

Heuristic confidence intervals for the Gamma test

by

Antonia J. Jones¹ and Samuel E. Kemp²

¹School of Computer Science
Cardiff University
PO Box 916, Cardiff CF24 3XF, UK

²Faculty of Advanced Technology
University of Glamorgan
Pontypridd, Wales, CF37 1DL, UK

First draft: 18 February 2006

Accepted by:
*The 2006 International Conference on Artificial Intelligence (ICAI'06:
June 26-29, 2006, Las Vegas, USA.*

Abstract

The Gamma test is a fast non-parametric algorithm for estimating the noise variance in an input/output dataset modulo the best smooth model. In this short note we communicate a heuristic method for computing confidence intervals for the noise estimate returned by a Gamma test analysis.

Keywords: Gamma test, noise estimates, smooth modelling, non-linear modelling, neural networks, confidence intervals.

1 Introduction

The Gamma test [Stefánsson et al., 1997, Končar, 1997] is a fast non-parametric algorithm for estimating the second moment of the noise distribution, σ^2 , in an input/output dataset modulo the best *smooth* model, even though this model is unknown. In general, we can assume the output y is determined by

$$y = f(\mathbf{x}) + r \quad (1)$$

where $y \in \mathbb{R}$ is the output, $\mathbf{x} \in \mathbb{R}^m$ is a vector of m inputs, f is a smooth unknown function, and r is a random variable with mean zero and a variance σ^2 representing the noise. In this context the noise is any component of the output that cannot be accounted for by a smooth transformation of the corresponding inputs. The causes of noise in a set of observations $\{(\mathbf{x}_i, y_i) | 1 \leq i \leq M\}$ may be attributed to one or more of the following.

- Measurement error
- Not all the relevant factors that influence the output are included in the inputs.
- The underlying relationship between input and output is not smooth.

The Gamma test estimate for σ^2 is called the Gamma statistic (denoted by Γ) and, despite the fact that f is unknown, can be computed in $O(M \log M)$ expected time using a *kd-tree*¹. The applications of having such an efficient technique for estimating the noise variance are wide ranging, particularly in the field of non-linear modelling and neural networks. For example, in neural networks the Gamma statistic can be used as a metric for deciding when to cease training (*i.e.* in order to prevent overfitting we cease when the mean squared error reaches Γ), as well as providing a criterion for input variable selection. For a comprehensive overview of the Gamma test and its application to non-linear modelling and prediction the reader may wish to consult [Jones, 2004].

This note describes a heuristic method for estimating confidence intervals for a Gamma statistic. Before doing so, we empirically examine the distribution of Gamma values, evaluated for the same input/output function and noise distribution, where M is fixed, and then examine how these distributions change as M is varied.

2 The Gamma distribution for fixed M

We imagine that for a fixed process² and fixed M we are able to construct many input/output data sets of size M . For each data set we compute a Γ statistic. Now we

¹Fast open-source implementations of the Gamma test in C, Mathematica and R can be download from <http://users.cs.cf.ac.uk/Antonia.J.Jones/GammaArchive/IndexPage.htm>.

²*i.e.* a fixed non-linear function f and a fixed noise distribution.

ask: *What is the distribution of Γ over the set of all possible data sets of size M ?* Although this distribution is theoretically somewhat inaccessible we can easily construct an experiment to get an approximate histogram for artificial data.

Consider the process $y = \sin(x) + r$, where x is a random variable taken over the range $[0, 2\pi]$, r is a Gaussian random variable with mean zero and $\sigma^2 = 0.075$, and $M = 1000$. We generate 1000 different datasets of this fixed process and compute a Gamma statistic for each. A plot of the Γ histogram for $M = 1000$ is shown in Figure 2, as a comparison the histogram for $M = 500$ is shown in Figure 1. For $M = 1000$, the mean Gamma statistic over the 1000 samples is 0.07493139 with a standard deviation of 0.003740105. From Figures 1 and 2 it seems reasonable to assume that the Γ distribution is tending to normality as M increases. Based on this assumption we can apply the Student's t -test to give a confidence interval of (0.07473666, 0.07512611) at the 90% level. The interval becomes larger as the confidence level increases e.g. at the 99% level the interval is (0.07462616, 0.07523662).

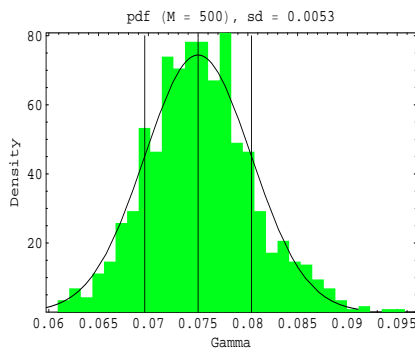


Figure 1: The Γ histogram of 1000 samples for $M = 500$ with an overlay plot of the corresponding normal distribution. The mean 0.075 and \pm one standard deviation are shown.

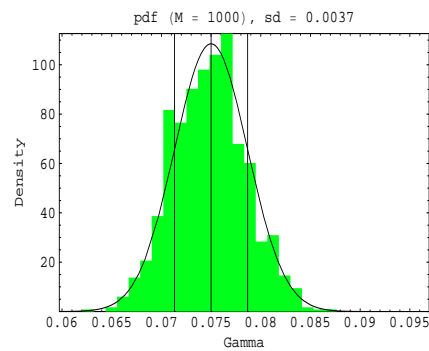


Figure 2: The Γ histogram of 1000 samples for $M = 1000$ with an overlay plot of the corresponding normal distribution. The mean 0.075 and \pm one standard deviation are shown.

By the Central Limit theorem we might expect that the standard deviation, SD, of the Γ distribution for fixed M scales like

$$\text{SD} = c/\sqrt{M} \quad \text{i.e.} \quad \log(\text{SD}) = \log(c) - \frac{1}{2} \log(M) \quad (2)$$

for some $c > 0$, as M becomes large. To check whether SD does scale according to (2) we used the same fixed process as before, but this time we vary M i.e. $M = 500$ to $M = 1000$ in steps of $\Delta M = 50$. Figure 3 shows the plot of $\log(\text{SD})$ against $\log(M)$ with the regression line calculated as

$$\log(\text{SD}) = -2.0605 - 0.5091 \log(M) \quad (3)$$

The slope of (3) is close to the theoretical value, $-\frac{1}{2}$, from (2). By taking the exponential of the intercept in (3) we find that in this particular case the SD scales like $0.1274/\sqrt{M}$.

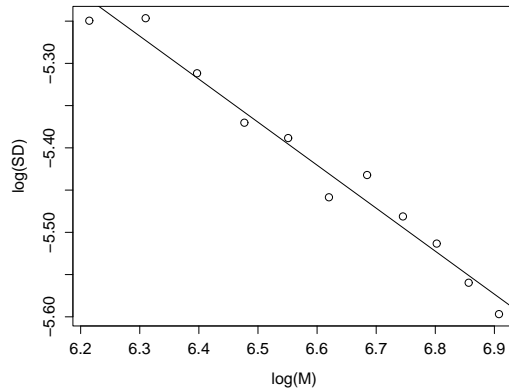


Figure 3: The $\log(\text{SD})$ plotted against $\log(M)$ with the regression line fit $\log(\text{SD}) = -2.0605 - 0.5091 \log(M)$.

3 Computing confidence intervals

Given L samples from a fixed, approximately normal, distribution the standard small sample Student t -test used to generate confidence intervals can be considered as a procedure `StudentTCI[mean, se, dof]`, where `mean` and `se` are the sample mean and standard error. The standard error is computed from the sample standard deviation `sd` as `sd/√L`. The number of degrees of freedom `dof` in this context is the number of data samples minus one, i.e. $L - 1$. The procedure should take an option which specifies the confidence level and return the left and right endpoints of the confidence interval. Thus

```
StudentTCI[mean, se, dof, ConfidenceLevel → 0.9]
```

would return the 90% confidence interval.

For example on the data sets for $M = 1000$ generated in the previous section this procedure returns $(0.07473666, 0.07512611)$ at the 90% confidence level. This is a precise estimate for the confidence interval, but it is based on 1000 Gamma tests on independent data sets of size $M = 1000$, i.e. we needed 10^6 input/output data points to arrive at this figure. With this amount of data we could almost certainly derive an extremely accurate estimate for the Gamma statistic. Plainly we need a more practical method for computing confidence intervals.

Our heuristic method of estimating confidence intervals, derives from the M -test (*i.e.* computing a Γ over an increasing number of data points) and is based on some assumptions. We shall assume that the Gamma distribution for fixed M is approximately normal. This is manifestly not the case at the tails of the distribution but, as illustrated in the previous sub-section, may serve for the purposes of the present section. We saw that the standard deviation of the Gamma distribution scales approximately like c/\sqrt{M} , for some $c > 0$, as M becomes large. We can exploit this observation to construct a heuristic algorithm to compute confidence intervals based on an M -test.

Student's t -test is based on the idea of taking independent samples from a *fixed* distribution. In our case the Gamma distribution varies as M increases in the M -test; the mean remains fixed at the variance σ^2 of the noise, but `sd` scales like c/\sqrt{M} as M increases.

In order to use `StudentTCI` to compute confidence intervals for an M -test we need to have estimates for the mean and standard deviation of the Gamma distribution derived from a limited L number of sample Gamma test calculations in which M varies. Suppose for increasing $M = M_1, M_2, \dots, M_L$ we have computed (M_i, Γ_i) ($1 \leq i \leq L$).

The problem we have is that our samples $\Gamma_1, \dots, \Gamma_L$ are each computed on data sets of *different* sizes, and so are actually drawn from *different* Gamma distributions. These distributions have the same underlying mean (the true noise variance σ^2), but *different* standard deviations, which we now *assume* are of the form $c/\sqrt{M_i}$ ($1 \leq i \leq L$) for some unknown but fixed³ $c > 0$.

To correct for this we linearly ‘normalise’ each sample value Γ_i so that it can be considered to come from *the same* distribution. We do this by multiplying by $\sqrt{M_i}$ for calculating the *normalised* mean and standard deviation, and then *re-normalising* to the distribution for M_L by multiplying by $M_L^{-1/2}$. In this way we endeavour to ensure that we are averaging comparable quantities. Thus we compute

$$\begin{aligned} \text{mean} &= \frac{1}{M_L^{1/2} L} \sum_{i=1}^L \sqrt{M_i} \Gamma_i \\ \text{sd}^2 &= \frac{1}{M_L(L-1)} \sum_{i=1}^L (\sqrt{M_i} \Gamma_i - \sqrt{M_i} \text{mean})^2 \\ \text{se} &= \frac{\text{sd}}{\sqrt{L}} \\ \text{CI} &= \text{StudentTCI}[\Gamma_L, \text{se}, L-1, \text{ConfidenceLevel} \rightarrow 0.9] \end{aligned}$$

which returns the confidence interval `CI` associated with the L^{th} Gamma statistic Γ_L in an M -test which computes Γ_i for $M = M_1, M_2, \dots, M_L$. Note that Γ_L is given as the `mean` in `StudentTCI` because this is the best estimate we have of the distribution noise variance σ^2 .

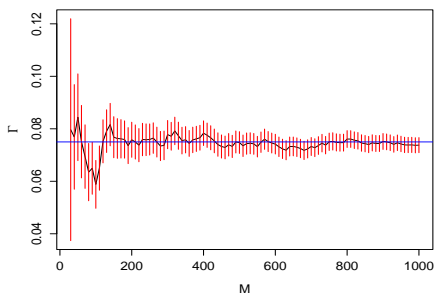


Figure 4: An M -test with the heuristic confidence intervals at the 90% level shown as red bars for $f(x) = \sin(x) + r$ with $\sigma^2 = 0.075$. The true noise variance (blue line) is almost always inside the confidence interval.

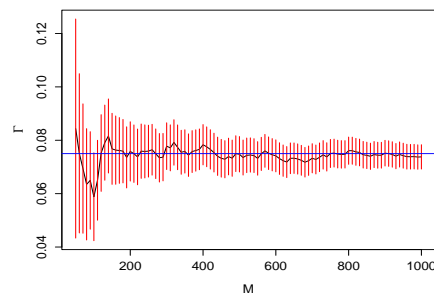


Figure 5: An M -test with the heuristic confidence intervals at the 99% level shown as red bars for $f(x) = \sin(x) + r$ with $\sigma^2 = 0.075$. The true noise variance (blue line) is *always* inside the confidence interval.

³Of course, the value of c might be expected to be problem dependent.

Figure 4 shows the heuristic confidence intervals at the 90% level for $M_L = 20$ to $M_L = 1000$ in steps of 10. The confidence interval returned by the heuristic for $M_L = 1000$ is (0.07089257, 0.07665589), which compares with our earlier accurate confidence interval of (0.07473666, 0.07512611). The latter was computed using 10^6 data samples and 1000 Gamma tests, whereas the former, our heuristic, used $M_L = 1000$ data samples and $L = 99$ Gamma tests.

For comparison, Figure 5 shows the intervals at the 99% level for $M_L = 20$ to $M_L = 1000$ in steps of 10. The confidence interval returned by the heuristic for $M_L = 1000$ is (0.06921554, 0.07833291) compared with (0.07462616, 0.07523662) for the accurate intervals.

As we should expect from a heuristic confidence interval, our method is a *conservative estimate* because it contains as a subset the more accurate interval.

4 Conclusions

In this short note we have illustrated, using artificial data, that for fixed M the Γ distribution can be considered approximately normal, except of course at the tails. Using data generated by the same process, but for varied M , we saw the standard deviation of the Γ distribution scaled as c/\sqrt{M} , as M becomes large. This approximate normality allows us to use the Student's t -test to construct accurate Γ confidence intervals for fixed M .

The successive Gamma statistics computed in an M -test are not independent, because the data used to compute Γ_i is a subset of the data used to compute Γ_{i+1} . However, it seems to be the case (for a reasonable step size) that this dependence is sufficiently weak that, after linear normalisation, the distribution of L such samples behaves as the Central Limit theorem might suggest and that the standard error scales as sd/\sqrt{L} .

Thus, as we have illustrated, by normalising and then re-normalising, the L different Γ values from an M -test can be treated as if they were L independent samples from the *same* approximately normal distribution, so that Student's t -test can be applied to return confidence intervals. This makes very efficient use of the available data and allows us to enhance the standard M -test by adding heuristic confidence intervals at very little extra computational cost.

Comparisons between the accurate confidence intervals, computed on the basis of data sets of size 10^6 , and the heuristic confidence intervals, computed using 10^3 data points, demonstrated that the heuristic method presented is a close *conservative* approximation to the accurate confidence intervals for fixed M .

References

- [Jones, 2004] Jones, A. J. (2004). New tools in non-linear modelling and prediction. *Computational Management Science*, 1(2):109–149. ISSN 1619-697.
- [Končar, 1997] Končar, N. (1997). *Optimisation methodologies for direct inverse neuro-control*. PhD thesis, Department of Computing, Imperial College of Science, Technology and Medicine, University of London.
- [Stefánsson et al., 1997] Stefánsson, A., Končar, N., and Jones, A. J. (1997). A note on the gamma test. *Neural Computing Applications*, 5:131–133.