Special Issue Article

JSA

J Strain Analysis
2016, Vol. 51(1) 63–71

DOI: 10.1177/0309324715575445
sdj.sagepub.com
SAGE

Model tuning in engineering: Uncovering the logic

Katie Steele[1] and Charlotte Werndl[1,2]

Abstract

In engineering, as in other scientific fields, researchers seek to confirm their models with real-world data. It is common practice to assess models in terms of the distance between the model outputs and the corresponding experimental observations. An important question that arises is whether the model should then be 'tuned', in the sense of estimating the values of free parameters to get a better fit with the data, and furthermore whether the tuned model can be confirmed with the same data used to tune it. This dual use of data is often disparagingly referred to as 'double-counting'. Here, we analyse these issues, with reference to selected research articles in engineering (one mechanical and the other civil). Our example studies illustrate more and less controversial practices of model tuning and double-counting, both of which, we argue, can be shown to be legitimate within a Bayesian framework. The question nonetheless remains as to whether the implied scientific assumptions in each case are apt from the engineering point of view.

Keywords

Tuning, updating, confirmation, evidence, Bayesianism

Date received: 29 October 2014; accepted: 6 February 2015

Introduction

The practice of *model tuning* (sometimes also referred to as *calibration* or *model updating*) – estimating the free parameters of models that represent the world to get a better fit with empirical observations – perplexes scientists, engineering researchers being no exception. On one hand, it is surely good scientific practice to admit improvements of one's theory (in this case a theoretical model of some phenomena) in response to the data. This is part and parcel of the empiricist approach championed in science. On the other hand, there is an uneasiness about model tuning being ad hoc rather than principled, and moreover, there is a worry that tuning compromises the logic of scientific testing, leading to data being *double-counted* in the broader scientific enterprise of theory construction and theory confirmation. In particular, a common worry is that it is illegitimate to use the same data both to tune the free parameters of a model and to confirm the model and this is often referred to as the illegitimacy of double-counting (the underlying intuition seems to be that data for confirmation need to be new; if they are not because they have already been used to estimate the parameters, then the data will lead to wrong judgements about whether or not a model is confirmed).

Reconciling these conflicting reactions to model tuning, particularly in the context of engineering research, is the problem we address in this article. We analyse two example studies (one from mechanical engineering and the other from civil) that involve more and less controversial practices of model tuning and so-called double-counting. We argue that both cases of model tuning can be shown to be legitimate in a Bayesian framework for assessing scientific theories/hypotheses in the light of evidence. The Bayesian rationalisation reveals, however, just what scientific commitments underlie these specific cases of model tuning; it is a question for engineering researchers as to whether the scientific commitments in each case are apt.

The article proceeds as follows: section 'The Bayesian wisdom on double-counting' gives an overview of model tuning and confirmation in the Bayesian context. We introduce basic Bayesian logic of hypothesis/theory assessment and the kind of hypothesis space that allows us to conceptualise model tuning. The Bayesian framework further permits a distinction between two different kinds of 'double-counting', or in

[1]London School of Economics, London, UK
[2]University of Salzburg, Salzburg, Austria

Corresponding author:
Charlotte Werndl, University of Salzburg, Franziskanergasse 1, Salzburg 5020, Austria.
Email: c.s.werndl@lse.ac.uk

other words, two different senses in which data are used for both tuning and confirmation. Sections 'Example 1: double-counting I' and 'Example 2: double-counting I and II' turn to the Bayesian analysis of our engineering case studies.

The Bayesian wisdom on double-counting

A Bayesian framework helps to make model tuning less mysterious and clarifies the issue of double-counting. The reason for appealing to a Bayesian framework is that it is a comprehensive yet simple logic for assessing hypotheses or theories in the light of evidence. In other words, the Bayesian framework is a very plausible logic of *confirmation* of hypotheses, where confirmation has to do with the amount of confidence that is warranted in a hypothesis. Note that the Bayesian approach to confirmation is not uncontroversial, with so-called Classical approaches being the main contenders; see Howson and Urbach[1] for discussion (there are also divisions within the Bayesian camp, but these detailed disputes about confirmation logic need not concern us here). The divisions become rather nuanced for the confirmation of hypotheses pertaining to models, which is the kind of setting that concerns us here; see Burnham and Anderson[2] for a detailed overview of this special field of confirmation known as *model selection theory*. We do not here argue for a particular stance in these disputes. Our aim is rather to demonstrate how researchers can analyse the practice of model tuning in terms of a plausible confirmation logic, in this case Bayesian.

This section outlines the basics of Bayesian confirmation and provides a very general way of conceptualising model tuning in Bayesian terms (section 'Basic Bayesian updating'). We then examine the issue of double-counting evidence in the light of this Bayesian conceptualisation (sections 'Double-counting I: Tuning' and 'Double-counting II: confirmation'). Our presentation of the Bayesian framework that grounds the discussion of tuning will be similar to the one in Steele and Werndl[3] (pp. 615–618), which addresses tuning as well as other issues of confirmation in climate science.

Before we proceed, a few clarifications about the meaning of the words 'confirmation' and 'validation' are in order. As just mentioned, the Bayesian framework and this article are concerned with the question to what extent a given hypothesis is supported by the data, and in the philosophy literature this is standardly referred to as the question of *confirmation*.[1,4] Note that it is quite general: admittedly, the Bayesian framework, as employed here, concerns the relative standing, in terms of truth, of hypotheses associated with distinct scientific models. But this is not to say that the question is always which model faithfully represents the world in all respects. Model hypotheses may be specified in a variety of ways; the model hypotheses at issue may merely posit adequacy of the model for a particular purpose or application.[5] Needless to say, confirmation is also of central importance in engineering. Hence, many engineering articles, for example, the case study of Brownjohn et al.[6] discussed later, are concerned with confirmation, but the term confirmation is less widely used in engineering than in philosophy. *Validation* is a term that frequently appears in the engineering literature. Some equate model confirmation with validation. Indeed, our discussions (for instance, with workshop participants at the 2011 International Workshop on the Validation of Computational Mechanics Models at the British Museum hosted by the VANESSA Project) with engineers back this up, and Oreskes et al.[7] comment that what is often most relevant is confirmation: 'what typically passes for validation and verification is at best confirmation' (p. 643). Some also consider validation as different from confirmation. In particular, sometimes validation is understood as verification, that is, as showing that a certain claim is true. However, it is impossible to verify hypotheses about models in science because such hypotheses can never be shown to be true with certainty, and hence verification is only of limited relevance.[7] Notwithstanding these issues about the meaning of words, for what follows all that matters is that we have clarified what we mean by confirmation, and that confirmation, thus understood, is also an important topic in engineering.

Basic Bayesian updating

As noted, the Bayesian framework is essentially a logic for adjusting confidence in the truth of hypotheses in the light of data/evidence. It is effectively a special application of probability logic. The guiding question is as follows: what is the bearing of evidence E on the truth of hypothesis H? The Bayesian response involves a probability function over a proposition space that includes H and E, where the probabilities represent degrees of confidence in the truth of the respective propositions (where a belief of 0.7 in H shows stronger confidence in the truth of H than does a belief of 0.3 in H).

Let us now first focus on *comparative confirmation*, that is, on the case where one hypothesis H is compared with another hypothesis H^*. The following equations state the Bayesian rule for adjusting confidence when comparing hypotheses. It gives an expression for the ratio of the degrees of confidence in H relative to H^* having observed E. This expression is referred to as the ratio of *posterior* probabilities (where the relevant probability function is denoted Pr_{post} below). The new or posterior ratio is effectively the ratio of the *prior* conditional probabilities of H given E and H^* given E (where the relevant probability function is denoted Pr below). This latter expression can be expanded as follows

$$\frac{Pr_{post}(H)}{Pr_{post}(H^*)} = \frac{Pr(H|E)}{Pr(H^*|E)} = \frac{Pr(E|H)}{Pr(E|H^*)} \times \frac{Pr(H)}{Pr(H^*)} \qquad (1)$$

The above implies that if the new evidence E is more probable given H as compared to H^* (these conditional probabilities are referred to as *likelihoods*), then learning E increases the probability or confidence in H as compared to H^*. This is arguably the core tenet of Bayesian confirmation logic.

The shift from a prior probability function to a posterior probability function in line with Bayes' rule, as per the expression above, is referred to as Bayesian updating or Bayesian conditionalisation. In the case that the ratio of the probability of H relative to the probability of H^* increases upon updating, we say that H is *confirmed* relative to H^* by E. If the ratio decreases upon updating, we say that H is *disconfirmed* relative to H^* by E.

Let us now focus on *non-comparative confirmation*. Here, the question is whether a certain hypothesis H is confirmed *tout court* (and not just relative to some specified other hypotheses). Arguably, in many circumstances what scientists and decision-makers are interested in is non-comparative confirmation. In the Bayesian framework, non-comparative confirmation amounts to comparing the hypothesis H with its complement, $\neg H$, which refers to all other alternatives distinct from H. Although this formally means that H is compared to all other possibilities, this should not deter from the fact that what is determined is whether H is confirmed *tout court*. The following equations state the Bayesian rule for adjusting confidence for non-comparative confirmation

$$\frac{Pr_{post}(H)}{Pr_{post}(\neg H)} = \frac{Pr(H|E)}{Pr(\neg H|E)} = \frac{Pr(E|H)}{Pr(E|\neg H)} \times \frac{Pr(H)}{Pr(\neg H)} \tag{2}$$

Again, this shift from a prior probability function to a posterior probability function is referred to as Bayesian updating or Bayesian conditionalisation. If this shift due to E involves an increase in the unconditional probability for the hypothesis H (i.e. the ratio of probabilities for the hypothesis H relative to $H^* = \neg H$ increases), then we say that H is *confirmed*. If this shift due to E leads to a decrease in the unconditional probability for the hypothesis H, then we say that H is *disconfirmed*.

The key to conceptualising model tuning in Bayesian terms is to depict the hypothesis space as comprising one or more families of hypotheses, each family corresponding to a base model that has one or more 'free parameters' that may take a number of different values. The individual hypotheses within the base-model family correspond to all possible combinations of values for the free parameters. These individual model hypotheses (which we will hereafter simply refer to as 'model instances') make claims about the respective fully specified models that may be true or false – claims about what the models are supposed to represent in the world. This is often merely implicit in research involving models (and the relevant community of researchers understands what information the model is supposed to provide about the world) but it can be useful to make it explicit. As noted above, any of a variety of claims may be analysed regarding the way and the extent to which a model represents the world. For some applications, scientists may only be interested in whether the trend of one particular variable in the model matches the trend of its real-world counterpart. For other applications, a more comprehensive representation of the world may be of concern. Note that the notion of a model's 'fit with the evidence', formalised in the likelihood functions, in turn depends on the claims or hypotheses under consideration.

Notwithstanding this flexibility in defining model hypotheses, some critics of Bayesian methods argue that focusing on the truth of model hypotheses, whatever their content, is misguided. The claim is that scientists are and should be interested, rather, in the long-run predictive accuracy of tuning given base models, as per Classical model selection methods. While an important dispute in confirmation theory, and certainly relevant to the question of how to *best* interpret and justify scientific reasoning about models, including model tuning, it is beyond the scope of this article to engage in this substantial dispute. As noted at the outset, our aim here is more modest: to show how scientists may rationalise practices of model tuning in terms of *one plausible* logic of confirmation.

Given a Bayesian hypothesis space as outlined, the way to think of model tuning is as follows: it is not about constructing a model or changing a model in an ad hoc way, but rather about updating probabilities or levels of confidence for all the model instances in light of the data. One or more of the model instances will emerge as most probably true, or warranting most confidence, given the data. These model instances can therefore be singled out as superior.

In the remainder of this section, we illustrate model tuning in the Bayesian setting in terms of a very simple linear (and later also quadratic) base model. Section 'Double-counting I: Tuning' discusses tuning itself (which we refer to as double-counting I); section 'Double-counting II: confirmation' discusses model confirmation in the process of tuning (which we refer to as double-counting II). The distinction between the two different kinds of double-counting is also made in Steele and Werndl.[3] Both double-counting I and the more controversial double-counting II are shown to be legitimate possibilities within a Bayesian framework.

Double-counting I: tuning

Consider just one base model, L, with a very simple structure: a linear relationship between the variable y and the variable t. We will assume a probabilistic model error term that is distributed normally with standard deviation σ (the error term could also be interpreted as observational error or as a combined term for observational error and model error)

$$L : y(t) = \alpha t + \beta + N(0, \sigma) \quad (3)$$

As mentioned, the Bayesian account of model tuning depends on the assumption that there is a whole *family of specific instances* of the base model L (each specific model instance has particular values for the unknown parameters α and β). For example, when the possible values for α and β are {1, 2, 3, 4}, then the scientist associates with L a (discrete, in this case) set of specific model instances that can be labelled $L_{1,1}$, $L_{1,2}$,... (the subscripts indicate the values for α and β).

A brief comment on our assumptions about the space of model instances and corresponding Bayesian probabilities: we appeal to discrete probability distributions over model instances in our illustrative examples, but our points equally apply to the more realistic continuous cases, described by *probability density functions* (pdfs). (In this case, all individual model instances, that is, points on the continuum of hypotheses, have probability 0, but the shape of the pdf is informative, and intervals or ranges of model instances will have positive probability.) We consider prior probability distributions to represent scientifically plausible beliefs about the hypotheses (combinations of parameter values) in question, prior to learning the real-world data constituting the evidence. One may be concerned that problematic cases will arise, especially if the parameter space is continuous and unbounded and the most appropriate state of belief seems to be complete ignorance or impartiality among possible parameter values. Indeed, the obvious candidate for an *uninformative* prior – the uniform distribution – is not in fact a proper probability distribution in such a case. Notwithstanding the various interesting responses to this problem of representing ignorance over an unbounded parameter space from Bayesian statisticians (see, for instance, Berger[8], and Kass and Wasserman[9]), our own position is that a state of complete impartiality over an unbounded continuum of hypotheses regarding a model parameter is not a plausible state of scientific opinion. In any practical application, scientists will have reason to favour some range of parameter values over values outside this range (and they may also have reason to discriminate further). Of course, it can be very difficult to determine what is the most plausible prior probability distribution over model hypotheses in any given case. We contend, however, that priors should be selected according to reasons pertaining to background scientific knowledge, rather than in an ad hoc way, and where the uniform distribution is appropriate, it will range over a bounded space of parameter values. This moreover corresponds to the way many scientists who explicitly employ the Bayesian model to represent their reasoning proceed.

Tuning of L amounts to comparing specific instances of the base model –$L_{1,1}$, $L_{1,2}$,... – with respect to the observed values for $y(t)$ (the data). Note that, strictly speaking, what is compared are model hypotheses that postulate that the model in question accurately describes the data generation process for $y(t)$. Note also that, given the probabilistic error term, the data cannot falsify any of the model instances $L_{1,1}$, $L_{1,2}$,..., even if the data are very far away from the specified line.

Model tuning, understood in this way, may well also lead to confirmation of $L_{i,j}$, say, with respect to $L_{k,l}$. As noted above, the extent of confirmation depends on the *likelihood ratio*: $Pr(E|L_{i,j})/Pr(E|L_{k,l})$, where $Pr(E|L_{i,j})$ is the probability, Pr, of the observed data points or evidence, E, given the model $L_{i,j}$. (If one wants to be as precise as possible, one should also state the background knowledge B in the likelihood expressions. That is, they should read $Pr(E|L_{i,j}\&B)$. In the interests of readability, we will refrain from using these more precise expressions.) If the likelihood ratio is greater than 1, $L_{i,j}$ is confirmed by the evidence relative to $L_{k,l}$, and if the likelihood ratio is less than 1, $L_{i,j}$ is disconfirmed by the evidence relative to $L_{k,l}$. In the special case when the likelihood ratio equals 1, neither hypothesis is confirmed relative to the other. Note that, even if $L_{i,j}$ is disconfirmed by the evidence E relative to $L_{k,l}$ (i.e. $L_{i,j}$ is less confirmed by the evidence E than is $L_{k,l}$), it may still be the case that both models are nonetheless confirmed relative to their respective complements.

The value of the ratio of *posterior* (post-evidence) probabilities of $L_{i,j}$ and $L_{k,l}$ is a further matter that depends also on their relative *prior* (initial) probabilities (in the next subsection, we will not refer to ratios but rather to the prior and posterior probabilities for model instances conditional on the base model being true; these are just two ways of saying the same thing). Where the prior probability distribution over model instances given a base model is *uniform*, the ratios of posterior probabilities for all pairs of model instances are equivalent to the corresponding likelihood ratios. This is to say that the posterior probability distribution over the model instances, conditional on the base model, has the same shape as the distribution of likelihoods (the distribution of probabilities representing fit with the data for each respective model instance). Both are peaked on the model instance that has best fit with the data. If the prior probability distribution over the model instances in question is otherwise non-uniform, then the ratios of posterior probabilities will be a trade-off between the corresponding likelihood ratio and the prior probability ratio, as per the Bayesian rule stated above.

We begin with this case to show that a certain kind of double-counting, that is, using the same data for both tuning and confirmation, is clearly implied by the Bayesian depiction of tuning: the whole point is to determine which model instances are confirmed relative to others in light of the data. Call this *double-counting I*; it is synonymous with tuning and can be rationalised in Bayesian terms.

Double-counting II: confirmation

The philosopher of science John Worrall[4] suggests that the kind of double-counting that is illegitimate is when

evidence is used to calibrate a base model such as L above and the same evidence is also used to confirm the base-model hypothesis itself (and not only specific instances of this base model relative to others):

> What those who thought that they were criticizing the 'UN' or 'no-double-use' rule were really doing was pointing out that the same manoeuvre – of using data to fix parameter values or particular theories within a given general framework – [...] is often also used positively within general theoretical frameworks. The manoeuvre will seem positive when the general framework that is being presupposed is supported independently of the particular data being used. [...] But, however positive the manoeuver looks, the evidence involved indeed does not – cannot! – supply any further support for the general framework. Instead that evidence simply (though importantly) transfers the support enjoyed by the general framework theory to the particular theory thus deducted from the evidence plus the general theory. (p. 144)

Bayesian confirmation logic, however, does not support this general position. Let us first focus on comparative confirmation. Here, the Bayesian framework tells us that *double-counting II is legitimate and can arise for two reasons: (1) 'average' fit with the evidence may be better for one base model relative to another and/or (2) the specific instances of one base model that are favoured by the evidence may be more plausible than those of the other base model that are favoured by the evidence* (cf. Howson[10]).

To illustrate this, let us introduce a second base-model hypothesis, a quadratic of the form

$$Q : y(t) = \alpha t^2 + \beta + N(0, \sigma) \tag{4}$$

Assume that the specific model instances are all given by combinations of α and β, where each may take any value in the set {1, 2, 3, 4}. The error standard deviation, σ, is again assumed to be fixed. $Q_{1,1}$, $Q_{1,2}$ and so on will denote specific model instances.

In the Bayesian framework, the confirmation of one base-model hypothesis, for example, L, with respect to another, for example, Q, is given by the likelihood ratio $Pr(E|L)/Pr(E|Q)$. As before, if the ratio is greater than 1, then L is confirmed relative to Q, and vice versa. (Note that the relative posterior probabilities of L and Q, that is, $Pr(L|E)/Pr(Q|E)$, depend also on their prior probability ratio.) However, the relevant likelihoods are not entirely straightforward. They are given by

$$\begin{aligned} Pr(E|L) = {} & Pr(E|L_{1,1}) \times Pr(L_{1,1}|L) + \ \ldots \ + \\ & Pr(E|L_{4,4}) \times Pr(L_{4,4}|L), \\ Pr(E|Q) = {} & Pr(E|Q_{1,1}) \times Pr(Q_{1,1}|Q) + \ \ldots \ + \\ & Pr(E|Q_{4,4}) \times Pr(Q_{4,4}|Q) \end{aligned} \tag{5}$$

Here, $Pr(L_{1,1}|L)$ is the prior probability (i.e. the probability before the evidence is received) of $y(t) = t + 1 + N(0, \sigma)$ being the true description of the data generation process for $y(t)$, given that the true model is linear. (In section 'Double-counting I: Tuning', we noted that these conditional prior probabilities for model instances are necessary for determining their conditional posterior probabilities.) The equations above are the formal expressions of our earlier statement that confirmation of base models depends on (1) fit with the data and (2) the conditional priors of all specific instances of these base models.

Let us first turn to the special case where the conditional prior probabilities of all specific instances of L and Q are equivalent

$$\begin{aligned} & Pr(L_{1,1}|L) = \cdots = Pr(L_{4,4}|L) = \cdots \\ & = Pr(Q_{1,1}|Q) = \cdots = Pr(Q_{4,4}|Q) = x \end{aligned} \tag{6}$$

Suppose that the observed data E yield on balance greater likelihoods for instances of L than Q. Then *there is confirmation of L relative to Q due to reason (1). That is, the 'average' fit with the evidence is better for base-model hypothesis L than for base-model hypothesis Q. Moreover, there is tuning* because the evidence E is used to determine the most likely parameter values of α and β.

Another special case arises when the base models would have equivalent fit with the data if all specific models were weighted equally, but the prior probabilities over the model instances are in fact not equal. To be more specific, suppose that the specific instances of L that have the higher likelihoods for E have higher conditional priors (i.e. are considered to be more plausible) than the specific instances of Q that have the higher likelihoods. Then *there is confirmation of L relative to Q because of reason (2). That is, the specific instances of L that are favoured by the evidence are more plausible than the specific instances of Q that are favoured by the evidence. Moreover, again, there is tuning* because E is used to determine the most likely parameter values of α and β. Next to these two special cases, there is also the case of double-counting that arises through a combination of (1) and (2).

Finally, let us turn to non-comparative confirmation. Here, the question is whether a hypothesis H is confirmed *tout court* (and not just relative to some other specified hypotheses). Recall that in the Bayesian framework non-comparative confirmation amounts to comparing the hypothesis H with its complement, $\neg H$, that refers to all other possible models distinct from the base model of H. Often, $\neg H$ cannot be specified as a set of explicit alternative model hypotheses, but rather involves an unspecified *catch-all* hypothesis. We will presume that $\neg H$ involves a catch-all here, as this is a common scenario. One can easily see that the confirmation of H relative to $\neg H$ in light of evidence E is difficult to assess in such a case. The relevant likelihood ratio is

$$\frac{Pr(E|H)}{Pr(E|\neg H)} \tag{7}$$

The likelihood $Pr(E|H)$ is calculated as above and depends on both conditional priors for model instances and their respective fit with the evidence. The problem that arises is that the likelihood associated with the hypothesis involving a catch-all, $Pr(E|\neg H)$, is difficult to evaluate. How should one estimate the probability of some evidence conditional on the truth of a hypothesis which we cannot fully specify? Determining the comparative likelihood ratio where a catch-all is concerned is indeed rather speculative. Yet, we think that often when a model hypothesis accords very well with evidence, scientists presume that the complement of the hypothesis would not accord so well with the evidence and that a rough estimate can be given of $Pr(E|\neg H)$. Hence, many studies still proceed under the assumption that base model hypotheses may be confirmed (or disconfirmed) to some degree in non-comparative terms, given evidence. In any case, what is relevant for this article is that *the conclusions about double-counting II carry over to non-comparative confirmation. Just like for comparative confirmation, double-counting II is legitimate and can arise because (1) the 'average' fit with the evidence is very good (and presumed to be better than the fit for $\neg H$) and/or (2) the specific instances of the base model that are favoured by the evidence are very plausible (and presumed to be more plausible than specific instances of $\neg H$).*

Worrall[4] discusses cases where data seem to be used for tuning and confirmation of a base-model hypothesis. He argues that what really happens in these cases is that only some of the data are needed to determine the values of the initial free parameters and that the rest of the data then confirm the hypothesis. Thus, he argues, there is no double-counting. However, this splitting of the data does not fit all cases; in particular, for model instances with stochastic error terms, splitting the data can throw away valuable information about the free parameters and is not in keeping with Bayesian logic of confirmation. Rather, as we have seen for our simple example cases, *all the data are used to determine the values of the free parameters and at the same time all the data are used to confirm base-model hypotheses*, and thus these are genuine cases of double-counting.

Example 1: Double-counting I

Let us now analyse our first engineering case study involving model tuning, and hence double-counting I. The aim is to show that the tuning can be rationalised in Bayesian terms, but that various scientific assumptions are thereby brought to light.

Consider a simple tie bar with a hole at the centre to which a load is applied. The distribution of stress is a function of the geometry, applied load and the material properties. When the applied loads are small, the plate will behave elastically. That is, if the load is released, the plate will return to its original shape, and for most engineering materials the stress varies linearly with the applied load. However, when the applied load exceeds a critical value, the plate behaves plastically, that is, the stress varies non-linearly with applied load and the deformation is permanent. While the elastic case is easy to model, the *plastic case* is difficult and often involves tuning of parameters.

So our case study of double-counting I will be about the plastic case, namely, an article by Wang et al.[11] The base model Wang et al. are using involves the fundamental relationship between stress and strain and has five free parameters, describing the material properties that define the elastic–plastic stress/strain curve and the geometric dimensions of the tie bar. More specifically, the parameters that need to be tuned are the Young's modulus, the initial yield stress, the plastic strain corresponding to 285 MPa, the plastic strain corresponding to the ultimate stress (at 310 MPa) and the thickness of the plate. For Wang et al., the base model is not in question and they want to find out the correct values of the free parameters.

The data that are used to estimate the values of the free parameters are full-field measurements of strain. They were obtained from a quasi-static tensile test using a digital image correlation system, where the tensile load was applied in small monotonic increments from 0 to failure. To measure the distance between model instances and observed data, the cosine distance was used (i.e. the cosine value of the angle between two vectors; unity cosine distance denotes collinear vectors and zero cosine distance is for perpendicular vectors; the higher the distance, the better) and additionally the length of the vectors was compared. The best parameter values are those that maximise the cosine distance and where the length of the vectors is as close as possible.

To determine the best parameter values, finite element modelling was applied, where one starts with initial values that are taken from the literature from previous experimental results or derived from theoretical considerations. The initial values of the Young's modulus, the initial yield stress, the plastic strain corresponding to 285 MPa, the plastic strain corresponding to the ultimate stress (at 310 MPa) and the thickness of the plate that Wang et al. started with were 69 GPa, 250 MPa, 0.008 m/m, 0.075 m/m and 1 mm, respectively. Through an iterative procedure, Wang et al. arrived at the values that lead to model instances that are closest to the data, namely, 77 GPa, 180 MPa, 0.0072 m/m, 0.074 m/m and 1.2 mm, respectively.

While Wang et al. do not apply the Bayesian framework, their model-tuning procedure can be reconstructed in it. More specifically, on a Bayesian reconstruction, the free parameters are those whose precise values are to some extent uncertain. The various combinations of free parameter values are the model instances under consideration. The distance between model output and data, for any model instance, can be translated into a probability measure of the data given

the model (the greater the distance between model output and data, the lower the likelihood). For instance, Gauss proved that a Gaussian probabilistic error distribution for model predictions of individual data yields a likelihood function that matches the least-squares assessment of 'relative fit with the data' of model instances.[12] Wang et al.'s assessment of model instances depends entirely on their fit with the data because the iterative procedure finds the model instance that is closest to the data – or at least approximately so given the numerical method. (A difference between Wang et al.'s procedure and the Bayesian analysis is that the former uses finite element modelling to arrive at an estimate of the best fitting model instance, or combination of parameter values, based on a set of initially plausible values, while in the Bayesian analysis all the infinite number of possible combinations of parameter values are compared to arrive at the best fitting one. While this is a difference in the way the best fitting model instance is determined, there is no difference in substance.) In the Bayesian setting, model instances are ultimately assessed in terms of their relative posterior probabilities. As noted in section 'Double-counting I: Tuning', the relative posterior probabilities for (a bounded set of) model instances match their relative likelihoods only when the prior probabilities of the model instances are equal. This is to say that Wang et al.'s implicit prior/initial probability distribution over the combinations of parameter values is effectively a flat one. In sum, if we assume a flat prior probability distribution over a suitable range of model instances and appropriate likelihoods (probability measures for the data given the model instances), then the Bayesian analysis yields a posterior probability distribution over the model instances that matches Wang et al.'s result, that is, a distribution for the model instances given the data that is peaked around the model instance that is 'closest' to the data.

At this point, a comment is in order. In a Bayesian analysis of tuning as outlined in section 'The Bayesian wisdom on double-counting', the result obtained is a probability distribution over the model instances, that is, combinations of free parameter values, given the data. In contrast to this, Wang et al. just present the best model instance and not a probability distribution over the space of model instances. In Bayesian terms, the selection of one model instance at the end presumably amounts to choosing the most probable one. This is a further step that goes beyond Bayesian logic per se; it is nonetheless the standard final output in model selection theory (cf. Burnham and Anderson[2]). While this further step lies outside of Bayesian logic, the model-tuning procedure of Wang et al. can nonetheless be understood and justified, as just explained, in terms of Bayesian updating.

Finally, let us return to the issue of the flat prior distribution over candidate model instances. As noted, this makes the model selection dependent on fit with the data; indeed, Wang et al. replace the theoretical estimates of the free parameter values with the combination of estimates that gives closest fit with the data. Some might argue that this gives too much attention to the data and amounts to simply discarding the pre-experimental estimates of the parameter values. A Bayesian analysis illuminates these worries by making salient alternative possibilities for carrying out the tuning. Those who criticise the overriding role of the data in this tuning procedure are questioning this flat prior probability distribution over model instances. They would perhaps argue that the prior distribution over initial combinations of parameter values is not flat because certain combinations are favoured by theoretical considerations. In this case, the best model instances are those that perform best in a trade-off between prior plausibility and fit with the data (as discussed earlier in section 'Double-counting I: Tuning'). Indeed, for the non-elastic stress engineering example, a flat prior distribution seems somewhat counterintuitive as prior reasoning seems to privilege a particular set of parameter values that are derived from theoretical considerations.

Example 2: Double-counting I and II

Let us now discuss an example of double-counting I and II. Singapore has a continuing programme of highway bridge upgrading for refurbishing and strengthening bridges to allow for increasing traffic. In the context of this programme, the need arose to come up with a reliable model of Pioneer bridge (a bridge taking a busy main road across a coastal inlet near a major port facility) before and after refurbishing it. Brownjohn et al.[6] attempt to come up with such a reliable model and hence their concern is non-comparative confirmation, that is, the question to what extent their model is confirmed by the data. In what follows, we concentrate on their model of the bridge before the refurbishment work and the case of the bridge simply supported on abutment.

Brownjohn et al. model the bridge using three-dimensional (3D) beam elements supported on abutments. The free parameters of the model that were tuned in response to the data are the *vertical bending stiffness of the T-beams* EI_T and the *vertical bending stiffness of the diaphragms* EI_D. So their base model hypotheses are as follows:

H: model instances hypotheses of the bridge model using 3D beam elements supported on abutments.
$\neg H$: catch-all.

The data that were used to estimate the free parameters are the first five *natural frequencies of the bridge* (a bridge has the tendency to undergo resonance, that is, to oscillate at larger amplitudes at certain frequencies, and these are called the natural frequencies of the bridge). Finite element modelling was used to find the values of the parameters that fit the data best, where

the best model was understood to be the model with a minimum least square distance between the observed and the modelled data. Finite element modelling starts with initial values taken from previous experimental results in the literature or from theoretical considerations (in this case $EI_T = 1.06\times10^9\,\mathrm{Nm}^2$ and $EI_D = 4.9\times10^7\,\mathrm{Nm}^2$), and then uses a numerical procedure to arrive at the parameter values that fit the data best. This tuning procedure yielded the following parameter values: $EI_T = 1.32\times10^9\,\mathrm{Nm}^2$ and $EI_D = 6.87\times10^7\,\mathrm{Nm}^2$.

Brownjohn et al. also use the same data about the natural frequencies of the bridge to confirm the base model H in non-comparative terms. More specifically, consider the following comments by Brownjohn et al.:[6] 'While FEM is an efficient analytical tool, creating a model capable to reproduce the measured dynamic characteristics of the prototype is a challenge' (p. 168), followed later by 'The results of updating are summarized in Table 3, with pairing of experimental and analytical frequencies shown in Figure 11. Significant improvement in agreement between the first five analytical frequencies and their experimental counterparts can be seen after updating' (p. 169),[6] and 'Beyond that, a FE model of the bridge, before and after the upgrading, is now available that can estimate load carrying capacity more reliably than would a pure desk exercise' (pp. 170–171).[6] Comments such as these ones are suggestive of Brownjohn et al. regarding their model as being confirmed in non-comparative terms.

While Brownjohn et al. do not employ the Bayesian framework, their procedure can be roughly reconstructed in Bayesian terms as follows. To begin with, the free parameters are those whose values are uncertain, at least within some range. Indeed, Brownjohn et al.[6] (p. 168) make a comment to this effect (note that in the end they do not include the Young's modulus as a free parameter; this is because the type of vibrations considered is governed by vertical bending stiffnesses (EI_T and EI_D), and while Young's modulus (E) is a contributing factor to these stiffnesses, alongside the second moments of area (I), they always appear together as bending stiffness ($EI = E^*I$) since attempts to decouple them often lead to ill-posed problems with non-unique solutions):

> It seems a truism that the parameters for updating must be uncertain in the model. For example, Young's modulus for concrete is uncertain, while for steel it would be nonsense to look for a small change in a well-established value.

The various plausible combinations of free parameter values correspond to the set of model instances under consideration. As for the first example, the distance between model output and data can be translated into a probability measure of the data given the model instances or combinations of free parameter values (and the greater the distance, the lower the probability). Furthermore, as for the first example, the prior probability distribution over the plausible range of parameter values of the T-beam stiffness and diaphragm stiffness conditional on model M being true is effectively a flat one (because the updating method finds the model instance that is closest to the data). Given this uniform prior probability distribution and the probability measure of the data given the model instances, the result obtained from the Bayesian analysis is a probability distribution over the model instances given H and the data peaked around the parameter values that give the closest fit to the data – see the discussion of double-counting I in section 'Double-counting I: Tuning'. So, just like in the first example, the model-tuning procedure of Brownjohn et al. can be conceptualised in a Bayesian framework: they assess which combination of parameter values for the base model (within some bounded range) is best given the data, arriving at a probability distribution over the parameter values peaked on the combination of parameter values that correspond to the model instance closest to the data.

However, unlike for the first example, there is a further element present in Brownjohn et al.'s analysis. Namely, recall the quotations of Brownjohn et al. given above. These comments suggest that in addition to using the data for tuning, they also seem to use the data for *confirming* the base model in non-comparative terms. In particular, these comments are suggestive of Brownjohn et al. assigning to the probability of the natural frequencies data E given H a much greater value than is assigned to the probability of the data E given $\neg H$ (i.e. $Pr(E|H) > Pr(E|\neg H)$), implying that the base model H is confirmed by the evidence E, that is, the data warrant increased confidence in H relative to its complement $\neg H$. To be sure, Brownjohn et al. do not explicitly assign probabilities to the evidence given the model H and to the evidence given the catch-all $\neg H$. Yet from their claim that there is confirmation we can infer that they are committed to the claim that $Pr(E|H) > Pr(E|\neg H)$.

To conclude, just like in the Bayesian analysis sketched in section 'Double-counting II: confirmation', this second example plausibly illustrates double-counting II because Brownjohn et al. seem to be comparing, in terms of fit with the evidence, their base model hypothesis H with its complement $\neg H$, in this case an unspecified 'catch-all'. In other cases in science, H is compared not to its entire complement but rather to an alternative base model hypothesis H^* or to several alternative base model hypotheses $H1^*$,..., Hn^* (this is the case, for example, in Beck and Yuen[13] and Oh et al.[14]). In all these cases, whether it be a case where H is (dis)confirmed relative to its complement or to a set of explicit alternative hypotheses, we have double-counting II. Note that, as for the first example, in a Bayesian analysis as presented above, the result obtained is a probability distribution over the combinations of parameter values conditional on the base model H and the data. In contrast to this, Brownjohn et al. just present the best values for the free parameters. This is a further step that goes beyond the Bayesian analysis and is common in model selection theory. Even if this further step is taken, Brownjohn et al.'s model-tuning account can

be justified within a Bayesian framework that legitimates both double-counting I and II.

Concluding remarks

We have shown how paradigmatic model-tuning practices in engineering can be understood and justified within a Bayesian framework. In particular, we justified cases of so-called double-counting, both I and II.

Given that tuning (or calibration or model updating) is about comparing the plausibility of various candidate model instances in the light of data, it comes under the domain of confirmation logic. Broadly speaking, the logic of confirmation is about having well-formed and consistent judgments of confidence in hypotheses, before and after learning new evidence. Here, we have appealed to Bayesian logic of confirmation, but, as noted from the outset, we do not pretend that this choice is uncontroversial; one might argue that an alternative logic of confirmation is more appropriate. The question would then arise as to whether and how practices of model tuning in engineering can be squared with any such alternative logic of confirmation. The basic Classical approach, for instance, would insist that it does not make sense to identify the best (most probable) instance of a given base model; rather, tuning can at best falsify or rule out certain combinations of values for free parameters that yield excessively poor fit with the data. Alternative Classical model selection methods explicitly address the greater possibility of 'over-fitting to data', the more free parameters there are in a base model; moreover, these methods focus on the long-run predictive accuracy of tuned base models, as opposed to the truth of hypotheses associated with model instances.

Notwithstanding the nuanced debates in confirmation and model selection theory, the basic point here is as follows: if one wants to argue that a given practice of model tuning/confirmation is justified, one must refer to an appropriate logic of confirmation. Of course, even if the Bayesian logic is the most appropriate guide to scientific reasoning, we are not suggesting that researchers should always frame and report their model updating practices in terms of, say, probability distributions over sets of model instances and so on. The common practice of shifting the values of free (uncertain) parameters so that the model is closer to the data, according to some salient measure of closeness, is an intuitive and appropriate way to proceed in many cases. It is rather that, if one wants to properly argue that tuning is not ad hoc, one must turn to the fundamentals of confirmation. Moreover, in some applications, it may not be obvious what weight should be given to fit with the data as compared to the prior plausibility of various values for free parameters, or even what parameters of a model should be regarded as free; in such cases, there is also practical reason to appeal to a more fundamental confirmation framework.

Acknowledgements

The authors are listed alphabetically; this work is fully collaborative.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Arts and Humanities Research Council (grant number AH/J006033/1) and the Economic and Social Research Council (grant number ES/K006576/1). Katie Steele was also supported by a Research Fellowship at the Swedish Collegium for Advanced Study (September to December 2014).

References

1. Howson C and Urbach P. *Scientific reasoning: the Bayesian approach*. 3rd ed. Chicago, IL: Open Court, 2005.
2. Burnham KP and Anderson DR. *Model selection and multimodal inference: a practical information-theoretic approach*. New York: Springer, 1998.
3. Steele K and Werndl C. Climate models, confirmation and calibration. *Brit J Philos Sci* 2013; 64: 609–635.
4. Worrall J. Error, tests, and theory confirmation. In: Mayo DG and Spanos A (eds) *Error and inference: recent exchanges on experimental reasoning, reliability, and the objectivity and rationality of science*. Cambridge: Cambridge University Press, 2010, pp.125–154.
5. Parker WS. Confirmation and adequacy for purpose in climate modelling. *P Aristotelian Soc* 2009; 83: 233–249.
6. Brownjohn JM, Moyo P, Omenzetter P, et al. Assessment of highway bridge upgrading by dynamic testing and finite-element model updating. *J Bridge Eng* 2003; 8(3): 162–172.
7. Oreskes N, Shrader-Frechette K and Belitz K. Verification, validation, and confirmation of numerical models in the earth sciences. *Science* 1994; 263(5147): 641–646.
8. Berger JO. *Statistical decision theory and Bayesian analysis*. Berlin: Springer, 1984.
9. Kass RE and Wasserman L. The selection of prior distributions by formal rules. *J Am Stat Assoc* 1996; 91(435): 1343–1370.
10. Howson C. Accommodation, prediction and Bayesian confirmation theory. *PSA: Proc Bienn Meet Philos Sci Assoc* 1988; 1988: 381–392.
11. Wang W, Mottershead JE, Sebastian CM, et al. Shape features and finite element model updating from full-field strain data. *Int J Solids Struct* 2011; 48: 1644–1657.
12. Forster MR. Key concepts in model selection: performance and generalizability. *J Math Psychol* 2000; 44: 205–231.
13. Beck JL and Yuen K. Model selection using response measurements: Bayesian probabilistic approach. *J Eng Mech* 2004; 130(2): 192–203.
14. Oh CK, Beck JL and Yamada M. Bayesian learning using automatic relevance determination prior with an application to earthquake early warning. *J Eng Mech* 2008; 134(12): 1013–1020.