Performance of estimators of the coefficient of variation under distributional specification of the data

Raydonal Ospina[1] Jorge I. Vélez[2,3] Fernando Marmolejo-Ramos[4,5,#]

[1] *Department of Statistics, CAST - Computational Agriculture Statistics Laboratory, Federal University of Pernambuco, Recife/PE, Brazil.*
[2] *Universidad del Norte, Barranquilla, Colombia.*
[3] *The Arcos-Burgos Group, Department of Genome Sciences, John Curtin School of Medical Research, The Australian National University, Canberra, ACT, Australia.*
[4] *Gösta Ekman Laboratory, Department of Psychology, Stockholm University, Stockholm, Sweden.*
[5] *School of Psychology, The University of Adelaide, Adelaide, Australia.*
[#] *Corresponding author:* fernando.marmolejo.ramos@psychology.su.se

1 Simulation study

Here, we present another simulations to compare the behavior of each estimator of the CV. This results are not included in the paper once that the graphic representation can confuse the lector.

Identification of simulation scenarios: Without lost of generality, we generated samples for the Uniform distribution with the set the parameters $a = 0$ and $b = 1$; thus the mean is $1/2$, the variance is $1/12$ and $\theta = 1/\sqrt{3}$. A total of 50 simulation scenarios were studied.

In the Binomial distribution, $p = \{0.1, 0.2, 0.3, 0.4, \dots, 0.9\}$ making up a total of 54 simulation scenarios (combinations of p and n). As the mean and standard deviation are np and $\sqrt{np(1-p)}$, respectively. In particular, for $p = 0.5$ (value that maximize the variance) follows that $\theta = \sqrt{np(1-p)}/np = 1/\sqrt{n}$. Note that, independently of p, when $n \to \infty$, the population CV, θ tends to zero.

When working with the Poisson(λ) distribution, we considered $\lambda = \{4, 5, 6, \dots, 20\}$ (A total of 17 choices for λ) Because the mean and variance are equal to λ, it follows that $\theta = 1/\sqrt{\lambda}$. The normal distribution can also be used to approximate the Poisson distribution for large values of λ. Typically, $\lambda > 20$ produce suitable approximations and in this way we do not consider another scenarios. Note that for large values of λ, the θ converges to zero. We therefore evaluated 102 simulation scenarios.

Among all distributions considered in this study, the Exponential(λ) distribution is particularly interesting because, regardless of λ, the value of the classic CV is always equal to 1. In fact, the mean and standard deviation are $1/\lambda$ and $\sqrt{1/\lambda^2}$, respectively. We set $\lambda = \{2, 4, 6, \ldots, 20\}$ and used the same sample sizes mentioned above to compare the different estimators of the CV across a total of 60 simulation scenarios.

In order to evaluate the effect of the sample size and the parameter ν (Degrees of freedom) on the estimators of the CV when the data come from a χ^2_ν distribution, n was varied as previously described and $\nu = \{2, 4, 6, \ldots, 20\}$ was considered. Sixty simulation scenarios were evaluated in this case.

For the Beta(α, β) distribution, the parameters α and β took values in $\{2, 4, 6, 8, 10\}$. The mean and variance of the Beta(α, β) distribution are given by $\alpha/(\alpha+\beta)$ and $\alpha\beta/\{(\alpha+\beta)^2(\alpha+\beta+1)\}$, respectively; it is straightforward to show that $\theta = \beta^{1/2}\{\alpha(\alpha+\beta+1)\}^{-1/2}$. Letting $\alpha = \beta$, the expression for the mode simplifies to 1/2, showing that for $\alpha = \beta > 1$ the mode (respectively, anti-mode when $\alpha = \beta < 1$), is at the center of the distribution and the $\theta = 1/\sqrt{2\beta+1}$, is a decreasing function of β. A similar approach was used for the Gamma(α, β); the mean, variance and classic CV are given by $\alpha\beta^{-1}$, $\alpha\beta^{-2}$ and $\theta = \beta^{-1/2}$, respectively. Note that, θ is a function only of scale parameter. A total of 150 simulation scenarios were evaluated in both cases.

For the case of the Normal(μ, σ^2) distribution, we set $\mu = \{10, 15, 20, 25, 30\}$ and $\sigma = \{0.5, 1, 2, 3, 4, 5\}$ and used the sample size mentioned above. The total number of scenarios under evaluation was 180. It is worth mentioning these values of μ and σ^2 were chosen to guarantee that no observations would fall outside the $(\mu - 2\sigma, \mu + 2\sigma)$ limits and so the sample estimators of the CVs in Table **??** would always be positive. Let us recall that, for normally distributed data, $\approx 95\%$ of the distribution falls within two standard deviations around the mean.

The Ex-Gaussian distribution (also called as exponentially modified Gaussian (EMG)) is defined by the parameters μ, σ and ν. The ex-Gaussian approach offers a parsimonious parameterization that describes the shape of a typical Reaction Time (RT) distribution, which resembles a normal distribution but with heavy right tail. The ex-Gaussian model assumes that RT distribution can be approximated by convolution of a normal and an exponential function. The parameters μ and σ are the mean and the standard deviation of a Normal distribution. Here, ν is a decay rate and reflects extremes in performance. (More details about this distribution see Burbeck and Luce (1982); Dawson (1988) and (Palmer et al., 2011)). To generate observations from this distribution, we followed the strategy described by Marmolejo-Ramos et al. (2015) where $\mu = \{200, 300, 400, 500, 600\}$, σ was fixed at 20, and $\nu = \{400, 300, 200, 100, 50\}$. A total of 150 simulation scenarios were evaluated.

1.1 Results

The main results are presented in Figures 1–6. Figure 1 displays our findings for the Uniform$(0, 1)$ distribution.

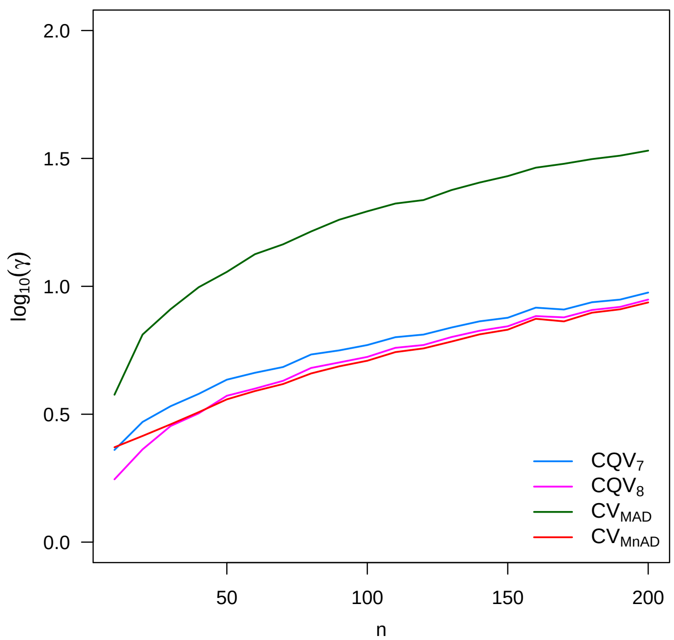

Figure 1: Values of $\log_{10}(\gamma)$ as a function n for the Uniform$(0, 1)$ distribution.

Compared to the classical estimator of the CV, the CV_{MnAD} estimator seems to be a plausible alternative, with highest relative efficiency compared to the baseline, regardless of n, followed by the CQV_8, CQV_7 and CV_{MAD} estimators. In fact, for $n = 100$ (considered a large sample size) we have estimates of the MSE of CVs given by $CV = 0.0018$, $CQV_7 = 0.0106$, $CQV_8 = 0.0095$, $CV_{\text{MAD}} = 0.0355$, and $CV_{\text{MnAD}} = 0.0092$. Note there is a slight difference in performance between the CQV_7 and CQV_8 estimators of the CV, which highlights the importance of carefully selecting the `type` of quantile estimator to be used. It is indisputable that the CV_{MnAD} estimator outperforms better that the CV_{MAD}, and that it also outperforms the others estimators.

Figure 2 displays γ for the Binomial, Poisson, Exponential and $\chi^2 u$ distributions. In the Binomial distribution, the CQV_7 and CQV_8 estimators of the CV present the higher values of γ among all estimators evaluated. The MSE of these two estimators behaves in a similar way, and is up to 50 times higher than that of the classical estimator of the CV for this distribution. Although there are some combinations of n and p for which γ is low (i.e., $\gamma < 10$), none of these estimators performs equally well (or better) than the classical estimator of the CV. That is, there is no combination of n and p for which $\gamma \leq 1$. The 3D surface plots in top panel of the CQV_7, CQV_8 and CV_{MnAD} shows spike-like patterns, indicating unstable of these estimators. In contrast, the CV_{MAD} estimator is more stable and show lower values of γ regardless of n and p. Overall, out of the four estimators under evaluation, the CV_{MAD} estimator seems to be a plausible alternative to the classical estimator of the CV despite its maximum value of $\gamma \approx 8$ when $n = 200$ and $p = 0.7$.

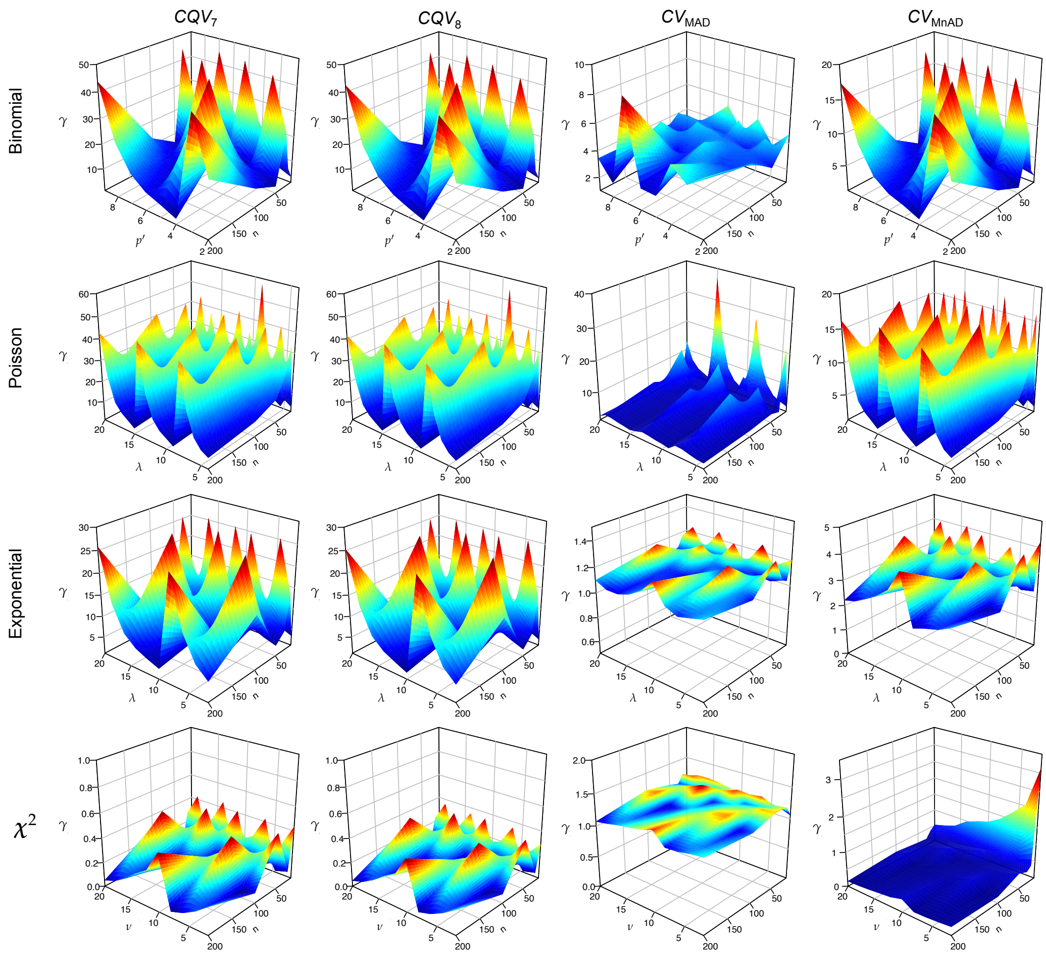

Figure 2: γ as a function of n and the corresponding parameters of the Binomial(n, p), Poisson(λ), Exponential(λ) and χ^2_ν distributions. Higher values of γ are presented in dark red, and lower values in dark blue. Note that, in the first row, $p' = 10p$ is shown. Conventions as in Figure 1.

Regardless of n and λ, the CQV_7 and CQV_8 estimators present higher values of γ than the CV_{MAD} and CV_{MnAD} estimators of the CV in the Poisson(λ) distribution. This indicates that the former two estimators have higher MSEs than the latter two (Figure 2, second row). In practical terms, this implies that the median-based estimators evaluated herein perform better than the CQV_7 and CQV_8 estimators. Despite some specific higher values of γ when n is quite small, the CV_{MAD} estimator is a plausible alternative to the classical estimator of the CV, especially for large sample sizes (that is, $n > 100$). Although regardless of n and λ the CV_{MnAD} estimator is a better alternative than the CQV_7 and CQV_8 estimators of the CV, its performance is not comparable to that of the CV_{MAD} estimator.

Similar results to those described above were also observed for the Exponential(λ) distribution. Whilst the CQV_7 and CQV_8 estimators perform poorly, the CV_{MAD} and CV_{MnAD} estimators have a better performance when compared to the classical estimator of the CV (Figure 2, third row). Note that $\gamma \in (1.02, 1.28)$ for the CV_{MAD} estimator; the minimum value is reached when $n = 50$ and $\lambda = 8$, and the maximum when $n = 20$ and $\lambda = 6$. This result indicates that the CV_{MAD} estimator is a feasible alternative to the classical estimator of the CV, especially when a small-to-large sample size is available and the time between events is relatively large. In practical terms, this implies that, compared to that of the classical estimator of the CV, the MSE of the CV_{MAD} estimator is $2 - 28\%$ higher. Close inspection revealed that, on average, $\gamma = 1.11$, and $\gamma = 1.10$ when $n = n_{\max} = 200$ and $\lambda = \lambda_{\max} = 20$.

When samples of size n are drawn from a χ^2_ν distribution, the CQV_7 and CQV_8 estimators behave similarly (Figure 2, bottom). Surprisingly, $\gamma \leq 1$ for both estimators, which implies that the classical estimator of the CV is outperformed by either of them regardless of n and ν. In particular, $\gamma_{CQV_7} \in (0.04, 0.47)$ and $\gamma_{CQV_8} \in (0.04, 0.42)$; these values are reached when $n = 20$ and $\nu = 16$, and $n = 30$ and $\nu = 10$ in the former, and when $n = 20$ and $\nu = 16$, and $n = 200$ and $\nu = 10$ in the latter. Detailed analysis of the results from the median-based estimators of the CV revealed that the CV_{MnAD} estimator outperforms the CV_{MAD} estimator when either $n > 10$ regardless of ν, or when $n = 10$ provided that $12 < \nu \leq 20$. Overall, our findings indicate that the CQV_8 is a better alternative to the classical estimator of the CV estimator for this particular distribution.

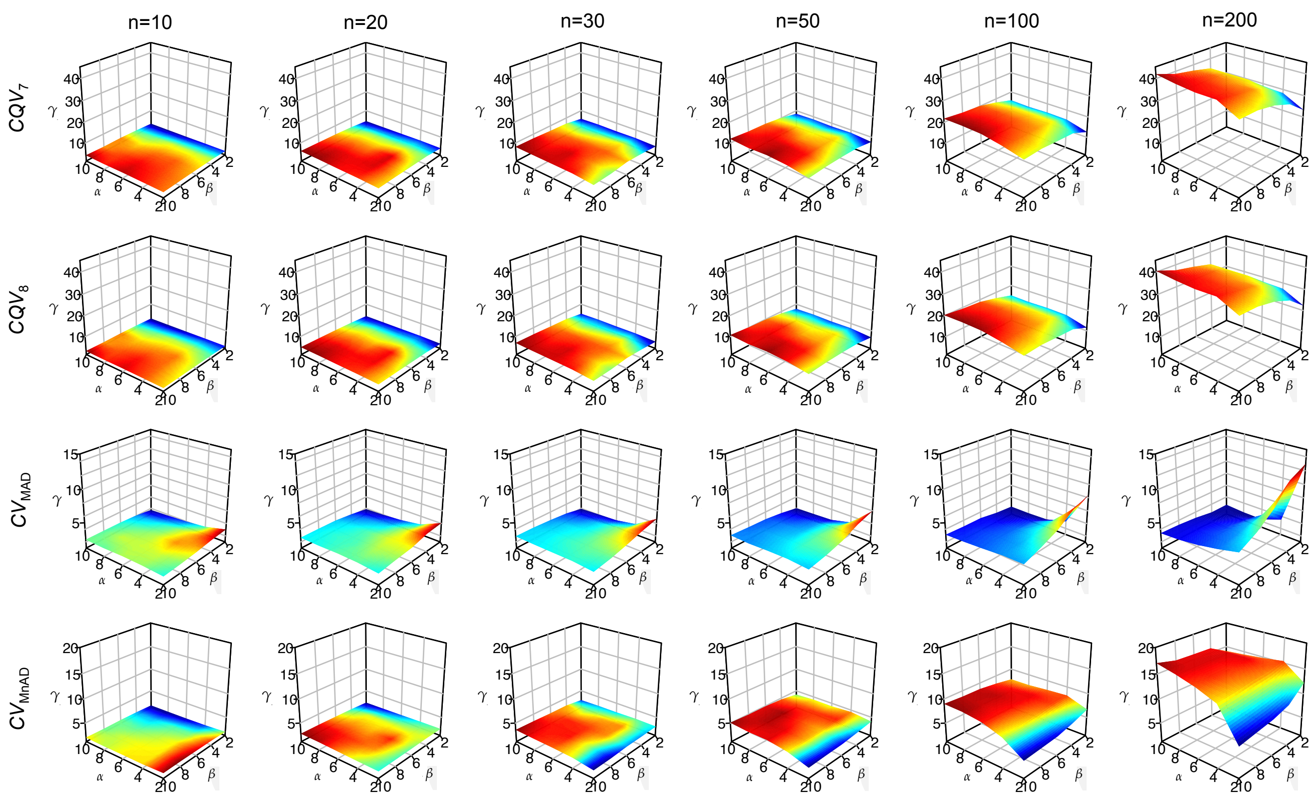

Figure 3: γ as a function of n for the Beta(α, β) distribution. Conventions as in Figure 1.

Figure 3 depicts the values of γ for the Beta(α, β) distribution as a function α, β, n and the CV estimator. Although our results indicate that none of the estimators of the CV provided values of $\gamma \leq 1$ for this particular distribution, there are several aspects worth highlighting. First, none of the

CV estimators evaluated has comparable MSE values (that is, $\gamma > 1$ in all cases) to the classical CV estimator. Secondly, the CQV_7 and CQV_8 estimators have the poorest performance among the four estimators under evaluation regardless of α and β, especially when $n \geq 50$. Third, γ rapidly increases with n for the CQV_7 and CQV_8 estimators, but the same does not seem to occur for the CV_{MAD} or CV_{MnAD} estimators. This particular result implies that, as n increases, the performance of the two former estimators worsens in favor of using the classical estimator of the CV instead. Fourth, γ seems not to be considerably affected by α and β for the CQV_7 and CQV_8 estimators, at least for $n < 100$ fixed; however, whilst $\gamma_{CV_{\mathrm{MAD}}}$ increases when α and β are small, $\gamma_{CV_{\mathrm{MnAD}}}$ decreases only when α is small.

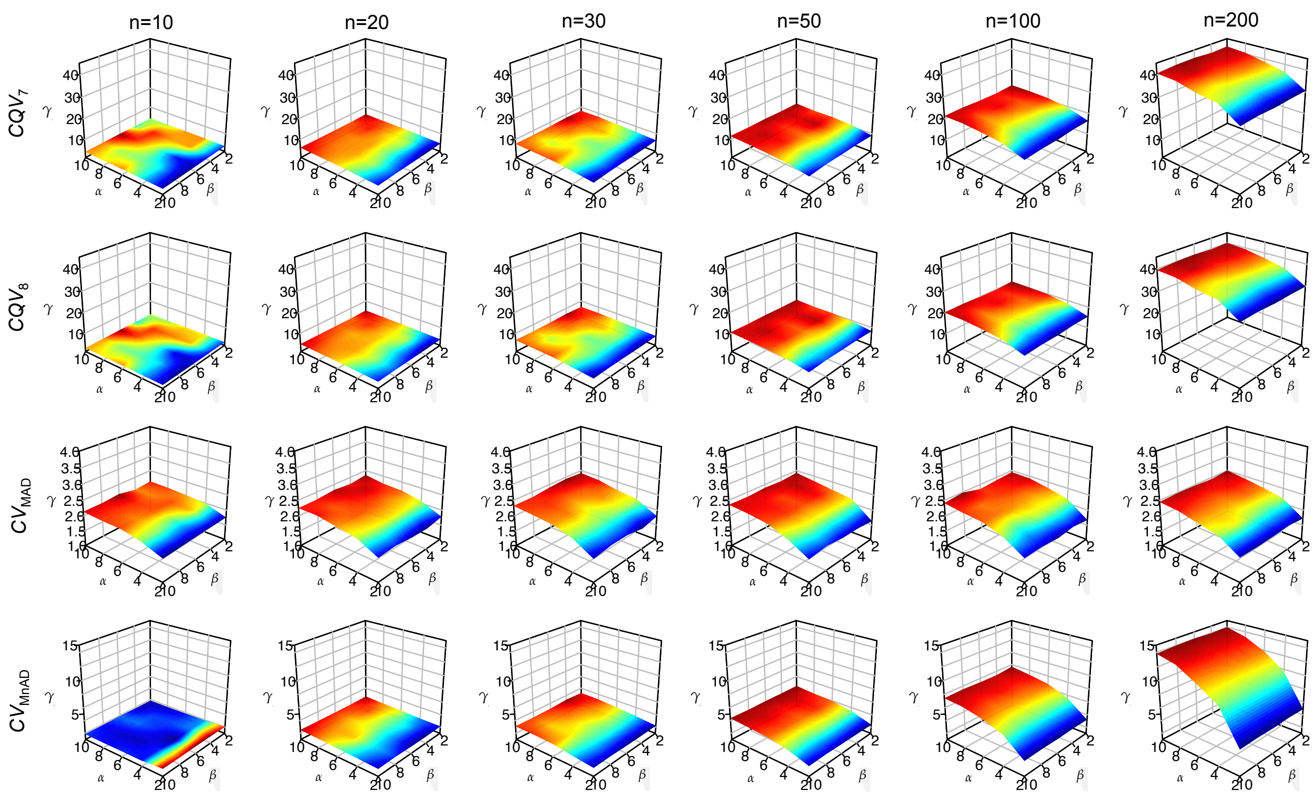

Figure 4: γ as a function of n for the Gamma(α, β) distribution. Conventions as in Figure 1.

The results for the Gamma(α, β) distribution as a function α, β, n and the CV estimator are shown in Figure 4. Similarly to what was found in the Beta(α, β) distribution, in this case the values of γ for the CQV_7 and CQV_8 estimators perform equally poorly, and the value of γ significantly increases as n increases regardless of α and β. This result implies that the CQV-based estimators do not perform as well as the classical estimator when n increases. Though the lowest values of γ are obtained when n is small (that is, when $n = 10$), these estimators are far from being equivalent to the classical estimator of the CV for this particular distribution (i.e., $\gamma > 1$ in all cases) regardless of n, α and β.

On the other hand, the CV_{MAD} and CV_{MnAD} estimators seem to be a better alternative to the CQV-based estimators above, with the former outperforming the latter regardless of n, α and β (Figure 4, third and fourth rows). In particular, the values of $\gamma_{CV_{\rm MAD}}$do not change as n increases, and reach their lowest value when $\alpha = 2$ regardless of β. Although $\gamma_{CV_{\rm MAD}} > 1$ in all cases, this latter result implies that the performance of the $CV_{\rm MAD}$ estimator is not affected by the sample size n. In contrast, the performance of the $CV_{\rm MnAD}$ is always affected by n, with the lowest and highest values of $\gamma_{CV_{\rm MnAD}}$ being obtained when $n = 10$ and $n = 200$, respectively (Figure 4, bottom). For this particular distribution, our results point out that, although $\gamma > 1$, the $CV_{\rm MnAD}$ estimator is a feasible alternative to the classical estimator of the CV when the sample size is small (that is, $n \leq 20$), and the $CV_{\rm MAD}$ estimator should be the natural choice when mid-to-large sample sizes (that is, $n > 30$) are available.

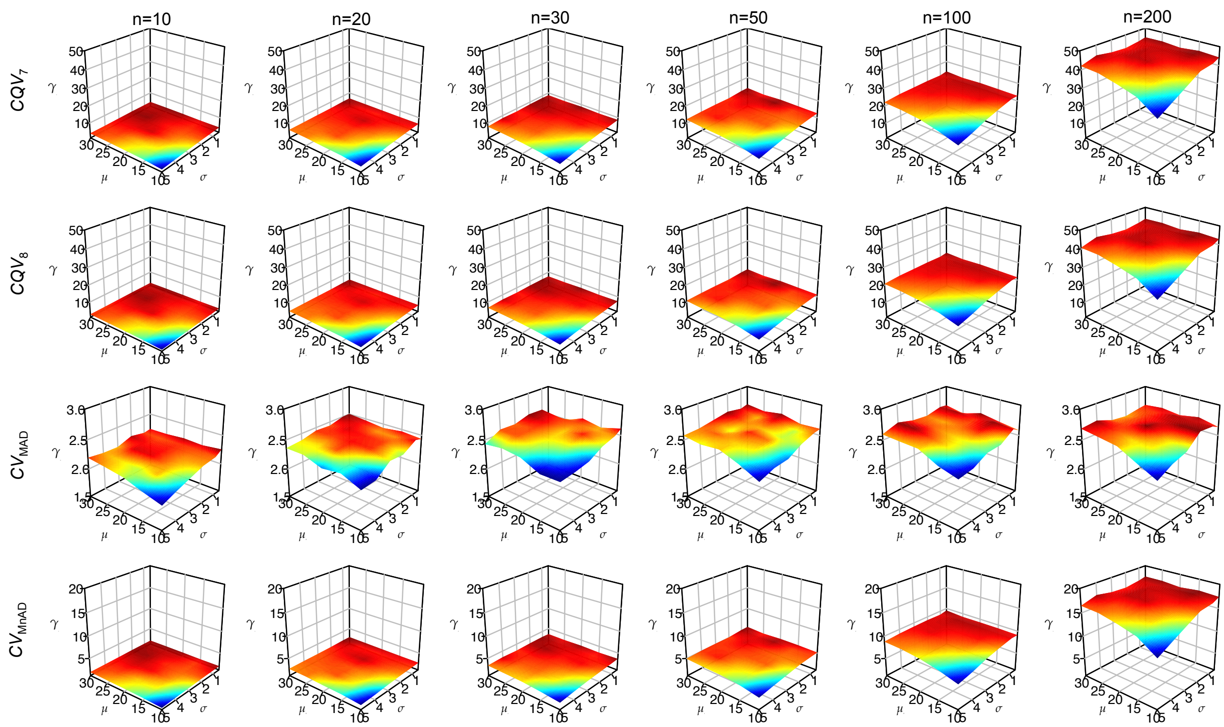

Figure 5: γ as a function of n for the Normal(μ, σ^2) distribution. Conventions as in Figure 1.

Figure 5 shows the results of γ for the Normal(μ, σ^2) distribution. The values of γ for the CQV_7 estimator increase as the sample size increase, with a minimum value of $\gamma_{\rm min} = 2.46$ when $n = 10$, $\mu = 10$ and $\sigma = 10$, and a maximum value of $\gamma_{\rm max} = 44.98$ when $n = 200$, $\mu = 15$ and $\sigma = 0.5$ (Figure 5, first row). In practical terms, this result implies that, as n increases, the CQV_7 will produce higher MSE values than the classical CV estimator. In other words, this finding plays against using the CQV_7 estimator instead of the classical CV estimator when the data comes from a Normal(μ, σ^2) distribution, especially when the sample size is large. A similar result was obtained for the CQV_8 estimator ($\gamma_{\rm min} = 1.86$ at $n = 10$, $\mu = 10$ and $\sigma = 5$; $\gamma_{\rm max} = 43.60$ at $n = 200$, $\mu = 15$ and $\sigma = 0.5$), (Figure 5, second row). For the CQV-based estimators, our findings suggest that the $CV_{\rm MAD}$ estimator performs better than its $CV_{\rm MnAD}$ counterpart, especially when $n > 30$. In particular, our results indicate a more consistent behaviour of the former estimator over the latter, which may favour its use over the classical estimator of the CV despite that the condition $\gamma_{CV_{\text{MAD}}} \leq 1$ does not hold (see Equation **??** and the text underneath for more details). Close inspection shows that $\gamma_{CV_{\text{MAD}}} \in (1.92, 2.73)$ and $\gamma_{CV_{\text{MnAD}}} \in (1.39, 17.62)$, from which it is straightforward to conclude that the MAD-based estimator is a better choice. It is also quite surprising that the $\gamma_{CV_{\text{MAD}}}$ slightly increases with n. On a side note, it is interesting that the lowest values of γ were obtained when μ and σ are small regardless of which CV estimator is used.

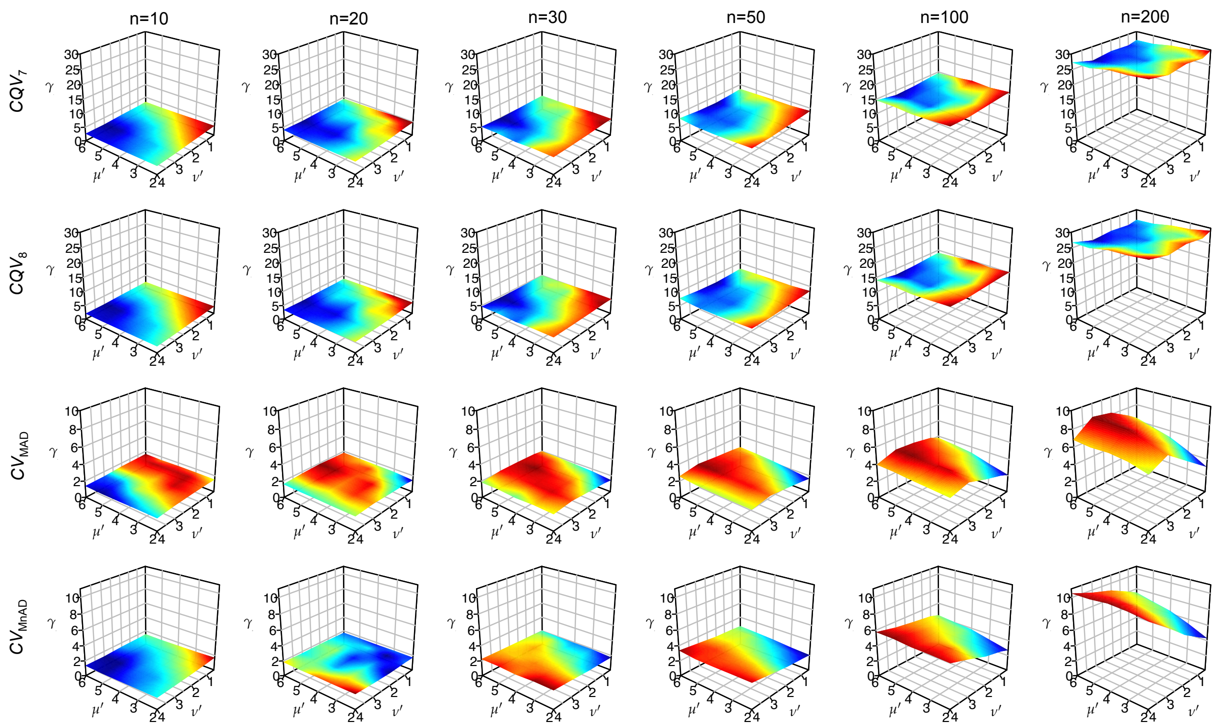

Figure 6: γ as a function of n, $\mu' = \mu/100$ and $\nu' = \nu/100$ for the Ex-Gaussian(μ, σ, ν) distribution with $\sigma = 20$ fixed. Conventions as in Figure 1.

In Figure 6 we show the values of γ for the ExGaussian(μ', σ, ν') distribution as a function of the sample size n and the parameters $\mu' = \mu/100$ and $\nu' = \nu/100$ when $\sigma = 20$ is fixed. Although none of the estimators evaluated showed an equivalent or better performance than the classical estimator of the CV (see Equation **??**), there are some aspects that deserve to be described.

Firstly, the values of γ for the CQV_7 and CQV_8 estimators, as observed in the previous distributions, increase as a function of n regardless of μ' and ν'. In general, $\gamma_{CQV_7} \in (2.50, 29.90)$; the minimum and maximum values were reached at $\mu' = 6$ and $\nu' = 1$ when $n = 10$, and at $\mu' = 2$ and $\nu' = 0.5$ when $n = 200$, respectively. Similarly, $\gamma_{CQV_8} \in (1.93, 29.29)$; the minimum and maximum values were reached at the combinations of n, μ' and ν' described above. When $n > 50$, the MSE of these two estimators is higher than that of the classical estimator of the CV, making them, in practice, a less feasible alternative to the latter (Figure 6, first and second rows). Altogether, these results indicate how similar the performance of the CQV-based estimators is, and that they do not represent a suitable choice to the classical CV estimator for this particular distribution.

Secondly, there seems to be little to no difference in performance between the MAD- and MnAD-based estimators of the CV when $n < 50$. However, the CV_{MAD} performs slightly better than the CV_{MnAD} estimator regardless of μ' and ν' when $n > 50$. It is quite surprising that, in contrast with the results obtained for most of the distributions being tested, $\gamma_{CV_{\mathrm{MAD}}}$ increases with n as samples of this are drawn from the ExGaussian(μ, σ, ν) distribution.

References

Burbeck, S. L. and Luce, R. D. (1982) Evidence from auditory simple reaction times for both change and level detectors. *Perception & Psychophysics*, **32**, 117–133.

Dawson, M. R. W. (1988) Fitting the ex-gaussian equation to reaction time distributions. *Behavior Research Methods, Instruments, & Computers*, **20**, 54–57.

Marmolejo-Ramos, F., Vélez, J. I. and Romão, X. (2015) Automatic outlier detection of discordant outliers using ueda's method. *Journal of Statistical Distribution and Applications*, **2**, 1–14.

Palmer, E. M., Horowitz, T. S., Torralba, A. and Wolfe, J. M. (2011) What are the shapes of response time distributions in visual search? *Journal of experimental psychology. Human perception and performance*, **37**, 58–71. URLhttp://dx.doi.org/10.1037/a0020747.