Zbigniew Ziembik

Independent Chair of Biotechnology and Molecular Biology
Opole University, Poland

Misinterpretation of correlation coefficients - false covariabilities in compositional data

Correlation between concentrations – expected outcome

Identification of the pollution source on the base of the **deposit composition** and **relationships between concentrations** of the components.

Assumption

Concentrations of components in environment are strictly related to abundances (masses) of the materials emitted from source.

The source can be identified through analysis of the deposit composition changes in space and/or in time.

Limitation

A part (sample) of the whole object can be investigated. Only **concentration of the components can be determined**.

Compositiona Data (CoDa)

Compositiona data – any vector $\mathbf{x}$ with non-negative elements x_1, …, x_D representing proportions of some whole, subjected to the constraint:

$$x_1 + x_2 + \ldots + x_{D-1} + x_D = 1$$

The closure operation C applied on quantity $\mathbf{w}$:

$$\mathrm{C}(\mathbf{w}) = \left[\frac{w_1}{\sum_{j=1}^{D} w_j}, \frac{w_2}{\sum_{j=1}^{D} w_j}, \ldots, \frac{w_{D-1}}{\sum_{j=1}^{D} w_j}, \frac{w_D}{\sum_{j=1}^{D} w_j} \right] = \left[x_1, x_2, \ldots, x_{D-1}, x_D \right] = \mathbf{x}$$

Units

Amount, abundance (w) – mass, volume, mol, count describe absolute quantity [kg, m^3, mol, -]

Content – fraction, ratio [%, mg/kg, ppm].
Concentration [mol/dm^3, g/dm^3, mol/kg, Bq/kg] describe relative quantity. All concentrations can be transformed to fractions.

Fractions of all D components in a system contain common denominator:

$$\sum_{j=1}^{D} \text{amount}_j$$

Influence of common divisor on correlation

Karl Pearson 1897

Covariance, variance and correlation

The covariance between two jointly distributed real-valued random variables x and y is defined:

$$\operatorname{cov}(x,y) = E[(x - E[x])(y - E[y])]$$
$$\operatorname{cov}(x,x) = E[(x - E[x])^2] = \operatorname{var}(x)$$

Pearson's correlation coefficient ρ :

$$\rho(x,y) = \frac{\operatorname{cov}(x,y)}{\sqrt{\operatorname{var}(x)\operatorname{var}(y)}}$$

To estimate ρ in population the R parameter is calculated from sample.

Testing $\rho = 0$

It is assumed that the x and y values originate from a bivariate normal distribution, and that the relationship (if exists) is linear. Given a sample of n points (x_i, y_i) the estimated correlation coefficient R is calculated from the formula:

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

Testing H_0: $\rho = 0$ (H_1: $\rho \neq 0$) at confidence level 1-α. The test statistics t_R is calculated and it follows Student's t-distribution with $n-2$ degrees of freedom.

$$t_R = R\sqrt{\frac{n-2}{1-R^2}}$$

If $|t_R| < t(1-0.5\alpha, n-2)$ then accept H_0, else reject null hypothesis.

Testing ρ =0

For a given α and n the critical value R_α can be calculated.

If the calculated from data $|R| < R_\alpha$ then H_0 is accepted (at the α confidence level).

If the calculated from data $|R| > R_\alpha$ then H_0 is rejected and H_1 is accepted.

Let A, B, and C be some quantities. They are normally distributed and uncorrelated, i.e. $\mathrm{E}(\mathrm{cor}(A,B))=\mathrm{E}(\mathrm{cor}(A,C))=\mathrm{E}(\mathrm{cor}(B,C))=0$.

What if the absolute quantities are transformed to the relative?

Does common divisor affect inference about correlation?

$$\mathrm{cor}(A,B) \leftrightarrow \mathrm{cor}\left(\frac{A}{C},\frac{B}{C}\right)$$

Simulation

In demonstration the simulated data were used. The number N samples composed of 3 groups with n normally distributed numbers were considered. The A/C and B/C ratios were calculated.

Schemes of data structures

no	A	B	C
1	a_1	b_1	c_1
2	a_2	b_2	c_2
...	...	...	...
n-1	$a_{n\text{-}1}$	$b_{n\text{-}1}$	$c_{n\text{-}1}$
n	a_n	b_n	c_n

no	A/C	B/C
1	a_1/c_1	b_1/c_1
2	a_2/c_2	b_2/c_2
...	...	...
n-1	$a_{n\text{-}1}/c_{n\text{-}1}$	$b_{\text{n-}1}/c_{n\text{-}1}$
n	a_n/c_n	b_n/c_n

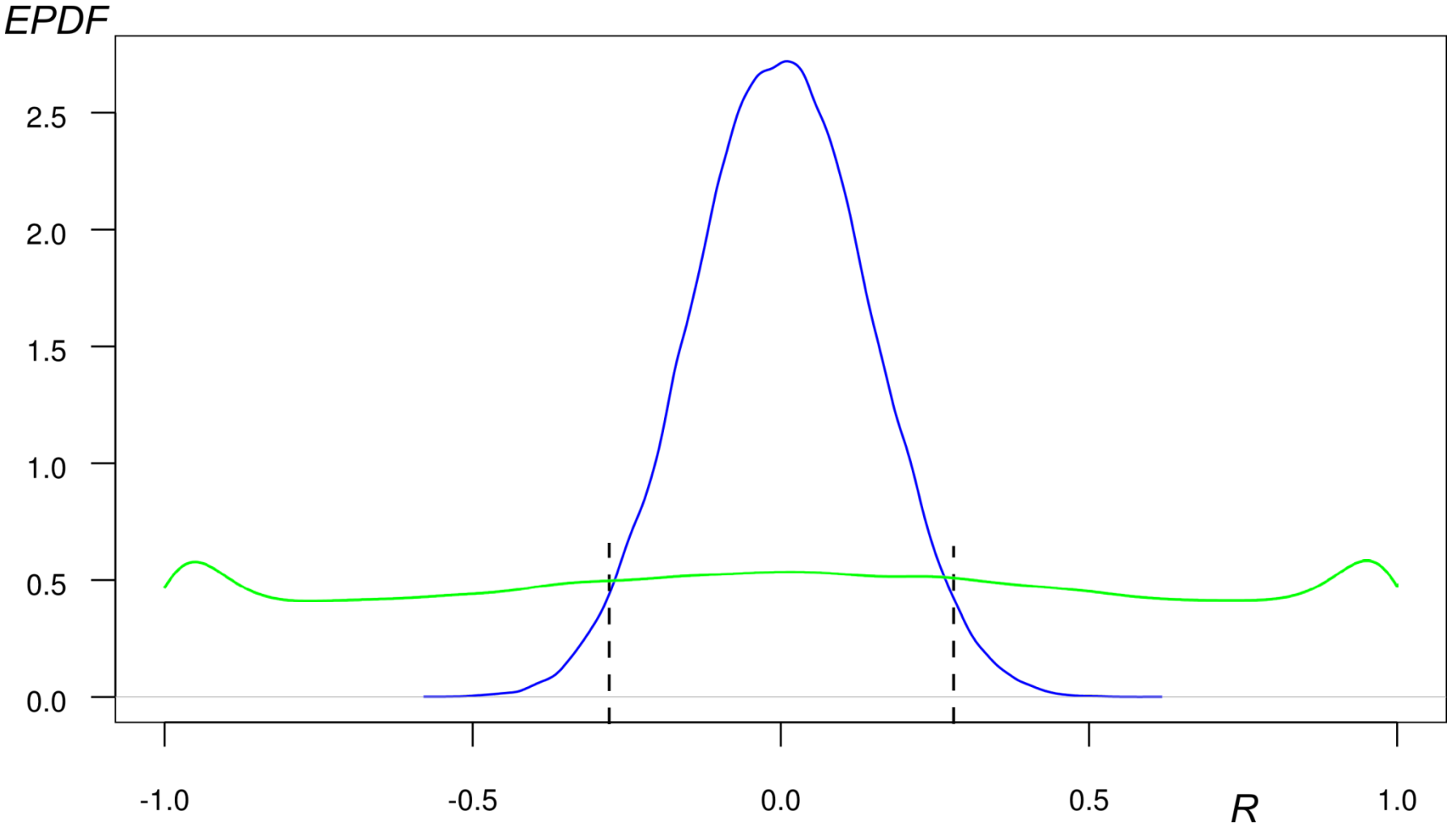
Repeated N times

Estimated Probability Distribution Function of ρ for the true H_0: ρ=0

Number of cases in measurement n=50, number of repetitions N=10^5

$|\rho_{0.05}|$=0.279, cumulated EPDF in 0.025 – 0.975

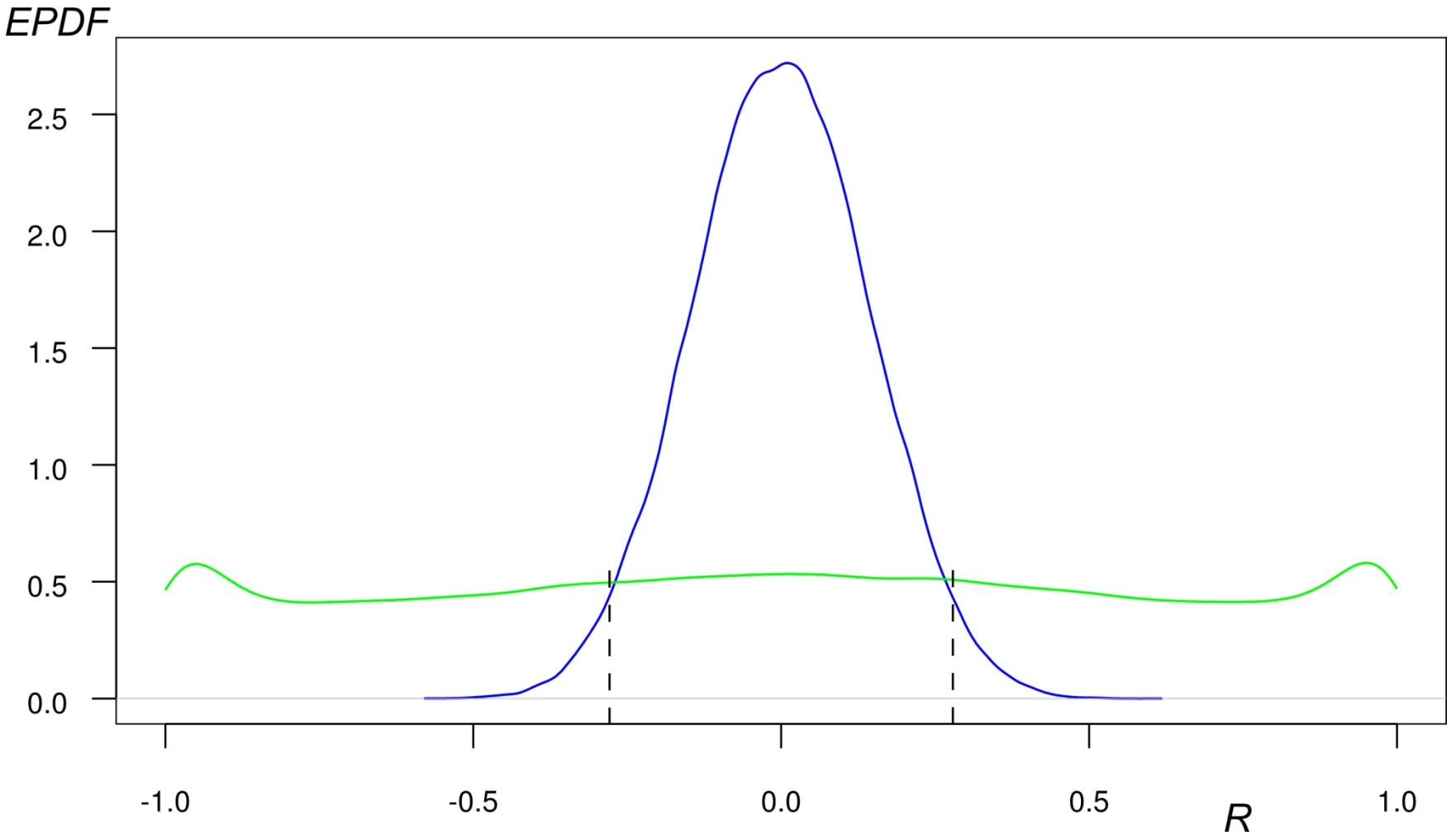
cumulated EPDF in 0.354 – 0.643



Number of cases in measurement n=30, number of repetitions N=10^5

$|\rho_{0.05}|$=0.361, cumulated EPDF in 0.024 – 0.975

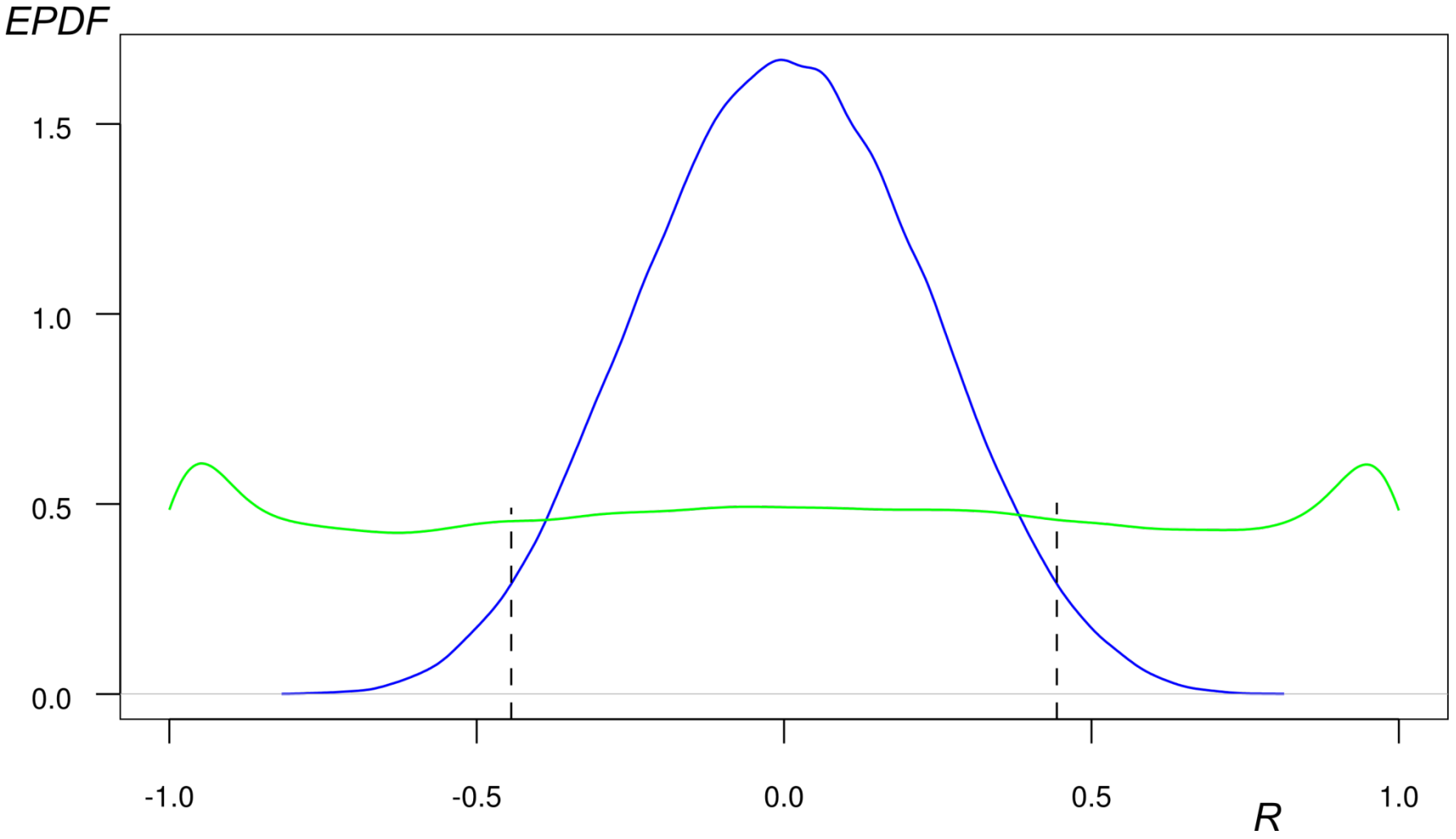
cumulated EPDF in 0.321 – 0.680



Number of cases in measurement n=20, number of repetitions N=10^5

$|\rho_{0.05}|$=0.444, cumulated EPDF in 0.025 – 0.974

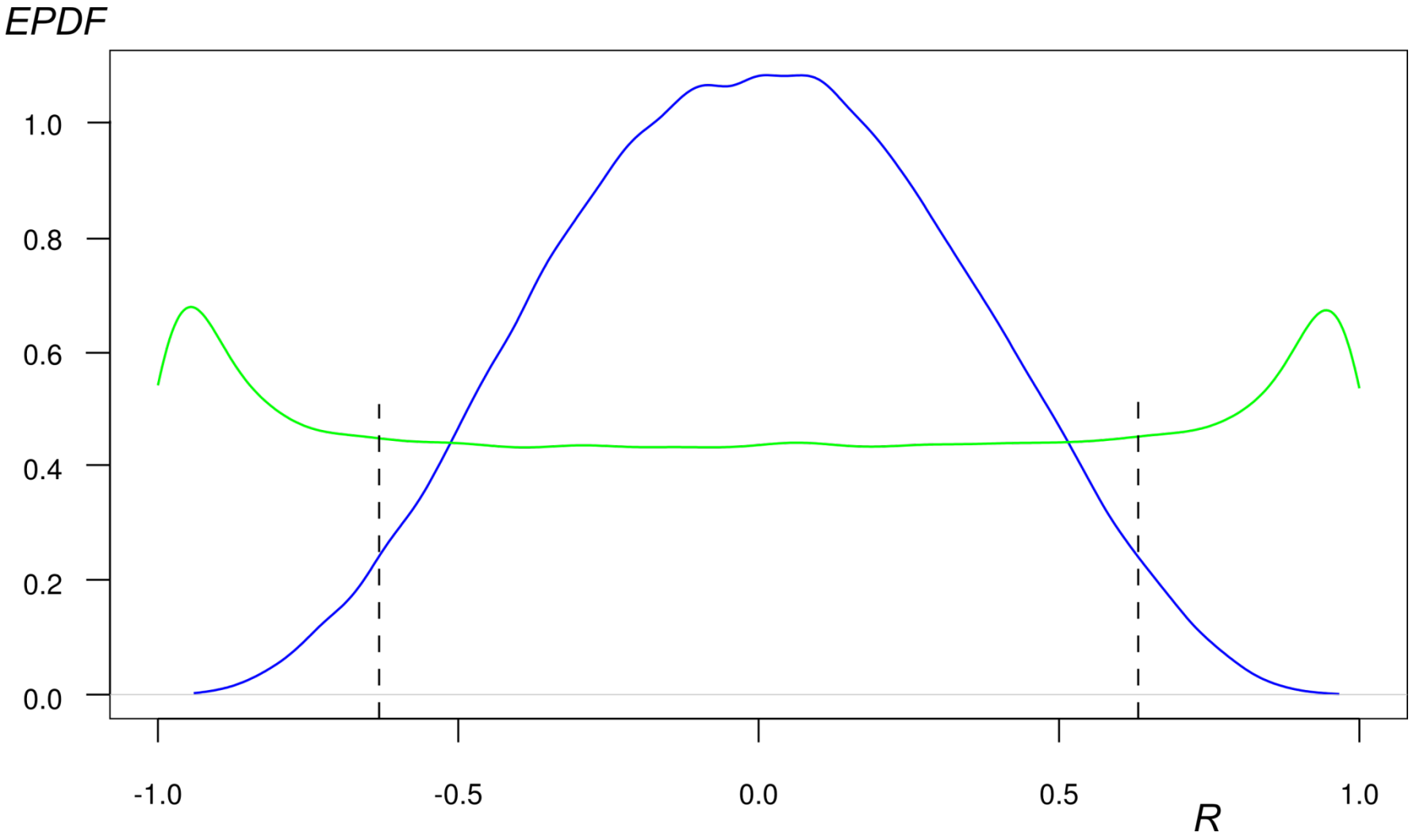
cumulated EPDF in 0.288 – 0.713



Number of cases in measurement n=10, number of repetitions N=10^5

$|\rho_{0.05}|$=0.631, cumulated EPDF in 0.025 – 0.975

cumulated EPDF in 0.226 – 0.775



1st Conclusion

Artificial correlations are observed in compositional data.

It is the effect of:

$$\mathrm{cov}(x_k, x_1) + \ldots + \mathrm{cov}(x_k, x_D) = -\mathrm{var}(x_k)$$

More than 2 variables

5 variables x and subcompositions of 4 variables are considered in simulations, with 3 different relationships:

- all variables are uncorrelated E(ρ)=0,
- variables number 1 and 2 are positively correlated E(ρ)=1, for the remaining ones E(ρ)=0,
- variables number 1 and 2 are negatively correlated E(ρ)=-1, for the remaining ones E(ρ)=0.

Variables 1-4 were chosen from normal distribution with mean 10 and standard deviation 1. Variable 5 was chosen from normal distribution with mean 1000 and standard deviation 100 – it represents the main component of the investigated system.

Uncorrelated variables E(ρ)=0

	$x1$	$x2$	$x3$	$x4$	$x5$
$x1$	1.00				
$x2$	0.08	1.00			
$x3$	-0.04	0.00	1.00		
$x4$	0.05	-0.03	-0.20	1.00	
$x5$	0.08	0.01	0.04	-0.15	1.00

Correlations coefficients between variables $x1$.. $x5$

	$x1$	$x2$	$x3$	$x4$	$x5$
$x1$	1.00				
$x2$	0.55	1.00			
$x3$	0.51	0.48	1.00		
$x4$	0.62	0.53	0.47	1.00	
$x5$	-0.83	-0.80	-0.76	-0.82	1.00

Correlations coefficients between variables C($x1$ .. $x5$)

	$x1$	$x2$	$x3$	$x4$
$x1$	1.00			
$x2$	-0.32	1.00		
$x3$	-0.35	-0.33	1.00	
$x4$	-0.20	-0.36	-0.43	1.00

Correlations coefficients between variables C($x1$ .. $x4$)

Critical R values for n =100:

$R_{0.05}$=0.20

$R_{0.01}$=0.26

$R_{0.001}$=0.32

Correlated variables $x1$ and $x2$, E(ρ)=1

	$x1$	$x2$	$x3$	$x4$	$x5$
$x1$	1.00				
$x2$	1.00	1.00			
$x3$	0.11	0.11	1.00		
$x4$	-0.10	-0.10	-0.07	1.00	
$x5$	0.00	0.00	0.07	-0.09	1.00

Correlations coefficients between variables $x1$ .. $x5$

	$x1$	$x2$	$x3$	$x4$	$x5$
$x1$	1.00				
$x2$	1.00	1.00			
$x3$	0.53	0.53	1.00		
$x4$	0.47	0.47	0.47	1.00	
$x5$	-0.91	-0.91	-0.76	-0.73	1.00

Correlations coefficients between variables C($x1$ .. $x5$)

	$x1$	$x2$	$x3$	$x4$
$x1$	1.00			
$x2$	1.00	1.00		
$x3$	-0.61	-0.61	1.00	
$x4$	-0.67	-0.67	-0.18	1.00

Correlations coefficients between variables C($x1$ .. $x4$)

Critical R values for n =100:

$R_{0.05}$=0.20

$R_{0.01}$=0.26

$R_{0.001}$=0.32

Correlated variables $x1$ and $x2$, E(ρ)=-1

	$x1$	$x2$	$x3$	$x4$	$x5$
$x1$	1.00				
$x2$	-1.00	1.00			
$x3$	0.08	-0.08	1.00		
$x4$	-0.04	0.04	0.08	1.00	
$x5$	0.13	-0.13	0.11	0.14	1.00

Correlations coefficients between variables $x1$ .. $x5$

	$x1$	$x2$	$x3$	$x4$	$x5$
$x1$	1.00				
$x2$	-0.45	1.00			
$x3$	0.49	0.22	1.00		
$x4$	0.39	0.29	0.50	1.00	
$x5$	-0.59	-0.35	-0.85	-0.83	1.00

Correlations coefficients between variables C($x1$ .. $x5$)

	$x1$	$x2$	$x3$	$x4$
$x1$	1.00			
$x2$	-0.86	1.00		
$x3$	-0.06	-0.13	1.00	
$x4$	-0.23	0.00	-0.69	1.00

Correlations coefficients between variables C($x1$ .. $x4$)

Critical R values for n =100:

$R_{0.05}$=0.20

$R_{0.01}$=0.26

$R_{0.001}$=0.32

2nd Conclusion

The result of R calculations depend on selection of the variables. This phenomenon is called the **subcompositional incoherence.**

Compositional data transformation

Basic transformation - centered logratio clr

$$\mathrm{clr}(\mathbf{x}) = \left[\frac{x_1}{\mathrm{g}(\mathbf{x})}, \ldots, \frac{x_D}{\mathrm{g}(\mathbf{x})} \right] \qquad \mathrm{g}(\mathbf{x}) = \sqrt[D]{\prod_{j=1}^{D} x_j}$$

Advantages and disadvantages:
+ apropriate projection of distances between compositional points,
+ simple interpretation of computation results (one-to-one correspondence with original variables,
- singular covariance matrix (disables utilization of some methods),
- subcompositional incoherent.

Problem of false correlation retains

Solution … ?

Variation coefficient h_{jl}:

$$h_{jl} = \mathrm{var}\left(\ln \frac{x_j}{x_l} \right), \quad j,l = 1,\ldots,D$$

For w_l increasing linearly with increase in w_j (positive correlation), the h_{jl} value is small (0).

Disadvantages of h_{jl}:

- its statistical properties are not universal, the h_{jl} values can be compared only within a given data set
- h_{jl} is not appropriate for co-variability estimation of negatively correlated $\mathbf{w}$ components, its big values do not indicate uniquely negative correlation

Final conclusions

- Correlation between raw concentrations does not likely reflect correlation between abundances.
- The standard correlation test is useless in E(ρ=0) verification.
- Though the functional substitute of correlation coefficient remains unknown (for me), the h_{jl} parameter provides appropriate information about positive co-variability (parallel increase in values of both variables).
- Application of statistical methods, which demand correlation matrix in data analysis, will provide erroneous and/or delusive conclusions.
- Results of concentration determination in samples can not be mapped one-to-one on composition of population.