Distribution and robustness of a distance-based multivariate coefficient of variation

AERTS Stéphanie HEC-ULg, University of Liège stephanie.aerts@ulg.ac.be
Joint work with HAESBROECK Gentiane and RUWET Christel

Université de Liège

P7/06 StUDyS

Measuring relative variability

Univariate setting:
Univariate coefficient of variation : ratio of the standard deviation to the mean

$$\mathrm{CV} = \frac{\sigma}{\mu}$$

This relative dispersion measure is advocated when comparing variability of populations with variables expressed in different units or having really different means.

Multivariate setting:
When the data are intrinsically multivariate, comparing relative variability marginally may lead to controversial results.

Goal: Summarize multivariate relative variability in **one single index**.

Applications in

- External Quality Assesment programs (to assess the reproducibility of measurement methods)
- Biostatistics (comparison of different species on the basis of several traits)
- Finance (comparison of the performance of several portfolios)
- ...

Multivariate coefficients of variation

Let $X \in \mathbb{R}^p \sim F_p(\mu, \Sigma)$ with mean vector $\mu \neq 0$ and covariance matrix $\Sigma \in \mathcal{S}_p^+$.
Several propositions of **multivariate coefficients of variation** exist in the literature (see Albert and Zhang, 2010 for a review):

Reyment (1960): $\gamma_R = \sqrt{\frac{(\det \Sigma)^{1/p}}{\mu^t \mu}}$

Van Valen (1974): $\gamma_{VV} = \sqrt{\frac{\mathrm{tr}\,\Sigma}{\mu^t \mu}}$

Voinov & Nikulin (1996): $\gamma_{VN} = \sqrt{\frac{1}{\mu^t \Sigma^{-1} \mu}}$

Albert & Zhang (2010): $\gamma_{AZ} = \sqrt{\frac{\mu^t \Sigma \mu}{(\mu^t \mu)^2}}$

In practice, these coefficients can be estimated by plugging any location and covariance estimators, T_n and C_n, in expressions above.

Focus on Voinov and Nikulin's CV

Voinov and Nikulin's CV

- makes use of the whole correlation structure
- has an intuitive definition (Mahalanobis distance between the origin of the design space and the mean vector)
- is scale invariant

Sample distribution under elliptical symmetry
Under **elliptical distributions** and if V_n is an estimator of γ_{VN} computed with **equivariant** estimators of location and covariance, the distribution of V_n depends on the parameters (μ, Σ) only through γ_{VN}.

Sample distribution under normality:
Under **normality** and if V_n^{cl} is the **sample estimator** of γ_{VN}, then

$$\frac{n-p}{p}\frac{1}{(V_n^{cl})^2} \sim F_{p;n-p}\left(\frac{n}{\gamma_{VN}^2}\right).$$

This allows to

- construct **exact confidence intervals** for the parameter γ_{VN} (inversion method)
- study the **bias** of the sample estimator

Bias of the sample estimator

In finite samples, V_n^{cl} underestimates the relative dispersion.

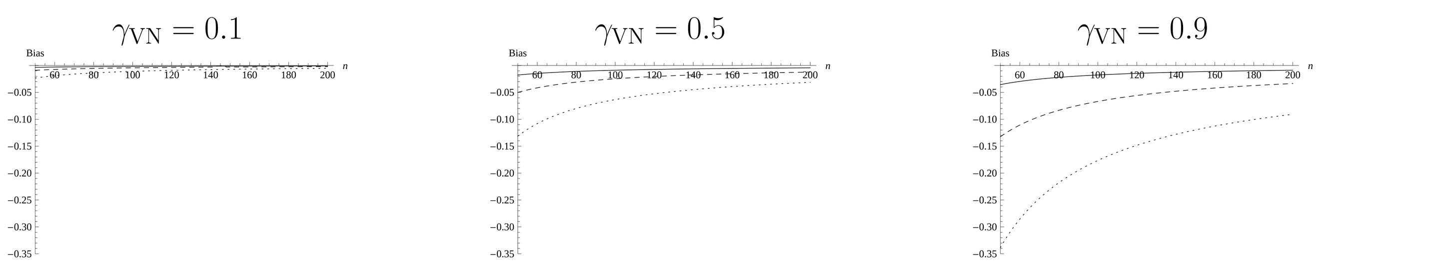

Fig. 1: Bias of the estimator V_n^{cl} w.r.t. the sample size n (solid line: $p = 3$, dashed line: $p = 7$ and dotted line: $p = 20$)

Bias correction

Bias-correction 1: Plugging unbiased estimators
The first advocated bias correction consists in taking, when it is possible, the square root of the inverse of an unbiased estimator for $1/\gamma_{VN}^2$, i.e.

$$V_n' = \sqrt{\frac{1}{\frac{p}{n}\left(\frac{n-p-2}{p}\frac{1}{(V_n^{cl})^2} - 1\right)}}$$

Bias correction 2: Inversion
$g : \gamma \longmapsto \mathrm{E}_\gamma[V_n^{cl}]$ and $b(\gamma) = g(\gamma) - \gamma$

For an observed value v_n, let $\gamma_1 = g^{-1}(v_n)$
The bias is $b(\gamma_1) = g(\gamma_1) - \gamma_1 = v_n - \gamma_1$
Thus, the estimator corrected for bias is given by:

$$V_n'' = V_n^{cl} - b\left(g^{-1}(V_n^{cl})\right) = g^{-1}(V_n^{cl})$$

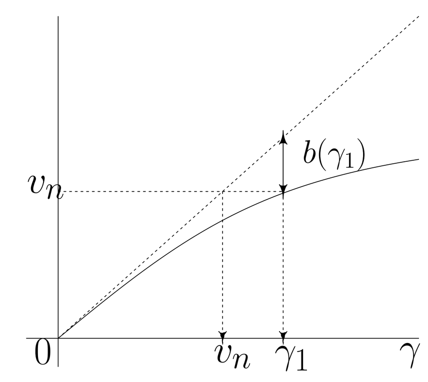

Fig. 2: Expectation (solid-line curve) of V_n^{cl} w.r.t γ, for $n = 50$ and $p = 7$

Simulations suggest that both corrections tend to reduce the bias. The first one tends to overestimate the parameter γ_{VN} but the second one allows a considerable improvement.

Robustness - Influence function

The statistical functional related to γ_{VN} is given by $V(F,T,C) = (T(F)^t C(F)^{-1} T(F))^{-1/2}$ where T and C are any statistical functionals of multivariate location and covariance.

The **influence function** of the statistical functional V at the model F is defined by

$$\mathrm{IF}(x; V, F) = \frac{\partial}{\partial \varepsilon} V((1-\varepsilon)F + \varepsilon \Delta_x)\Big|_{\varepsilon=0}$$

where Δ_x is the Dirac distribution having all its mass at $x \in \mathbb{R}^p$.

Provided that $T(F) = \mu$ and $C(F) = \Sigma$, we have

$$\mathrm{IF}(x; V, F) = \frac{\gamma_{VN}^3}{2}\left(\mu^t \Sigma^{-1} \mathrm{IF}(x; C, F) \Sigma^{-1} \mu - 2\, \mu^t \Sigma^{-1} \mathrm{IF}(x; T, F)\right)$$

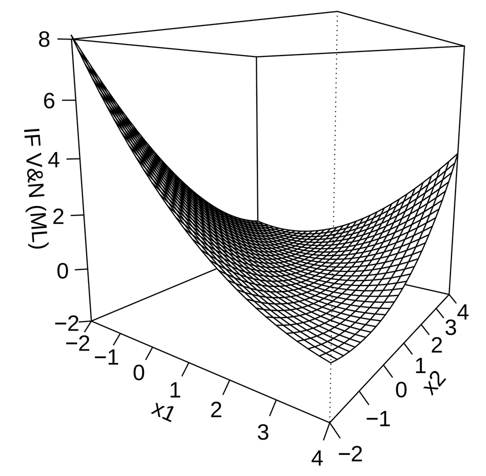

Fig. 3: IF with classical estimators

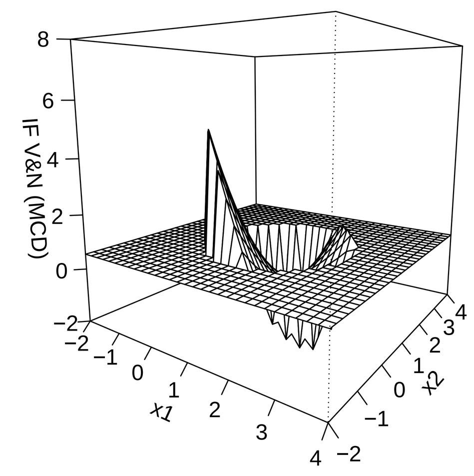

Fig. 4: IF with MCD estimators

The sample estimator is extremely sensitive to local contamination (unbounded IF).
One should use **robust estimators** of location and covariance (MCD, M, S,...) to obtain a robust CV estimator.

Computation of IF allows to:

- study local robustness
- construct a **diagnostic tool** to detect influential observations (Pison & Van Aelst, 2004)
- derive a general expression for the **asymptotic variance** of several estimators for γ_{VN}.

Ongoing research

Testing procedures for the equality of multivariate coefficients of variation

- Using asymptotic properties (**Wald-type test**)
- Study of the stability of level and power of these tests under contamination thanks to the IF's

References

[1] Aerts, S., Haesbroeck, G. and Ruwet, C. (2014a). Multivariate coefficients of variation: comparison and influence functions. *Submitted for publication.*
[2] Aerts, S., Haesbroeck, G. and Ruwet, C. (2014b). Distribution under elliptical symmetry of a distance-based multivariate coefficient of variation. *Submitted for publication.*
[3] Albert, A. and Zhang, L. (2010). A novel definition of the multivariate coefficient of variation. *Biometrical Journal*, 52, 667–675.
[4] Pison, G. and Van Aelst, S. (2004). Diagnostic plots for robust multivariate methods. *J. Comput. Graph. Statist.*, 13, 310–329.
[5] 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.
[6] Van Valen, L. (1974). Multivariate structural statistics in natural history. *Journal of Theoretical Biology*, 45, 235–247.
[7] Voinov, V.G. and Nikulin, M.S. (1996). *Unbiased estimators and their applications.* Vol. 2, multivariate case. Dordrecht: Kluwer.