Bayesian model comparison applied to the Explorer-Nautilus 2001 coincidence data

P. Astone,[1] **G. D'Agostini,**[1] **S. D'Antonio**[2]
1) INFN and University of Rome "La Sapienza", Rome, Italy
2) INFN and University of Rome "Tor Vergata", Rome, Italy
(Presented by P. Astone)

Abstract

Bayesian reasoning is applied to the data by the ROG Collaboration, in which gravitational wave (g.w.) signals are searched for in a coincidence experiment between Explorer and Nautilus. The use of Bayesian reasoning allows, under well defined hypotheses, even tiny pieces of evidence in favor of each model to be extracted from the data. The combination of the data of several experiments can therefore be performed in an optimal and efficient way. Some models for Galactic sources are considered and, within each model, the experimental result is summarized with the likelihood rescaled to the insensitivity limit value ("$\mathcal{R}$ function"). The model comparison result is given in in terms of Bayes factors, which quantify how the ratio of beliefs about two alternative models are modified by the experimental observation.

1 Introduction

A recent analysis of data from the resonant g.w. detectors Explorer and Nautilus [1] has shown some hints of a possible signal over the background expected from random coincidences. The indication appears only when the data are analyzed as a function of the sidereal time. Reference [1] does not contain statements concerning the probability that some of the observed coincidences could be due to g.w.'s rather than background. Only bottom plots of Fig. 5 and Fig. 7 of that paper gives p-values (the meaning of 'p-value', to which physicists are not accustomed, will be clarified later) for each bin in sidereal time, given the average observed background at that bin. But p-values are not probabilities that the 'only background' hypothesis is true, though they are often erroneously taken as such, leading to unpleasant consequences in the interpretation of the data [2]. Indeed, in this case too, Fig. 5 and Fig. 7 of Ref. [1] might have produced in some reader sentiments different from those of the members of the ROG Collaboration, who do not believe with high probability to have observed g.w.'s. However, the fact remains that the data are somewhat intriguing, and it is therefore important to quantify how much we can reasonably believe the hypothesis that they might contain some g.w. events. The aim of this paper is to show how to make a quantitative assessment of how much the experimental data prefer the different models in hand.

The choice of the Bayesian approach is quite natural to tackle these kind of problems, in which we are finally interested in the comparison of the probabilities that different models could explain the observed data. In fact, the concept of probability of hypotheses, probability of 'true values', probability of causes, etc., are only meaningful in this approach. The alternative ('frequentistic') approach forbids to speak about probability of hypotheses. Frequentistic 'hypothesis test' results are given in terms of 'statistical significance', a concepts which notoriously confuses most practitioners,

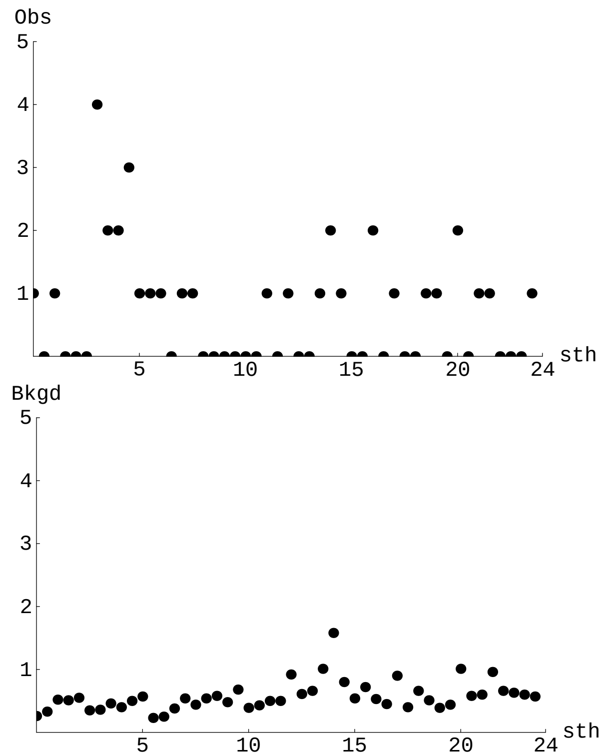


Figure 1: Explorer-Nautilus coincidence events (upper plot) and background estimates (lower plot) as a function of the sidereal time in 1/2 hour bins.

since it is commonly (incorrectly!) interpreted as it would be the probability of the 'null hypothesis' [2]. Moreover, this approach provides only 'accepted/rejected' conclusions, and thus it is not suited to extract evidence from noisy data and to combine it with other evidence provided by other data.

In the next section we present shortly the experimental data, referring to Ref. [1] and references therein for details. Then we review how the problem is approached in conventional statistics, explaining the reasons why we think that is unsatisfactory. Finally, we illustrate the Bayesian alternative for parametric inference and model comparison, and apply it to the ROG data.

2 Experimental data

This analysis has been performed on a data set of Explorer-Nautilus coincidences with an energy filter veto (i.e. requiring agreement between the event energies of the two antennae) and with a fixed time window of ± 0.5 s. Data we are referring to are those obtained using runs longer than 12 hours. The data are grouped in half hour bins of sidereal time, as shown in Fig. 1. The upper plot of the figure reports the number of observed coincidences (n_c), while the lower plot gives the average number of the background events estimated by off-time techniques. It is worth remarking that the method we are going to use does not depend critically on the width of the bins, provided that the width is small enough to assure a good resolution of the antenna pattern. (To state it clearly, contrary to other methods in which some binning is required and the resulting significance depends dramatically on the choice of the binning, in our method we could have, virtually, bins of arbitrary small width. Rebinning does not spoils the quality of the information, as long as the binning is finer than the structures exhibited by the antenna pattern and there are no clustering of events within a bin. The latter possibility is excluded by inspecting the arrival time of the individual events, as shown in Ref. [1] for the events around 4:00.)

As far as the background is concerned, we recall that the random coincidence background is well described by a Poisson distribution [1], and that the sidereal hour fluctuations of the averages is compatible with the grand average over the 24 hours of 0.57 ± 0.03 events/hour. For these reasons, we believe that the the value of $\lambda_B = 0.57$ is the most reasonable value to use as parameter of the Poisson distribution which models the background fluctuation in the 0.5-hour bins.

3 P-value analysis of the 'statistical significance' of the data

P-value is the term preferred in modern statistics to describe what physicists call, in simple words, "probability of the tail(s)," or "probability to observe the events actually observed, or rarer ones, *given* a null hypothesis" (note 'given': the probability of whatever *has been* observed, without the specification of a particular condition, is always unity). In the frequentistic approach, the null hypothesis is rejected with a significance level α if the p-value gets below α, where α is typically chosen to be 5% or 1%. Besides the recognized misinterpretation of the p-value result (see e.g. [2]), there are often disputes about how this reasoning should be applied, because it is easy to show that there is much arbitrariness in the kind of test to be performed (it is well known that practitioner often seeks for the test that tells what they like, moving for χ^2-test, to run-test and to other tests with fancy names, if the previously tried tests were "not sensitive to the effect") and in the data to include in the test, as it is sketched in the following subsections.

3.1 P-value based on the overall number of events

The expected number of events due to the null hypothesis $H_0 =$ "only background" is 27.4 ($= 0.57 \times 48$). Having observed 34 events we get:

$$\text{p-value}|_{\text{integral}} = P(n_c \geq 34 \,|\, \mathcal{P}_{\lambda_B=27.4}) = 12\% \,, \tag{1}$$

a value that it is not considered 'significative'. However, the obvious criticism to this procedure is that we have only used the integrated number of coincidences, losing completely the detailed information provided by the time distribution. The problem can be better understood in the limiting case of 1000 bins, an expected background of 1 event/bin, and an experimental result in which 999 bins have contents which 'nicely' (Poisson) fluctuate around 1, and a single bin exhibiting a spike of 31 counts. The p-value would be of 16%, accepting the hypothesis that the data are explained well by background alone.

3.2 P-value based on the bin presenting the highest excess

The naïve solution to this paradox is to calculate a p-value using only the bin presenting the highest fluctuation. This approach would give a very small number (0.5×10^{-36}) in our 1000 bin example, and would remain below the 1% threshold even if the spike has only 5 events over a background of 1. Applying this reasoning to the ROG data we get

$$\text{p-value}|_{\text{max}} = P(n_c \geq 4 \,|\, \mathcal{P}_{\lambda_B=0.57}) = 2.8 \times 10^{-3} \,, \tag{2}$$

a p-value which may be considered 'significative'. (Note that, if we used the observed background of Fig. 1, that we do not believe it is the correct number to use, the p-value would be 5.2×10^{-4} .)

3.3 P-value based on the argument that the highest excess could have shown up everywhere in the time distribution

Again, the previous procedure can be easily criticized, because "the bin to which the test has been applied has been chosen after having observed the data, while a peak would have been arisen, a priori, everywhere in the plot." The standard procedure to overcome this criticism is to calculate the probability that a peak of that or higher value would have shown up everywhere in the data, i.e.

$$\text{p-value}|_{\text{scan}} = 1 - \prod_{i=1}^{n_{bin}} F(3 \,|\, \mathcal{P}_{\lambda_B})\,, \tag{3}$$

where n_{bin} is the number of bins and $F(\cdot)$ stands for the cumulative distribution (the product in Eq. (3) is based on the assumption of independence of the bins). In our case we get 13% or 23% depending whether a constant or varying background is assumed, i.e. p-values above any over-optimistic choice of the p-value threshold.

It is interesting to note that the 13% p-value can be reobtained approximately as p-value$|_{\text{scan}} \approx n_{bin} \times$ p-value$|_{\text{max}}$, showing that even a very pronounced excess can be considered not significant if a large number of observational bins are involved in the experiment (and practitioners restrict arbitrary the region to which the test is applied, if they want the test to state what they would like…). The dependence of the result of the method on observations far from the region where there could be a good physical reason to have a signal is annoying (and for this reason, practitioners who choose a suitable region around the peak do, intuitively, something correct…). On the other hand, the reasoning does not take into account that other bins could be interested by the signal.

We shall see in Sec. 4.3 how to use properly the prior knowledge that a (physically motivated) signal could have appeared everywhere in the histogram.

3.4 Why not to use p-values

To conclude this section, let us summarize the reasons for not to use procedures based on p-values.

- The interpretation of p-values is misleading, because they do not provide probabilities of hypotheses, though they sound and are commonly interpreted as such.

- Methods based on p-values pretend to provide answers only based on the statistical properties of the null hypothesis, without taking into account if other hypotheses are conceivable, and how the alternative hypotheses describe the data. For example, these methods do not take into account the fact that a supposed signal appears at a given place rather than elsewhere, which bins could be affected by a physical model and how reasonable a model is.

- These methods provide only binary answers, accepted/rejected. As a consequence they are not efficient enough to analyze rare phenomena, which can only be discovered by a proper combination of (even very) small pieces of evidence.

4 The Bayesian way out: how to use of experimental data to update the credibility of hypotheses

We think that the solution to the above problems consists in changing radically our attitude, instead of seeking for new 'prescriptions' which might cure a trouble but generate others. The so called Bayesian approach, based on the natural idea of probability as 'degree of belief' and on the rules of logic, seems to us to be the proper way to deal with our problem. A key role in this approach is played by Bayes' theorem, which, apart from normalization constant, can be stated as

$$P(H_i \,|\, \mathit{Data}, I_0) \propto P(\mathit{Data} \,|\, H_i, I_0) \cdot P(H_i \,|\, I_0)\,, \tag{4}$$

where H_i stand the hypotheses that could produce the *Data* with *likelihood* $P(\mathit{Data} \,|\, H_i, I_0)$. $P(H_i \,|\, \mathit{Data}, I_0)$ and $P(H_i \,|\, I_0)$ are, respectively, the *posterior* and *prior* probabilities, i.e. with or without taking into account the information provided by the *Data*. I_0 stands for the general status of information, which is usually considered implicit and will then be omitted in the following formulae.

The presence of priors, considered a weak point by opposers of the Bayesian theory, is one of the points of force of the theory. First, because priors are necessary to make the 'probability inversion' of Eq. (4). Second, because in this approach all relevant conditions must be clearly stated, instead of being hidden in the method or left to the arbitrariness of the practitioner. Third, because prior knowledge can be properly incorporated in the analysis to integrate missing or deteriorated experimental information (and whatever it is done should be stated explicitly!). Finally, because the clear separation of prior and likelihood in Eq. (4) allows to publish the results in a way independent from $P(H_i \,|\, I_0)$, if the priors might differ largely within the members of the scientific community. In particular, the Bayes factor, defined as

$$BF_{ij} = \frac{P(\mathit{Data} \,|\, H_i)}{P(\mathit{Data} \,|\, H_j)}\,, \tag{5}$$

is the factor which changes the 'bet odds' (i.e. probability ratios) in the light of the new data. In fact, dividing member to member Eq. (4) written for hypotheses H_i and H_j, we get

$$\text{posterior odds}_{ij} = BF_{ij} \cdot \text{prior odds}_{ij}\,. \tag{6}$$

Since we shall speak later about models $\mathcal{M}_i$, the odd ratio updating is given by

$$\frac{P(\mathcal{M}_i \,|\, \mathit{Data})}{P(\mathcal{M}_j \,|\, \mathit{Data})} = \underbrace{\frac{P(\mathit{Data} \,|\, \mathcal{M}_i)}{P(\mathit{Data} \,|\, \mathcal{M}_j)}}_{\textit{Bayes factor}} \cdot \frac{P_\circ(\mathcal{M}_i)}{P_\circ(\mathcal{M}_j)} \tag{7}$$

Some general remarks are in order.

- Conclusions depend only on the observed data and on the previous knowledge. In particular they do not depend on unobserved data which are rarer than the data really observed (that is what p-values imply).
- At least two models have to be taken into account, and the likelihood for each model must be specified.
- There is no need to consider 'all possible models' (for which we can only wait the end of Humanity or of other intelligent beings…), since what matters are relative beliefs.
- Similarly, there is no need that the model must be declared before the data are taken, or analyzed. What matters is that the *initial* beliefs should be based on general arguments about the plausibility of each model and on agreement with other experimental information, other than *Data*.

An analogue of Eq. (4) applies to the parameters of a model. For example, if, given a model $\mathcal{M}$, we are interested to the rate of g.w. on Earth, r, Bayes' theorem gives

$$f(r \mid Data, \mathcal{M}) \approx f(Data \mid r, \mathcal{M}) \times f_\circ(r, \mathcal{M})\,, \qquad (8)$$

where $f()$ stand for probability density functions (pdf) Also in this case, a prior independent way of reporting the result is possible. The difficulty of dealing with an infinite number of Bayes factors (precisely ∞^2, given each r_i and r_j) can be overcome defining a function $\mathcal{R}$ of r which gives the Bayes factor with respect to a reference $r_\circ$. This function is particularly useful if $r_\circ$ is chosen to be the asymptotic value at which the experiment looses completely sensitivity. For g.w. search this asymptotic value is simply $r \to 0$. In other cases it could be an infinite particle mass [3] or an infinite mass scale [4]. In the case of g.w. rate r, extensively discussed in Ref. [5], we get

$$\mathcal{R}_\mathcal{M}(r) = \frac{f(Data \mid r, \mathcal{M})}{f(Data \mid r = 0, \mathcal{M})} = \frac{\mathcal{L}_\mathcal{M}(r)}{\mathcal{L}_\mathcal{M}(r = 0)}\,, \qquad (9)$$

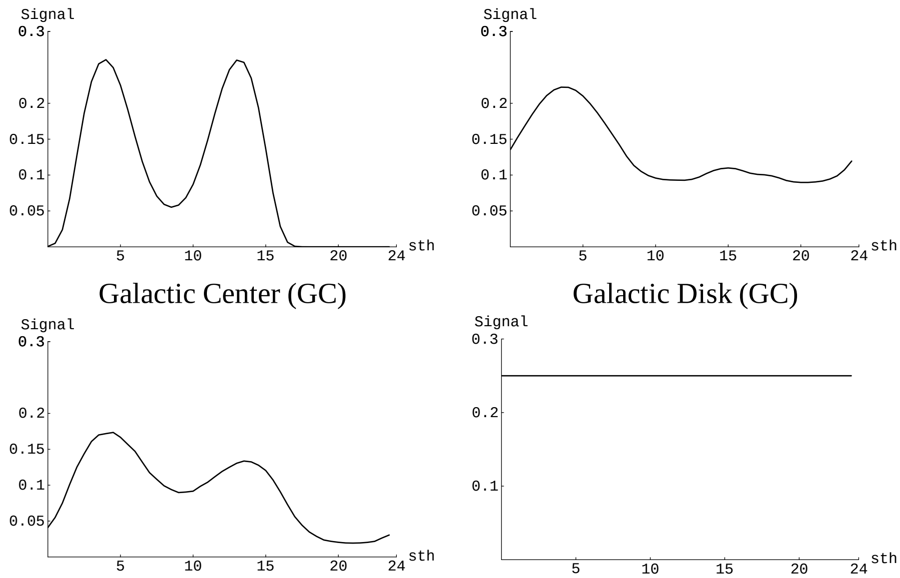
where $\mathcal{L}_\mathcal{M}(r)$ is the model dependent likelihood. [Note that, indeed, in the limit of $r \to 0$ the likelihood depends only on the background expectation and not on the specific model. Therefore $\mathcal{L}_\mathcal{M}(r = 0) \to \mathcal{L}_{\mathcal{M}_\circ}$, where $\mathcal{M}_\circ$ stands for the model "background alone".] This $\mathcal{R}$ function has the meaning of *relative belief updating factor* [5], since it tells us how we *must* modify our beliefs of the different values of r, given the observed data. In the region where $\mathcal{R}$ vanishes, the corresponding values of r are excluded. On the other hand, in the region where $\mathcal{R}$ is about unity, the data are unable to change our beliefs, i.e. we have lost sensitivity. The region of transition between 0 and 1 defines the *sensitivity bound*, a concept that does not have a probabilistic meaning and, since it does not refers to terms such as 'confidence', does not cause the typical misinterpretations of the frequentistic 'confidence upper/lower limits' (for a recent example of results using these ideas see Ref. [8]). Values of r preferred by the data are spotted by large value of $\mathcal{R}$. We shall in the sequel how a plot of the $\mathcal{R}$ function gives an immediate representation of what the data tell about a parameter (Figs. 3 and 4). Another interesting feature of this function is that, if several independent data sets are available, each providing some information about model $\mathcal{M}$, the global information is obtained multiplying the various $\mathcal{R}$ functions:

$$\mathcal{R}_\mathcal{M}(r\,;\, \textit{All data}) = \prod_i \mathcal{R}_\mathcal{M}(r\,;\, \textit{Data}_i)\,. \qquad (10)$$

4.1 Models for Galactic sources of gravitational waves

Having seen that the analysis has to be based on models for the emission of g.w.'s, let us focus on some popular models within Galaxy. This limitation is due to the sensitivity of the ROG detectors. The models taken into account are: emission only from sources concentrated in the Galactic Center (GC); emission from sources uniformly distributed over the Galactic Disk (GD); emission from sources distributed as the known visible mass of Galaxy (GMD). In addition we have also included the non-Galactic model of sources isotropically distributed around Earth (ISO). In this context, this latter model can just be seen as an academic example to give a feeling of what the analysis method would produce in such a scenario.

The pattern $\mathcal{P}_\mathcal{M}(i)$ of the two detectors Explorer and Nautilus to signals due to these models are shown in Fig. 2. These patterns are all given as a function of the sidereal time. $\mathcal{P}_\mathcal{M}(i)$ depends on the space and energy distributions of g.w. sources described by a model, taking also into account the variation of the detector efficiency with the sidereal time, due to the varying orientation of the antennae respect to the sources. It depends also on the (unknown) energies of g.w. signals, on the noise, and on the coordinates (latitude,longitude,azimuth) of the detectors on Earth. Simply



Galactic Center (GC)

Galactic Disk (GC)

Galactic mass distribution (GMD)

Isotropic (ISO)

Figure 2: Antenna patterns for the various models considered in the analysis.

speaking, for various models of the space distribution of sources of g.w., we expect different responses from the coincidence experiment, i.e. different 'antenna patterns', seen as a function of the sidereal time.

The calculation of $\mathcal{P}_{\mathcal{M}}(i)$ depends on the expected g.w. energy and some assumptions are needed. The signal's amplitude h is unknown, and thus we have evaluated the patterns by integrating over a uniform distribution of h values, ranging – on Earth – from 1×10^{-18} up to 3.0×10^{-18}. Different reasonings may be done here, leading to different choices for the amplitudes range and for the distributions of the signals. We have done the simplest choice, supposing that we do not know anything on the signals, but the fact that no signals have been observed at the detectors with amplitude greater than 3.0×10^{-18}. The lowest limit has been chosen considering the fact that the detector's efficiencies below 1×10^{-18} is very small (less than $\approx 10\%$). Note that the considered h values are based on 'standard' assumptions about the g.w. energy release in cryogenic bars. The Galactic Disk (GD) model has been constructed considering g.w. sources uniformly distributed over the Galactic plane, which means a distribution of sources which is not uniform around the Earth, given the fact that the Earth is 8.5 kpc from the Center of Galaxy, which is the center of the disk (whose radius is 21 kpc). The GMD distribution, taking into account the mass distribution in Galaxy, is much more interesting than the GD model. In fact we do not expect a uniform distribution of the sources over the GD, but a distribution which is concentrated near the GC [7].

4.2 Relative belief updating factor for a given model

Having introduced the general ideas, let us apply them to the data of our interest. In the i-th bin in sidereal time (see Fig. 1) we have $n_c(i)$ observed coincidences, with an expected number λ_B due to background and an expected number

$$\lambda_S^{\mathcal{M}}(i) = \alpha \times \mathcal{P}_{\mathcal{M}}(i) \tag{11}$$

due to the detectors response to g.w. signals, which depends on the model. The parameter α is proportional to the rate r of g.w. events, expressed as events/day through

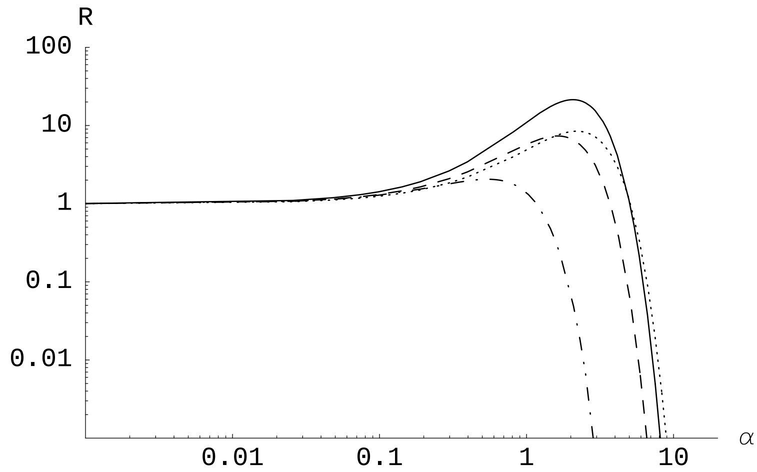


Figure 3: $\mathcal{R}$ function for the four models considered: Galactic Center (continuous); Galactic Disk (dashed); GMD (dotted); isotropic (dot-dashed).

the overall efficiency of detection, ϵ, and the total observation time, T_{obs}:

$$\alpha = r \times T_{obs} \times \epsilon \tag{12}$$

The efficiency of detection is calculated from the antenna pattern as

$$\epsilon = \frac{\sum_i^{n_{bins}} \mathcal{P}_{\mathcal{M}}(i)}{n_{bins} \times 1} \tag{13}$$

We have then the following likelihood for each model:

$$\mathcal{L}_{\mathcal{M}}(\alpha) = f(Data \mid \alpha, \mathcal{M}) = \prod_i \frac{e^{-\lambda_{\mathcal{M}}(i)}\, \lambda_{\mathcal{M}}(i)^{n_c(i)}}{n_c(i)!}\,, \tag{14}$$

with

$$\lambda_{\mathcal{M}}(i) = \lambda_B + \lambda_S^{\mathcal{M}}(i)\,. \tag{15}$$

$\lambda_S^{\mathcal{M}}(i)$ is the parameter of the Poisson distribution that describes the g.w. signal given model $\mathcal{M}$. It can be said, in simple words, to be the number of g.w. events that could be present in each bin in a coincidence experiment having performance and exposure time as that described in Ref. [1]. Hence,

$$n_{gwc} = \alpha \times \sum_i \mathcal{P}_{\mathcal{M}}(i) \tag{16}$$

gives the total number of such g.w. events. Our interest is, via α, to infer n_{gwc} and the rate r. However, given the strong dependence of the inference from the priors, typical for this kind of frontier measurements, we prefer to report the result in terms of $\mathcal{R}$ functions, as discussed above. The resulting $\mathcal{R}_{\mathcal{M}}(\alpha)$'s are shown in Fig. 3. Note that the figures are in log-log scale to make it clear that many orders of magnitude of α are involved.

The results are summarized in Tab. 1. α_{ML} is the value that maximizes $\mathcal{R}$. The symbol ML reminds that α_{ML} is the value that maximes the likelihood (likelihood and $\mathcal{R}$ function differ by a factor). Indeed, this is the result that a Maximum Likelihood (ML) analysis would produce for α. The parameter α is turned into g.w. rate using

Table 1: Summary of $\mathcal{R}$ function for each model, together with the parametric inference based on Maximum Likelihood or on the Bayes formula assuming a uniform prior for α.

Model	GC	GD	GMD	ISO
$\sum_i \mathcal{P}(i)$	4.7	6.1	4.4	12
ϵ (%)	9.8	13	9.2	25
$\mathcal{R}_{max}$	21	7.2	8.2	1.9
	(events)			
α_{ML}	2.1	1.6	2.2	0.5
$\sigma_{ML}(\alpha)$	1.0	0.9	1.2	0.5
$\mathrm{E}[\alpha \mid f_\circ(\alpha) = k]$	2.3	1.8	2.5	0.7
$\sigma[\alpha \mid f_\circ(\alpha) = k]$	1.0	0.9	1.2	0.4
	(events)			
$n_{gwc_{ML}}$	10	10	10	7
$\sigma_{ML}(n_{gwc})$	5	5	5	6
$\mathrm{E}[n_{gwc} \mid f_\circ(\alpha) = k]$	11	11	11	9
$\sigma[n_{gwc} \mid f_\circ(\alpha) = k]$	5	6	5	5
	(events/day)			
r_{ML}	1.1	0.9	1.2	0.3
$\sigma_{ML}(r)$	0.5	0.5	0.7	0.3
$\mathrm{E}[r \mid f_\circ(\alpha) = k]$	1.2	1.0	1.3	0.4
$\sigma[r \mid f_\circ(\alpha) = k]$	0.6	0.5	0.6	0.2

Eq. (12). $\mathcal{R}_{max}$ gives how much the belief for $\alpha = \alpha_{ML}$ increases with respect to $\alpha = 0$. The higher relative belief updating factor is obtained for the Galactic Center model. Within this model, the pdf for r around 1.2 events/day gets enhanced by a factor 21 with respect to $r = 0$.

Figure 3 shows clearly how the initial beliefs about α (and therefore on r) are updated, within each model. We want to stress that the final conclusion depends still on the prior beliefs. If someone thought that α had to be above 10 this person had to reconsider completely his/her beliefs, independently from the model; if another person believed that only values below 0.01 were reasonable, the experiment would not affect at all his/her beliefs, independently of the model. For this reason, the ML value α_{ML} could be misleading if erroneously associated, as it often happens, to the value around which our confidence is finally concentrated, independently from any prior knowledge. Nevertheless, and with these warnings, we report in Tab. 1 also the results obtained from a ML analysis and from a naïve Bayesian inference that assumes a uniform prior on α (and therefore on r and n_{gwc}, since they differ by factors). $\sigma_{ML}(\alpha)$ has been evaluated from the curvature of the minus-log-likelihood around its minimum, i.e. $\sigma_{ML}^{-2} = \partial^2/\partial\alpha^2\,(-\ln\,\mathcal{L}(\alpha))|_{\alpha_{ML}}$. The results of the 'naïve Bayesian inference' are reported as expected values $\mathrm{E}[\cdot]$ and standard deviations evaluated from the final distribution. The condition $f_\circ(\alpha)$ has been written explicitly in $\mathrm{E}[\cdot]$ and $\sigma[\cdot]$, according to the Bayesian spirit. Note that, for obvious reasons, the mode of the posterior calculated using a uniform prior is exactly equivalent to the ML estimate. This observation is important to understand the slightly different results obtained with the two methods. The posterior expected value is always larger than the ML one, simply because of the asymmetry of $\mathcal{L}$.

Perhaps n_{gwc} is the most interesting quantity to understand the conclusions of these *model dependent* analyses that, we like to repeat it, *do not take properly into* *account prior knowledge*. The three physical model suggest about 10 coincidences due to g.w.'s, with a 50% uncertainty. Instead, for the unphysical model (ISO) less events are found and with larger uncertainty. Note that, for this model, the mode of the posterior (or, equivalently, the ML estimate) gives a number of candidate events that is the difference between the total number of observed events and that expected from the background alone. Instead, for the three Galactic models, a number of events larger than this difference is attributed to the signal, as a consequence of a 'possibly good' time modulation recognized in the data (in other words, the method 'likes to think' that, given a time distribution shape that reminds the pattern of the Galactic models, the background has most likely underfluctuated within what is reasonably allowed by its probability distribution).

To summarize this subsection, the three Galactic models show good agreement in indicating for which values of g.w. events, or event rate, we *must increase* our beliefs. But the final beliefs depend on our initial ones, as explained introducing the Bayesian approach. If *you* think that, given your best knowledge of the models of g.w.'s sources and of g.w. interaction with cryogenic detectors, a g.w. rate on Earth of up to $\mathcal{O}(1)$ event/day is quite possible, the data make you to believe that this rate is $\approx 1.0 \pm 0.5$ event/day and that they contain $\approx 10 \pm 5$ genuine g.w. coincidences.

4.3 Model comparison taking into account the a priori possible values of the model parameters

While in the previous subsection we have been interested to learn about α or r *within* a model (and then, since all results are conditioned by that model, it makes no sense from that perspective to state if the model is right or wrong), let us see now how to *modify* our beliefs on each model. This is a delicate question to be treated with care. Intuitively, we can imagine that we have to make use of the $\mathcal{R}$ values, in the sense that the higher is the value and the most 'the hypothesis' increases its credibility. The crucial point is to understand that 'the hypothesis' is, indeed, a *complex* (somewhat *multidimensional*) hypothesis. Another important point is that, given a non null background and the properties of the Poisson distribution, we are never *certain* that the observations are not due to background alone (this is the reason why the $\mathcal{R}$ function does not vanish for $\alpha \to 0$).

The first point can be well understood making an example based on Fig. 3 and Tab. 1. Comparing $\mathcal{R}_{ML}$ for the different models one could come to the rash conclusion that the Galactic Center model is enhanced by 21 with respect to the non g.w. hypothesis, or that the Galactic Center model is enhanced by a factor $21/7$ with respect to the hypothesis of signals from sources uniformly distributed over the Galactic Disk. However these conclusions would be correct only in the case that each model would admit *only* that value of the parameter which maximizes $\mathcal{R}$, i.e.

$$\begin{aligned} BF_{GC(\alpha=2.1),\,GD(\alpha=1.6)} &= \frac{f(\mathit{Data}\,|\,GC,\,\alpha=2.1)}{f(\mathit{Data}\,|\,GD,\,\alpha=1.6)} \\ &= \frac{\frac{f(\mathit{Data}\,|\,GC,\,\alpha=2.1)}{f(\mathit{Data}\,|\,\mathit{BKGD\ alone})}}{\frac{f(\mathit{Data}\,|\,GD,\,\alpha=1.6)}{f(\mathit{Data}\,|\,\mathit{BKGD\ alone})}} = \frac{\mathcal{R}_{GC}(2.1)}{\mathcal{R}_{GD}(1.6)} \approx \frac{21}{7} = 3. \end{aligned} \tag{17}$$

But we are, indeed, interested in $BF_{GC\,,GD}$ and not in $BF_{GC(\alpha=2.1),\,GD(\alpha=1.6)}$. We must take into account the fact that a wide range of α values could be associated to each model.

Let us take the Bayes factor defined in Eq. (7). The probability theory teaches us promptly what to do when each model depends on parameters:

$$P(\mathit{Data}\,|\,\mathcal{M}) = \int P(\mathit{Data}\,|\,\mathcal{M},\boldsymbol{\theta})\, f(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}\,, \tag{18}$$

where $\boldsymbol{\theta}$ stands for the set of the model parameters and $f(\boldsymbol{\theta})$ for their pdf. Applying this formula to the Bayes factors of our interest we get

$$BF_{\mathcal{M}_i,\mathcal{M}_j} = \frac{P(Data\,|\,\mathcal{M}_i)}{P(Data\,|\,\mathcal{M}_j)} \quad = \quad \frac{\int \mathcal{L}_{\mathcal{M}_i}(\alpha\,;\,Data)\, f_{\circ_{\mathcal{M}_i}}(\alpha)\,\mathrm{d}\alpha}{\int \mathcal{L}_{\mathcal{M}_j}(\alpha\,;\,Data)\, f_{\circ_{\mathcal{M}_j}}(\alpha)\,\mathrm{d}\alpha} \tag{19}$$

where $f_\circ(\alpha)$ is the (model dependent) prior about α. Note that the Bayes factors with respect to $\mathcal{M}_0 =$"background alone" get the simple expression

$$BF_{\mathcal{M},\mathcal{M}_0} = \int \mathcal{R}_{\mathcal{M}}(\alpha)\, f_{\circ_{\mathcal{M}}}(\alpha)\,\mathrm{d}\alpha\,. \tag{20}$$

Equation (19) shows that the 'goodness' of the model depends on the integrated likelihood

$$\int \mathcal{L}_{\mathcal{M}}(\alpha\,;\,Data)\, f_{\circ_{\mathcal{M}}}(\alpha)\,\mathrm{d}\alpha \tag{21}$$

which is sometimes called 'evidence' (in the sense that "the higher is this number, the higher is the evidence that the data provide in favor of the model"). It is important to note that $\mathcal{L}_{\mathcal{M}}(\alpha\,;\,Data)$ has its maximum value around the ML point α_{ML}, but Eq. (21) takes into account all prior possibilities of the parameter. Thus, in general, it is not enough that one model fits the data better than its alternative (think, e.g., at the minimum χ^2 as a measure of fit goodness) to prefer finally that model. First there are the model priors, which we have to take into account. Second, the evidence (21) takes into account the parameter space preferred by the likelihood (i.e. the values around the ML point) with respect to the parameter space allowed a priori by the model. In the extreme case, one could have a model that can fit 'perfectly' the experimental data after having adjusted dozens of parameters, but this model yields a very small 'evidence' and it is therefore disregarded. This automatic filtering against complicated models is a nice feature of the Bayesian theory and reminds the Ockham' Razor criterion [9].

To better understand the role of the parameter prior in Eq. (21), let us take the example of a model (which we do not consider realistic and, hence, we have discarded a priori in our analysis) that gives a signal only in one of the 1/2 hours bins, being all bins a priori equally possible. This model $\mathcal{M}_s$ would depend on two parameters, α and t_s, where t_s is the center of the time bin. Considering α and t_s independent, the parameter prior is $f_\circ(\alpha, t_s) = f_\circ(\alpha)\cdot f_\circ(t_s)$, where $f_\circ(t_s) = 1/48$ is a probability function for the discrete variable t_s. The 'evidence' for this model would be

$$\sum_{t_s} \int \mathcal{L}_{\mathcal{M}_s}(\alpha, t_s\,;\,Data)\, f_\circ(\alpha)\, f_\circ(t_s)\,\mathrm{d}\alpha = \sum_{t_s} \frac{1}{48} \int \mathcal{L}_{\mathcal{M}_s}(\alpha, t_s\,;\,Data)\, f_\circ(\alpha)\,\mathrm{d}\alpha$$

If the data show a very large peak in correspondence of $t_s = t_{s_{ML}}$, we have that $\mathcal{L}_{\mathcal{M}_s}(\alpha, t_{s_{ML}}\,;\,Data) \ggg \mathcal{L}_{\mathcal{M}_s}(\alpha, t_s \neq t_{s_{ML}}\,;\,Data)$ and then

$$\sum_{t_s} \int \mathcal{L}_{\mathcal{M}_s}(\alpha, t_s\,;\,Data)\, f_\circ(\alpha)\, f_\circ(t_s)\,\mathrm{d}\alpha \approx \frac{1}{48} \int \mathcal{L}_{\mathcal{M}_s}(\alpha, t_{s_{ML}}\,;\,Data)\, f_\circ(\alpha)\,\mathrm{d}\alpha$$

This model is automatically suppressed by a factor ≈ 48 with respect to other models that do not have the time position as free parameter. Note that this suppression goes in the same direction of the reasoning described in Sec. 3.3. But the Bayesian approach tells us when and how this suppression has to be applied. Certainly not in the Galactic models we are considering.

As we have seen, while the Bayes factors for simple hypotheses ('simple' in the sense that they have no internal parameters) provide a prior-free information of how to modify the beliefs, in the case of models with free parameters Bayes factors remain independent from the beliefs about the models, but do depend on the priors about the model parameters. In our case they depend on the priors about α, which might be different for different models. If we were comparing different models, each with its $f_\circ(\alpha)$ about which there is full agreement in the scientific community, all further calculations would be straightforward. However, we do not think to be in such a nice text-book situation, dealing with open problems in frontier physics (for example, note that α, and then r and n_{gwc}, depend on the g.w. cross section on cryogenic bars, and we do not believe that the understanding of the underlying mechanisms is completely settled). In principle every physicist which have formed his/her ideas about some model and its parameters should insert his/her functions in the formulae and see from the result how he/she should change his/her opinion about the different models. Virtually *our task ends here*, having given the $\mathcal{R}$ functions, which can be seen as the best summary of an experimental fact, and having indicated how to proceed (for recent examples of applications of this method in astrophysics and cosmology see Refs. [10, 11, 12]). Indeed, we proceed, showing how beliefs can change given some possible scenarios for $f_\circ(\alpha)$.

The first scenario is that in which the possible value of α are considered so small that $f_\circ(\alpha)$ is equal to zero for $\alpha > 0.01$. The result is simple: the data are irrelevant and beliefs on the different models are not updated by the data.

Other scenarios might allow the possibility that $f_\circ(\alpha)$ is positive for values up to $\mathcal{O}(1)$ and more. We shall use three different pdf's for α as examples of prior beliefs, that we call 'sceptical', 'moderate' and 'uniform' (up to $\alpha = 10$). The 'moderate' pdf corresponds to a rate which is rapidly going to zero around the value which we have measured. The initial pdf is modelled with a half-Gaussian with $\sigma = 1$. The 'sceptical' pdf has a σ ten times smaller. The 'uniform' considers equally likely all α up to the last decade in which the $\mathcal{R}$ functions are sizably different from zero. Here are the three $f_\circ(\alpha)$:

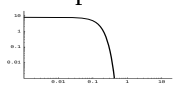
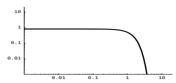
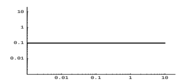
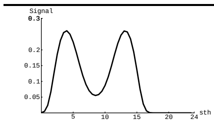
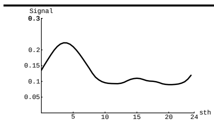
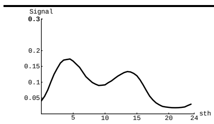
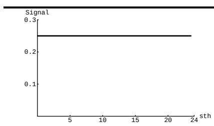
$$f_\circ(\alpha\,|\,\text{sceptical}) \propto \mathcal{N}(0,\,0.1) \qquad (\alpha > 0) \tag{22}$$
$$f_\circ(\alpha\,|\,\text{moderate}) \propto \mathcal{N}(0,\,1) \qquad (\alpha > 0) \tag{23}$$
$$f_\circ(\alpha\,|\,\text{uniform}) = k \qquad (0 < \alpha < 10)\,, \tag{24}$$

where $\mathcal{N}(\mu,\sigma)$ stands for a Gaussian distribution. For simplicity, we use the same sets of priors for all models, though they could, and probably should, be different for each model. But we think that this is sufficient for the purpose of this exercise, which is that of illustrating the method.

Using these three pdf's for the parameter α, we can finally calculate all Bayes factors. We report in Tab. 2 the Bayes factors of the models of Fig. 2 with respect to model $\mathcal{M}_0$ = "only background", using Eq. (20). All other Bayes factors can be calculated as ratio of these. The interpretation of the numbers is straightforward, remembering Eq. (5). If the preference was for α values below 0.1 (the 'sceptical'), the data produce a Bayes factor just above 1 for all models, indicating that the experiment has slightly increased our conviction, but essentially there is no model particularly preferred. If, instead, we think, though with low probability that even values of α above 1 are possible (i.e. $r \gtrsim 0.5$ event/day), then Bayes factors are obtained which that can sizably increase our suspicion that some events could be really due to one of these models.[1] Within this 'moderate' scenario there is some preference for the Galactic Center model with a Bayes factor about 2 with respect to each other model. This result contradict the naïve judgment based on observation of a 'peak' at around 4:00. The response of the Bayesian comparison takes into account all features of the model pattern, including the width of the peaks.

[1] To understand the quantitative role of the Bayes factors, let us make some examples: 8.4 means that if someone was in serious doubt to believe or not (i.e. $P = 50\%$), in the light of the data his/her belief increases to 89%. A BF of 100 would make the same person 'practically sure' (99%), or would bring into serious doubt a sceptical person (who had an initial belief of 1%).

Table 2: Bayes factors, for the four models of Fig. 2 with respect to model $\mathcal{M}_0 =$ "only background" depending on three choices for $f_\circ(\alpha)$. The thumbnails showing $f_\circ(\alpha)$ are log-log plots with abscissa scales exactly as in Fig. 3.

	'sceptical'	'moderate'	'uniform'
$f_0(\alpha)$			
Model			
	1.3	8.4	5.4
	1.4	4.1	1.7
	1.2	3.9	2.6
	1.2	1.4	0.2

5 Conclusion and discussion

This paper is mainly on methodological issues related to model comparisons in critical, frontier physics cases where the prior knowledge is relevant.

We have given reasons of why 'conventional statistics' (i.e. the collection of frequentistic prescriptions) does not adequately approach the problem of model comparison, mainly because of impossibility of classify hypotheses in a probability scale and of the pretension that good criteria to state what is 'significant' can be derived from the properties of the null hypothesis alone, without considering the details of the alternative hypotheses.[2]

Within the so called Bayesian framework, we have started from the basic observation that the most a probability theory should do is to provide rules to modify our beliefs on the light of experimental data. Beliefs can be about the values of the parameters of a model or about alternative models. As far as beliefs on model parameters are concerned, we have shown that the likelihood, rescaled to its insensitivity limit value (the $\mathcal{R}$ function, or 'relative belief updating factor'), represents a good, prior independent way of summarizing the information contained in the data with respect to a given model. Indeed, when this method is applied to the Explorer-Nautilus data, from the visual inspection of the $\mathcal{R}$ function the reader gets, for each model, an immediate overview of what the data say about the number of events involved in the observation ($\alpha = r \times T_{obs} \times \epsilon$). The α values for which the relative belief updating factor is maximum correspond to a total number of g.w. events in the data (n_{gwc}) about 10 for all three Galactic models. For those who share beliefs that numbers of this order of

[2] If you are puzzled by the question "why do frequentistic hypothesis test often work?", you might give a look at Chap. 10 of Ref. [2].

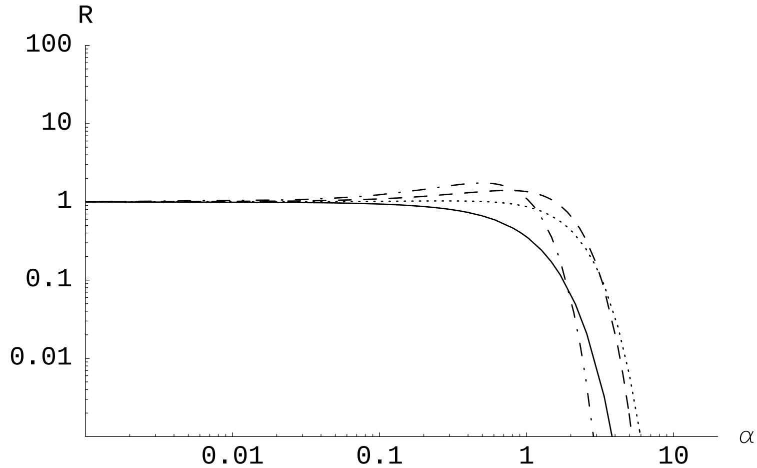


Figure 4: Same as Fig. 3, but for data grouped in solar time bins

.

magnitude or more are possible (and that one of the three models is the correct one), the $\mathcal{R}$ can be translated into a result $n_{gwc} \approx 10 \pm 5$ (or a rate on Earth of $\approx 1.0 \pm 0.5$).

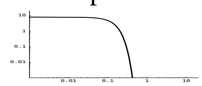
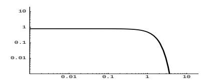
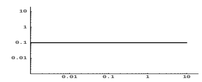
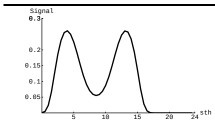
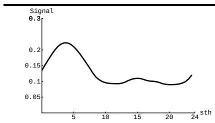
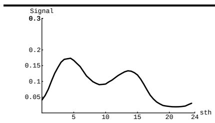
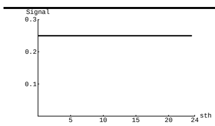
Going to the model comparison, we have shown the unavoidable complication due to the fact that each model depends on a free parameter (α) and, hence, the Bayes factors depend on the prior pdf of this parameter, i.e. $f_{\circ_{\mathcal{M}}}(\alpha)$. Since the models used do not come with a kind of reference $f_{\circ_{\mathcal{M}}}(\alpha)$ (we hope that more work will be done in this direction) we had to do some choices and we have given the results under different scenarios, from the most negative one ("there is no chance that the models produce something observable, given the present energy sensitivity") to some others in which α above 1 are conceivable (described by the priors we have called in the text 'moderate' and 'uniform'). Given these scenarios the Galactic Center model gets preferred over the others by a Bayes factor of about 2:1.

We would like to end replying to the objection, arisen often in discussions, that "the plot with coincidences grouped in bins of sidereal time provides the same information of that in which coincidences are grouped in bins of solar time". This might be true if one is blindly looking for "statistical significance", following strictly frequentistic prescriptions, which, as explained above, we do not consider the proper way to go [many igNobel prices should have been given in the past decades to (especially particle-)physicists who finally found a good "statistical significance"…].

To answer this objection we have done the exercise of applying exactly the same analysis with the same models to the data grouped in bins of solar time. The results are given in Fig. 4 and Tab. 3, which have the analogous meaning of Fig. 3 and Tab. 2. The $\mathcal{R}$ functions do not show values of α particularly preferred by the data: they have approximately a smoothed step-function shape which divides the orders of magnitudes of α (on the right side) that are excluded from those in which the experiment loses sensitivity. This is more ore less what we could get distributing 34 entries in 48 bins according to a multinomial distribution: small differences in the shape of $\mathcal{R}_{\mathcal{M}}(\alpha)$ depend on the individual occurrence of the multinomial data set (but never forget that the multinomial distribution does not *forbid* strong clustering of the entries around one bin!).

The lesson from this exercise is that the Explorer-Nautilus 2001 data, plotted as a function of a sensible physical quantity and compared with physically motivated models, does not provide the same information of any random sample. Indeed, the evidence in support of the models is not enough to modify strongly our beliefs, but it

Table 3: Same as Tab. 2, but for data grouped in solar time bins

	‘sceptical’	‘moderate’	‘uniform’
$f_0(\alpha)$			
Model			
	1.0	0.5	0.1
	1.1	1.2	0.3
	1.0	0.9	0.2
	1.2	1.2	0.2

is certainly at the level of “stay tuned”, waiting for results of the 2003 run.

6 Acknowledgments

The authors thank the ROG collaboration for having provided the 2001 data of Explorer and Nautilus. We also thank G. Giordano who has given us the information to compute the mass distribution model of Galaxy (GMD). Finally, P.A. and S.D. thank warmly the organizers of GWDA for such a vital and fruitful workshop.

References

[1] P. Astone et al., Class. Quantum Grav. **19** (2002) 5449.

[2] G.D’Agostini *Bayesian reasoning in data analysis: A critical introduction*, World Scientific Publishing 2003 (under publication, ISBN 981-238-356-5).

[3] G. D’Agostini and G. Degrassi, Eur. Phys. J. **C10** (1999) 633.

[4] ZEUS Collaboration, J. Breitweg et al., Eur. Phys. J **C14** (2000) 239.

[5] P.Astone and G.D’Agostini, CERN-EP/99-126 and hep-ex/9909047.

[6] G. D’ Agostini, Nuclear Physics B (Proc. Suppl.) **109B** (2002) 148.

[7] G. Paturel, Y. Baryshev, A&A **398** (2003) 377.

[8] P. Astone at al., Phys. Rev. **66** (2002) 102002.

[9] J.O. Berger and W.H. Jefferys, Am. Scientist **89** (1992) 64.

[10] T.J. Loredo and D.Q. Lamb, Phys. Rev. **D65** (2002) 063002.

[11] M.V. John and J.V. Narlikar, Phys.Rev. **D65** (2002) 043506.

[12] M.P. Hobson, S.L. Bridle and O. Lahav, astro-ph/0203259.