The Influence of Lag-1 Autocorrelations on Goodness-of-Fit Tests for Reverberation Chambers

Ramiro Serra*, Andrea Cozza†, Florian Monsef†
*Eindhoven University of Technology, Eindhoven, The Netherlands.
Email: r.serra@tue.nl
†L2S, UMR8506, Département de Recherche en Electromagnétisme
SUPELEC, Univ Paris-Sud, CNRS, Gif-sur-Yvette, France

February 18, 2014

ABSTRACT

Goodness-of-fit (GoF) tests are often used in reverberation chambers to check whether the chamber is operating in an overmoded regime. The tests are usually performed with the prerequisite that the field sampled values must be uncorrelated. In the present paper we recall that no theoretical background is necessarily enforced for the samples to be independent when using such statistical tools. We analyze the sensitivity of the most common GoF tests used in electromagnetic compatibility to the degree of sample autocorrelation.

Index Terms—Autocorrelation coefficient, goodness-of-fit tests, random fields, reverberation chambers, statistical electromagnetism.

1 INTRODUCTION

Knowing and assessing the performance of reverberation chambers (RC) [1] [2] for a large number of operating conditions is of crucial importance for many electromagnetic compatibility (EMC) applications. To this end, several performance indicators have been proposed and implemented by RC users. There exist a significant number of such performance indicators, each one with its advantages and disadvantages. Generally speaking, the robustness and ease to implement are the two most valued characteristics when choosing which indicator to use.

Nevertheless, one must be aware of the nature of each indicator, since they all represent a different "look" at RC performance. For instance, if undergoing an optimization process, only one indicator is taken into account, the general conclusions could be misleading [3].

Furthermore, some of them are deeply interrelated and rooted on the same fundamentals. As an example: field uniformity and field inhomogeneity, as defined in [4], are different indicators both based on the fact that, in a well-performing RC, the mean energy density is independent of position [1].

In this paper we focus on the relation between two of the most important performance indicators used in literature and suggested in [4]: the autocorrelation coefficient and the goodness-of-fit (GoF) tests. For those, is commonly accepted that data to be tested must be *independent*. Data uncorrelation then seems like a prerequisite before applying the different GoFs. But a closer look at their specific null hypotheses suggests this is likely not the case. Indeed, as far as we have investigated, no theoretical argument was found that justifies the samples to be independent when applying GoF tests. This not withstanding, one can find some source [5] explaining the sensitivity to the way data are formatted depends on the GoF nature, i.e., if those are parametric or non-parametric.

As results, it remains obscure if in order to adequately check for the performance of an RC by using GoF tests, one should assure first, or not, uncorrelation of data. In other words, if these GoF tests *require* or not independent and identically distributed (i.i.d.) data as input. What is then the practical impact when this prerequisite is not met? There is a need to better clarify what this impact could mean in practice, since the autocorrelation coefficient estimator ρ is significantly ill-behaved for autocorrelation values close to zero. Correlation is seldom (or never) total. Are we then willing to throw away useful data, that cost time and efforts to obtain, just because their information is not "absolutely new"?

Our findings show up to what extent different GoF tests usually implemented within the EMC community are sensitive to data autocorrelation.

2 AUTOCORRELATION COEFFICIENT AND THRESHOLDS

In an ideal RC the field distribution inside the working volume between one stir state and the following one would be expected to change drastically and keep no similarities between them. The autocorrelation coefficient, as defined in [4] and referred to herein as $\rho(r)$, provides a measure of this effect by quantifying the similarity, i.e. the covariance, between a set of measured data (e.g. the power received by an antenna, the electric field, etc) at a fixed spatial position, and the same set of data but circularly shifted a variable number of 'lags' such that,

$$\rho(r) = \frac{1}{N} \frac{\sum_{i=1}^{N} (x_i - \langle x \rangle)(x_{i+r} - \langle x \rangle)}{s_x^2}, \qquad (1)$$

where N is the total number of stirrer positions and $x_1, x_2, x_3, \ldots x_N$ is the field magnitude of interest (x) at consecutive stirrer positions. s_x^2 is the sample variance estimator and r is the number of lags. The subindex $i+r$ is *modulo* N [4]. If the stirring process is efficient at a given frequency, then the autocorrelation coefficient ρ should be low with relatively low stir state lags r.

Even though these autocorrelation coefficient provides us with a consistent and fairly reliable way for RC characterization, it bears two significant inconveniences in its practical application. Firstly, because it is difficult to compare between two (or more) different performances, since $\rho(r)$ depends on the specific number of lags. Secondly,

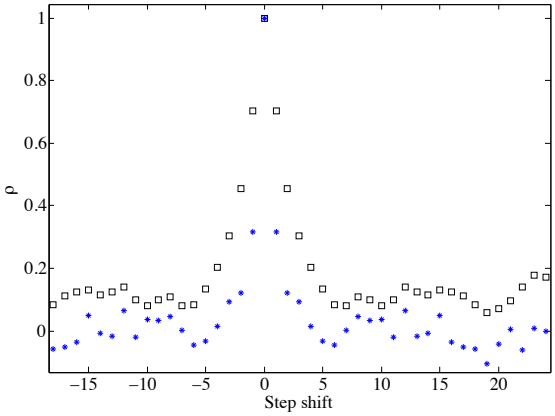

Fig. 1: Autocorrelation function obtained by using the model given by (2) for $\rho = 0.3$ (asterisks) and $\rho = 0.7$ (squares), respectively.

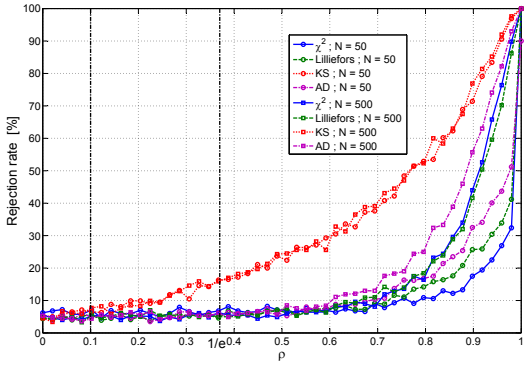

Fig. 2: Rejection rates ($\alpha = 5\%$) for different samples sizes and GoF tests as a function of data autocorrelation. The vertical lines highlight the two most commonly used thresholds for ρ.

since it is in practice very unlikely to obtain exactly $\rho = 0$ for a finite number of samples $N < \infty$ out of finitely long ensembles (even for ideally uncorrelated samples). Different thresholds are defined in order to decide for data uncorrelation. For instance, [4] assumes the commonly accepted value $|\rho| \leq \frac{1}{e} \approx 0.37$. Sometimes the more stringent value of $|\rho| \leq 0.1$ is used as in [6], [7]. In [8] it is emphasized the importance of understanding the role of this threshold in the estimation of the number of independent samples.

Generally speaking, the choice of a meaningful threshold for ρ depends on two basic criteria: the sample size N and the level of significance for accepting the null hypothesis that data are uncorrelated. For instance, as stated in [7], a threshold of $\frac{1}{e}$ may be acceptable for $N = 30$ with a 5% level of significance. For the same threshold, more samples are needed ($N = 50$) if one would like a 1% level of significance. It is mentioned in the conclusions of [7] that the proper choice of this threshold must be based on a combination of statistical analysis and engineering judgements, in other words, in the trade-off between measurement uncertainty and time (cost). The aim of our paper is to provide another view on the definition and role of this threshold. We investigate its impact on the outcome of different GoF tests.

3 Goodness-of-Fit Tests

It is commonly accepted that under good reverberation conditions, the various field quantities follow known probability density functions (pdf) for an ensemble of stir states. The basic underlying quantities, i.e. the *per axis* in-phase and quadrature components of the electric and magnetic fields, follow independent Gauss-normal distributions [1], [2] with zero mean and equal variances [9]. All other field quantities based on these elementary ones (i.e. field strength, phase, energy densities, etc) thus belong to the circular-Gauss family of distributions.

It is therefore quite natural to assess RC performance by comparing a set of measured samples with its theoretical pdf. This comparison, however, cannot be done just by "the eye". Comparing a histogram with its fitted pdf in this manner could potentially be misleading and biased. A friendly eye would conclude differently than an unfriendly one, and a trained eye would assess contrarily to a neophyte in such matters. In order to reduce subjectivity, GoF tests help in providing a quantified results on how good a pdf fits a given histogram. They still keep some subjective aspects, such as the choice of bin sizes, critical values, etc, but all together they allow everybody to draw the same conclusions on the same sets of data if the parameters are kept the same.

Nevertheless, there is often the question of which GoF test to apply. Indeed, there is a wide family of GoF tests spanning from the laxest ones in one extreme up to the most stringent ones on the opposite extreme. Some of them give more weight to deviations in the tails of the pfds, some of them give more importance to the converge towards the central tendency of a data set, etc. Our main research question, leading to this contribution is: How much importance GoF tests give to the autocorrelation of data?

4 The Influence of Lag-1 Autocorrelation on Goodness-of-Fit Tests

In order to shed some light on this topic, we choose to focus our research on lag-1 autocorrelation (i.e. $r = 1$) and the following GoF tests of normality, commonly encountered in RC literature:

- the χ^2 test [2];
- The Lilliefors (L) test [10]
- the Kolmogorov-Smirnov (KS) test [11]; and,
- the Anderson-Darling (AD) test [8], [12],;

All the aforementioned GoF tests are able to test for normality of data, and the order in which they are mentioned approximately ranks them from laxest to most stringent. In our context, GoF test stringency refers to how severe such a test would be in rejecting the null hypothesis. Stringent tests are known as "high power" tests [13], in reference to their power to detect even subtle deviations from normality. Even though they share their applicability to normality testing, their different null hypotheses H_0 differ slightly:

- $H_0^{\chi^2}$: the data are a random sample from a normal distribution.

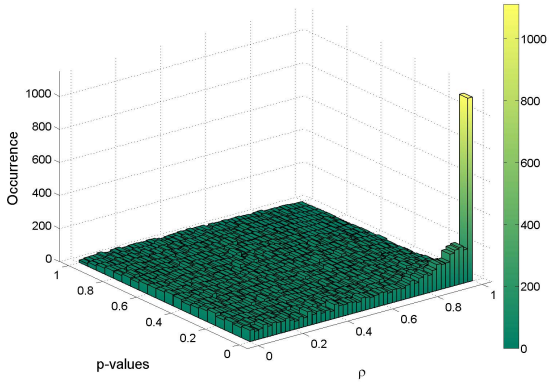

Fig. 3: Two-dimensional histogram for the p-values and ρ results for $H_0^{\chi^2}$

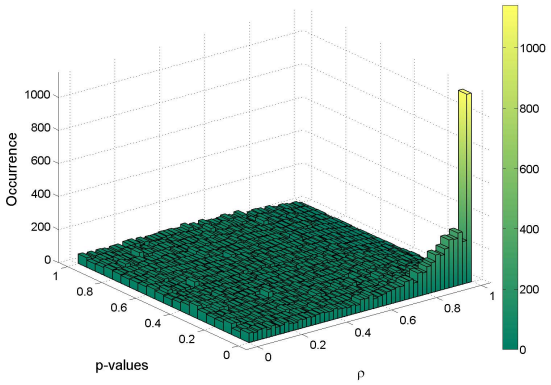

Fig. 4: Two-dimensional histogram for the p-values and ρ results for H_0^{L}

- H_0^{L}: the data come from an unspecified normal distribution.
- H_0^{KS}: the data is a random sample of a standard normal distribution $\mathcal{N}(0,1)$.
- H_0^{AD}: the data come from a normal distribution.

Subtle changes in their different null hypotheses can provide, apart from their different power, different results when testing goodness-of-fit to the same set of data. Particular attention must be paid to the KS test, since it requires data standardization before application, in order to have a sample with zero mean and unit variance. By using simulated data, we analyze the influence of these violations of the (apparent) i.i.d.-prerequisite on the rejection rates of GoF tests. We also indirectly investigate whether or not can GoF tests be used to assess the correlation of data sets, in particular in case of perfectly normal data.

4.1 Simulation Study

Data for the study were computer-generated using a simple model for generating $\mathcal{N}(0,1)$ data samples with a known *a priori* autocorrelation ρ, according to the following process:

$$\begin{aligned} x_1 &= \eta_1 \\ x_i &= \rho x_{i-1} + \sin(\arccos(\rho))\eta_i\,, \end{aligned} \qquad (2)$$

where the random variable η is strictly normal with zero mean and unit variance $\eta \sim \mathcal{N}(0,1)$. For such correlation model, we report in Fig. 1 the autocorrelation function obtained for $\rho = 0.3$ and $\rho = 0.7$, respectively.

For all simulations we choose positive values of ρ in the interval $[0,1]$. With equation (2), repeatedly, data sets x_i are generated that are used as input to the GoF tests. The simulation study is carried out by generating sets of x_i for $N = 50$ and $N = 500$ and for 50 values of $\rho \in [0,1]$. To reduce statistical uncertainties, each realization was repeated 1000 times.

4.2 Results

Fig. 2 summarizes the percentage of rejected null hypothesis of normality for different GoF tests under study and different number of samples, for a significance level

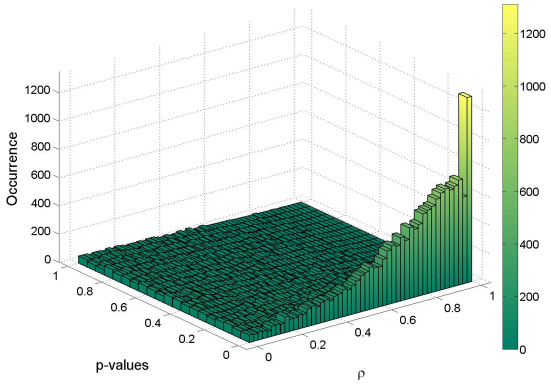

Fig. 5: Two-dimensional histogram for the p-values and ρ results for H_0^{KS}

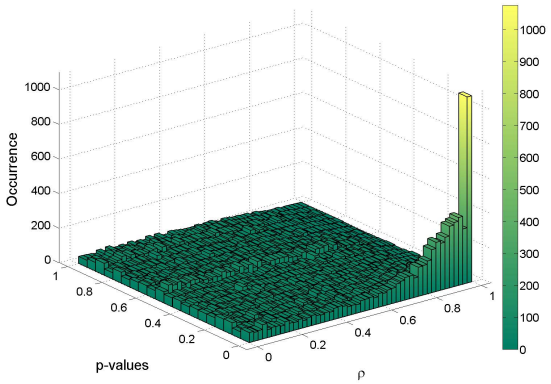

Fig. 6: Two-dimensional histogram for the p-values and ρ results for H_0^{AD}

$\alpha = 5\%$. We stress the fact that these rejection rates are solely coming from data correlation, since the tested data sets come from a correct normal parent distribution, as described by the process in Eq . (2). It can be seen that the KS test shows a higher sensitivity to data autocorrelation than the other GoF tests. The KS test is also more robust against different sample sizes. It seems that the KS test can be potentially suggested as a way of identifying residual autocorrelation on data, even for low values of ρ. Two lines have been added to Fig 2 indicating the two most commonly used values for autocorrelation thresholds.

The χ^2, Lilliefors and Anderson-Darling GoF tests seem to be significantly indifferent even for high autocorrelations, while the Kolmogorov-Smirnov test is susceptible to react on data autocorrelation for lower values of ρ. This result seems to suggest that most GoF tests are more robust against the autocorrelations found in practice than what is intuitively sound. Indeed, one would expect that it would be possible to even setup a criterion for autocorrelation thresholds based on desired levels of rejections, but our work proves that this is not possible.

A closer look on this aspect is represented in Figs. 3 - 6, where two-dimensional histograms show how the p-values resulting from the GoF tests are distributed for different bins of autocorrelation values. The case for N=50 is considered here. We can see that when correlation is low, the p-values resulting from the GoF tests are almost uniformly distributed. When the correlation rises, we observe that the number of occurrence of small p-values increases, i.e., that the null hypothesis is more often rejected. This effect occurs for relatively high correlation values, i.e., for values of ρ greater than 0.6, except for the KS test for which the number of rejects occurs sooner. This observed behavior of the p-values is in perfect accordance to what is known in literature, viz. that under the null hypothesis, p-values are uniformly distributed, and under the alternative hypothesis the distribution of p-values is not uniform, but with a tendency towards zero [14].

4.3 Discussion

It is commonly accepted that for proper RC operation, data must be independent. Correlated data, indeed, provides only partial information and might represent a waste of time and resources [8]. The data used in this study largely violates this apparent assumption of independence of statistical tests. Nevertheless, most of the GoF tests studied in the present work show significant robustness against correlated data, for two different sample sizes.

4.4 Conclusion

Our simulation study has shown that, in many cases, the rejection rates of GoF tests for normally-distributed data are surprisingly indifferent to non-i.i.d. data input. This finding questions the commonly accepted necessity of data uncorrelation.

However, it cannot be excluded that our study is based on only lag-1 autocorrelations. This implicitly assumes that the autocorrelation coefficient is 'well-behaved', i.e. monotonically declining with increasing lag distance. More general lag-r autocorrelations might be investigated in the future. Another possible improvement is the use of Portmanteau tests, such as the Ljung-Box test, in order to test for data autocorrelation. The problem with this kind of tests is that they tend to be too severe [15].

The results compiled in this article represent a step forward in the understanding of RC performance. The results are somehow counterintuitive and challenge our established habits.

The fact GoF tests are generally robust except in extreme cases of autocorrelation enable RC users to perform some statistical analysis and interpretation of data, independently of the independence (or not) of data.

REFERENCES

[1] D. Hill, "Plane-wave integral representation for fields in reverberation chambers," *IEEE Transactions on Electromagnetic Compatibility*, vol. 40, pp. 209–217, 1998.

[2] J. Kostas and B. Boverie, "Statistical Model for a Mode-Stirred Chamber," *IEEE Transactions on Electromagnetic Compatibility*, vol. 33, no. 4, pp. 366–370, 1991.

[3] R. Serra and F. Leferink, "Optimizing the Stirring Strategy for the Vibrating Intrinsic Reverberation Chamber," in *IEEE International Symposium on Electromagnetic Compatibility (EMC Europe), 2010*, 2010.

[4] *Reverberation chamber test methods*, International Electrotechnical Commission (IEC), Std. 61 000-4-21, 2011.

[5] E. Whitley and J. Ball, "Statistical review 6: Nonparametric methods," *Critical Care*, vol. 6, no. 6, pp. 509–513, 2002. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC153434

[6] L. Musso, "Assessment of Reverberation Chamber Testing for Automotive Applications," Ph.D. dissertation, Politecnico di Torino, Feb 2003.

[7] O. Lundén and M. Backstrom, "Stirrer Efficiency in FOA Reverberation Chambers. Evaluation of Correlation Coefficients and chi-squared tests," in *IEEE International Symposium on Electromagnetic Compatibility*, vol. 1. IEEE, 2000, pp. 11–16.

[8] C. Lemoine, P. Besnier, and M. Drissi, "Estimating the effective sample size to select independent measurements in a reverberation chamber," *IEEE Transactions on Electromagnetic Compatibility*, vol. 50, no. 2, pp. 227–236, 2008.

[9] R. Serra, F. Leferink, and F. Canavero, ""Good-but-imperfect" electromagnetic reverberation in a VIRC," in *IEEE International Symposium on Electromagnetic Compatibility (EMC)*, 2011, pp. 954–959.

[10] F. Monsef and A. Cozza, "Goodness-of-fit tests in radiated susceptibility tests," in *Workshop on Aerospace EMC, 2012 Proceedings ESA*. IEEE, 2012, pp. 1–5.

[11] P. Corona, G. Ferrara, and M. Migliaccio, "Reverberating chambers as sources of stochastic electromagnetic fields," *IEEE Transactions on Electromagnetic Compatibility*, vol. 38, no. 3, pp. 348–356, 1996.

[12] V. Mariani Primiani and F. Moglie, "Numerical simulation of reverberation chamber parameters affecting the received power statistics," *IEEE Transactions on Electromagnetic Compatibility*, vol. 54, no. 3, pp. 522–532, 2012.

[13] C. Lemoine, P. Besnier, and M. Drissi, "Investigation of reverberation chamber measurements through high-power goodness-of-fit tests," *IEEE Transactions on Electromagnetic Compatibility*, vol. 49, no. 4, pp. 745–755, 2007.

[14] D. J. Murdoch, Y.-L. Tsai, and J. Adcock, "P-values are random variables," *The American Statistician*, vol. 62, no. 3, 2008.

[15] R. Serra and A. C. Rodriguez, "The Ljung-Box test as a performance indicator for VIRCs," in *IEEE International Symposium on Electromagnetic Compatibility, (EMC Europe 2012)*, Sept 2012, pp. 1–6.