British Journal of Mathematical and Statistical Psychology (2013), 66, 45–56

The British Psychological Society

www.wileyonlinelibrary.com

Commentary

Posterior predictive checks can and should be Bayesian: Comment on Gelman and Shalizi, 'Philosophy and the practice of Bayesian statistics'

John K. Kruschke*

Department of Psychological and Brain Sciences, Indiana University, Bloomington, USA

Bayesian inference is conditional on the space of models assumed by the analyst. The posterior distribution indicates only which of the available parameter values are less bad than the others, without indicating whether the best available parameter values really fit the data well. A posterior predictive check is important to assess whether the posterior predictions of the least bad parameters are discrepant from the actual data in systematic ways. Gelman and Shalizi (2013) assert that the posterior predictive check, whether done qualitatively or quantitatively, is non-Bayesian. I suggest that the qualitative posterior predictive check might be Bayesian, and the quantitative posterior predictive check should be Bayesian. In particular, I show that the 'Bayesian *p*-value', from which an analyst attempts to reject a model without recourse to an alternative model, is ambiguous and inconclusive. Instead, the posterior predictive check, whether qualitative or quantitative, should be consummated with Bayesian estimation of an expanded model. The conclusion agrees with Gelman and Shalizi regarding the importance of the posterior predictive check for breaking out of an initially assumed space of models. Philosophically, the conclusion allows the liberation to be completely Bayesian instead of relying on a non-Bayesian *deus ex machina*. Practically, the conclusion cautions against use of the Bayesian *p*-value in favour of direct model expansion and Bayesian evaluation.

1. Introduction

Bayesian inference is conditional on the space of models assumed by the analyst. Within that assumed space, the posterior distribution only tells us which parameter values are relatively less bad than the others. The posterior does not tell us whether the least bad parameter values are actually any good. Assessing the goodness of the least bad parameter values is the job of the *posterior predictive check*. In a posterior predictive check, the analyst assesses whether data simulated from credible parameter values resemble the

*Correspondence should be addressed to John K. Kruschke, Department of Psychological and Brain Sciences, Indiana University, 1101 E. 10th St., Bloomington, IN 47405-7007, USA (e-mail: kruschke@indiana.edu).

DOI:10.1111/j.2044-8317.2012.02063.x

actual data, with 'resemblance' measured in any way that is meaningful in the applied context. If the resemblance is not good enough, then the analyst changes the model and does Bayesian inference on the modified model. This cycle repeats until the resemblance of the predicted data and the actual data is good enough for purposes of the application.

The posterior predictive check allows the analyst to solve the problem of being confined within the initially assumed space of models. Gelman and Shalizi (2012, 2013) emphasized that the posterior predictive check is a non-Bayesian process: 'It is by this non-Bayesian checking of Bayesian models that we solve our … problem' (Gelman & Shalizi, 2013, p. 17). In particular, the goodness of the resemblance, between simulated and actual data, is assayed in either of two non-Bayesian ways, qualitative or quantitative.

In the *qualitative* way of assessing resemblance between simulated and actual data, the analyst can visually examine graphical or tabular displays to look for structured patterns in the residuals between actual and simulated data. If there appears to be structure in the residuals that meaningfully informs the interpretation of the model, then the analyst can change the model so that it better captures the revealed trends. This intuitive assessment uses no explicit, formal Bayesian calculations.

Although intuitive assessment of pattern is not formally Bayesian, some leading theories in cognitive science assert that perception is well described as Bayesian inference. Essentially, these theories propose that the mind has a vast library of candidate perceptible patterns, with a distribution of prior credibilities across those patterns, and the observed residuals are used to infer, in a Bayesian manner, the posterior credibilities of candidate patterns for the residuals. Thus, when we perceive a pattern in the residuals, it is because that pattern has a reasonably high posterior credibility among the various patterns we have available in our perceptual space.

In the *quantitative* way of assessing resemblance between simulated and actual data, the analyst defines a formal measure of the magnitude of discrepancy between observed data y and predicted values $\hat{y}$, denoted $T(y,\hat{y})$. The observed data may be the actual data from the empirical research and denoted y^{act}, or the observed data may be simulated from the model and denoted y^{rep}, where the superscript rep refers to 'replication'. With many replications of data simulated from posterior parameter values, a sampling distribution of $T(y^{rep},\hat{y})$ is created. From that sampling distribution we compute the probability of obtaining a value of T as big as or bigger than the actual one: $p(T(y^{rep},\hat{y}) \geq T(y^{act},\hat{y}))$. This probability is also known as the 'Bayesian p-value'. If the Bayesian p-value is very small, then we reject the model and search for something better. Gelman and Shalizi (2012, 2013) point out that this procedure can reject a model without specifying an alternative model.

While this quantitative process is non-Bayesian, I show that its results are ambiguous and undertaking it is unnecessary. Instead, a Bayesian procedure can yield clearer results. Specifically, the analyst should create a formal model that addresses the perceived discrepancy, and the expanded model can be assessed in a Bayesian fashion. This approach avoids ambiguity in T, which could be a signature of many different underlying structures. The expanded model is assessed by Bayesian parameter estimation, and does not necessarily rely on Bayesian model comparison, which has problems of hypersensitivity to priors (as pointed out by Gelman & Shalizi, 2013).

The rest of this article expands on the two-pronged argument outlined above. Examples of regression analysis are provided to illustrate ambiguous implications from T and p, but clearer conclusions from Bayesian estimation of specific expanded models. The argument agrees that a posterior predictive check is an important step in Bayesian data analysis, but avers that a posterior predictive check need not be inherently non-Bayesian. Whether the check of resemblance is qualitative or quantitative, it should be consummated by a formal

specification of structure in an expanded or new model, with parameters estimated in a Bayesian fashion. Thus, a posterior predictive check can and should be Bayesian.

2. A qualitative posterior predictive check can be Bayesian

In a qualitative posterior predictive check, the analyst displays the original data along with the posterior predictions in some way that highlights potentially systematic discrepancies. The display could be tabular or graphical, and can accentuate or attenuate different aspects of the data and posterior predictions. Regardless of the exact nature of the display, the human analyst must perceive systematic patterns in the discrepancies. What are the possible patterns that a human can perceive? And of all those possible patterns, which ones are most likely to be perceived when observing a display?

Some leading theories in cognitive science describe perception as Bayesian inference (e.g., Kersten, Mamassian, & Yuille, 2004; Knill & Richards, 1996; Shams & Beierholm, 2010; Yuille & Kersten, 2006). According to this theoretical perspective, the mind has a vast repertoire of possible descriptions of the world, with innate or previously learned knowledge providing a prior distribution over that space of perceptible patterns. When new stimuli impinge upon the senses, the mind infers the most likely distal objects that may have produced the sensory stimulus. The inference relies heavily on prior knowledge, and formal Bayesian models have successfully accounted for many aspects of human perception.

One of the simplest examples of prior knowledge deployed in perception is the interpretation of three-dimensional shape from observable shading on the object. Consider Figure 1, which shows two circular regions spanned by gradients of grey. When the light end of the gradient is at the top, we perceive the circular region as a protuberance, but when the light end of the gradient is at the bottom, we perceive the circular region as an indentation. This difference in perceptual interpretation is explained by the mind applying prior knowledge: illumination usually comes from above, as from the sun and sky. As another example, consider the learning of functional relationships between input and output values, such as drug dosage (input value) and symptom severity (output value). People are tasked with learning the relationship between the variables by observing many examples, and then their learned responses are used to teach new learners. After a few generations, the learned and retaught function evolves into a linear relationship, regardless

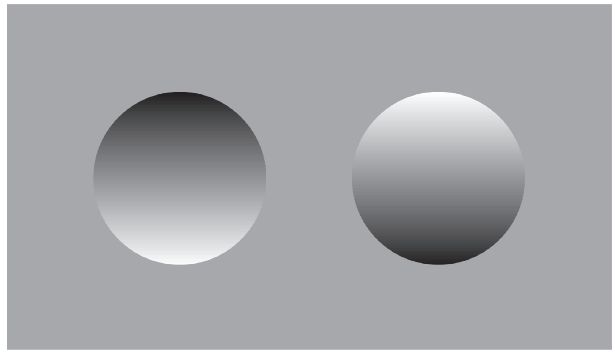

Figure 1. The circular region on the left is perceived as an indentation, while the circular region on the right is perceived as a protuberance, even though the gradients of grey are identical except for orientation. Perception apparently employs prior knowledge that illumination comes from above, and that the surface itself has constant colour.

of how it started, which reveals that linear relations are weighed heavily in learner's prior knowledge (Kalish, Griffiths, & Lewandowsky, 2007). While recent theories have given explicit formal expression to the idea of perception as Bayesian inference, informal theories of perception as inference go back at least to Helmholtz (1867), although it is doubtful that Helmholtz had any explicitly Bayesian notions (Westheimer, 2008).

A variety of other aspects of cognition and learning have been modelled as Bayesian inference (for overviews, see Chater, Tenenbaum, & Yuille, 2006; Jacobs & Kruschke, 2010). Recent work has shown that human perception of accidental coincidences versus causes can be modelled as Bayesian inference (Griffiths & Tenenbaum, 2007), and human interpretation of many different data structures can be modelled as Bayesian inference (Tenenbaum, Kemp, Griffiths, & Goodman, 2011). The issue of how the mind or brain might implement Bayesian inference is one of current discussion and debate. Some theorists suggest that the mind merely approximates Bayesian inference (Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010), but this particular approach may be unsatisfying because many non-Bayesian algorithms are 'approximately' Bayesian without having any necessary relation to Bayesian computation (Kruschke, 2010a). Another approach suggests that the mind might be well described as Bayesian only within certain levels of analysis, while larger-scale behaviour is not (Kruschke, 2006). Whatever the domain or level of analysis, the goal for genuinely Bayesian models of cognition is discovering functional forms and priors that closely mimic human behaviour.

Regardless of the ultimate veracity of any specific Bayesian model of perception or cognition, the point of this section is that intuitive assessment of patterned discrepancies could be Bayesian. There is nothing necessarily non-Bayesian in a qualitative posterior predictive check. On the other hand, I am not claiming that qualitative posterior predictive checking is in fact, or must be, well described as Bayesian inference. Indeed, even if human perception and cognition – the engine of qualitative posterior predictive checking – ultimately proves to be impossible to adequately model as Bayesian inference, it is still appropriate to formally analyse scientific data with Bayesian methods (Kruschke, 2010b), and it is still the case that *quantitative* posterior predictive checking can and should be Bayesian, as the next section illustrates.

3. A quantitative posterior predictive check should be Bayesian

As a concrete example to frame discussion, consider the data displayed in Figure 2. For every individual, we measure a criterion value y that we wish to predict from a value x. The conventional first approach would be simple linear regression with normally distributed noise, expressed formally as

$$\hat{y} = \beta_0 + \beta_1 x, \tag{1}$$

$$y \sim N(\hat{y}, \sigma), \tag{2}$$

where $\hat{y}$ is the predicted value of y for predictor value x, β_0 is the intercept, β_1 is the slope, and σ is the standard deviation of the normal distribution.

For Bayesian estimation of the three parameters in equations (1) and (2), I began with vague priors that had minimal influence on the posterior distribution. The analysis used Markov chain Monte Carlo (MCMC) sampling by JAGS (Plummer, 2003) called from R (R Development Core Team, 2011) via package rjags, with programs created in the style of

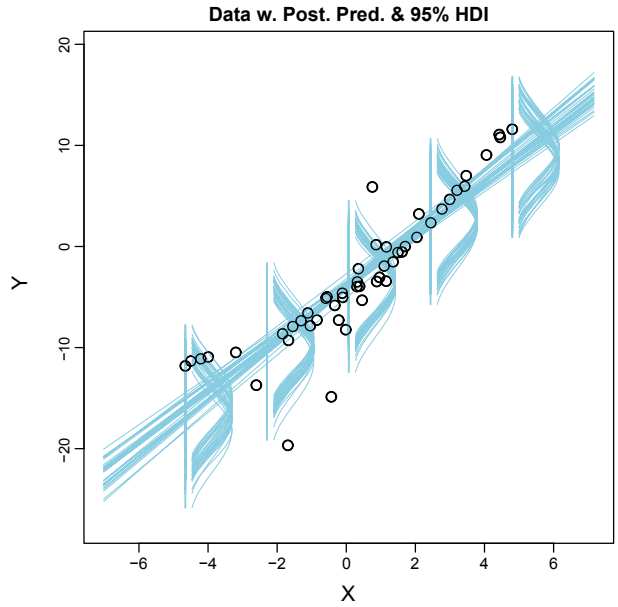

Figure 2. Data with posterior predictions, using linear regression with normally distributed likelihood as defined in equations (1) and (2). The lines extending from left to right show a smattering of credible regression lines from the MCMC chain. The vertical segments show 95% highest density intervals (HDIs) with normal density functions (plotted sideways) having corresponding credible standard deviations. The data appear to be too tightly clustered within vertical slices, relative to the spread of the posterior predicted normal distributions. The data also appear to have a slight non-linear trend.

Kruschke (2011b). Figure 2 shows plots of 30 credible regression lines superimposed on the data. Displayed with each line are sideways plots of a normal distribution with the corresponding standard deviation. The slope, intercept, and standard deviation of the 30 plots came from every (200,000/30)th step in the MCMC chain of 200,000 steps.

Visual inspection of the posterior estimates in Figure 2 suggests at least two discrepancies between data and model. First, the data appear to be too tightly clustered within vertical slices, relative to the spread of the posterior predicted normal distributions. Second, the data also appear to have a slight upward curvature relative to the linear predictions of the model.

Having noticed possible systematic discrepancies between the data and the posterior predictions, what should we do next? One possibility is to create some measure of discrepancy, $T(y, \hat{y})$, that somehow captures the seemingly anomalous discrepancy. The measure T does not need to express an alternative model; it merely needs to quantify the discrepancy. We then generate the sampling distribution of T from the posterior distribution, and assess whether $p(T(y^{rep}, \hat{y}) \geq T(y^{act}, \hat{y}))$ is sufficiently small that we are justified to look for a better model of the data. Gelman and Shalizi (2013, footnote 11) say 'the tail-area probabilities are relevant [because] they make it possible to reject a Bayesian model without recourse to a specific alternative' and 'What we are advocating, then, is what Cox and Hinkley (1974) call "pure significance testing", in which certain of the model's implications are compared directly to the data, rather than entering into a contest with some alternative model' (Gelman & Shalizi, 2013, p. 20).

For example, suppose we want to define a measure of upward curvature for the data in Figure 2. For purposes of defining the measure of discrepancy, we will index the 50 observations from smallest x-value to largest x-value. Thus, $\langle x_1, y_1 \rangle$ is the leftmost point, and $\langle x_{50}, y_{50} \rangle$ is the rightmost point. Upward curvature implies that the left-hand end and right-hand end points tend to be above the linear prediction, while middle points, namely $\langle x_{25}, y_{25} \rangle$ and $\langle x_{26}, y_{26} \rangle$, tend to be below the linear prediction. This signature of curvature could be formalized as, say,

$$T(y, \hat{y}) = (y_1 - \hat{y}_1) + (y_{50} - \hat{y}_{50}) - (y_{25} - \hat{y}_{25}) - (y_{26} - \hat{y}_{26}). \quad (3)$$

Defining T in terms of ranked data has precedents in Gelman, Carlin, Stern, and Rubin (2004). For example, when modelling a set of data with a normal distribution and assessing leftward skew or outliers, one definition for T was simply $T(y, \hat{y}) = y_1 = \min(y)$ (Gelman *et al.*, 2004, p. 160). For the same set of data, another definition for T was $T(y, \hat{y}) = |y_{61} - \hat{y}| - |y_6 - \hat{y}|$ (Gelman *et al.*, 2004, p. 164). Thus, the form of definition of T in equation (3) is consistent with standard practice.

The value of T in equation (3) for the actual data in Figure 2 is greater than zero. In fact, across all the credible parameter values in the 200,000-step MCMC chain, the average value of $T(y^{act}, \hat{y})$ is 7.52, as shown in the left panel of Figure 3. This distribution and the others in Figure 3 were created by generating a complete set of random data from the model at every step in the 200,000-step MCMC chain, and computing $T(y^{rep}, \hat{y})$, $T(y^{act}, \hat{y})$, and $T(y^{rep}, \hat{y}) - T(y^{act}, \hat{y})$ at every step. The fact that $T(y^{act}, \hat{y})$ is robustly greater than zero indicates that it is a plausible signature of upward curvature. The expected value of $T(y^{rep}, \hat{y})$, however, must be zero, because randomly generated values will be above or below the prediction line equally often. This expected value is verified in the middle panel of Figure 3. The right panel of Figure 3 shows the sampling distribution of $T(y^{rep}, \hat{y}) - T(y^{act}, \hat{y})$, where it can be seen that the Bayesian p-value is .116, which is not very small. In other words, from the definition of curvature in equation (3), we would *not* reject the linear model.

What are we to conclude about curvature in the data, in light of this failure to reject the linear model using a Bayesian p-value? Not much (in my opinion), because the visual impression of discrepancy is very strong, and we can always try some other definition of T.

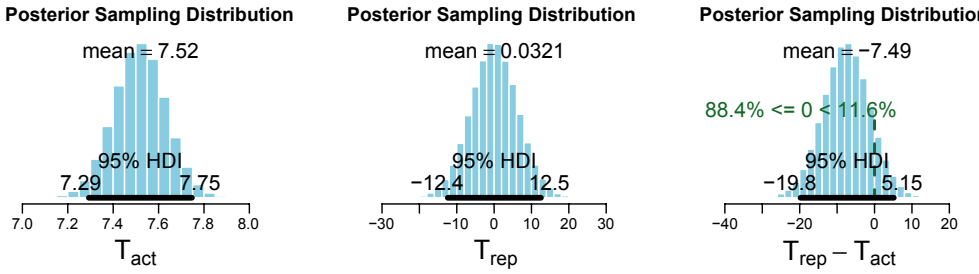

Figure 3. Posterior sampling distributions of $T(y^{act}, \hat{y})$, $T(y^{rep}, \hat{y})$, and $T(y^{rep}, \hat{y}) - T(y^{act}, \hat{y})$ for T defined in equation (3), from the posterior and data of Figure 2. 'HDI' denotes highest density interval. In the right panel, the text '88.4% $\leq$ 0 < 11.6%' means that 88.4% of the distribution falls below zero, and 11.6% of the distribution falls above zero. (Theoretically, $T(y^{rep}, \hat{y})$ is symmetric with a mean of 0.0. The histogram in the middle panel deviates slightly from the theoretical characteristics because of random sampling noise.)

Analysts who harbour a desire to reject the model can keep trying until they find a definition of T for which p is small, while analysts who harbour a desire not to reject the model can stop when they find a definition of T for which p is not very small.

Importantly, the goal of posterior predictive checking is not merely to reject the model, because, as Gelman and Shalizi (2012, 2013) and Gelman *et al.* (2004) have emphasized, we know in advance that the descriptive model is almost surely wrong for real data. The goal of posterior predictive checking is to come up with a more satisfying descriptive model of the data. Therefore we can simply side-step the process of arbitrarily defining T, generating its sampling distribution and struggling with its ambiguous implications. Instead, we should expand the descriptive model with explicit structural terms that capture the trends in which we are interested.

The apparent discrepancy in Figure 2 can be directly expressed in an expanded model that allows for non-linear trend and outliers. For example, we can directly express a quadratic trend and a heavy-tailed distribution as

$$\hat{y} = \beta_0 + \beta_1 x + \beta_2 x^2, \tag{4}$$

$$y \sim t(\hat{y}, \sigma, \nu), \tag{5}$$

where β_2 is the coefficient of quadratic trend and $\nu \geq 1$ is the degrees of freedom parameter for the t distribution. The t distribution is often used as a convenient descriptive distribution for data with outliers (e.g., Damgaard, 2007; Jones & Faddy, 2003; Lange, Little, & Taylor, 1989; Meyer & Yu, 2000; Tsionas, 2002). When ν is large (e.g., 100), the t distribution is very nearly normal. When ν gets close to 1, the t distribution is strongly kurtotic.

Figure 4 shows the results from Bayesian estimation of the five parameters in equations (4) and (5). As before, the prior distributions were minimally informed, and the analysis used MCMC sampling by JAGS (Plummer, 2003) called from R (R Development Core Team, 2011) with programs in the style of Kruschke (2011b). Visual inspection of the posterior estimates suggests that the model describes the data well: the data tend to be tightly clustered near the quadratic curve, with only a few outliers accommodated by the heavy-tailed distribution. (In fact, the data were randomly generated from exactly such a model, and the Bayesian estimates recovered the generating values well. But we never know the true generating model for real data.)

How do we know that the expanded model is better than the original model? In principle, we could do Bayesian model comparison. But in practice, Bayesian model comparison can be hypersensitive to the choice of prior distributions in the models, as Gelman and Shalizi (2013) remind us. Therefore Bayesian model comparison is to be avoided unless we have well-informed priors that put the two models on equal footing (e.g., Kruschke, 2011a; Liu & Aitkin, 2008; Vanpaemel, 2010), which we do not have in this case. Instead, because the models are nested in this case, we can simply see whether the posterior estimates of the additional parameters are credibly non-zero. Figure 5 displays the marginals of the posterior distribution, where it can be seen that the quadratic coefficient β_2 is robustly non-zero. Thus, despite the fact that the Bayesian p-value did *not* reject the linear model (recall Figure 3), an expanded model with an explicit quadratic trend strongly *does* implicate non-linearity in the data.

As another illustration of the peril of allowing arbitrary definitions of T without a specific alternative model, suppose we look at Figure 4 with the aim of finding even more discrepancy and rejecting the expanded model (perception of pattern is linked to motivation; e.g.,

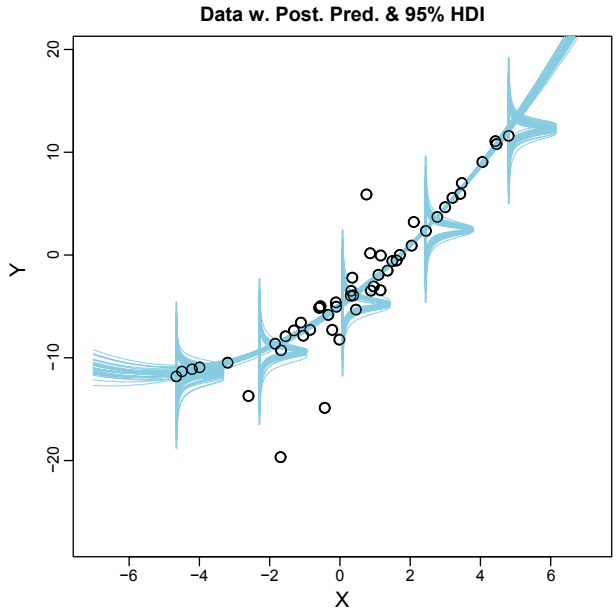

Figure 4. Data with posterior predictions from quadratic regression with *t*-distributed likelihood as defined in equations (4) and (5). The curves extending from left to right show a smattering of credible regression lines from the MCMC chain. The vertical lines show 95% highest density intervals (HDIs) and *t* density functions with corresponding standard deviations and degrees of freedom. The data appear to be well described by the posterior prediction (which is fortunate, because this form of model actually generated the data).

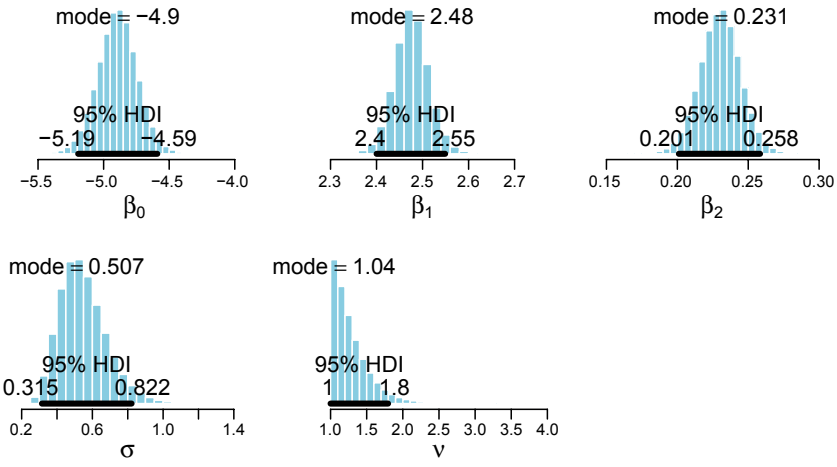

Figure 5. Marginals of the posterior distribution for the five parameters of equations (4) and (5), for the data displayed in Figure 4. The histograms summarize an MCMC chain of 200,000 steps. The top right-hand histogram shows that the quadratic coefficient β_2 is credibly greater than zero. (The displayed modes are approximated by a kernel density smoother.)

Whitson & Galinsky, 2008). It appears that there is counterclockwise 'torsion' in the residuals, such that there are more outliers in the lower-left and upper-right quadrants near the median of x. Because I am not sure what I mean by this, in terms of an actual structural trend expressed in a functional form, I will define a signature of the torsion as

$$T(y,\hat{y}) = -\sum_{i=6}^{17}(y_i - \hat{y}_i) + \sum_{i=28}^{29}(y_i - \hat{y}_i). \tag{6}$$

The expression in equation (6) merely sums the residuals in a particular range below the median of x and subtracts the result from the sum of residuals in a particular range above the median of x. A posterior predictive check produces the posterior sampling distributions shown in Figure 6. The Bayesian p-value is small, just .032. According to conventional p-value criteria, this result should lead us to reject the model, without recourse to a specific alternative.

But this conclusion seems unwarranted. In this case, we know that the data were actually generated by the model that has been rejected, but this conflict is not the reason for being sceptical, because for real data we do not know the true generator of the data. The scepticism arises because the definition of T was cherry-picked from a universe of all possible definitions of T without any motivation other than trying to prove the model wrong.

If I were forced to define a functional form for the structural trend of 'torsion in outliers', I might attempt to use a likelihood distribution that has a skew parameter, with the skew parameter functionally linked to the value of x, so that the skew is negative when x is just below its median, but positive when x is just above its median. This expanded form involves new parameters for skew and for the functional relation between skew and x, and we would also have to specify a prior on the parameters of the expanded model. A prior that would be agreeable to a sceptical audience might favour null values on the expanded parameters because the model is so unusual. Even without a sceptical prior on the extra parameters, there is increased uncertainty in the higher-dimensional parameter space, hence it is less likely that the estimates of the extra parameters would be credibly non-zero. Even though an expanded model might be deemed arbitrary like T, Bayesian evaluation of the expanded model incorporates penalties for arbitrariness, unlike T. There

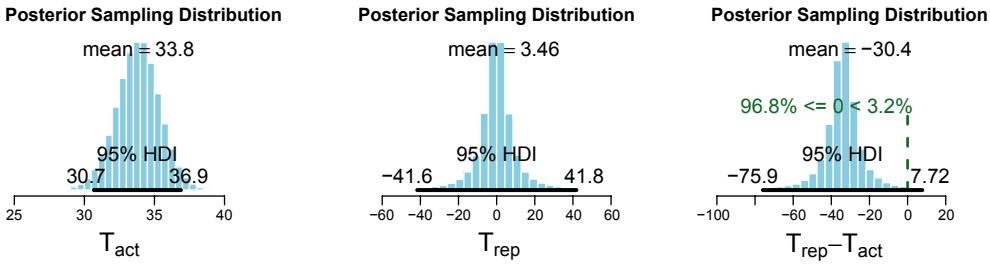

Figure 6. Posterior sampling distributions of $T(y^{act},\hat{y})$, $T(y^{rep},\hat{y})$, and $T(y^{rep},\hat{y}) - T(y^{act},\hat{y})$ for T defined in equation (6), from the posterior and data of Figure 4. 'HDI' denotes highest density interval. In the right panel, only 3.2% of the distribution falls above zero. (Theoretically, $T(y^{rep},\hat{y})$ is symmetric with a mean of 0.0. The histogram in the middle panel deviates slightly from the theoretical characteristics because of random sampling noise in the extreme tails of the distribution.)

is a penalty from a sceptical prior and from increased uncertainty in a higher-dimensional parameter space. Moreover, if Bayesian model comparison is undertaken with appropriate caution, the diluted prior on the higher-dimensional parameter space automatically penalizes the more complex model to fend off overfitting (often referred to as the Bayesian Occam's razor effect, e.g., MacKay, 2003).

I have presented two examples in which the conclusion from a Bayesian p-value conflicted with the conclusion from a Bayesian estimation of an expanded model. In general, the conclusions from Bayesian estimation of an expanded model supersede the conclusions of a corresponding Bayesian p-value. If the conclusions agree, the expanded model and explicit posterior distribution provide rich structural definition that is more specific than the ambiguous signature expressed by T. If the conclusions disagree, then again we look to the explicit structural form of the expanded model, and its estimated parameters, to better understand the data. If a Bayesian p-value is small and rejects a model, it merely confirms a foregone conclusion, and we still need an explicit structural form to understand why. If a Bayesian p-value is large and does not reject a model, it might be merely because the definition of T does not capture the structural form of the discrepancy which would be apparent when estimated in an explicit expanded model.

4. Summary and conclusion

In typical research, the models we use to describe data are selected because of their familiarity from previous training, tractability in computation, and prior probability of describing trends we care about in the specific application. But we know in advance that the models are merely descriptive, and that the data were almost surely *not* generated by such a model. Gelman and Shalizi (2013, p. 20) say 'The goal of model checking, then, is not to demonstrate the foregone conclusion of falsity as such, but rather to learn how, in particular, this model fails'. My argument above is completely consistent with this perspective. The argument, bolstered with examples, said merely that the *ad hoc* construction of a measure T such that $p(T^{rep} \geq T^{act})$ is an exercise in a foregone conclusion. Moreover, the implications are ambiguous because the measure T does not entail a specific structural form for an expanded model. Instead of going through the foregone conclusion and ambiguous implication of Bayesian p values, we should instead define an expanded model and evaluate it with Bayesian estimation.

I have also suggested that a qualitative posterior predictive check may be Bayesian, insofar as perception and cognition themselves may be Bayesian. There is no inherent necessity for model checking to be non-Bayesian. Formal Bayesian calculations are conditional on a particular model space, but there are a variety of ways to provoke the analyst to consider other model spaces. The provocation can come from a posterior predictive check, or the provocation can come from learning about other types of models in other applications and wondering whether there is an analogous application, or the provocation can come from simply wanting to prove a competing theorist wrong. But whatever the provocation, the space of possible alternatives is still governed by the mental prior in the analyst's mind.

References

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006, July). Special issue: Probabilistic models of cognition. *Trends in Cognitive Sciences*, *10*(7), 287–344.

Cox, D. R., & Hinkley, D. V. (1974). *Theoretical statistics*. London: Chapman & Hall.

Damgaard, L. H. (2007). Technical note: How to use WinBUGS to draw inferences in animal models. *Journal of Animal Science*, *85*, 1363–1368.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). Boca Raton, FL: CRC Press.

Gelman, A., & Shalizi, C. R. (2012). Philosophy and the practice of Bayesian statistics in the social sciences. In H. Kincaid (Ed.), *The Oxford handbook of philosophy of social science*. Oxford: Oxford University Press.

Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, *66*, 8–38. doi:10.1111/j.2044-8317.2011.02037.x

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: exploring representations and inductive biases. *Trends in Cognitive Sciences*, *14*(8), 357–364.

Griffiths, T. L., & Tenenbaum, J. B. (2007). From mere coincidences to meaningful discoveries. *Cognition*, *103*(2), 180–226.

Helmholtz, H. L. (1867). *Handbuch der physiologischen Optik*. Leipzig: L. Voss.

Jacobs, R. A., & Kruschke, J. K. (2010). Bayesian learning theory applied to human cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*, 8–21.

Jones, M. C., & Faddy, M. J. (2003). A skew extension of the *t*-distribution, with applications. *Journal of the Royal Statistical Society, Series B*, *65*(1), 159–174.

Kalish, M. L., Griffiths, T. L., & Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin & Review*, *14*(2), 288–294.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology*, *55*, 271–304.

Knill, D. C., & Richards, W. (Eds.) (1996). *Perception as Bayesian inference*. Cambridge: Cambridge University Press.

Kruschke, J. K. (2006). Locally Bayesian learning with applications to retrospective revaluation and highlighting. *Psychological Review*, *113*(4), 677–699.

Kruschke, J. K. (2010a). Bridging levels of analysis: comment on McClelland et al. and Griffiths et al. *Trends in Cognitive Sciences*, *14*(8), 344–345.

Kruschke, J. K. (2010b). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences*, *14*(7), 293–300.

Kruschke, J. K. (2011a). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, *6*(3), 299–312.

Kruschke, J. K. (2011b). *Doing Bayesian data analysis: A tutorial with R and BUGS*. Burlington, MA: Academic Press/Elsevier.

Lange, K. L., Little, R. J. A., & Taylor, J. M. G. (1989). Robust statistical modeling using the *t* distribution. *Journal of the American Statistical Association*, *84*(408), 881–896.

Liu, C. C., & Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, *52*, 362–375.

MacKay, D. J. C. (2003). *Information theory, inference & learning algorithms*. Cambridge: Cambridge University Press.

Meyer, R., & Yu, J. (2000). BUGS for a Bayesian analysis of stochastic volatility models. *Econometrics Journal*, *3*(2), 198–215.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (dsc 2003)*, Vienna.)

R Development Core Team (2011). *R: A language and environment for statistical computing* [computer software manual]. Vienna: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Shams, L., & Beierholm, U. R. (2010). Causal inference in perception. *Trends in Cognitive Sciences*, *14*, 425–432.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285.

Tsionas, E. G. (2002). Bayesian inference in the noncentral Student-*t* model. *Journal of Computational and Graphical Statistics*, *11*(1), 208–221.

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*, *54*, 491–498.

Westheimer, G. (2008). Was Helmholtz a Bayesian? a review. *Perception*, *37*, 642–650.

Whitson, J. A., & Galinsky, A. D. (2008). Lacking control increases illusory pattern perception. *Science*, *322*, 115–117.

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: Analysis by synthesis? *Trends in Cognitive Sciences*, *10*(7), 301–308.

Received 30 January 2012