Multivariate coefficients of variation: comparison and influence functions

S. Aerts*, G. Haesbroeck† and C. Ruwet‡

Abstract

In the univariate setting, coefficients of variation are well-known and required when one wants 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 widespread 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 their sample versions is illustrated by means of influence functions and a robust counterpart based on the Minimum Covariance Determinant estimator is advocated. Simulations compare the performance of the two proposals. Then, we focus on two particular extensions and use influence functions to derive the asymptotic variances of their classical estimators under the hypothesis of normality. Finally, a diagnostic tool for detecting observations having an unduly large effect on these classical estimators is proposed.

Keywords: Coefficient of Variation, Influence Functions, Robust Estimation

*HEC-ULg, University of Liege (ULg, N1), Rue Louvrex 14, 4000 Liege, Belgium; stephanie.aerts@ulg.ac.be

†Department of Mathematics, University of Liege (ULg, B37), Grande Traverse 12, 4000 Liege, Belgium; g.haesbroeck@ulg.ac.be

‡Department of Mathematics, University of Liege (ULg, B37), Grande Traverse 12, 4000 Liege, Belgium; cruwet@ulg.ac.be, and Haute Ecole de la Province de Liege, Service de mathematiques, 6, quai Gloesner, 4000 Lige, Belgium

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 to compare the variability of populations characterized by variables expressed in different units. Moreover, as the coefficient of variation is 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 coefficient of variation is considered as a more informative quantity than the standard deviation. For example, coefficients of variation are often used to assess the reproducibility of measurement methods or equipments. The lower the CV, the better the precision of the method is considered to be. External Quality Assessment (EQA) programs often deal with data measured by means of different equipments. EQA organisers are interested in getting statistical evidence about the best performing techniques as far as reproducibility is concerned. The coefficient of variation may be helpful in this context and is recommended nowadays in EQA analyses (Karkalousos and Evangelopoulos, 2011).

In practice, coefficients of variation are often estimated by the ratio of the sample standard deviation to the sample mean. As 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 a given serum from which the concentration of glucose (in mmol/L) had to be measured. Table 1 below summarizes the main characteristics of the measurements obtained by these laboratories on the first sample, taking into account the fact that five out of seven official techniques to measure the concentration of such an analyte were used by some laboratories.

Looking at the results reported in Table 1, several remarks come to mind. First of all, these data are intrinsically bivariate as there are two sera to measure 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 coefficients of variation for each marginal variable is usually not appropriate as the results may be controversial. The

		Serum 1			Serum 2		
	n	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

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

second remark is related to the lack of robustness of the sample coefficient of variation. Indeed, as mentioned before, methods 2 and 3 seem to perform really poorly with respect to the others. However, Figure 1 may explain the relative poor performance of these two methods as their measurements are heavily contaminated, as is often the case for EQA data (Healy 1979, Zhou et al 2006). 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 coefficient of variation. More specifically, Section 2 summarizes the main multivariate extensions of the coefficient of variation, as reviewed by Albert and Zhang (2010). To measure the robustness of their sample versions, influence functions are computed for all the proposals and illustrated graphically in Section 3. The influence functions are then used as a diagnostic tool four outlying observations in Section 4 and as a mean to compute asymptotic variances under normality in Section 5. A simulation study focusing both on robustness and variability is reported in section 6 while Section 7 outlines some conclusions.

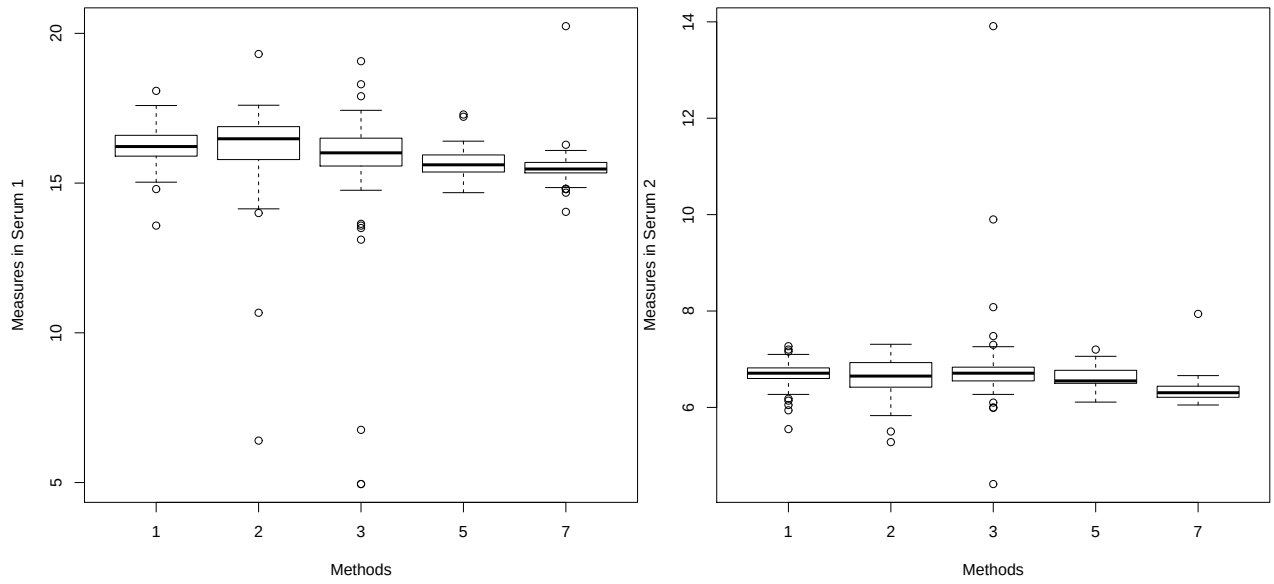


Figure 1: Boxplots of the measures of glucose concentration in serum 1 and in serum 2, the measurements being splitted with respect to the measuring device.

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 Σ (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 (Bennett 1977, Underhill 1990, Boik and Shirvani 2009) suggest to work with a $p \times p$ matrix called the *coefficient of variation matrix*, Ψ say, with element (i, j) given by $\Psi_{ij} = \Sigma_{ij}/\mu_i\mu_j, i, j = 1, \ldots, p$, 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 γ. Albert and Zhang (2010) reviewed the existing definitions and added a novel one. Using their notations, the multivariate coefficients of variation that are considered throughout the paper are listed

here:

$$\text{Reyment's CV (1960):}\quad \gamma_{\mathrm{R}} = \sqrt{\frac{(\det\Sigma)^{1/p}}{\mu^t\mu}}$$

$$\text{Van Valen's CV (1974):}\quad \gamma_{\mathrm{VV}} = \sqrt{\frac{\operatorname{tr}\Sigma}{\mu^t\mu}}$$

$$\text{Voinov \& Nikulin's CV (1996):}\quad \gamma_{\mathrm{VN}} = \sqrt{\frac{1}{\mu^t\Sigma^{-1}\mu}}$$

$$\text{Albert \& Zhang's CV (2010):}\quad \gamma_{\mathrm{AZ}} = \sqrt{\frac{\mu^t\Sigma\mu}{(\mu^t\mu)^2}}$$

It is worth noting that all the coefficients listed above reduce to the univariate coefficient of variation when $p = 1$ but, as soon as p is bigger than 1, they do not measure the same quantity anymore. Albert and Zhang (2010) proved that $\gamma_{\mathrm{R}} \leq \gamma_{\mathrm{VV}}$ for all $\mu \neq 0$ and for any symmetric and positive (semi-)definite matrix Σ. Moreover, if $\det\Sigma > 0$, $\gamma_{\mathrm{VN}} \leq \gamma_{\mathrm{AZ}} \leq \gamma_{\mathrm{VV}}$ and the equality between γ_{VN} and γ_{AZ} holds if μ is an eigenvector of Σ. As four non equivalent proposals exist, one can wonder which coefficient should be used in practice.

First, let us note that Van Valen's coefficient does not depend on the correlation structure but only takes the total variance into account. As most multivariate techniques make use of the whole correlation structure to measure variability, it seems to us that this coefficient of variation is not as adequate as the others. Then, γ_{AZ} has an advantage with respect to the two remaining coefficients. Indeed, the reason why Albert and Zhang (2010) introduced this novel coefficient and advocated its use is that they worked with compositional data for which the covariance matrix is singular. In that case, γ_{VN} and γ_{R} are either not defined or useless while γ_{AZ} is properly defined. Now, if this particular degenerate case is discarded, the invariance behaviour of the coefficients as well as their geometric interpretation might be interesting to take into account.

Concerning the invariance behaviour, recall first that in the univariate case, the coefficient of variation is dimensionless. This means that the random variables X and kX, for a positive constant k, have the same coefficient of variation. In the multivariate setting, all the coefficients of variation reviewed by Albert and Zhang (2010) remain unchanged if all variables are multiplied by the same constant. However, under a more general scale

transformation like AX, for any $p \times p$ non-singular matrix A, the invariance does not hold anymore for all coefficients (even if A is a diagonal matrix other than a multiple of the identity matrix). Proposition 1 below (the proof is outlined in the Appendix) shows that only the coefficient of variation defined by Voinov and Nikulin satisfies this more general invariance property. Nevertheless, when the matrix A is orthogonal, the property also holds for the other coefficients of variation.

Proposition 1 *Let the p-dimensional vector X be distributed according to a distribution F with mean vector μ (with $\mu \neq 0$) and covariance matrix $\Sigma \in \mathcal{S}_p^+$). The coefficient of variation of Voinov and Nikulin is scale invariant for any $p \times p$ non-singular matrix A. The other coefficients of variation satisfy the scale invariance property when A is orthogonal.*

Let us now turn to the consideration of the geometric interpretation of these coefficients. As already mentioned, they do note measure dispersion in the same way. For example, Reyment's coefficient would yield the same value for any distribution with a fixed covariance structure Σ and any mean vector lying on the sphere of radius R while γ_{VN} and γ_{AZ} would vary quite drastically for varying positions of μ on the sphere (see Albert and Zhang 2010 for further discussion). It is therefore difficult to pinpoint one of the four proposals as the best to use in applications. Nevertheless, the invariance property of γ_{VN} and the simple definition of Reyment lead us to focus mainly on these two coefficients even though most developments will be performed for all four multivariate CV's.

3 Influence functions of the Multivariate coefficients of variation

The definitions given above only rely on the mean vector μ, the covariance matrix Σ and the dimension p. Even if it is natural to estimate these unknown quantities by the sample mean and sample covariance matrix, any estimator of multivariate location and dispersion may be used.

As illustration, let us come back to the EQA data with the consideration of the two sera simultaneously. The estimations of the coefficients of variation computed with the

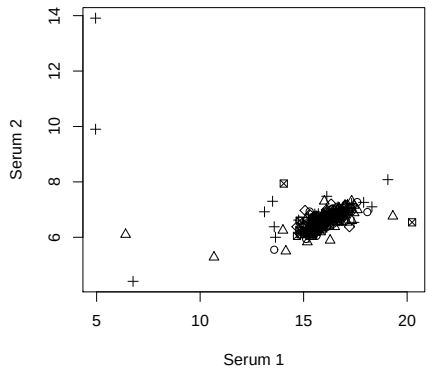


Figure 2: Scatter plot of the bivariate EQA data using different plotting symbols to characterize the different measuring devices.

sample mean and covariance matrix are reported in the left panel of Table 2. Whatever coefficient, the values computed on the measurements obtained by means of methods 2 and 3 are much bigger than those derived for the other measurements. As shown on Figure 2, it is not surprising as the observations drawn as triangles or crosses correspond to these two methods.

	Classical estimations				Robust estimations			
Method	$\hat{\gamma}_{\mathrm{R}}$	$\hat{\gamma}_{\mathrm{VV}}$	$\hat{\gamma}_{\mathrm{VN}}$	$\hat{\gamma}_{\mathrm{AZ}}$	$\hat{\gamma}_{\mathrm{R}}$	$\hat{\gamma}_{\mathrm{VV}}$	$\hat{\gamma}_{\mathrm{VN}}$	$\hat{\gamma}_{\mathrm{AZ}}$
1	0.016	0.035	0.033	0.034	0.012	0.028	0.025	0.027
2	0.046	0.111	0.064	0.106	0.021	0.040	0.038	0.038
3	0.081	0.143	0.074	0.111	0.017	0.037	0.032	0.035
5	0.020	0.036	0.030	0.033	0.015	0.027	0.025	0.025
7	0.028	0.049	0.032	0.043	0.011	0.020	0.018	0.018

Table 2: Estimated multivariate coefficients of variation on the EQA data when using the sample mean and covariance matrix or the MCD estimator.

As pointed out by Albert and Zhang (2010), resorting to a robust estimator of location and scatter, like the Minimum Covariance Determinant (MCD) estimator of Rousseeuw (1985), should be advocated when the data contain outliers. The right panel of Table 2

yields the coefficients of variation computed on the MCD estimator with 25% of breakdown point. Even though the values related to methods 2 and 3 remain bigger than the others, the difference is much less pronounced.

Influence functions can be computed to formalize the sensitivity analysis of these estimators and to compare their robustness when working with classical or with robust estimators of location and scatter.

The influence function of a statistical functional CV at the model F is defined, for those distributions for which the derivative makes sense, by

$$\mathrm{IF}(x;\mathrm{CV},F) = \lim_{\varepsilon\to 0}\frac{\mathrm{CV}(F_\varepsilon)-\mathrm{CV}(F)}{\varepsilon} = \frac{\partial}{\partial\varepsilon}\mathrm{CV}(F_\varepsilon)\bigg|_{\varepsilon=0}$$

where $F_\varepsilon = (1-\varepsilon)F + \varepsilon\Delta_x$, Δ_x being the Dirac distribution having all its mass at the point $x \in I\!R^p$ (see Hampel et al, 1986).

The functionals of interest here are those corresponding to the four coefficients of variation listed above. They are defined by

$$\begin{aligned}
\mathrm{CV}_{\mathrm{R}}(F;T,C) &= \sqrt{\frac{(\det C(F))^{1/p}}{T(F)^t T(F)}}\\
\mathrm{CV}_{\mathrm{VV}}(F;T,C) &= \sqrt{\frac{\mathrm{tr}\, C(F)}{T(F)^t T(F)}}\\
\mathrm{CV}_{\mathrm{VN}}(F;T,C) &= (T(F)^t C(F)^{-1} T(F))^{-1/2}\\
\mathrm{CV}_{\mathrm{AZ}}(F;T,C) &= \sqrt{\frac{T(F)^t C(F) T(F)}{(T(F)^t T(F))^2}}
\end{aligned}$$

where T and C are respectively location and covariance statistical functionals defined on the set of distributions in $I\!R^p$ for which the second moment is well defined.

When the location and scatter functionals are Fisher consistent, i.e. $T(F) = \mu$ and $C(F) = \Sigma$, where μ and Σ are the expectation and covariance matrix of F, Fisher consistency also holds for the statistical functionals of the coefficients of variation, i.e. $\mathrm{CV}_i(F;T,C) = \gamma_i$, where i stands for $\mathrm{R}, \mathrm{VV}, \mathrm{VN}$ or AZ. When the statistical functionals T and C are simply the expectation and covariance matrix, then the statistical functional CV_i evaluated at the empirical distribution F_n yields the natural estimation of the corresponding coefficient of variation, denoted as $\overline{\mathrm{CV}}_i$ from now on.

Proposition 2 below derives the influence functions of the statistical functionals defined for the four multivariate coefficients of variation. The proof is sketched in the Appendix and only requires straightforward application of classical differentiation rules.

Proposition 2 *For any distribution F with mean vector $\mu \neq 0$ and covariance matrix $\Sigma \in \mathcal{S}_p^+$, the influence functions of Reyment, Van Valen, Voinov & Nikulin and Albert & Zhang multivariate coefficients of variation are given by*

$$IF(x; CV_{\mathrm{R}}, F) = \frac{\gamma_{\mathrm{R}}}{2}\left(\frac{\mathrm{tr}(\Sigma^{-1}IF(x;C,F))}{p} - 2\,\frac{\mu^t IF(x;T,F)}{\mu^t\mu}\right)$$

$$IF(x; CV_{\mathrm{VV}}, F) = \frac{\gamma_{\mathrm{VV}}}{2}\left(\frac{\mathrm{tr}(IF(x;C,F))}{\mathrm{tr}(\Sigma)} - 2\,\frac{\mu^t IF(x;T,F)}{\mu^t\mu}\right)$$

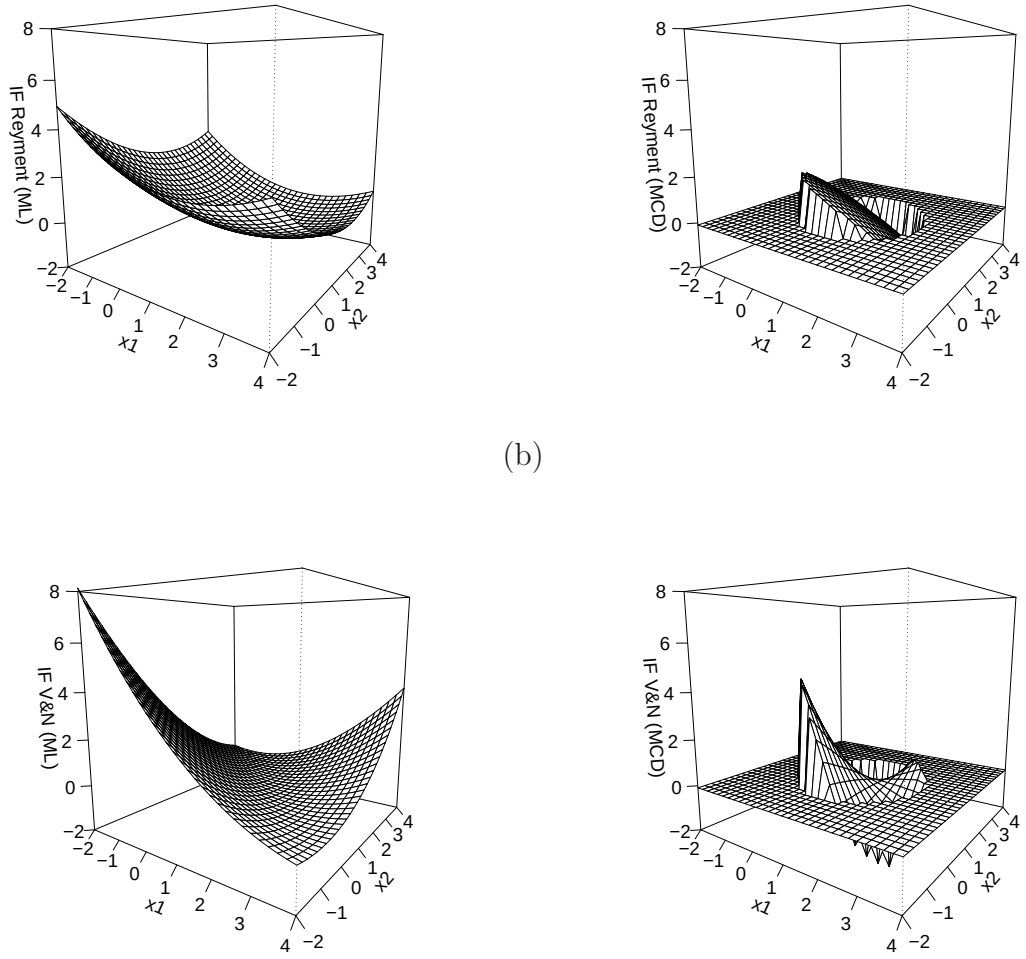
$$IF(x; CV_{\mathrm{VN}}, F) = \frac{\gamma_{\mathrm{VN}}^3}{2}\left(\mu^t\Sigma^{-1}IF(x;C,F)\Sigma^{-1}\mu - 2\,\mu^t\Sigma^{-1}IF(x;T,F)\right)$$

$$IF(x; CV_{\mathrm{AZ}}, F) = \frac{1}{2\gamma_{\mathrm{AZ}}\,(\mu^t\mu)^2}\left(\mu^t IF(x;C,F)\mu + 2\,\mu^t\left(\Sigma - 2\,\gamma_{\mathrm{AZ}}^2(\mu^t\mu)I_{p\times p}\right)IF(x;T,F)\right)$$

where $IF(x;T,F)$ and $IF(x;C,F)$ are the influence functions of the location and scatter functionals used in the definition of the multivariate coefficients of variation, assuming that $T(F) = \mu$ and $C(F) = \Sigma$, and $I_{p\times p}$ is the p-dimensional identity matrix.

If the influence functions of the location and dispersion estimators are bounded, the influence functions of these multivariate coefficients of variation will be bounded also. Figure 3 shows the influence functions of Reyment's and Voinov & Nikulin's coefficient of variation when using the classical estimators (left panel) and the MCD estimators (right panel) under a bivariate normal distribution with mean $\mu = (1,1)^t$ and identity covariance matrix. One can see that both coefficients based on the MCD estimator will be influenced by an infinitesimal contamination located in the neighborhood of the mean but the influence is close to zero as the contamination gets further away. When using the classical estimator, the influence is clearly unbounded.

As already pointed out, all multivariate coefficients of variation reviewed by Albert and Zhang (2010) simplify to the univariate CV characterized by the statistical functional $CV(F) = S(F)/T(F)$ where $S(F)$ is a scale functional and $T(F)$ a location one. The influence function of this univariate functional has been derived by Groeneveld (2011) but can also be obtained by setting $p = 1$ in the influence functions given in proposition



(a) (b)

(c) (d)

Figure 3: Influence functions of Reyment's (in plots (a) and (b)) and Voinov & Nikulin's (in plots (c) and (d)) coefficient of variation when using the classical estimators (left column) and the MCD estimators (right column).

2 noting that $2S(F)IF(x;S,F) = IF(x;C,F)$ if $C(F) = S(F)^2$. It takes the nice and closed following form:

$$IF(x;CV,F) = \frac{IF(x;S,F)\mu - \sigma IF(x;T,F)}{\mu^2} \tag{1}$$

where $IF(x;T,F)$ and $IF(x;S,F)$ are the influence functions of T and S.

4 Detection of influential observations

One potential application of influence functions is through the construction of a diagnostic plot in which influential observations may be spotted easily. The aim is to detect those observations having an unduly large effect on relative multivariate dispersion when the classical estimators are used. In Section 3, the influence functions of the coefficients of variation have been computed for any estimator of multivariate location and scatter. When using the empirical mean and covariance matrix, the influence functions of Reyment and Voinov and Nikulin's coefficients simplify as follows:

$$IF(x; \overline{\mathrm{CV}}_{\mathrm{R}}, F) = \frac{\gamma_{\mathrm{R}}}{2}\left(\frac{1}{p}(x-\mu)^t\Sigma^{-1}(x-\mu) - 2\frac{x^t\mu}{\mu^t\mu} + 1\right) \tag{2}$$

$$IF(x; \overline{\mathrm{CV}}_{\mathrm{VN}}, F) = \frac{\gamma_{\mathrm{VN}}^3}{2}\left(\left(\mu^t\Sigma^{-1}(x-\mu) - 1\right)^2 - \mu^t\Sigma^{-1}\mu - 1\right) \tag{3}$$

Computing these influence functions on each observation x_i yields so-called empirical influence measures which may then be plotted with respect to the indexes of the observations to visually determine those yielding (too) small or (too) big values, as suggested by Pison and Van Aelst (2004). To decide how big an influence measure must be to consider that it is too big (or too small), a cutoff value is needed. Usually, such a critical value is given by an extreme quantile of the distribution of the influence function under some underlying distribution. Proposition 3 derives the distribution of the IF of the two considered coefficients of variation under normality when using the empirical mean and covariance matrix.

Proposition 3 *Let $\chi^2(p,\delta)$ denote the non-central χ^2 distribution with p degrees of freedom and non-centrality parameter $\delta > 0$. When $X \sim \Phi_{\mu,\Sigma} = N_p(\mu,\Sigma)$,*

$$IF(X; \overline{CV}_{\mathrm{R}}, \Phi_{\mu,\Sigma}) = \frac{\gamma_{\mathrm{R}}}{2}\left(\frac{Y}{p} - \frac{p}{(\mu^t\mu)^2}\mu^t\Sigma\mu - 1\right) \tag{4}$$

where $Y \sim \chi^2(p,\delta_R)$ with $\delta_R = \frac{p^2}{(\mu^t\mu)^2}\mu^t\Sigma\mu$ and

$$IF(X; \overline{CV}_{\mathrm{VN}}, \Phi_{\mu,\Sigma}) = \frac{\gamma_{\mathrm{VN}}}{2}\left(W - (1+\gamma_{\mathrm{VN}}^2)\right) \tag{5}$$

where $W \sim \chi^2(1,\delta_{VN})$ with $\delta_{VN} = \gamma_{\mathrm{VN}}^2$.

The distributions derived in Proposition 3 are only valid when parameters μ and Σ are known. In practice, one needs to estimate them. Approximate cutoff values are then derived by plugging the estimations in the expressions 4 and 5. As the classical estimators may suffer from the masking effect, Pison and Van Aelst (2004) advocate using robust estimations for μ and Σ. In Figure 4, the empirical influence measures for the EQA data based on $\overline{\mathrm{CV}}_{\mathrm{VN}}$ are plotted with respect to those based on $\overline{\mathrm{CV}}_{\mathrm{R}}$. Here the reweighted MCD estimator with 25% breakdown point has been used for computing the estimated values of μ and Σ (computed with `covMcd` in `R` software).

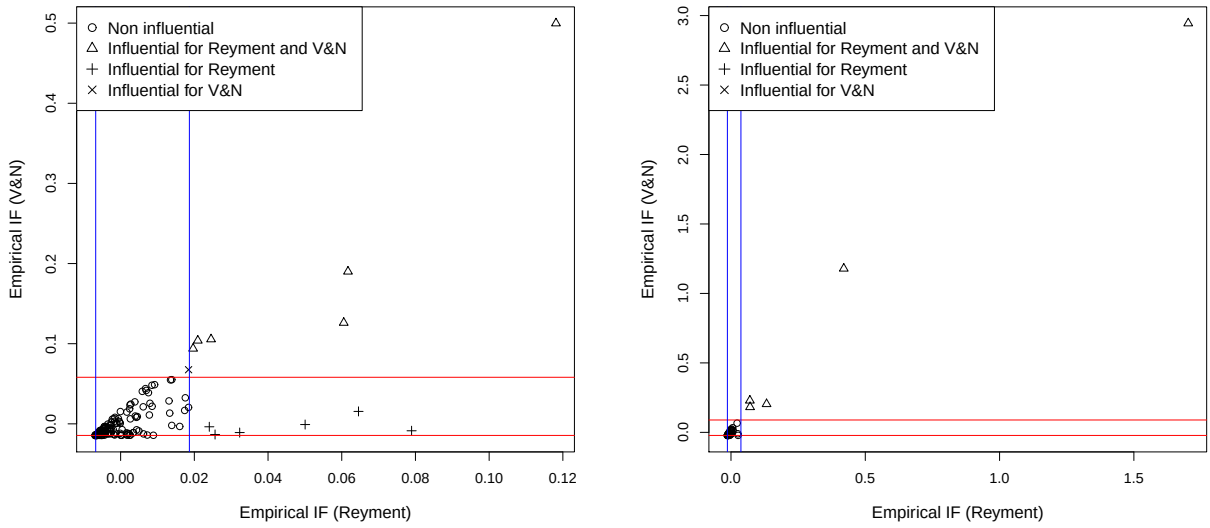


Figure 4: Empirical influence measures based on $\overline{\mathrm{CV}}_{\mathrm{R}}$ (horizontal axis) and $\overline{\mathrm{CV}}_{\mathrm{VN}}$ (vertical axis) when estimating the parameters μ and Σ in a robust way (for the EQA data measured by method 1 on the left panel and by method 2 on the right panel).

As $\overline{\mathrm{CV}}_{\mathrm{R}}$ and $\overline{\mathrm{CV}}_{\mathrm{VN}}$ differ in their geometrical interpretation, their do not necessarily detect the same observations as influential. The plane may then be divided into different parts as illustrated on figure 4: depending on its position with respect to the cutoffs represented by the horizontal and vertical lines, an observation may be a regular observation (symbol $\circ$), an influential observation for both measures of relative dispersion (symbol $\triangle$) or influential for one of them and not for the other (symbols $\times$ and $+$).

It is also worth noting that observations with large robust distances do not automat-

ically have a large influence on the relative dispersion. This can easily be understood from expression (2) where the projection on μ is also taken into account. To emphasize the difference between robust distances and influence measures, figure 5 plot the original data together with the robust $97,5\%$ tolerance ellipse. Different plotting symbols allow to distinguish between outliers (observations lying beyond the tolerance ellipse) that are influential or not as well as regular observations that might be influential.

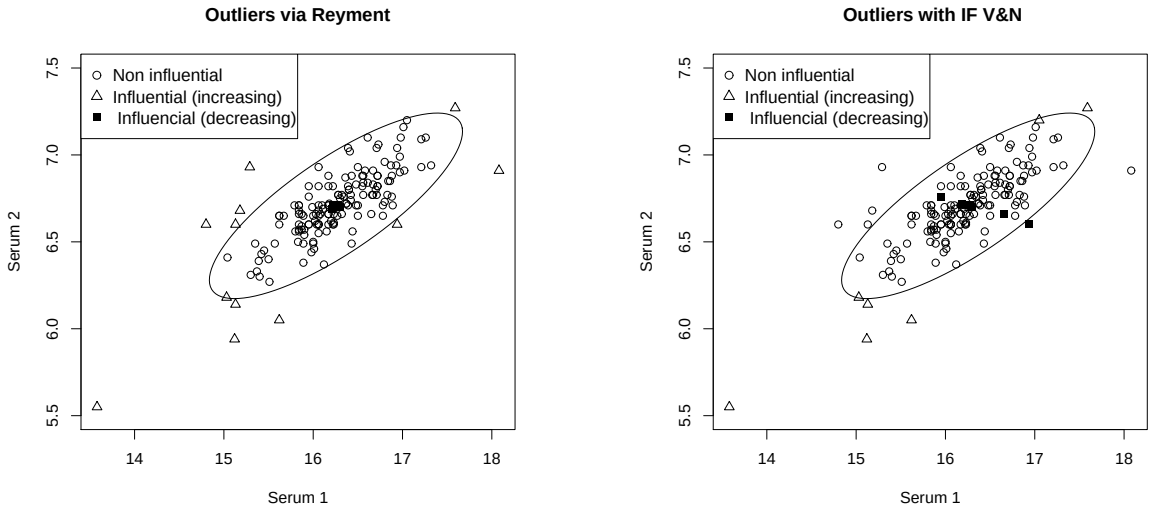


Figure 5: Location of outlying or influential observations based on $\overline{\mathrm{CV}}_{\mathrm{R}}$ (left panel) or $\overline{\mathrm{CV}}_{\mathrm{VN}}$ (right panel) (for the EQA data measured by method 1.

As it can be seen in figure 5, this detection method allows to pinpoint not only the observations that tend to increase relative dispersion as measured by γ_{VN} or γ_{R} (represented by the symbol $\triangle$) but also those that tend to decrease it (represented by the symbol ■).

5 Asymptotic variances

Influence functions are useful in their own right, e.g. for detecting influential observations as illustrated in Section 4. Moreover, when the estimator of interest is asymptotically normal, the influence function may also be used to derive the asymptotic variance of the

estimator, by means of (see Hampel et al, 1986)

$$\mathrm{ASV}(CV, F) = \int_{\mathbb{R}^p} IF^2(x; CV, F) dF(x). \tag{6}$$

In the univariate case and under normality, Ahmed (1995) proved the asymptotic normality of the sample coefficient of variation with asymptotic variance given by $\frac{\sigma^2}{2\mu^2} + \frac{\sigma^4}{\mu^2}$ where μ and σ^2 are respectively the mean and variance of the underlying distribution. Under the same assumptions, computing (6) with $IF(x; \bar{X}, F) = x - \mu$ and $IF(x; S; F) = \frac{1}{2}\left((x-\mu)^2 - \sigma^2\right)$ yields the same asymptotic variance.

Under a Pareto-type distribution, Albrecher et al (2010) proved the asymptotic normality of the sample coefficient of variation with an asymptotic variance given by $\frac{\sigma_*^2 \mu^2}{4\sigma^2}$ where

$$\sigma_*^2 = \frac{\mu_4}{\mu^4} - \left(\frac{\mu_2}{\mu^2}\right)^2 + 4\left(\frac{\mu_2}{\mu^2}\right)^3 - \frac{4\mu_2\mu_3}{\mu^5}$$

μ_i being the non-centered moment of order i and σ^2 the variance of the underlying distribution...

In the multivariate setting, if joint asymptotic normality holds for the vector[1] of estimators $V_n = (T_n, \mathrm{vec}(C_n))^T$ under the underlying distribution F, where T_n and C_n are some estimators of location and covariance-matrix respectively, the Delta method (see Van der Vaart, 1998) ensures the asymptotic normality of all four sample multivariate coefficients, as they can be expressed by differentiable functions of V_n.

The asymptotic variance of an asymptotically normal estimator of γ_{R} and γ_{VN} obviously depends on the estimators of location and covariance used through their influence function. However, under normality, provided that the estimators of location and scatter are affine equivariant, the following lemma characterizes the general form of their influence functions, which makes it possible to derive asymptotic variances of CV's in full generality.

Lemma 1 *For any affine equivariant location and scatter matrix functionals T and C possessing an influence function, there exist three functions ξ_T, α_C, β_C : $[0, +\infty[\to \mathbb{R}$*

[1] *vec(.)* is the operator which stacks the columns of a matrix on top of each other.

such that

$$IF(x;T,F) = \xi_T[d(x)](x-\mu) \tag{7}$$

and

$$IF(x;C,F) = \alpha_C[d(x)](x-\mu)(x-\mu)^t - \beta_C[d(x)]\Sigma \tag{8}$$

with $d^2(x) = (x-\mu)^t\Sigma^{-1}(x-\mu)$ *and* $F = N(\mu,\Sigma)$.

From now on, only location and covariance estimators that fulfill these conditions will be considered. This includes classical estimators and robust estimators S, M, MCD and reweighted MCD. Proposition 4 characterize the general form of the asymptotic variances of estimators of γ_{R} and γ_{VN} in this case.

Proposition 4 *Let* CV_R *and* CV_{VN} *denote estimators of* γ_{R} *and* γ_{VN} *computed with affine equivariant estimators of location* C *and covariance* T. *Under normality, their asymptotic variances are given by*

$$\begin{aligned} ASV(CV_{\mathrm{R}},\Phi_{\mu,\Sigma}) = \frac{\gamma_{\mathrm{R}}^2}{4}\Bigg[& \frac{4}{(\mu^t\mu)^2}\ \mu^t E\left[\xi_T^2[d(X)]\ (X-\mu)^2\right]\mu \\ & + \frac{4}{p(\mu^t\mu)}E\left[(\alpha_C[d(X)]d^2(X) - p\beta_C[d(X)])\ \xi_T[d(X)]\ (X-\mu)^t\right]\mu \\ & + \frac{1}{p^2}E\left[\left(\alpha_C[d(X)]d^2(X) - p\beta_C[d(X)]\right)^2\right]\Bigg] \end{aligned}$$

and

$$\begin{aligned} ASV(CV_{\mathrm{VN}},\Phi_{\mu,\Sigma}) = \frac{\gamma_{\mathrm{VN}}^6}{4}\Bigg[& E\left[\alpha_C^2[d(X)]\ \left(\mu^t\Sigma^{-1}(X-\mu)\right)^4\right] \\ & - 4E\left[\alpha_C[d(X)]\gamma_T[d(X)]\ \left(\mu^t\Sigma^{-1}(X-\mu)\right)^3\right] \\ & - \frac{2}{\gamma_{\mathrm{VN}}^2}E\left[\left(\alpha_C[d(X)]\beta_C[d(X)] - 2\gamma_{\mathrm{VN}}^2\ \xi_T^2[d(X)]\right)\ \left(\mu^t\Sigma^{-1}(X-\mu)\right)^2\right] \\ & + \frac{4}{\gamma_{\mathrm{VN}}^2}E\left[\beta_C[d(X)]\xi_T[d(X)]\ \left(\mu^t\Sigma^{-1}(X-\mu)\right)\right] \\ & + \frac{1}{\gamma_{\mathrm{VN}}^2}E\left[\beta_C^2[d(X)]\right]\Bigg] \end{aligned}$$

Plugging the corresponding functions α_C, β_C and ξ_T in expressions above allow to derive the desired asymptotic variances. For example, asymptotic variances of the classical estimators $\overline{\mathrm{CV}}_{\mathrm{R}}$ and $\overline{\mathrm{CV}}_{\mathrm{VN}}$, for which $\xi_T(.) = \alpha_C(.) = \beta_C(.) = 1$, simplify as follow :

$$\mathrm{ASV}(\overline{\mathrm{CV}}_{\mathrm{R}}, \Phi_{\mu,\Sigma}) = \frac{\gamma_{\mathrm{R}}^2}{2p}\left(1 + 2\frac{p}{(\mu^t\mu)^2}\mu^t\Sigma\mu\right). \quad (9)$$

$$\mathrm{ASV}(\overline{\mathrm{CV}}_{\mathrm{VN}}, \Phi\mu, \Sigma) = \frac{\gamma_{\mathrm{VN}}^2}{2} + \gamma_{\mathrm{VN}}^4. \quad (10)$$

Let us first notice that these ASV increase monotonically with γ_i for both coefficients. For fixed γ_i, the asymptotic variance of $\overline{\mathrm{CV}}_{\mathrm{VN}}$ does not depend on the dimension while the ASV of $\overline{\mathrm{CV}}_{\mathrm{R}}$ decreases as p increases. The asymptotic variances (9) and (10) are often used to construct large sample confidence intervals.

6 Simulations

In this section, we first show by means of simulations that the derived ASV of Section 5 for $\overline{\mathrm{CV}}_{\mathrm{R}}$ and $\overline{\mathrm{CV}}_{\mathrm{VN}}$ are confirmed with finite sample results. In a second time, non robustness of classical estimates of Reyment's coefficient and Voinov and Nikulin's coefficient is illustrated by means of simulations.

In order to confirm the results on ASV, data were generated from a p-variate normal for which the mean vector μ and the covariance matrix Σ were designed in order to limit the parameters to one, this parameter being the theoretical coefficient of variation γ_{R}, i.e. $\mu = \frac{1}{\gamma_{\mathrm{R}}}e_1$ and $\Sigma = I_{p\times p}$. Several sample sizes and dimensions were investigated: $n = 50, 100, 500$ or 1000 and $p = 3, 7$ or 20. For each sample size, dimension and theoretical value of γ_{R}, $m = 1000$ samples were generated according to this configuration. Then, the classical estimations $\overline{\mathrm{CV}}_{\mathrm{R}}$ were computed on these samples. By definition of the asymptotic variance, one should observe a convergence of the values $n\mathrm{Var}(\overline{\mathrm{CV}}_{\mathrm{R}})$ to $\mathrm{ASV}(\overline{\mathrm{CV}}_{\mathrm{R}}, \Phi)$. The same simulations were performed for Voinov and Nikulin's coefficient.

The left columns of this table contain the finite samples variances multiplied by the sample size n, which turn out to be very close to the corresponding asymptotic variances. However, the convergence seems to be slower as γ_i increases. The same pattern arises for other values of p.

		$n\text{Var}\left[\overline{\text{CV}}_\text{i}\right]$		$\text{ASV}(\overline{\text{CV}}_\text{i})$	
	n	R	VN	R	VN
$\gamma_i = 0.1$	50	0.00086	0.00505	0.00081	0.00510
	100	0.00077	0.00466		
	500	0.00078	0.00484		
	1000	0.00084	0.00515		
$\gamma_i = 0.5$	50	0.07672	0.18225	0.08036	0.18750
	100	0.07111	0.19998		
	500	0.07671	0.19015		
	1000	0.07882	0.17958		
$\gamma_i = 0.9$	50	0.55814	0.08679	0.71396	1.06110
	100	0.68084	1.00132		
	500	0.68637	1.05734		
	1000	0.71217	1.04107		

Table 3: Finite sample variances for $\overline{\text{CV}}_\text{R}$ and $\overline{\text{CV}}_\text{VN}$ under normality assumption ($p = 7$).

Influence functions derived in Section 3 have already shown the non-robustness of multivariate coefficients of variation if location and covariance estimates used to compute them are not adequate. To illustrate this on finite samples, we compare the classical estimates and MCD estimates for γ_R and γ_VN under contamination of simulated data. In fact, three estimates are under consideration: the classic estimates $\overline{\text{CV}}_\text{i}$ for i= R or VN and their 50% and 25% breakdown reweighted MCD counterparts (computed with `covMcd` in `R`), denoted here as follow: $\text{CV}_{i;MCD50}$ and $\text{CV}_{i;MCD75}$. The contaminated model $F_\varepsilon = (1-\varepsilon)F + \varepsilon G$, was used to generate, for several contamination percentages ε, $m = 1000$ samples of size n, with F being the p-variate normal distribution with parameters given by configuration above and G being a p-variate normal distribution with $\mu_G = \mu_F$ and $\Sigma_G = 100 * \Sigma_F$ (scale contamination). To asses the performance of these estimators, the mean squared error was computed. In figure 6 the MSE for $n = 100, p = 7$ and $\gamma_i = 0.1$ are plotted versus the percentage of contamination.

At the uncontaminated normal model, the classical estimates have the lowest MSE but as soon as contamination is introduced, these estimators break down, contrary to MCD estimates which resist better under every non null contamination percentages. For small contamination percentages, $\mathrm{CV}_{i;MCD75}$ perform better than $\mathrm{CV}_{i;MCD50}$ since the choice of a higher breakdown point results in a loss of efficiency for location and covariance MCD estimates used to compute them. However, as soon as the contamination percentage is higher than 25%, high bias of $\mathrm{CV}_{i;MCD75}$ counterbalance this efficiency gain resulting in higher MSE than those of $\mathrm{CV}_{i;MCD50}$.

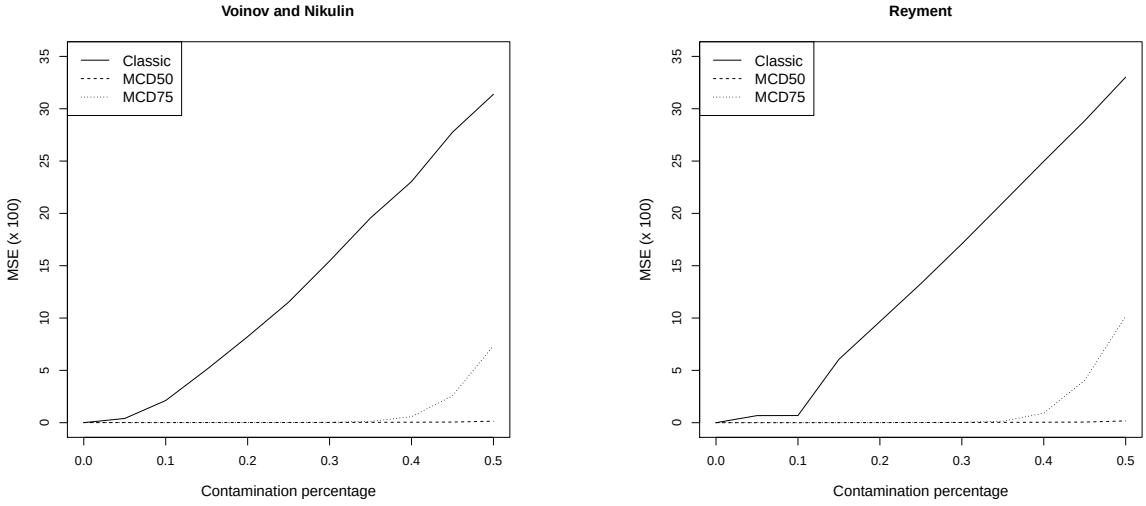


Figure 6: MSE of several estimates of γ_{R} (left panel) and γ_{VN} (right panel) for sample size $n = 100$ from a 7-variate normal distribution.

In order to take into account heavier tailed distributions, the same simulations have been computed for a multivariate Student distribution with df = 5 degrees of freedom and with mean vector and covariance matrix given by $\mu = \frac{1}{\gamma_i} e_1$ and $\Sigma = I_{p\times p}$. Figure 7 shows that conclusions remain similar for this distribution.

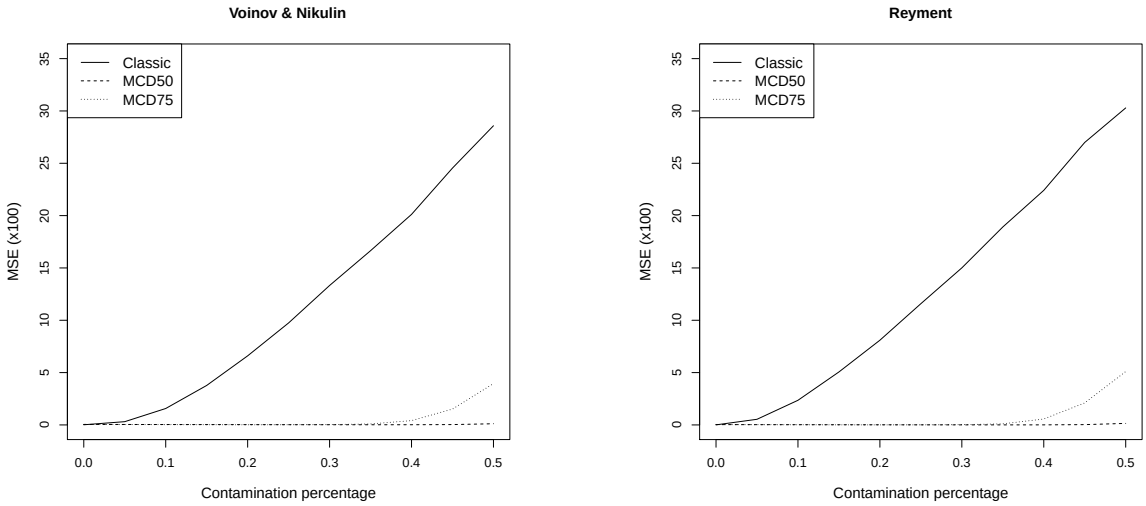


Figure 7: MSE of several estimates of γ_{R} (left panel) and γ_{VN} (right panel) for sample size $n = 100$ from a 7-variate Student distribution with $\mathrm{df} = 5$ degrees of freedom.

Conclusion

Acknowledgements:

This work was partially supported by the IAP Research Network P7/06 of the Belgian State.

Appendix

Proof of Proposition 1:

Let μ' and Σ' denote the mean vector and covariance matrix of the random vector $X' = AX$. It is well known that

$$\mu' = A\mu \text{ and } \Sigma' = A\Sigma A^t.$$

Then, the following transformations hold:

$$\begin{aligned}
&\mu'^t\mu = \mu^t A^t A\mu \\
&\operatorname{tr}\Sigma' = \operatorname{tr}(\Sigma A^t A) \\
&\det\Sigma' = \det\Sigma \times (\det A)^2 \\
&\mu'^t\Sigma'^{-1}\mu' = \mu^t\Sigma^{-1}\mu \\
&\mu'^t\Sigma'\mu' = \mu^t(A^t A)\Sigma(A^t A)\mu
\end{aligned}$$

and the proof may be completed by noting that if A is an orthogonal matrix, $A^t A = I$. □

Proof of Proposition 2: Under condition of existence, the influence function of CV_i where $i = \mathrm{R}$, VV, VN or AZ is the derivative of $\mathrm{CV}_i((1-\varepsilon)F + \varepsilon\Delta_x)$ w.r.t. ε evaluated in $\varepsilon = 0$.

For the coefficient of Reyment, one gets

$$\begin{aligned}
\mathrm{IF}(x;\mathrm{CV}_\mathrm{R},F) &= \frac{\partial}{\partial\varepsilon}\left(\frac{(\det C(F_\varepsilon))^{1/p}}{T(F_\varepsilon)^t T(F_\varepsilon)}\right)^{1/2}\Bigg|_{\varepsilon=0} \\
&= \frac{1}{2}\left(\frac{(\det C(F))^{1/p}}{T(F)^t T(F)}\right)^{-1/2}\frac{\partial}{\partial\varepsilon}\left(\frac{(\det C(F_\varepsilon))^{1/p}}{T(F_\varepsilon)^t T(F_\varepsilon)}\right)\Bigg|_{\varepsilon=0} \\
&= \frac{1}{2}\frac{1}{\gamma_\mathrm{R}}\left[\frac{\frac{\partial}{\partial\varepsilon}\left((\det C(F_\varepsilon))^{1/p}\right)\Big|_{\varepsilon=0} T(F)^t T(F) - (\det C(F))^{1/p}\,\frac{\partial}{\partial\varepsilon}\left(T(F_\varepsilon)^t T(F_\varepsilon)\right)\Big|_{\varepsilon=0}}{\left(T(F)^t\,T(F)\right)^2}\right] \\
&= \frac{1}{2}\frac{1}{\gamma_\mathrm{R}}\left[\frac{\frac{1}{p}(\det\Sigma)^{1/p}\operatorname{tr}\left(\Sigma^{-1}\,\frac{\partial}{\partial\varepsilon}C(F_\varepsilon)\big|_{\varepsilon=0}\right)}{\mu^t\mu} - \frac{2.(\det\Sigma)^{1/p}\,\mu^t\,\frac{\partial}{\partial\varepsilon}T(F_\varepsilon)\big|_{\varepsilon=0}}{(\mu^t\mu)^2}\right] \\
&= \frac{\gamma_\mathrm{R}}{2}\left[\frac{1}{p}\operatorname{tr}\left(\Sigma^{-1}\mathrm{IF}(x,C,F)\right) - 2\,\frac{\mu^t\mathrm{IF}(x,T,F)}{\mu^t\mu}\right],
\end{aligned}$$

noting that

$$\frac{\partial}{\partial\varepsilon}\det C(F_\epsilon)|_{\varepsilon=0} = (\det C(F))\ \operatorname{tr}\left(\frac{\partial}{\partial\varepsilon}\,C(F_\epsilon)|_{\varepsilon=0}\,C(F_\epsilon)^{-1}\right) = (\det\Sigma)\ \operatorname{tr}\left(\Sigma^{-1}\frac{\partial}{\partial\varepsilon}\,C(F_\epsilon)|_{\varepsilon=0}\right)$$

.

For Van Valen's coefficient of variation, one gets

$$\begin{aligned}
\mathrm{IF}(x;\mathrm{CV}_{\mathrm{VV}},F) &= \frac{\partial}{\partial\varepsilon}\left(\frac{\mathrm{tr}\left(C(F_\varepsilon)\right)}{T(F_\varepsilon)^tT(F_\varepsilon)}\right)^{1/2}\Bigg|_{\varepsilon=0} \\
&= \frac{1}{2}\frac{1}{\gamma_{\mathrm{VV}}}\,\frac{\partial}{\partial\varepsilon}\left(\frac{\mathrm{tr}\left(C(F_\varepsilon)\right)}{T(F_\varepsilon)^tT(F_\varepsilon)}\right)\Bigg|_{\varepsilon=0} \\
&= \frac{1}{2}\frac{1}{\gamma_{\mathrm{VV}}}\left[\frac{(T(F)^tT(F))\ \mathrm{tr}\left(\frac{\partial}{\partial\varepsilon}C(F_\varepsilon)\big|_{\varepsilon=0}\right)-\mathrm{tr}\left(C(F)\right)\ \ \frac{\partial}{\partial\varepsilon}\left(T(F_\varepsilon)^tT(F_\varepsilon)\right)\big|_{\varepsilon=0}}{\left(T(F)^tT(F)\right)^2}\right] \\
&= \frac{1}{2}\frac{1}{\gamma_{\mathrm{VV}}}\left[\frac{(\mu^t\mu)\ \mathrm{tr}\left(\mathrm{IF}(x,C,F)\right)-2\mathrm{tr}(\Sigma)\ \mu^t\mathrm{IF}(x,T,F)}{\left(\mu^t\mu\right)^2}\right] \\
&= \frac{\gamma_{\mathrm{VV}}}{2}\left[\frac{\mathrm{tr}(IF(x;C,F))}{\mathrm{tr}(\Sigma)}-2\,\frac{\mu^tIF(x;T,F)}{\mu^t\mu}\right]
\end{aligned}$$

For Voinov and Nikulin's coefficient of variation, one gets

$$\begin{aligned}
\mathrm{IF}(x;\mathrm{CV}_{\mathrm{VN}},F) &= \frac{\partial}{\partial\varepsilon}\left(T(F_\varepsilon)^tC(F_\varepsilon)^{-1}T(F_\varepsilon)\right)^{-1/2}\Big|_{\varepsilon=0} \\
&= -\frac{1}{2}\gamma_{\mathrm{VN}}^3\,\frac{\partial}{\partial\varepsilon}\left(T(F_\varepsilon)^tC(F_\varepsilon)^{-1}T(F_\varepsilon)\right)\Big|_{\varepsilon=0} \\
&= -\frac{1}{2}\gamma_{\mathrm{VN}}^3\Bigg[\ \frac{\partial}{\partial\varepsilon}T(F_\varepsilon)^t\Big|_{\varepsilon=0}C(F)^{-1}T(F) \\
&\qquad\qquad + T(F)^t\,\frac{\partial}{\partial\varepsilon}C(F_\varepsilon)^{-1}\Big|_{\varepsilon=0}T(F)+T(F)^tC(F)^{-1}\,\frac{\partial}{\partial\varepsilon}T(F_\varepsilon)\Big|_{\varepsilon=0}\Bigg] \\
&= -\frac{1}{2}\gamma_{\mathrm{VN}}^3\Bigg[\mathrm{IF}(x;T,F)^t\Sigma^{-1}\mu+\mu^t\left(-C(F)^{-1}\,\frac{\partial}{\partial\varepsilon}C(F_\varepsilon)\Big|_{\varepsilon=0}C(F)^{-1}\right)\mu \\
&\qquad\qquad + \mu^t\Sigma^{-1}\mathrm{IF}(x;T,F)\Bigg] \\
&= -\frac{1}{2}\gamma_{\mathrm{VN}}^3\left[2\mathrm{IF}(x;T,F)^t\Sigma^{-1}\mu-(\Sigma^{-1}\mu)^t\mathrm{IF}(x;C,F)(\Sigma^{-1}\mu)\right]
\end{aligned}$$

And finally, for Albert and Zhang's coefficient of variation,

$$\begin{aligned}
\mathrm{IF}(x;\mathrm{CV}_{\mathrm{AZ}},F) &= \frac{\partial}{\partial\varepsilon}\left(\frac{T(F_\varepsilon)^t C(F_\varepsilon)T(F_\varepsilon)}{\left(T(F_\varepsilon)^t T(F_\varepsilon)\right)^2}\right)^{1/2}\Bigg|_{\varepsilon=0} \\
&= \frac{1}{2}\frac{1}{\gamma_{\mathrm{AZ}}}\,\frac{\partial}{\partial\varepsilon}\left(\frac{T(F_\varepsilon)^t C(F_\varepsilon)T(F_\varepsilon)}{\left(T(F_\varepsilon)^t T(F_\varepsilon)\right)^2}\right)\Bigg|_{\varepsilon=0} \\
&= \frac{1}{2}\frac{1}{\gamma_{\mathrm{AZ}}}\left[\frac{\left(T(F)^t T(F)\right)^2 \frac{\partial}{\partial\varepsilon}\left(T(F_\varepsilon)^t C(F_\varepsilon)T(F_\varepsilon)\right)\big|_{\varepsilon=0}}{\left(T(F)^t T(F)\right)^4}\right. \\
&\qquad\left. - \frac{\left(T(F)^t C(F)T(F)\right)\frac{\partial}{\partial\varepsilon}\left(\left(T(F_\varepsilon)^t T(F_\varepsilon)\right)^2\right)\Big|_{\varepsilon=0}}{\left(T(F)^t T(F)\right)^4}\right] \\
&= \frac{1}{2}\frac{1}{\gamma_{\mathrm{AZ}}}\left[\frac{\left(\mu^t\mu\right)^2\left(2\ \mathrm{IF}(x,T,F)^t\Sigma\mu + \mu^t\mathrm{IF}(x,C,F)\mu\right)}{\left(\mu^t\mu\right)^4}\right. \\
&\qquad\left. - \frac{\left(\mu^t\Sigma\mu\right)4\left(\mu^t\mu\right)\mathrm{IF}(x,T,F)^t\mu}{\left(\mu^t\mu\right)^4}\right] \\
&= \frac{1}{2\gamma_{\mathrm{AZ}}\ (\mu^t\mu)^2}\left(\mu^t IF(x;C,F)\mu + 2\ \mu^t\left(\Sigma - 2\ \gamma_{\mathrm{AZ}}^2(\mu^t\mu)I_{p\times p}\right)IF(x;T,F)\right)
\end{aligned}$$

□

Bibliography

Albert, A. and Zhang, L. (2010). A novel definition of the multivariate coefficient of variation. *Biometrical Journal*, 52, 667–675.

Ahmed S.E. (1995). A Pooling Methodology for Coefficient of Variation. *Sankhya: The Indian Journal of Statistics*, 57, 57–75.

Bennett, B.M. (1977). On multivariate coefficients of variation. *Statistische Hefte*, 18, 123–128.

Boik, R.J. and Shirvani, A. (2009), Principal components on coefficient of variation matrices, *Statistical Methodology*, 6, 21–46.

Groenveld, R. (2011). Influence Functions for the Coefficient of Variation, Its Inverse, and CV Comparisons, *Communications in Statistics : Theory and Methods*, 40, 4139–4150.

Hampel, F.R., Rousseeuw, P.J., Ronchetti, E.M. and Stahel, W.A. (1986). Robust Statistics : The Approach based on Influence Functions, Wiley

Healy, M.J.R. (1979). Outliers in clinical chemistry quality-control schemes, *Clinical Chemistry*, 25, 675–677.

Karkalousos, P. and Evangelopoulos A. (2011), Quality Control in Clinical Laboratories, in *Applications and Experiences of Quality Control*, Ed. Ivanov O., InTech.

Pison, G. and Van Aelst, S. (2004). Diagnostic plots for robust multivariate methods, *J. Comput. Graph. Statist.*, 13, 310–329.

Rousseeuw P.J. and A.M. Leroy (1987). *Robust regression and Outlier Detection* , Wiley.

Underhill, L.G. (1990). The coefficient of variation biplot. *Journal of Classification*, 7, 241–256.

Van der Vaart A.W. (1998). *Asymptotic statistics*. New-York: Cambridge University Press.

Voinov V.G. and Nikulin M.S. (1996). *Unbiased estimators and their applications*. Vol. 2, multivariate case. Dordrecht: Kluwer.

Zhou Q., Li S., Li X., Wang W. and Wang Z. (2006). Detection of outliers and establishment of target in external quality assessment programs, in *Clinica Chimica Acta*, 372, 94–97.