Multivariate coefficient of variation for functional data

Mirosław Krzyśko[1] and Łukasz Smaga[2]

[1]Interfaculty Institute of Mathematics and Statistics
The President Stanisław Wojciechowski State University of Applied Sciences in Kalisz
[2]Faculty of Mathematics and Computer Science
Adam Mickiewicz University

XLVIII International Biometrical Colloquium
September 9-13, 2018, Szamotuły

Coefficient of variation

- The coefficient of variation (CV) being the ratio of the standard deviation to the population mean is widely used relative variation measure.
- The CV is a dimensionless quantity, which can be expressed in percent.
- This quantity is usually used to compare the variability of several populations, even when they are characterized by variables expressed in different units as well as have really different means. In particular, the CV is often used to assess the performance or reproducibility of measurement techniques or equipments.

Multivariate coefficient of variation

- For multivariate data, computing the CV for each variable is a common practice, although this ignores the correlation between them, and this does not summarize the variability of the multivariate data into a single index.
- The known multivariate extensions of the CV are less considered in the literature. Perhaps it is due to the fact that generalizing the univariate CV to the multivariate setting is not straightforward, and the multivariate CV's do not generally measure the same quantity, when the number of variables is greater than one. Nevertheless, they all reduce to the CV in the univariate case.

- Let $\mathbf{X} = (X_1, \ldots, X_p)^\top$ be p-dimensional random vector with mean vector $\mathbf{u} \neq \mathbf{0}_p$ and covariance matrix $\mathbf{\Sigma}$.
- The multivariate coefficients of variation (MCV) by Reyment (1960), Van Valen (1974), Voinov and Nikulin (1996, p. 68) and Albert and Zhang (2010) are:

$$\mathrm{MCV}_{\mathrm{R}} = \sqrt{\frac{(\det \mathbf{\Sigma})^{1/p}}{\mathbf{u}^\top \mathbf{u}}}, \ \mathrm{MCV}_{\mathrm{VV}} = \sqrt{\frac{\mathrm{tr}\mathbf{\Sigma}}{\mathbf{u}^\top \mathbf{u}}}, \ \mathrm{MCV}_{\mathrm{VN}} = \sqrt{\frac{1}{\mathbf{u}^\top \mathbf{\Sigma}^{-1} \mathbf{u}}}, \ \mathrm{MCV}_{\mathrm{AZ}} = \sqrt{\frac{\mathbf{u}^\top \mathbf{\Sigma} \mathbf{u}}{(\mathbf{u}^\top \mathbf{u})^2}},$$

  respectively.
- The $\mathrm{MCV}_{\mathrm{R}}$ and $\mathrm{MCV}_{\mathrm{VV}}$ are based on the generalized variance $\det \mathbf{\Sigma}$ and the total variance $\mathrm{tr}\mathbf{\Sigma}$, respectively. In the $\mathrm{MCV}_{\mathrm{VN}}$, the Mahalanobis distance $\mathbf{u}^\top \mathbf{\Sigma}^{-1} \mathbf{u}$ appears to be a natural extension of the CV. Finally, the $\mathrm{MCV}_{\mathrm{AZ}}$ is derived based on a matrix generalizing the square of the CV.

- We have:

$$\mathrm{MCV}_{\mathrm{AZ}} = \frac{\sqrt{\mathrm{Var}(\mathbf{u}_*^\top \mathbf{X})}}{\|\mathbf{u}\|}, \tag{1}$$

  where $\mathbf{u}_* = \mathbf{u}/\|\mathbf{u}\|$, i.e., $\mathrm{MCV}_{\mathrm{AZ}}$ is the univariate coefficient of variation for $\mathbf{u}_*^\top \mathbf{X}$.
- Let $\mathbf{X}(t) = (X_1(t), \ldots, X_p(t))^\top$, $t \in [a, b]$, $a, b \in \mathbb{R}$ be p-dimensional random process with mean function $\boldsymbol{\mu}(t) = (\mu_1(t), \ldots, \mu_p(t))^\top \neq \mathbf{0}_p$. We also assume that $\mathbf{X}(t)$, $t \in [a, b]$ belongs to the Hilbert space $L_2^p[a, b]$ of p-dimensional vectors of square integrable functions on $[a, b]$. Let $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ denote the inner product and the norm in $L_2^p[a, b]$.

Definition 1

The functional multivariate coefficient of variation for $\mathbf{X}(t)$, $t \in [a, b]$ is defined as follows ($\boldsymbol{\mu}_*(t) = \boldsymbol{\mu}(t)/\|\boldsymbol{\mu}\|$, $t \in [a, b]$):

$$\mathrm{FMCV} = \frac{\sqrt{\mathrm{Var}(\langle \boldsymbol{\mu}_*, \mathbf{X} \rangle)}}{\|\boldsymbol{\mu}\|}. \tag{2}$$

Theorem 1

If $X_i(t)$, $t \in [a, b]$, $i = 1, \ldots, p$, are square integrable, i.e.,

$$\mathrm{E}\|X_i\|^2 = \mathrm{E}\int_a^b X_i^2(t)\, dt < \infty,$$

then

$$\mathrm{Var}(\langle \boldsymbol{\mu}_*, \mathbf{X} \rangle)$$

exists. Furthermore, the FMCV defined in (2) *is the CV of the random variable*

$$\langle \boldsymbol{\mu}_*, \mathbf{X} \rangle.$$

- Let $\mathbf{X}(t)$ belong to a finite dimensional subspace $\mathcal{L}_2^p[a,b]$ of $L_2^p[a,b]$, where the components of $\mathbf{X}(t)$ can be represented by a finite number of basis functions, i.e.,

$$X_k(t) = \sum_{l=1}^{B_k} \alpha_{kl}\varphi_{kl}(t), \tag{3}$$

where $k = 1,\ldots,p$, $t \in [a,b]$, $B_k \in \mathbb{N}$, α_{kl} are random variables with finite variance and $\{\varphi_{kl}\}_{l=1}^{\infty}$, $k = 1,\ldots,p$ are bases in the space $L_2^1[a,b]$. The equations (3) can be expressed in the following matrix notation:

$$\mathbf{X}(t) = \mathbf{\Phi}(t)\boldsymbol{\alpha}, \tag{4}$$

where

$$\mathbf{\Phi}(t) = \operatorname{diag}\left(\boldsymbol{\varphi}_1^\top(t),\ldots,\boldsymbol{\varphi}_p^\top(t)\right)$$

is the block diagonal matrix of $\boldsymbol{\varphi}_k^\top(t) = (\varphi_{k1}(t),\ldots,\varphi_{kB_k}(t))$, $k = 1,\ldots,p$ and $\boldsymbol{\alpha} = (\alpha_{11},\ldots,\alpha_{1B_1},\ldots,\alpha_{p1},\ldots,\alpha_{pB_p})^\top$.

- $$\mathrm{FMCV} = \frac{\sqrt{\mathrm{Var}\left(\frac{\mathbf{a}^\top \mathbf{J}_{\boldsymbol{\Phi}} \boldsymbol{\alpha}}{\|\mathbf{J}_{\boldsymbol{\Phi}}^{1/2}\mathbf{a}\|}\right)}}{\|\mathbf{J}_{\boldsymbol{\Phi}}^{1/2}\mathbf{a}\|} = \sqrt{\frac{\mathbf{a}^\top \mathbf{J}_{\boldsymbol{\Phi}} \boldsymbol{\Sigma}_{\boldsymbol{\alpha}} \mathbf{J}_{\boldsymbol{\Phi}} \mathbf{a}}{(\mathbf{a}^\top \mathbf{J}_{\boldsymbol{\Phi}} \mathbf{a})^2}}, \tag{5}$$

  where $\mathbf{a} = \mathrm{E}(\boldsymbol{\alpha})$, $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}} = \mathrm{Cov}(\boldsymbol{\alpha})$ and $\mathbf{J}_{\boldsymbol{\Phi}} = \mathrm{diag}(\mathbf{J}_{\boldsymbol{\varphi}_1}, \ldots, \mathbf{J}_{\boldsymbol{\varphi}_p})$ and $\mathbf{J}_{\boldsymbol{\varphi}_k} = \int_a^b \boldsymbol{\varphi}_k(t)\boldsymbol{\varphi}_k^\top(t)\, dt$ is the $B_k \times B_k$ cross product matrix corresponding to the basis $\{\varphi_{kl}\}_{l=1}^{\infty}$, $k = 1, \ldots, p$.

Theorem 2

Under the above assumptions and notation, the functional multivariate coefficient of variation for the random process $\mathbf{X}(t)$, $t \in [a, b]$, is the multivariate coefficient of variation of Albert-Zhang type for random vector $\mathbf{J}_{\boldsymbol{\Phi}}^{1/2}\boldsymbol{\alpha}$, if the matrix $\mathbf{J}_{\boldsymbol{\Phi}}^{1/2}$ exists.

- Although the FMCV is defined for univariate and multivariate functional data, we note that even when $p = 1$, the FMCV reduces to the $\mathrm{MCV}_{\mathrm{AZ}}$ (no to the CV), since $B_1 > 1$ usually.

- In practice, we have to estimate the unknown vector $\boldsymbol{\alpha}$ in (4) as well as its parameters $\mathbf{a}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}$ appearing in the FMCV given in (5).
- Let $\mathbf{x}_1(t), \ldots, \mathbf{x}_n(t)$, $t \in [a, b]$ be a random sample containing realizations of the process $\mathbf{X}(t)$. These observations are represented similarly as in (4), i.e.,

$$\mathbf{x}_i(t) = \boldsymbol{\Phi}(t)\boldsymbol{\alpha}_i,$$

  where $t \in [a, b]$ and $i = 1, \ldots, n$.
- Then, the vectors $\boldsymbol{\alpha}_i$, $i = 1, \ldots, n$ can be estimated by the least squares method or the roughness penalty approach.
- The expansion lengths B_k in (3) can be selected deterministically or by using information criteria as the Akaike and Bayesian information criteria.

- Using the estimators of $\boldsymbol{\alpha}_i$, say $\hat{\boldsymbol{\alpha}}_i$, $i = 1, \ldots, n$, we can estimate the mean vector $\mathbf{a}$ and the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}$.
- The classical estimators are the sample mean and the sample covariance matrix, i.e.,

$$\hat{\mathbf{a}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{\alpha}}_i, \ \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\alpha}} = \frac{1}{n}\sum_{i=1}^{n}(\hat{\boldsymbol{\alpha}}_i - \hat{\mathbf{a}})(\hat{\boldsymbol{\alpha}}_i - \hat{\mathbf{a}})^{\top}. \tag{6}$$

- However, these estimators may break down, when the data contain outliers. Thus, many authors recommend the use of robust estimators of location and scatter in the presence of outlying observations.
- Similarly to Aerts et al. (2015), we will mainly use the two commonly used ones, i.e., the minimum covariance determinant (MCD) estimator and the S-estimator.

- For a given breakdown point α, the MCD estimator is based on a subset of $\{\hat{\boldsymbol{\alpha}}_1, \ldots, \hat{\boldsymbol{\alpha}}_n\}$ of size $h = \lfloor n(1-\alpha) \rfloor$ minimizing the generalized variance (i.e., the determinant of covariance matrix) among all possible subsets of size h. Then, the MCD estimators of $\mathbf{a}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}$ are the sample mean and the sample covariance matrix (multiplied by a consistency factor) computed from this subset.
- The location and scatter S-estimators are the vector $\mathbf{a}_n$ and the positive definite symmetric matrix $\boldsymbol{\Sigma}_n$ which minimizes $\det(\boldsymbol{\Sigma}_n)$ subject to

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\sqrt{(\hat{\boldsymbol{\alpha}}_i - \mathbf{a}_n)^\top \boldsymbol{\Sigma}_n^{-1} (\hat{\boldsymbol{\alpha}}_i - \mathbf{a}_n)}\right) = b_0,$$

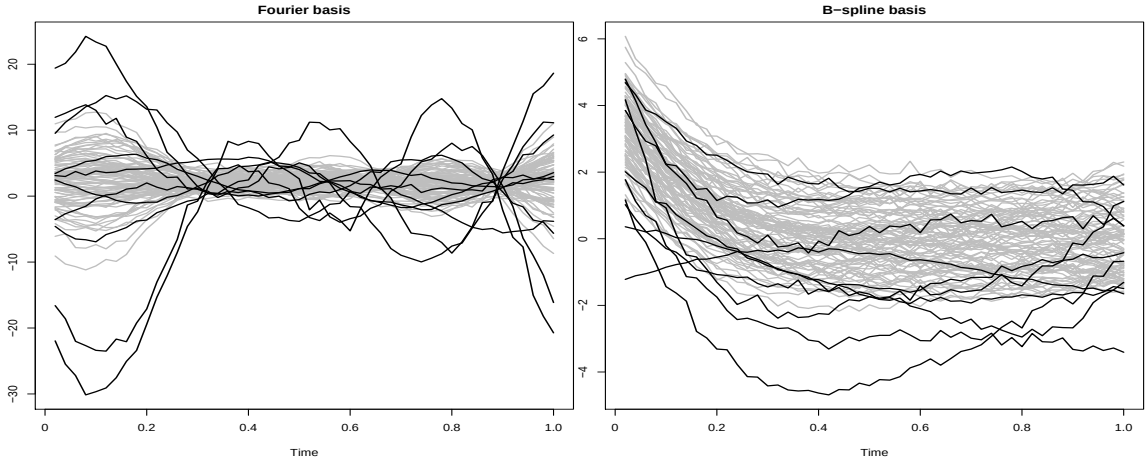
  where $\rho : \mathbb{R} \to [0, \infty)$ is a given non-decreasing and symmetric function (e.g., Tukey's biweight) and b_0 a constant needed to ensure consistency of the estimator.
- The (classical and robust) estimators of the FMCV are obtained by substituting the parameters $\mathbf{a}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}$ in (5) by their estimators $\hat{\mathbf{a}}$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\alpha}}$.

- The functional sample $\mathbf{x}_1(t), \ldots, \mathbf{x}_n(t)$ of size $n = 100$ contains realizations of $\mathbf{X}(t)$, $t \in [0, 1]$ with $p = 5$. These observations are generated as follows:

$$\mathbf{x}_i(t_j) = \mathbf{\Phi}(t_j)\boldsymbol{\alpha}_i + \boldsymbol{\epsilon}_{ij},$$

where $i = 1, \ldots, n$, t_j, $j = 1, \ldots, 50$ are equally spaced design time points in $[0, 1]$, the matrix $\mathbf{\Phi}(t)$ contains basis functions with $B_k = 5$, $k = 1, \ldots, p$, $\boldsymbol{\alpha}_i$ are $5p$-dimensional random vectors, and $\boldsymbol{\epsilon}_{ij} = (\epsilon_{ij1}, \ldots, \epsilon_{ijp})^\top$ are the measurement errors such that $\epsilon_{ijk} \sim N(0, 0.025 r_{ik})$ and r_{ik} is the range of the k-th row of $(\mathbf{\Phi}(t_1)\boldsymbol{\alpha}_i \ldots \mathbf{\Phi}(t_{50})\boldsymbol{\alpha}_i)$, $k = 1, \ldots, p$. We use the Fourier and B-spline bases.

- The vectors $\boldsymbol{\alpha}_i$, $i = 1, \ldots, n$ were generated from multivariate normal or t_5- distributions with mean $\mathbf{a}$ and covariance matrix $\mathbf{\Sigma}_{\boldsymbol{\alpha}}$. Similarly to Aerts et al. (2015), we set $\mathbf{a} = \mathbf{a}_1 := a\mathbf{e}_1$ or $\mathbf{a} = \mathbf{a}_2 := (a/(5p)^{1/2})\mathbf{1}_{5p}$ and $\mathbf{\Sigma}_{\boldsymbol{\alpha}} = (1 - \rho)\mathbf{I}_{5p} + \rho\mathbf{1}_{5p}\mathbf{1}_{5p}^\top$, where a is chosen to get a given value of the FMCV, $\mathbf{e}_1 = (1, 0, \ldots, 0)^\top$ and $\rho = 0, 0.5, 0.8$. We set $\mathrm{FMCV} = 0.1, 0.5, 0.9$. Moreover, to obtain uncontaminated and contaminated functional data, $\varepsilon\%$ of the observations are generated with $10\mathbf{\Sigma}_{\boldsymbol{\alpha}}$, where $\varepsilon = 0, 10, 20, 30, 40, 50$.



Exemplary realizations of the first functional variables of simulated data with 10% of outlying observations. The uncontaminated (resp. contaminated) data are depicted in gray (resp. black).

Simulation studies - mean squared error

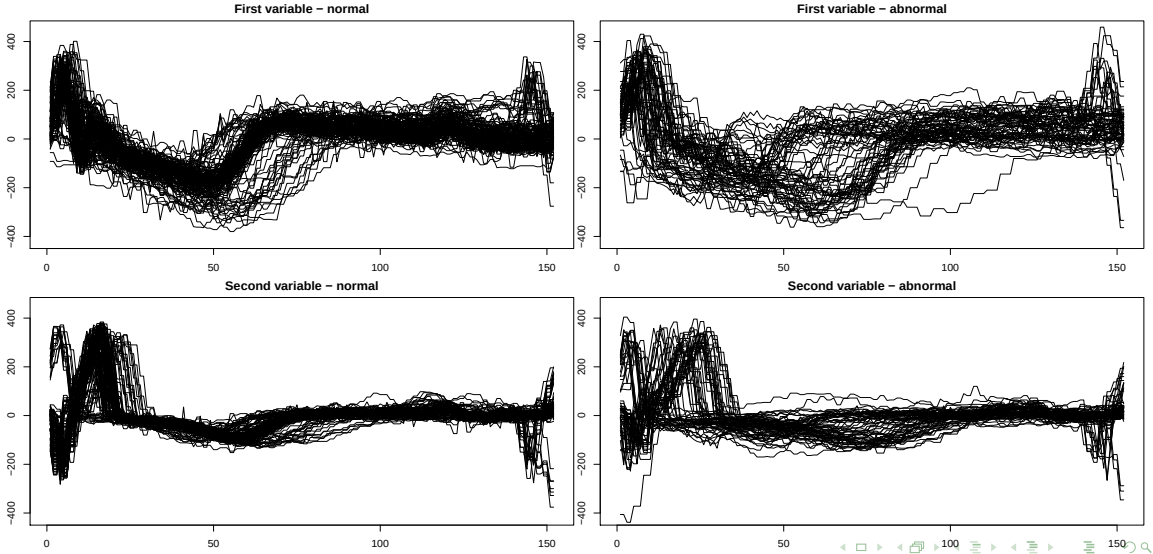
		class.	MCD	S	class.	MCD	S	class.	MCD	S
	ε	FMCV = 0.1			FMCV = 0.5			FMCV = 0.9		
F	0	0.0001	0.0002	0.0001	0.0209	0.0314	0.0212	0.1197	0.1654	0.1216
	10	0.0019	0.0002	0.0002	0.1278	0.0347	0.0304	0.6283	0.1763	0.1662
	20	0.0053	0.0002	0.0006	0.2919	0.0368	0.0519	1.1743	0.1909	0.2820
	30	0.0095	0.0003	0.0019	0.4902	0.0412	0.1177	1.7589	0.2138	0.5742
	40	0.0147	0.0011	0.0069	0.7307	0.0936	0.3261	2.3889	0.4412	1.4175
	50	0.0204	0.0065	0.0140	0.9976	0.3683	0.6583	2.9514	1.3361	2.5026
B	0	0.0002	0.0004	0.0002	0.0433	0.0616	0.0441	0.2508	0.3492	0.2533
	10	0.0023	0.0004	0.0003	0.2143	0.0644	0.0582	1.0780	0.3563	0.3268
	20	0.0060	0.0004	0.0007	0.4704	0.0687	0.0882	2.0930	0.3842	0.4986
	30	0.0109	0.0005	0.0022	0.7978	0.0717	0.1672	3.2243	0.3956	0.9228
	40	0.0166	0.0013	0.0074	1.1598	0.1435	0.4648	4.3543	0.7577	2.2992
	50	0.0227	0.0078	0.0153	1.5075	0.6046	0.9543	5.3330	2.3050	4.1944

- MSE's are usually similar for different bases, but greater differences can appear for greater FMCV. Moreover, MSE for the B-spline basis is often greater than for the Fourier basis.

Application to ECG data

- We consider the ECG data set originated from Olszewski (2001).
- During each of 200 heartbeats, two electrodes were used to measure the ECG. For one heartbeat and one electrode, the ECG was measured in 152 design time points, and the resulting values of ECG form a curve, which can be treated as discrete functional observation.
- We have 200 two-dimensional discrete functional data observed in 152 design time points ($n = 200$, $p = 2$, $m_i = 152$, $i = 1, \ldots, n$).
- The heartbeats were assigned to normal or abnormal group.
- Abnormal heartbeats are representative of a cardiac pathology known as supraventricular premature beat.
- The normal and abnormal groups consist of 133 and 67 functional observations, respectively.

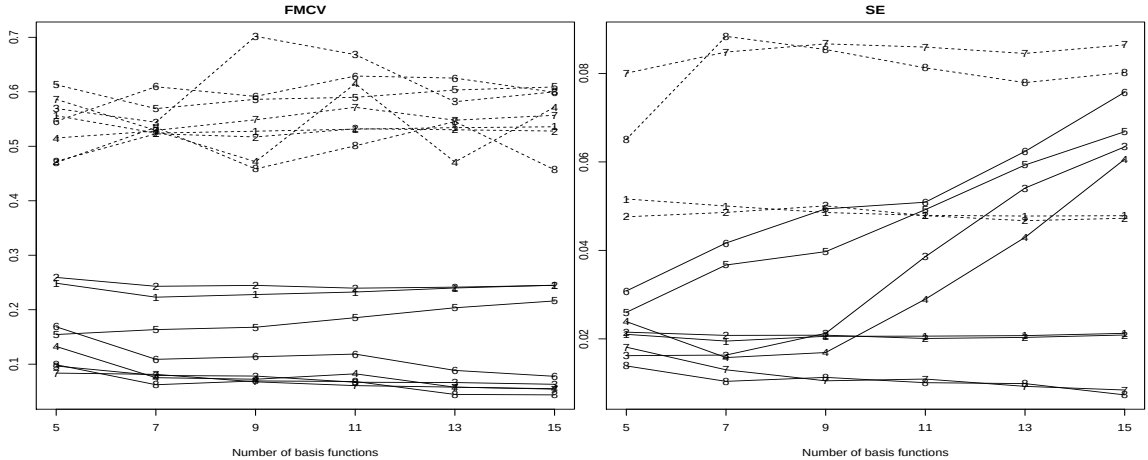
Application to ECG data



Application to ECG data

- The ECG database was used to discriminate between normal and abnormal heartbeats (Olszewski, 2001).
- For illustrative purposes, we show that this has also sense from variability point of view. To do this, we compute the FMCV for both normal and abnormal heartbeats separately.
- The basis functions representation of the data was obtained by using the Fourier and B-spline bases and $B_1 = B_2 = 5, 7, 9, 11, 13, 15$, if it was possible.
- To estimate the FMCV, we used the same estimators as in simulation experiments, i.e., the classical, MCD and S estimators, as well as the the pairwise estimator (OGK).
- The standard errors (SE) were obtained by the bootstrap method, based on 1000 bootstrap samples.

Application to ECG data



normal - solid line, abnormal - dashed line, 1 - classical, Fourier, 2 - classical, B-spline, 3 - MCD, Fourier, 4 - MCD, B-spline, 5 - S, Fourier, 6 - S, B-spline, 7 - OGK, Fourier, 8 - OGK, B-spline

References

1. Aerts, S., Haesbroeck, G., Ruwet, C. (2015). Multivariate coefficients of variation: comparison and influence functions. Journal of Multivariate Analysis 142, 183–198.
2. Albert, A., Zhang, L. (2010). A novel definition of the multivariate coefficient of variation. Biometrical Journal 52, 667–675.
3. Olszewski, R.T. (2001). Generalized Feature Extraction for Structural Pattern Recognition in Time-Series Data. Ph.D. Thesis, Carnegie Mellon University, Pittsburgh, PA.
4. Reyment, R.A. (1960). Studies on Nigerian Upper Cretaceous and Lower Tertiary Ostracoda: part 1. Senonian and Maastrichtian Ostracoda. Stockholm Contributions in Geology 7, 1–238.
5. Van Valen, L. (1974). Multivariate structural statistics in natural history. Journal of Theoretical Biology 45, 235–247.
6. Voinov, V.G., Nikulin, M.S. (1996). Unbiased Estimators and Their Applications, Vol. 2, Multivariate Case. Kluwer, Dordrecht.

Thank you for your attention