Assessing the Performance of Uncertainty Calculations by Simulation

B. D. Hall

Measurement Standards Laboratory of New Zealand,
Industrial Research Ltd., PO Box 31-310,
Lower Hutt 5040, New Zealand

Abstract

Some RF and microwave measurements pose difficulties for uncertainty calculations. A case in point is when the magnitude of the complex reflection coefficient of a weakly reflecting device is measured, such as a load or power sensor. To investigate this, simulations are used to generate a large number of independent sets of data for testing different calculation methods. We find that, while several standard methods of calculating uncertainty perform badly, a simple *ad hoc* modification can overcome these problems. The study shows that the simulation approach is a pragmatic way of assessing uncertainty calculation performance.

1 Introduction

Different methods have been proposed for calculating measurement uncertainty [1, 2, 3]. In most well-designed measurements, there is an almost linear relationship between the quantity to be measured and any quantities that contribute to the measurement result. In such cases the standard methods of uncertainty calculation usually perform well and are in agreement. However, there are situations where different methods of uncertainty calculation do not produce equivalent results when applied to the same set of data. The question is then: which, if any, of the uncertainty statements is correct?

To assess the performance of uncertainty calculation methods, simulated sets of data can be used [4]. The idea is quite simple. When experiments are simulated, the quantity intended to be measured (the measurand) is known, whereas it is not known in actual experiments. So, having simulated some experimental data and carried out the uncertainty calculation, one can check whether the uncertainty interval obtained has been 'successful' in bracketing the measurand. Moreover, it is possible to simulate many independent experiments and assess the long-run success rate of a method when applied to data typical of a particular measurement procedure. The success rate provides a measure of the 'level of confidence', or 'coverage probability', which is the performance parameter usually specified for methods of uncertainty calculation. An acceptable method of uncertainty calculation should generate intervals that contain the measurand on most occasions (typically, a level of confidence of 95% is used in metrol-

ogy). A procedure that cannot do this under test conditions is surely of little use in real measurements.

In this paper, the simulation method is applied to measurements of the complex reflection coefficient of a weakly reflecting object. The case is of practical interest to the RF measurement community because devices are often designed specifically to have small reflection coefficients. For instance, low-reflection loads are often used as calibration standards for network analyzers, and power sensors have a low reflection coefficient to reduce mismatch errors. Measurements are made to characterize the magnitude of such devices, and it is also of interest in some calibration work to characterize the squared magnitude of reflection coefficients.

Low reflection coefficient measurements pose difficulties for uncertainty calculations for two reasons. First, the quantity of interest has a lower bound, because the magnitude of a reflection coefficient cannot be negative. Second, the non-linearity of the transformation from rectangular to polar coordinates near the origin may violate assumptions made by some methods.

The intent of the paper is to show that simulation of measurement scenarios is a pragmatic way of assessing the performance of uncertainty calculations. The author has no wish to discourage use of the general methods of uncertainty calculation investigated. On the contrary, the verification process described here can be used to enhance confidence in their use, or, if problems are found, it allows *ad hoc* changes to a method to be investigated so that performance can be improved in specific types of measurement.

2 The Measurement Scenario

We consider measurements of a complex quantity $\mathbf{\Gamma} = \Gamma_1 + \mathrm{j}\Gamma_2$ and are interested in the values of $|\mathbf{\Gamma}|$ and $|\mathbf{\Gamma}|^2$. Measurement data $\mathbf{z} = z_1 + \mathrm{j}z_2$ are simulated, where z_1 and z_2 are considered to be independent, each with a standard uncertainty u that is associated with a Gaussian distribution. Estimates of the quantities of interest, $|\mathbf{\Gamma}|$ and $|\mathbf{\Gamma}|^2$, are obtained from

$$|\mathbf{z}| = \sqrt{z_1^2 + z_2^2} \tag{1}$$

$$|\mathbf{z}|^2 = z_1^2 + z_2^2 \ . \tag{2}$$

This scenario is simple to avoid other complicating aspects of the measurement problem, such as a finite number of observations (leading to uncertainty in the value of u) and correlation between z_1 and z_2. These could easily be included in a more detailed simulation study.

2.1 Uncertainty Calculations

Two methods of uncertainty calculation are considered. One is described in the *Guide to the Expression of Uncertainty in measurement* (GUM) [1], the other is a Monte Carlo method described in the first Supplement to the GUM (SUP) [2].

2.1.1 GUM Method

Following the *Law of Propagation of Uncertainty* (LPU) in the GUM, the standard uncertainty in the measurement of $|\mathbf{\Gamma}|$ is $u(|\mathbf{z}|) = u$, which does not depend on the observed data z_1 and z_2 (see [4] for details). An uncertainty interval for $|\mathbf{\Gamma}|$ with a 95% coverage level is[1]

$$[\, |\mathbf{z}| - 1.96u, |\mathbf{z}| + 1.96u \,] \,. \tag{3}$$

For measurements of $|\mathbf{\Gamma}|^2$, the LPU gives a standard uncertainty $u(|\mathbf{z}|^2) = 2|\mathbf{z}|u$. So an uncertainty interval with a 95% coverage level is[2]

$$[\, |\mathbf{z}|^2 - 3.92|\mathbf{z}|u, |\mathbf{z}|^2 + 3.92|\mathbf{z}|u \,] \,. \tag{4}$$

2.1.2 SUP Method

The SUP method of calculating uncertainty used here can be described in a series of steps [2]. For the measurement of $|\mathbf{\Gamma}|$ we

1. Generate sequences of samples z_{1i} and z_{2i}, where $i = 1, \cdots, L$ and L is a large number, by drawing from a pair of Gaussian distributions with means z_1 and z_2, and variance u^2.
2. Calculate $|\mathbf{z}_i| = \sqrt{z_{1i}^2 + z_{2i}^2}$, for $i = 1, \cdots, L$.
3. Sort these values in ascending order.
4. Take the $0.025L^{\text{th}}$ smallest value as the lower bound of the uncertainty interval.
5. Take the $0.975L^{\text{th}}$ smallest value as the upper bound of the uncertainty interval.

The result is an interval said to have 95% probability of containing $|\mathbf{\Gamma}|$.

A similar procedure applies to the measurement of $|\mathbf{\Gamma}|^2$ except that at step 2 we calculate $|\mathbf{z}_i|^2 = z_{1i}^2 + z_{2i}^2$. The result is an interval said to have 95% probability of containing $|\mathbf{\Gamma}|^2$.

2.2 The evaluation method

To assess the long-run success rates of the GUM and SUP methods, observations are repeatedly simulated and used as inputs to the uncertainty calculations as if they had been obtained from independent measurements of a fixed measurand. The procedure is as follows:

1. A value for $\mathbf{\Gamma} = \Gamma_1 + \mathrm{j}\Gamma_2$ is selected (the measurand is then the fixed value $|\mathbf{\Gamma}|$, or $|\mathbf{\Gamma}|^2$).
2. A sequence of pairs, $(z_1[j], z_2[j])$, $j = 1, \cdots, N$, is drawn from independent Gaussian distributions with means Γ_1 and Γ_2, respectively, and variances u^2.

[1] The ± 1.96 factors are the 2.5% and 97.5% points of the normal distribution.

[2] The ± 3.92 factors are simply $2 \times \pm 1.96$.

3. The GUM and SUP methods are used to obtain 95% uncertainty intervals for the measurand from each pair.

4. An interval containing the measurand is considered a success.

5. The success-rate of each method is reported after a large number of simulations N.

These steps assess the success-rate of a method at one fixed value of the measurand. In order to investigate performance over a range of values, the procedure can be repeated with different values at step 1. This is the approach taken in the next section.

The symmetry of the measurement scenario means that performance will be independent of the radial (azimuthal) coordinate of the measurand in the complex plane. So, without loss of generality, seven $\mathbf{\Gamma}/u$ values lying along the real axis were chosen for this study.

3 Initial Results

The performance results for $|\mathbf{\Gamma}|$ and $|\mathbf{\Gamma}|^2$ are summarised in Table 1. The columns labeled GUM and SUP report the number of successes out of $N = 1000$ for the respective methods.[3] In carrying out the SUP method, a Monte Carlo (MC) sample size of $L = 10^4$ was used. The MC samples generated for $|\mathbf{\Gamma}|$ and $|\mathbf{\Gamma}|^2$ were independent.

Table 1: Successes out of 1000 for $|\mathbf{\Gamma}|$ and $|\mathbf{\Gamma}|^2$

$\mathbf{\Gamma}/u$	$\lvert\mathbf{\Gamma}\rvert$		$\lvert\mathbf{\Gamma}\rvert^2$	
	GUM	SUP	GUM	SUP
0.1	884	0	1000	0
0.2	880	0	999	0
0.5	920	735	999	765
1.0	955	908	984	913
2.0	956	934	943	959
5.0	948	945	942	950
10.0	943	943	948	950

Although we see that the success-rates of both methods are satisfactory when $\mathbf{\Gamma}/u \geq 2$, something goes wrong at lower values of $\mathbf{\Gamma}/u$. For the SUP method, the trend in poor performance appears the same in both measurements. When $\mathbf{\Gamma}/u < 0.5$, the success-rates drop sharply to zero coverage. On the other hand, the GUM method behaves quite differently in the two cases. In measurements of $|\mathbf{\Gamma}|$, GUM success-rates drop below nominal when $\mathbf{\Gamma}/u < 0.5$, while in measurements of $|\mathbf{\Gamma}|^2$, the success-rates are much higher than nominal, almost perfect in fact.

A particularity of the SUP method used here is that none of the MC uncertainty intervals generated will ever include the origin. For example, Figure 1 shows the histogram of

[3] Some variability in the number of successes observed can be expected. For a success-rate of $p = 0.95$, the standard deviation of the number of successes observed is roughly $\sqrt{Np(1-p)}$, which is approximately equal to 7 with $N = 1000$.

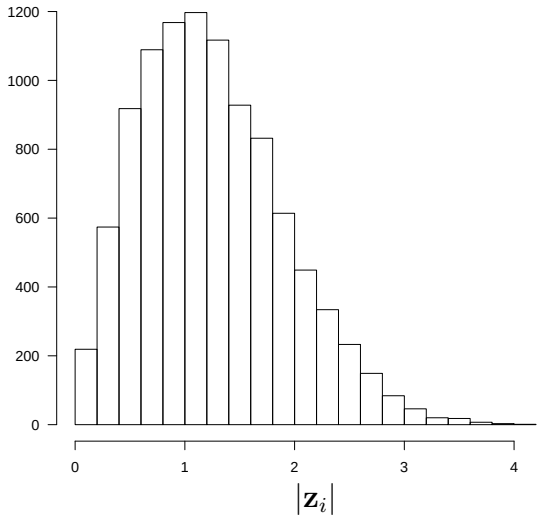

Figure 1: Histogram of a Monte Carlo sample of magnitudes $|\mathbf{z}_i|$ generated for an observed value $\mathbf{z} = 0 + \mathrm{j}0$ and $u = 1$. The lower bound of the uncertainty interval found in this sample is at 0.224.

a MC sample of $|\mathbf{z}_i|$ generated when $\mathbf{z} = 0 + \mathrm{j}0$. For this sample, the SUP method finds a lower bound to the right of the origin at $|\mathbf{z}_i| = 0.224$, because, as described in section 2.2, the smallest 250 values in the MC sample are excluded from the interval. Since no other value of $\mathbf{z}$ can produce a MC sample with a lesser lower bound, we see that the SUP method cannot perform well with small measurands.

The GUM method, on the other hand, can generate uncertainty intervals with negative lower bounds. This is sometimes interpreted as a symptom of method-failure in particular situations. However, a negative lower bound does not present a problem: we may simply reset that bound to zero, which narrows the interval [5, §2.3].

Nonetheless, the GUM method is based on a linear approximation to the measurement function in the region of interest, which is a concern in this problem. We believe that the high success-rates observed in the measurement of $|\mathbf{\Gamma}|^2$ are due to the breakdown of this approximation. Indeed, in both measurement scenarios one might have expected that the SUP method would show superior performance to the GUM method, because the SUP method does not make this linear approximation.

4 Single-sided intervals

It appears that a bias away from the origin in the lower bound is affecting the performance SUP-method when $\mathbf{\Gamma}/u < 0.5$. We therefore consider an adaptive modification to that algorithm (SUP-II) as follows. At the start of each uncertainty calculation we determine $t = |\mathbf{z}|/u$. Then, if $t < \tau$ for some fixed value of τ, we set the lower bound of the uncertainty interval to zero and take the upper bound as the $0.95L^{\mathrm{th}}$ smallest value, instead of carrying out steps 4 and 5 of the SUP method.

This modification is common-sense, but the value to be used for τ is not obvious. If τ is too small, we can expect the performance problems to continue. If τ is too large, the new method will probably generate wide uncertainty intervals making measurement results less accurate. Simulation can be used to look for suitable τ values.

Table 2 reports the number of successes, for different values of τ, in $N = 1000$ simulations of measurements of fixed $|\mathbf{\Gamma}|$.[4] A good choice of this control parameter is seen to be $\tau \approx 2.5$. It is interesting to note that there is a drop in performance in the $\tau = 2.5$ results at around $\mathbf{\Gamma}/u = 1$.

Table 2: Results for $|\mathbf{\Gamma}|$ using SUP-II

$\mathbf{\Gamma}/u$	$\tau = 0$	$\tau = 1$	$\tau = 2$	$\tau = 2.5$	$\tau = 3$
0.1	0	369	874	937	993
0.2	0	377	858	947	991
0.5	760	781	830	945	981
1	907	897	910	919	960
2	958	956	941	951	941
5	937	958	957	952	938
10	956	953	948	959	951

Table 3 shows the number of successes under the same simulation conditions for measurements of $|\mathbf{\Gamma}|^2$. Again a value of $\tau \approx 2.5$ is suitable and we can also see a dip in performance around $\mathbf{\Gamma}/u = 1$.

Table 3: Results for $|\mathbf{\Gamma}|^2$ using SUP-II

$\mathbf{\Gamma}/u$	$\tau = 0$	$\tau = 1$	$\tau = 2$	$\tau = 2.5$	$\tau = 3$
0.1	0	398	865	961	989
0.2	0	397	859	950	989
0.5	759	806	825	936	983
1	911	904	922	886	962
2	932	950	942	951	949
5	951	943	949	964	957
10	960	948	946	955	950

Having improved the SUP method substantially, a similar modification can be made to the GUM method (GUM-II). We again determine $t = |\mathbf{z}|/u$, at the start of each calculation and use it to select between the conventional GUM algorithm and a modified method. If $t < \tau$, we choose the interval $[\,0, |\mathbf{z}| + 1.64u\,]$, instead of the interval (3).[5] The simulation results shown in Table 4 suggest again that $\tau \approx 2.5$ is a suitable value for the control parameter.

We do not report results for the GUM-II method applied to measurements of $|\mathbf{\Gamma}|^2$, because no significant change in performance was observed. Table 1 showed that the

[4] A more detailed analysis of the adaptive method applied to a measurement of $|\mathbf{\Gamma}|$ is given in [6].

[5] The factor 1.64 is the 95% point of the normal distribution.

Table 4: Results for $|\mathbf{\Gamma}|$ using GUM-II

$\mathbf{\Gamma}/u$	$\tau = 0$	$\tau = 1$	$\tau = 2$	$\tau = 2.5$	$\tau = 3$
0.1	874	875	881	937	993
0.2	899	875	889	947	991
0.5	935	928	935	945	981
1	946	939	953	957	960
2	967	960	956	954	943
5	942	956	956	954	952
10	956	953	948	959	951

GUM method is conservative in this problem, with high success-rates that suggest the intervals generated may be too wide. We believe that this is due to a bias introduced by the calculation of $|\mathbf{z}|$, which is used in (4). The non-linear transformation of the simulated coordinates z_1 and z_2 to a magnitude in equation (4) will produce values of $|\mathbf{z}|$ that tend to be greater than $|\mathbf{\Gamma}|$, thereby increasing the average width of the uncertainty intervals generated.

5 Which method is better for $|\Gamma|$ and $|\Gamma|^2$?

For measurements of $|\mathbf{\Gamma}|$, there is not much in the data presented so far to choose between the SUP-II and the GUM-II methods. However, in the case of $|\mathbf{\Gamma}|^2$ our simulations suggest that the GUM and GUM-II methods are conservative, so we might be inclined to choose the SUP-II method as best.

Nevertheless, our objective in selecting a method of uncertainty calculation must be to obtain the most accurate results possible without sacrificing the level of confidence. So, it is also important to look at the widths of the intervals generated. Table 5 reports the mean widths of intervals generated for $N = 1000$ simulations at each measurand, using $\tau = 2.5$. Here we see that the GUM-II method produces narrower intervals for $|\mathbf{\Gamma}|$, so the GUM-II method is superior in this case. We also see that GUM-II produces narrower intervals for measurements of $|\mathbf{\Gamma}|^2$. In that case, we should definitely prefer GUM-II over SUP-II, because the former generates narrower intervals (more accurate results) at a higher success-rate than SUP-II (more likely to capture the measurand in an uncertainty interval).

Table 5: Mean widths for $|\mathbf{\Gamma}|$ and $|\mathbf{\Gamma}|^2$ using adaptive methods

$\mathbf{\Gamma}/u$	$\|\mathbf{\Gamma}\|$		$\|\mathbf{\Gamma}\|^2$	
	GUM-II	SUP-II	GUM-II	SUP-II
0.1	2.9	3.2	6.3	10.5
0.2	2.9	3.2	6.4	10.6
0.5	2.9	3.2	6.9	11.0
1.0	3.1	3.3	8.4	12.3
2.0	3.6	3.6	15.1	17.9
5.0	3.9	3.9	40.6	41.0
10.0	3.9	3.9	78.7	78.8

6 Discussion

Methods of evaluating measurement uncertainty are specified at some level of confidence, or coverage probability, that indicates how they are to perform over many different applications. A method's performance cannot be determined from an application to just one set of data. Nor can the validity of a method be determined by comparing single uncertainty intervals obtained by applying different methods to the same set of data.

It may be possible to verify a method's specified level of confidence by mathematical analysis (e.g., [6]), although in general such a task is likely to be too difficult. On the other hand, the simulation method is relatively simple to apply in specific scenarios.

While simulation does not offer the same potential for insight and generalisation as a mathematical analysis, it does permit an investigation to be carried out under measurement conditions that are of interest to the metrologist. This study has shown the value of such an approach when general-purpose methods of uncertainty calculation are challenged by certain types of measurement. It has also shown that a simulation framework can help to assess *ad hoc* adaptations of a more general method that may better suit particular measurement.

It must be acknowledged that the SUP method used above is only one of several algorithms described in [2]; the single-sided algorithm incorporated in our SUP-II method is another (see also [7]). Some metrologists may be inclined to make an 'expert' decision, based on the data they obtain, choosing one or other of these algorithms to evaluate an uncertainty statement. In so doing they are following a procedure analogous to what we have called the SUP-II method, albeit without necessarily fixing a value for the control parameter τ. This study found that the performance of SUP-II is quite sensitive to τ, which suggests that such informal practices might benefit from a similar analysis.

7 Conclusion

In this study, two standard methods of uncertainty calculation failed to achieve their nominal coverage levels when calculating uncertainty in the magnitude, and the magnitude squared, of a device with a small complex reflection coefficient. The problem was only identified after looking at the long run success rates of each method when applied to many simulated data sets. It was then possible to modify the methods and achieve improved performance in most cases.

The use of simulated data in this way is proposed as a general-purpose tool for the verification of uncertainty calculation methods.

Acknowledgement

This work was funded by the New Zealand Government as part of a contract for the provision of national measurement standards.

References

[1] BIPM, IEC, IFCC, ISO, IUPAC, IUPAP, and OIML, *Evaluation of measurement data Guide to the expression of uncertainty in measurement JCGM 100:2008 (GUM 1995 with minor corrections)*, 1st ed. Paris, Sèvres: BIPM Joint Committee for Guides in Metrology, 2008.

[2] ——, *Evaluation of measurement data Supplement 1 to the "Guide to the expression of uncertainty in measurement" Propagation of distributions using a Monte Carlo method JCGM 101:2008*, 1st ed. Paris, Sèvres: BIPM Joint Committee for Guides in Metrology, 2008.

[3] C. M. Wang and H. K. Iyer, "Propagation of uncertainties in measurements using generalized inference," vol. 42, pp. 145–153, 2005.

[4] B. D. Hall, "Evaluating methods of calculating measurement uncertainty," *Metrologia*, vol. 45, pp. L5–L8, 2008.

[5] R. Willink, "Principles of probability and statistics for metrology," *Metrologia*, vol. 43, pp. S211–S219, 2006.

[6] ——, "On the validity of methods of uncertainty calculation," *Metrologia*, submitted.

[7] ——, "Uncertainty in repeated measurement of a small non-negative quantity: explanation and discussion of Bayesian methodology," *Accreditation and Qualtity Assurance*, 2009, at press, DOI 10.1007/s00769-009-0595-7.