high swimming speed) and travelling for longer distances in the pool.

PC3, beside its relatively low amount of explained variance (Eigenvalue less than one), has only one relevant loading with *probe*, thus it looks like an idiosyncratic property of this measure, while both *latency* and *distance*, the two main MWM descriptors, are totally uncorrelated with PC3.

The above reasoning implies we can safely collapse the initial five dimension data set into a two dimensional component space spanned by PC1 and PC2 that in turn verify the initial hypothesis of 'learning' and 'motor activity' as the two hidden factors shaping MWM test.

Each original variable is a mixture, blending with different proportions learning (PC1) and motility (PC2) hidden factors. *Latency* is an 'almost pure' learning descriptor, *Distance* is a balanced mixture of learning and motility.

As for *Distance*, a multiple regression gives the following 'recipe' disentangling the contributions of the two main components:

$$\text{Distance} = 1040.34 + 203.42(\text{PC1}) + 172.73(\text{PC2}) \qquad (1)$$

The fit of this reconstruction is almost perfect (see Fig. 2 bottom panel) reaching a Pearson *r* near to unity (*r* = 0.995). *Distance* can thus be considered as a cocktail obtained by mixing 203 parts of Learning with 173 parts of Motility, the intercept (1040.34) being a scale factor necessary for passing from *z*-scores (the component have by construction zero mean and unit standard deviation on the whole set) to actual *Distance* values. This almost perfect fit comes from the fact *Distance* has null correlations with the minor components (0.029, 0.071, −0.064; for PC3, PC4 and PC% respectively) and is thus completely defined in terms of the two major components.

This allows to separate the 'learning' and 'motility' components of *Distance*.

Fig. 2 reports (bottom panel) the almost perfect fit between distance as estimated by PCA and the observed distance values. The two EE and SC groups are scattered along the regression line without any neat separation. This is consistent with the lack of any statistical significant difference between EE and SC rats as for Distance (Table 3).

When the animals (Fig. 2, top panel) are projected in the PC1, PC2 space, the two groups are neatly discriminated (Table 3) allowing us to grasp the effect exerted by the treatment.

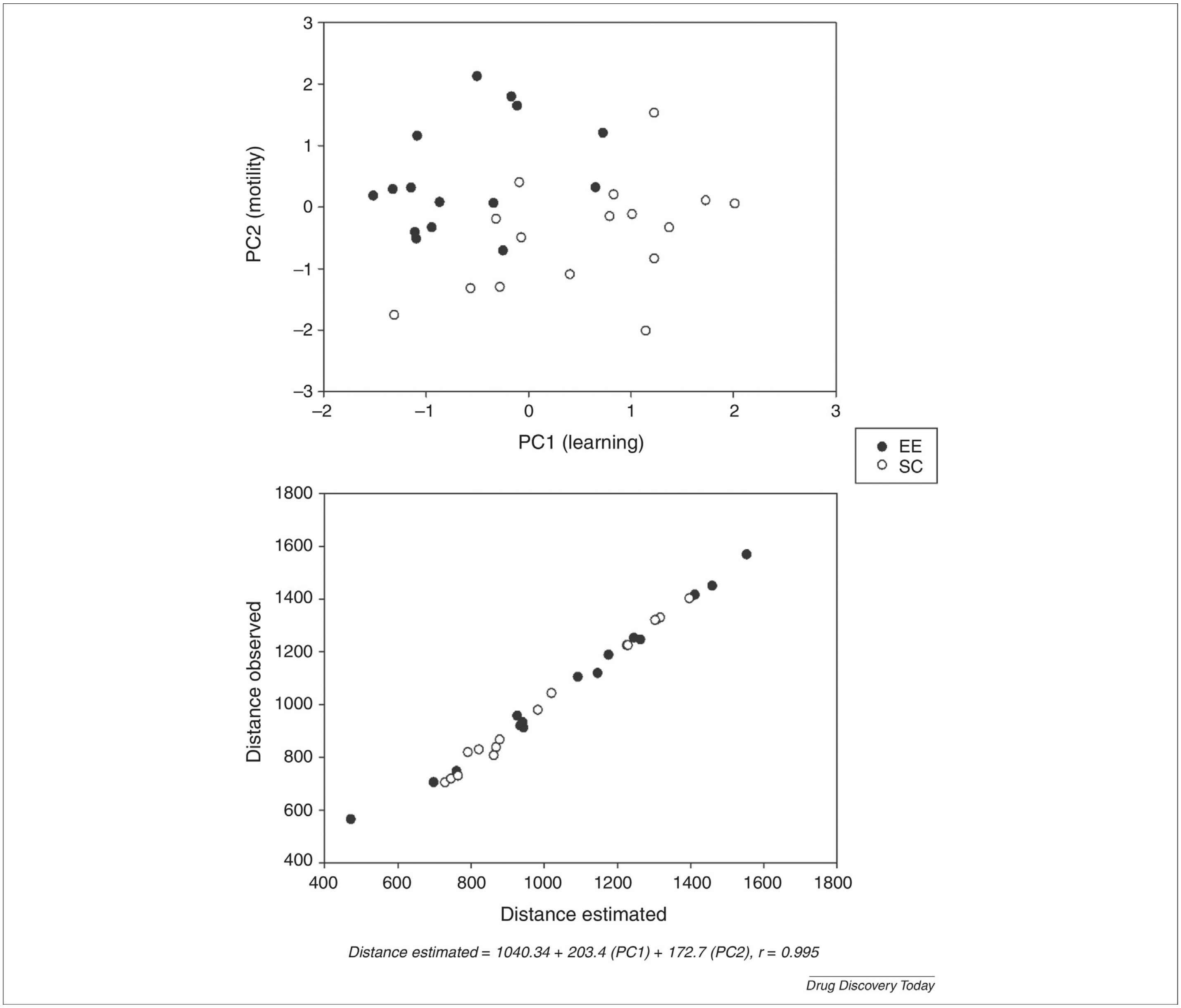

FIGURE 2

Component space and original variables reconstruction. The distance travelled by the rat in MWM is almost perfectly ($r = 0.995$) estimated as a linear function of learning (PC1) and motility (PC2) components. The EE (black dots) and SC (white dots) are not discriminated in terms of distance (bottom panel). The two EE and SC groups are very well discriminated in the PC1 vs. PC2 component space (top panel, PC1 is positively correlated with Latency, thus it is an inverse measure of learning, see text).

Table 3 reports the statistical comparisons between EE and SC groups (statistical significant results bolded).

EE animals show both better learning (lower PC1 values) and higher motility (higher PC2 scores) than SC ones. These two effects exert an opposite effect on *distance*, this balance eliminates any statistical significance as for *distance* variable as such. When the two effects are disentangled by PCA (Fig. 2, top panel) we can appreciate the between groups differences in terms of latent factors.

TABLE 3

Descriptive and inferential statistics.

Variable	EE (mean and SD)	SC (mean and SD)	*t*-Value	*p*
Distance	996 (256)	1092 (281)	0.89	0.381
PC1	−0.606 (0.683)	0.568 (0.923)	***4.01***	***0.0004***
PC2	0.484 (0.891)	−0.454 (0.897)	***−2.92***	***0.0068***

Comparison between SC (Standard Cage) and EE (Enriched Environment) groups in terms of descriptive (mean and standard deviation) and inferential (*t*-value and statistical significance) statistics for *Distance*, PC1 and PC2. The statistically significant values are given in bold italics.

The above results indicate how important is to disentangle [15,16] the different latent factors embedded into a complex (even if apparently direct) measurement: not taking into consideration this inherent complexity gives a false impression of lack of effect out of the combination of two opposite statistically significant modulations.

The procedure described above is nothing else than an example of 'scientific wisdom' supported by PCA, it is worth noting how the quantitative aspects are strictly intermingled with the 'qualitative reasoning' allowing for a 'soft' approach not imposing any (often unjustified) a priori hypothesis on the data.

Enlarging the view

The above way of reasoning can be applied to any experimental or descriptive study in science. In biomedical sciences it is extremely common to interpret any biochemical (or pharmacological) mechanism in terms of pathways in which different agents (enzymes, genes, transcription factors, drugs, etc.) mutually interact to generate regulation networks. Such networks are formally equivalent to correlation matrices in which the agents (nodes) are the variables and the strength of their relationship (links) are proportional to their correlation coefficients [15,16]. Being the principal components the eigenvectors of the correlation matrix among a set of variables, it is straightforward to interpret the PCA solution of an interaction network as the 'optimal decomposition' into independent pathways of a set of interactions.

Still more cogent is the relation linking PCA to the extraction of relevant information in metabolomics [11,12] where a complex NMR or MS spectrum is de-convolved into its principal components, each one corresponding to an 'hidden factor' of the studied complex system. These factors, depending on the specific application, can correspond to metabolic modules, specific pathways or drug mode-of-actions [17,18].

The components act as 'signatures' of a specific complex entity (a mode-of-action, a metabolic pathway.) that cannot be easily traced back to a classical 'if-then' mechanistic explanation. The above considerations hold true for any other –omics dealing with extremely high dimensional data sets.

The limits of PCA approach are strictly linked with its merits and derive from the intrinsic context dependent character of correlations.

As aptly explained by Huang [19] with specific reference to Systems Biology, complex systems (here we adopt the Weaver [20] definition of 'organized complexity' for all those systems whose behavior depends upon non-trivial relations among its parts) show an exacerbated context dependence.

Context dependence can be a very serious limitation of PCA approach. Like any other correlation based approach, PCA is strictly dependent upon the data variability range. Fig. 3 reports the strict correlation existing between the gene expressions profiles relative to two independent samples of the same kind of tissue (macrophages in this case). The points (approximately 25000) are the logarithm of the expression level of different genes; the axes correspond two independent macrophage population samples [21].

The figure suggests a striking order governing gene expression regulation at large: the two samples have a Pearson correlation coefficient equal to 0.98 on the whole genome expression profile. This organization does not originate from strict mechanistic relations but from a thermodynamic (and thus statistical) emergent behavior: the degree of order (correlation) is thus strictly dependent on the considered scale. From the top left inset of the figure is evident how the amount of correlation is a function of *d*, the expression range taken into consideration. This means that the entity of the correlation (in PCA terms the amount of variance explained by PC1) increases at larger scales while it vanishes at smaller gene expression variability ranges.

This implies that changing the reference data set could change the nature (loading pattern) and relative importance (eigenvalues, percentage of explained variance) of the extracted components [15,22]. This lack of generality of PCA can be turned into an advantage when using PCA for deriving thermodynamic-like descriptors of complex biological systems.

Collective parameters from PCA: biological statistical mechanics

In effective physical models, many microscopic details collapse into few coarse-grained parameters. Thus, while three-dimensional liquids have enormous microscopic diversity, in many conditions their behavior only depends upon their viscosity and density [23]. This is the main reason of the success of thermodynamics that allows for getting rid of very complex systems behavior by means of few collective parameters without entering into detailed (and impossible to manage) microscopic details.

In the following I will comment on two extremely interesting 'thermodynamics-like' avenues of research in biomedicine, holding promises to surmount the actual crisis of molecular ultra-reductionist approaches [24]: (1) complex networks analysis and (2) generalized order and organization quantification. In both cases PCA plays a leading role.

Networks

In 1952, the Dutch electrical engineer Bernard Tellegen [25] developed a conservation principle (tailored upon Kirchoff's laws of electrical circuits) of both potential and flux across a network. The flux does not need to be an electrical current and the same holds for the potential: any system that can be modeled by a set of nodes linked by edges has similar emerging properties independently of the physical nature of nodes and edges. This result opened the way to a sort of 'network thermodynamics' strictly dependent from wiring architecture while largely independent of the constitutive laws governing the single elements of the network [26]. PCA naturally fits into this paradigm: a correlation matrix corresponds to a graph (network) having the variables as nodes and the edges labeled by pairwise correlation coefficients [15]. The principal components (eigenvectors) of a network allow for the optimal 'modular dissection' of the network itself [27]. This property is widely used to dissect protein structures into modules (domains) [27] or to identify 'preferred pathways' in metabolic networks [28].

On a drug discovery perspective, it is worth noting how the emerging field of 'Network Pharmacology' [29–31] is drastically changing the notion of what is a 'target' of drug action. The long-lasting crisis in the number of new developed drugs [32] urges scientists to re-consider the notion of 'ideal biological profile' of a drug molecule.

For decades, the dominant paradigm of pharmacology was fully reductionist: the goal of research efforts was the quest for the main molecular determinant of a given disease. This determinant, generally a protein molecule, was considered the 'target' of the drug (receptor) and the candidate molecules were screened for their ability to bind selectively to the receptor [29–31,33]. The 'best binders' candidates entered into subsequent phases in which their efficacy was tested on animal models of the disease, and eventually go into clinical trails.

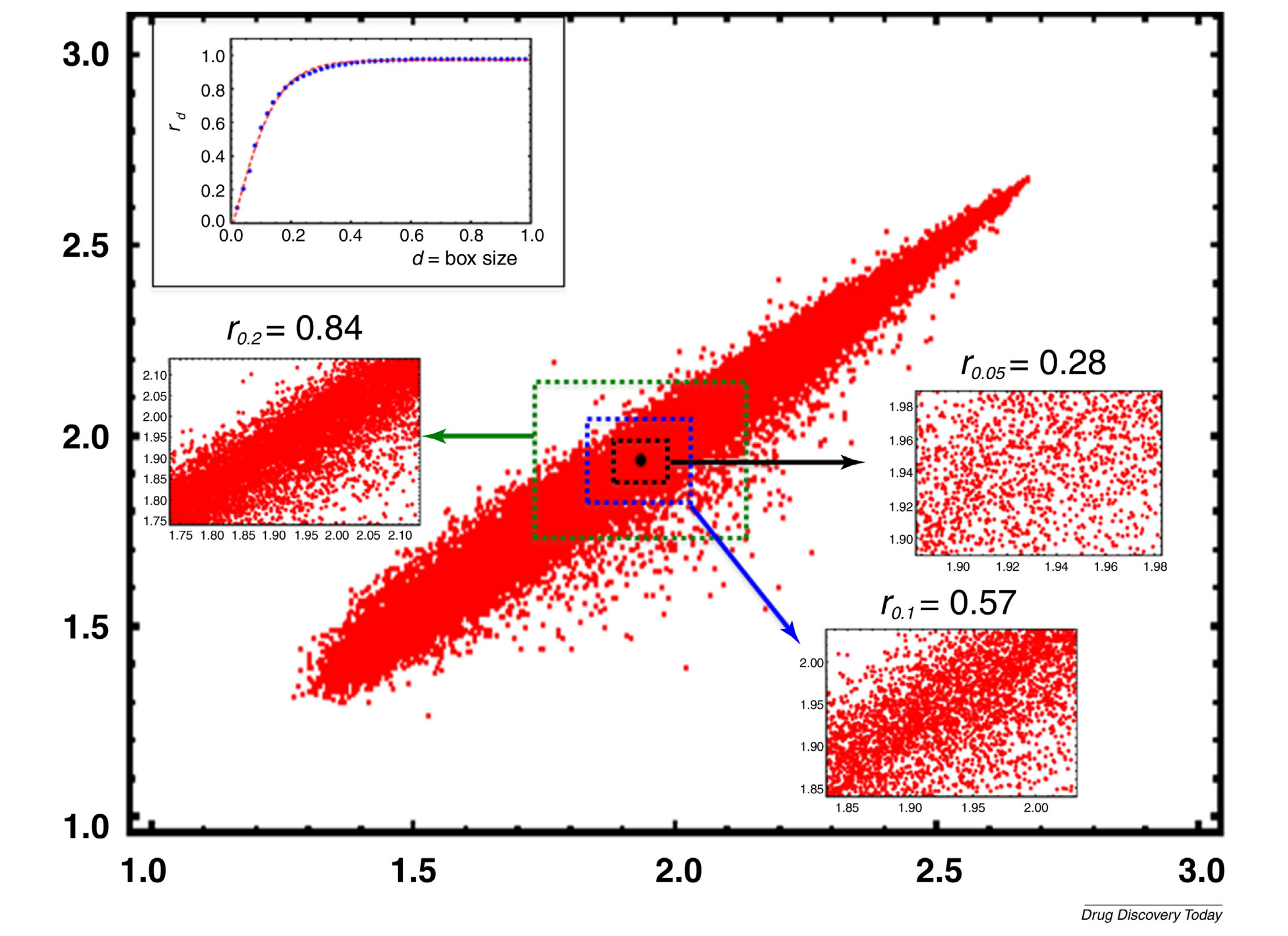

FIGURE 3

Genome regulation at large: scale dependency of correlation. The *X* and *Y* axes of the figure correspond to the whole genome expression profile of two independent macrophage cell populations. The approximately 25 000 vector points are the expression values (logarithmic units) of distinct genes simultaneously measured by an Affymetrix microarray platform [21]. The striking ($r = 0.98$) global correlation observed at the whole genome scale, drastically reduces when zooming on more restricted areas correspondent to the snapshots in the figure: r_d is the observed correlation at different values of *d* (box size of the gene selection expression range). The link between d and the observed Pearson correlation is reported in the top-left inset of the figure. The range restriction effect is a well-known property of correlation [22] that, in this particular case, is instrumental for selecting the minimum variability span where global genome control is detectable (onset of the correlation plateau in the top left graph).

This strategy worked remarkably well for many years then, almost abruptly, around the eighties of the last century, entered a deep crisis provoking the paradox of an exponential growth of basic knowledge going hand-in-hand with a drastic fall of newly marketed drugs. This crisis is very well described in [32], in which the authors sketch an approximate estimate of 76% of drugs discovered in the last twenty years referring to receptor molecules discovered around the fifties, while only the 6% bind to recently discovered targets; for the remnants no reasonable hypothesis of mechanism of action does hold.

The network pharmacology paradigm tries to overcome the above reductionist view posing that biological regulation, at any level, must be intended as a relational paradigm in which the observed action is the resultant behavior of a network of mutually interacting agents [33].

Network paradigm, instead of looking for strong and selective binders, focuses on candidate drugs weakly interacting with a multitude of different receptors. These ‘weak multiple binders’ are more efficient network modifiers than too selective agents [29–33].

Principal components correspond to the collective modes of network dynamics, this makes PCA the method of choice for both the prediction of ‘most influential nodes’ of a network [33] and to quantitatively summarize into a single score the results of candidate drugs across a battery of tests [34].

Order and organization

The degree of order of a system can be defined in many different ways, all the different definitions sharing the same basic concept: ‘The more ordered a system is, the more correlated are its constituent elements’ [35]. This feature is consistent with the Kolmogorov-Chaitin’s definition of algorithmic complexity [36]:

‘The information content or complexity of an object can be measured by the length of its shortest description. For instance the string ‘01010101010101010101010101010101’ has the short description '16 repetitions of 01', while '11001000011000011101111011101100' presumably has no simpler description other than writing down the string itself. More formally, the Algorithmic Complexity (AC) of a string x is defined as the length of the shortest program that computes or outputs x, where the program is run on some fixed reference universal computer'.

This formal definition of complexity, coming from information theory, can be easily translated in PCA terms: the more components are needed to get rid of a given system variance, the more complex the system is [35]. The formal link between the 'length of the shortest description' and PCA can be found in [37].

The amount of variance explained by a fixed number of components correlates with the 'amount of order' of the studied system and negatively scales with complexity [35,37]. The quantitation of the 'amount of order' of brain metabolism (in terms of between areas correlation estimated by PET or MRI) is widely used in neuroscience, allowing to predict clinically relevant outcomes as the probability of transition toward Alzheimer disease of a population of mild-cognitive impaired patients [38], or the response to treatment in acutely psychotic [39] and major depression [40] patients.

The use of emerging collective properties (as the amount of order) as primary variables allows for a robust approach to complex systems impossible to manage in terms of microscopic-level descriptors.

Conclusions

The Karl Pearson [3] 'least squares optimal' summary of multidimensional data, still occupies the frontiers of science. This is due by the fact the investigation of the correlation structure of the natural systems is an invariant of the scientific knowledge.

In these last years, the quest for 'unification' of natural sciences shifted from the search of fundamental laws (all the natural entities are made of the same matter) to the focus on systemic invariants (all the natural entities can be considered in terms of interactions among their parts). Scientists are discovering 'organization principles' shared by economical, ecological and physiological systems [41]. The present work, adopting a pharmacological perspective, tries to convey the important position PCA occupies in this emerging 'systems science'.

Appendix A. An intuitive geometrical look at PCA

We can imagine a multidimensional data set as a cloud of points (statistical units) embedded in a space spanned by n variables (the n descriptors attached to each statistical unit): the goal of PCA is to identify the 'directions' of the space where the elongation (inertia, variance) of points is maximal [6,8]. Fig. A1 reports a simple three-dimensional picture explaining this concept.

The cube is spanned by the original variables (panel a). PCA corresponds to the rotation of the reference set along the direction of maximal elongation of the data cloud that, keeping invariant the mutual positions of the points, projects them along three mutually orthogonal new axes (PC1, PC2, PC3).

PC1 corresponds to the maximal elongation direction, PC2 to the direction orthogonal to PC1 that explains the maximal variance and PC3 to the maximal variance direction orthogonal to both PC1 and PC2 (b). The first two components account for the by far major part of the information present in the original data set, consequently, the projection of the points (statistical units) on the PC1 vs. PC2 plane, allows for an exhaustive summary of the original three-dimensional information (c). After this procedure, each statistical unit is defined by two coordinates (PC1 and PC2 scores) correspondent to two mutually independent descriptors that are the 'hidden variables' at the basis of the directly measured descriptors.

PCA can be thought as the fitting of an n-dimensional ellipsoid to the data, where each axis of the ellipsoid represents a principal component. To find the axes of the ellipse, we must first subtract the mean of each variable from the dataset to center the data cloud on the origin. Then, we compute the covariance matrix of the data set, and calculate its eigenvalues and corresponding eigenvectors. The next steps are the orthogonalization and normalization of the set of eigenvectors to become unit vectors. Once this is done, each of the mutually orthogonal unit eigenvectors can be interpreted as an axis of the ellipsoid fitted to the data. The proportion of the variance each eigenvector accounts for can be calculated by dividing the eigenvalue corresponding to that eigenvector by the sum of all eigenvalues (see also: https://en.wikipedia.org/wiki/Principal_component_analysis).

In biomedical applications the original variables are often expressed in different measurement units (time, space,

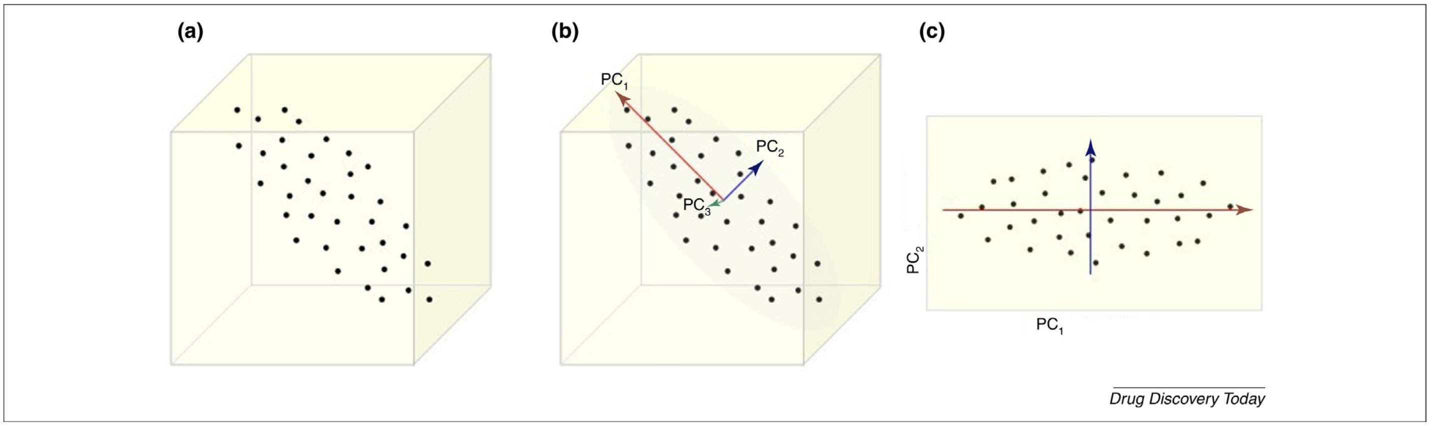

FIGURE A1

PCA: a geometrical view. The figure reports a geometrical sketch of PCA. Panel (a) depicts a three-dimensional data cloud, the dimensions of the box correspond to the original variables. In Panel (b) the three principal components of the data set are drawn: the length of the corresponding vectors is proportional to the variance explained by each principal component (PC). In panel (c) the statistical units are projected in the space spanned by the two major components.

concentrations.) thus, after being subtracted of the mean, the original variables are divided by their standard deviation so to work with z-scores. The use of z-scores corresponds to the extraction of the eigenvalues and eigenvectors of the original data correlation instead of covariance matrix (correlations correspond to the covariances of standardized variables).

All the commercial statistical software packages (SAS, Statistica, SPSS, StatView) do have routines for PCA computation, the same holds true for general purpose software packages like MATLAB.

R open-source freeware has many routines to compute PCA and cognate techniques like SVD, SSA, Factor Analysis.

References

1 Preisendorfer, R.W. (1988) In *Principal component analysis in meteorology and oceanography*, (vol. 425) (Mobley, C.D., ed.), Elsevier, Amsterdam

2 Beltrami, E. (1873) Sulle funzioni bilineari. *Giornale Mat. Uso degli Stud. Univ.* 11 (2), 98–106

3 Pearson, K. (1901) On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, Dublin Philos. Mag. J. Sci.* 2 (11), 559–572

4 (1993) Chemical pattern recognition and multivariate analysis for QSAR studies. *TrAC Trends Anal. Chem.* 12 (2), 50–60

5 Eriksson, L. *et al.* (2013) *Multi- and megavariate data analysis basic principles and applications*. Umetrics Academy, Umea

6 Christie, O.H. (1995) Introduction to multivariate methodology, an alternative way? *Chemomet. Intell. Lab. Syst.* 29 (2), 177–188

7 Ghil, M. *et al.* (2002) Advanced spectral methods for climatic time series. *Rev. Geophys.* 40 (1), 3–40

8 Lewi, P.J. (1995) Pattern recognition, reflections from a chemometric point of view. *Chemomet. Intell. Lab. Syst.* 28 (1), 23–33

9 Hwang, J. *et al.* (2013) Fast and sensitive recognition of various explosive compounds using Raman spectroscopy and principal component analysis. *J. Mol. Struct.* (1039), 130–136

10 Langfelder, P. and Horvath, S. (2007) Eigengene networks for studying the relationships between co-expression modules. *BMC Systems Biol.* 1 (1), 54

11 Keun, H.C. (2006) Metabonomic modeling of drug toxicity. *Pharmacol. Therapeut.* 109 (1), 92–106

12 Nicholson, J.K. *et al.* (2002) Metabonomics: a platform for studying drug toxicity and gene function. *Nat. Rev. Drug Discov.* 1 (2), 153–161

13 Zuena, A.R. *et al.* (2016) Maternal exposure to environmental enrichment before and during gestation influences behaviour of rat offspring in a sex-specific manner. *Physiol. Behav.* 163, 274–287

14 Bro, R. and Smilde, A.K. (2014) Principal component analysis. *Anal. Methods* 6 (9), 2812–2831

15 Aste, T. and Di Matteo, T. (2006) Dynamical networks from correlations. *Phys. A: Stat. Mech. Appl.* 370 (1), 156–161

16 De la Fuente, A. *et al.* (2004) Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics* 20, 3565–3574

17 Keun, H.C. *et al.* (2004) Geometric trajectory analysis of metabolic responses to toxicity can define treatment specific profiles. *Chem. Res. Toxicol.* 17 (5), 579–587

18 Yi, Z.B. *et al.* (2007) Evaluation of the antimicrobial mode of berberine by LC/ESI-MS combined with principal component analysis. *J. Pharmaceut. Biomed. Anal.* 44, 301–304

19 Huang, S. (2004) Back to the biology in systems biology: what can we learn from biomolecular networks? *Brief. Funct. Genom. Proteom.* 2 (4), 279–297

20 Weaver, W. (1948) Science and complexity. *Am. Scient.* 36, 536–549

21 Tsuchiya, M. *et al.* (2016) Self-organizing global gene expression regulated through criticality: mechanism of the cell-fate change. *PLoS ONE* 11 (12), e0167912http://dx.doi.org/10.1371/journal.pone.0167912

22 Giuliani, A. *et al.* (2004) Invariant features of metabolic networks: a data analysis application on scaling properties of biochemical pathways. *Phys. A: Stat. Mech. Appl.* 337 (1), 157–170

23 Transtrum, M.K. *et al.* (2015) Perspective: sloppiness and emergent theories in physics, biology and beyond. *J. Chem. Phys.* 143, 01091

24 Joyner, M. *et al.* (2016) What happens when underperforming big ideas in research become entrenched? *JAMA* http://dx.doi.org/10.1001/jama.2016.11076

25 Tellegen, B. (1952) A general network theorem with application. *Phillips Res. Rep.* 7, 259–269

26 Mickulecki, D. (2001) Network thermodynamics and complexity: a transition to relational systems theory. *Comput. Chem.* 25, 369–391

27 Tasdighian, S. *et al.* (2013) Modules identification in protein structures: the topological and geometrical solutions. *J. Chem. Inform. Model* 54, 159–168

28 Price, N.D. *et al.* (2003) Analysis of metabolic capabilities using singular value decomposition of extreme pathway matrices. *Biophys. J.* 84, 794–804

29 Hopkins, A.L. (2008) Network pharmacology: the next paradigm in drug discovery. *Nat. Chem. Biol.* 4 (11), 682–690

30 Ligeti, B. *et al.* (2015) A network-based target overlap score for characterizing drug combinations: high correlation with cancer clinical trial results. *PLoS ONE* 10, e0129267

31 Csermely, P. *et al.* (2005) The efficiency of multi-target drugs: the network approach might help drug design. *TiPS Trends Pharmacol. Sci.* 26, 178–182

32 Overington, J.P. *et al.* (2006) How many drug targets are there? *Nat. Rev. Drug Discov.* 5 (12), 993–996

33 Csermely, P. *et al.* (2013) Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. *Pharmacol. Therapeut.* 138, 333–408

34 Taylor, G. *et al.* (2015) Efficacy of a virus-vectored vaccine against human and bovine respiratory syncytial virus infections. *Sci. Transl. Med.* 7 300ra127-300ra127

35 Giuliani, A. *et al.* (2001) A complexity score derived from principal components analysis of nonlinear order measures. *Phys. A: Stat. Mech. Appl.* 301, 567–588

36 Li, M. and Vitányi, P. (2013) *An introduction to Kolmogorov complexity and its applications*. Springer Science & Business Media

37 Soofi, E. (1994) Capturing the intangible concept of information. *J. Am. Stat. Ass.* 89 (428), 1243–1254

38 Pagani, M. *et al.* (2016) Predicting the transition from normal aging to Alzheimer's disease: a statistical mechanistic evaluation of FDG-PET data. *NeuroImage* 141, 282–290

39 Sarpal, D.K. *et al.* (2015) Baseline striatal functional connectivity as a predictor of response to antipsychotic drug treatment. *Am. J. Psychiatry* 173, 69–77

40 Salomons, T.V. *et al.* (2014) Resting-state cortico-thalamic-striatal connectivity predicts response to dorsomedial prefrontal rTMS in major depressive disorder. *Neuropsychopharmacology* 39 (2), 488–498

41 Gorban, A.N. *et al.* (2010) Correlations, risk and crisis: from physiology to finance. *Phys. A: Stat. Mech. Appl.* 389 (16), 3193–3217