CORRELATO – User guide

INTRODUCTION

A highly forceful and simple statistical method for creating predictive models based on structure-property relationships is presented. The described methodology can be used for a wide range of chemical substances to solving a key problem – the search for the structure of a substance with a given activity or any characteristics.

Within the framework of fundamental research, the search for quantitative relationships between the elements of any complicated system allows one to create forecasting models to achieve a specific result without going deep into routine procedures, which is necessary for significant savings in resources. In the focus of chemical sciences, in addition to the development of synthetic strategies and the accumulation of data on a number of practically important characteristics of objects, there is a search for correlations between the structures of substances and their properties.

ABOUT ALGORITHM

The developed algorithm allows researchers, without special knowledge in the field of mathematical statistics and the logic of creating machine learning algorithms, to find relationships between the properties of chemical objects. It is based on the statistical analysis of data summarized in the input ASCII table. The rows and columns of this table correspond to a set of chemicals and their properties, respectively. Based on such a table, at each iteration, two arrays Y and X are formed: $Y = \{Y_n\}_{n=1}^{N}$ and $X = \{X_n\}_{n=1}^{N}$, which are sequences with an equal number of elements N corresponding to the number of rows (or substances in the table). Through a random combination of values at the intersection of the current n-th row with selected columns, the following calculations are performed at each iteration:

$$Y_n = \prod_i {y_i}^a \text{, and } X_n = \prod_j {x_j}^b, \qquad (1)$$

where y_i and x_j are the properties of the system (real numbers; for example, the photocatalytic activity in the experiment or the band gap in quantum-chemical calculations); i,j – are the numbers of columns in the table (integers; for example: 1,2,3, etc.); a,b – are the degrees that are randomly selected from the limited set of real numbers specified by the researcher.

Now it is needed to estimate how much the obtained sequences $\{Y_n\}$ and $\{X_n\}$ correlate with each other. The simplest way to do this is to calculate the r-Pearson coefficient:

$$r_j = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{(\sum_{i=1}^{N}(X_i - \bar{X})^2(Y_i - \bar{Y})^2)^{1/2}}, \qquad (2)$$

where i – is the index for the Y and X arrays, and j – is the iteration number (the total number of iterations is set by the researcher). The Pearson correlation coefficient r can take values from +1 to -1. The stronger the relationship between data sets (the points are near the line), the greater r, and an r value close to zero indicates no correlation (the points are scattered on the plane). When implementing the described numerical algorithm, iterations with negative values of r are ignored, and among others, a series is selected for which r corresponds to the task (recommended threshold: $r > 0.85$).

ABOUT PROGRAM

The input file for the calculation is an ASCII table, the columns in which are separated by a tab character. The first line is headings, which are recommended to use Latin letters. At the intersection of rows and columns are the properties of substances (activity, equilibrium constant, wavelength, etc.), i.e. the array of considered substances is associated with rows, and the columns with properties.

	Demo Version	Full Version
Distribution	Binary executable for 64-bit Windows (ZIP archive)	PHP source code for Linux (TAR.GZ archive)[1]
Suggestion	Free	Request[2]
Maximum number of columns	2	Unlimited
Maximum number of rows	10	Unlimited
Maximum number of iterations	50	Unlimited
An array of numbers for raising components to a power[3]	–3 to +3 in steps of 1, including 0	Any
Option to exclude a set of elements from the analysis	No	Yes
Special analysis of "outliers"	Simple	Statistical
Open MPI interface support	No	Yes[4]

[1] Requires CLI PHP interpreter to be installed; the source code can be also embedded in web applications;
[2] Contact the developer - Prof. Alexander Yu. Tolbin: tolbin@ipac.ac.ru;
[3] See Eqn. 1;
[4] It is possible to run one or multiple processes per node; requires an installed MPI interface.

The Demo version of CORRELATO is downloaded as a ZIP archive which contains a 64-bit Windows executable correlatto_win.exe, configuration file correlatto_win.ini, and example input file input.txt. The input file for calculations is an ASCII table, with the columns separated by a TAB character. The first row is the column headings. Before starting the program, the user should configure the settings (correlatto_win.ini file):

Parameter	Sample values	Description
Y_cols	1,3	Column numbers (counting from 1) to form X or Y assets. In the demo version, user can use no more than two columns for preparing both X and Y assets (comma delimiter without spaces).
X_cols	2,4	
pearson_threshold	0.6	The threshold for Pearson coefficient. Below this value, results are skipped.
iter_count	50	The maximum number of iterations. For the demo version, no more than 50 iterations are allowed.

Now run correlatto_win.exe. A command window is opening to proceed with calculations. File plot.txt will appear to contain the results of combinations of the selected columns (a series of XY ASCII tables). Below is an example of the calculation (plot.txt):

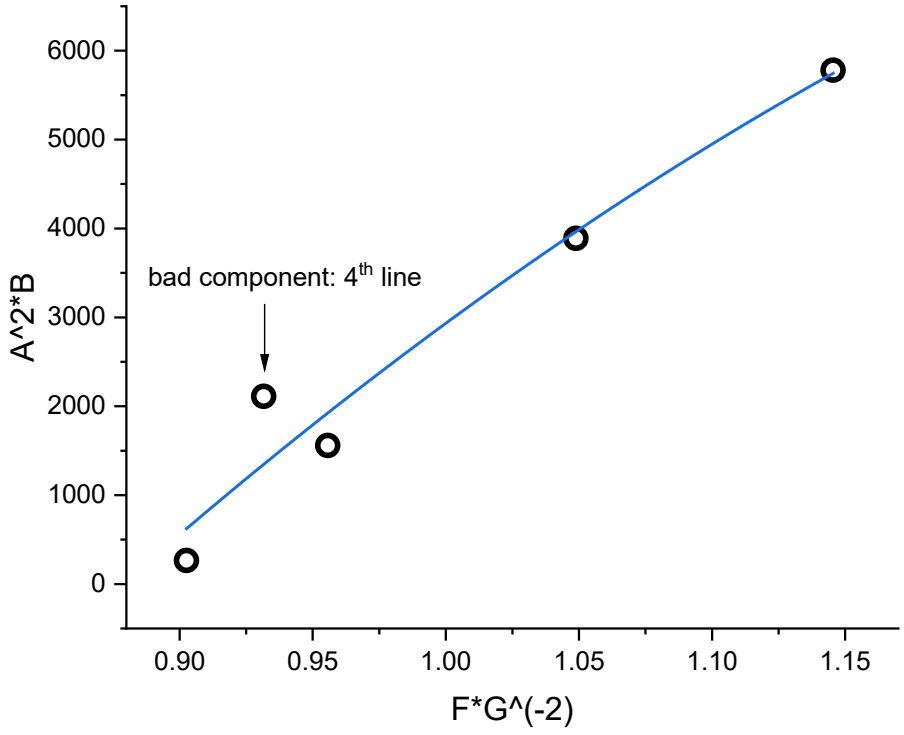
```
Iteration # 10, Pearson: 0.9233, bad component: 2

A^(-2)*B^3 vs. log(G^3)
0.41016170146922    1.0813148788927E-9
0.40061672511065    2.0833333333333E-10
0.41963725920371    5.1939058171745E-10
0.42904440076229    8.4526008002462E-5
0.45686503314917    0.00048979591836735

Iteration # 11, Pearson: 0.9752, bad component: 4

A^2*B vs. F*G^(-2)
1.1455058873675     5780
1.048875432526      3888
0.95568157950011    1559.52
0.93162879768128    2111.85
0.90259869073597    264.6
```

When logarithmic scale is considered, a decimal logarithm is used. The chart for Iteration #11 would look like presented in the following Figure:

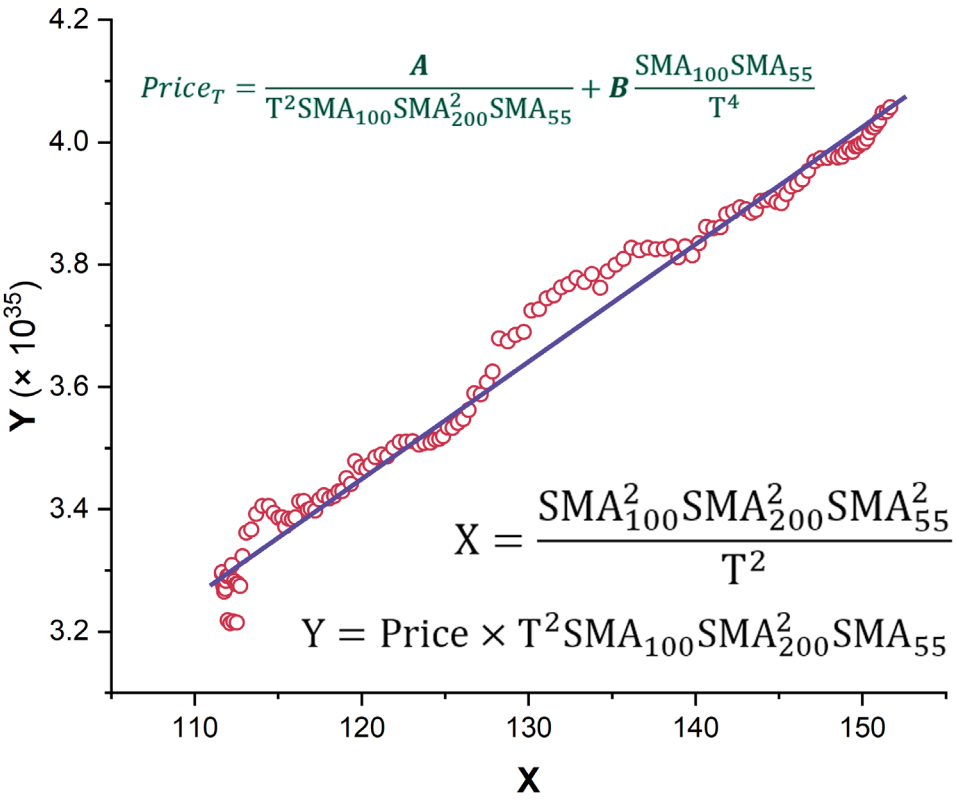


An example of a correlation when looking for relationships between the data presented in the input file (input.txt). Captions of X and Y axes show the analytical form of random combinations between the properties of substances.

Bad components are the rows that can be excluded to improve the relationship. The full version of the CORRELATO program is

available for Linux OS, has no restrictions, and is periodically improved.

The following example demonstrates the ability to predict the stock price of a certain company in the short term:



An example of the correlation of SMA values with the stock price; A and B are the fitting coefficients.

Herein, the time period has been set linearly (15-minute interval) and SMA are the simple moving averages that can be predicted with high accuracy. Closed prices were used in the analysis.

A very important detail should be noted. Thus, correlations provided for functional analysis (high *r*-Pearson) relevant only for the data for which they were established. Using other arrays may give incorrect results. However, if we consider zonal parametric analysis (low *r*-Pearson), there is the possibility to select elements according to their characteristics, including new data that did not participate in the correlation analysis.

CERTIFICATE OF COMPUTER PROGRAM (RU)
==========

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2022613888

Поиск корреляций между неограниченным набором данных – Correlato

Правообладатель: ***Федеральное государственное бюджетное учреждение науки Институт физиологически активных веществ Российской академии наук (RU)***

Автор(ы): ***Толбин Александр Юрьевич (RU)***

Заявка № **2022612994**
Дата поступления **09 марта 2022 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ ***15 марта 2022 г.***

Руководитель Федеральной службы по интеллектуальной собственности

ДОКУМЕНТ ПОДПИСАН ЭЛЕКТРОННОЙ ПОДПИСЬЮ
Сертификат 68b80077c14e40f0a94edbd24145d5c7
Владелец **Зубов Юрий Сергеевич**
Действителен с 2.03.2022 по 26.05.2023

Ю.С. Зубов