Journal of Multivariate Analysis 142 (2015) 183–198

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

# Multivariate coefficients of variation: Comparison and influence functions

S. Aerts [a,*], G. Haesbroeck [b], C. Ruwet [c]

[a] HEC-ULg, University of Liege (ULg, N1), Rue Louvrex 14, 4000 Liège, Belgium
[b] University of Liege (ULg, Polytech 1), Allée de la Découverte 12, 4000 Liège, Belgium
[c] Haute Ecole Prov. de Liège (Service de Math), Quai Gloesener 6, 4200 Liège, Belgium

ARTICLE INFO

*Article history:*
Received 8 April 2014
Available online 28 August 2015

*AMS subject classifications:*
62H12
62F35

*Keywords:*
Multivariate coefficient of variation
Influence functions
Minimum Covariance Determinant estimator
S estimator

ABSTRACT

In the univariate setting, coefficients of variation are well-known and used to compare the variability of populations characterized by variables expressed in different units or having really different means. When dealing with more than one variable, the use of such a relative dispersion measure is much less common even though several generalizations of the coefficient of variation to the multivariate setting have been introduced in the literature. In this paper, the lack of robustness of the sample versions of the multivariate coefficients of variation (MCV) is illustrated by means of influence functions and some robust counterparts based either on the Minimum Covariance Determinant (MCD) estimator or on the S estimator are advocated. Then, focusing on two of the considered MCV's, a diagnostic tool is derived and its efficiency in detecting observations having an unduly large effect on variability is illustrated on a real-life data set. The influence functions are also used to compute asymptotic variances under elliptical distributions, yielding approximate confidence intervals. Finally, simulations are conducted in order to compare, in a finite sample setting, the performance of the classical and robust MCV's in terms of variability and in terms of coverage probability of the corresponding asymptotic confidence intervals.


## 1. Introduction

The ratio of the standard deviation to the population mean is known as the coefficient of variation (CV). It is a pure number free from any unit of measurement and as such, it allows the comparison of the variability of populations characterized by variables expressed in different units. As a relative dispersion measure with respect to the mean, it may also be useful when comparing the variability of populations with really different means.

In some applications, the CV is considered as a more informative quantity than the standard deviation. For example, it is often used to assess the reproducibility of measurement methods or equipments. The coefficient of variation is recommended in External Quality Assessment (EQA) programs (see [14]), which often deal with data measured by means of different equipments. EQA organizers are interested in getting statistical evidence about the best performing techniques as far as reproducibility is concerned. The lower the CV, the better the precision of the method is considered to be.

In practice, coefficients of variation are often estimated by using the sample standard deviation and the sample mean. As an illustrative example, we will focus on the following real EQA data. In 1996, $n = 371$ Belgian laboratories took part to such an EQA scheme and received two samples of two distinct sera from which the concentration of glucose (in mmol/L) had to

* Corresponding author.
*E-mail addresses:* stephanie.aerts@ulg.ac.be (S. Aerts), g.haesbroeck@ulg.ac.be (G. Haesbroeck), christel.ruwet@hepl.be (C. Ruwet).

http://dx.doi.org/10.1016/j.jmva.2015.08.006
0047-259X/

**Table 1**
Summary statistics for the concentration of glucose measured in two sera by 371 laboratories in Belgium in 1996.

| | $n$ | Serum 1 | | | Serum 2 | | |
|---|---|---|---|---|---|---|---|
| | | Mean | SD | CV | Mean | SD | CV |
| All methods | 371 | 15.942 | 1.350 | 0.085 | 6.658 | 0.528 | 0.079 |
| Method 1 | 163 | 16.220 | 0.571 | 0.035 | 6.697 | 0.227 | 0.034 |
| Method 2 | 47 | 16.060 | 1.883 | 0.117 | 6.616 | 0.426 | 0.064 |
| Method 3 | 79 | 15.655 | 2.231 | 0.143 | 6.813 | 0.973 | 0.143 |
| Method 5 | 32 | 15.690 | 0.566 | 0.036 | 6.637 | 0.228 | 0.034 |
| Method 7 | 50 | 15.541 | 0.782 | 0.050 | 6.340 | 0.276 | 0.044 |

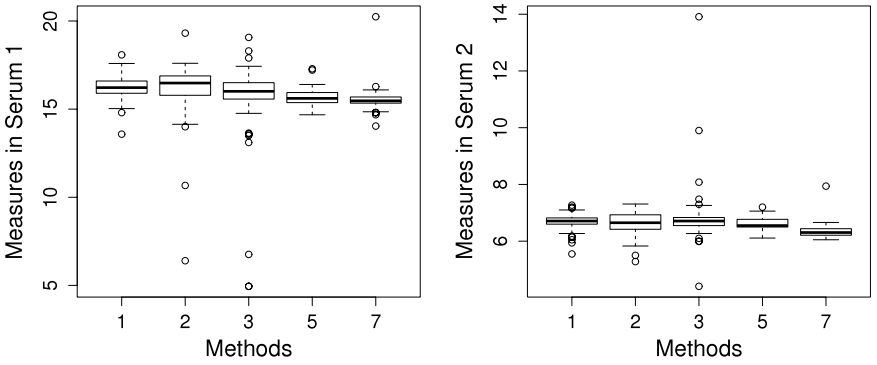


**Fig. 1.** Boxplots of the measures of glucose concentration in serum 1 and in serum 2, the measurements being split with respect to the measuring device.

be measured. Table 1 summarizes the main characteristics of the measurements, taking into account the fact that five out of seven official techniques to measure the concentration of such an analyte were used by the laboratories.

Looking at the results reported in Table 1, several remarks come to mind. First of all, these data are intrinsically bivariate, since for each individual (laboratory) two measurements (one for each serum) were performed by the same measuring device, under the same conditions, and sent to the EQA center, while the coefficient of variation is univariate. In this example, one could argue that comparing the coefficients of variation computed on the two sera separately seems to lead to the same overall conclusion: methods 1 and 5 are quite similar, closely followed by method 7. Methods 2 and 3 behave poorly. Nevertheless, computing only marginal coefficients of variation is usually not appropriate as the results may be controversial. The second remark is related to the lack of robustness of the sample coefficient of variation. Indeed, the poor performance of methods 2 and 3, may be explained, as illustrated in Fig. 1, by the presence of atypical measurements, as is often the case for EQA data (see e.g. [13,28]). If the observations lying beyond three median absolute deviations from the median are removed, then the coefficients of variation of the concentrations in serum 1 become 0.038 for both methods, yielding a value much more comparable with the other reported values.

The aim of this paper is to study the robustness of multivariate extensions of the CV. More specifically, Section 2 summarizes the main multivariate extensions reviewed by Albert and Zhang [3]. To measure the robustness of their sample versions, influence functions are computed and illustrated graphically in Section 3. The influence functions are then used as diagnostic tools for influential observations in Section 4 and as a means to compute asymptotic variances in Section 5. A simulation study focusing both on robustness and variability is reported in Section 6. Some conclusions follow.

## 2. Multivariate coefficients of variation

Let $X = (X_1, \ldots, X_p)^t$ be a $p$-variate random vector distributed according to a given distribution $F$ with mean vector $\mu \neq 0$ and covariance matrix $\Sigma$ (assumed to be symmetric and positive definite, i.e. $\Sigma \in \mathcal{S}_p^+$). Extending the univariate definition of the coefficient of variation to the multivariate setting is not as straightforward as one could imagine. Some authors [6,22,7] suggest to work with a $p \times p$ matrix called the *coefficient of variation matrix*, with element $(i, j)$ given by $\Sigma_{ij}/(\mu_i\mu_j)$, $i, j = 1, \ldots, p$, assuming $\mu_i \neq 0$ for all $i$. However, it is not easy to compare $p \times p$ matrices and controversial results may appear.

In this paper, we will focus on multivariate extensions of the coefficient of variation that summarize multivariate relative variation into a single index $\gamma$. Albert and Zhang [3] reviewed the existing definitions and added a novel one. Using their notations, the multivariate coefficients of variation (MCV's) that are considered throughout the paper are listed here:

$$\text{Reyment's CV [18]}: \ \gamma_{\text{R}} = \sqrt{\frac{(\det \Sigma)^{1/p}}{\mu^t \mu}}$$