LIGO SCIENTIFIC COLLABORATION
VIRGO COLLABORATION

Document Type	**LIGO-T1500597-v1**	2015/12/04
Quoting parameter-estimation results		
Christopher P. L. Berry *et al*. for the PE Group		

WWW: http://www.ligo.org/ and http://www.virgo.infn.it

Processed with LaTeX on 2015/12/04

Abstract

Measuring the properties of a gravitational-wave signal is a question of parameter estimation. The end result of these studies is a set of samples drawn from the posterior probability distribution. The posterior contains all the information that we have, but may not be easily digestible; in many cases, it is desirable to quote summary statistics to concisely describe findings. In most practical cases, it is impossible to compress a complete description of a distribution down to a single number; therefore, any point estimate could miss key pieces of information. Here, we discuss various possibilities for summary statistics. We also include a discussion of how to quote systematic errors on parameter estimates. While there is no perfect answer, our suggestion is to use X^{+Y}_{-Z}, where X is the median Y and Z are estimates for the statistical error (measurement precision) from the bounds of a symmetric credible interval, and then add in estimates for systematic error from the range of X, Y and Z, which could be presented as $X^{+Y\pm y}_{-Z\pm z}$ or $(X \pm x)^{+Y\pm y}_{-Z\pm z}$.

1 Introduction

When making inferences using data, we encode our belief about the value of parameters into probability density functions. Using the information encoded in our data $\boldsymbol{s}$ (and assuming a model that predicts what the data should be given a set of parameters), the best understanding of possible parameter values $\boldsymbol{\theta}$ is given by the posterior distribution

$$P(\boldsymbol{\theta}|\boldsymbol{s}) = \frac{P(\boldsymbol{s}|\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(\boldsymbol{s})}. \tag{1}$$

Here, we have used Bayes' theorem to relate the posterior $P(\boldsymbol{\theta}|\boldsymbol{s})$ to the likelihood $P(\boldsymbol{s}|\boldsymbol{\theta})$; the prior $P(\boldsymbol{\theta})$, which gives our best understanding of the parameters before we considered this data, and the evidence $P(\boldsymbol{s})$.

To map out the posterior probability distribution, we employ stochastic sampling algorithms that return random samples drawn for the posterior distribution. From these posterior samples we can reconstruct the shape of the posterior, calculate expectation values, find credible intervals, etc. The posterior samples can be regarded as the end product of the inference and the best representation of what we believe to be the parameter values. However, considering an entire distribution is not easy. We will discuss means of summarising key aspects of the our parameter estimates; since the posterior distributions can be intricate, it is difficult to preserve an accurate description in a few numbers. All approaches have advantages and disadvantages, and we seek a balance between simple-to-interpret and faithful descriptions.

2 Best estimates

Of primary importance is a single point estimate X that gives the best estimate for a parameter value. Here, best would depend upon the particular application, which will vary between readers; hence, we look for generic quantities that may be useful, if not perfect, in a wide range of circumstances.

2.1 Maximum likelihood

The maximum likelihood (ML) values gives the peak of likelihood. This is the point where the data best fits the model, and the point of maximum signal-to-noise ratio.

This point is independent of our choice of priors, so it may be argued that this is useful for readers who wish to use their own. However, should a reader want to try this in practise, they would need the entire distribution and not a point estimate. The likelihood is not a probability distribution for the parameters, and excludes potentially important information from the priors (for example that the spin magnitude is less than one or that sources are expected to be uniformly distributed in volume), so we do not consider it further.

The ML estimate also shares a number of disadvantages with the maximum posterior estimate (section 2.2).

2.2 Maximum posterior

The maximum posterior (maximum a postori; MAP) value is the peak of the posterior probability distribution, it is the modal value. It is attractive as it gives the most probable point.

There can be some ambiguity in its definition, specifically in which dimension we find the maximum. The obvious candidates are the single global maximum in the full parameter space, or the maxima of each one-dimensional distribution. For any given parameter, the two need not coincide. If a reader were only interested in one parameter, for example mass, and not interested in another, say polarization, they should consider the maximum of the one-dimensional distribution. However, if they were considering multiple parameters, say both masses, then combining the one-dimensional MAP estimates could land in a point of the two-dimensional parameter space that is highly improbable. The correct thing to do in this case would be to give the maximum in the two-dimensional space, but giving maxima for all possible combinations would be extremely unwieldy.

The maximum is not invariant under reparametrization. When converting parameters, factors from the Jacobian can move the position of the peak. For example, the peak for the two component masses does not need to correspond to the peak for the chirp mass and the mass ratio, or the peak in the luminosity distance does not need to correspond to the peak for the reciprocal of the luminosity distance.

The MAP value also does not need to represent a typical value: it may be the the peak lies well outside the region that contains most of the posterior mass. Furthermore, it is of little use for multimodal distributions. Therefore, the MAP value may not give a useful summary of the distribution.

2.3 Posterior mean

The posterior mean is the expectation value of the distribution. For a Gaussian distribution, this would also be the MAP value. The mean better traces the position of the posterior mass than the MAP. However, the mean position does not need to be a probable location; for example, with a bimodal distribution the mean may lie between the two peaks, in a region that is improbable.

Like the MAP value (section 2.2), the posterior mean is not invariant under reparametrization. The posterior means are also constructed using the marignalised one-dimensional distributions. It does not makes sense to combine means for different parameters (which is in effect a reparametrization): averages over the full distribution can always be calculated using the posterior samples.

2.4 Posterior median

The posterior median is the position of the 50% quantile. This gives a good indication of the centre of the distribution and the position of the posterior probability mass. It is less influenced by the tails of the distribution than the posterior mean (section 2.3). Like the posterior mean, the posterior median does not need to coincide with a probable posterior value.

The posterior median is invariant under reparametrization, provided that this is a monotonic mapping. For example, if given the median luminosity distance D, it is possible to find the median $1/D$, median luminosity D^2, median $\ln D$, etc. However, as with the posterior mean, it does not makes sense to combine medians for different parameters.

2.5 Illustration

As a simple illustration, Figure 1 shows a two-dimension distribution for parameters θ_1 and θ_2, along with the marginalised one-dimensional distributions. The MAP in the two-dimensional space would be in the peak at $(\theta_1 = 2, \theta_2 = -12)$, but this is clearly not a typical value: only 4% of the total posterior probability mass is in this peak. This effect is greater is higher dimensions. The one-dimensional MAPs occur at $\theta_1 = 15$ and $\theta_2 = 10$. In this case, combining them would give a sensible point, but this does not have to happen. The peak at $\theta_1 = 10$ is only slightly less probable that the MAP, so ignoring it neglects an important part of the distribution. The median traces the centre of the distribution, but lies in the valley between peaks for θ_1. The mean values are $\theta_1 \simeq 12$, which is again in the valley, and $\theta_2 \simeq 9$; these values are offset from the median because of the pull of the outlying peak.

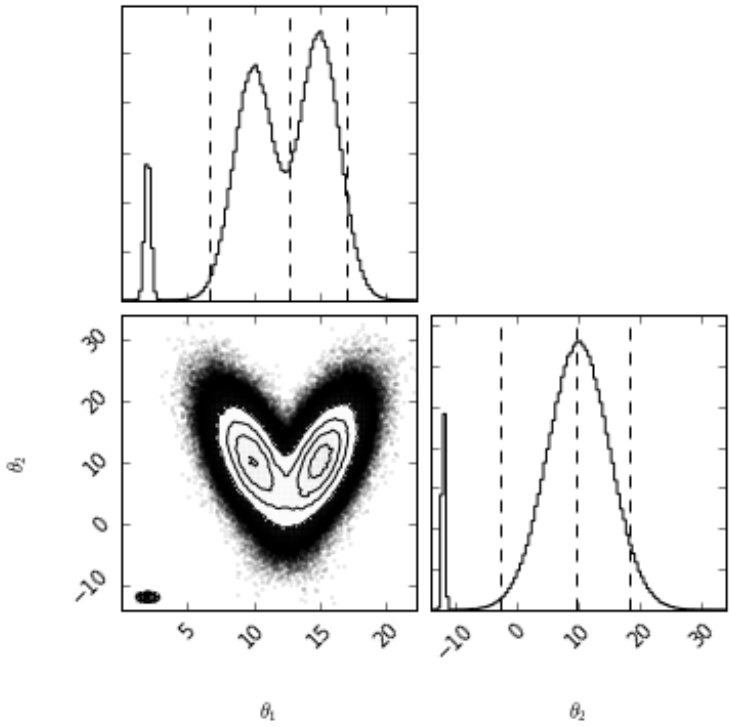

Figure 1: Probability distributions recreated from samples for parameters θ_1 and θ_2. The dotted lines for the one-dimensional distributions indicate the postions of the 5%, 50% and 95% quantiles and hence indicate the symmetric 90% credible interval (section 3.2.1) and the median (section 2.4).

3 Statistical error

Together with any point estimate, there must be an indication of uncertainty. We wish to quote this as X^{+Y}_{-Z}. Statistical uncertainty gives an indication of the width of the distribution.

3.1 Standard deviation

The standard deviation is commonly used to quantify the width of a distribution. It is determined by the second moment of the distribution, with the first moment giving the mean (section 2.3).

For Gaussian distributions it has a simple interpretation in terms of the enclosed probability, but this does not translate to other distributions. Furthermore, the standard deviation (as a single number) is of little use for skewed or multimodal distributions. Since we will not be typically considering Gaussian distributions, we will not consider the standard deviation further.

3.2 Credible interval

A credible interval describes an interval (in one-dimension or appropriate volume in higher dimensions) that encloses a given total posterior probability. For example, the 90% credible interval covers a total posterior probability of 0.9 such that we believe there is a 90% chance that the true value is inside the interval and a 10% chance that it is outside.

There are multiple ways of constructing credible intervals, which are detailed below. When using a credible interval, there is always the choice of which probability to use. The 50% credible interval is simple to interpret, but may not be considered broad enough. The $68.269\ldots\%$ credible interval gives the equivalent of a Gaussian one-sigma interval; it is familiar but not necessarily particularly meaningful for non-Gaussian distributions: quoting it may encourage improper use as if the distribution were Gaussian.

The 90% credible interval is simple to interpret and includes most of the potential range. The 95% credible interval would approximately coincide with the two-sigma interval for a Gaussian distribution, but may start to suffer from the tails of distributions being difficult to accurately estimate.

3.2.1 Symmetric credible intervals

Symmetric credible intervals are centred on the median and extend outwards such that there is an equal probability in each tail of the distribution. For example, the 90% credible interval has its lower bound at the 5% quantile and its upper bound on the 95% quantile, so that it contains the central 90% of the posterior probability. The 50% symmetric credible interval gives the interquartile range.

Symmetric credible regions naturally complement the posterior median (section 2.4). Together they give the position of the 50% quantile with the range to the the desired maximum and minimum quantiles. This marks the central region of the posterior distribution and gives a good indication of the range of parameter values that are consistent with our distribution; however, it can exclude highly probable values if these occur at the edges of parameter space, as can be the case for mass ratio.

3.2.2 One-sided credible regions

One-sided credible regions start from one edge of parameter space and continue until they contain the desired probability, such that there is a single tail outside the interval. For example, the 90% credible interval could extend from the minimum value to the 90% quantile or from the maximum value to the 10% quantile.

One-sided credible regions are only applicable for parameters that have a definite bound, for example the mass ratio, the spin magnitude, or component masses if starting from zero. They cannot be used for all parameters. They may be most useful when considering distributions that are peaked towards one bound, for example spin magnitudes may be peaked around zero. However, the same information could be simply presented by quoting an upper or lower credible bound.

3.2.3 Greedy credible intervals

Credible intervals can also be constructed by starting at the most probable position (the MAP value, section 2.2) and then working outwards including the next most probable positions. This can be done by starting at the MAP and adding the next adjacent element to the current interval, or by adding the next highest probability element. The latter results in the smallest possible credible interval, but it is not necessarily contiguous. Hence, it may be useful if considering how much of the parameter space should be searched to reach the desired total probability (as for sky location), but is less useful as a simple error bound of the form X^{+Y}_{-Z}. The former does give a continuous range, and gives a sensible error range about the MAP. However, for multimodal distributions (particularly if secondary modes contain greater total probability mass than the MAP peak) they can still exclude regions of high probability like other choices of credible intervals.

4 Systematic error

Assessing systematic error is notoriously difficult. To be able to do this properly we must know a correct value, and we do not have this luxury. The best option we have is a comparison between results assuming different approximants (potentially including numerical relativity in the future). Given posterior distributions calculated assuming several different waveform models, we must consider how to combine these to provide an estimate for the systematic error and hence to total uncertainty.

4.1 Combining ranges

A conservative option is to quote the maximum and minimum values of all possible statistical uncertainties as an overall error range. The problem with this is that it does not yield a range of parameter values that has an simple interpretation. It does not correspond to a credible interval and has no statistical meaing. It

therefore cannot be used for further analysis, other than to highlight regions that are thought implausible (although we cannot specify how implausible).

4.2 Averaging posteriors

Given several models we can marginalise over our uncertainty by averaging the posterior distributions. Each approximant can be assigned a probability to reflect our belief that it is correct, and using these we combine all the samples into a single distribution. The probability for each model could include the evidence for that model, to quantify how well the data are predicted, and a prior for our degree of belief that it is correct; for example, this could favour more accurate computational methods or weight models that include spin are applicable over a wider range of parameters more than those that do not. The simplest option would be to remain completely agnostic and give each approximant equal weighting.

Having computed the averaged posterior, we can quote a point estimate and uncertainty as usual X^{+Y}_{-Z}. These will then naturally fold in potential systematic errors: we do not quote an additional error, just give results having marginalised over uncertainty with regards to the model.

In the event that the models considered have different numbers of parameters (for example one with spin and one without), it is not clear that averaging gives a useful result. We can still consider the marginalised distributions for the common parameters, but not for those which are only free in a sub-set of models. In model averaging, these extra quantities become nuisance parameters that must be marginalised out. This restricts the final answer to the subspace of the common parameters.

This approach is straight-forward and gives robust results as it best summarises the knowledge that we have. However, it only allows us to consider the subset of models that we have considered: it does not construct an estimate for the typical difference between models. This is usually what is used to estimate the difference between our models and the true results. Furthermore, in the event that we get distinct modes in parameter values corresponding to different models, we might expect that the true value could lie somewhere between these peaks, even though our averaged posterior assigns little probability to this region.

4.3 Comparing posterior estimates

We can consider the problem of estimating systematic error as a question of parameter estimation given these different results. Within our assumption that the variation between approximants gives us an impression of the potential systematic error, we can try to infer the properties that describe the scatter in different estimates.

If we start with what we consider is our best estimate (this could be either the approximant that includes the most complete picture of the physics or the approximant-averaged posterior), then potential systematic error should have zero mean. This is because the systematic error is equally probable to be positive of negative; if we had evidence otherwise we could subtract this error to make a new best estimate, and then the systematic error would have zero mean. With the mean in hand, we are left trying to determine the shape of the systematic error. We know the scatter between estimates using different approximants, and for a given set of points with known scatter the maximum-entropy (least constraining) distribution is a Gaussian. Therefore, it is appropriate (given our state of ignorance) to use the root-mean-square scatter (relative to our best estimate) to estimate the standard deviation of this Gaussian. This can then be used as an approximate model for systematic error.

As an example, we could consider the scatter in the point estimate and the size of the credible intervals and quote $(X \pm x)^{+Y\pm y}_{-Z\pm z}$. Here, X is our best estimate and x is the estimated systematic error on this; Y is the best estimate for the range to the upper statistical-error bound and y is the estimated systematic error on this, etc. This clearly separates our systematic and statistical error, but is potentially confusing. For example, it must be made clear if the systematic error on the upper bound is estimated from variation in the size of the range Y or from the position of the bound $X + Y$. This could potentially be disambiguated by using $(X \pm x)^{+Y\pm y}_{-Z\pm z}$ for the former and $X^{+Y\pm y}_{-Z\pm z}$ for the latter. The second option has the interpretation that X is the single best estimate, and that the quoted ranges give the amount of variation about this.

As a concrete illustration, let us consider two cases. First, consider the case that two approximants give the same shape distribution, but simply offset by a constant a. This could be quoted as $(X \pm a)^{+Y\pm 0}_{-Z\pm 0}$

or as $X^{+Y\pm a}_{-Z\pm a}$ using the two interpretations outlined above. Second, let us consider the case that two approximants give the same central estimate, but one is wider by a factor of 2. This could be quoted as $(X \pm 0)^{+Y\pm Y/2}_{-Z\pm Z/2}$ or $X^{+Y\pm Y/2}_{-Z\pm Z/2}$. In both of these examples, we have just used the size of the scatter, which would be one-sigma estimate for the systematic error, when in practise we may wave to use a different width (say that which corresponds to 90% probability), but this amounts to a simple scaling of the values of x, y and z.

If one or more parameters is only included in a single model (for example if we only have one model with eccentricity), then we cannot bound systematic error from this approach.

5 Conclusion

Describing a multidimensional distribution in a few numbers clearly and in a useful manner is difficult. There is no way to cater for all readers, which is why it is vital to supply the full set of posterior samples. Then readers who need particular numbers can calculate them for themselves.

In order to supply simple numbers for a publication, each point estimate has its advantages and disadvantages.

- The posterior mean is simple to understand, but for generic distributions there is no good way to quote an accompanying statistical uncertainty, making it unappealing.
- The full-dimensional MAP point gives our single best estimate. This is perhaps the best result to give if the reader wants a consistent set of all parameters. However, this location is not invariant under a change of parameters, meaning that we should not place special significant to it. Furthermore, there is no good way to quote uncertainties for individual parameters.
- One-dimensional MAP points can be bounded by greedy credible intervals to give the most probable range of parameters. These are not invariant under a change of variables. However, if the most commonly used parameter values are quoted, the values may still be useful in most cases.
- The median with a symmetric credible interval gives a simple description of the range of parameters. This neatly describes the distribution, but does not have to give a probable value for the single point estimate.

The simplicity of the median and symmetric credible interval make it particularly appealing; it is sufficiently transparent that its shortcomings are easy to appreciate. It may not give the best single estimate, but it does reflect the idea that one should consider the entire distribution when using these results for further investigations.

Systematic error is difficult to quantify and communicate efficiently. Providing estimates (and posterior samples) from multiple approximants is an open and comprehensive means of showing how much results change. However, listing multiple numbers make it confusing for the reader who is looking for the best answer. It is necessary to provide a single result that best encodes what we believe about the source.

Marginalising over the approximant is one means of accounting for our uncertainty. However, this is implicitly assuming that the true answer is included amongst the set of approximants (or that the approximants make a dense covering of model space such that the true model is encompassed within their range).

The alternative, given our set of different results, is an attempt to boot-strap an estimate for the systematic error from the difference between approximants. This may give a better divide between estimated statistical and systematic error, but the resulting notation can be confusing. If notation like $(X \pm x)^{+Y\pm y}_{-Z\pm z}$ or $X^{+Y\pm y}_{-Z\pm z}$ is used, meaning needs to be carefully explained. However, the idea of splitting statistical and systematic error is common in many areas (such as particle physics), and allows readers to judge how accuracy could potentially be improved either by an increase in signal-to-noise ratio (decreasing statistical uncertainty) or by improved waveforms (decreasing systematic error).

Marginalising over the approximant restricts the final posterior to the sub-space of common parameters. This causes a problem if we wish to make statements about these parameters; however, it is not an issue if we cannot measure these extra parameters (they can already be considered nuisance parameters). Following the other approach, we cannot give an estimate for the systematic error for parameters which

are only included in a single model, but we can still quote a best estimate, assuming we adopt the highest-dimensional model for this purpose. If there is no clear reason to prefer one model over the others, picking it as the best model would not e a best representation of our understanding. A potential solution is to average the models to produce a best estimate, and then look at the scatter to estimate the systematic error. In this approach we are adopting the philosophy that we have made a measurement (say of the position of the 5% quantile) by several different methods which give different results: we construct an average as a best guess and use the scatter of possible results to produce an estimate of the uncertainty.

If posterior estimates are combined to boot-strap the systematic error, then both the credible interval and the systematic error must be quoted at the same significance, for example both should correspond to $\sim 68\%$ or 90%. The one-sigma $\sim 68\%$ errors may be most traditional, and what most readers from other fields would expect; this would also make the explanation of the systematic errors simplest, as it is just the root-mean-square of difference between approximants. However, using 90% or 95% uncertainties would give a fuller indication of the size of distributions, and naturally would include a greater probability of enclosing the true results. Avoiding the equivalent of the one-sigma probability would also discourage the mistaken idea that uncertainties can be treated as for Gaussian distributions.

Previous parameter-estimation studies have used the symmetric 90% credible interval, so for comparison it may be desirable to continue using this.