Easy Computation of the Bayes Factor to Fully Quantify Occam's Razor: Least-squares Fitting and Policy Decisions

D. J. Dunstan (d.dunstan@qmul.ac.uk)
Queen Mary University of London
J. Crowne
Queen Mary University of London
A. J. Drew
Queen Mary University of London

Research Article

Keywords: hypothesis selection and parameter estimation, computationally very intensive, physically-meaningful parameters

Posted Date: June 25th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-638470/v1

Easy computation of the Bayes factor to fully quantify Occam's razor: Least-squares fitting and policy decisions

D.J. Dunstan,* J. Crowne, and A.J. Drew

School of Physics and Astronomy, Queen Mary University of London,
London E1 4NS, UK

* Correspondence to **d.dunstan@qmul.ac.uk**

Abstract: The Bayes factor is the gold-standard figure of merit for comparing fits of models to data, for hypothesis selection and parameter estimation. However, it is little-used because it has been considered to be subjective, and to be computationally very intensive. A simple computational method has been known for at least 30 years, but has been dismissed as an approximation. We show here that all three criticisms are misplaced. The method should be used with all least-squares fitting, because it can give very different, better outcomes than classical methods. It can discriminate between models with equal numbers of parameters and equally good fits to data. It quantifies the Occam's Razor injunction against over-fitting, and it demands that physically-meaningful parameters rejected by classical significance testing be included in the fitting, to avoid spurious precision and incorrect values for the other parameters. It strongly discourages the use of physically-meaningless parameters, thereby satisfying the Occam's Razor injunction to use existing entities for explanation rather than multiplying new ones. More generally, being a relative probability, the Bayes factor combines naturally with other quantitative information to guide action in the absence of certain knowledge.

1. **Introduction**

"If your experiment needs statistics, you ought to have done a better experiment." – attributed to E. Rutherford. Nevertheless, almost every practising scientist, engineer, economist, etc, uses least-squares (LS) statistical methods to fit analytic expressions to data points. This is done for parameter estimation (uncertainties as well as values) and for hypothesis or model selection.[1] However, LS fitting poses questions. How to know if the fit is as good as may be? How to choose between models which all fit well? How to detect over-fitting and under-fitting? These questions require quantitative tests based on statistical theory. There are well-known statistical tools – significance tests – such as the traditional *p*-number or the 3-σ test, and the more recent AIC and BIC. Such tools are however inadequate, because they do not use the *prior* knowledge that we have.[2] The Bayes factor, derived from Bayes' theorem, does do this, and so has been described as the gold-standard figure-of-merit for comparing models. However, it is rarely used, not least because it can be computationally very demanding. Here we present an easy way of calculating it so that it can be routinely used with all least-squares fitting. We demonstrate its use – and usefulness – on three datasets from the literature. Outcomes can be very different from those both of significance testing and of the BIC. Moreover, when considering not merely whether a theory – a model – is true or not, but, as a practical matter, deciding what action should be taken given the outcomes of the fitting, the Bayes factor can quantitatively support intuition.

Bayes' Theorem explicitly includes prior knowledge in its calculation of the probability of a hypothesis given data. It was an unexceptionable part of probability theory in the nineteenth century. However, the increasing formalisation of probability theory and statistics in the twentieth century led to its sidelining, on the grounds that it introduces a subjective element, our state of knowledge, or grounds for belief, about future events. It was considered that probabilities should be purely objective. Jeffreys' seminal book in 1939 began the rehabilitation of Bayesian statistics.[3] This has been slow and controversial. For an entertaining historical survey, see the article by Leonard,[4] and for a non-technical discussion see Jaynes.[5] For an early technical account, see Kass and Raftery.[6]

Occam's Razor ("Entities should not be postulated without necessity"), in the context of least-squares fitting, demands that we should not use more fitting parameters than are necessary. That is, we should not overfit data. Classical – twentieth century – statistics scarcely quantifies this. In 1974, Akaike introduced the AIC (Akaike Information Criterion) which quantified this issue by preferring the model with the highest log-likelihood (see equation (2) in Section 2) less a penalty of n, the number of parameters.[7] This has now been largely supplanted by the BIC (Bayesian Information Criterion, or Schwartz criterion,[8] SBIC), like the AIC except that the penalty for n parameters is $\frac{1}{2}n\ln m$ where m is the number of data points. (Both the AIC and the BIC are usually presented after multiplication of these definitions by –2.) The BIC is now widely used.[9-11] Thousands of papers per year now cite BIC values to justify the choice of one model rather than another, e.g. in ecology.[12] However, the AIC and the BIC and many related criteria (DIC, FIC . . . WAIC) are gross approximations to the Bayes factor. Indeed, despite its name, the BIC is not Bayesian, and nor are the various related criteria. This is because they do not take into account the *prior* probabilities of the models. The Bayes factor does. In so doing, it quantifies two further intuitions, or corollaries, of Occam's razor. The first is that fits to data that use physically-meaningful parameters are preferable, if they fit, to fits that use physically-meaningless parameters such as coefficients in a polynomial or Fourier series. The latter introduce new entities while the former use entities that already exist. The second, closely related intuition, is that a model that is not capable of fitting all possible datasets (that does not span the data space) yet does fit the actual dataset is preferable to a model that could fit any data presented (that does span the data space).

Despite being the gold standard, the Bayes factor is little known and less used. It has been considered to be computationally massively intensive.[6,8-11] Except in simple problems of models with one fitting parameter, evaluating the Bayes factors of the models has required multi-dimensional integrals over parameter space. Fitting, for example, a multi-peak spectrum with tens of parameters, this requires computationally-heavy techniques such as Markov-chain Monte-Carlo integration. Because of the taint of subjectivism, in its use of what we know, many Bayesians have preferred to avoid prior knowledge and use in its place information obtained from the data, such as unit information priors. [6,13,14] Yet this concern is misplaced. What we know before analysing data is as objective, in the usual scientific sense, as the data themselves.

Here we present a formula for easy calculation of the Bayes factor after every LS fit with much less computational effort than the fit itself. This formula has been known since at least 1992,[14] and perhaps earlier. Its use in routine LS fitting has not been widely advocated, because of the subjectivity issue, and because it bypasses the computational difficulties of the Bayes factor by what has been widely thought to be an approximation, the Laplace approximation[15] – although McKay already in 1992 recognised this as exact in most situations.[14] Perhaps also because the value of the Bayes factor in quantifying the two further intuitions of Occam's razor mentioned above, and its value as a guide to action, have not been sufficiently appreciated. We present the method in Section 2. In Section 3 we briefly discuss the underlying theory, and in the Supplementary Information (SI §4) we give a derivation of the formula which we hope makes the underlying ideas clearer than they were in the older literature. Then in Section 4 we apply it to three examples of data-fitting in which the use of the Bayes factor leads to very different – and better – outcomes than traditional methods. Finally, in Section 5, we discuss the main outcomes, and consider the relevance of the Bayes factor to two live controversies. On significance (*p*-values etc) in fitting, we find that reliance on significance and the rejection of physically-meaningful parameters that do not pass significance tests will normally give incorrect results. On vitamin D and Covid-19, we see how the Bayes factor can combine with other information to provide quantitative support for actions that lack significant evidential support.

2. Methods

A least-squares fitting routine normally returns the parameter fitted values and their uncertainties, the fit residuals r_i and their standard deviation σ_r, and perhaps the parameter covariance matrix $\mathbf{Cov_p}$, the BIC, etc. The formula we apply uses the marginal likelihood integral (MLI) calculated for each LS fit. See Section 3. Calculating the MLI is done by,

$$\text{MLI} = (2\pi)^{n/2} L_{max} \frac{\sqrt{\det \mathbf{Cov}_p}}{\prod_{i=1}^{n} \Delta p_i} \tag{1}$$

where *n* is the number of parameters.[14] Then the Bayes factor between two models is the ratio of their MLI values. The first step in applying it is to calculate the quantity L_{max}, which is the value of the likelihood *L* at the fitted parameter values whether LS or ML fitting is used. *L* is the product of the probability densities of all the *m* datapoints given the fit. If it is not returned by an LS routine, it is readily calculated (see SI §S1). With perhaps hundreds of datapoints, *L* can be a very large or a very small number, depending on the value of the standard deviation of the residuals, σ, so it is more convenient to work with the log-likelihood, ln*L*. Equation (S1) shows that for a Gaussian distribution of residuals, maximising ln*L* is equivalent to minimising the sum of the squares of the residuals (the SSR). If the LS routine returns the SSR, then it is particularly easy to calculate ln*L*.

Next, we need $\mathbf{Cov_p}$. With software such as Mathematica, Matlab, Origin, etc, this is returned by the LS routine. If it has to be calculated, we show how in SI §S2.

The remaining term in equation (1) is the product of the *n* parameter ranges, Δp_i. These have to be decided upon and input by the user. There is nothing subjective about this, determined as they should be objectively by our *prior* knowledge, and open to reasoned debate and justification like any scientific knowledge or data. See SI §S3, and the examples in Section 4.

When we have the MLI values for two or more fits, their ratios give the relative probabilities for the models given the data – the Bayes factors (BF) between the models. It is more convenient to work with the logarithms, and then it is the difference of the lnMLI values, lnBF, which matters. Jeffreys and many subsequent authors have given verbal descriptions of the meaning of values of lnBF, in terms of the strength of the evidence in favour of the model with the higher lnMLI, such as <1 – barely worth considering, 1-2 – substantial, 2-5 – strong evidence, >5 – decisive.[3,6] More important than the verbal descriptions is that the Bayes factor simply expresses the relative probabilities of the models. The lnBF value corresponds to odds of $e^{\ln\mathrm{BF}}$ to 1 on the preferred model, or against the other model. The descriptions and the odds also apply to comparing models by differences in $\ln L_{max}$ between models with the same number of parameters, and by the Schwartz BIC (SBIC = –½BIC, which we use here for easy comparison with ln*L*, lnMLI and lnBF).

3. Theory

Equation (1) for the marginal likelihood integral has been given by many authors. Following Gull[16] we consider it first for a problem involving just one parameter λ distinguishing two versions of a theory (*The Story of Mr A and Mr B*, proposed originally by Jeffreys[3] and discussed by many authors). Mr A advocates the null hypothesis, **A,** in which this parameter does not appear. Mr B advocates the hypothesis, **B**, in which λ appears; least-squares fitting to the data **D** yields the fitted value $\lambda_0 \pm \delta\lambda$. Occam's razor tells us that the extra parameter λ should only be included if it is necessary. Then Bayes' theorem gives for the value of the Bayes factor, BF, for B against A,

$$\mathbf{BF} = \frac{Pr(\mathbf{B}|\mathbf{D})}{Pr(\mathbf{A}|\mathbf{D})} = \frac{Pr(\mathrm{B})}{Pr(\mathrm{A})} \times \frac{Pr(\mathbf{D}|\mathbf{B}, \boldsymbol{\lambda_0})}{Pr(\mathbf{D}|\mathbf{A})} \times \frac{\sqrt{2\pi}\,\delta\lambda}{\lambda_{\max}-\lambda_{\min}} \tag{3}$$

where Gull explains the first term in the RHS as having nothing to do with the theories or the data; it will normally be unity. Perhaps slightly tongue-in-cheek, Gull proposed that it could be adjusted to reflect the past performances of Mr A and Mr B. We take this term as unity here but we return to it in Section 5. The second term in the RHS is the ratio of the maximum likelihoods, which will normally favour **B** because adding fitting parameters will normally improve the fit to data. For **B**, it is the likelihood evaluated at the fitted value, λ_0. The third term in the RHS is the Occam factor, which will provide the penalty for the extra parameter in **B**. As Gull explains it, Mr B had to spread his probability *Pr*(**B**) over the *prior* range that he will have specified of possible values of λ from $\lambda_{\min}$ to $\lambda_{\max}$, with some pdf, that is assumed to be flat from $\lambda_{\min}$ to $\lambda_{\max}$.[6,14,16] When the data are given, the probability of the model becomes the integral (the MLI) of the product of this pdf and the function *L*(λ), treated as a Gaussian (this is the Laplace approximation[6,14,15]). Most of these possible parameter values perish and only the range described by the Gaussian of width σ = δλ survive. The integral thus becomes the area of the Gaussian times the flat value of Mr B's pdf, $1/(\lambda_{\max} - \lambda_{\min})$. Moreover, the width of this Gaussian, δλ, is the uncertainty or error σ_λ returned by the LS routine for λ.[2,16]

For models differing from the null hypothesis in more than one extra parameter, one might think that equation (3) could be generalised by multiplying the Occam's factors (the third term) for all the extra parameters together. That, however, normally grossly overestimates the MLIs, because of correlation or covariance between the parameters in the fits. The remedy is to use the square-root of the determinant of the parameter covariance matrix in place of the product of the uncertainties of the fitted parameter values, as in equation (1). See SI §4 for an explanation.

The ranges define a volume in parameter space, known as the prior parameter volume. Similarly, the square-root of the determinant of the covariance matrix defines another, smaller volume in the same space, the posterior parameter volume. The ratio of these two volumes is termed the Occam Factor.[16-18]

Our equation (1) is well-known in the literature, for example, it is equation (6) of MacKay's 1992 paper,[14] and equation (10.123) of Gregory's 2005 book.[19] However, in the rest of McKay's paper, and in most of the subsequent literature, the prior parameter volume in the denominator is not determined from our knowledge of the parameters and what values are physically realistic, but from the data (e.g. unit information priors). Indeed, that is the key step in using equation (1) to derive the BIC,[14] and is the reason the BIC treats all parameters alike. Gull[16] discusses the selection of the volume in the special case of one fitting parameter only, where the covariance matrix is not needed. Sivia and Skilling[2] also consider it but in the context of maximum likelihood fitting and apparently much more complicated calculations, in which our equation (1) is their equation (4.20). Much of the discussion of choice of priors is on mathematical, not physical grounds.[13,19,20] For a very recent survey, see Rougier and Priebe.[18]

It is worth noting that equation (1) is never analytically exact, because of the truncation of the integrals of the Gaussian functions $L(p_i)$ at the edges of the parameter prior volume, and eventually if $L(p_i)$ are not Gaussians. It is not difficult to check whether these issues are significant, nor to make reasonable corrections to the MLI when they are. See the example in Section 4.3, SI §8 Fig. S3.

These methods are applicable to Maximum Likelihood (ML) fitting. In contrast to LS fitting, ML fitting can easily handle the simultaneous fitting of multiple data sets, and datasets with different uncertainties σ_i on different residuals r_i, and it can handle outliers in a rigorous and respectable way.[21-23] See Example 2 (Section 4.2 and SI §8) for both these issues.

4. Examples of fitting data

4.1 *How many parameters best describe data in muon spectroscopy*?

Here we find that the Bayes factor demands the inclusion of more physically-meaningful parameters than the BIC or significance tests. We present some data that might reasonably be fitted with as few as three or as many as 22 physically-meaningful parameters. We find that the Bayes factor encourages the inclusion of all these parameters until the onset of over-fitting. Even though many of them have fitted values that fail significance tests, their omission distorts the fitting results severely.

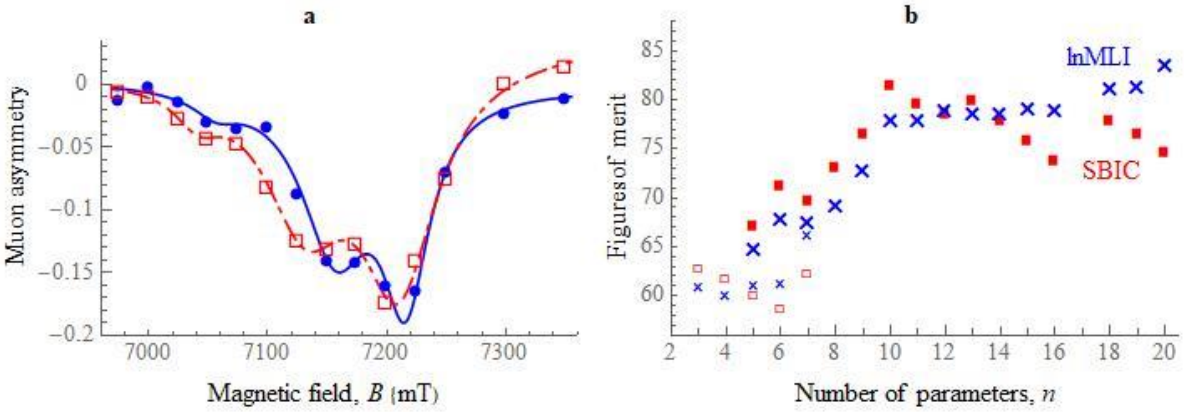

Fig.1. Muon-spin spectroscopy. Data from an experiment, muon polarisation as a function of magnetic field,[24] is shown in (a). Error bars on the data are estimated at ± 0.015. Linear background functions due to positrons have already been subtracted from the data. The blue solid-circle datapoints (•) were recorded

in the dark, while the red open-square datapoints (□) were photo-excited. The blue and red solid lines show 19-parameter fits of three Lorentzian peaks and two linear backgrounds, separately for the data in the dark (blue solid line) and photo-excited (red chain-dotted line). In (b), the evolution of the figures of merit of the fit with the number n of fitting parameters is shown (■ SBIC, × lnMLI). The open or small data points from three to seven parameters are for a single peak. The solid or large datapoints from five to 16 parameters are for two peaks, and from 18 to 20 parameters for three peaks.

Fig.1a shows an anti-level crossing spectrum observed in photo-excited muon-spin spectroscopy[24] from an organic molecule.[25] These spectra are expected to be Lorentzian peaks. Theory permits optical excitation to affect the peak position, the width and the strength (photosensitivity). In the field region over which the measurements are carried out, there is a background from detection of positrons, which has been subtracted from the data presented.[25] Wang *et al.*[25] did not attempt to fit the data; they did report a model-independent integration of the data, which demonstrated a change in area.

Fig.1b shows the evolution of the SBIC and the lnBF as the number of fitting parameters is increased. For a single peak, as the photosensitivity parameters $\Delta_L P$, $\Delta_L W$ and $\Delta_L A$ are introduced to the fit, (open and small data points for $n = 3 - 6$), the SBIC decreases and lnMLI scarcely increases. It is only with the inclusion of one background term ($n = 7$) that any figure of merit shows any substantial increase. There is no evidence here for photosensitivity. The weak peak around 7050 mT does not seem worth including in a fit, as it is evidenced by only two or three data points and is scarcely outside the error bars. Fitting with two peaks ($P_1 \sim 7210$ mT, $P_2 \sim 7150$ mT; solid or large data points for $p = 5 - 16$ in Fig.1b) gives substantial increases in the SBIC and lnMLI, further increased when the photosensitivity parameters $\Delta_L P_2$, $\Delta_L W_2$ and $\Delta_L A_2$ are included. The SBIC reaches its maximum here, at $n = 10$, and then decreases substantially. Additional parameters fail the significance tests as well as decreasing the SBIC (Fig.1b). Conventionally, the $n = 10$ fit would be accepted as best. The outcome would be reported as two peaks, with significant photo-sensitivities $\Delta_L P_2$, $\Delta_L W_2$ and $\Delta_L A_2$ for all three of the 7150 mT peak parameters, but no photosensitivity for the 7210 mT peak (Table I).

Table I. Photosensitivity results of fitting the data of Fig.1a with 10, 16 and 19 parameters. Parameter units as implied by Fig.1a.

	$\Delta_L P_1$	$\Delta_L W_1$	$\Delta_L A_1$	$\Delta_L P_2$	$\Delta_L W_2$	$\Delta_L A_2$
10-parameter fit	-	-	-	-14 ± 4 $p = 0.002$	21 ± 6 $p = 0.003$	-9 ± 2 $p = 0.0002$
16-parameter fit	-5 ± 3.5 $p = 0.18$	12 ± 8 $p = 0.16$	-7 ± 5 $p = 0.20$	-24 ± 8 $p = 0.008$	16 ± 13 $p = 0.24$	-5 ± 6 $p = 0.45$
19-parameter fit	-6 ± 2.8 $p = 0.07$	14 ± 6.2 $p = 0.05$	-9 ± 3.8 $p = 0.05$	-25 ± 4.8 $p = 0.0006$	10 ± 9 $p = 0.3$	-2.3 ± 4 $p = 0.6$

The Bayes factor gives a very different outcome. From 10 to 16 parameters, the Bayes factor between any two of these seven models is close to unity {Fig.1b). That is, they have approximately equal probability. The Bayes factor shows that what the conventional $n = 10$ analysis would report is false. Specifically, it is not the case that $\Delta_L P_2$, reported as -14 ± 4 mT, has a roughly ⅔ probability of lying between –10 and –18 mT. That is not consistent with the roughly equal probability that it lies in the $n = 16$ range (-24 ± 8). Table I shows that at $n = 16$, $\Delta_L P_2$ is the only photosensitivity parameter to pass significance tests. $\Delta_L A_2$, which had the highest significance level at $n = 10$, is now the parameter most consistent with zero. The other four are suggestively (about 1½σ) different from zero.

Since the Bayes factor has already radically changed the outcome by encouraging more physically-meaningful parameters, it is appropriate to try the 7050 mT peak parameters in the fit. With only 28 data-points, we should be alert to over-fitting. We can include P_3 and A_3 ($n = 18$), and $\Delta_L P_3$ ($n = 19$), but W_3 and $\Delta_L A_3$ do cause overfitting. Fig.1b shows substantial increases of both the SBIC and the lnMLI for $n = 18$ to $n = 20$, where the twentieth parameter is in fact $\Delta_L A_3$. The symptom of over-fitting that we observe here is an increase in the logarithm of the Occam Factor (lnMLI – ln$L$), the values of which decrease, –26.9, –33.5, –34.8, and then increase, –33.4, for $n = 16$, 18, 19 and 20 respectively. Just as ln$L$ must increase with every additional parameter, so should the Occam factor decrease, as the prior parameter volume must increase more with a new parameter than the posterior parameter volume. So we stop at $n = 19$. The outcome, Table I, is that the uncertainties on the $n = 16$ parameters have decreased markedly. This is due to the better fit, with a substantial increase in ln$L$ corresponding to reduced residuals on all the data. The 7210 mT peak 2 now has photosensitivities on all its parameters, significant to at least the 2σ or p-value ~0.05 level. And the photosensitivities $\Delta_L W_2$ and $\Delta_L A_2$, both so significant at $n = 10$, and already dwindling in significance at $n = 16$, are both now taking values quite consistent with zero. In the light of Table I, we see that stopping the fit at $n = 10$ results in completely incorrect results – misleading fitted values, with certainly false uncertainties.

4.2 *Fitting nearly linear data for the pressure dependence of the GaAs bandgap.*

The main purpose of this example is to show how the Bayes factor can be used to decide between two models which have equal goodness of fit to the data (equal values of ln$L$ and BIC, as well as p-values, etc). This illustrates the distinction it makes between physically-meaningful and physically meaningless parameters. This example also shows how ML fitting can be used together with the Bayes factor to obtain better results. For details, see SI §7.

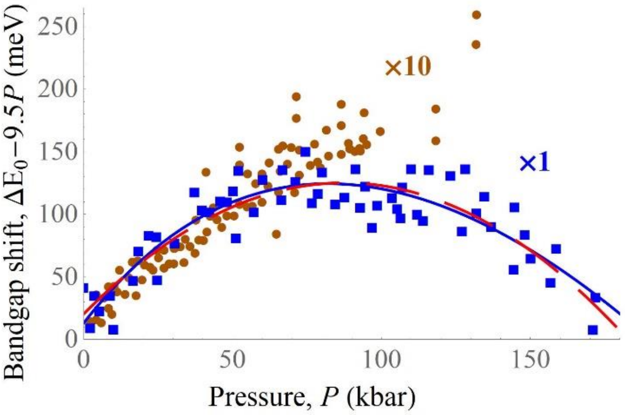

Fig.2. Data for $E_g(P)$ in GaAs from Goñi *et al.*[26] (■) and from Perlin *et al.*[27] (●) are shown after subtraction of the straight line $E_0 + 9.5P$ to make the curvature more visible. The Perlin data is expanded ×10 on both axes for clarity. Two least-squares fits to the Goñi data are shown, polynomial (dashed red line) and Murnaghan (solid blue line).

Fig.2 shows two datasets for the pressure dependence of the bandgap of GaAs. The original authors published quadratic fits, $E_g(P) = E_0 + bP + cP^2$, with $b = 10.8 \pm 0.3$ meV kbar^{-1} (Goñi *et al.*[26]) and 11.6 ± 0.2 meV kbar^{-1} (Perlin *et al.*[27]). Other reported experimental and calculated values for b ranged from 10.02 to 12.3 meV kbar^{-1}.[28] These discrepancies of about ±10% were attributed to experimental errors in high-pressure experimentation. However, from a comparison of six such datasets, Frogley *et al.* were able to show that the discrepancies arose from fitting the data with the quadratic formula. The different datasets were reconciled by using the Murnaghan equation of state and supposing the band-gap to vary linearly with the density (see SI, §7, equations (S4) and (S5).[28] The curvature c of the quadratic is constant, while the curvature of the density, due to the pressure dependence B' of the bulk modulus B_0, decreases with pressure – and the six datasets were recorded over very different pressure ranges, as in Fig.2. So the fitted values of c, c_0, were very different, and the correlation between b and c resulted in the variations in b_0.

Here, using the Bayes factor, we obtain the same result from a single dataset, that of Goñi *et al.*[26] The two fits are shown in Fig.2. They are equally good, with values of ln$L$ and SBIC the same to 0.01. The key curvature parameters, c and B$'$, are both returned as non-zero by 13.5σ (SI, §7, Table S1), consequently both with p-values less than 10^{-18}. However, c is a physically-meaningless parameter. The tightest constraint we have for setting its range is the values previously reported, ranging from 0 to 60 μeV kbar^{-2}, so we use Δc = 100 μeV kbar^{-2}. In contrast, B$'$ is known for GaAs to be 4.49.[29] For many other materials and from theory the range 4 to 5 is expected, so we use $\Delta \mathrm{B}' = 1$. The other ranges are same for both models (see SI §7). This difference gives a lnBF of 3.8 in favour of the Murnaghan model against the quadratic, which is strong evidence for it. Moreover, the value of B$'$ returned is 4.47 ± 0.33, in excellent agreement with the literature value. Had it been out of range, the model would have to be rejected. The quadratic model is under no such constraint; indeed, a poor fit might be handled by adding cubic and higher terms *ad lib*. This justifies adding about 5 to lnBF (see Section 4.3), giving a decisive preference to the Murnaghan model, and the value of b it returns, 11.6 ± 0.3. Note the good agreement with the value from Perlin *et al.*[27] If additionally we fix B$'$ at its literature value of 4.49,[29] lnBF is scarcely improved, because the Occam factor against this parameter is small, but the uncertainty on the pressure coefficient, Ξ/B_0, is much improved.

When we fit the Perlin data, the Murnaghan fit returns B$'$ = 6.6 ± 2.4. This is outside range, and indicates that this data cannot give a reliable value – attempting it is over-fitting. However, it is good to fit this data together with the Goñi data. The Perlin data, very precise but at low pressures only, will complement the Goñi data with its lower precision but large pressure range. We notice also that the Perlin data has a proportion of outlier data points. Maximum Likelihood fitting handles both issues. We construct ln$L$ using different pdfs $P(r)$

for the two datasets, and with a non-Gaussian pdf for the Perlin data (see SI §7 equation (S6). Fixing B′ at 4.49, fitting with the same Ξ/B_0 returns 11.42 ± 0.04 meV kbar^{-1}. Separate Ξ/B_0 parameters for the two datasets give an increase of ln$L$ of 4.6, with values 11.28 ± 0.06 and 11.60 ± 0.04 meV kbar^{-1} – a difference in b of 0.32 ± 0.07 meV kbar^{-1}, which is significant at 4½σ. This difference could be due to systematic error, e.g. in pressure calibration. Or it could be real. Goñi *et al.*[26] used absorption spectroscopy to measure the band-gap; Perlin *et al.*[27] used photoluminescence. The increase of the electron effective mass with pressure might give rise to the difference. In any case, it is clear that high-pressure experimentation is much more accurate than previously thought, and that ML fitting exploits the information in the data much better than LS fitting.

4.3 *Carbon nanotube Raman spectrum*

This example demonstrates how the Bayes Factor provides a quantitative answer to the problem, whether we should accept a lower quality of fit to the data if the parameter set is intuitively preferable. It also provides a simple example of a case where the Bayes factor calculated by equation (1) is in error and can readily be corrected (see SI §8 Fig.S3).

The dataset is a Raman spectrum of the radial breathing modes of a sample of carbon nanotubes under pressure, previously fitted in the traditional way.[30] The key issue in the fitting was to get the intensities of the peaks as accurately as possible, to help understand the evolution with pressure. Here, we take a part of the spectrum recorded at 0.23 GPa and monitor the quality of fit and the Bayes factor while parameters are added in four models. This part of the spectrum has seven sharp pseudo-Voigt peaks (Fig3a; the two strong peaks are clearly doublets). With seven peak positions P_i, peak widths W_i and peak intensities A_i, and a factor describing the Gaussian content in the pseudo-Voigt peak shape, there are already 22 parameters (for details, see SI §8). This gives a very poor fit, with ln$L$ = –431, SBIC = and lnMLI = –546. The ranges chosen for these parameters for calculating the MLI (see SI §8) are not important because they are used in all the subsequent models, and so they cancel out in the Bayes factors between the models.

To improve the fit, in the Fourier model we add a Fourier background $y = \sum c_i \cos x + s_i \sin x$, and in the Polynomial model, we add $y = \sum a_i x^i$ for the background. In both, the variable x is centred (x = 0) at the centre of the fitted spectrum and scaled to be ±π or ±1 at the ends. In the Peaks model we add extra broad peaks as background, invoking extra parameter triplets (P_i, W_i, A_i). These three models all gave good fits; at the stage shown in Fig.3a they gave ln$L$ values of –65, –54 and –51 and BIC values of –156 , –153 and –148 respectively. Thus there is not much to choose between the three models, but it is noteworthy that they give quite different values for the intensities of the weaker peaks, with the peak at 265 cm^{-1} at 20.5 ± 1.1, 25.5 ± 1.3 and 27 ± 1.7 respectively (this is related to the curvature of the background function under the peak). So it is important to choose.

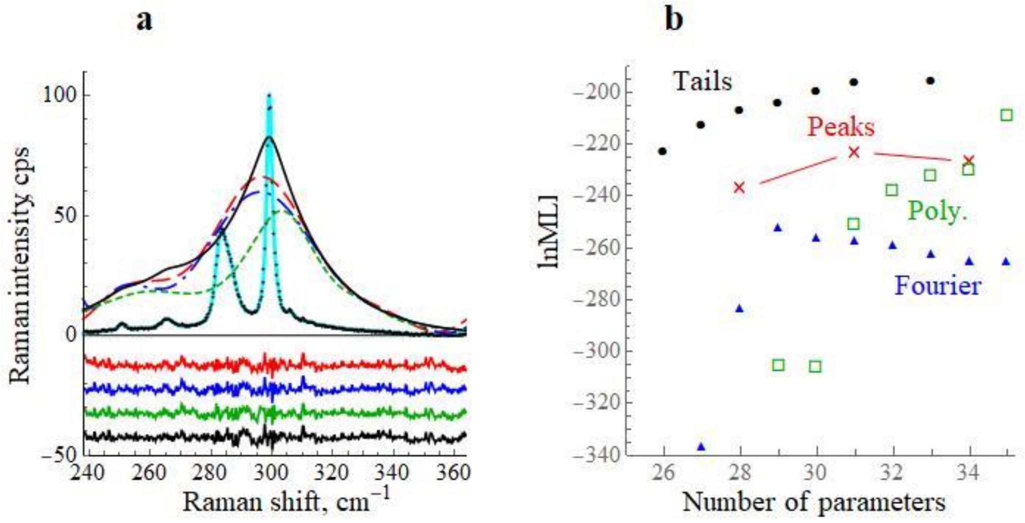

Fig.3. Carbon nanotube spectrum. In (a), the carbon nanotube Raman spectrum is plotted (black datapoints) with a fit (cyan solid line) using the Fourier model. The residuals for four good fits are shown, ×10 and displaced successively

downwards (Fourier, Polynomial, Peaks and Tails; see text), and the backgrounds are shown, ×8 (long dashed, chain-dotted, short dashed and solid, respectively. In (b), the evolution of the MLIs is shown against the number of parameters for these four models.

However, a fourth model was motivated by the observation that the three backgrounds all look as if they are related to the sharp peaks, rather like heavily broadened replicas (see Fig.3a). Accordingly, in a fourth model, we use no background, and instead modify the peak shape, giving it stronger, fatter tails than the pseudo-Voigt peaks (Tails model). This was done by adding to the Lorentzian peak function a second Lorentzian term with much greater width, or, still better, a smooth function approximating to exponential tails on both sides of the peak position (for details, see SI §8) with widths and amplitudes as fitting parameters. What is added may be considered as background and is shown in Fig.3a. This model, at the stage of Fig.3a, returned ln$L$ = –62, BIC = –146, and yet another very different value of 15.5 ± 1.0 for the intensity of the 265 cm^{-1} peak.

The Tails model is intuitively preferable to the other three because it does not span the data space – e.g. if there was really were broad peaks at the three positions identified by the Peaks model, or elsewhere, the Tails model could not fit them. That it does fit the data is intuitively strong evidence for its correctness. The Bayes factor confirms this intuition quantitatively. At the stage of Fig.3a, the lnMLI values are –251, –237 and –223 for the Fourier, Poly and Peaks models, and –211 for the Tails model. This gives a lnBF value of 11 for the Tails model over the Peaks model – decisive – and still larger lnBF values for these models over the Fourier and Poly models.

All models can be taken further, with more fitting parameters – more Fourier or polynomial terms or more peaks, and for the Tails model more parameters distinguishing the tails attached to each of the seven Lorentizian peaks. In this way, the three background models can improve to a ln$L$ ~ –20; the Tails model does not improve above ln$L$ ~ –50. However, as seen in Fig.3b, the MLIs worsen with too many parameters – except when over-fitting occurs, as seen for the Poly model at 35 parameters. The Tails model retains its positive lnBF > 10 over the other models.

The other models can have an indefinite number of additional parameters – more coefficients or more peaks, to fit any data set. It is in this sense that they span the data space. The actual number used is therefore itself a fitting parameter, with an uncertainty perhaps of the order of ±1, and a range from 0 to perhaps a quarter or a half of the number of data points d. We may therefore add to their lnMLIs the factor ~ln $4d^{-1}$ ~ –5 for a few hundred data points. This takes Tails to a lnBF > 15 over the other models – overwhelmingly decisive. This quantifies the intuition that a model not guaranteed to fit the data, but which does, is preferable to a model that certainly can fit the data because it spans the data space. It quantifies the question, how much worse a quality of fit should we accept for a model that is intuitively more satisfying; here we accept –30 on ln$L$ for a greater gain in the Occam factor. It quantifies the argument that the Tails model is the most worthy of further investigation because the fat tails probably have a physical interpretation worth seeking (in this context, it is interesting that in Fig.3a tails have been added only to the 255, 265 and 299 cm^{-1} peaks; adding tails to the others did not improve the fit; however, a full analysis and interpretation is outside the scope of this paper). In the Peaks model it is not probable (though possible) that the extra peaks would have physical meaning. In the other two models it is certainly not the case that their coefficients will have physical meaning.

5. **Discussion and Conclusions:**

The most surprising outcome of Section 4 is the desirability of including in models some parameters that fail significance tests, and reporting the outcomes. This is relevant to the controversy about significance tests such as *p*-values.

In the story of Mr A and Mr B, the two models are explicitly given equal *a priori* probabilities, $p(\mathbf{A}) = p(\mathbf{B}) = ½$ if there are no other models in contention, and before any data the lnBF between them is zero. Suppose that the fit using model **A** has given a set of parameter values $\mathbf{V}_\mathrm{A} = (p_{i0} \pm \delta p_i)$, defining the posterior parameter volume. With model **B**, including the extra parameter, correlations between parameters result in giving $\mathbf{V}_\mathrm{B} = (p'_{i0} \pm \delta p'_{i0}, \lambda_0 \pm \delta\lambda)$, defining a different posterior parameter volume. The uncertainties $\delta p'_i$ will generally be larger than δp_i, and the values p'_{i0}will generally be different from p_{i0}. For illustration, suppose that λ_0 is non-zero but fails significance tests, being perhaps just 1 or 2σ away from zero, and that the MLIs come out equal (i.e. the improvement in ln*L* in Model **B** is offset by the Occam factor, and lnBF remains at zero). Now to reject λ and to report only the fit to model **A** is to assert that the true values p_i have each a ⅔ chance of lying within $\mathbf{V}_\mathrm{A}$, within the 1σ ranges δp_i. However, that assertion is conditional on λ actually having the value zero; that is, it is conditional on the truth of the null hypothesis **A**. And that is a condition that we do not know to be true. The failure of **B** to attain significance is often mistakenly described as evidence for the null hypothesis **A**. Amrhien *et al.* report that around half of a large number of articles surveyed in five major journals make this mistake.[31] It is not just a scientific mistake.[10] It can be a disastrous guide to action.

According to the Bayes factor, the models **A** and **B** have equal probabilities, ½, and so what we know is that the parameters of model **A** have each a ⅓ chance of lying within their 1σ ranges δp_i around p_i and a ⅓ chance of lying within the 1σ ranges $\delta p'_i$ around p'_i. In fact, in this situation (and especially if a significant non-zero λ_0 would be an exciting result – see Ref. 32 and discussion below for a current example) the usual reaction to finding that λ_0 is 2σ away from zero is to repeat the experiment, to take more data. Of course, that has some chance of finding a λ_0 closer to zero, but it also has a good chance of confirming a non-zero λ_0. So the Bayes factor is a guide to action; the significance test is not.

Truth is not within the remit of probability theory. From its origins in Pascal's and Fermat's advice to the gambler the Chevalier de Méré (1654),[33] probability is fundamentally about how to act when we do not know what will happen (or what is true). Whether it be the turn of a card in poker, the weather forecast, or the administration of an untried medicament, we can write the value or profit of a success or win as $Pr(\text{win}) = p(\text{win}) \times \text{winnings}$, and similarly for a failure or a loss. In poker, the Expected Value of an action is defined as EV = $Pr(\text{win}) - Pr(\text{loss})$, and it is used to guide decisions how to act – whether to bet, or fold. The Bayes factor is the ratio of the probabilities of competing theories given the data. So it lends itself directly to multiplication by the financial or other quantifiable estimates of outcomes to guide actions.

Consider the current controversy about vitamin D and Covid-19. Model **A** (the null hypothesis) recommends inaction (action A), Model **B** recommends mass medication with vitamin D as a prophylactic (action B), and further research on the question (action C) may also be considered. The evidence for Model **B** is weak, but it is not insubstantial. A recent editorial in the BMJ concluded that it is strong enough to make the case for action C "compelling."[34] Martineau summarised the case for action B as ". . . it's not the highest level of evidence. I guess there's a philosophical question – if you have an intervention [action B] that has a good chance of working and is completely safe, why not implement it?"[35]

Of course, there are answers to Martineau's seemingly rhetorical question. There is the cost. Paying for action B means that something else won't be paid for, and if that would have worked and action B does not then action B will – at least in hindsight – have been a poor decision. There is the question, which of perhaps an unlimited number of equivalent actions B′ might be chosen – intravenous bleach or homeopathy as well as vitamin D? If one, why not all the others? Martineau's "if completely safe" is also important, since virtually nothing is completely safe. These points are important complexities, but citing them does not definitively answer the question.

Using the Bayes factor, Martineau's question can be answered quantitatively. A "good chance" implies a lnBF in the range ±1 for Model **B** against Model **A**. Crudely, the benefit of taking no action, A, is the saving on the cost of actions B and C. Maybe some £10^8. The benefit of action B at once, if Model **B** is true, is, crudely, some £10^{11} in the avoidance of unnecessary deaths and lockdowns. The benefit of action C alone is much more complex, delayed, even negative, if it displaces research into other therapies, but, crudely, it delays action B so its best return is smaller. So the contributions of ln $Pr(\mathbf{B})/Pr(\mathbf{A})$ to add to lnBF are about ln1000 = +7 for B and (less certainly) about +5 for C alone. A full analysis should of course refine these costs and benefits by costing the complexities. And of course it could use other quantitative data than financial, such as numbers of deaths. But if it were to confirm these outcomes, both B and C should be undertaken urgently.

The issue of bleach and homeopathy is readily dealt with. With an unlimited number of putative actions B' based on models $\mathbf{B}'$ to consider, their *a priori* probabilities should be rated as very small, except when there is evidence for them that is rated as not insubstantial. Then the factor $P(\mathbf{B}')/P(\mathbf{A})$ will outweigh – negatively – the factor $Pr(\mathbf{B}')/Pr(\mathbf{A})$.

For a simpler example, consider the example of Ref.32. They find evidence (from the LHCb experiment at CERN) for the violation of lepton universality (Model **B**), at the 3.1 sigma level, that is, a probability of 0.997, and a lnBF = ln $P(\mathbf{B})/P(\mathbf{A})$ against the null hypothesis (Model **A**) of –ln 0.003 = 6. This is sufficient to encourage further work. It may be further increased by ln $Pr(\mathbf{B})/Pr(\mathbf{A})$, if the value of physics beyond the Standard Model can be estimated, and the costs of the further work. The value is presumably of the order of the total cost of the Large Hadron Collider, as this is what is was built to find.

In conclusion, calculation of Bayes factors should be a routine part of all data fitting. It gives advice that is the opposite of much standard practice, but which satisfies Occam's Razor intuitions, and enables robust model selection and parameter estimation. Bayes factors, being the ratio of probabilities, are readily multiplied by financial or other quantitative data to quantify intuitive or philosophical arguments for actions.

References

1. W.A. Fuller. *Measurement Error Models*. (Wiley-Blackwell, Oxford, 1987).

2. D.S. Sivia and J. Skilling, J. *Data Analysis*: *A Bayesian Tutorial*. (Oxford University Press, Oxford, 2006).

3. H. Jeffreys. *Theory of Probability* (Oxford University Press, 1939, 1948, 1961, 1979).

4. T.H. Leonard. 2014. A personal history of Bayesian statistics. *WIREs Comput. Stat.* **6**, 80-115. (doi 10.1002/wics.1293)

5. E.T. Jaynes, *Bayesian methods: General background. An introductory tutorial.* In *Maximum Entropy and Bayesian Methods in Applied Statistics* (ed. Justice, J.H.) pp. 1-25 (Cambridge University Press, 1985).

6. R.E. Kass and A.E. Raftery. 1995. Bayes Factors. *J. Am. Stat. Assoc.* **90**, 773-795.

7. H. Akaike. 1974. A new look at the statistical model identification. *IEEE Trans. Automatic Control* **19**, 716-723.

8. G.E. Schwarz. 1978. Estimating the dimension of a model. *Ann. Stat.* ***6**, 461-464.*

9. T.J.Faulkenberry. 2018. Computing Bayes factors to measure evidence from experiments: An extension of the BIC approximation. *Biometrical Lett.* **55**, 31-43.

10. E.-J. Wagenmakers. 2007. A practical solution to the pervasive problem of *p* values. *Psychonomic Bull. Rev*. **14**, 779-804.

11. A.F. Jarosz & J. Wiley. 2014. What are the odds? A practical guide to computing and reporting Bayes Factors. *J. Problem Solving* **7**, Art.2 (2014).

12. D.J. Dunstan & D.J. Hodgson. 2014. Snails home. *Phys. Scr.* **89**, 068002.

13. T.S. Eicher, C. Papageorgiou & A.E. Raftery. 2011. Default priors and predictive performance in Bayesian model averaging, with application to growth determinants. *J. Appl. Econ.* **26**, 30-55 (2011).

14. D.J.C. MacKay. 1992. Bayesian interpolation. *Neural Computation* **4**, 448-472.

15. L. Tierney and J.B. Kadane. 1986. Accurate approximations for posterior moments and marginal densities. *J. Am. Stat. Assoc.* **81**, 82-86.

16. S.F. Gull. Bayesian inductive inference and maximum entropy. In *Maximum Entropy and Bayesian Methods in Science and Engineering* (Kluwer Academic Publishers, 1988), ed. Erickson G.J. and Smith C.R., vol.1, pp 53-74.

17. C.E. Rasmussen, Z. Ghahramani. Occam's Razor. In *Advances in Neural Information Processing Systems* **13** (MIT Press, Cambridge MA, 2001), ed. T.K. Leen, T.G. Dietterich & V. Tresp.

18. J. Rougier & C.E. Priebe. 2020. The exact form of the "Ockham Factor" in model selection. *Amer. Statistician*, DOI 10.1080/00031305.2020.1764865.

19. P.C. Gregory. *Bayesian Logical Data Analysis for the Physical Sciences* (Cambridge University Press, 2005)

20. D. Lunn, C. Jackson, N. Best, A. Thomas, A. & D. Spiegelhalter. *The BUGS Book – A Practical Introduction to Bayesian Analysis* (CRC Press / Chapman and Hall, 2012).

21. H.J. Motulsky & R.E. Brown. 2006. Detecting outliers when fitting data with nonlinear regression – a new method based on robust nonlinear regression and the false discovery rate. *BMC Bioinformatics* **7**, 123.

22. Y. Li, A.J. Bushby & D.J. Dunstan. 2018. Factors determining the magnitude of grain-size strengthening in polycrystalline metals. *Materialia* **4**, 182-191.

23. A. Daemi, H. Kodamana & B. Huang. 2019. Gaussian process modelling with Gaussian mixture likelihood. *J. Process Control* **81**, 209-220.

24. K. Yokohama, J.S. Lord, P. Murahari, K. Wang, D.J. Dunstan, S.P. Waller, D.J. McPhail, A.D. Hillier, J. Henson, M.R. Harper, P. Heathcote, & A.J. Drew. 2016. The new high field photoexcitation muon spectrometer at the ISIS pulsed neutron and muon source. *Rev. Sci. Instrum.* **87**, 125111.

24. K. Wang, P. Murahari, K. Yokoyama, J.S. Lord, F.L. Pratt, J. He, L. Schulz, M. Willis, J.E. Anthony, N.A. Morley, L. Nuccio, A.J. Misquitta, D.J. Dunstan, K. Shimomura, I. Watanabe, S. Zhang, P. Heathcote & A.J. Drew, A.J. 2017. Temporal mapping of photochemical reactions and molecular excited states with carbon specificity. *Nature Materials* **16**, 467-473.

26. A.R. Goñi, R.K. Strössner, K. Syassen & M. Cardona. 1987. Pressure dependence of direct and indirect optical absorption in GaAs. *Phys. Rev.* B**36**, 1581-1587.

27. P. Perlin, W. Trzeciakowski, E. Litwin-Staszewska, J. Muszalski, & M. Micovic. 1994. The effect of pressure on the luminescence from GaAs/AlGaAs quantum wells. *Semicond Sci Technol.* **9**, 2239-2246.

28. M.D. Frogley, J.L. Sly & D.J. Dunstan. 1998. Pressure dependence of the direct band-gap in tetrahedral semiconductors. *Phys. Rev.* B **58**, 12579-12582.

29. H.J. McSkimin, A. Jayaraman, A. & P. Andreatch. 1969. Elastic moduli of GaAs at moderate pressures and the evaluation of compression to 250 kbar. J. Appl. Phys, **38**, 2362.

30. A.C. Torres-Dias, T.F.T. Cerquiera, W. Cui, M.A.L. Marques, S. Botti, D. Machon, M.A. Hartmann, Y.W. Sun, D.J. Dunstan & A. San-Miguel. 2017. From mesoscale to nanoscale mechanics in single-wall carbon nanotubes. *Carbon* **123**, 145-150.

31. V. Amrhein, S. Greenland, and B. McShane. 2019. Retire statistical significance. *Nature* **567**, 305-307.

32. LHCb collaboration. 2021. Test of lepton universality in beauty-quark decays. arXiv:2103.11769v1 [hep-ex].

33. T.M. Apostol. *Calculus, Volume* II (2nd ed. John Wiley & Sons, 1969).

34. K.S. Vimaleswaran, N.G. Fohouri, and K. Khunti. 2021. Vitamin D and covid-19. BMJ 2021;372:n544.

35. A. Martineau. 2021. Quoted in *The Guardian*, 9th March 2021, https://www.theguardian.com/world/2021/mar/09/vitamin-d-supplements-may-offer-no-covid-benefits-data-suggests (accessed 15/03/21).

Acknowledgements: We thank Brian Smith, Ian Bell and colleagues at Renishaw plc, and Adrian Martineau and Yize Wan of the School of Medicine and Dentistry at QMUL for discussions. Graham Thompson, Adrian Bevan and Jonathan Hays of the Particle Physics

Research Group at QMUL assisted with ML methods. A.J.D. acknowledges funding from the European Research Council, from the project MuSES.

Author Contributions: D.J.D. initiated this study and completed it. J.C. found the Bayes factor literature and implemented many of the calculations. A.J.D. provided the muon example, and critically discussed all results presented. All authors contributed to the final manuscript.

Competing Interests: The authors declare no competing interests.

Figure Legends

Fig.1. Muon-spin spectroscopy. Data from an experiment, muon polarisation as a function of magnetic field,[24] is shown in (a). Error bars on the data are estimated at ± 0.015. Linear background functions due to positrons have already been subtracted from the data. The blue solid-circle datapoints (●) were recorded in the dark, while the red open-square datapoints (□) were photo-excited. The blue and red solid lines show 19-parameter fits of three Lorentzian peaks and two linear backgrounds, separately for the data in the dark (blue solid line) and photo-excited (red chain-dotted line). In (b), the evolution of the figures of merit of the fit with the number *n* of fitting parameters is shown (■ SBIC, × lnMLI). The open or small data points from three to seven parameters are for a single peak. The solid or large datapoints from five to 16 parameters are for two peaks, and from 18 to 20 parameters for three peaks.

Fig.2. GaAs Band-Gap. Data for $E_g(P)$ in GaAs from Goñi *et al.*[26] (■) and from Perlin *et al.*[27] (●) are shown after subtraction of the straight line $E_0 + 9.5P$ to make the curvature more visible. The Perlin data is expanded ×10 on both axes for clarity. Two least-squares fits to the Goñi data are shown, polynomial (dashed red line) and Murnaghan (solid blue line).

Fig.3. Carbon Nanotube Raman Spectrum. In (a), the carbon nanotube Raman spectrum are plotted (black datapoints) with a fit (cyan solid line) using the Fourier model. The residuals for four good fits are shown, ×10 and displaced successively downwards (Fourier, Polynomial, Peaks and Tails; see text). In (b), the evolution of the MLIs is shown against the number of parameters for these four models.

Tables

Table I. Photosensitivity results of fitting the data of Fig.1a with 10, 16 and 19 parameters. Parameter units as implied by Fig.1a.

	$\Delta_L P_1$	$\Delta_L W_1$	$\Delta_L A_1$	$\Delta_L P_2$	$\Delta_L W_2$	$\Delta_L A_2$
10-parameter fit	-	-	-	-14 ± 4	21 ± 6	-9 ± 2

				$p = 0.002$	$p = 0.003$	$p = 0.0002$
16-parameter fit	–5 ± 3.5 $p = 0.18$	12 ± 8 $p = 0.16$	–7 ± 5 $p = 0.20$	–24 ± 8 $p = 0.008$	16 ± 13 $p = 0.24$	–5 ± 6 $p = 0.45$
19-parameter fit	–6 ± 2.8 $p = 0.07$	14 ± 6.2 $p = 0.05$	–9 ± 3.8 $p = 0.05$	–25 ± 4.8 $p = 0.0006$	10 ± 9 $p = 0.3$	–2.3 ± 4 $p = 0.6$

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- BayesSciRepSuppInfosubmission.docx