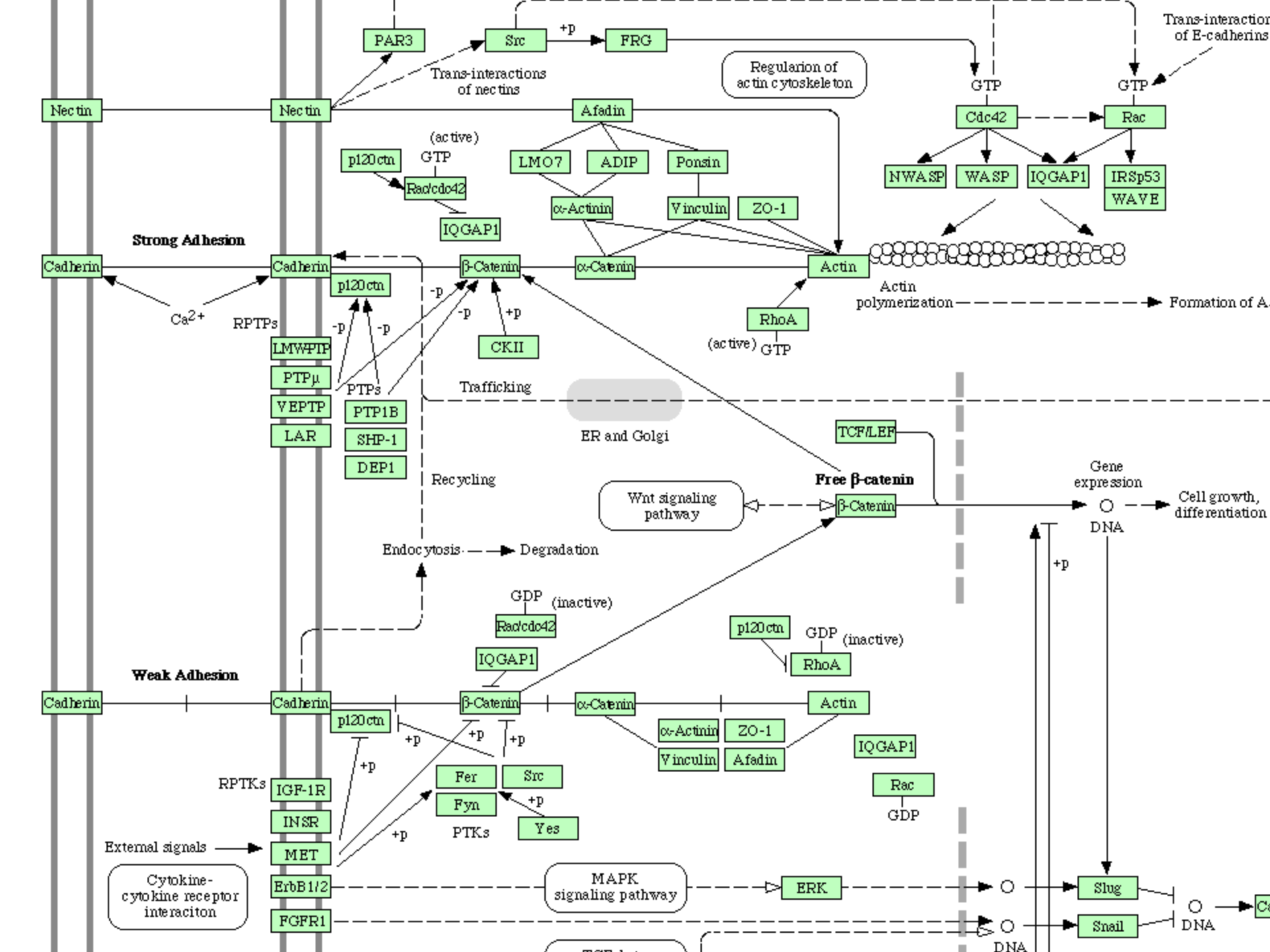
Analisi di Correlazioni

Alessandro Giuliani


PAR3
Src
+p
FRG
Trans-interaction of E-cadherins
Trans-interactions of nectins
Regularion of actin cytoskeleton
GTP
GTP
Nectin
Nectin
Afadin
Cdc42
Rac
(active)
p120ctn
GTP
Rac/cdc42
LMO7
ADIP
Ponsin
NWASP
WASP
IQGAP1
IRSp53
WAVE
IQGAP1
α-Actinin
Vinculin
ZO-1
Strong Adhesion
Cadherin
Cadherin
p120ctn
β-Catenin
α-Catenin
Actin
Actin polymerization
Formation of A
Ca2+
RPTPs
-p
-p
-p
-p
+p
CKII
RhoA
(active)
GTP
LMW-PTP
PTPμ
VEPTP
LAR
PTPs
PTP1B
SHP-1
DEP1
Trafficking
ER and Golgi
TCF/LEF
Recycling
Free β-catenin
Wnt signaling pathway
β-Catenin
Gene expression
DNA
Cell growth, differentiation
Endocytosis
Degradation
+p
GDP
(inactive)
Rac/cdc42
IQGAP1
p120ctn
GDP
(inactive)
RhoA
Weak Adhesion
Cadherin
Cadherin
β-Catenin
α-Catenin
Actin
p120ctn
+p
+p
+p
α-Actinin
ZO-1
Vinculin
Afadin
IQGAP1
Rac
GDP
RPTKs
IGF-1R
INSR
MET
Fer
Src
Fyn
+p
Yes
PTKs
+p
+p
External signals
Cytokine-cytokine receptor interaciton
ErbB1/2
FGFR1
MAPK signaling pathway
ERK
Slug
Snail
DNA
DNA

Del Rigore della Scienza

... In quell'impero, l'Arte della Cartografia raggiunse una tale Perfezione che la mappa di una sola provincia occupava tutta una Città e la mappa dell'Impero tutta una Provincia. Col tempo codeste Mappe Smisurate non soddisfecero e i Collegi dei Cartografi eressero una mappa dell'Impero che uguagliava in grandezza l'Impero e coincideva puntualmente con esso. Meno Dedite allo studio della cartografia, le Generazioni Successive compresero che quella vasta Mappa era inutile e non senza Empietà la abbandonarono all'Inclemenze del Sole e degl'Inverni. Nei deserti dell'Ovest rimangono lacere rovine della mappa, abitate da Animali e Mendichi; in tutto il paese non è altra reliquia delle Discipline Geografiche. (Suarez Miranda, Viaggi di uomini prudenti, libro quarto, cap. XLV, Lérida, 1658)

Da Jorge Luis Borges, *L'artefice* Ed. Mondadori i Meridiani vol. 1, pg. 1253

nature human behaviour

PERSPECTIVE

PUBLISHED: 10 JANUARY 2017 | VOLUME: 1 | ARTICLE NUMBER: 0021

OPEN

A manifesto for reproducible science

Marcus R. Munafò[1,2*], Brian A. Nosek[3,4], Dorothy V. M. Bishop[5], Katherine S. Button[6], Christopher D. Chambers[7], Nathalie Percie du Sert[8], Uri Simonsohn[9], Eric-Jan Wagenmakers[10], Jennifer J. Ware[11] and John P. A. Ioannidis[12,13,14]

Improving the reliability and efficiency of scientific research will increase the credibility of the published scientific literature and accelerate discovery. Here we argue for the adoption of measures to optimize key elements of the scientific process: methods, reporting and dissemination, reproducibility, evaluation and incentives. There is some evidence from both simulations and empirical studies supporting the likely effectiveness of these measures, but their broad adoption by researchers, institutions, funders and journals will require iterative evaluation and improvement. We discuss the goals of these measures, and how they can be implemented, in the hope that this will facilitate action toward improving the transparency, reproducibility and efficiency of scientific research.

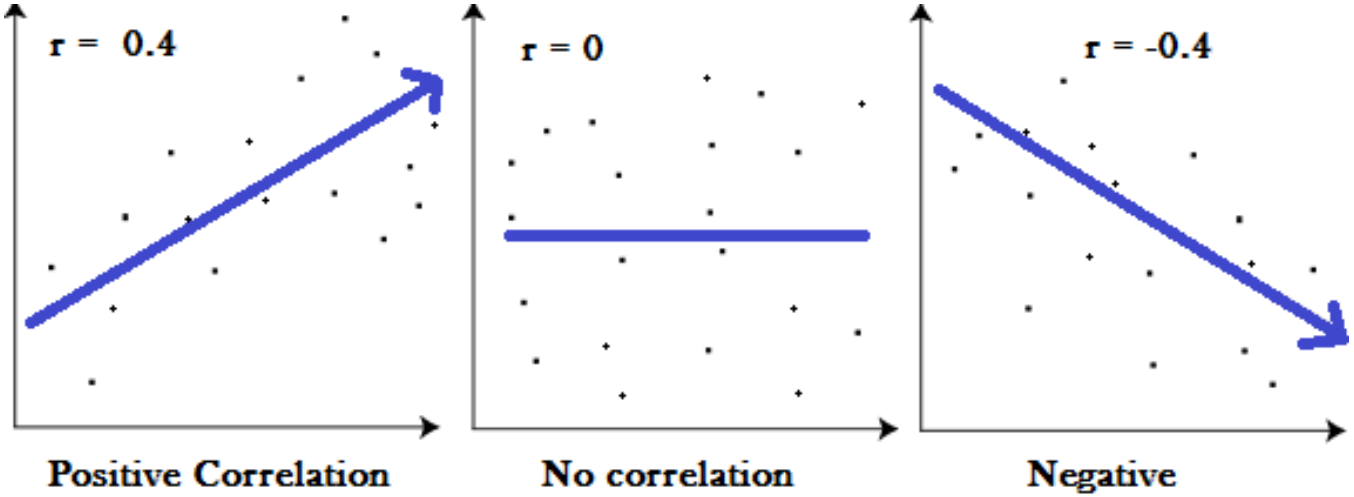
THE JOURNAL OF CHEMICAL PHYSICS **143**, 010901 (2015)

Perspective: Sloppiness and emergent theories in physics, biology, and beyond

Mark K. Transtrum,[1] Benjamin B. Machta,[2] Kevin S. Brown,[3,4] Bryan C. Daniels,[5] Christopher R. Myers,[6,7] and James P. Sethna[6]
[1]*Department of Physics and Astronomy, Brigham Young University, Provo, Utah 84602, USA*
[2]*Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, New Jersey 08544, USA*
[3]*Departments of Biomedical Engineering, Physics, Chemical and Biomolecular Engineering, and Marine Sciences, University of Connecticut, Storrs, Connecticut 06269, USA*
[4]*Institute for Systems Genomics, University of Connecticut, Storrs, Connecticut 06030-1912, USA*
[5]*Center for Complexity and Collective Computation, Wisconsin Institute for Discovery, University of Wisconsin, Madison, Wisconsin 53715, USA*
[6]*Laboratory of Atomic and Solid State Physics, Cornell University, Ithaca, New York 14853, USA*
[7]*Institute of Biotechnology, Cornell University, Ithaca, New York 14853, USA*

(Received 2 February 2015; accepted 4 June 2015; published online 1 July 2015)

As a young physicist, Dyson paid a visit to Enrico Fermi[1] (recounted in Ditley, Mayer, and Loew[2]). Dyson wanted to tell Fermi about a set of calculations that he was quite excited about. Fermi asked Dyson how many parameters needed to be tuned in the theory to match experimental data. When Dyson replied there were four, Fermi shared with Dyson a favorite adage of his that he had learned from Von Neumann: "with four parameters I can fit an elephant, and with five I can make him wiggle his trunk." Dejected, Dyson took the next bus back to Ithaca.



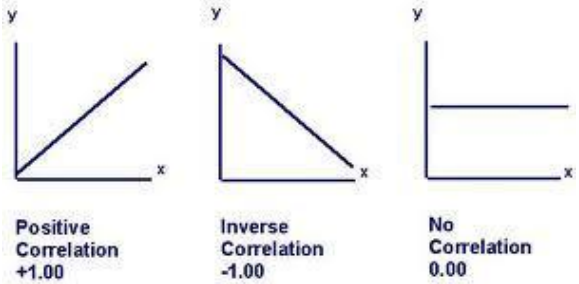
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

+ * + → +

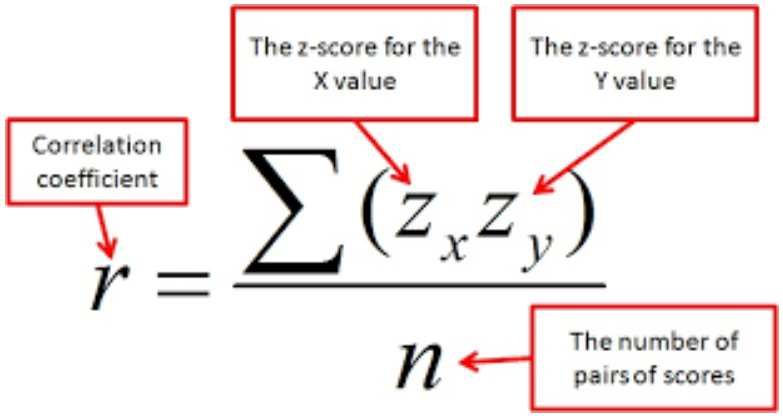
+ * - → -

- * - → +

- * + → -

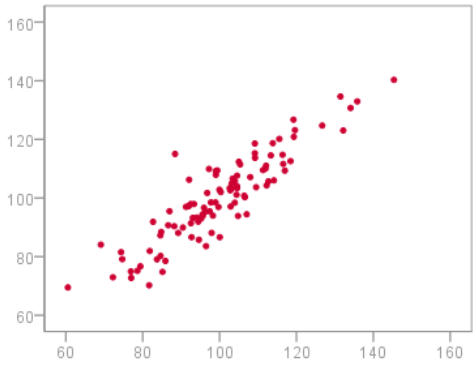


Correlation coefficient variation

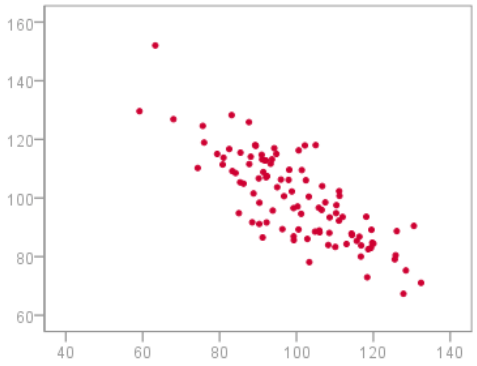


$$r = r_{xy} = \frac{\mathrm{Cov}(x, y)}{S_x \times S_y}$$

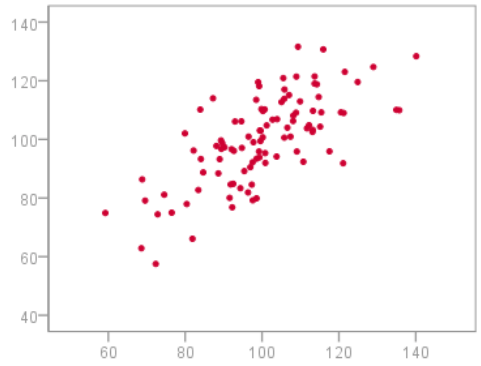
(PEARSON) CORRELATIONS VISUALIZED AS SCATTERPLOTS



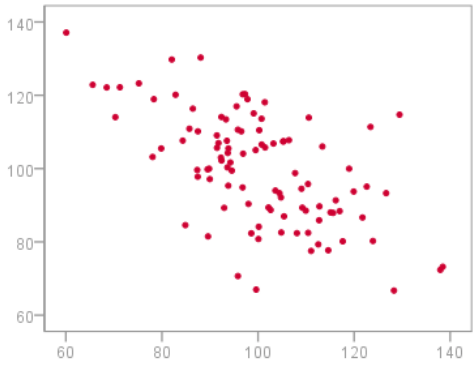
r = 0.9



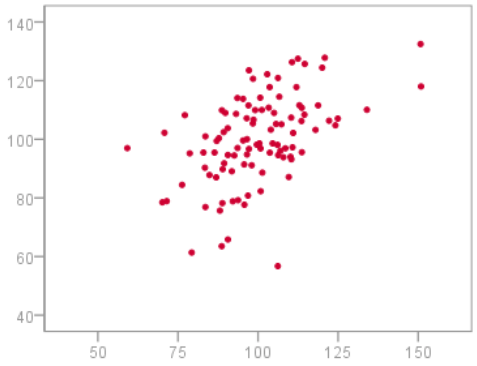
r = -0.8



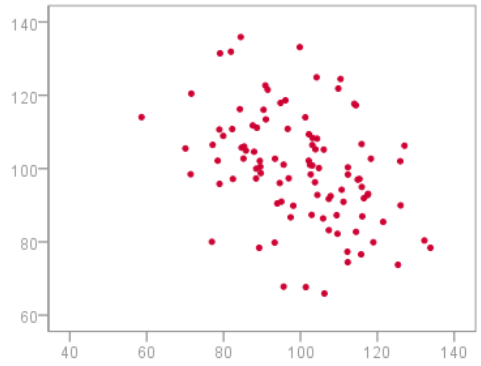
r = 0.7



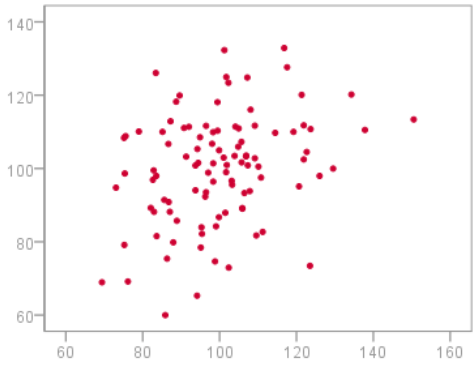
r = -0.6



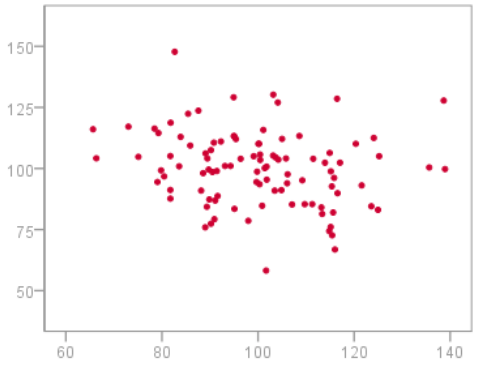
r = 0.5



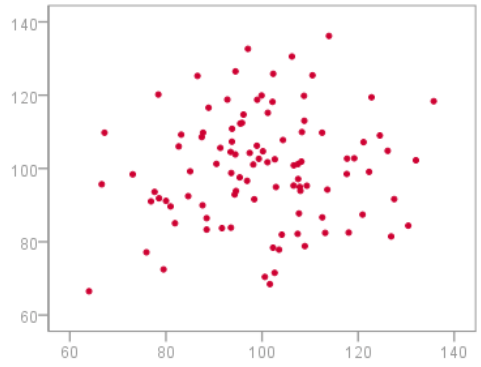
r = -0.4



r = 0.3



r = -0.2



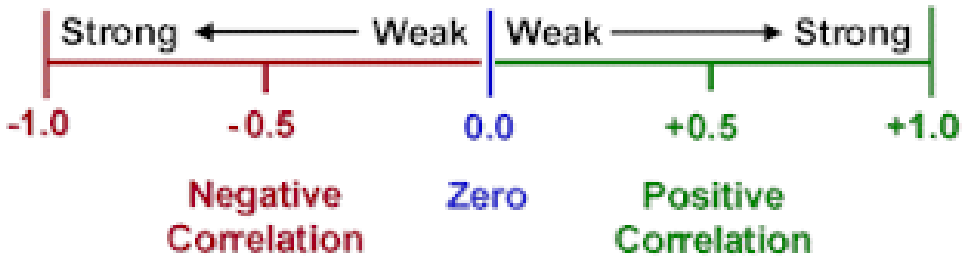
r = 0.1

Going from univariate to bivariate a quantum-leap in our view of the data happens: order (in both space and time) matters.

A	B	C
10	40	30
20	30	20
30	20	10
40	10	40

All the three vectors have identical statistical descriptors, but while A and B are negatively correlated A –C and B-C are each other independent

Correlation Coefficient
Shows Strength & Direction of Correlation



Correlation Coefficient (r)

- Correlation Coefficient (r) is a measure of association between two variables
- Varies from -1 to +1
- r is a ratio of variability in X to that of Y.

0 = no relationship;

1 = perfect relationship

$$r = \frac{\sum x_1 x_2}{\sqrt{(\sum x_1^{\,2})(\sum x_2^{\,2})}}$$

Detecting a correlation between two X and Y variables means that the knowledge of the value of X decreases the uncertainty about the corresponding value of Y.

This is a necessary (but not sufficient) condition for a causal link between X and Y.

Published online: September 10, 2015

Science & Society

EMBO reports

Could Big Data be the end of theory in science?

A few remarks on the epistemology of data-driven science

Fulvio Mazzocchi

The Deluge of Spurious Correlations in Big Data

Cristian Calude, Giuseppe Longo

Foundations of Science, pp. 1-18, March, 2016

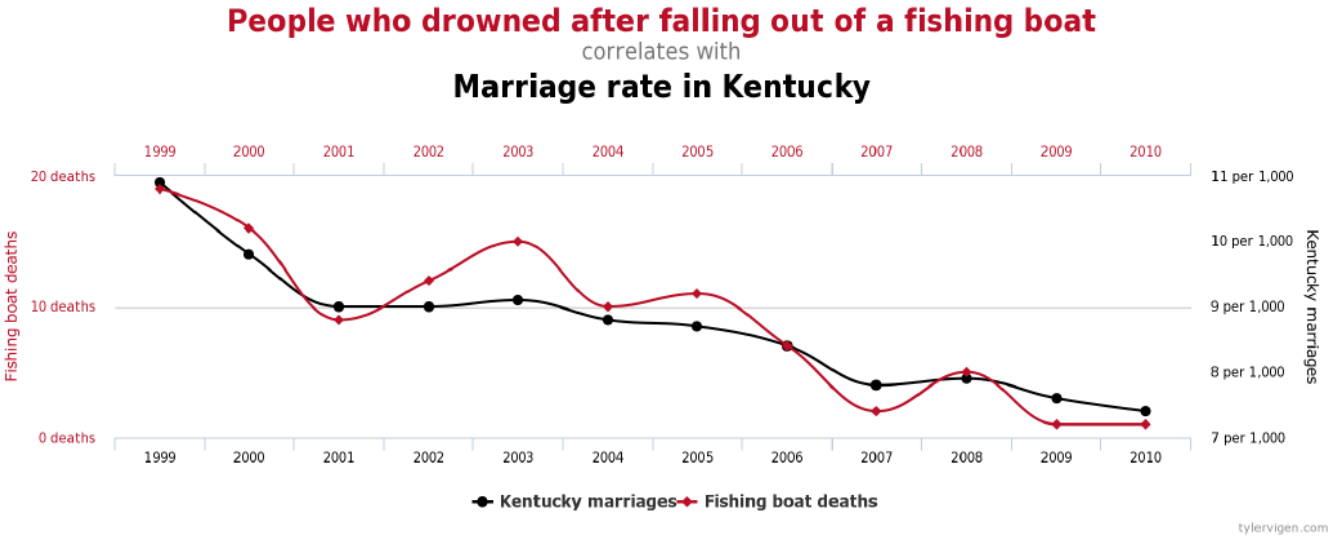
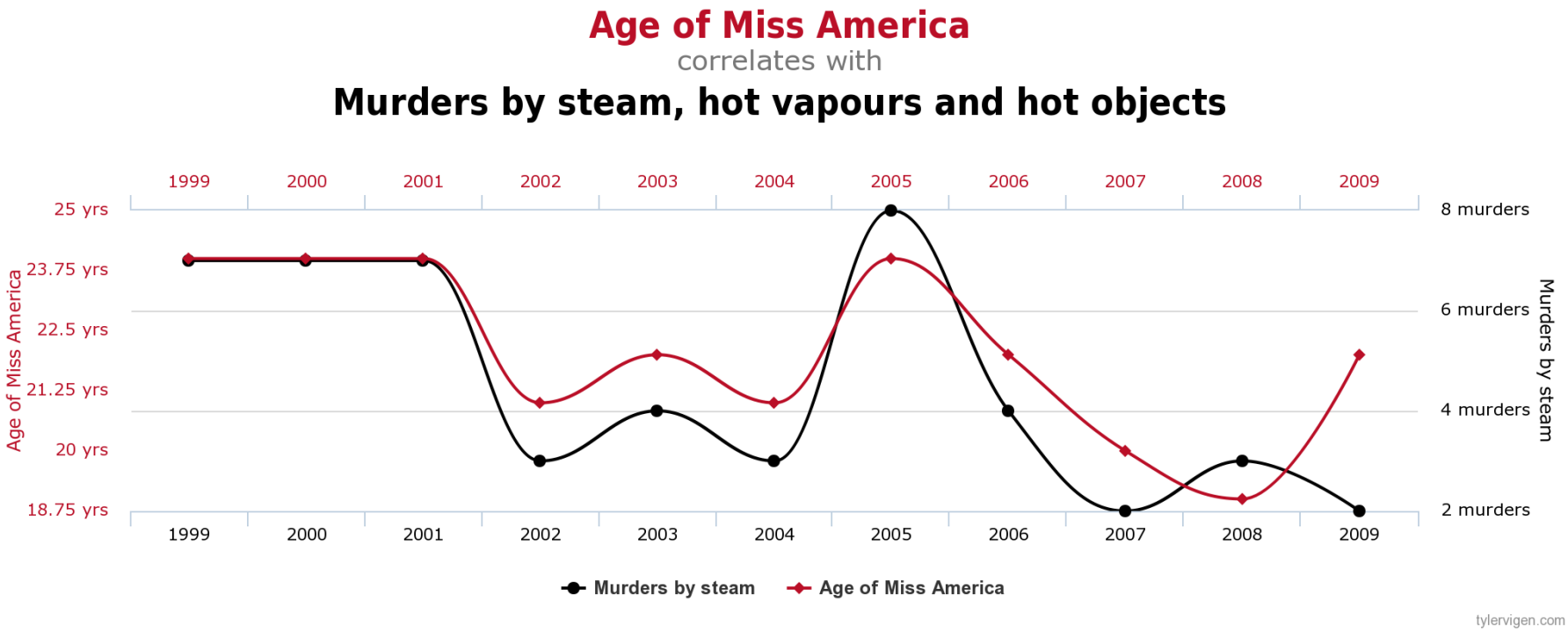
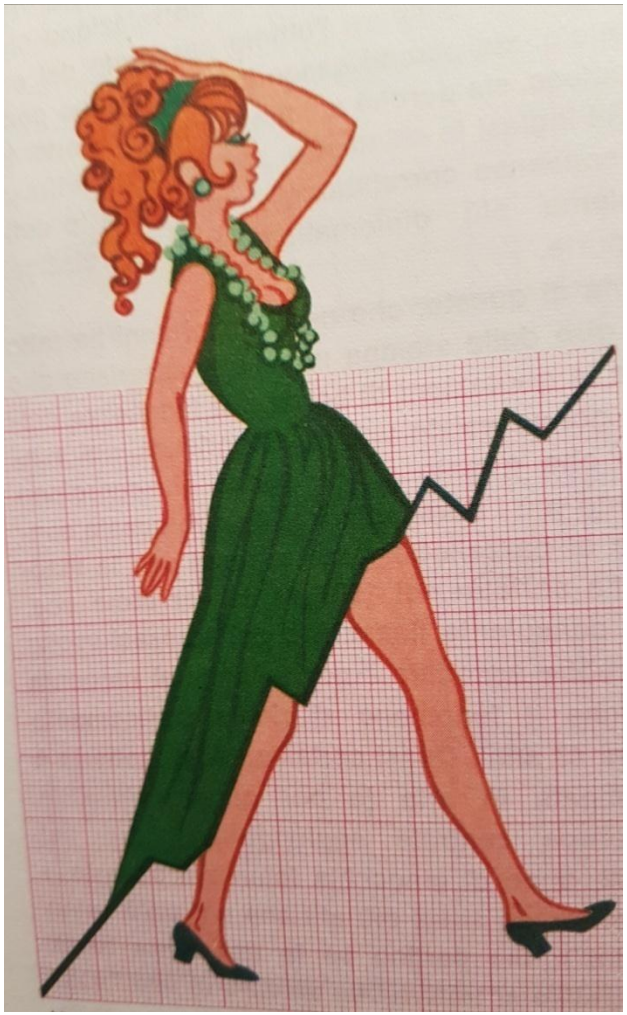


Figure 1: A correlation with $r = 0.952407$ [2].

http://www.tylervigen.com/spurious-correlations



Correlazioni spurie



La correlazione (spuria) tra l'accorciamento delle gonne e congiuntura economica positiva è stata presa sul serio da molti sociologi

Barber, N. (1999). Women's dress fashions as a function of reproductive strategy. *Sex Roles*, *40*(5), 459-471.

Docherty, C. A., & Hann, M. A. (1994). Stylistic Change in Womenswear Products Part II: The Relationship Between Hem Length and Various Economic Indicators. *Journal of the Textile Institute*, *85*(2), 283-287.

Detecting the optimal scale for the analysis is the most crucial problem in science.

Environmental Practice 16: 281–286 (2014)

Defining Appropriate Spatial and Temporal Scales for Ecological Impact Analysis[1]

Harriet L. Nash

Levin argues that scale is "the fundamental conceptual problem in ecology, if not in all of science" (Levin, 1992,

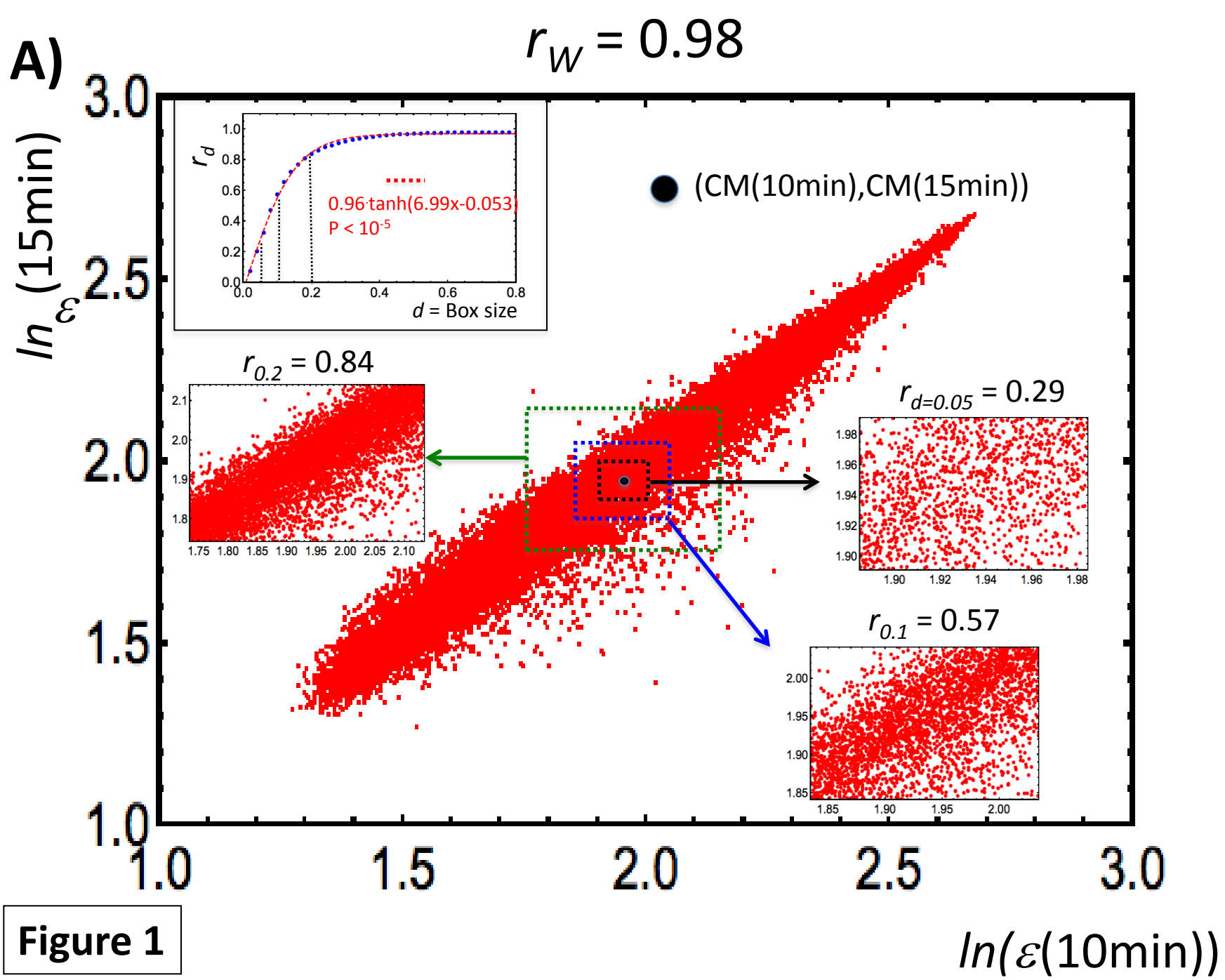
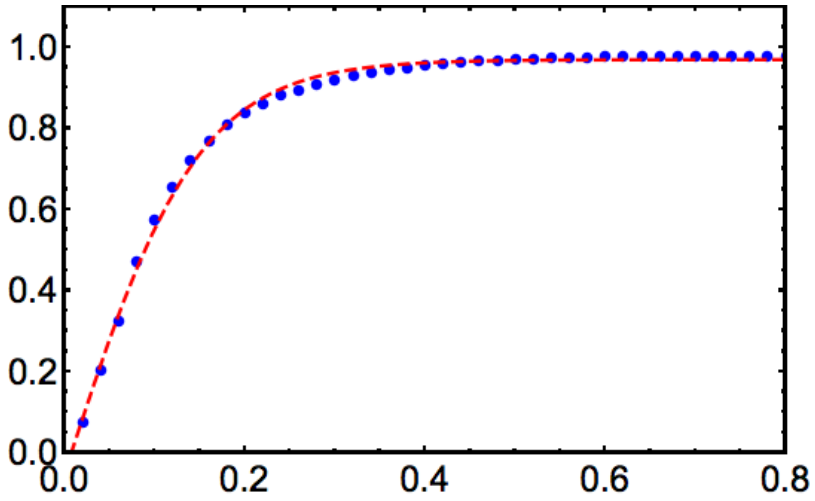

A)
r_W = 0.98
ln_ε(15min)
$ln(\varepsilon$(10min))
(CM(10min),CM(15min))
r_d
0.96·tanh(6.99x-0.053)
P < 10^{-5}
d = Box size
$r_{0.2}$ = 0.84
$r_{d=0.05}$ = 0.29
$r_{0.1}$ = 0.57

Figure 1

Ecology

Cell Biology



d = Box size

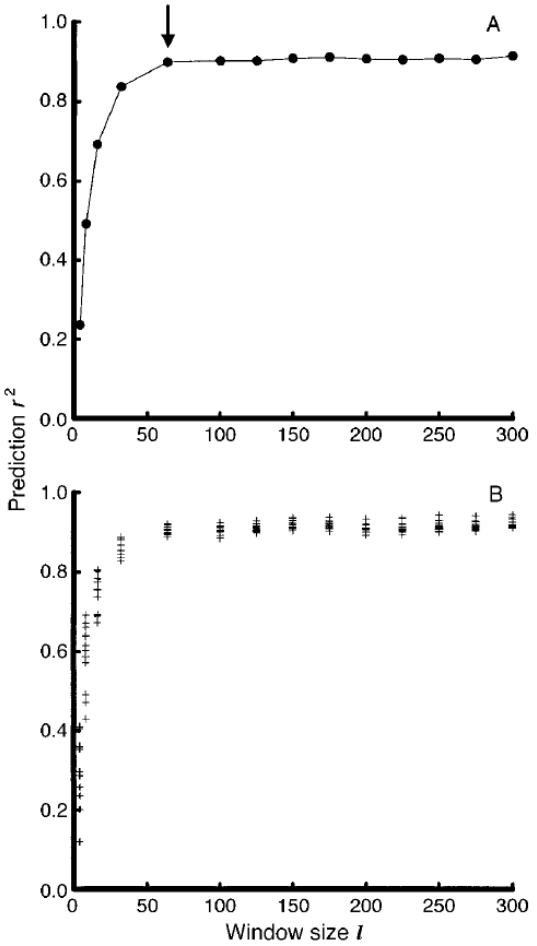


FIG. 3. Prediction $r^2 = 1 - E_l^2/\sigma_l^2$ as a function of window size l. The analyzed time series of prey densities contains 5000 data points, sampled at each time unit after transients have died out. The lag $\tau = 11$, and the prediction horizon $h = \tau$. The value $\tau = 11$ is chosen as the lag at which the autocorrelation function first crosses zero (this lag falls in the range 10–11 for different window sizes). Similar results were obtained for smaller values $\tau = 3$ and $\tau = 7$. (A) Embedding dimension $d = 5$; number of neighbors $k = 3$. (B) Embedding dimensions $d = 5, 7, 9$, and 11; for each d, a range of k values ($k = d - 2, d, d + 2$) was used.

The two graphs have corresponding axes.

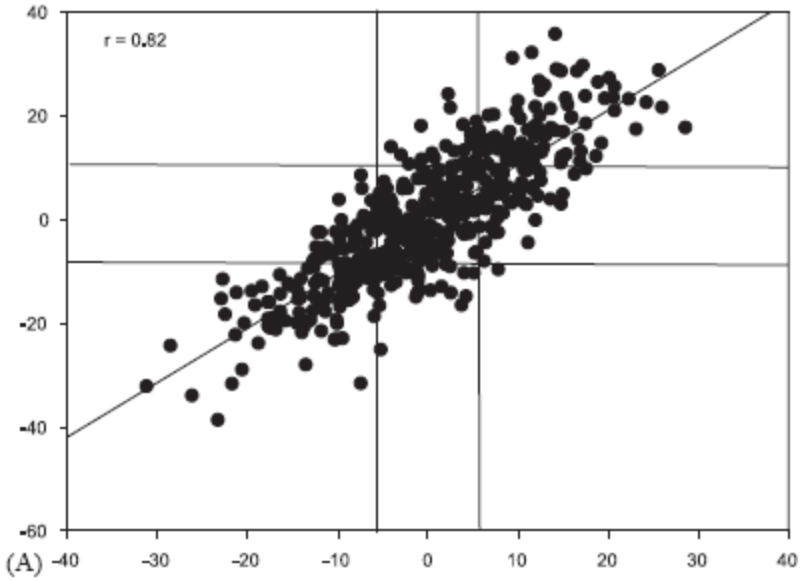
X axes: range of gene expression values (cell biology)
range of sampled territory in a sample. (ecology)

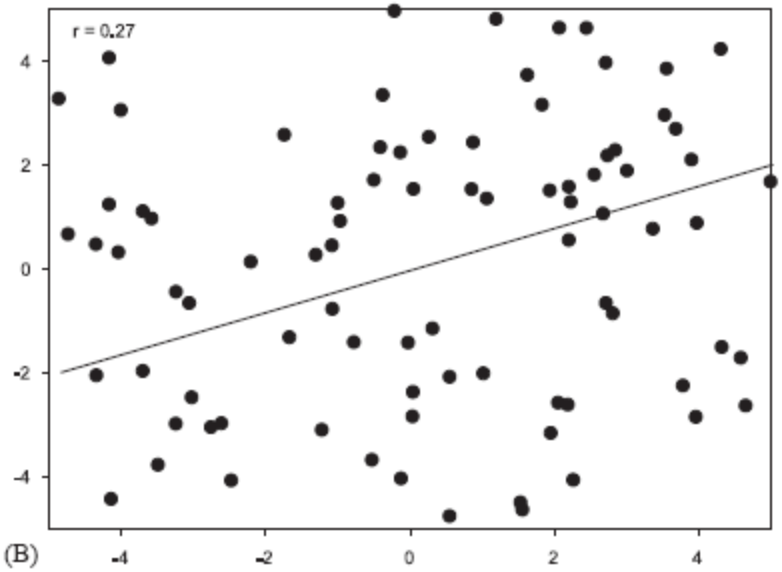
Y axes: mutual correlation between gene expression profiles (cell biology)
mutual correlation between species distribution (ecology)

Il modello:

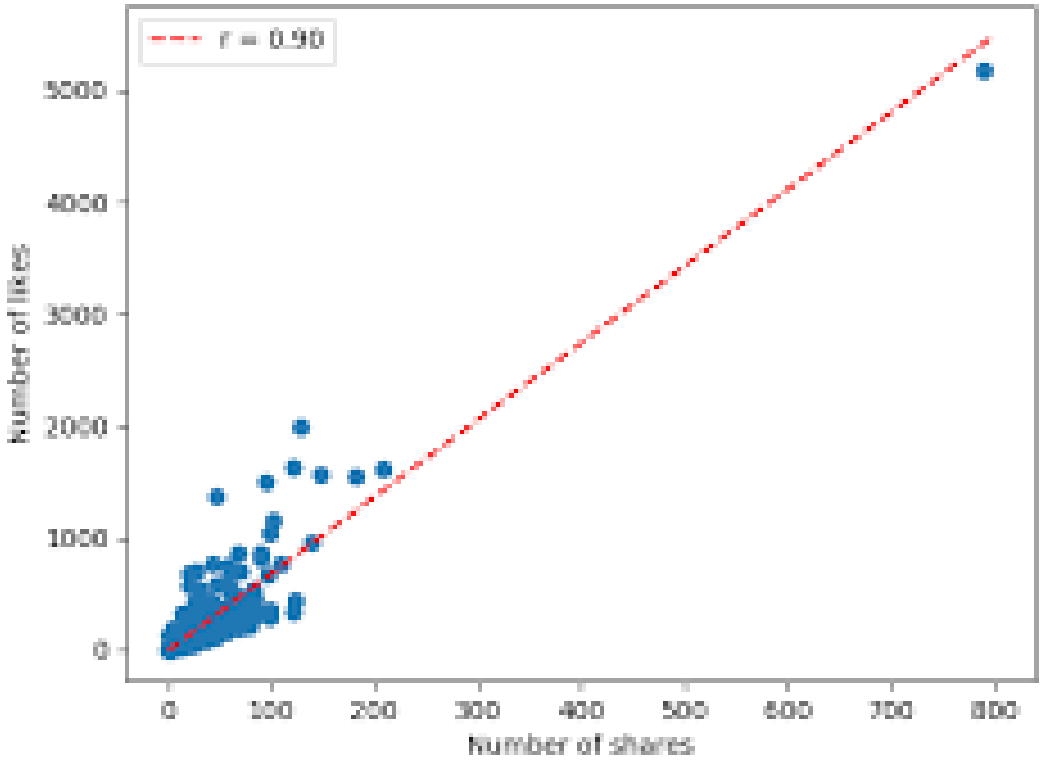
$$X = N(0, 10) \ ,$$

$$Y = X + N(0, 7) \ ,$$





A single outlier can create a spurious correlation (and consequently a spurious linear model, by constraining the model to be a ‘two-points’ regression (for each two points a straight line can be exactly drawn)



Chance Correlations in Structure–Activity Studies Using Multiple Regression Analysis

John G. Topliss* and Robert J. Costello

Departments of Medicinal Chemistry and Computer and Scientific Information, Research Division, Schering Corporation, Bloomfield, New Jersey 07003. *Received February 17, 1972*

Table IV. Relationship, for 30 Variables, between Number of Observations, r^2, and Average Number of Variables Included

No. of observations	r^2	Average no. of variables included
20	0.80	6.30
30	0.68	4.80
60	0.48	3.70
120	0.32	3.60
240	0.23	3.00

Figure 1. The relationship between r^2 and the number of observations.

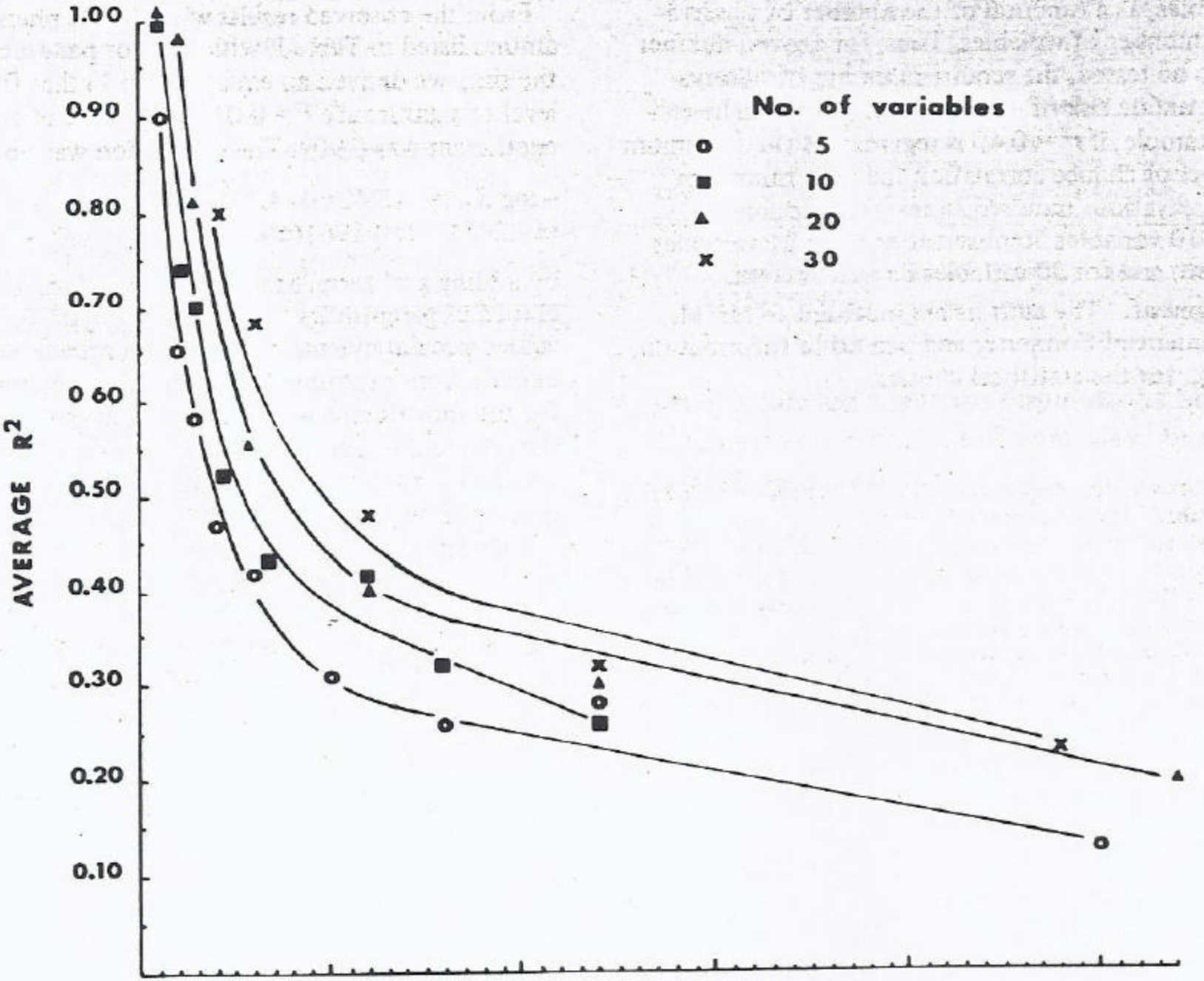
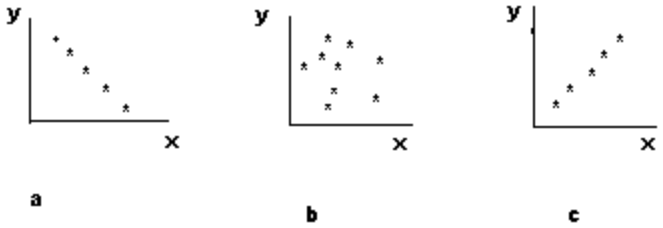


Table III Descriptive statistics and correlation matrix for study variables – correlation matrix

	MO	FP	MP	IM	IC	FM	FE	FI	SPC	DSC	DST
MO	1.00										
FP	0.31[a]	1.00									
MP	0.32[a]	0.71[a]	1.00								
IM	0.36[a]	0.12[c]	0.14[c]	1.00							
IC	0.39[a]	0.18[b]	0.21[a]	0.62[a]	1.00						
FM	0.26[a]	0.21[a]	0.14[c]	0.30[a]	0.27[a]	1.00					
FE	0.47[a]	0.21[a]	0.18[b]	0.38[a]	0.28[a]	0.24[a]	1.00				
FI	0.53[a]	0.26[a]	0.22[a]	0.36[a]	0.37[a]	0.29[a]	0.47[a]	1.00			
SPC	0.32[a]	0.22[a]	0.31[a]	0.51[a]	0.47[a]	0.32[a]	0.37[a]	0.35[a]	1.00		
DSC	−0.12[c]	0.03[c]	0.05[c]	0.17[b]	0.08[c]	0.18[b]	−0.05[c]	0.06[c]	0.01[c]	1.00	
DST	−0.02[c]	−0.01[c]	0.05[c]	0.24[a]	0.14[c]	0.05[c]	−0.05[c]	0.05[c]	0.05[c]	0.56[a]	1.00
DM	0.05[c]	0.144	0.136[c]	0.199[a]	0.169[b]	0.247[a]	0.08[c]	0.11[c]	0.14[c]	0.46[a]	0.71[a]

Notes: [a] correlation is significant at the 0.01 level (two-tailed); [b] correlation is significant at the 0.05 level (two-tailed); [c] correlation is non-significant



$$r = \frac{\langle xy \rangle - \langle x \rangle \langle y \rangle}{\sqrt{\langle (x_i - \langle x \rangle)^2 \rangle} \sqrt{\langle (y_i - \langle y \rangle)^2 \rangle}}$$

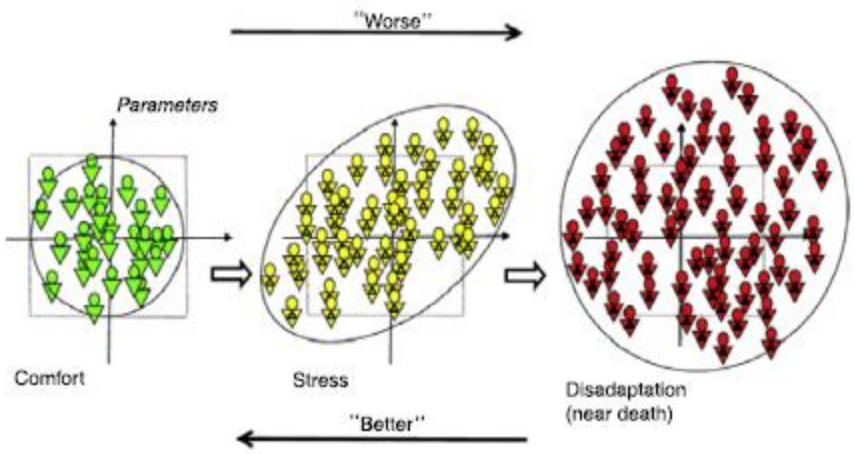
Pearson correlation coefficient is the basic metrics for approaching complexity

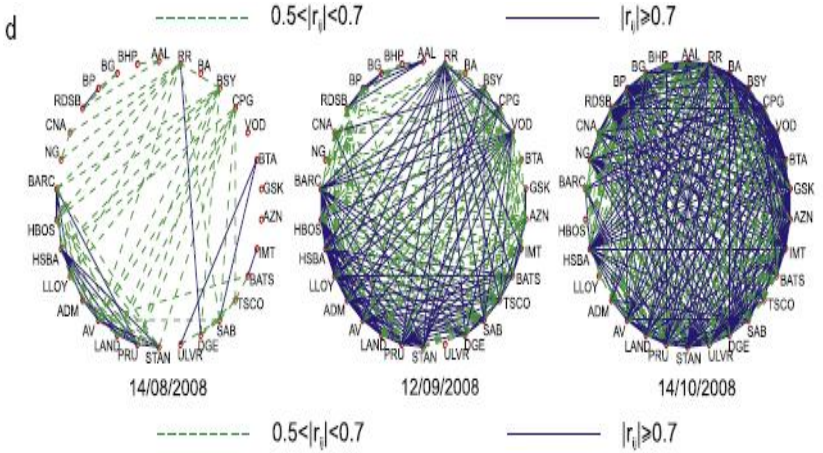
Physica A 389 (2010) 3193–3217

Correlations, risk and crisis: From physiology to finance

Alexander N. Gorban [a,*], Elena V. Smirnova [b], Tatiana A. Tyukina [a]

[a] University of Leicester, Leicester, LE1 7RH, UK
[b] Siberian Federal University, Krasnoyarsk, 660041, Russia

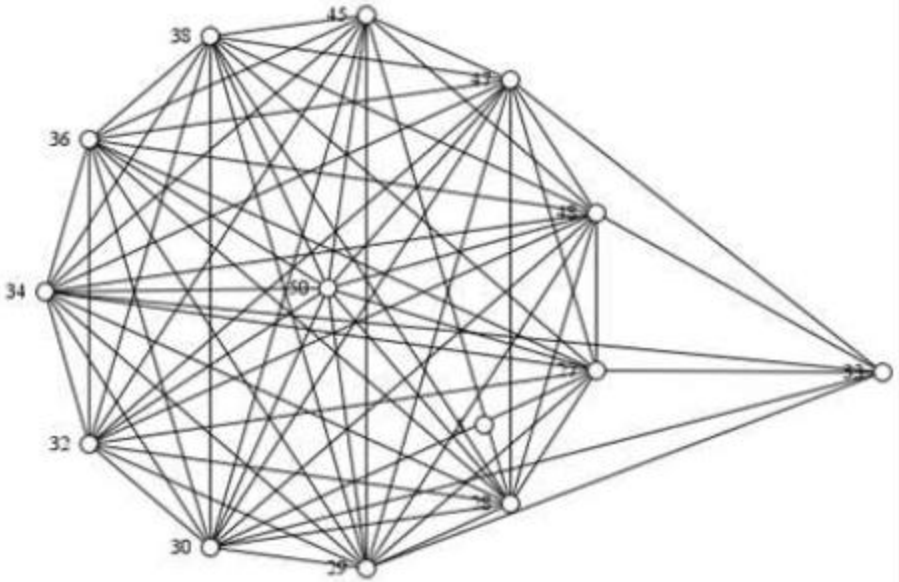


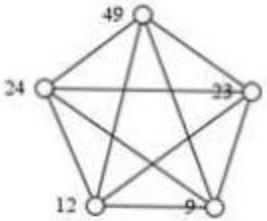


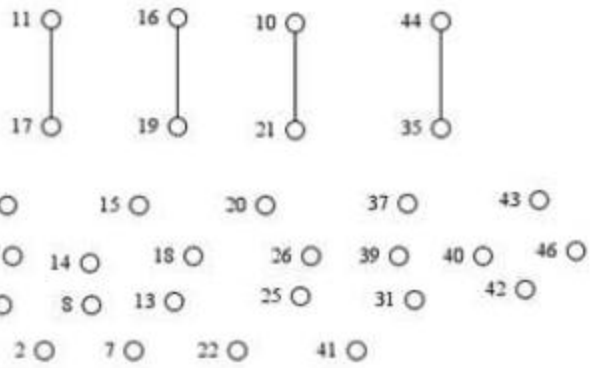
A Multi Scale Graph Theoretical Approach To Gene Regulation Networks: a Case Study In Atrial Fibrillation

Federica Censi, Alessandro Giuliani, Pietro Bartolini, Giovanni Calcagnini

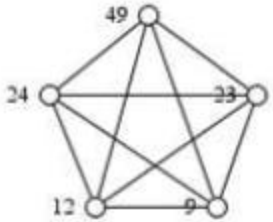
Atrial fibrillation patients

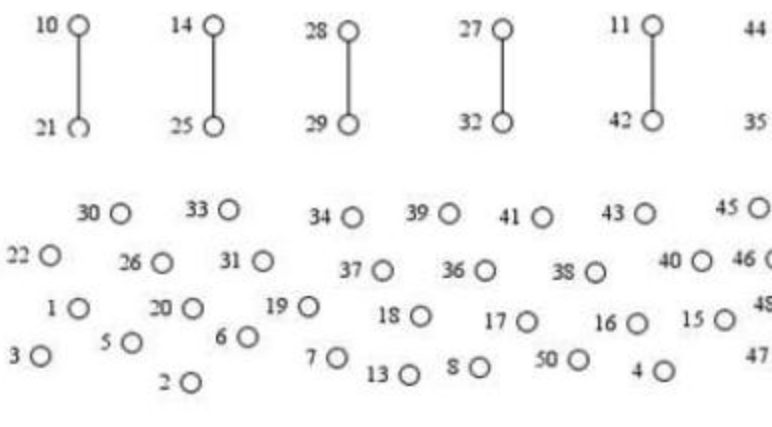




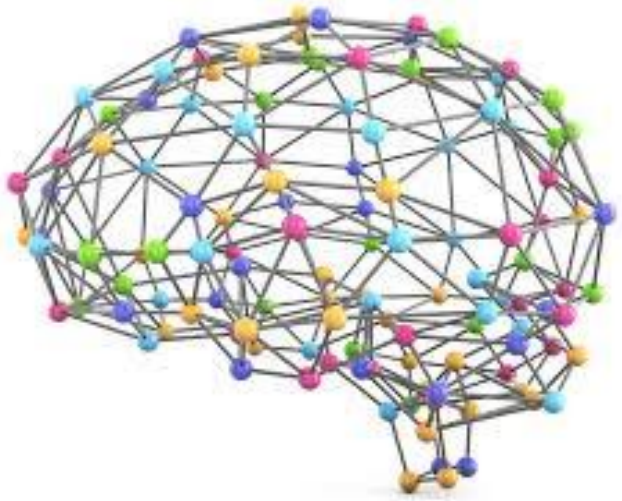


Controls





Brain has an incredibly complex connectivity structure that can be appreciated by many view points. One of this is metabolism: the metabolic rate (glucose consumption) is measured by PET at different brain areas (ROI = Regions Of Interest) and their average degree of correlation is estimated.



Predicting the transition from normal aging to Alzheimer's disease: A statistical mechanistic evaluation of FDG-PET data

Marco Pagani [a,b,*], Alessandro Giuliani [c], Johanna Öberg [d], Andrea Chincarini [e], Silvia Morbelli [f], Andrea Brugnolo [g], Dario Arnaldi [g], Agnese Picco [g], Matteo Bauckneht [f], Ambra Buschiazzo [f], Gianmario Sambuceti [f], Flavio Nobili [g]

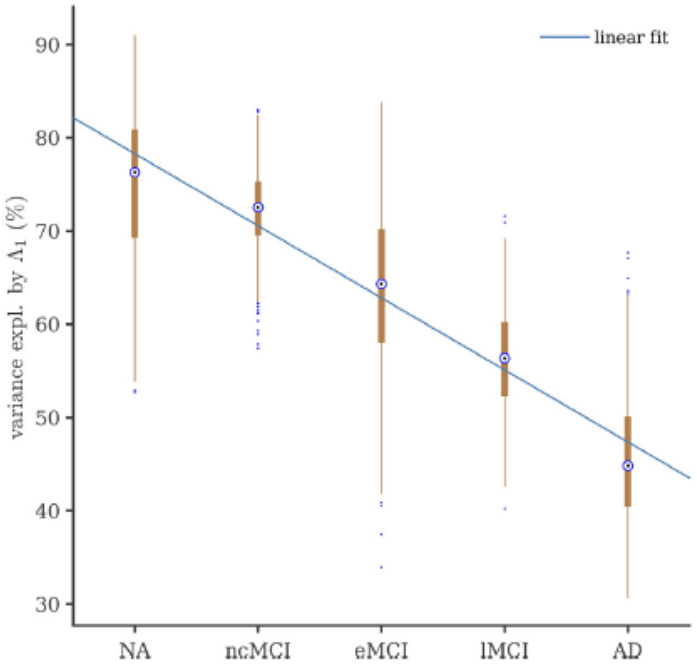


Fig. 1. The dynamics of the loss of order along the clinical status. Y-axis: variance explained by the first component; X-axis: disease severity. NA: normal aging; ncMCI: MCI patients not converting to AD at 5 years follow up; eMCI: MCI patients that converted to AD later than 2 years; lMCI: MCI patients that converted to AD within 2 years; AD: patients with mild AD dementia. The point distribution around the center of mass corresponds to bootstrap simulation.

Partial Correlation

- Note the subscripts in the symbol for a partial correlation coefficient:

$$r_{xy \bullet z}$$

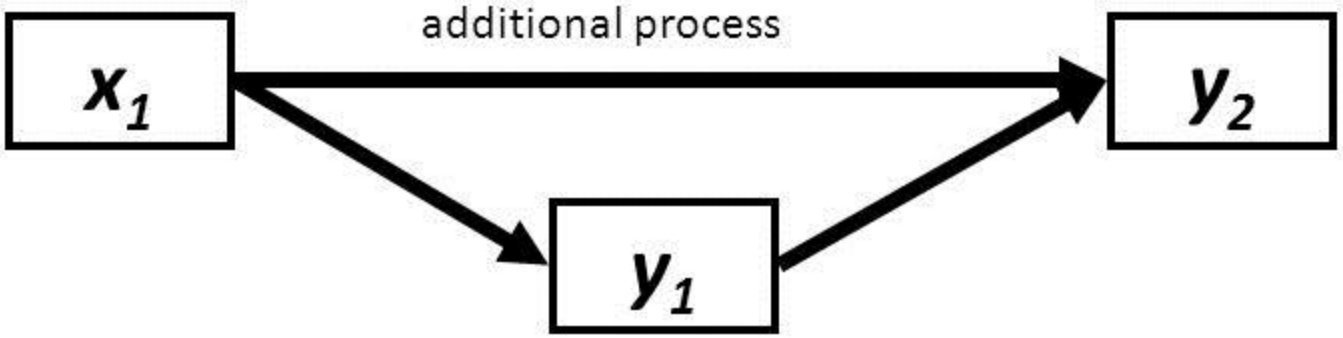
which indicates that the correlation coefficient is for X and Y *controlling for* Z

FORMULA 16.1

$$r_{yx.z} = \frac{r_{yx} - (r_{yz})(r_{xz})}{\sqrt{1 - r_{yz}^2}\sqrt{1 - r_{xz}^2}}$$

Note that you must first calculate the zero-order coefficients between all possible pairs of variables (variables X and Y, X and Z, and Y and X) before solving this formula.

The inequality implies that the true model is



Fourth Rule of Path Coefficients: when variables are connected by more than one causal pathway, the path coefficients are "partial" regression coefficients.

Which pairs of variables are connected by two causal paths?

answer: x_1 and y_2 (obvious one), but also y_1 and y_2, which are connected by the joint influence of x_1 on both of them.

BIOINFORMATICS

Vol. 20 no. 18 2004, pages 3565–3574
doi:10.1093/bioinformatics/bth445

Discovery of meaningful associations in genomic data using partial correlation coefficients

Alberto de la Fuente, Nan Bing†, Ina Hoeschele and Pedro Mendes*

Virginia Polytechnic Institute and State University, Virginia Bioinformatics Institute, 1880 Pratt Drive, Blacksburg, Virginia, 24061 USA

Received on June 2, 2004; revised on July 15, 2004; accepted on July 24, 2004
Advance Access publication July 29, 2004

Motivation: A major challenge of systems biology is to infer biochemical interactions from large-scale observations, such as transcriptomics, proteomics and metabolomics. We propose to use a partial correlation analysis to construct approximate Undirected Dependency Graphs from such large-scale biochemical data. This approach enables a distinction between direct and indirect interactions of biochemical compounds, thereby inferring the underlying network topology.

zeroth-order correlation: $$r_{xy} = \frac{\mathrm{cov}(xy)}{\sqrt{\mathrm{var}(x)\mathrm{var}(y)}} \quad (1)$$

first-order correlation: $$r_{xy.z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}} \quad (2)$$

second-order correlation: $$r_{xy.zq} = \frac{r_{xy.z} - r_{xq.z}r_{yq.z}}{\sqrt{(1 - r_{xq.z}^2)(1 - r_{yq.z}^2)}} \quad (3)$$

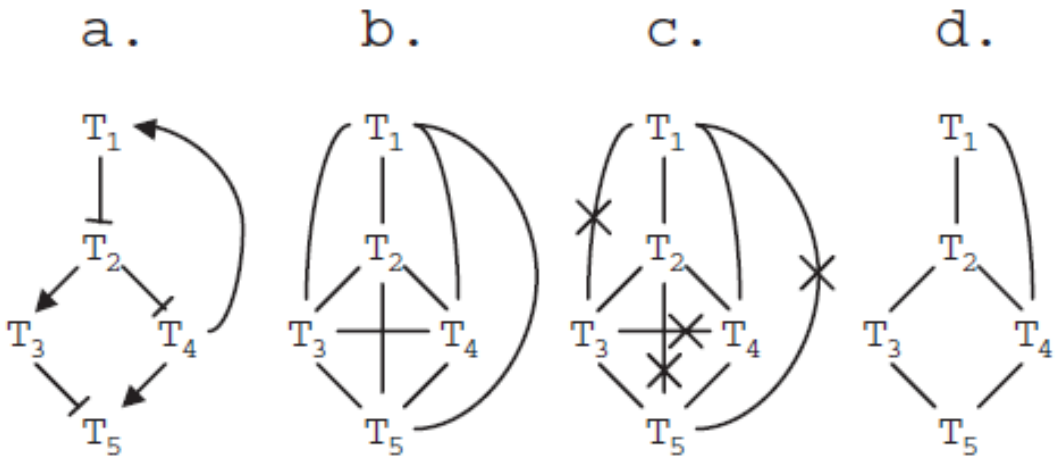


Fig. 1. Stepwise description of the method. (**a**) The actual causal network. (**b**) The UDG based on zero-order correlation. (**c**) Edges with zero partial correlation coefficients are eliminated. (**d**) The resulting UDG.

Correlation beyond Pearson…..

1. Spearman rank correlation coefficient

$$\rho = 1 - \frac{6\sum d^2}{n\left(n^2 - 1\right)}$$

n= number of paired ranks

d= difference between the paired ranks

(when two or more observations of one variable are the same, ranks are assigned by averaging positions occupied in their rank order)

Correlation beyond Pearson…

2. Kendall's Tau

- Kendall's Tau $\tau=\frac{C-D}{C+D}$
- Or $\tau=\frac{C-(N-C)}{N}=\frac{4C}{n(n-1)}$-1
- C:#of total concordant pairs
- D:#of total discordant pairs
- N:#of comparisons=$\binom{n}{2}=\frac{1}{2}$n(n-1)
- n:#of pairs

Correlation beyond Pearson…

3. Chi-Square

	A	Non A
B	********** ********** **********	* * *
Non B	* * *	********** ********** **********

10.3 - Sensitivity, Specificity, Positive Predictive Value, and Negative Predictive Value

In this example, two columns indicate the actual condition of the subjects, diseased or non-diseased. The rows indicate the results of the test, positive or negative.

Cell A contains true positives, subjects with the disease and positive test results. Cell D subjects do not have the disease and the test agrees.

A good test will have minimal numbers in cells B and C. Cell B identifies individuals without disease but for whom the test indicates 'disease'. These are false positives. Cell C has the false negatives.

If these results are from a population-based study, prevalence can be calculated as follows:

- **Prevalence of Disease= $T_{disease}$/ Total × 100**

The population used for the study influences the prevalence calculation.

Sensitivity is the probability that a test will indicate 'disease' among those with the disease:

- **Sensitivity: A/(A+C) × 100**

Specificity is the fraction of those without disease who will have a negative test result:

- **Specificity: D/(D+B) × 100**

Sensitivity and specificity are *characteristics of the test*. The population does not affect the results.

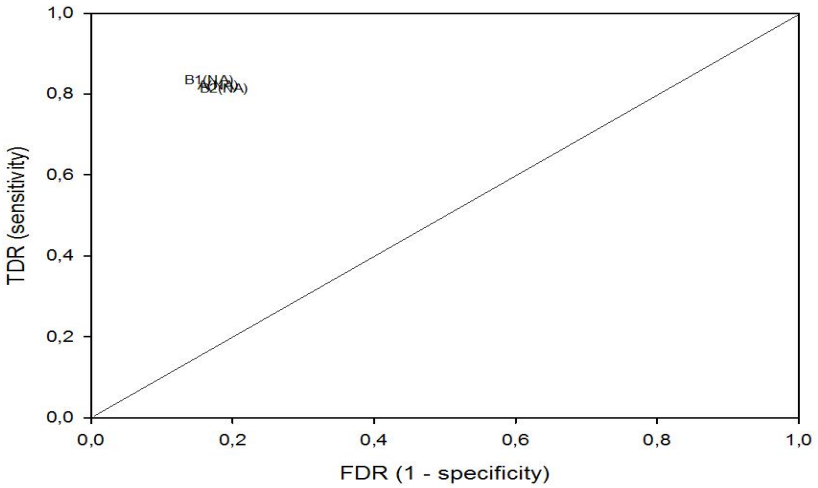
A clinician and a patient have a different question: what is the chance that a person with a positive test truly has the disease? If the subject is in the first row in the table above, what is the probability of being in cell A as compared to cell B? A clinician calculates across the row as follows:

- **Positive Predictive Value: A/(A+B) × 100**
- **Negative Predictive Value: D/(D+C) × 100**

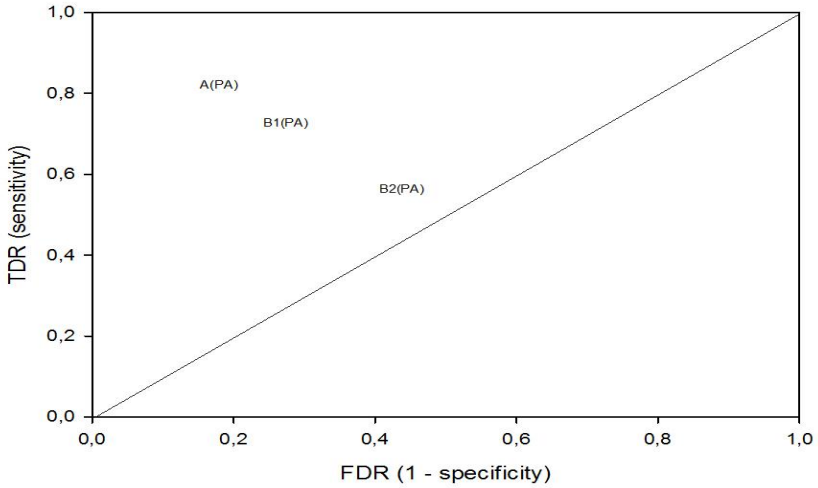
Positive and negative predictive values are influenced by the prevalence of disease in the population that is being tested. If we test in a high prevalence setting, it is more likely that persons who test positive truly have disease than if the test is performed in a population with low prevalence..

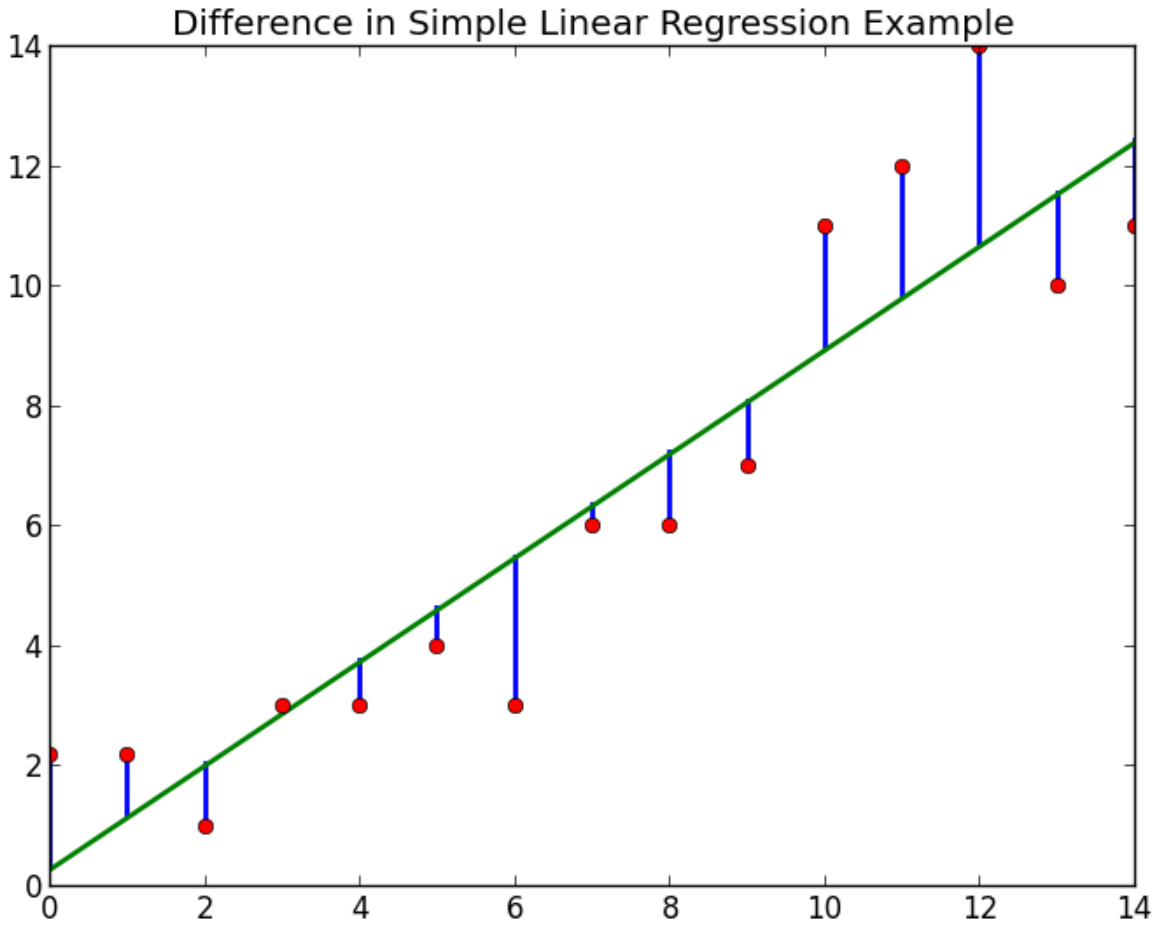
Test Result	Truth: Disease (number)	Truth: Non Disease (number)	Total (number)
Positive (number)	A (True Positive)	B (False Positive)	$T_{Test\ Positive}$
Negative (number)	C (False Negative)	D (True Negative)	$T_{Test\ Negative}$
	$T_{Disease}$	$T_{Non\ Disease}$	Total

ROC plane for NA



ROC plane for PA





To regress one dependent variable (Y) on and indpendent one (X) generates a model Y = a + bX where a and b parameters derive by the minimization of the squared distance (r (i) or residuals) of the observed points by the model (least squares optimization)

Solutions for a and b parameters when imposing:

S $(r(i))^2$ = minimum

$$a = \frac{\sum y - b \sum x}{n}$$

$$b = \frac{n \sum (xy) - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2}$$

La soluzione ai minimi quadrati, quella cioè che rende minima
la distanza quadratica dei valori osservati dalla loro stima
è riportata di seguito (n = numero di casi), notare la 'somiglianza'
della formula di b (slope) con quella del coefficiente di correlazione

.. riscrivendo la formula:

$$b = \frac{N \sum_i x_i y_i - \sum_i x_i \sum_i y_i}{N \sum_i x_i^2 - (\sum_i x_i)^2} = \frac{S_{xy}}{S_{xx}} = \frac{\operatorname{cov}(x, y)}{\operatorname{var}(x)}$$

http://it.wikipedia.org/wiki/Regressione_lineare

Quando invece di una sola variabile indipendente ne ho molte, ognuna che apporta un suo contributo alla spiegazione della varianza di Y (variabile dipendente), il modello visto in precedenza rimane sostanzialmente immutato:

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i,$$

$$e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_1 - \cdots - \hat{\beta}_p x_p$$

$$\sum_{i=1}^{N} \sum_{k=1}^{p} X_{ij} X_{ik} \hat{\beta}_k = \sum_{i=1}^{N} X_{ij} y_i, \ j = 1, p$$

$$\left(\mathbf{X^T X}\right) \hat{\boldsymbol{\beta}} = \mathbf{X^T y}$$

Quantitative Structure–Activity Relationships of Mutagenic and Carcinogenic Aromatic Amines

Romualdo Benigni,*,† Alessandro Giuliani,† Rainer Franke,‡ and Andreas Gruska‡

Istituto Superiore di Sanità, Laboratory of Comparative Toxicology and Ecotoxicology, Viale Regina Elena 299, I-00161 Rome, Italy, and Consulting in Drug Design GbR, Gartenstr. 14, D-16352 Basdorf, Germany

Scheme 1. Pattern of Metabolic Activation Pathways of Aromatic Amines. The Scheme Sketches the Most Representative Pathways (see details in the text)

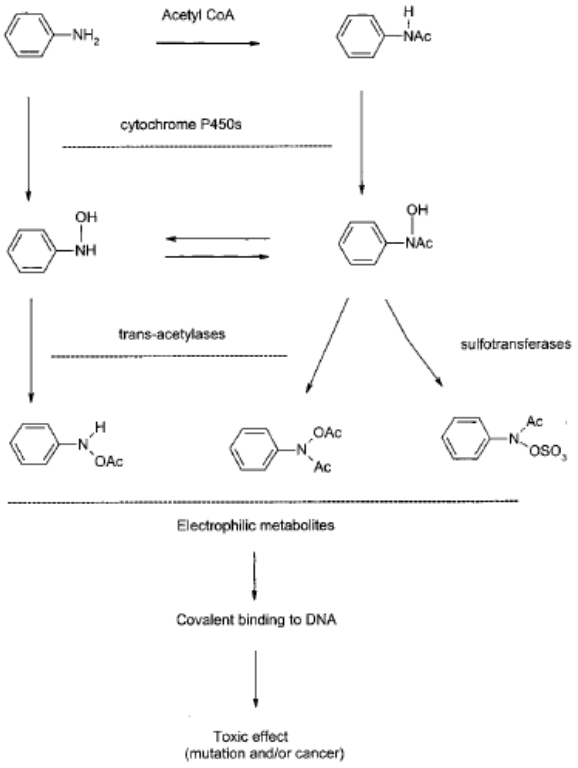


Table 2. Chemical Descriptors and Carcinogenic Potencies (BRR, rats; BRM, mice) of the Carcinogenic Compounds in Table 1

									BRR		BRM	
no.	MR_3	$\Sigma MR_{2,6}$	$\Sigma\chi_{2,6}$	$\Sigma\chi^v_{2,6}$	E_S (R)	EHOMO	ELUMO	log P	observed	predicted[a]	observed	predicted[b]
1	0.8	0.2	0	0	0	−8.3724	−0.3671	2.27	0.37	0.31	0.59	0.97
2	0.1	0.2	0	0	0	−8.2341	0.0005	2.16	2.03	2.21	0.97	0.86
3	0.56	0.2	0	0	3	−8.4704	−0.3989	2.61	2.26	1.58	1.47	0.85
4	0.1	0.7	0.42	−0.19	0	−8.5595	0.0261	3.60	1.14	1.65		
5	0.1	0.66	0.42	−0.19	0	−8.5528	0.4059	1.73	0.39	0.13	−0.89	−0.45
6	0.1	0.2	0	0	2	−8.3483	−0.0286	3.02	1.39	1.45	0.63	0.59
7	0.1	2.64	0.12	−0.13	0	−8.5626	0.1005	2.95			−0.82	−1.11
8	0.1	1.2	0.84	−0.38	0	−8.2613	−0.089	1.52			−0.66	−0.32
9	0.1	0.84	0.65	0.13	1	−8.7212	−1.0742	0.34	−0.44	−0.36	0.47	0.14
10	0.1	0.2	0	0	0	−8.4373	0.3141	2.56	1.00	1.29	0.79	0.29
11	0.1	0.2	0	0	5	−8.7767	−0.1195	1.64	0.18	0.09		
12	0.1	0.2	0	0	0	−8.3253	0.1629	1.91	1.32	1.06	0.78	0.45
13	0.1	1.35	0.26	−0.5	0	−8.717	−0.1664	0.2			−1.03	−1.39
14	0.56	0.2	0	0	0	−8.0133	−0.2496	2.39	0.57	1.51	0.74	1.04
15	0.74	0.2	0	0	0	−9.039	−0.7543	0.93	−0.30	−0.15		
16	0.1	0.2	0	0	0	−8.6088	0.4153	1.26	−0.46	−0.04		
17	0.1	0.89	0.29	−0.55	0	−8.5707	0.2834	1.01	0.62	−0.13	−0.89	−1.19
18	0.1	0.2	0	0	0	−8.5906	0.1051	1.78			0.15	0.42
19	0.1	0.7	0.42	−0.19	0	−8.3803	0.1761	1.00	−0.34	−0.13	−0.94	−0.44
20	0.1	0.64	0.08	−0.74	0	−8.3819	0.111	1.00	−0.18	−0.13	−0.97	−0.3
21	0.1	0.66	0.01	−0.18	0	−8.5193	0.2607	1.48	−0.53	0.04		
22	0.1	0.89	0.29	−0.55	0	−8.5439	0.2527	1.48	0.15	0.04		
23	0.1	0.2	0	0	0	−9.0261	−0.3937	1.31	1.04	0.85		
24	0.1	0.89	0.29	−0.55	0	−8.1651	0.3817	0.23	−0.12	−0.4	−0.82	−0.83
25	0.1	0.2	0	0	2	−8.2965	0.3429	3.71	1.19	1.69	0.09	0.39
26	0.1	0.2	0	0	2	−8.6024	−0.2779	2.85	1.68	1.39	0.50	0.51
27	0.8	0.9			0	−7.9895	−0.3544	1.48	0.36	0.04	−0.01	−0.07
28	0.74	0.2	0	0	3	−9.5438	−1.1895	0.94			−1.01	−1.29
29	0.1	0.89	0.29	−0.55	0	−9.1996	−1.1264	0.96			−0.34	−0.3

Supervised Learning

$$\begin{aligned}\mathrm{BRM} = {} & 0.56\ (\pm 0.18)\ \log P + \\ & 1.03\ (\pm 0.74)\ \mathrm{EHOMO} - 1.19\ (\pm 0.58)\ \mathrm{ELUMO} - \\ & 0.79\ (\pm 0.37)\ \textstyle\sum \mathrm{MR}_{2,6} - 0.93\ (\pm 0.90)\ \mathrm{MR}_3 - \\ & 0.22\ (\pm 0.19)\ E_S(\mathrm{R}) + 8.51\ (\pm 6.31) \quad (8)\end{aligned}$$

$$n = 37 \quad r = 0.845 \quad r^2 = 0.714 \quad s = 0.485$$
$$F = 12.5 \quad p < 0.001$$

$$\begin{aligned}\mathrm{BRM} = {} & 1.03\ (\pm 0.31)\ \log P + \\ & 3.37\ (\pm 1.11)\ \mathrm{EHOMO} - 0.97\ (\pm 0.89)\ \mathrm{ELUMO} - \\ & 0.96\ (\pm 0.36)\ \textstyle\sum \mathrm{MR}_{2,6} - 1.41\ (\pm 0.92)\ \mathrm{MR}_3 + \\ & 2.21\ (\pm 0.89)\ I(\mathrm{NO}_2) + 27.73\ (\pm 9.48) \quad (10)\end{aligned}$$

$$n = 17 \quad r = 0.968 \quad r^2 = 0.937 \quad s = 0.281$$
$$F = 25.0 \quad p < 0.001$$

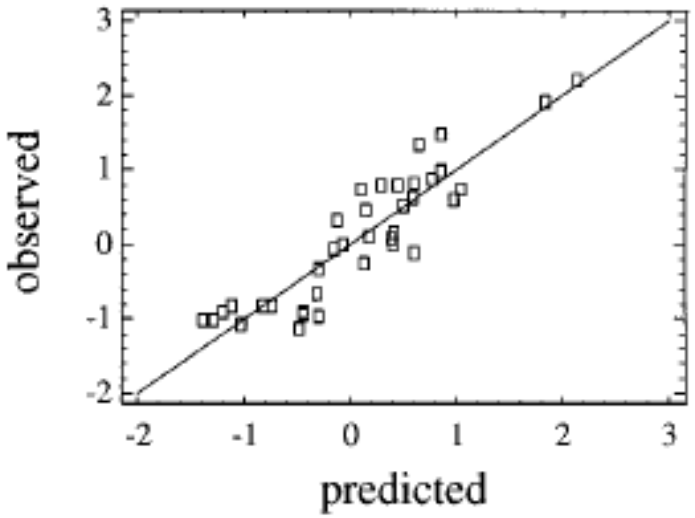


Figure 4. Plot of observed values of BRM against predicted values from eq 13.

$$\begin{aligned} \mathrm{BRM} = {} & 1.03\ (\pm 0.31)\ \log P + \\ & 3.37\ (\pm 1.11)\ \mathrm{EHOMO} - 0.97\ (\pm 0.89)\ \mathrm{ELUMO} - \\ & 0.96\ (\pm 0.36)\ \textstyle\sum \mathrm{MR}_{2,6} - 1.41\ (\pm 0.92)\ \mathrm{MR}_3 + \\ & 2.21\ (\pm 0.89)\ I(\mathrm{NO}_2) + 27.73\ (\pm 9.48) \quad (10) \end{aligned}$$

$$n = 17 \quad r = 0.968 \quad r^2 = 0.937 \quad s = 0.281$$
$$F = 25.0 \quad p < 0.001$$

La regressione logistica è un caso particolare di modello lineare avente una particolare trasformata come variabile dipendente detta *logit* . Si tratta di un modello di regressione applicato nei casi in cui la variabile dipendente sia di tipo dicotomico.

Il modello si indica come: $\text{logit}(p) = a0 + a1x1 + a2x2 + a3x3\ldots$

Essendo $\text{logit}(p) = \ln(p/(1-p))$ e, conseguentemente

$p = \exp(a0 + a1x1 + a2x2 + a3x3..)/1 + \exp(a0 + a1x1 + a2x2 + a3x3..)$

Non c’è alcun motivo per non usare direttamente, la variabile dicotomica (y) tal quale (Y prende due valori 1,0 per presente e assente), in tal caso avremmo l’usuale regressione lineare:

$y = a0 + a1x1 + a2x2 + a3x3\ldots$

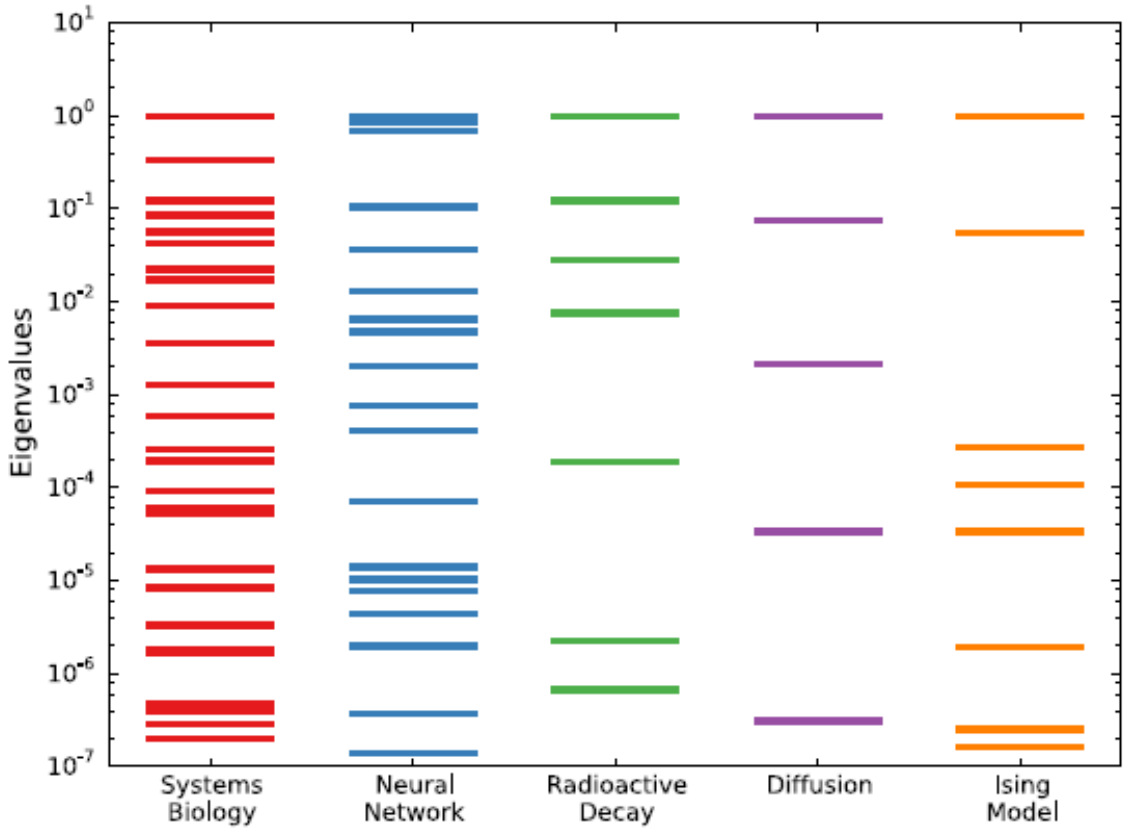
Solo che così avremmo delle stime di y nello spazio dei reali e non in quello di probabilità (che va da 0 a 1) che in molti casi è ciò che ci occorre ……

FIM: Fisher Information Matrix

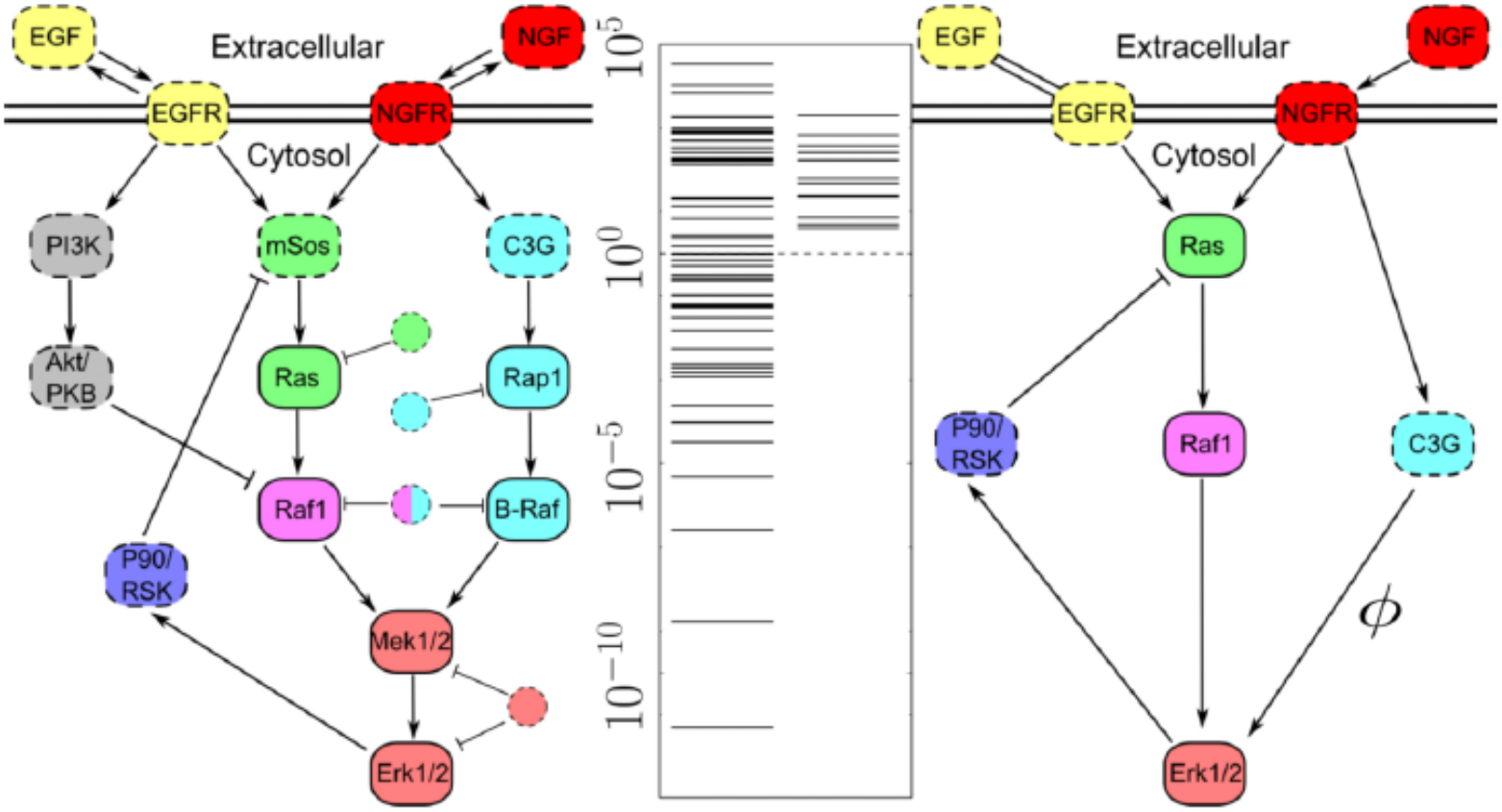
To what extent the knowledge of variable X decrease the uncertainty about Y

methods, as we detail in Section II. In particular, the Fisher Information Matrix (FIM) can be used to estimate the uncertainty in each parameter in our model. The result for the sum of exponentials is that each parameter is *almost completely undetermined*. Any parameter can be varied by an infinite amount and the model could still fit the data. This does not mean that all parameters can be varied independently of the others. Indeed, while the statistical uncertainty in each individual parameter might be infinite, the data place constraints on *combinations* of the parameters.

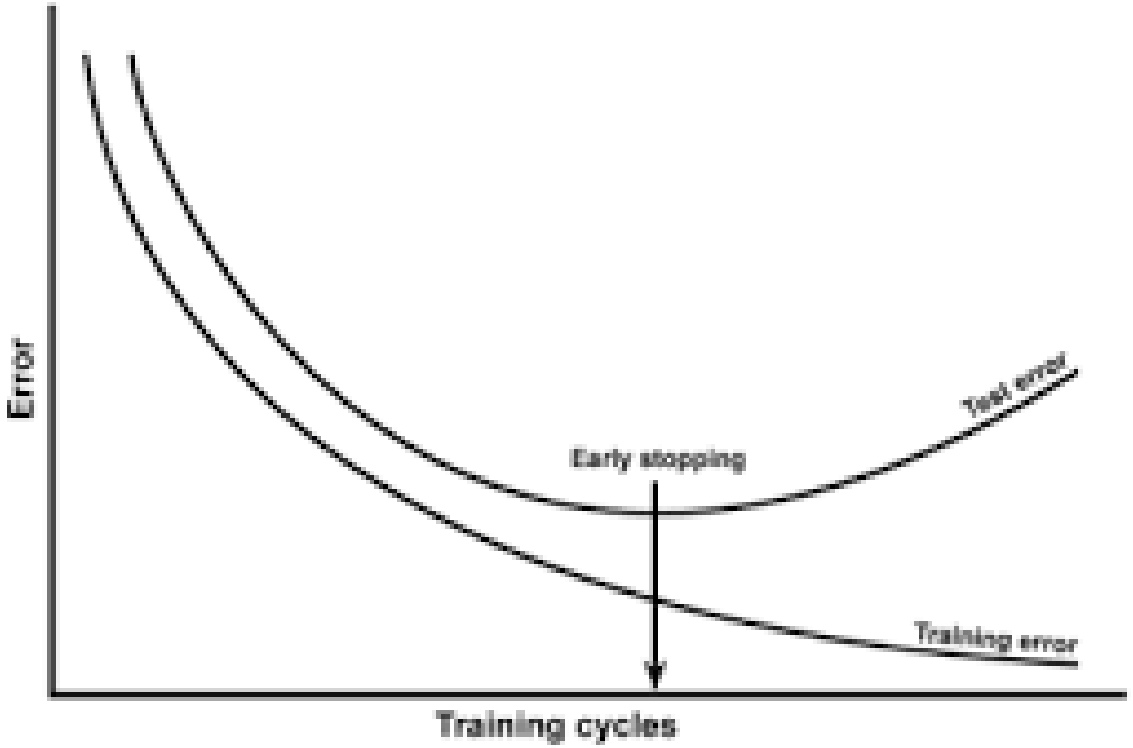
The eigenvalues of the FIM tell us which parameter combinations are well-constrained by the data and which are not. Most of the FIM eigenvalues are very small, corresponding to combinations of parameters that have little effect on model behavior. These unimportant parameter combinations are designated *sloppy*. A small number of eigenvalues are relatively large, revealing the few parameter combinations that are important to the model (known as *stiff*). It is generally



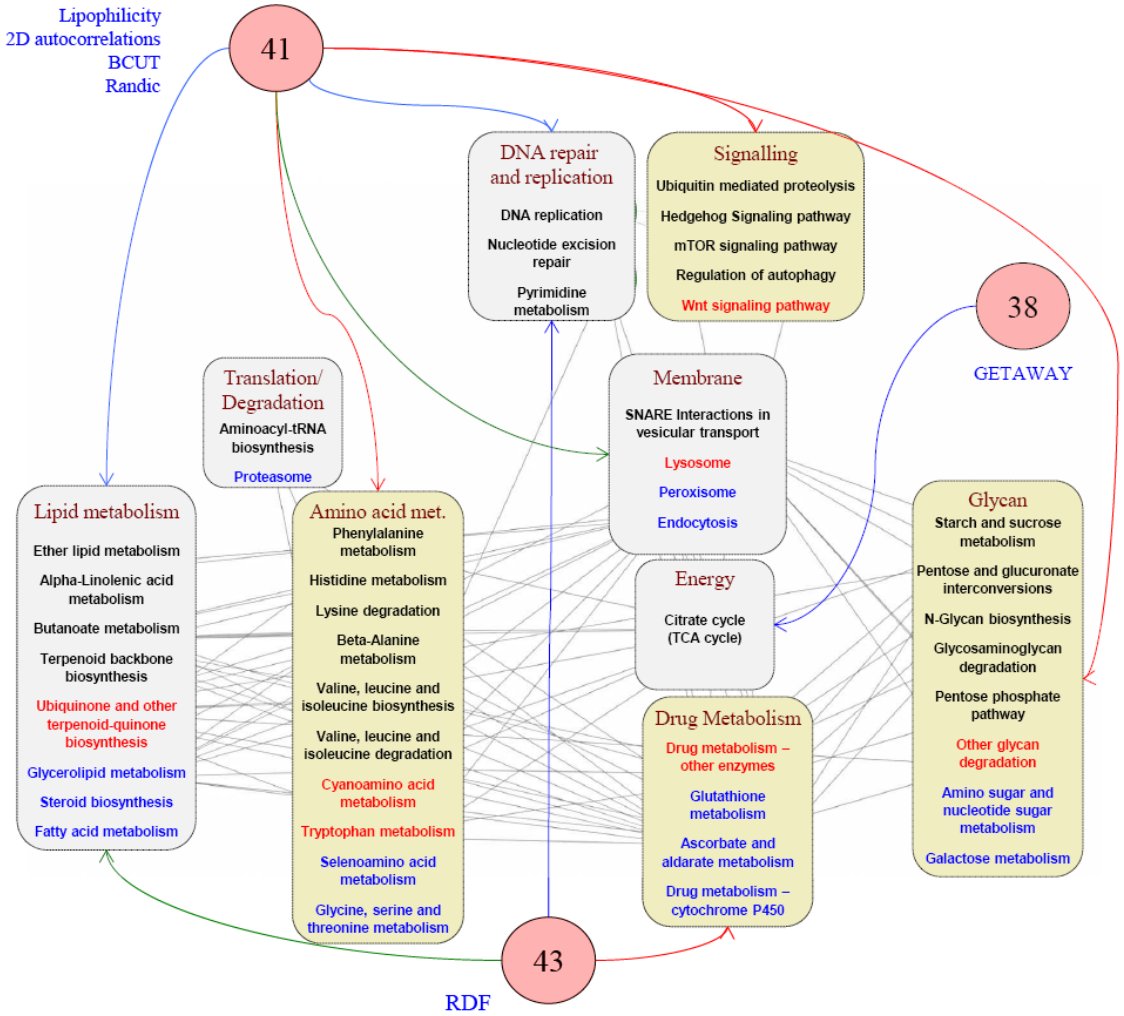
We noted previously that the characteristic eigenvalue spectrum of the FIM suggests a simpler, lower-dimensional "theory" embedded within larger, more complex "models," and in this section, we make this notion explicit. We will see that


EGF
Extracellular
NGF
EGFR
NGFR
Cytosol
PI3K
mSos
C3G
Akt/
PKB
Ras
Rap1
Raf1
B-Raf
P90/
RSK
Mek1/2
Erk1/2
10^{-5}
10^{0}
10^{-5}
10^{-10}
EGF
Extracellular
NGF
EGFR
NGFR
Cytosol
Ras
P90/
RSK
Raf1
C3G
ϕ
Erk1/2

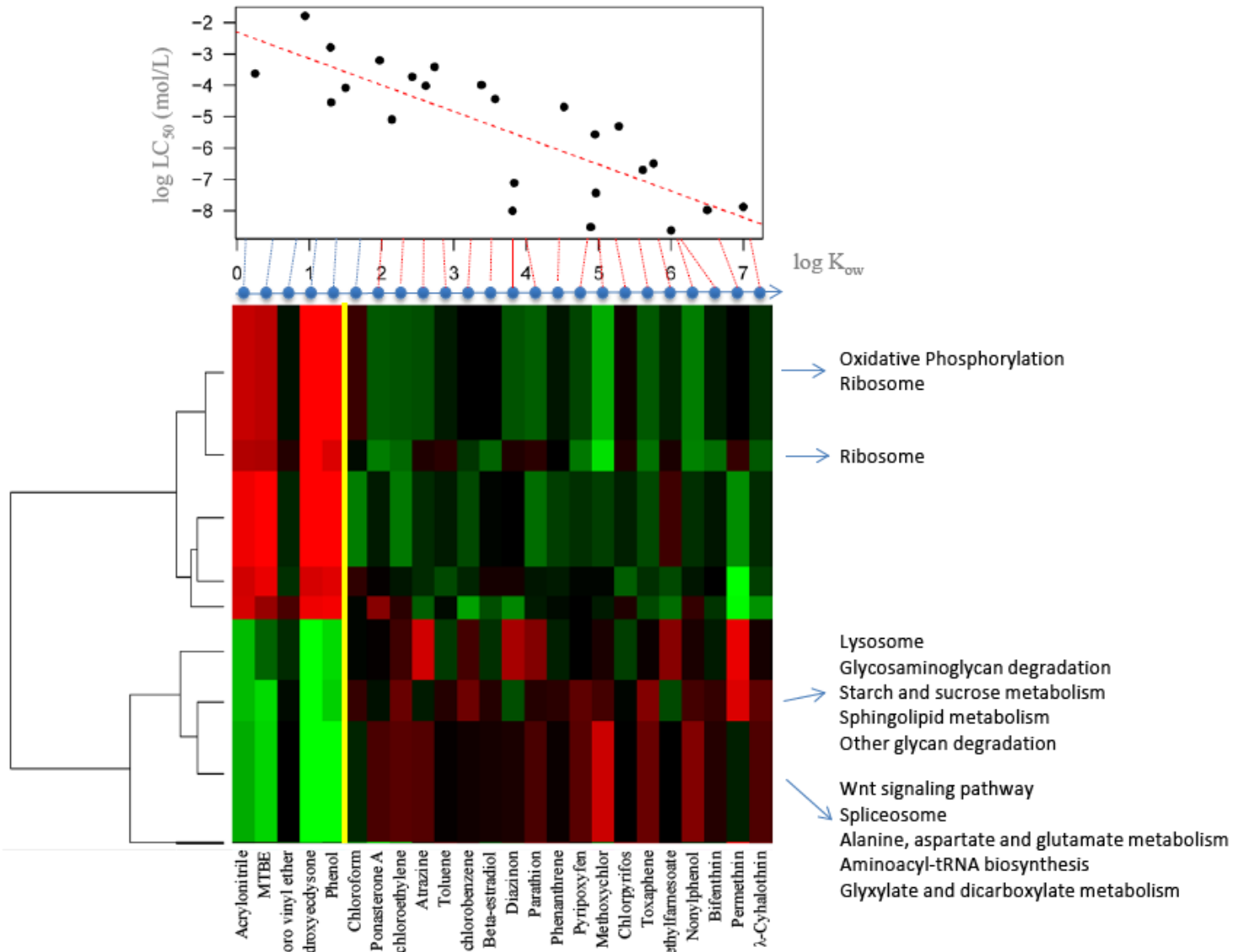
Overfitting refers to a model that fits the training data too well. **Overfitting** happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data.



log(LC50) = -0.8438 log(Kow) – 2.3078. (validated on hundreds of molecules)



can be applied to 24 molecules


log LC_{50} (mol/L)
−2
−3
−4
−5
−6
−7
−8
0
1
2
3
4
5
6
7
log K_{ow}
Acrylonitrile
MTBE
2-chloro vinyl ether
20-hydroxyecdysone
Phenol
Chloroform
Ponasterone A
Trichloroethylene
Atrazine
Toluene
Dichlorobenzene
Beta-estradiol
Diazinon
Parathion
Phenanthrene
Pyripoxyfen
Methoxychlor
Chlorpyrifos
Toxaphene
Methylfarnesoate
Nonylphenol
Bifenthrin
Permethrin
λ-Cyhalothrin
Oxidative Phosphorylation
Ribosome
Ribosome
Lysosome
Glycosaminoglycan degradation
Starch and sucrose metabolism
Sphingolipid metabolism
Other glycan degradation
Wnt signaling pathway
Spliceosome
Alanine, aspartate and glutamate metabolism
Aminoacyl-tRNA biosynthesis
Glyxylate and dicarboxylate metabolism

SULLE FUNZIONI BILINEARI

DI

E. BELTRAMI

LIII. *On Lines and Planes of Closest Fit to Systems of Points in Space.* *By* KARL PEARSON, *F.R.S., University College, London* *.

(1) 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. Analytically this consists in taking

$$y = a_0 + a_1 x, \quad \text{or} \quad z = a_0 + a_1 x + b_1 y,$$

$$\text{or} \quad z = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 + \ldots + a_n x_n,$$

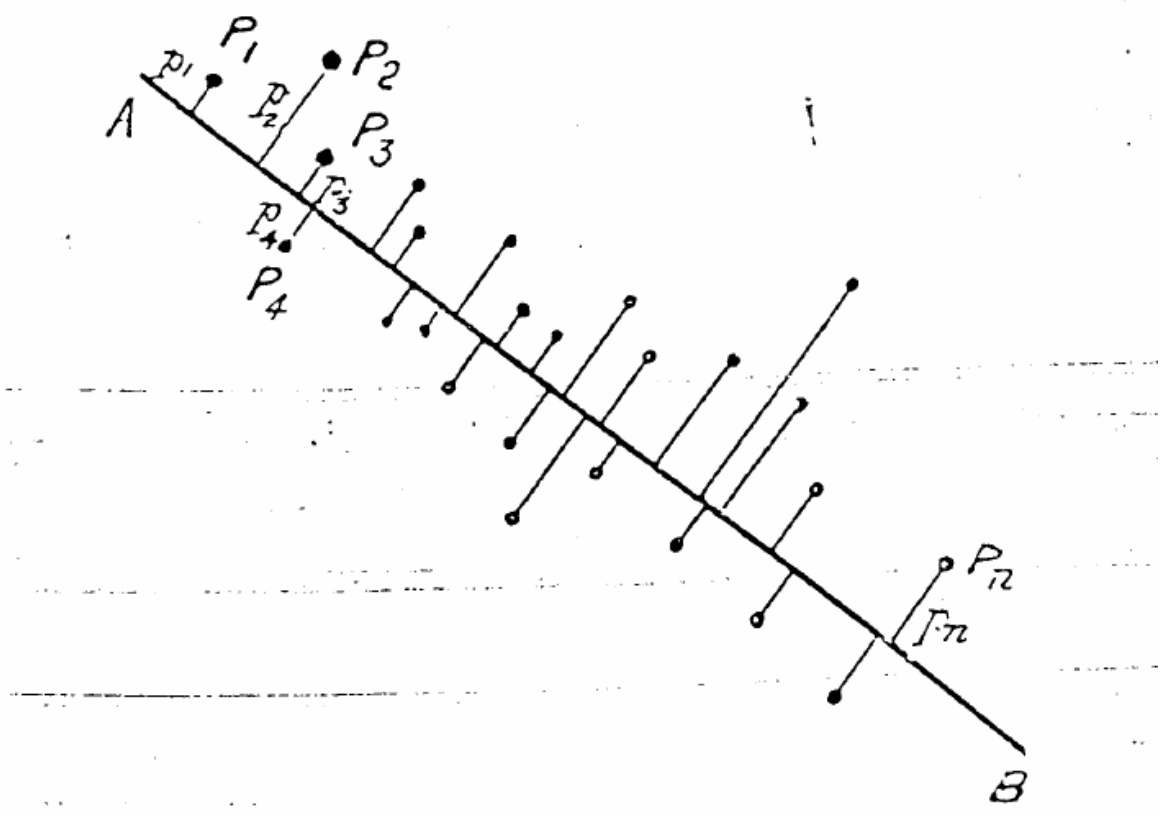
where y, x, z, x_1, x_2, $\ldots x_n$ are variables, and determining the "best" values for the constants a_0, a_1, b_1, a_0, a_1, a_2, a_3, $\ldots a_n$

For example:—Let $P_1, P_2, \ldots P_n$ be the system of points with coordinates $x_1, y_1; x_2, y_2; \ldots x_n y_n$, and perpendicular distances $p_1, p_2, \ldots p_n$ from a line A B. Then we shall make

$$U = S(p^2) = \text{a minimum.}$$

If y were the dependent variable, we should have made

$$S(y' - y)^2 = \text{a minimum}$$



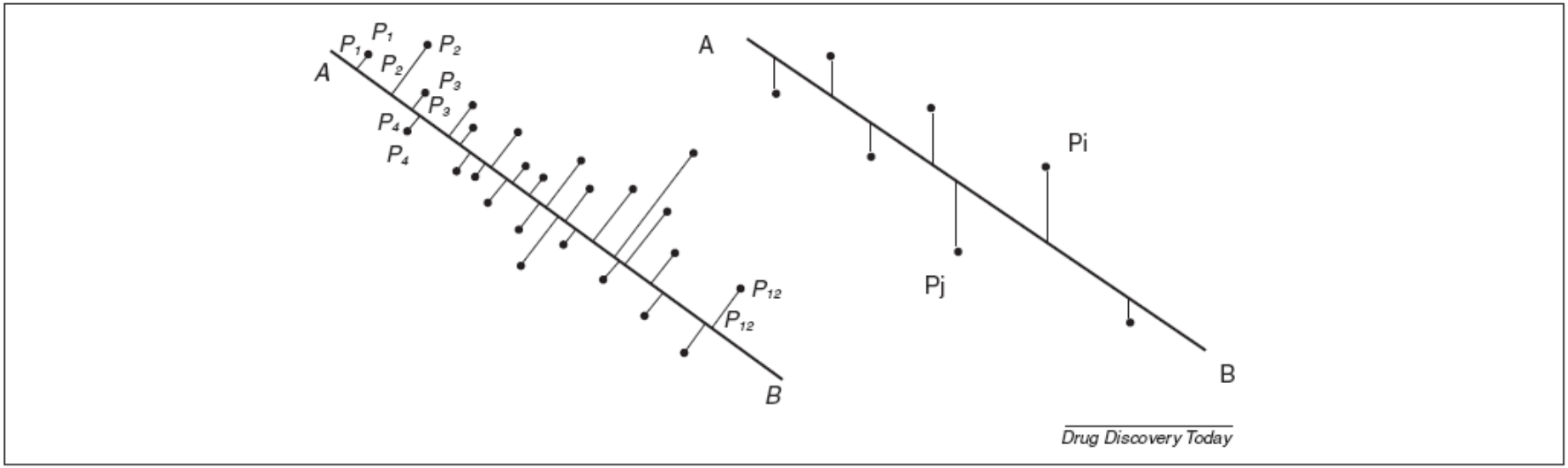
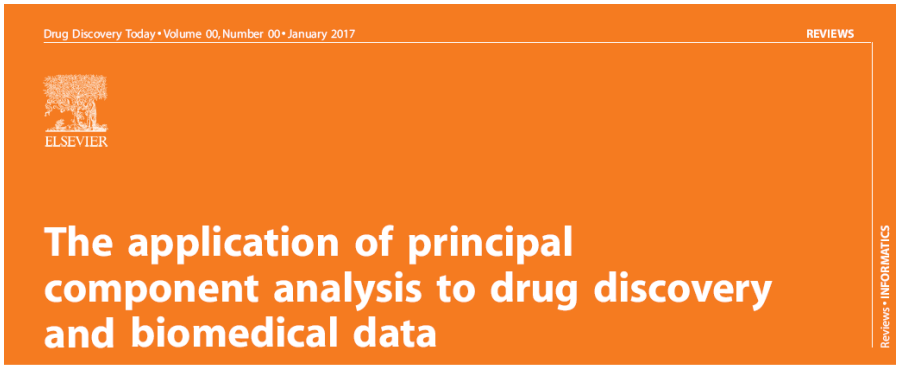


FIGURE 1

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.



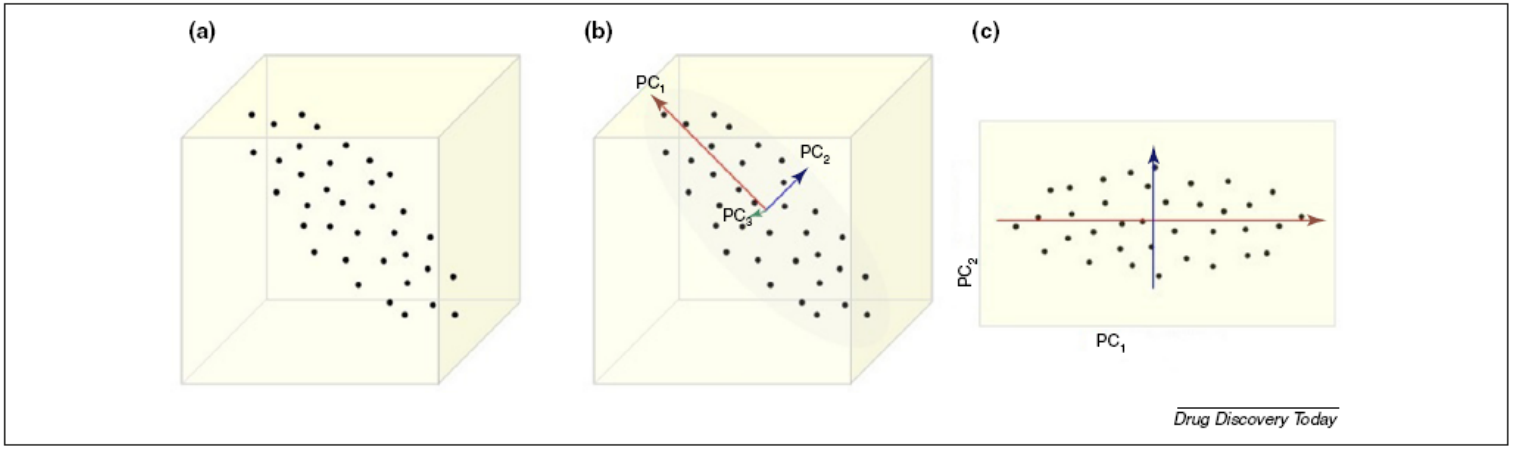


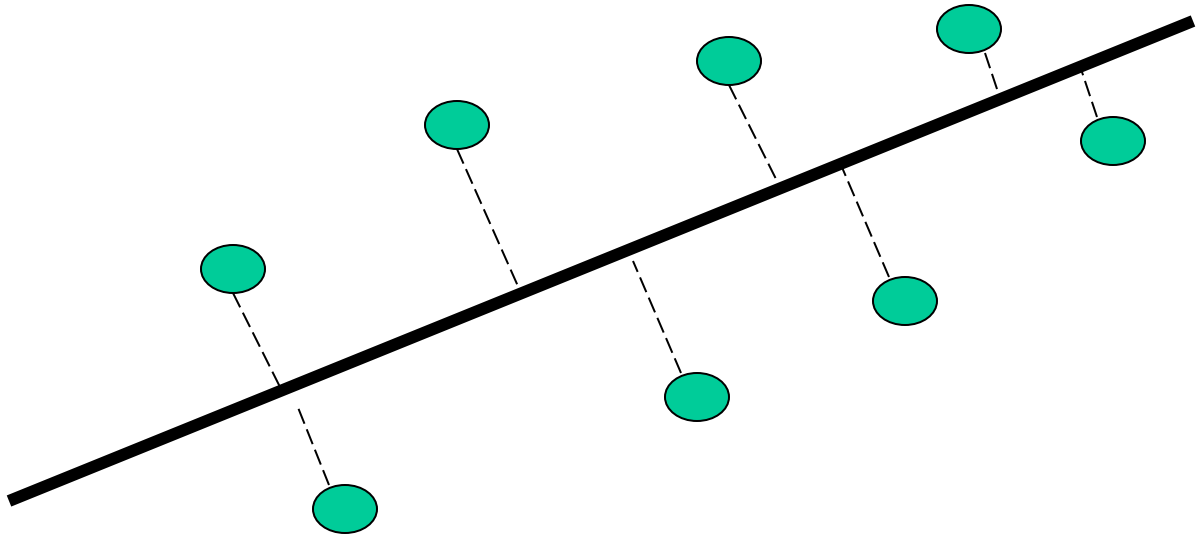
FIGURE A1

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.

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

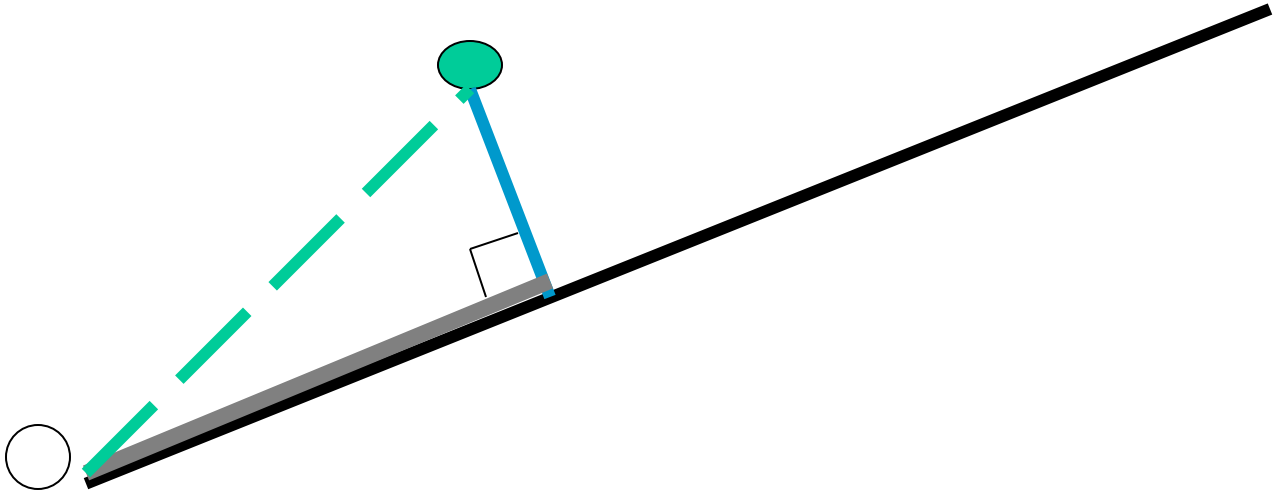
Algebraic Interpretation

- Formally, minimizing the sum of squares of distances to the line…

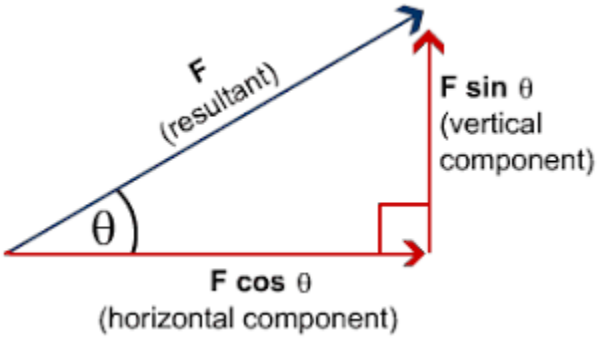


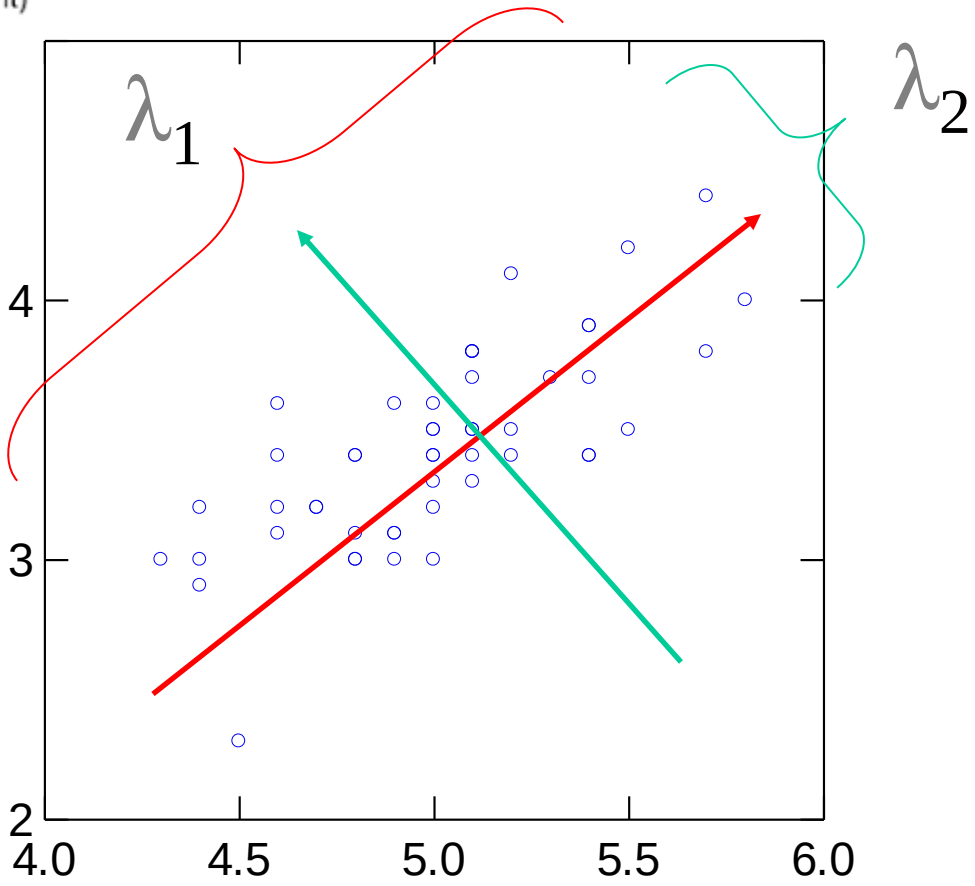
Algebraic Interpretation

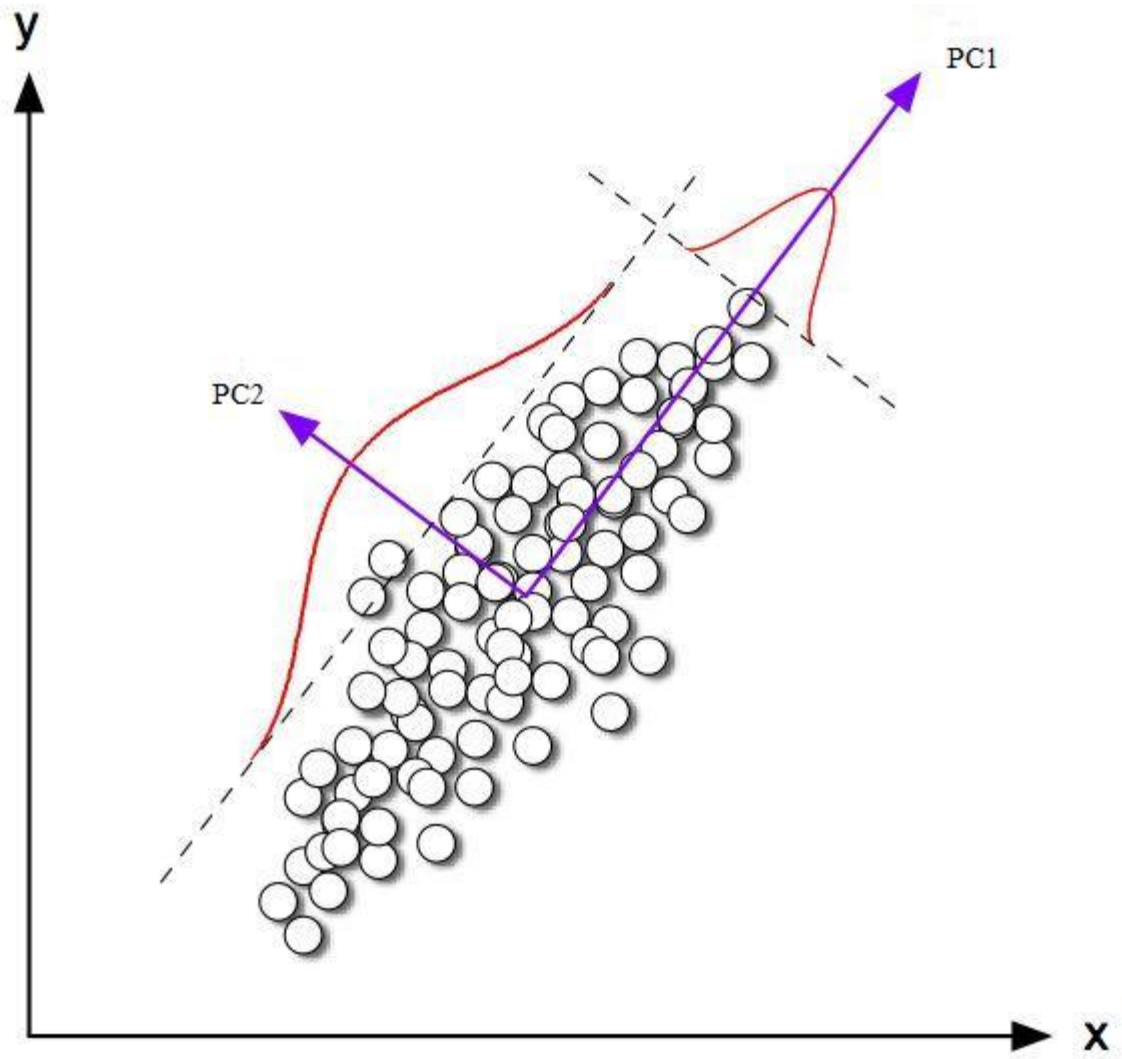
- … is the same as maximizing the sum of squares of the projections on that line, thanks to Pythagoras.



PCA Eigenvalues






y
PC1
PC2
x

Morris Water Maze

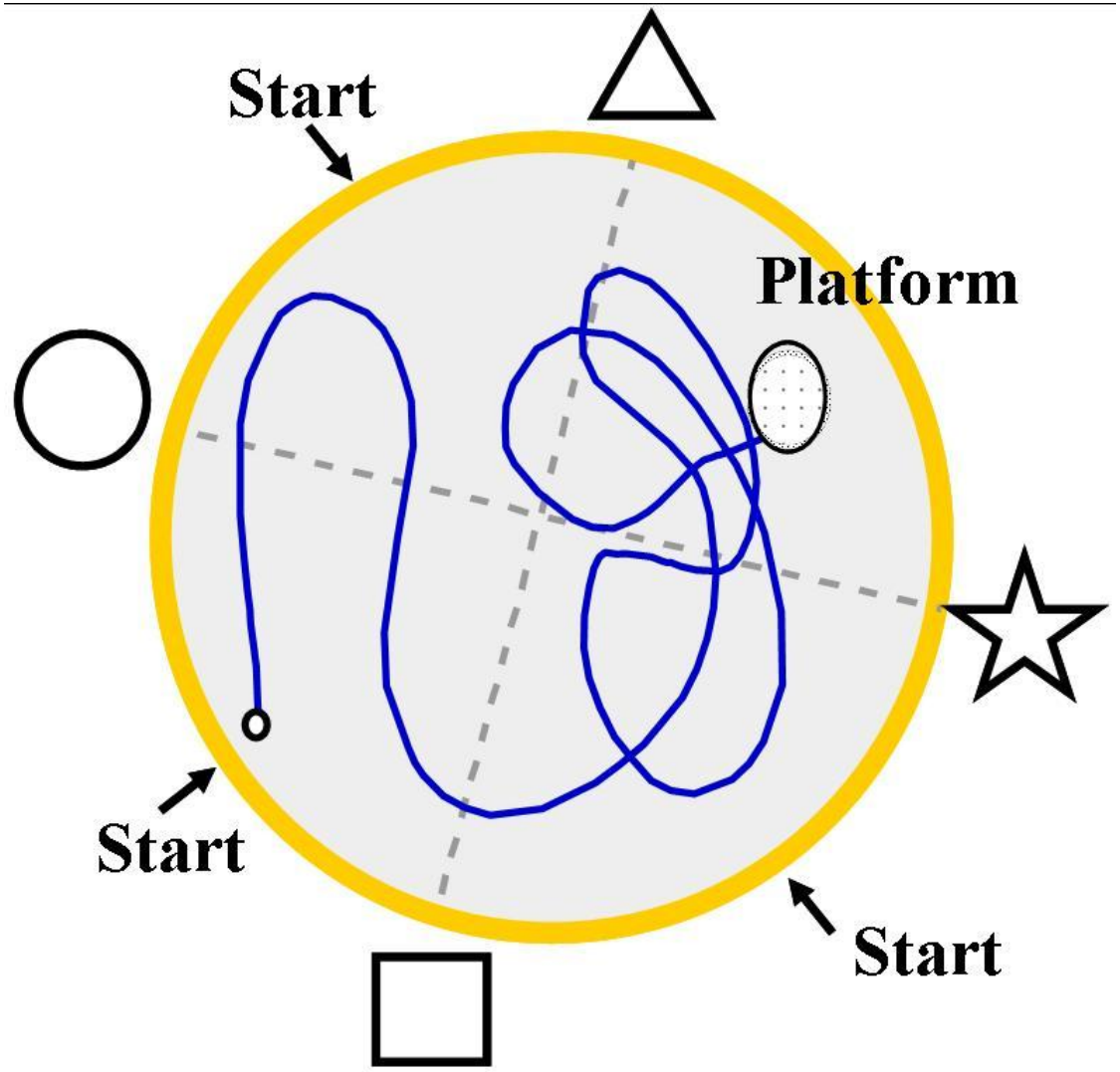


TABLE 1

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	

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.

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.

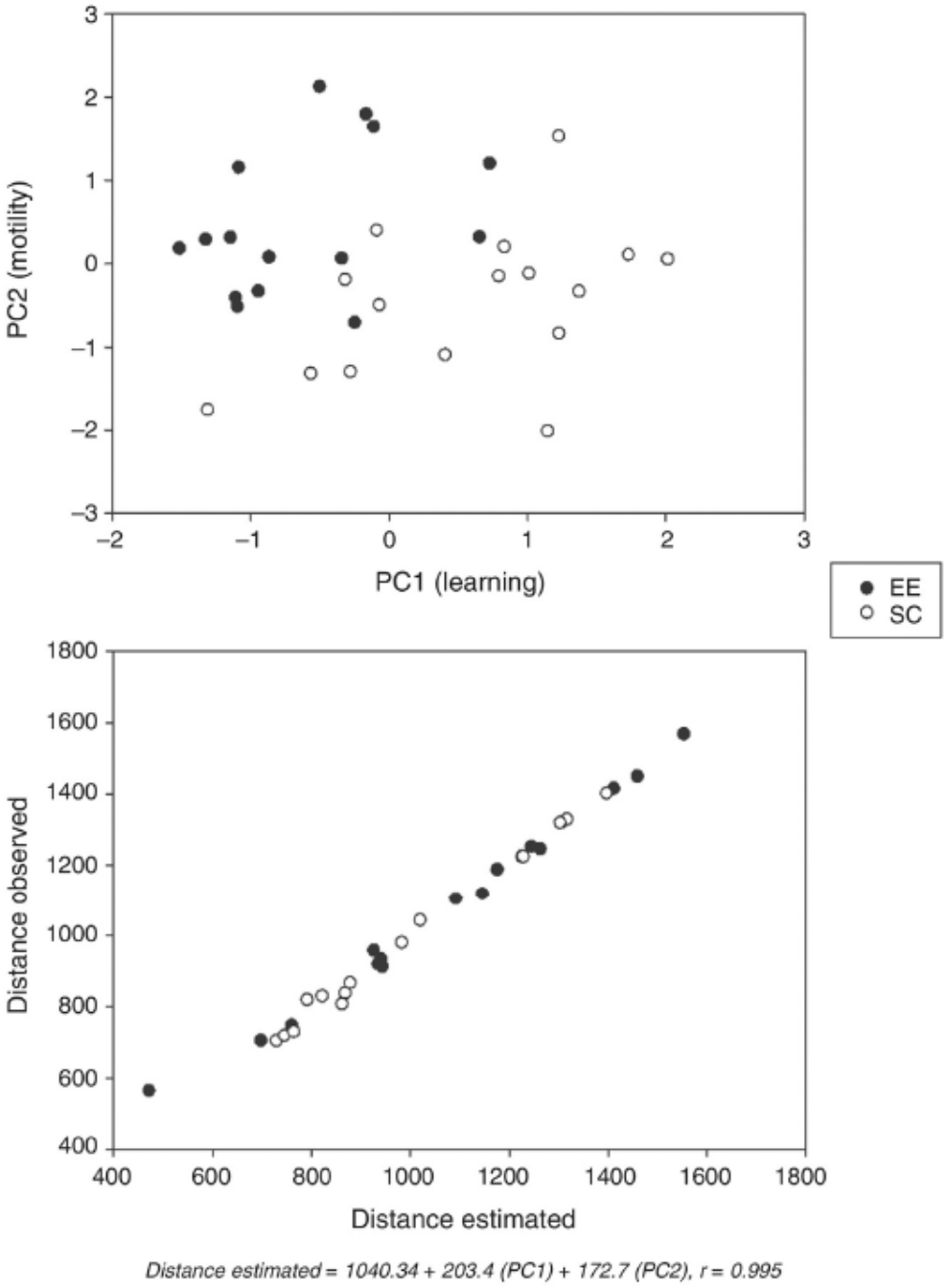


Table 1
Distances of European cities (km) from the main cities of Latium

	Rome	Latina	Frosinone	Viterbo	Rieti
Amsterdam	430	447	449	415	409
Athens	347	321	331	346	364
Barcelona	283	305	293	292	271
Beograd	227	222	236	220	238
Berlin	393	400	409	374	373
Bern	227	249	247	220	205
Bonn	353	370	372	339	330
Bruselles	388	406	406	371	365
Bucharest	364	355	368	359	378
Budapest	268	261	274	246	259
Calais	418	448	446	418	405
Copenhagen	510	522	527	492	491
Dublin	622	645	641	615	600
Edinburgh	637	655	655	625	615
Frankfurt	318	333	336	302	295
Hamburg	435	448	453	417	414
Helsinki	727	729	739	706	713
Istanbul	452	430	443	443	464
Lisbon	615	637	622	624	604
London	474	494	493	464	456
Luxembourg	325	346	346	315	307
Madrid	449	470	458	460	440
Marseille	200	223	213	202	183
Moscow	782	773	785	759	774
Munich	230	245	250	216	213
Oslo	664	675	682	646	645
Paris	365	386	383	357	343
Prague	305	313	320	286	290
Sofia	294	273	286	280	301
Stockholm	653	658	668	632	636
Warsaw	435	433	444	413	421
Vienna	255	254	265	233	240
Zurich	227	246	246	214	205

On the constructive role of noise in spatial systems

Alessandro Giuliani [a], Alfredo Colosimo [b], Romualdo Benigni [a], Joseph P. Zbilut [c,1]

[a] *Laboratory of Comparative Toxicology and Ecotoxicology, Istituto Superiore di Sanità, Viale Regina Elena 299, 00161 Rome, Italy*
[b] *Department of Biochemistry, University of Rome "La Sapienza", Rome, Italy*
[c] *Department of Molecular Biophysics and Physiology, Rush University, 1653 W. Congress, Chicago, IL 60612, USA*

Received 28 May 1998; revised manuscript received 13 July 1998; accepted for publication 13 July 1998
Communicated by C.R. Doering

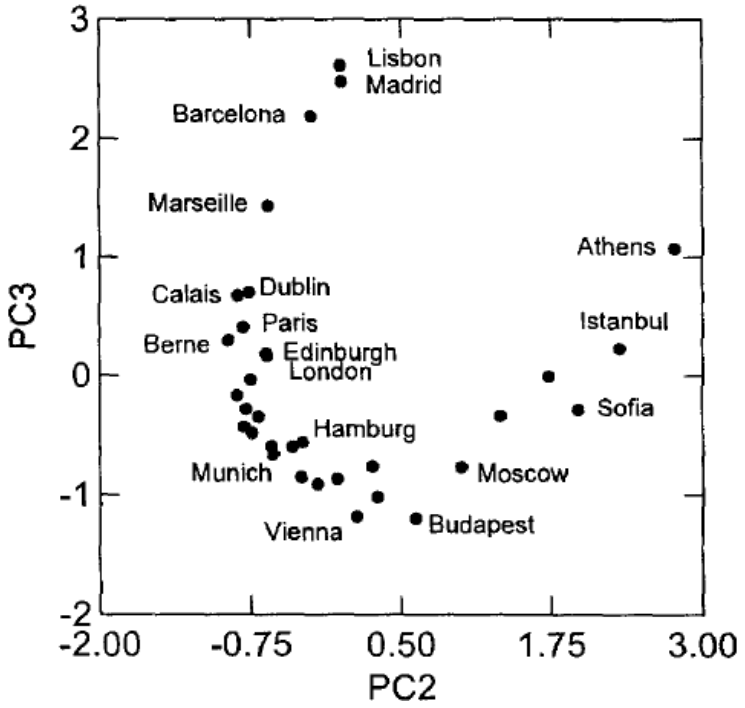
Table 2
Factor loadings and proportions of explained variance

Variables	Components				
	PC1	PC2	PC3	PC4	PC5
Rome	0.9997	0.0137	−0.0184	−0.0120	0.0001
Frosinone	0.9973	−0.0715	0.0132	0.0011	0.0029
Latina	0.9987	−0.0420	−0.0272	0.0058	−0.0024
Rieti	0.9909	0.0162	0.0393	−0.0009	−0.0023
Viterbo	0.9964	0.0837	−0.0070	0.0060	0.0017
Explained variance	0.9965	0.0029	0.000569	0.000043	0.000005

$$\theta = -4.8306 + 41.097(\mathrm{PC2}) - 27.086(\mathrm{PC3}),$$

$$r = 0.97, \quad p < 0.0001. \qquad (2)$$

Thus, PCA broke down the information into "size" (PC1) and "shape" (PC2, PC3) components [6], and separated effects relative to different measurement scales. PC4 and PC5 were not correlated with any meaningful characteristic of the European map, and were considered to be noise derived from the measurement of distances on a European map using a ruler.



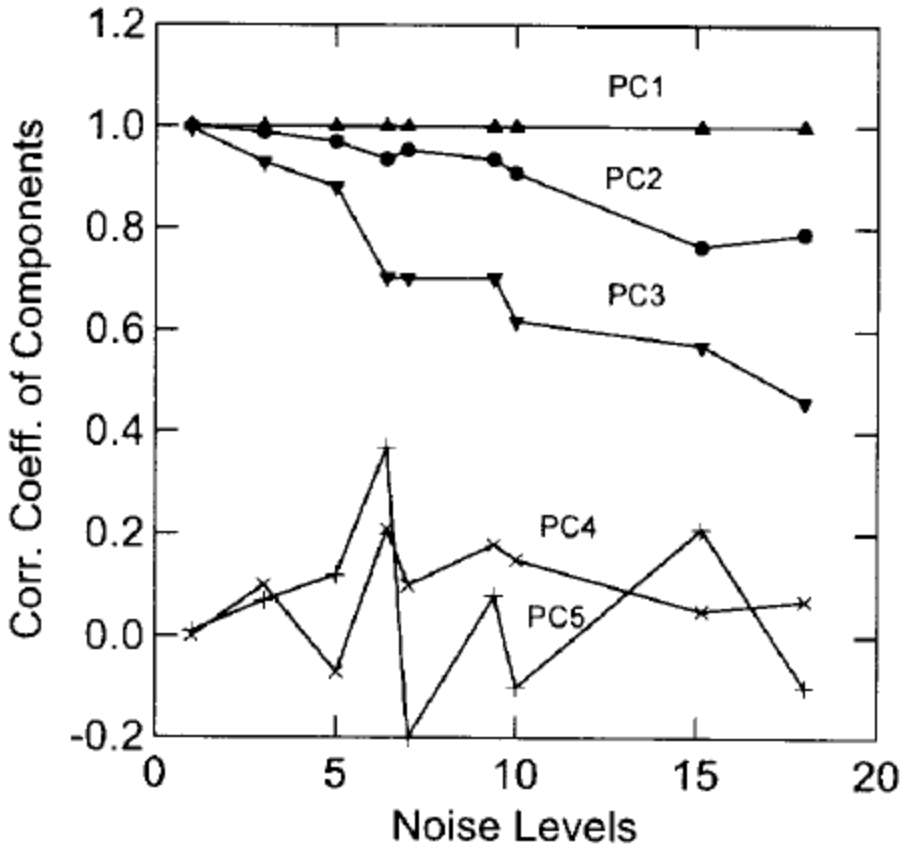
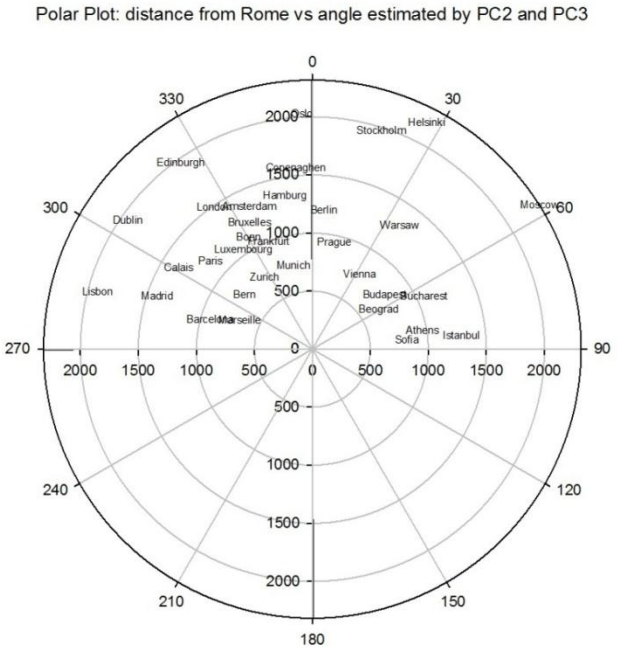
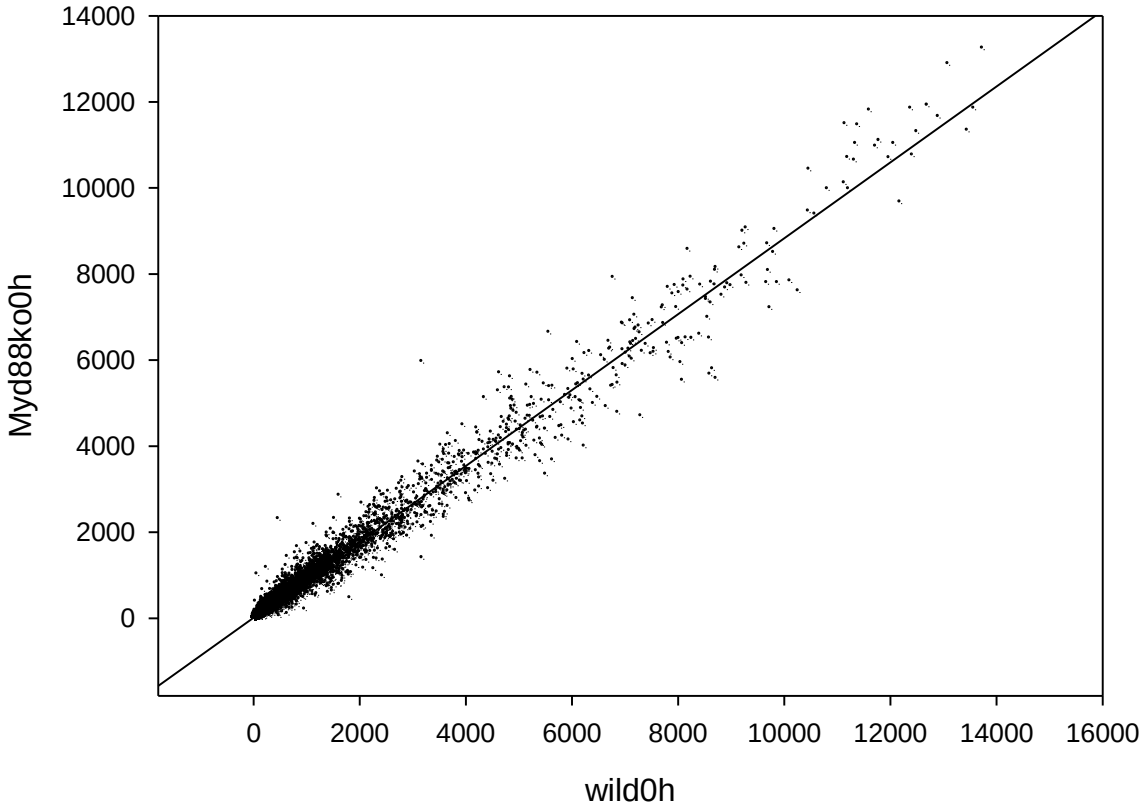


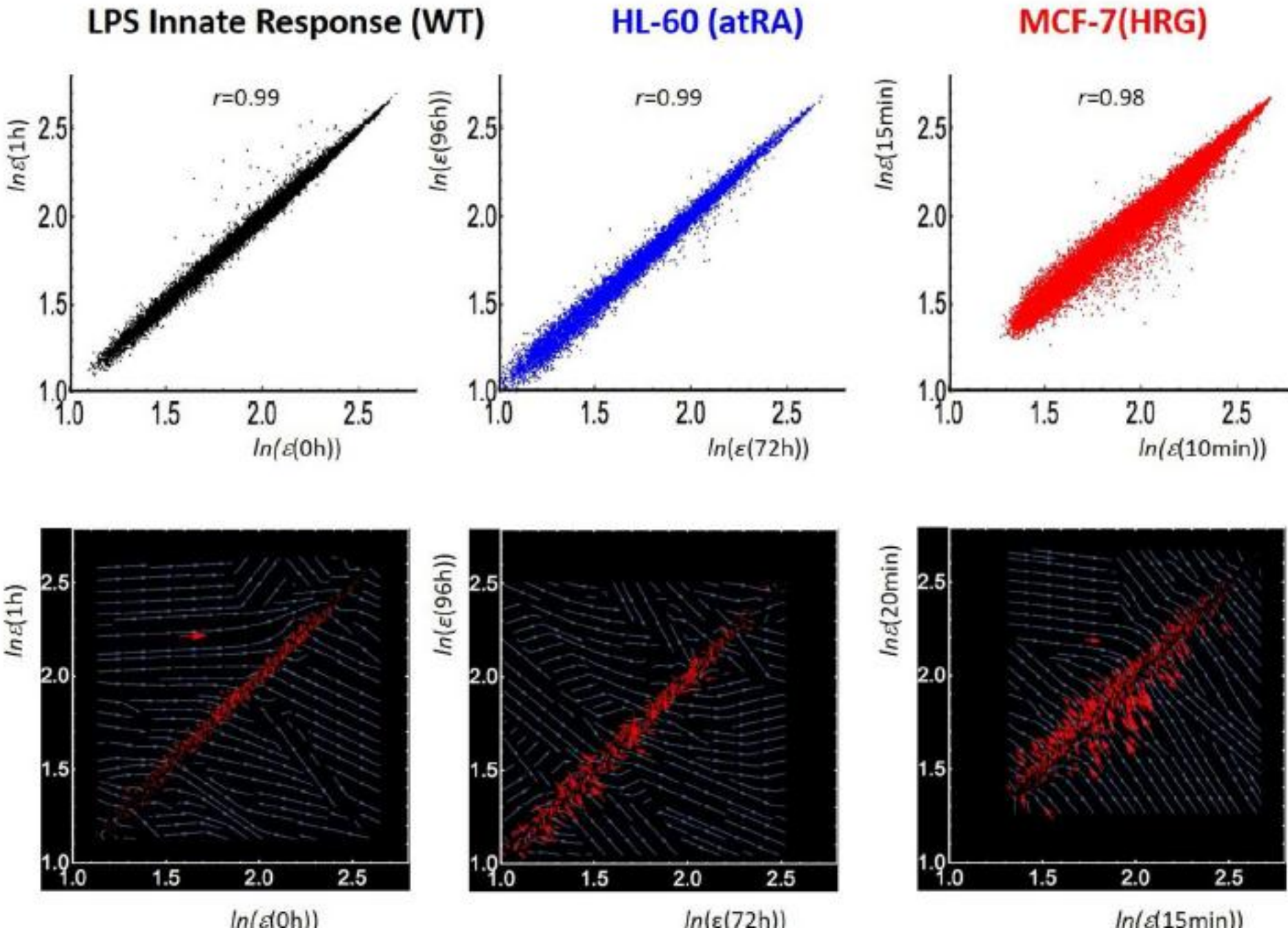
Fig. 3. The figure displays the degree of recognition (Pearson's correlation coefficient) between the noise-corrupted PCs and their original counterparts, for different amounts of noise. The abscissa in expressed in SD units (mm) (1 mm = 3 km).

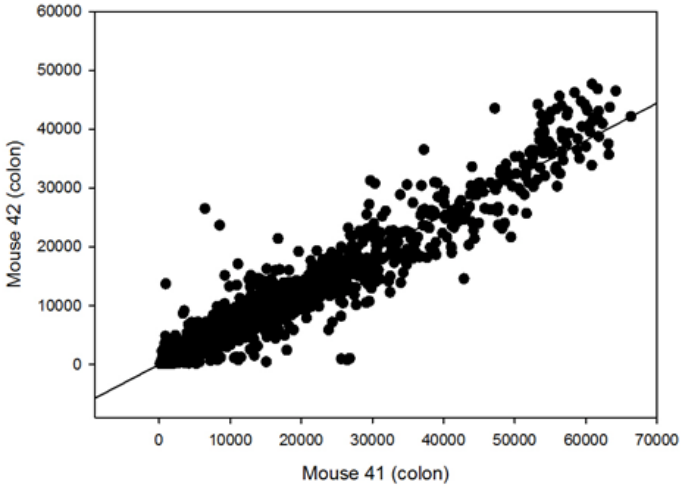


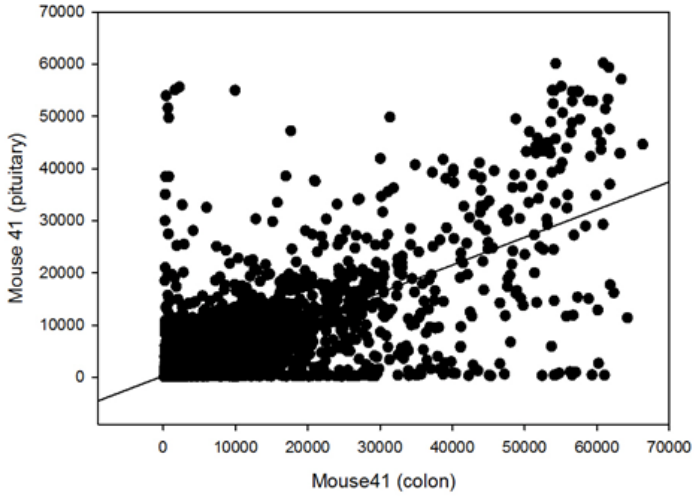
Europe (as seen from Latium) is reconstructed by the polar plot on PC2,PC3 (minor) components

Relation between global gene expression




LPS Innate Response (WT)
HL-60 (atRA)
MCF-7(HRG)
r=0.99
r=0.99
r=0.98
ln ε(1h)
ln(ε(96h))
ln ε(15min)
ln(ε(0h))
ln(ε(72h))
ln(ε(10min))
ln ε(1h)
ln(ε(96h))
ln ε(20min)
ln(ε(0h))
ln(ε(72h))
ln(ε(15min))
1.0
1.5
2.0
2.5


Mouse 42 (colon)
Mouse 41 (colon)
60000
50000
40000
30000
20000
10000
0
0
10000
20000
30000
40000
50000
60000
70000


Mouse 41 (pituitary)
Mouse41 (colon)
70000
60000
50000
40000
30000
20000
10000
0
0
10000
20000
30000
40000
50000
60000
70000