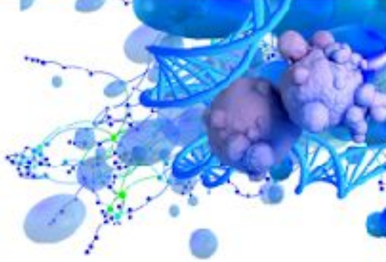


BGRS/SB-2022

13th International Multiconference on Bioinformatics of Genome Regulation and Structure/Systems Biology

Symposium "Big genetic Data Analysis, deep learning, mathematical modeling and supercomputing"

Multidimensional analysis, dimension reduction, categorization with statistical approach - stability and reproducibility

Karpenko Dmitriy Vladimirovich

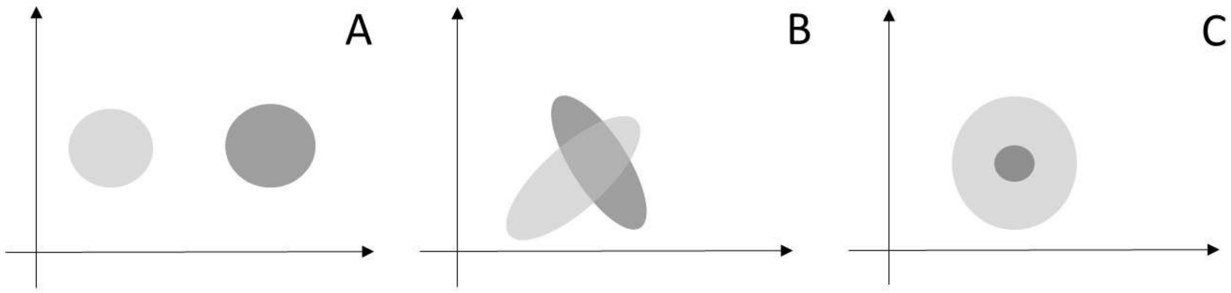
National Medical Research Center for Hematology, Moscow, Russian Federation
Scientist of Laboratory of physiology of hematopoiesis

Moscow Institute of Physics and Technology
Master of applied physics and mathematics (2007)

Medical data

Peculiarities :

- *multiple parameters*
- *sparse data sets*
- *mosaic data*
- *fragmentary data*
- *misleading conventions*
- *individual cases*



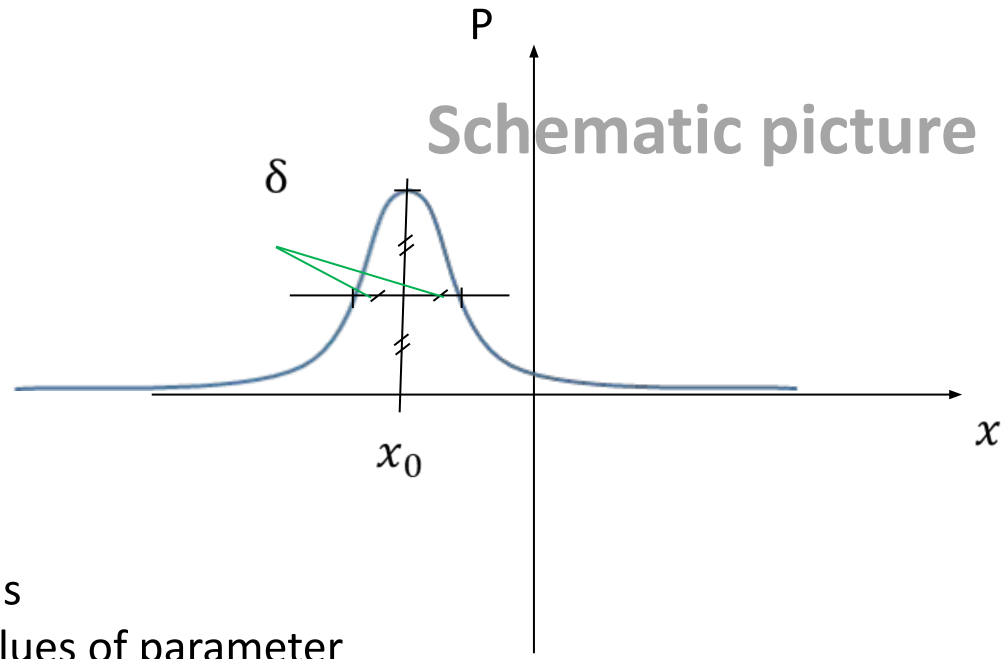
Aims :

- *determine parameters distinguishing groups*
- *predict affiliation of new element*
- *testify individual hypotheses*

Normal distribution

sparse data sets

$$P = N2^{\left(\frac{x-x_0}{\delta}\right)^2}, N = f(\delta)$$



Medical or biological parameters usualy are

- restricted in values as more 0
- ranged in values for several orders

Normal distribution is

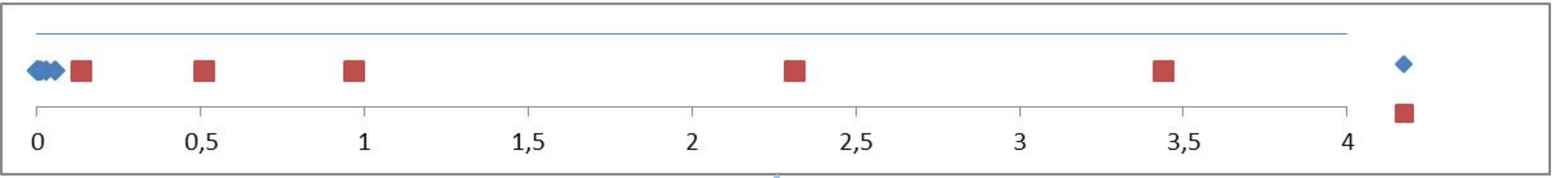
- unrestricted in values of parameter

- Such parameters should be studied in log scale in matter of groups comparison
- That could not be upplied if parameter got values ≤ 0

Groups are determined by two values for each parameter

- geometric mean
- geometric standard deviation factor

Example



Student's *t*-test **p = 0,077 (>0,05)**

◆	🟥
0,002259	0,131861
0,005752	0,968072
0,029597	2,31227
0,057017	3,439049
0,001134	0,507352
0,012853	

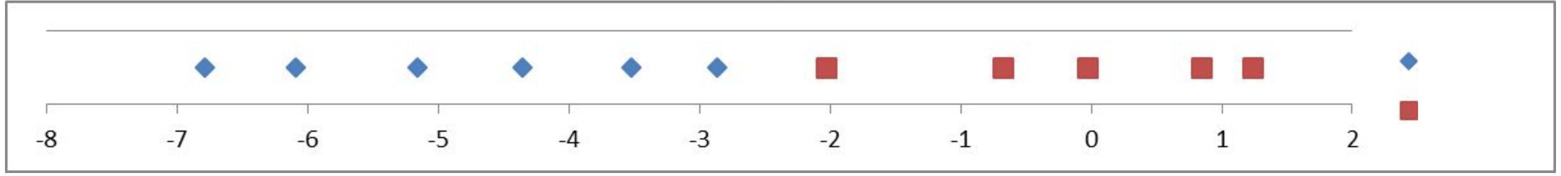
(0;+∞)

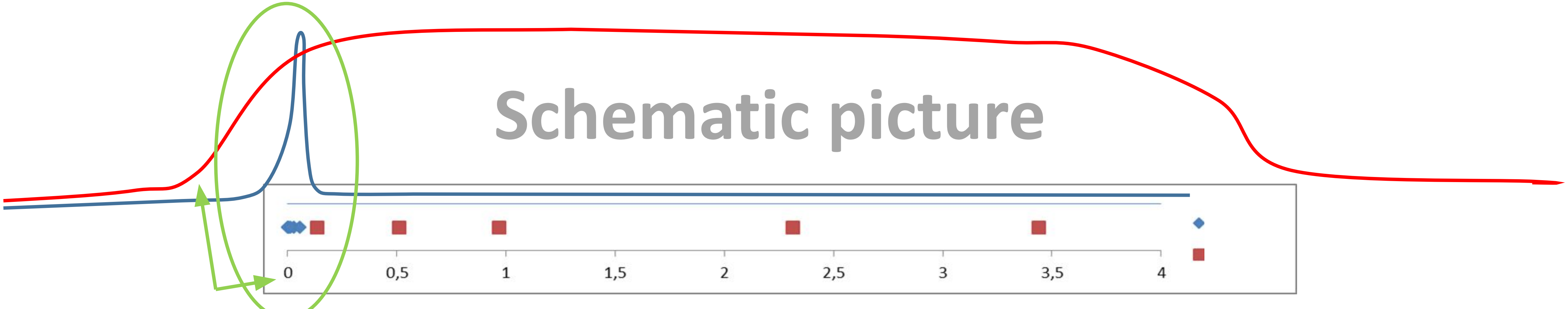
Log(data)

(-∞;+∞)

Student's *t*-test **p = 0,00037 (<0,05)**

◆	🟥
-6,09283	-2,026
-5,15812	-0,03245
-3,52009	0,83823
-2,86441	1,235195
-6,78168	-0,67855
-4,35416	





Robustness and statweight

We can find out that some elements of groups or their parameters could be

- out of place
- false affiliated

$$S_i = A^{\frac{|x_i - G| + \delta_i}{2\triangle}}$$

Statweight is used to make penalty for outrunned values

- Normal-like one-humped function
- Maximal value a bit lower than 1
- Exponentially penalties
  - Deviations from group mean
  - Inaccuracy in each value
- Interpreted as effective number of measurements

- A – adjusting parameter (A = 2)
- x_i - parameter value of element
- G – group mean for parameter
- δ_i - margin for parameter value of element
- $\triangle$ - group mean for $|x_i - G| + \delta_i$

Implementation of Statweight

1. Put S^0 = 1 for all valid values; 0 for other, S = S^0
2. Calculate group mean according to statweight $G = \frac{\sum x_i S_i}{\sum S_i}$
3. Calculate $\triangle = \frac{\sum(|x_i - G| + \delta_i) S_i^0}{\sum S_i^0}$
4. Calculate statweight

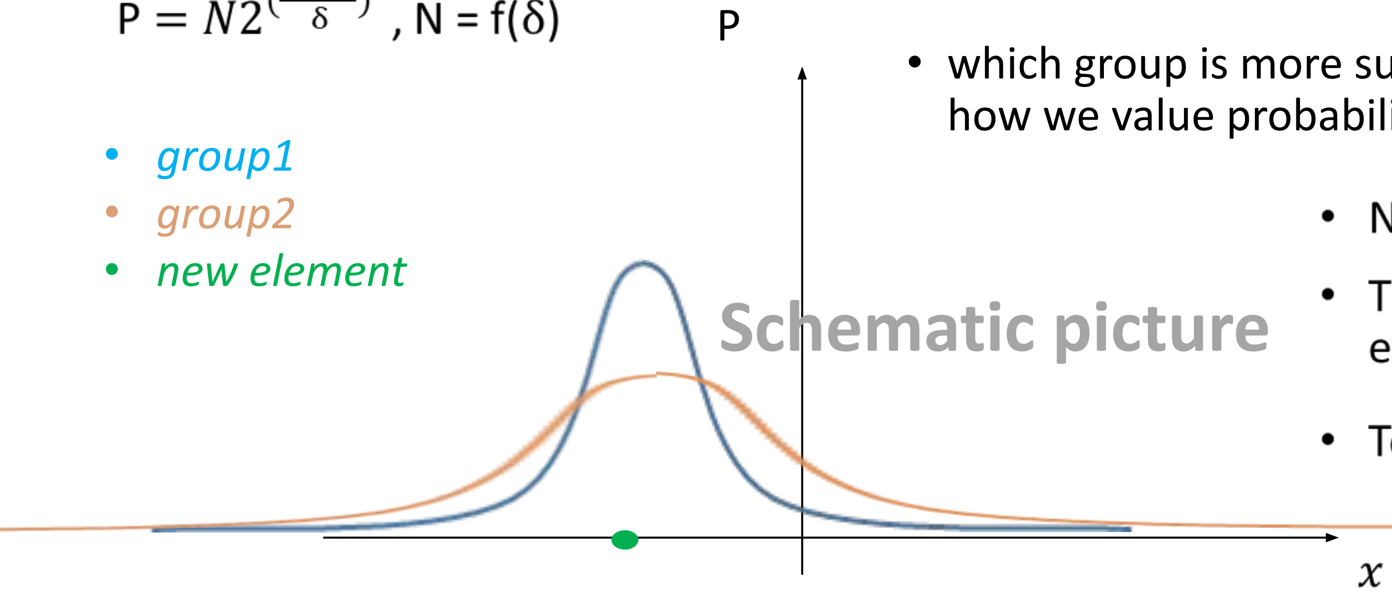
- Recursively proceed to 2-4 until change of G between iterations becomes less than set value

- This allows to utilize margin for each individual value
- Algorithm works the same for group and elements

Binary classification

$$P = N2^{\left(\frac{x-x_0}{\delta}\right)^2}, N = f(\delta)$$

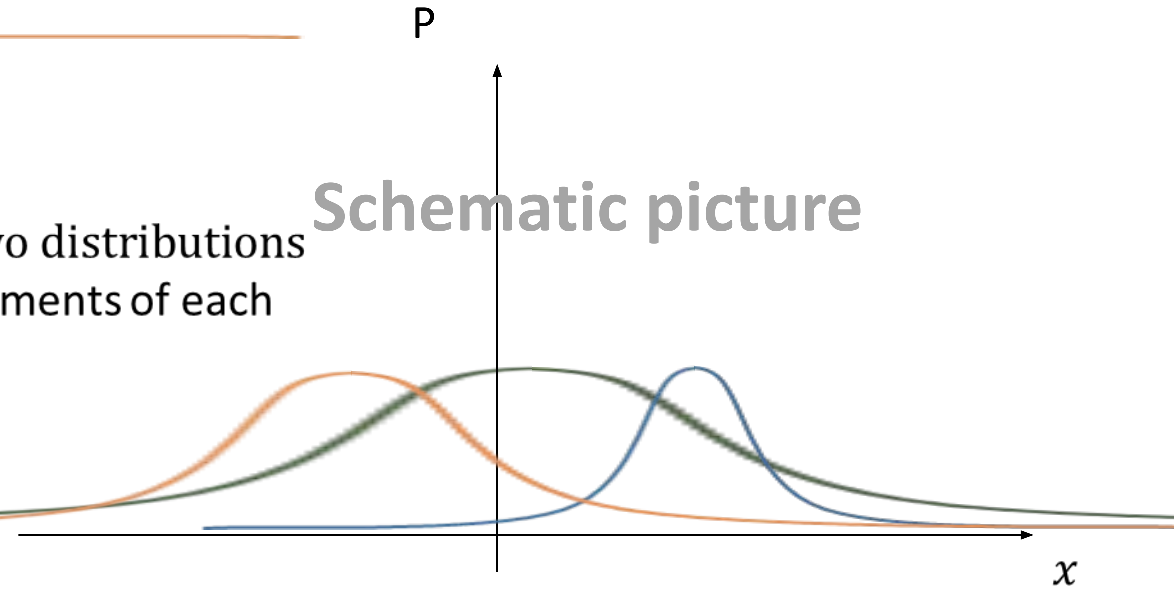
- *group1*
- *group2*
- *new element*



- which group is more suitable for the element and how we value probabilities??

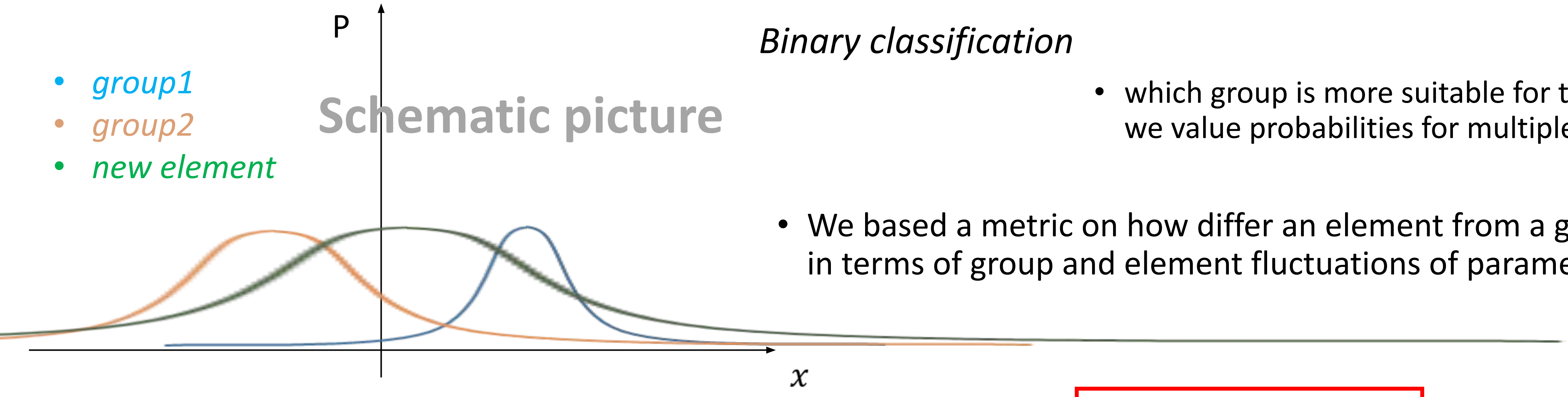
- N should be negated
- That way we compare relative possibilities that the element could appear in appropriate groups
- To simplify, we compare $l = \frac{|x-x_0|}{\delta}$ and $l = \frac{|x-x_0|}{\delta}$

- $l = \frac{|x_1-x_2|}{\delta}$ is a metric
- Considering δ = $\delta_1 + \delta_2$, we can measure distance between two distributions
- That way we can utilize knowledge about perception of measurements of each element
- groups and elements are the same by mean of data proceeding



By comparing l and l we answer the question about affilation of a new element according to given paramater

Binary classification



- which group is more suitable for the element and how we value probabilities for multiple dimensions??

- We based a metric on how differ an element from a group in terms of group and element fluctuations of parameter

multiple parameters

fragmentary data

Summarizing for multiple dimensions:

- Should not be effect from *fragmentary data* so numbers of dimensions
- No one dimension should take domination over others

To meet criteria lets find *relative affilation*

- ratio $R = l/l$ was scaled from [0;+ ∞) to [0;1]
- Mean R was calculated for all dimensions

- Should be taken mean value for classification
- Value of classification for each dimension should be restricted

$$R \leq 1 : R = 0.5 * R$$

$$R \geq 1 : R = 1 - 0.5/R$$

- R value represents numerical classification of new element between two preset groups in range [0;1]
- The R value is an approximation and should not be considered as probability. But it can serve as a certain factor.
- All parameters are putted at same scale

Non numerical data

Discrete data

- for binary state parameters arbitrary pair numbers(>0) could be given
- For multistate parameters each state could be set as a parameter

Data out of measurements range

- Zero or undetected level of parameter could be replaced by estimation of minimal detection level divided by a method accuracy. Enlarged deviation value should be assigned to such cases
- Values, which exceeded maximum value, could be processed the same way

These substitutions should be done with new data with precautious because it can lead to certain artifacts and mistakes in interpretations.

Creating new dependable *parameters*

Certain experimental models and conditions allow to derive definite assumptions that can be formed as new parameters

- difference between control group and affected group
- time effect for same object of study
- etc.. individual cases

Layer of new parameters could also be derived out of data without any certain predictions

- pair linear correlations between parameters could be valued

- correlation of parameters is a distinctive object appearing from certain processes and thus should be described separately
- common pattern for calculations so it could be easily updated and scaled

- For each group element divide parameter value for group mean to normalize
- Normalized values plotting on two dimensional plane for each pair of parameters separately
- For each plot evaluating angles of lines between group center and each group elements
- Calculating mean and deviations for angle
- For each "pair plot" value of angle for new element could be calculated accordingly

Angles or such derived parameters should be calculated in usual , non-logarithmic scale

- Acquired set of new data could be processed as undependable parameters along with primary ones
- We can or can not understand mechanics that arise new parameters

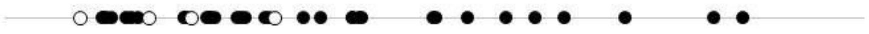
Data representation

- Each element of each group could be considered as a *new element*
- graph for groups can be formed with elements of groups

Graphs of groups elements can be made

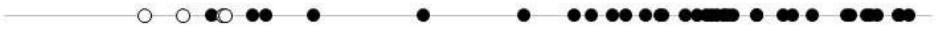
- For *relative affiliations by* each parameter
- For summarized *relative affiliation*

example

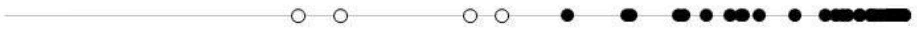


Scalar projection to line between centers of two groups

without fluctuation normalization



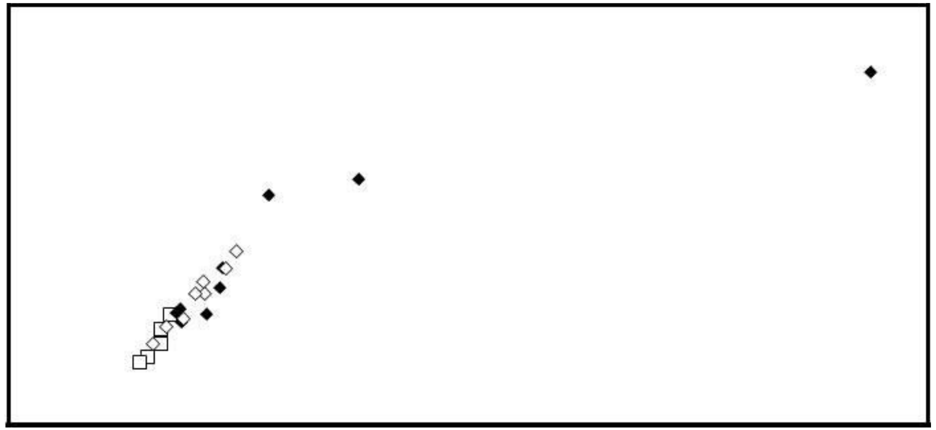
Using proposed metric



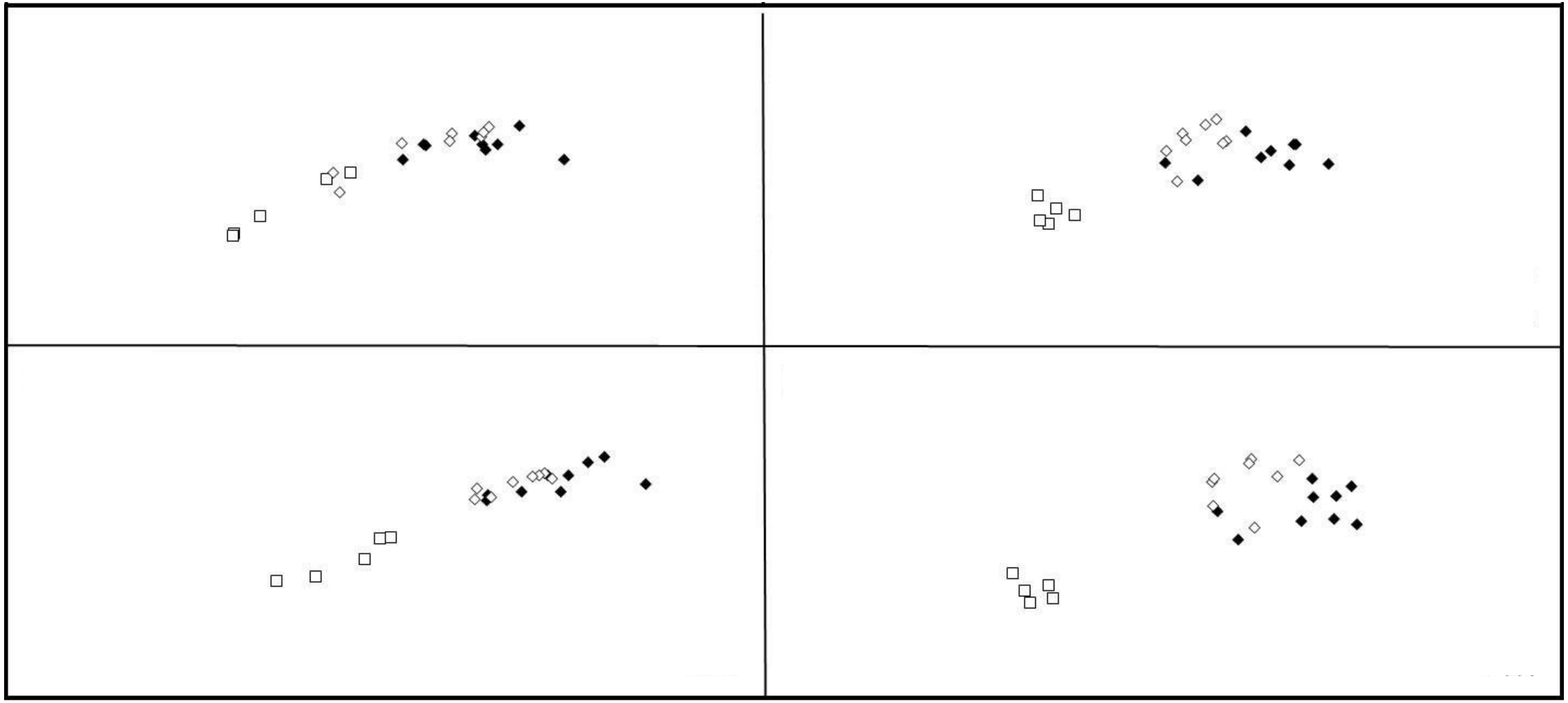
with fluctuation normalization

Scalar projection to line between centers of two groups, two pairs of groups are considered

example



Using proposed metric,
two pairs of groups are considered



without fluctuation normalization

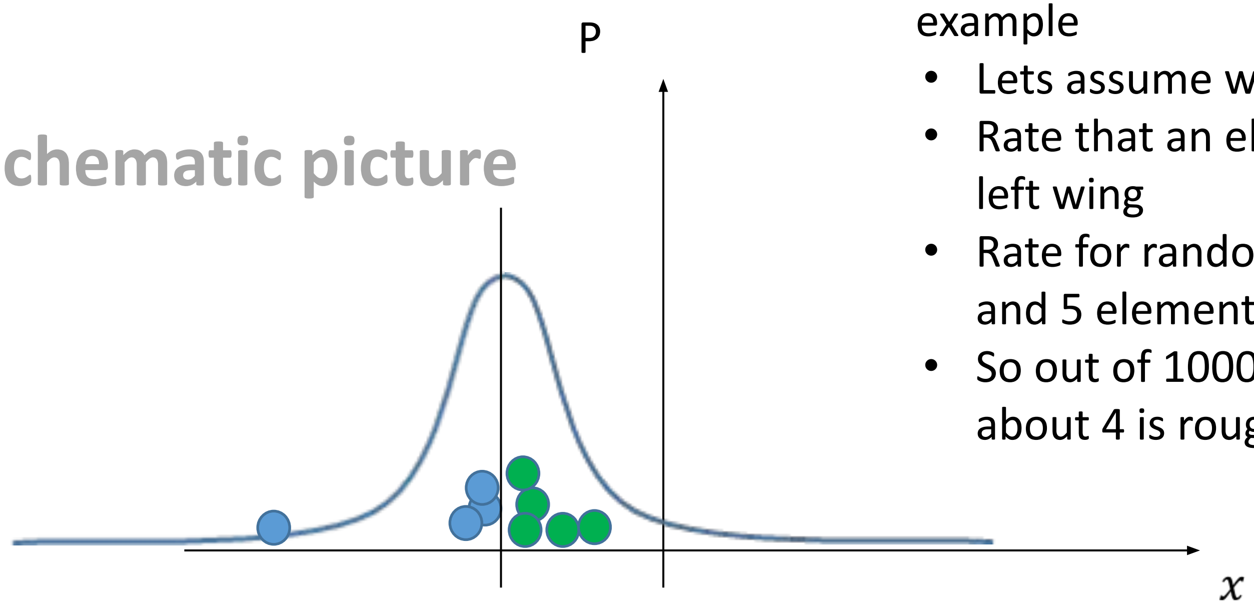
with fluctuation normalization

Without pair correlations

With pair correlations

artifacts

- Sparse data sets with multiple parameters are a perfect source for artifacts



example

- Lets assume we got one general distribution
- Rate that an element go to right wing is 0,5 same for left wing
- Rate for random assigned two colors (4 elements and 5 elements) would separate is $2/2^4 2^5$=0,003906
- So out of 1000 comparison of same distributions about 4 is roughly expected to be false discriminated

to distinct random values separation of groups from consistent

- Get more data
- Studding distribution of values inside groups
- Comparison with other parameters

Questioning single parameter

- Make rank of M parameters basing on P for null hypothesis
- For given parameter recalculate P_i as P_i*M/$rank_i$

Dimensions reduction

- Only part of all parameters discriminates given groups

- Values for *relative affiliations* can be used to test null hypothesis

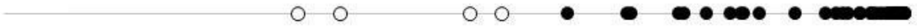
- Rank of M parameters can be formed by null hypothesis tests results for each parameter
- Only given T part of top ranks to be considered
- Result of null hypothesis P recalculate as P = P*M/T

Dimensions reduction

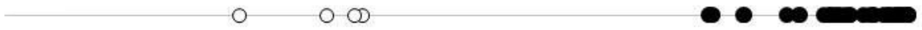
- Makes groups separation better
- Lessens cost of future measurements required for classification of new elements
- Unused parameters should not be forgotten but implied for multiple comparison attenuations

example

without dimension reduction



Using proposed metric
with fluctuation normalization



with dimension reduction

Reproducibility

sparse data sets

Groups rearrangement

- Take one element out from a group
- Find all key values for a data set
- Repeat for each element of each group

That way we acquire rearranged data sets

Rearranged data can be used for:

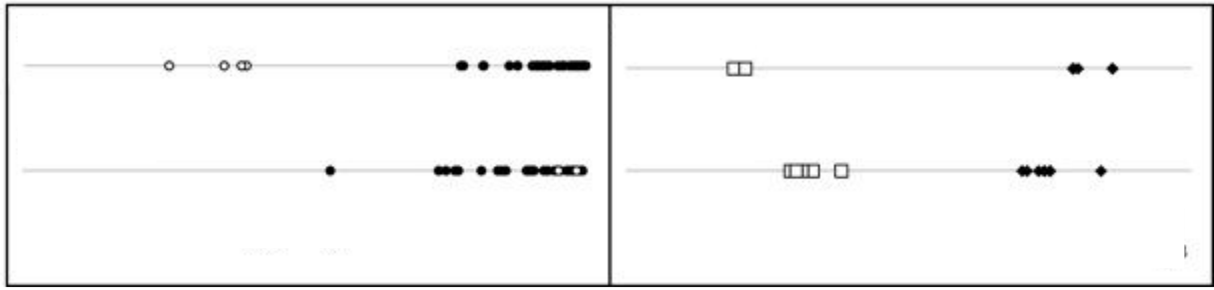
- Finding key parameters for dimensions reduction
  - based on reverse results of test of null hypothesis for rearranged data sets
- Representation of data

Groups rearrangement

- Provides significant revaluation in cases of sparse data sets
- Lessens effect of particular outrunning values of parameter
- Provides more reproducible results

example

without groups rearrangements



with groups rearrangements

fragmentary data

For each rearrangement:

- rearranged data sets are different
  - Set of parameters could be different, due to *fragmentary data*
- But metric for binary affiliation is the same
  - Its’ results could be compared directly

Data representation

- rearrangements
- dimension reductions

Evaluation of groups separation

- summarizing relative affiliations on rearranged data
  - For groups elements

affiliating new element

- summarizing relative affiliations on full dataset
  - For groups elements
  - For new element

Affiliation value

- Not probability
- It is nodded to 0,5 as dimensions grow by unseparating parameters

Lets consider affiliation as a parameter to calculate final affiliation

- It is done on parameter restricted to [0;1]
  - Adjusting it to (-∞;+∞)
- Edged values referred as definite for appropriate group
- So for affiliation is used Fermi-Dirac like step functions
- Affiliation for given point is ratio of f_i and f_i
  - And again adjusting from [0;+∞) to [0;1]

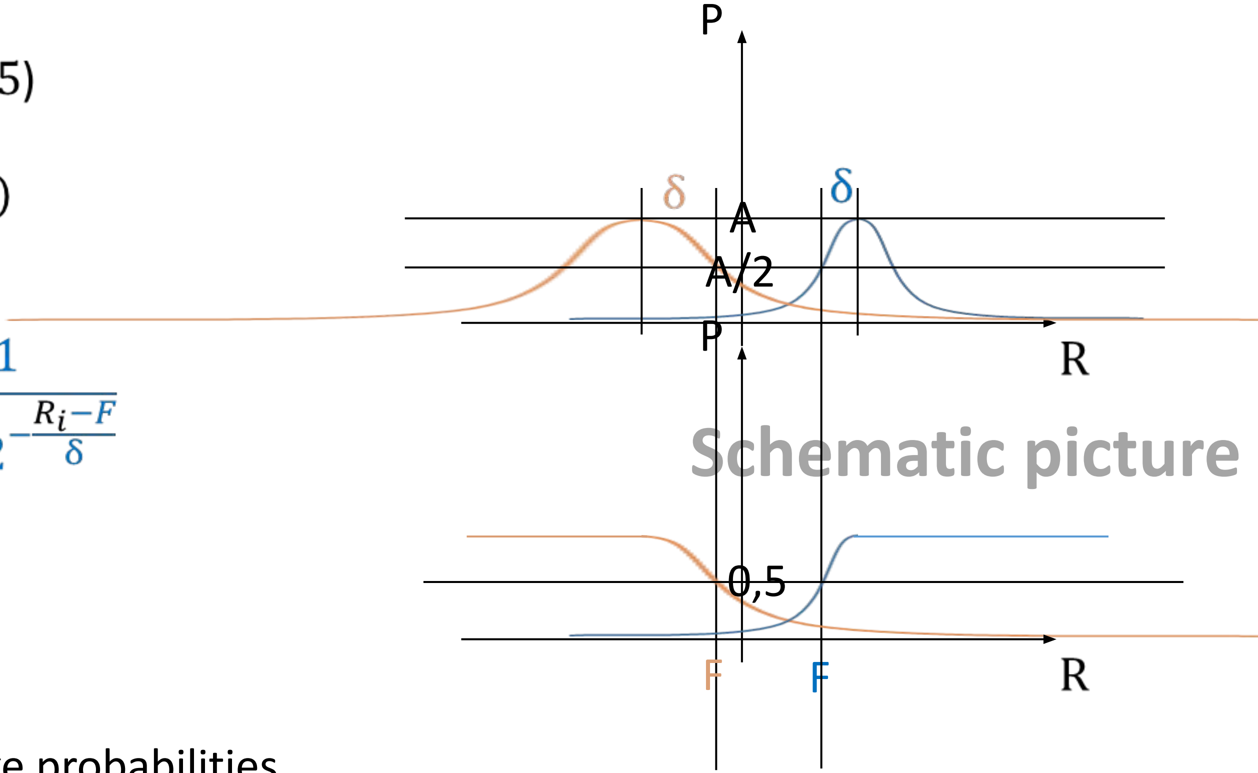
$$\mathrm{R} \leq 0{,}5 : \mathrm{R} = \log(\mathrm{R}/0{,}5)$$

$$R \geq 0{,}5 : \mathrm{R} = \log\left(\frac{0{,}5}{1-R}\right)$$

$$f_i = \frac{1}{1+2^{\frac{R_i-F}{\delta}}} \qquad f_i = \frac{1}{1+2^{-\frac{R_i-F}{\delta}}}$$

$$\mathrm{R} \leq 1 : \mathrm{R} = 0.5 * \mathrm{R}$$

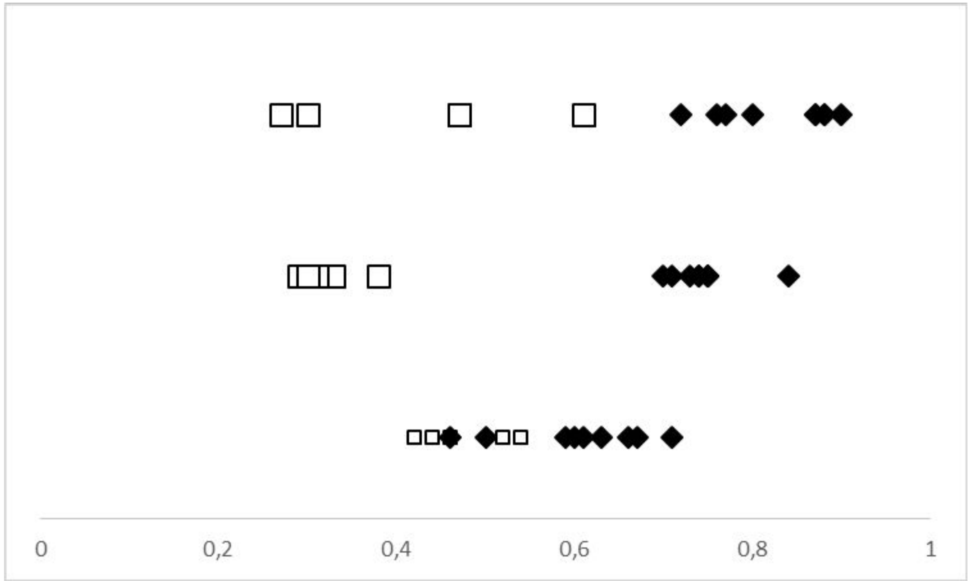
$$\mathrm{R} \geq 1 : \mathrm{R} = 1 - 0.5/\mathrm{R}$$



This values ranged from [0;1] could be considered like probabilities

Prevalence of groups is ignored , but should be also taking into final consideration *example*

Summarized relative affiliation to one of two groups for groups elements

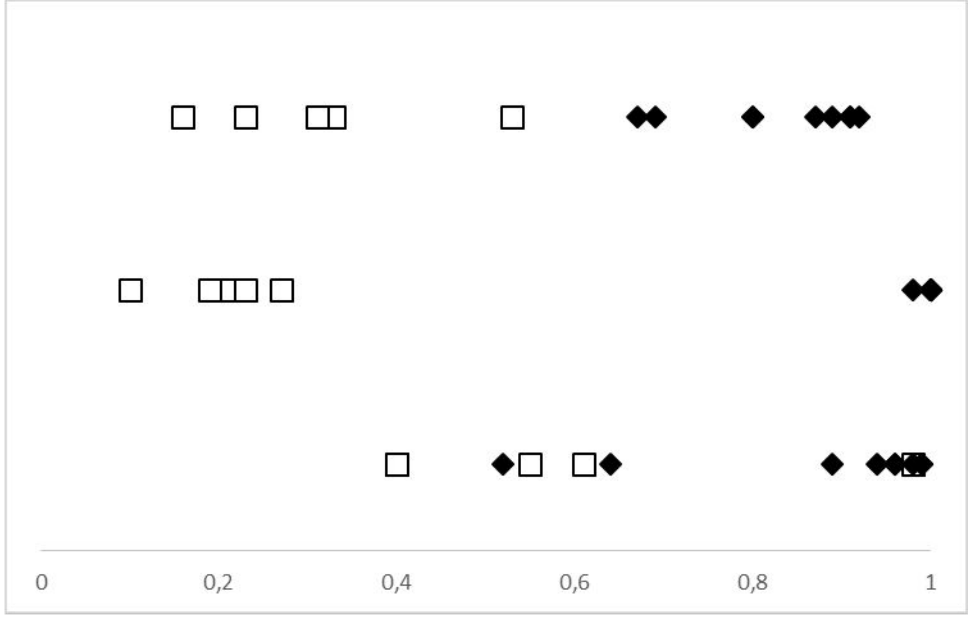


Top 5 primary parameters considered

51 parameters considered

All 378 parameters considered

Estimation for probabilities of affiliation to one of two groups for groups elements



Top 5 primary parameters considered

51 parameters considered

All 378 parameters considered

conclusions

Key approaches

- Statistical
- Suggested alternative metrics
- Implied statweight concept
- Usage uncertainties of individual values
- Ranged values
- Dependencies isolation
- Data rearrangement

Created tool

- Utilizing margins for individual values
- Robustness with statweight usage
- Work with fragmented data sets
- Work with sparse data sets
- Working with nonparametric data
- Allow to interpreter outranged data
- Allow to interpreter ranged values
- Allow to take dependable parameters in calculations
- Allow to distinguish each individual parameter including dependencies it self
- Parameters calculated the way not to shade each other
- Provide evaluation of two groups similarities based on multiple parameters
- Provide criteria to determine key parameters to distinguish two groups
- Categorizes new element between two groups
- Provide self check for categorizing
- Have parser to work with incoming data in excel format
- Can be run as an executable

Working environment

- Windows
- Python 3.7

Acknowledge

- Bigildeev Aleksey for work discussions
- Colleagues of laboratory for data sets to analyze

limitations

- Multimodal distributions
- Distributions with difficult topology

They should be dealt with appropriate methods, which require more than sparse data. But introduced approaches can be locally utilized to solve them.

contacts

- Karpenko Dmitriy (rus/eng)
- 89268784636 tel., whatsapp
- D_@list.ru e-mail