



Post-Processing Posterior Predictive p Values

Nils Lid Hjort, Fredrik A Dahl & Gunnhildur Högnadóttir Steinbakk

To cite this article: Nils Lid Hjort, Fredrik A Dahl & Gunnhildur Högnadóttir Steinbakk (2006) Post-Processing Posterior Predictive p Values, Journal of the American Statistical Association, 101:475, 1157-1174, DOI: [10.1198/016214505000001393](https://doi.org/10.1198/016214505000001393)

To link to this article: <https://doi.org/10.1198/016214505000001393>



Published online: 01 Jan 2012.



Submit your article to this journal [↗](#)



Article views: 456



View related articles [↗](#)



Citing articles: 11 View citing articles [↗](#)

Post-Processing Posterior Predictive p Values

Nils Lid HJORT, Fredrik A. DAHL, and Gunnhildur Högnadóttir STEINBAKK

This article addresses issues of model criticism and model comparison in Bayesian contexts, and focuses on the use of the so-called *posterior predictive p values* (ppp). These involve a general discrepancy or conflict measure and depend on the prior, the model, and the data. They are used in statistical practice to quantify the degree of surprise or conflict in data and to compare different combinations of prior and model. The distribution of such ppp values is far from uniform however, as we demonstrate for different models, making their interpretation and comparison a difficult matter. We propose a natural calibration of the ppp values, where the resulting cppp values are uniform on the unit interval under model conditions. The cppp values, which in general rely on a double-simulation scheme for their computation, may then be used to assess and compare different priors and models. Our methods also make it possible to compare parametric and nonparametric model specifications, in that genuine “measures of surprise” are put on the same canonical uniform scale. We illustrate our techniques for some applications to real data. We also present supplementing theoretical results on various properties of the ppp and cppp.

KEY WORDS: Calibration of ppp values; Dipper data; Double simulation; Model criticism; Posterior predictive p values; Prior construction; Prior predictive evaluation; Surprise.

1. INTRODUCTION

Bayesian inference involves selecting a prior, $\pi(\theta)$, for the unknown parameters, θ , and a model, $f(y|\theta)$, for the data y given θ . In complex situations, some model criticism analysis is often desired, exposing the model to certain posterior checks; this sometimes leads to modifications in the model or in the prior. In other situations there might be a need to consider several candidates for both the prior and the model. This leads to questions on how to meaningfully criticize, evaluate, compare, and select among these candidates.

1.1 Existing Approaches

There are by necessity several approaches to handling such general problems. Classic goodness-of-fit monitoring remains relevant; also Bayesians may need to check whether the data follow a normal distribution. Bayes factors (of which there are different related versions) are often used (see, e.g., Smith and Spiegelhalter 1980; Kass and Raftery 1995; Marden 2000). Related to these again is the so-called “Bayesian information criterion” (BIC) (see, e.g., Robert 2001, chap. 7, and the lucid discussion in Clyde and George 2004). A quite general model-evaluation strategy uses the deviance information criterion (DIC) of Spiegelhalter, Best, Carlin, and van der Linde (2002) (see also van der Linde 2004); this method is in widespread use because its computation is an easy byproduct of the Markov chain Monte Carlo (MCMC) simulations often used to simulate from the posterior distributions. Model adequacy evaluation tools geared in part specifically toward use in hierarchical models have been proposed and discussed by Gelfand and Dey (1994), Dey, Gelfand, Swartz, and Vlachos (1998), O’Hagan (2003), Lu, Hodges, and Carlin (2006), and Bayarri and Castellanos (2004). Versions of the Bayesian model selection problem may also be cast in decision theoretic terms involving utility or loss functions (see Gelfand and Ghosh 1998;

Gutiérrez-Peña and Walker 2001; Claeskens and Hjort 2003; Hjort and Claeskens 2003; Kadane and Lazar 2004).

In addition, various authors have attempted to construct “Bayesian p values,” which can be viewed as quantifying the degree of surprise from data, given the prior and the model, sometimes also focusing on certain hypotheses. The Bayesian p values come in many forms, ranging from the prior predictive p values of Box (1980) to the posterior predictive p values touched on by Guttman (1967) and Rubin (1984), a tool worked out more fully by Gelman, Meng, and Stern (1996) and Meng (1994). Important variations and alternative Bayesian p values were introduced by Bayarri and Berger (2000) and further discussed and analyzed by Robins, van der Vaart, and Ventura (2000), Bayarri and Berger (2004), and Bayarri and Castellanos (2004).

We note that the division line between “model criticism” (or “model evaluation”) and “model comparison” is not sharp. If model criticism tools show four out of five models to be faulted but accept the fifth, then a comparison is already taking place. Nevertheless, methods like the BIC and the DIC are naturally sorted under model comparison, or model selection, whereas p value methods, the main concern of this article, are meant as model evaluation methods.

1.2 The Posterior Predictive p Value

This article focuses on one of the aforementioned mechanisms for carrying out such evaluation and comparisons, namely the posterior predictive p value (ppp). It requires specification of a suitable discrepancy measure $D = D(y, \theta)$, reflecting aspects thought to be important for the final conclusions of the statistical analysis. The intention is to assess a posteriori the fit of the underlying model assumptions or to quantify the degree of surprise by observing what we actually observed, in view of prior and model. In the formulation of Gelman et al. (1996), the ppp is defined as

$$\text{ppp} = \text{ppp}(y^{\text{obs}}) = \Pr\{D(y^{\text{rep}}, \theta) \geq D(y^{\text{obs}}, \theta) | \text{data}\}. \quad (1)$$

Here y^{obs} signifies the observed data or, in some cases, a suitably relevant subset of the full dataset, whereas y^{rep} represents a new (“future”) y of the same type, drawn conjointly with θ ,

Nils Lid Hjort is Professor, Department of Mathematics, University of Oslo, N-0316 Oslo, Norway (E-mail: nils@math.uio.no). Fredrik A. Dahl is Senior Statistician, Health Services Research Unit, Akershus University Hospital, P.B. 300, 1471 Lorenskog, Norway (E-mail: fredrik.dahl@ahus.no). Gunnhildur Högnadóttir Steinbakk is Doctoral Student, Department of Mathematics, University of Oslo, N-0316 Oslo (E-mail: ghs@math.uio.no). The article has benefited from constructive suggestions from the associate editor, the editor Frank Samaniego, Alan Gelfand, Bent Natvig, Geir Storvik, and two anonymous reviewers. The research of Dahl and Steinbakk was supported in part by the Norwegian Science Council under the BeMatA program “Evaluation of Hierarchical Models,” led by Hjort.

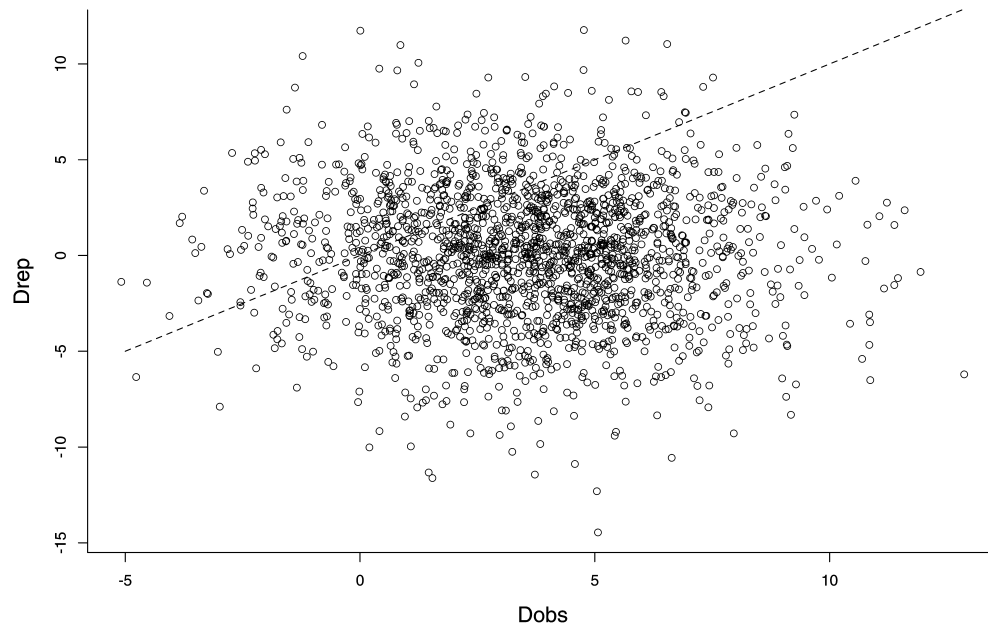


Figure 1. A Total of 2,000 Pairs of $D(\mathbf{y}^{\text{rep}}, \mu, \sigma)$ Against $D(\mathbf{y}^{\text{obs}}, \mu, \sigma)$ for the Speed of Light Data, Using the Normal Model With Discrepancy Measure $D = |y_{(61)} - \mu| - |y_{(6)} - \mu|$, With a Vague Uniform Prior for $(\mu, \log \sigma)$. The ppp value is .208.

which comes from the posterior distribution $\pi(\boldsymbol{\theta}|\text{data})$. More concretely,

$$(\boldsymbol{\theta}, \mathbf{y}^{\text{rep}}) \sim \pi(\boldsymbol{\theta}|\text{data})f(\mathbf{y}|\boldsymbol{\theta}), \quad (2)$$

so that in particular, $\mathbf{y}^{\text{rep}}$ comes from the posterior predictive distribution $\int f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\text{data})d\boldsymbol{\theta}$. Provided that we can simulate $\boldsymbol{\theta}_j$'s from the posterior and $\mathbf{y}_j^{\text{rep}}$ datasets from the model $f(\mathbf{y}|\boldsymbol{\theta}_j)$, we may evaluate the ppp as

$$\text{ppp}(\mathbf{y}^{\text{obs}}) \doteq \frac{1}{A} \sum_{j=1}^A I\{D(\mathbf{y}_j^{\text{rep}}, \boldsymbol{\theta}_j) \geq D(\mathbf{y}^{\text{obs}}, \boldsymbol{\theta}_j)\} \quad (3)$$

across a high number A of simulations. It is useful to plot $D(\mathbf{y}^{\text{obs}}, \boldsymbol{\theta}_j)$ versus $D(\mathbf{y}_j^{\text{rep}}, \boldsymbol{\theta}_j)$, whereby the ppp value is identified as the proportion of points above the diagonal; see Figure 1.

Example: Newcomb's Speed of Light Data. In an impressive 1882 experiment, Simon Newcomb attempted to measure the amount of time required for light to travel a distance of 7,442 meters. His 66 observations, recorded as deviations from 24,800 nanoseconds, were used by Gelman, Carlin, Stern, and Rubin (2004, sec. 6.3) to illustrate aspects of the ppp technology. There are two rather extreme values at -44 and -2 , with the remaining 64 points following a reasonably symmetric distribution from 16 to 40. (Incidentally, the accepted speed of light constant translates to a true value of 33.0 on Newcomb's scale.) Gelman et al. first used $\min\{y_1, \dots, y_{66}\}$ as a discrepancy to show that a $N(\mu, \sigma^2)$ model does not fit the data (there being two clear outliers), and then turned to the discrepancy measure $D(\mathbf{y}, \mu, \sigma) = |y_{(61)} - \mu| - |y_{(6)} - \mu|$, where $y_{(i)}$ denotes the i th-ordered data point. They computed $\text{ppp} = .26$ from 200 simulations from the traditional "objective posterior" for (μ, σ) that corresponds to a uniform prior for $(\mu, \log \sigma)$, and concluded from this that "any observed asymmetry in the middle of the distribution can easily be explained by sampling

variation." Figure 1 displays $(D^{\text{obs}}, D^{\text{rep}})$ pairs from 2,000 simulations. A precise ppp computation, with 1 million rather than 200 or 2,000 simulations, gives $\text{ppp} = .208$.

The appeal of the ppp apparatus is partly the generous flexibility afforded the statistician through choices of the discrepancy function, which can be set up to test or inspect different aspects of the model formulation. Unlike traditional test statistics, the discrepancy functions are also allowed to depend on the unknown parameters. Also typically presented as a bonus point is the apparent lack of strong dependence on the prior; the ppp may be computed even with vague or objective noninformative priors as long as the posterior is proper. Among the take-home messages of this article is that the apparent bonus point just described is deceiving; even though different priors may give rise to approximately the same posteriors, and therefore to approximately the same ppp numbers, the degree of surprise might be quite different. We argue in particular against the interpretation implicit in Gelman et al.'s analysis; we cannot know whether $\text{ppp} = .208$ is surprising or not until we have decided on a proper scale. As we discuss later, the uniform scale is not at all appropriate.

The ppp value of observed data $\mathbf{y}^{\text{obs}}$ is often viewed as being capable of checking adequacy of a model, separated from the prior (cf. the foregoing quote from Gelman et al.). Although this is true for some tailor-made discrepancies, it is our view that ppp measures in general might better be seen as ways of assessing the adequacy of the combined (model, prior). In the (relatively rare) cases where our prior is viewed as "the distribution Nature used when creating the world," we might prefer the prior predictive p value (prpp), suggested by Box (1980). Although this is also a bona fide p value [with null distribution uniform on $(0, 1)$], Box's prpp can only handle test statistics $D = D(\mathbf{y})$ that do not depend on the $\boldsymbol{\theta}$. Thus, even in cases when the prior is taken quite literally, there is a place for ppp values.

1.3 The Calibrated ppp

In a certain sense, the ppp calculation uses the data twice, first updating the prior to fit the data better and then estimating how surprising the data are relative to the posterior parameter distribution. Thus it is not surprising that its distribution across likely values of $\mathbf{y}^{\text{obs}}$ is not uniform; we can, in fact, demonstrate various extreme aspects of nonuniformity in several situations. This makes the interpretation and comparison of ppp values a difficult and risky matter. To alleviate this problem, we propose to post-Process or calibrate the ppp value to

$$\text{cphp}(\mathbf{y}^{\text{obs}}) = \Pr\{\text{ppp}(\mathbf{Y}) \leq \text{ppp}(\mathbf{y}^{\text{obs}})\}, \quad (4)$$

where the distribution of $\mathbf{Y}$ is that implied by the prior and model conditions. By construction, this cphp, the perfected ppp value, will now have a uniform null distribution, that is, it may be viewed as a genuine p value. The main message of our article is that the ppp values have limited information value in themselves, but that the naturally rescaled cphp versions are genuinely useful and interpretable across different combinations of priors and models.

That the ppp numbers sometimes do not convey very useful information was not picked up in the initial reactions when the method was introduced (see, e.g., the discussion contributions to Gelman et al. 1996). There have been later warnings in the literature, however. For example, Dey et al. (1998) complained that the ppp values were not able to pick out model inadequacies in a string of Bayesian logistic-normal regression setups, Sinharay and Stern (2003) similarly observed that the ppp numbers tended to cluster too tightly around $\frac{1}{2}$ to give clear signals of model distortions in hierarchical models, and Bayarri and Berger (2000) found $\text{ppp} = .409$ in a situation where data clash seriously with a model. We, in fact, demonstrate that these are fairly typical situations; when information content increases in relation to the complexity of a model, the ppp values will tend to cluster around $\frac{1}{2}$ for natural classes of discrepancy functions. Our contention is that these problems are solved by the calibration mechanism alluded to earlier, as discussed in more detail in Section 4. When properly calibrated, the cphp values are well able to give signals of surprise or conflict and may be used to screen away unfortunate combinations of prior and data model in a unified manner.

Unlike the ppp value, the ppp value can be applied in cases where the prior is vaguer than one's "true" subjective belief about the parameters in question. The prior may still be informative but does not need to be fine-tuned to reflect all aspects of our prior belief. Typically, the ppp value becomes less and less dependent on the prior, for a given data model, as data information accumulates; however, the cphp value remains critically dependent on the actual prior used. This is at it should be; the calibration transform (4) actively uses and assesses the implications of a given prior.

1.4 The Present Article

The layout of our article is as follows. In Section 2 we investigate the structurally simple situation where data are normal with a normal prior for the mean. Here we find a formula for the ppp and use this to study important special cases, corresponding to having a sharp prior, a flatter prior, and a fixed prior with increasing sample size. Such calculations and analysis may also be extended to normal-normal hierarchical mod-

els. We also include a brief large-sample analysis of the ppp for general parametric models. Further properties and aspects of the ppp are gleaned from studying situations with a mixture of subpriors in Section 3.

In Section 4 we propose and develop our perfected ppp values, the cphp, and discuss their computation and interpretation. The cphp may in general be computed through a double-simulation regime. The details of our general theory are then worked out in the context of general linear regression models in Section 5, illustrated for a real dataset, pertaining to the modeling of sprint speedskating results, in Section 6. These techniques could be used routinely in all Bayesian analyses of normal-linear models, to check for model adequacy and to monitor data for any serious conflicts with the prior used. We also illustrate our techniques for two models pertaining to survival patterns of the European dipper species in Section 7. Brooks, Catchpole, and Morgan (2000) analyzed the same data using the ppp apparatus, but through our cphp we reach somewhat different conclusions.

Our cphp values, by construction, have been transformed to a "canonical scale of surprise," namely the uniform on the unit interval. Therefore, observed cphp numbers enjoy a clear interpretation and can be compared across several proposed or imagined combinations of prior and model. We may even apply the techniques to comparison of parametric versus nonparametric model specifications, as illustrated in Section 8. Such comparisons are often quite difficult, conceptually and operationally, with other approaches.

The area of Bayesian model evaluation is evolving rapidly, with sometimes conflicting views expressed by different authors. The reviewers of our manuscript, for example, gave somewhat conflicting expert opinions and recommendations regarding our methods and their ramifications. For these reasons, we attempt to clarify and summarize our main findings and views in Section 9. We end the article in Section 10 with a list of ppp- and cphp-related topics and issues, some worthy of further research effort.

The success of our cphp analysis depends crucially on the choice of discrepancy measure, the $D(\mathbf{y}, \boldsymbol{\theta})$. This choice should be aided by context and aspects of the actual application, and in particular may be constructed to address those model aspects considered crucial for the principal conclusions of the statistical analysis. (In this connection, see the constructions of Dey et al. 1998; O'Hagan 2003; Sinharay and Stern 2003.) Bayesian statisticians need as always to care about prior, model, and loss function; the modern, conscientious members of the species also need to exhibit creative acuity for matters pertaining to model selection and screening of priors, which in the methodology and machinery of this article would mean choosing good and problem-relevant discrepancy measures, followed by appropriate cphp analysis.

2. THE NORMAL-NORMAL MODEL

Here we investigate the simple situation where data are normal with unknown mean and the prior for this mean parameter is also normal. We find an explicit formula for the ppp that provides certain general insights into its properties and behavior. Our calculations extend to the case with the traditional inverse-gamma times normal conjugate prior in the general normal model and may also be generalized to the case of normal-normal-type hierarchical models.

2.1 A Formula for the ppp Value

Assume that the data $\mathbf{y} = (y_1, \dots, y_n)$, conditional on θ , are iid from $N(\theta, \sigma^2)$, with known standard deviation σ , and let the prior be $\theta \sim N(\theta_0, \sigma_0^2)$. We choose to work with

$$D(\mathbf{y}, \theta) = \frac{n(\bar{y} - \theta)^2}{\sigma^2} = \text{monotone}(|\bar{y} - \theta|)$$

as a discrepancy measure, where $\bar{y}$ is the mean of y_i 's; any monotone increasing function of $|\bar{y} - \theta|$ gives the same ppp value, as seen from (1). This is a pivotal quantity, with distribution equal to a χ^2_1 for each given θ . For the following result, let $F_{1,1}(v, \kappa)$ be the cumulative distribution function of a noncentral Fisher variable with degrees of freedom (1, 1) and excentricity parameter κ , that is, of a variable of the form $(X + \kappa^{1/2})^2/Y^2$, where X and Y are independent and standard normals.

Proposition 1. For the situation described above, the ppp value, as a function of the observed data, may be expressed as

$$\text{ppp}(\mathbf{y}^{\text{obs}}) = F_{1,1}\left(1 + \frac{\sigma^2}{n\sigma_0^2}, \frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \frac{(\bar{y}^{\text{obs}} - \theta_0)^2}{\sigma_0^2}\right). \quad (5)$$

Proof. Using well-known techniques, the posterior distribution for θ can be found to be normal with mean $(1 - \rho_n)\theta_0 + \rho_n\bar{y}^{\text{obs}}$ and variance $\rho_n\sigma^2/n$, writing ρ_n for $n\sigma_0^2/(n\sigma_0^2 + \sigma^2)$. Conditional on θ , the $\bar{y}^{\text{rep}}$ is, of course, an $N(\theta, \sigma^2/n)$. Now write

$$(\theta|\text{data}) \sim (1 - \rho_n)\theta_0 + \rho_n\bar{y}^{\text{obs}} + \rho_n^{1/2}(\sigma/\sqrt{n})N_0$$

and

$$(\bar{y}^{\text{rep}}|\theta) \sim \theta + (\sigma/\sqrt{n})N,$$

in terms of two independent standard normals N_0 and N . We learn that $D(\mathbf{y}^{\text{rep}}, \theta) = N^2$ for given θ , independent of N_0 , whereas

$$\begin{aligned} D(\mathbf{y}^{\text{obs}}, \theta) &= \frac{n}{\sigma^2} \left\{ (1 - \rho_n)(\theta_0 - \bar{y}^{\text{obs}}) + \frac{\sigma}{\sqrt{n}} \rho_n^{1/2} N_0 \right\}^2 \\ &= \left\{ \rho_n^{1/2} N_0 - (1 - \rho_n) \frac{\sqrt{n}(\bar{y}^{\text{obs}} - \theta_0)}{\sigma} \right\}^2. \end{aligned}$$

Hence the ppp value becomes

$$\text{ppp} = \Pr \left[N^2 \geq \rho_n \left\{ N_0 - \frac{1 - \rho_n}{\rho_n^{1/2}} \frac{\sqrt{n}}{\sigma} (\bar{y}^{\text{obs}} - \theta_0) \right\}^2 \right],$$

from which the result follows.

2.2 Special Cases

The foregoing formulas offer insight into the ppp value. Here are some remarks and consequences, which are followed up in later sections.

Large n or Flat Prior, or Both. Suppose that $\sqrt{n}\sigma_0$ is large, that is, either n is large or the prior is flat, or both. Then ρ_n goes to 1, and $\text{ppp}(\mathbf{y}^{\text{obs}}) \rightarrow \Pr\{N^2 \geq N_0^2\} = \frac{1}{2}$ for all $\mathbf{y}^{\text{obs}}$. This occurs irrespective of the observed $\bar{y}^{\text{obs}}$ and is illustrated in Figure 2 for $n = 10$, $\sigma = 1$, $\theta_0 = 0$, and $\sigma_0 = 5$, giving $\rho_n = .996$; the ppp is very close to $\frac{1}{2}$ over a broad range of $\bar{y}^{\text{obs}}$.

Moderate n and Sharp Prior. Suppose, on the other hand, that $\sqrt{n}\sigma_0$ is small, which means that the prior knowledge about θ is reasonably sharp compared with sample size; that is, it is believed rather firmly that θ is close to θ_0 . Then ρ_n is close to 0, and, in the limit,

$$\text{ppp}(\mathbf{y}^{\text{obs}}) \rightarrow p^*(\bar{y}^{\text{obs}}) = \Pr \left\{ N^2 \geq \frac{n(\bar{y}^{\text{obs}} - \theta_0)^2}{\sigma^2} \right\}.$$

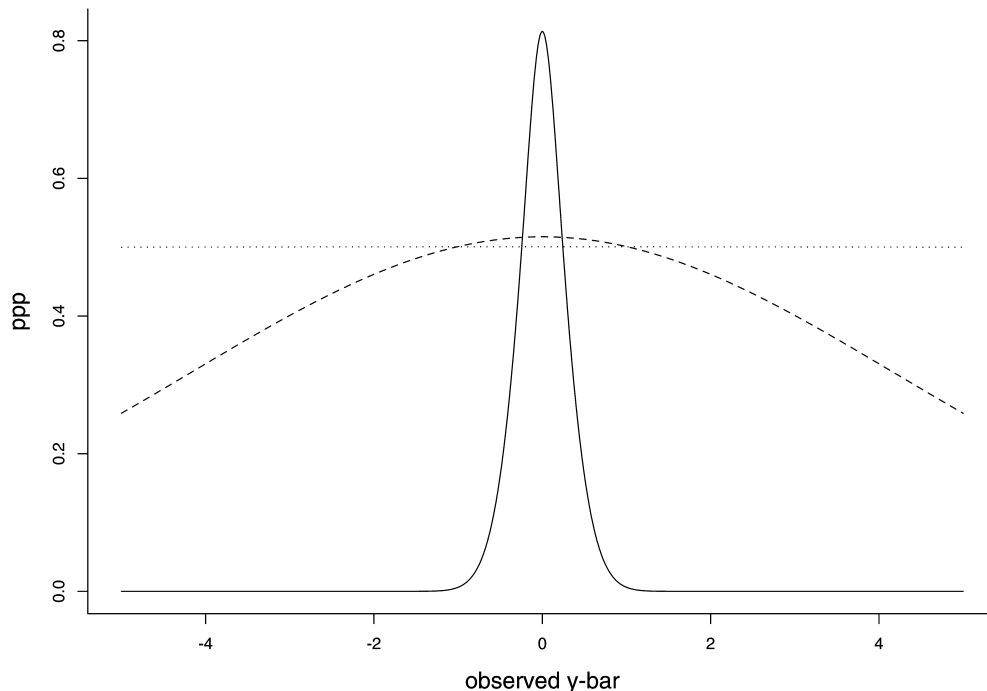


Figure 2. The $\text{ppp}(\mathbf{y}^{\text{obs}})$ of (5) Displayed as a Function of $\bar{y}^{\text{obs}}$, for $n = 10$, $\sigma = 1$, $\theta_0 = 0$, for the Three Cases σ_0 Equal to .1 (—), 1.0 (-----), and 5.0 (.....).

This is the classic p value for testing the hypothesis $\theta = \theta_0$; see Figure 2, for $n = 10$ and $\sigma_0 = .1$, with $\rho_n = .091$. Observe also that for given ρ_n , there is a maximum attainable value for ppp, namely $F_{1,1}(1/\rho_n)$, for $\bar{y}^{\text{obs}} = \theta_0$; this is the maximally unsurprising value we might have for $\bar{y}$ under the given prior. In particular, the ppp distribution is not symmetric around $\frac{1}{2}$. The figure also displays the ppp as a function of $\bar{y}^{\text{obs}}$ for an intermediate case of $n = 10$ and $\sigma_0 = 1$, for which $\rho_n = .909$. We observe that the distribution of ppp has a clear maximum value just above $\frac{1}{2}$ (actually, .5151), with most mass from, say, .25 to this max value.

2.3 Full Normal Analysis

Above, we took the data standard deviation to be known. More realistically, both parameters of the $N(\mu, \sigma^2)$ distribution are unknown. The traditional conjugate prior takes an inverse gamma for σ^2 and a normal for μ given σ . Agree to say that $(\lambda, \mu) = (1/\sigma^2, \mu)$ comes from the $\text{GN}(\frac{1}{2}a_0, \frac{1}{2}b_0, \mu_0, c_0)$ distribution if

$$\begin{aligned} \lambda &\sim \text{Gam}(\tfrac{1}{2}a_0, \tfrac{1}{2}b_0) \quad \text{and} \\ (\mu|\lambda) &\sim N(\mu_0, (c_0\lambda)^{-1}). \end{aligned} \quad (6)$$

Accordingly, its density is

$$\pi(\lambda, \mu) \propto \lambda^{a_0/2-1} \exp(-\tfrac{1}{2}b_0\lambda) \lambda^{1/2} \exp\{-\tfrac{1}{2}\lambda c_0(\mu - \mu_0)^2\}.$$

The likelihood for an observed dataset $y_1, \dots, y_n$ from the normal may be written as proportional to $\lambda^{n/2} \exp[-\frac{1}{2}\lambda\{Q_0^{\text{obs}} + n(\mu - \bar{y}^{\text{obs}})^2\}]$, where $Q_0^{\text{obs}} = \sum_{i=1}^n (y_i - \bar{y}^{\text{obs}})^2$. The posterior density becomes

$$\begin{aligned} &\propto \lambda^{(a_0+n)/2-1} \lambda^{1/2} \\ &\quad \times \exp\left[-\tfrac{1}{2}\lambda\{b_0 + Q_0^{\text{obs}} + n(\mu - \bar{y}^{\text{obs}})^2 + c_0(\mu - \mu_0)^2\}\right] \\ &= \lambda^{(a_0+n)/2-1+1/2} \\ &\quad \times \exp\left[-\tfrac{1}{2}\lambda\{b_0 + Q_0^{\text{obs}} + (c_0^{-1} + n^{-1})^{-1}(\bar{y}^{\text{obs}} - \mu_0)^2\right. \\ &\quad \left.+ (c_0 + n)(\mu - \tilde{\mu})^2\right\}], \end{aligned}$$

proving that

$$\begin{aligned} \{(\lambda, \mu)|\text{data}\} &\sim \text{GN}(\tfrac{1}{2}(a_0 + n), \\ &\quad \tfrac{1}{2}(b_0 + Q_0^{\text{obs}} + (c_0^{-1} + n^{-1})^{-1}(\bar{y}^{\text{obs}} - \mu_0)^2), \\ &\quad \tilde{\mu}, c_0 + n), \end{aligned} \quad (7)$$

in terms of $\tilde{\mu} = (c_0\mu_0 + n\bar{y}^{\text{obs}})/(c_0 + n)$; see also the more general analysis of Section 5.

We continue using the natural discrepancy function $D(\mathbf{y}, \theta) = n(\bar{y} - \mu)^2/\sigma^2$, for which, when $\theta = (\mu, \sigma)$ is given, $D(\mathbf{y}^{\text{rep}}, \theta) = N^2$ for a standard normal N , independent of

$$\begin{aligned} D(\mathbf{y}^{\text{obs}}, \theta) &= n\lambda\{\tilde{\mu} + (c_0 + n)^{-1/2}\lambda^{-1/2}N_0 - \bar{y}^{\text{obs}}\}^2 \\ &= \frac{n}{c_0 + n} \\ &\quad \times \left\{N_0 + (c_0 + n)^{1/2}\lambda^{1/2} \frac{c_0}{c_0 + n}(-\bar{y}^{\text{obs}} + \mu_0)\right\}^2, \end{aligned}$$

where again N_0 is standard normal. This yields an exact computable expression for the ppp in this case,

$$\begin{aligned} \text{ppp}(\mathbf{y}^{\text{obs}}) &= \int_0^\infty \Pr\left[N^2 \geq \frac{n}{c_0 + n} \left\{N_0 + \lambda^{1/2} \frac{c_0}{(c_0 + n)^{1/2}} \right. \right. \\ &\quad \left. \left. \times (-\bar{y}^{\text{obs}} + \mu_0)\right\}^2 \middle| \lambda\right] g_n(\lambda) d\lambda \\ &= \int_0^\infty F_{1,1}\left(\frac{c_0 + n}{n}, \frac{c_0^2\lambda}{c_0 + n}(\bar{y}^{\text{obs}} - \mu_0)^2\right) g_n(\lambda) d\lambda, \end{aligned} \quad (8)$$

in terms of the gamma density $g_n(\lambda)$ with parameters $(\frac{1}{2}a_n, \frac{1}{2}b_n)$, where, from the foregoing, $a_n = a_0 + n$ and $b_n = b_0 + Q_0^{\text{obs}} + c_0n(\bar{y}^{\text{obs}} - \mu_0)^2/(c_0 + n)$. The previous case of known σ corresponds to a_0 and b_0 going to infinity with $a_0/b_0 = 1/\sigma^2$ and $1/(c_0\lambda) = \sigma^2/c_0 = \sigma_0^2$, that is, $c_0 = \sigma^2/\sigma_0^2$. In this case the foregoing formula reduces to (5).

Remark 1. The ppp value of (5) converges rather rapidly to $\frac{1}{2}$ as sample size increases. Some analysis reveals that $F_{1,1}(1 + a/n, b/n) \doteq \frac{1}{2} + \frac{1}{2}(a - b)/(\pi n)$, with consequences for the (5) formula and approximations.

Remark 2. The formulas given by Meng (1994, sec. 3) relate to the ppp value for another discrepancy measure, namely $D(\mathbf{y}, \theta_0) = n(\bar{y} - \theta_0)^2/\sigma^2$, used as a test statistic for testing $\theta = \theta_0$. Our ppp measure uses the more general $D(\mathbf{y}, \theta)$.

Remark 3. Result (7) is not new, of course, but the direct quick derivation that we give takes less space than identifying a proper reference, which then would require careful translation of all quantities involved there to quantities in our framework. A similar comment applies to the more general result of Section 5.2.

2.4 ppp Behavior for General Parametric Models

Calculations similar to the foregoing can also be carried out for more complicated models, with additional effort. In work not reported on here, we have found, for example, explicitly computable formulas for the ppp value for the situation where data are exponential and the parameter has a gamma prior, as well as for the Poisson–gamma situation. In general we would not be able to have explicit formulas for the $\text{ppp}(\mathbf{y})$, however, and would need to resort to the simulation scheme of (3) to compute the ppp value.

Some general phenomena may be recognized from the foregoing analysis of the normal–normal and other cases mentioned. Suppose that data $\mathbf{y} = (y_1, \dots, y_n)$ are iid from a density $g(\mathbf{y}, \theta)$, conditional on θ . Assume also that a discrepancy measure of the type $D(\mathbf{y}, \theta) = H(\sqrt{n}(\hat{\theta}(\mathbf{y}) - \theta), \theta)$ is used, where $\hat{\theta}(\mathbf{y})$ is the maximum likelihood estimator. We might take, say, $H(\mathbf{v}, \theta) = \mathbf{v}^t \mathbf{J}(\theta) \mathbf{v}$, with the Fisher information matrix $\mathbf{J}(\theta)$; in this case $D(\mathbf{y}, \theta)$ is close to a χ_p^2 for large n , where p is the dimension of θ . In general,

$$\begin{aligned} \text{ppp}(\mathbf{y}^{\text{obs}}) &= \Pr\{H(\sqrt{n}(\hat{\theta}(\mathbf{y}^{\text{rep}}) - \theta), \theta) \\ &\quad \geq H(\sqrt{n}(\hat{\theta}(\mathbf{y}^{\text{obs}}) - \theta), \theta) | \text{data}\}. \end{aligned}$$

Two cases of interest are as follows. Robins et al. (2000) gave results of a similar nature.

First, consider the case of a “sharp prior,” tightly concentrated around some θ_0 . Then this also goes for the posterior, and

$$\begin{aligned} \text{ppp}(\mathbf{y}^{\text{obs}}) &\doteq \Pr\{H(\sqrt{n}(\hat{\theta}(\mathbf{y}^{\text{rep}}) - \theta_0), \theta_0) \\ &\geq H(\sqrt{n}(\hat{\theta}(\mathbf{y}^{\text{obs}}) - \theta_0), \theta_0)\} \\ &= p^*(\mathbf{y}^{\text{obs}}). \end{aligned}$$

This is the classic p value for testing $\theta = \theta_0$ with the test $D(\mathbf{y}, \theta_0) = H(\sqrt{n}(\hat{\theta}(\mathbf{y}) - \theta_0), \theta_0)$. This result requires only that $D(\mathbf{y}, \theta)$ is continuous in θ .

Next, study the large-sample scenario where data really follow the $g(\mathbf{y}, \theta_{\text{tr}})$ model, for a suitable true parameter value, and n grows. Under mild regularity conditions, the distribution of $\sqrt{n}(\hat{\theta}(\mathbf{y}^{\text{rep}}) - \theta)$ is close to, say, $\mathbf{V}(\theta)$, which is $N_p(\mathbf{0}, \mathbf{J}(\theta)^{-1})$, and this approximation statement holds uniformly in a neighborhood around the θ_{tr} value. In particular, $\sqrt{n}(\hat{\theta}(\mathbf{y}^{\text{rep}}) - \theta) \rightarrow_d \mathbf{V}$, which is $N_p(\mathbf{0}, \mathbf{J}(\theta_{\text{tr}})^{-1})$. Second, from Bernshtein–von Mises type theorems (see, e.g., Lehmann 1983, chap. 6), the posterior distribution of $\sqrt{n}(\theta - \hat{\theta}^{\text{obs}})$ is with probability 1 coming close to that of $\mathbf{V}_0$, another and independent $N_p(\mathbf{0}, \mathbf{J}(\theta_{\text{tr}})^{-1})$ variable. All of this implies that $D(\mathbf{y}^{\text{rep}}, \theta) \rightarrow_d H(\mathbf{V}, \theta_{\text{tr}})$ and $D(\mathbf{y}^{\text{obs}}, \theta) \rightarrow_d H(\mathbf{V}_0, \theta_{\text{tr}})$, with probability 1. Therefore, as long as $H(\mathbf{v}, \theta)$ is continuous,

$$\text{ppp}(\mathbf{y}^{\text{obs}}) \rightarrow \Pr\{H(\mathbf{V}, \theta_{\text{tr}}) \geq H(\mathbf{V}_0, \theta_{\text{tr}})\} = \frac{1}{2} \quad \text{a.s.}$$

This is the precise description of a phenomenon that has occasionally been noted in the literature but perhaps not well understood; see, for example, comments by Sinharay and Stern (2003) about the ppp values clustering around $\frac{1}{2}$.

3. THE ppp VALUE WHEN THE PRIOR IS A MIXTURE

Here we study the ppp value for normal data under a mixture prior for its mean, and use insights thus revealed to make some general comparisons with the so-called prior predictive p values advocated by Box (1980). Assume as in the previous section that data $y_1, \dots, y_n$ conditional on θ are iid from $N(\theta, \sigma^2)$ and that the same discrepancy measure $D(\mathbf{y}, \theta)$ is used, but that the prior is a mixture of two different hypotheses about nature, $\theta \sim p_1 N(\theta_{0,1}, \sigma_{0,1}^2) + p_2 N(\theta_{0,2}, \sigma_{0,2}^2)$. Then

$$\begin{aligned} (\theta, \bar{y}) &\sim p_1 \pi_1(\theta) f(\bar{y}|\theta) + p_2 \pi_2(\theta) f(\bar{y}|\theta) \\ &= p_1 \pi_1(\theta|\text{data}) f_1(\bar{y}) + p_2 \pi_2(\theta|\text{data}) f_2(\bar{y}), \end{aligned}$$

in terms of the posterior densities $\pi_j(\theta|\text{data})$ for θ and of the marginal densities $f_j(\bar{y})$ under the two prior hypotheses in question. In fact, $f_j(\bar{y})$ is a normal with mean $\theta_{0,j}$ and variance $\sigma_{0,j}^2 + \sigma^2/n$. This leads to

$$(\theta|\text{data}) \sim \tilde{p}_1(\mathbf{y}^{\text{obs}}) \pi_1(\theta|\text{data}) + \tilde{p}_2(\mathbf{y}^{\text{obs}}) \pi_2(\theta|\text{data}),$$

where $\tilde{p}_j(\mathbf{y}^{\text{obs}}) = p_j f_j(\bar{y}^{\text{obs}}) / \{p_1 f_1(\bar{y}^{\text{obs}}) + p_2 f_2(\bar{y}^{\text{obs}})\}$ for $j = 1, 2$.

This may now be used to find a formula for the ppp. As in Section 2, $D(\mathbf{y}^{\text{rep}}, \theta)$ may be represented as N^2 , where N is a standard normal. For $D(\mathbf{y}^{\text{obs}}, \theta)$, it is with probability $\tilde{p}_j(\mathbf{y}^{\text{obs}})$

of the type worked with in Section 2.1, with appropriate parameters $\rho_{n,j} = n\sigma_{0,j}^2 / (n\sigma_{0,j}^2 + \sigma^2)$ and $\theta_{0,j}$, for $j = 1, 2$. Hence

$$\begin{aligned} \text{ppp}(\mathbf{y}^{\text{obs}}) &= \sum_{j=1}^2 \tilde{p}_j(\mathbf{y}^{\text{obs}}) F_{1,1} \left(\frac{1}{\rho_{n,j}}, \frac{(1 - \rho_{n,j})^2}{\rho_{n,j}} \frac{n(\bar{y}^{\text{obs}} - \theta_{0,j})^2}{\sigma^2} \right). \end{aligned}$$

The formula easily generalizes to a mixture across a wider spectrum of hypotheses about θ .

Several general aspects of ppp values can be learned from applying this formula in various mixture prior settings. A variety of possible shapes for the $\text{ppp}(\bar{y}^{\text{obs}})$ curves emerge from different combinations of the parameters. Among findings of interest are the following points:

1. Suppose that the two prior hypotheses are both sharp, situated at $\theta_{0,1}$ and $\theta_{0,2}$ with small standard deviations $\sigma_{0,1}$ and $\sigma_{0,2}$. Then we learn that the resulting $\text{ppp}(\mathbf{y}^{\text{obs}})$ function is relatively unaffected by the balance parameters p_1 and p_2 . The ppp tool essentially works as a classic frequentist p value for testing $\theta = \theta_{0,1}$ if data indicate that it is the first prior component that is the real one and for testing $\theta = \theta_{0,2}$ if the second prior component is picked out by the data.

For comparison, consider the prpp value advocated by Box (1980) with respect to the test statistic that sorts the $\mathbf{y}$ values according to their prior likelihood, and take p_2 very small but positive. This $\text{prpp}(\mathbf{y}^{\text{obs}})$ would give a value close to 0 for $\bar{y}^{\text{obs}}$ close to $\theta_{0,2}$, because $\bar{y}$ values close to $\theta_{0,2}$ are highly unlikely under the given prior. In fact, we may easily prove that

$$\lim_{p_2 \rightarrow 0} \text{ppp}(\bar{y}^{\text{obs}} = \theta_{0,2}) = 1$$

and

$$\lim_{p_2 \rightarrow 0} \text{prpp}(\bar{y}^{\text{obs}} = \theta_{0,2}) = 0$$

in the situation with two sharp hypotheses, that is, $\sigma_{0,j} = 0$ for $j = 1, 2$. Thus the ppp value may in this respect be seen as a highly noncontinuous operator, because the two $\text{ppp}(\mathbf{y}^{\text{obs}})$ curves are highly different for $p_2 = 0$ and $p_2 = .001$.

2. Suppose that $\sigma_{0,1}$ is moderate while $\sigma_{0,2}$ is small. Then when $\bar{y}^{\text{obs}}$ is close to $\theta_{0,2}$, the ppp value acts as a classic p value for testing $\theta = \theta_{0,2}$; if, on the other hand, $\bar{y}^{\text{obs}}$ is some distance away from $\theta_{0,2}$, then the ppp value is close to what it would be for the first prior component alone. Again, this behavior is relatively independent of the parameters p_1 and p_2 . It is interesting and perhaps mildly contrainuitive that even when p_2 is very small, the ppp indicates nonsurprise of $\bar{y}^{\text{obs}}$ close to $\theta_{0,2}$; thus, in this situation, even some events that have very low prior probability are deemed completely acceptable (see Fig. 3).
3. Let us now hold p_2 fixed (but small), while $\theta_{0,2}$ varies. We have seen that when $\theta_{0,2}$ is large, $\text{ppp}(\bar{y}^{\text{obs}} = \theta_{0,2})$ is close to 1. Also, if $\theta_{0,2} \rightarrow 0$, then we clearly get $\text{ppp}(\bar{y}^{\text{obs}} = \theta_{0,2}) \rightarrow 1$. But if $\theta_{0,2}$ is at a moderate distance from $\theta_{0,1}$, then $\text{ppp}(\bar{y}^{\text{obs}} = \theta_{0,2})$ may be close to 0. This is a bit of a paradox; we have two competing models, represented

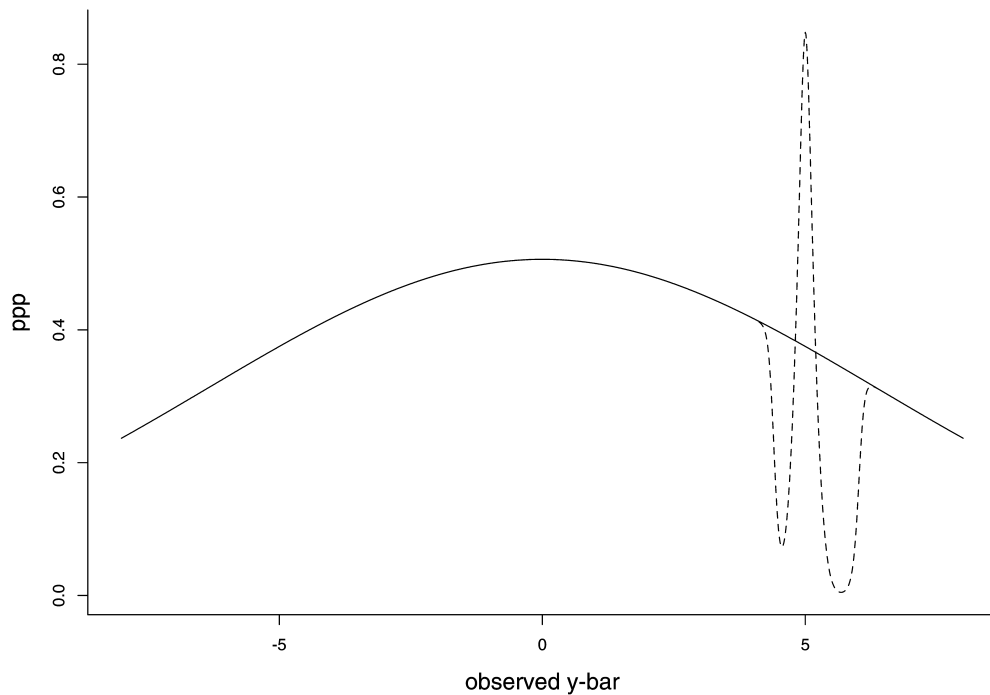


Figure 3. The $\text{ppp}(\mathbf{y}^{\text{obs}})$ as a Function of $\bar{y}^{\text{obs}}$, for $n = 25$ and $\sigma = 1$, Displayed for Two Priors, for the Normal $(0, 1)$ (—) and a Sharp Bi-Mixture With $p_1 = .999$ and $p_2 = .001$, of Two Normals $(0, 1)$ and $(5, .05^2)$ (-----).

through a prior with two peaks, and observe a value that fits the a priori unlikely model well. If the models are either very different or very similar, then ppp gives a high value, whereas if the models are moderately different, we get a low ppp value.

These examples illustrate the fact that the ppp value is relatively insensitive to the magnitude of the probability mass that the prior assigns to parts of the parameter space that are distant from one another. This is particularly true if the different parameter values give very different distributions for the observed $\mathbf{y}^{\text{obs}}$. If this is the case, then conditioning on the observation will essentially eliminate parameter values that fail to explain the data, making their prior likelihood irrelevant.

4. ppp VALUE CALIBRATION AND THE DOUBLE-LEVEL SIMULATION

Suppose that one computes $\text{ppp}(\mathbf{y}^{\text{obs}}) = .28$ for one's dataset, for a given prior and model. This is a well-defined probability, as in (1), but to judge the significance of the .28 number, possibly in comparison with other ppp values for other combinations of prior and model for the same dataset, we are forced to define the underlying probability scale: How rare, or how common, are values $< .28$?

There has perhaps been a certain Pavlovian tendency in applied statistics work to interpret the ppp values on the uniform scale, as for classic p values. We have already seen that the distribution of ppp is often quite nonuniform, however, and we provide more precise information later in the article. We see that an observed $\text{ppp} = .28$ (say) may be extremely surprising in one situation, whereas the value $\text{ppp} = .06$ (say) may not be very surprising in another situation. Like frequentists, who compute p values as probabilities of events involving outcomes that did

not occur, here also the Bayesian (if a priori willing to work with the ppp in the first place) is forced to consider the values of $\mathbf{y}$ that were not observed. In this section we define and study the appropriate distribution of $\text{ppp}(\mathbf{Y})$ across values of $\mathbf{Y}$, and use this to introduce our calibrated ppp value, the $\text{cPPP}(\mathbf{y}^{\text{obs}})$. We illustrate its actual use with real data in Sections 6 and 7.

4.1 The Null-Null Distribution of the ppp

Classical Bayes analysis, of course, operates conditional on $\mathbf{y}^{\text{obs}}$. When one also wishes to test the validity of aspects of the model and the prior, however, one needs to take into account the distribution of $\mathbf{y}^{\text{obs}}$ as well. This motivates defining the null-null distribution of $\text{ppp}(\mathbf{Y})$, corresponding to the distribution of $\text{ppp}(\mathbf{y})$ across precisely those $\mathbf{y}$ values that occur by the combined mechanism of the prior and the model. The null-null distribution of $U = \text{ppp}(\mathbf{Y})$ under a perfect prior and a perfect model is

$$G(u) = \Pr\{\text{ppp}(\mathbf{Y}) \leq u\}, \quad \text{where } \mathbf{Y} \sim \int f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) d\boldsymbol{\theta}. \quad (9)$$

This is the distribution that should be used to calibrate the observed ppp value. We propose using

$$\text{cPPP}(\mathbf{y}^{\text{obs}}) = G(\text{ppp}(\mathbf{y}^{\text{obs}})) = \Pr\{\text{ppp}(\mathbf{Y}) \leq \text{ppp}(\mathbf{y}^{\text{obs}})\}. \quad (10)$$

The distribution of $\text{cPPP}(\mathbf{Y})$, across values of $\mathbf{Y}$ as before, is then, by construction, a uniform on $(0, 1)$ (as long as G is continuous). In other words, the cPPP is a proper p value.

To illustrate, consider the normal-normal situation of Section 2 and suppose that the world is exactly as we have imagined, in terms of both Nature's distribution of θ and our own modeling of data given θ . Then $\bar{Y}$ is normal $(\theta_0, \sigma_0^2 + \sigma^2/n)$,

expressible as $\theta_0 + (\sigma_0^2 + \sigma^2/n)^{1/2}M$ with a standard normal M . The consequent null–null distribution of ppp becomes

$$\begin{aligned} G(u) &= \Pr\left\{F_{1,1}\left(\frac{1}{\rho_n}, \frac{(1-\rho_n)^2 n(\bar{Y} - \theta_0)^2}{\sigma^2}\right) \leq u\right\} \\ &= \Pr\left\{\frac{(1-\rho_n)^2 n(\bar{Y} - \theta_0)^2}{\rho_n \sigma^2} \geq q(u)\right\} \\ &= \Pr\left\{\frac{n}{\sigma^2}\left(\sigma_0^2 + \frac{\sigma^2}{n}\right)M^2 \geq \frac{\rho_n}{(1-\rho_n)^2}q(u)\right\} \\ &= \Pr\left\{\chi_1^2 \geq \frac{q(u)\rho_n}{1-\rho_n}\right\} \end{aligned}$$

for $0 \leq u \leq F_{1,1}(1/\rho_n)$. Here $q(u)$ is the excenter parameter that makes $F_{1,1}(1/\rho_n, q) = u$.

The ppp distribution can also easily be displayed through simulation of

$$\begin{aligned} U &= F_{1,1}\left(\frac{1}{\rho_n}, \frac{(1-\rho_n)^2}{\rho_n}\left(1 + \frac{n\sigma_0^2}{\sigma^2}\right)M^2\right) \\ &= F_{1,1}\left(\frac{1}{\rho_n}, \frac{1-\rho_n}{\rho_n}M^2\right), \end{aligned} \quad (11)$$

where $M \sim N(0, 1)$. To compute the $\text{cPPP}(\mathbf{y}^{\text{obs}})$, we may simply check the relative frequency of such simulated U 's that fall below the observed $\text{ppp}(\mathbf{y}^{\text{obs}})$. For this particular situation, we may even get an explicit expression,

$$\begin{aligned} \text{cPPP}(\mathbf{y}^{\text{obs}}) &= \Pr\left\{F_{1,1}\left(\frac{1}{\rho_n}, \frac{(1-\rho_n)^2 n(\bar{Y} - \theta_0)^2}{\sigma^2}\right) \leq F_{1,1}\left(\frac{1}{\rho_n}, \frac{(1-\rho_n)^2 n(\bar{y}^{\text{obs}} - \theta_0)^2}{\sigma^2}\right)\right\} \\ &= \Pr\left\{\frac{n(\bar{Y} - \theta_0)^2}{\sigma^2} \geq \frac{n(\bar{y}^{\text{obs}} - \theta_0)^2}{\sigma^2}\right\} \\ &= \Pr\left\{\chi_1^2 \geq \frac{n(\bar{y}^{\text{obs}} - \theta_0)^2/\sigma^2}{1 + n\sigma_0^2/\sigma^2}\right\}. \end{aligned} \quad (12)$$

Whereas the classic p value statistic is uniform on the unit interval under the null hypothesis, the present null–null distribution is quite far from such a form. The U variable of (11) is confined to $[0, F_{1,1}(1/\rho_n)]$, with a sharp upper threshold, and is highly skewed to the left. For $\sqrt{n}\sigma_0$ moderate or large (sample size is moderate or prior is moderately flat, or both), ρ_n is close to 1, and the ppp distribution is quite tightly concentrated around $\frac{1}{2}$. Only for $\sqrt{n}\sigma_0$ quite small, with ρ_n close to 0, does the ppp distribution come close to the uniform distribution on the unit interval (which is the limit case as $\rho_n \rightarrow 0$). But the statistical distribution of $\text{cPPP}(\mathbf{Y})$ of (12) is, by construction, exactly a uniform on $(0, 1)$, provided that the prior and the model are correct. Figure 4 shows the partly extreme nature of the ppp distribution, and Figure 5 shows the ppp and cPPP curves, along with a third variant treated in a later section.

Remark 4. When n grows, the observed $\bar{y}^{\text{obs}}$ will converge to the true underlying mean value θ_{tr} , and from (12) we see that $\text{cPPP}(\mathbf{y}^{\text{obs}})$ tends a.s. to $\Pr\{|N(0, 1)| \geq |\theta_{\text{tr}} - \theta_0|/\sigma_0\}$. Thus

$$\text{conf} = \frac{|\theta_{\text{tr}} - \theta_0|}{\sigma_0} = c_0^{1/2} \frac{|\theta_{\text{tr}} - \theta_0|}{\sigma}$$

emerges as a natural measure of conflict between the real data distribution and the $N(\theta_0, \sigma_0^2)$ prior. If $\text{conf} \geq 1.96$, then $\text{cPPP}(\mathbf{y}^{\text{obs}})$ will, for large n , be below the critical value .05, and so on. The second representation of conf uses $c_0 = \sigma^2/\sigma_0^2$, which has interpretation as prior sample size. See Remark 5 for a generalization and Section 10.1 for a constructive prior calibration idea.

4.2 The Double-Simulation Method to Calibrate the ppp Value

The foregoing reasoning invites the following simulation method to compute the calibrated ppp value. Simulate values (θ_k, y_k) for $k = 1, \dots, B$, for a high number B , where $\theta_k \sim \pi(\theta)$ and the full dataset y_k is drawn from the model given θ_k . Then compute

$$\text{cPPP}(\mathbf{y}^{\text{obs}}) \doteq \frac{1}{B} \sum_{k=1}^B I\{\text{ppp}(y_k) \leq \text{ppp}(\mathbf{y}^{\text{obs}})\}. \quad (13)$$

It is the perhaps chief claim of our article that although the $\text{ppp}(\mathbf{y}^{\text{obs}})$ of (3) may have a difficult interpretation and sometimes a low information value, the $\text{cPPP}(\mathbf{y}^{\text{obs}})$ of (13) has a clear meaning and can be highly informative.

Although clear in interpretation and natural qua strategy, the (13) operation may, of course, be both cumbersome and computationally costly from an operational standpoint, because in general it must amount to a double simulation, with AB operations in total, following (3). Therefore, one should look for ways to simplify the computational burden. Sometimes an explicit formula may be worked out for the $\text{ppp}(\mathbf{y}^{\text{obs}})$ with considerable benefit for the cPPP computations, as we discuss in the next section, or we may be helped by knowing the distribution of $D(\mathbf{y}^{\text{rep}}, \theta)$. In other cases we may look for variance reduction tricks or for ways to approximate the $\text{ppp}(\mathbf{Y})$ distribution, the benefit being that we may be allowed a moderate number rather than a large number B of repeated samplings. We have actually developed some methods of this kind in connection with the cPPP analysis of the bird survival data in Section 7, and will report on this elsewhere.

4.3 Alternative “What if” Scenarios

We judge the foregoing calibration to be the canonical one, transforming the ppp numbers under the proposed prior and model to the uniform scale. However, one may also consider other “what if” scenarios that from different perspectives lead to other potentially interesting distributions for $\mathbf{Y}$, and hence to other calibrations for $\text{ppp}(\mathbf{Y})$, in the formulation of (9)–(10). Suppose in general terms that a distribution for such $\mathbf{Y}^*$ datasets is being considered, where such $\mathbf{Y}^*$ are drawn from a mechanism different from the canonical one given in (9); this other distribution could take; for example, the form $\int f(\mathbf{y}|\theta)\pi^*(\theta) d\theta$ for a prior $\pi^*(\theta)$ different from $\pi(\theta)$. This defines an alternative ppp distribution, $G^*(u) = \Pr\{\text{ppp}(\mathbf{Y}^*) \leq u\}$, and in turn a differently calibrated ppp value,

$$\begin{aligned} \text{cPPP}^*(\mathbf{y}^{\text{obs}}) &= G^*(\text{ppp}(\mathbf{y}^{\text{obs}})) \\ &= \Pr\{\text{ppp}(\mathbf{Y}^*) \leq \text{ppp}(\mathbf{y}^{\text{obs}})\}. \end{aligned} \quad (14)$$

A ppp number is computed under a given model specification, say $\mathcal{M}$, and we may write $\text{ppp}(\mathbf{y}^{\text{obs}}, \mathcal{M})$ to indicate this. Whereas we earlier calibrated $\text{ppp}(\mathbf{y}^{\text{obs}}, \mathcal{M})$ through the

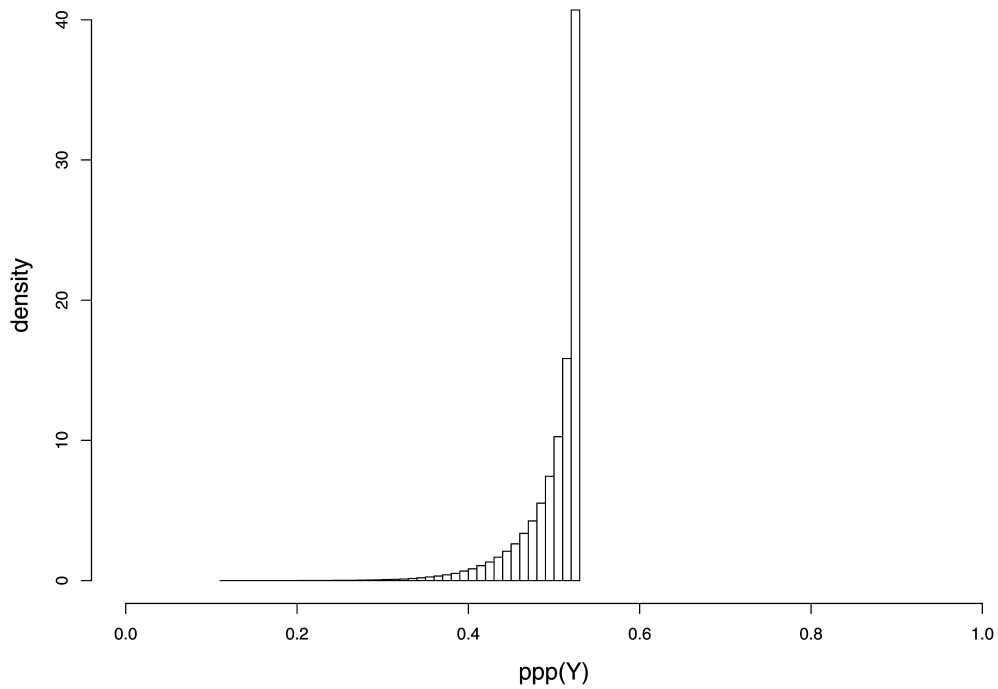


Figure 4. Density of $\text{ppp}(\mathbf{Y})$, Displayed as a Normalized Histogram With 1 Million Simulations, for $n = 5$, $\sigma = 1$, $\sigma_0 = 1$. The distribution has exact mean $\frac{1}{2}$, with sharp right cutoff point at $F_{1,1}(1/\rho_n)$.

same model $\mathcal{M}$, the operations as described earlier amount to calibrating $\text{ppp}(\mathbf{y}^{\text{obs}}, \mathcal{M})$ using an alternative model $\mathcal{M}^*$ for datasets $\mathbf{Y}^*$. The scheme (14) may then be thought of in terms of

$$\begin{aligned} \text{cPPP}^*(\mathbf{y}^{\text{obs}}) &= c[\text{ppp}(\mathbf{y}^{\text{obs}}, \mathcal{M}), \mathcal{M}^*] \\ &= \text{calibrated ppp}(\mathbf{y}^{\text{obs}}, \text{model } \mathcal{M}), \quad \text{under model } \mathcal{M}^*. \end{aligned} \quad (15)$$

The definition here is fully general and operational, and the cPPP^* may be computed for any well-defined $\mathbf{Y}^*$ distribution, through double simulation if necessary. The normal-normal model may make the implications of the (14) idea more clear; the reasoning that led to (12) now gives

$$\text{cPPP}^*(\mathbf{y}^{\text{obs}}) = \Pr \left\{ \frac{n(\bar{Y}^* - \theta_0)^2}{\sigma^2} \geq \frac{n(\bar{y}^{\text{obs}} - \theta_0)^2}{\sigma^2} \right\}, \quad (16)$$

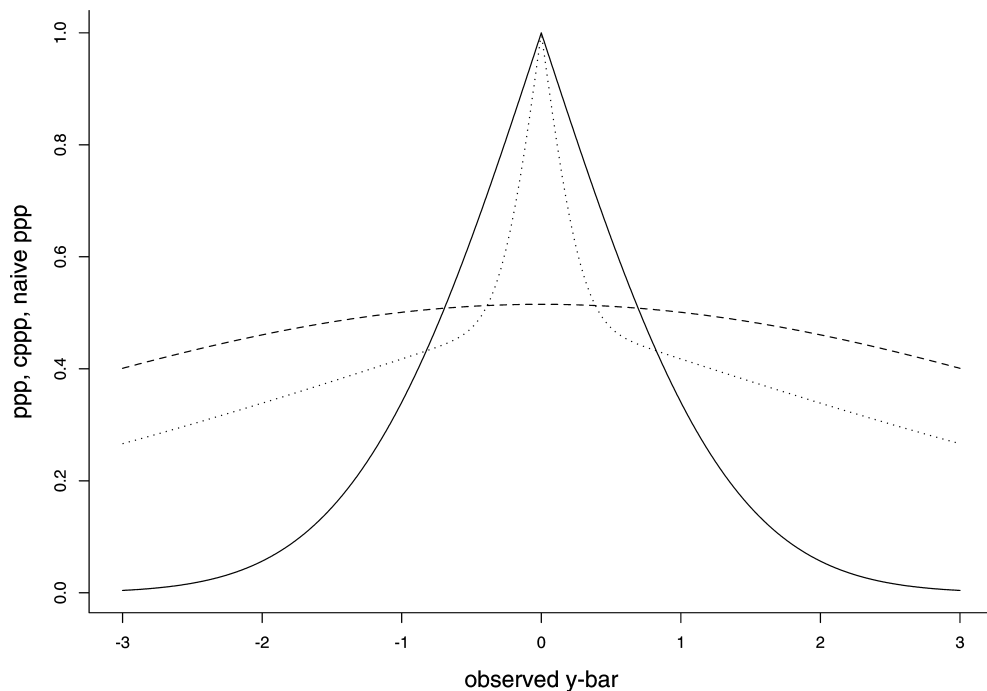


Figure 5. The Three Functions ppp (----), cPPP (—), and cPPP^* (.....) are displayed as functions of $\bar{y}^{\text{obs}}$, for $n = 10$, $\sigma = 1$, with an $N(0, 1)$ prior for θ .

in terms of the distribution of the sample average $\bar{Y}^*$ stemming from datasets $\mathbf{Y}^*$ drawn from the intended alternative what-if distribution. This gives formulas generalizing that of (12), for different models $\mathcal{M}^*$ for datasets $\mathbf{Y}^*$.

As a particular illustration, we may consider the somewhat naive data-twice-version that instead of sampling $\mathbf{y}$ from the prior predictive $\int f(\mathbf{y}|\theta)\pi(\theta) d\theta$ uses the posterior predictive $\int f(\mathbf{y}|\theta)\pi(\theta|\text{data}) d\theta$. This has some intuitive attraction in that it does consider the new knowledge about θ not available before the data; it would also be close to using $f(\mathbf{y}|\hat{\theta})$ as the model for data, with the maximum likelihood estimator. There are instances in the literature of such an analysis, sometimes carried out tentatively to illustrate consequences under different scenarios. The Bayesian clinical trials literature sometimes discusses “sampling density” versus “fitting density” (see, e.g., Gelfand and Wang 2002 for related discussion).

For the normal-normal situation, once more we know that $\theta|\text{data}$ is distributed as a normal with mean $(1 - \rho_n)\theta_0 + \rho_n\bar{y}^{\text{obs}}$ and variance $\rho_n\sigma^2/n$, with ρ_n as in Section 2.1. Sampling data Y_i^* from the $N(\theta, \sigma^2)$ for such θ , we find that $\bar{Y}^* - \theta_0$ is normal $(\rho_n(\bar{y}^{\text{obs}} - \theta_0), (1 + \rho_n)\sigma^2/n)$. Using the cppp* formula (16), this yields

$$\begin{aligned} \text{cppp}^*(\mathbf{y}^{\text{obs}}) &= \Pr\left[\left\{\rho_n(\bar{y}^{\text{obs}} - \theta_0) + (1 + \rho_n)^{1/2}(\sigma/\sqrt{n})N\right\}^2 \geq (\bar{y}^{\text{obs}} - \theta_0)^2\right] \\ &= \Pr\left[\frac{\sigma^2}{n}(1 + \rho_n)\left\{\frac{\sqrt{n}}{\sigma}\frac{\rho_n}{(1 + \rho_n)^{1/2}}(\bar{y}^{\text{obs}} - \theta_0) + N\right\}^2 \geq (\bar{y}^{\text{obs}} - \theta_0)^2\right] \\ &= \Pr\left\{\chi_1^2\left(\frac{\rho_n^2}{1 + \rho_n}z_n^{\text{obs}}\right) \geq \frac{z_n^{\text{obs}}}{1 + \rho_n}\right\}, \end{aligned}$$

in terms of $z_n^{\text{obs}} = n(\bar{y}^{\text{obs}} - \theta_0)^2/\sigma^2$. In the terminology and thinking of (15), this would be the formula for

$$\begin{aligned} \text{cppp}^*(\mathbf{y}^{\text{obs}}) &= \text{c}[\text{ppp}(\mathbf{y}^{\text{obs}}, \text{prior}), \text{posterior}] \\ &= \text{calibrated ppp}(\mathbf{y}^{\text{obs}}, \text{model prior}), \\ &\quad \text{under model posterior.} \end{aligned}$$

Our previous $\text{cppp}(\mathbf{y}^{\text{obs}})$ is, in this terminology, the same as $\text{c}[\text{ppp}(\mathbf{y}^{\text{obs}}, \text{prior}), \text{prior}]$.

Figure 5 illustrates the ppp, cppp, and cppp* curves for $n = 10$, $\sigma = 1$, with a normal $(0, 1)$ prior for θ , as a function of $\bar{y}^{\text{obs}}$. We see that the cppp* is not a satisfactory measure of surprise; we would scarcely ever be surprised if the numbers were interpreted on the uniform scale. This is due to the double use of data when we draw $\mathbf{Y}^*$ from the posterior predictive. There could be other choices for $\mathbf{Y}^*$ distributions with more relevance and with a less overcautious cppp* curve than here.

5. ppp AND cppp ANALYSIS FOR GENERAL REGRESSION

Here we study the general linear regression model for data $(\mathbf{x}_i, y_i)$ for which $y_i = \mathbf{x}_i^t\boldsymbol{\beta} + \varepsilon_i$, for $i = 1, \dots, n$, where $\mathbf{x}_i$ is a p -dimensional covariate vector for individual i , and the ε_i 's are

independent and mean-0 normal with standard deviation σ . In standard matrix formulation, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, with least squares estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}$ for $\boldsymbol{\beta}$; it is assumed that the $n \times p$ matrix $\mathbf{X}$ is of full rank. We develop theory for a canonical ppp measure that makes a good-quality evaluation of the underlying model assumptions. It is based on the discrepancy measure $D(\mathbf{y}, \boldsymbol{\theta}) = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^t\boldsymbol{\Omega}_n(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})/\sigma^2$, where $\boldsymbol{\Omega}_n = \mathbf{X}^t\mathbf{X} = \sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^t$ and where we write $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma)$ for the full parameter vector. It may also be expressed as

$$\begin{aligned} D(\mathbf{y}, \boldsymbol{\theta}) &= \frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^t\boldsymbol{\Omega}_n(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\sigma^2} = \frac{\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2}{\sigma^2} \\ &= \sum_{i=1}^n \frac{(\hat{\mu}_i - \mu_i)^2}{\sigma^2} \end{aligned} \quad (17)$$

in terms of the vector $\boldsymbol{\mu}$ of means $\mu_i = \mathbf{x}_i^t\boldsymbol{\beta}$ and their fitted values $\hat{\mu}_i = \mathbf{x}_i^t\hat{\boldsymbol{\beta}}$.

One may, of course, include other discrepancies for analysis, like $\max_{i \leq n} |y_i - \mathbf{x}_i^t\boldsymbol{\beta}|/\sigma$ or $\max_t |n^{-1} \sum_{i=1}^n I\{(y_i - \mathbf{x}_i^t\boldsymbol{\beta})/\sigma \leq t\} - \Phi(t)|$, writing Φ for the cumulative standard normal. Theory and computational schemes may be developed for these and other D functions, following the arguments and methods of this section. Any Bayesian regression analysis could in principle be supplemented with ppp and cppp analysis along the lines that we give here.

5.1 The Case of Known σ

Calculations for the ppp and cppp are rather easier and more immediately interpretable for the case where the data standard deviation σ is taken as known; thus we study that case first. Assume that $\boldsymbol{\beta}$ has a prior distribution of the form $N_p(\boldsymbol{\beta}_0, \sigma^2(c_0\boldsymbol{\Omega}_0)^{-1})$, with a matrix $\boldsymbol{\Omega}_0$ and a scalar c_0 . The parameterization is a bit redundant, because c_0 may be taken into the $\boldsymbol{\Omega}_0$, but it is useful to start with the covariance structure and then explore different levels of sharpness (c_0 large) or vagueness (c_0 small) for the prior. Using multivariate normal theory and some matrix algebra, we find that $\boldsymbol{\beta}$ given data is normal, with

$$E(\boldsymbol{\beta}|\text{data}) = \tilde{\boldsymbol{\beta}} = (c_0\boldsymbol{\Omega}_0 + \boldsymbol{\Omega}_n)^{-1}(c_0\boldsymbol{\Omega}_0\boldsymbol{\beta}_0 + \boldsymbol{\Omega}_n\hat{\boldsymbol{\beta}}^{\text{obs}})$$

and

$$\text{var}(\boldsymbol{\beta}|\text{data}) = \sigma^2(c_0\boldsymbol{\Omega}_0 + \boldsymbol{\Omega}_n)^{-1}.$$

In addition, $\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{\text{obs}} = (c_0\boldsymbol{\Omega}_0 + \boldsymbol{\Omega}_n)^{-1}c_0\boldsymbol{\Omega}_0(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}^{\text{obs}})$. Because $D(\mathbf{y}^{\text{rep}}, \boldsymbol{\theta}) \sim \chi_p^2$, independent of $D(\mathbf{y}^{\text{obs}}, \boldsymbol{\theta})$, this leads to the formula

$$\text{ppp}(\mathbf{y}^{\text{obs}}) = \Pr\{\chi_p^2 \geq (\mathbf{U} + \mathbf{f})^t\boldsymbol{\Omega}_n(\mathbf{U} + \mathbf{f})\},$$

where $\mathbf{U} \sim N_p(\mathbf{0}, (c_0\boldsymbol{\Omega}_0 + \boldsymbol{\Omega}_n)^{-1})$ and $\mathbf{f} = (c_0\boldsymbol{\Omega}_0 + \boldsymbol{\Omega}_n)^{-1}c_0\boldsymbol{\Omega}_0(\hat{\boldsymbol{\beta}}^{\text{obs}} - \boldsymbol{\beta}_0)/\sigma$. Computation would most often be simplest through simulation of $\mathbf{U}$ vectors. We note that when c_0 is large, $\mathbf{U}$ goes to $\mathbf{0}$ and $\mathbf{f}$ goes to $(\hat{\boldsymbol{\beta}}^{\text{obs}} - \boldsymbol{\beta}_0)/\sigma$, making $\text{ppp}(\mathbf{y}^{\text{obs}}) \rightarrow p^*(\mathbf{y})$, the traditional p value for testing $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ using the $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^t\boldsymbol{\Omega}_n(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)/\sigma^2$ test. On the other hand, when c_0 becomes small, $\mathbf{f} \rightarrow \mathbf{0}$ and $\text{ppp}(\mathbf{y}^{\text{obs}}) \rightarrow \frac{1}{2}$ for each dataset $\mathbf{y}^{\text{obs}}$.

To calibrate the ppp, we need to compute a number B of $\text{ppp}(\mathbf{y}_k)$ values from simulated datasets $\mathbf{y}_k$, as per (13). We would draw these sets by first drawing $\boldsymbol{\beta}_k$ from the posterior

distribution and then creating $\mathbf{y}_k = \mathbf{X}\boldsymbol{\beta}_k + \boldsymbol{\varepsilon}_k$, without changing or resampling the covariate vectors. For each of these B sets, we would find the $\text{ppp}(\mathbf{y}_k)$ number through simulation of A random vectors $\mathbf{U}$, that is,

$$\text{ppp}(\mathbf{y}_k) \doteq \frac{1}{A} \sum_{j=1}^A \Pr\{\chi_p^2 \geq (\mathbf{U}_{k,j} - \mathbf{f}(\mathbf{y}_k))^t \boldsymbol{\Omega}_n (\mathbf{U}_{k,j} - \mathbf{f}(\mathbf{y}_k))\}.$$

The only simple case is that of $\boldsymbol{\Omega}_0$ being proportional to $\boldsymbol{\Omega}_n$. In that case, we may without loss of generality, take $\boldsymbol{\Omega}_0 = n^{-1}\boldsymbol{\Omega}_n$, with appropriate prior sample size interpretation for the flexible factor c_0 in $\text{var } \boldsymbol{\beta} = \sigma^2(n/c_0)\boldsymbol{\Omega}_n^{-1}$; the $\boldsymbol{\Omega}_n$ matrix indeed grows linearly with sample size n . We then have

$$\begin{aligned} \tilde{\boldsymbol{\beta}} &= \frac{c_0}{c_0 + n} \boldsymbol{\beta}_0 + \frac{n}{c_0 + n} \hat{\boldsymbol{\beta}}^{\text{obs}}, \\ \mathbf{U} &\sim N_p\left(\mathbf{0}, \frac{n}{c_0 + n} \boldsymbol{\Omega}_n^{-1}\right), \quad \text{and} \\ \mathbf{f} &= \frac{c_0}{c_0 + n} \frac{\hat{\boldsymbol{\beta}}^{\text{obs}} - \boldsymbol{\beta}_0}{\sigma}. \end{aligned}$$

This leads to a representation for $(\mathbf{U} + \mathbf{f})^t \boldsymbol{\Omega}_n (\mathbf{U} + \mathbf{f})$, in terms of $\mathbf{W} \sim N_p(\mathbf{0}, \boldsymbol{\Omega}_n^{-1})$, as

$$\begin{aligned} &\frac{n}{c_0 + n} \left(\mathbf{W} + \left(\frac{c_0 + n}{n} \right)^{1/2} \frac{c_0}{c_0 + n} \frac{\hat{\boldsymbol{\beta}}^{\text{obs}} - \boldsymbol{\beta}_0}{\sigma} \right)^t \\ &\times \boldsymbol{\Omega}_n \left(\mathbf{W} + \left(\frac{c_0 + n}{n} \right)^{1/2} \frac{c_0}{c_0 + n} \frac{\hat{\boldsymbol{\beta}}^{\text{obs}} - \boldsymbol{\beta}_0}{\sigma} \right) \\ &\sim \frac{n}{c_0 + n} \chi_p^2 \left(\frac{c_0^2}{n} \frac{1}{c_0 + n} \frac{(\hat{\boldsymbol{\beta}}^{\text{obs}} - \boldsymbol{\beta}_0)^t \boldsymbol{\Omega}_n (\hat{\boldsymbol{\beta}}^{\text{obs}} - \boldsymbol{\beta}_0)}{\sigma^2} \right). \end{aligned}$$

Consequently, under the $\boldsymbol{\Omega}_0 = n^{-1}\boldsymbol{\Omega}_n$ scenario,

$$\text{ppp}(\mathbf{y}^{\text{obs}}) = F_{p,p} \left(1 + \frac{c_0}{n}, \frac{c_0^2}{c_0 + n} \frac{\kappa_n^{\text{obs}}}{\sigma^2} \right), \quad (18)$$

in terms of $\kappa_n^{\text{obs}} = n^{-1}(\hat{\boldsymbol{\beta}}^{\text{obs}} - \boldsymbol{\beta}_0)^t \boldsymbol{\Omega}_n (\hat{\boldsymbol{\beta}}^{\text{obs}} - \boldsymbol{\beta}_0)$ and the non-central F distribution function.

This also leads to an explicit formula for the cPPP ,

$$\begin{aligned} \text{cPPP}(\mathbf{y}^{\text{obs}}) &= \Pr\{\text{ppp}(\mathbf{Y}) \leq \text{ppp}(\mathbf{y}^{\text{obs}})\} \\ &= \Pr\left\{ \frac{(\hat{\boldsymbol{\beta}}(\mathbf{Y}) - \boldsymbol{\beta}_0)^t \boldsymbol{\Omega}_n (\hat{\boldsymbol{\beta}}(\mathbf{Y}) - \boldsymbol{\beta}_0)}{\sigma^2} \right. \\ &\quad \left. \geq \frac{(\hat{\boldsymbol{\beta}}^{\text{obs}} - \boldsymbol{\beta}_0)^t \boldsymbol{\Omega}_n (\hat{\boldsymbol{\beta}}^{\text{obs}} - \boldsymbol{\beta}_0)}{\sigma^2} \right\} \\ &= \Pr\left\{ \chi_p^2 \geq \frac{c_0}{c_0 + n} \frac{(\hat{\boldsymbol{\beta}}^{\text{obs}} - \boldsymbol{\beta}_0)^t \boldsymbol{\Omega}_n (\hat{\boldsymbol{\beta}}^{\text{obs}} - \boldsymbol{\beta}_0)}{\sigma^2} \right\}. \end{aligned}$$

Here we use that under prior and model conditions, $\hat{\boldsymbol{\beta}}(\mathbf{Y})$ is normal with mean $\boldsymbol{\beta}_0$ and variance $\sigma^2(1 + n/c_0)\boldsymbol{\Omega}_n^{-1}$.

5.2 Full Analysis With Unknown σ

The traditional conjugate prior in this model takes an inverse gamma for σ^2 and a normal for $\boldsymbol{\beta}$ given σ . In generalization of material of Section 2.3, we say that $(\lambda, \boldsymbol{\mu}) = (1/\sigma^2, \boldsymbol{\beta})$ comes from the $\text{GN}_p(\frac{1}{2}a_0, \frac{1}{2}b_0, \boldsymbol{\beta}_0, c_0\boldsymbol{\Omega}_0)$ distribution if

$$\begin{aligned} \lambda &\sim \text{Gam}\left(\frac{1}{2}a_0, \frac{1}{2}b_0\right) \quad \text{and} \\ (\boldsymbol{\beta}|\lambda) &\sim N_p(\boldsymbol{\beta}_0, \lambda^{-1}(c_0\boldsymbol{\Omega}_0)^{-1}). \end{aligned} \quad (19)$$

Thus $\boldsymbol{\beta}$ has prior mean $\boldsymbol{\beta}_0$ and prior variance $(E\lambda^{-1})c_0^{-1}\boldsymbol{\Omega}_0^{-1}$. We shall see that the posterior takes the form

$$\begin{aligned} &\{(\lambda, \boldsymbol{\beta})|\text{data}\} \\ &\sim \text{GN}_p\left(\frac{1}{2}(a_0 + n), \right. \\ &\quad \left. \frac{1}{2}(b_0 + Q_0^{\text{obs}} + (\hat{\boldsymbol{\beta}}^{\text{obs}} - \boldsymbol{\beta}_0)^t \mathbf{K} (\hat{\boldsymbol{\beta}}^{\text{obs}} - \boldsymbol{\beta}_0)), \right. \\ &\quad \left. \tilde{\boldsymbol{\beta}}, c_0\boldsymbol{\Omega}_0 + \boldsymbol{\Omega}_n\right), \end{aligned} \quad (20)$$

in which $Q_0^{\text{obs}} = \sum_{i=1}^n (y_i - \mathbf{x}_i^t \hat{\boldsymbol{\beta}}^{\text{obs}})^2 = \|\mathbf{y} - \hat{\boldsymbol{\mu}}^{\text{obs}}\|^2$ is the least sum of squares and

$$\begin{aligned} \tilde{\boldsymbol{\beta}} &= (c_0\boldsymbol{\Omega}_0 + \boldsymbol{\Omega}_n)^{-1}(c_0\boldsymbol{\Omega}_0\boldsymbol{\beta}_0 + \boldsymbol{\Omega}_n\hat{\boldsymbol{\beta}}^{\text{obs}}) \quad \text{and} \\ \mathbf{K} &= (c_0^{-1}\boldsymbol{\Omega}_0^{-1} + \boldsymbol{\Omega}_n^{-1})^{-1}. \end{aligned}$$

Proving this is accomplished partly through a little algebraic lemma that states that for vectors $\mathbf{a}_0$ and $\mathbf{a}_1$ and positive definite matrices $\mathbf{G}_0$ and $\mathbf{G}_1$, the following identity holds:

$$\begin{aligned} (\mathbf{x} - \mathbf{a}_0)^t \mathbf{G}_0 (\mathbf{x} - \mathbf{a}_0) + (\mathbf{x} - \mathbf{a}_1)^t \mathbf{G}_1 (\mathbf{x} - \mathbf{a}_1) \\ = (\mathbf{x} - \tilde{\mathbf{a}})^t (\mathbf{G}_0 + \mathbf{G}_1) (\mathbf{x} - \tilde{\mathbf{a}}) + (\mathbf{a}_1 - \mathbf{a}_0)^t \mathbf{M} (\mathbf{a}_1 - \mathbf{a}_0), \end{aligned}$$

where $\tilde{\mathbf{a}} = (\mathbf{G}_0 + \mathbf{G}_1)^{-1}(\mathbf{G}_0\mathbf{a}_0 + \mathbf{G}_1\mathbf{a}_1)$ and $\mathbf{M} = (\mathbf{G}_0^{-1} + \mathbf{G}_1^{-1})^{-1}$. To prove (20), start with the prior density for $(\lambda, \boldsymbol{\beta})$, which is proportional to

$$\lambda^{a_0/2-1} \exp(-\frac{1}{2}b_0\lambda) \lambda^{p/2} \exp\{-\frac{1}{2}c_0\lambda(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^t \boldsymbol{\Omega}_0 (\boldsymbol{\beta} - \boldsymbol{\beta}_0)\}.$$

The likelihood for data is proportional to $\lambda^{n/2} \exp[-\frac{1}{2}\lambda\{Q_0^{\text{obs}} + (\hat{\boldsymbol{\beta}}^{\text{obs}} - \boldsymbol{\beta})^t \boldsymbol{\Omega}_n (\hat{\boldsymbol{\beta}}^{\text{obs}} - \boldsymbol{\beta})\}]$. Using the foregoing algebraic identity, the posterior then becomes proportional to

$$\begin{aligned} &\lambda^{(a_0+n)/2-1+p/2} \\ &\times \exp\left[-\frac{1}{2}\lambda\{b_0 + Q_0^{\text{obs}} + (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})^t (c_0\boldsymbol{\Omega}_0 + \boldsymbol{\Omega}_n)^{-1} (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}) \right. \\ &\quad \left. + (\hat{\boldsymbol{\beta}}^{\text{obs}} - \boldsymbol{\beta}_0)^t \mathbf{K} (\hat{\boldsymbol{\beta}}^{\text{obs}} - \boldsymbol{\beta}_0)\}\right]. \end{aligned}$$

This is of the required form; see also Remark 3.

This leads to an accessible scheme for computing the ppp , under the (19) prior, as follows. Under model conditions, conditional on $(\boldsymbol{\beta}, \sigma)$, one knows that $\hat{\boldsymbol{\beta}}$ is an $N_p(\boldsymbol{\beta}, \sigma^2\boldsymbol{\Omega}_n^{-1})$, allowing $D(\mathbf{y}^{\text{rep}}, \boldsymbol{\theta})$ a representation of the form $\mathbf{V}^t \boldsymbol{\Omega}_n \mathbf{V}$, where $\mathbf{V} \sim N_p(\mathbf{0}, \boldsymbol{\Omega}_n^{-1})$; hence $D(\mathbf{y}^{\text{rep}}, \boldsymbol{\beta})$ is a χ_p^2 . To work with $D(\mathbf{y}^{\text{obs}}, \boldsymbol{\theta})$, we represent for the posterior distribution $\boldsymbol{\beta}$ for given λ as $\tilde{\boldsymbol{\beta}} + \lambda^{-1/2}\mathbf{U}$, where $\mathbf{U} \sim N_p(\mathbf{0}, (c_0\boldsymbol{\Omega}_0 + \boldsymbol{\Omega}_n)^{-1})$ and independent of the χ_p^2 distribution of $D(\mathbf{y}^{\text{rep}}, \boldsymbol{\theta})$. Thus

$$\begin{aligned} D(\mathbf{y}^{\text{obs}}, \boldsymbol{\theta}) &= \lambda(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{\text{obs}} + \lambda^{-1/2}\mathbf{U})^t \boldsymbol{\Omega}_n (\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{\text{obs}} + \lambda^{-1/2}\mathbf{U}) \\ &= (\lambda^{1/2}(c_0\boldsymbol{\Omega}_0 + \boldsymbol{\Omega}_n)^{-1}c_0\boldsymbol{\Omega}_0(\hat{\boldsymbol{\beta}}^{\text{obs}} - \boldsymbol{\beta}_0) + \mathbf{U})^t \boldsymbol{\Omega}_n \\ &\quad \times (\lambda^{1/2}(c_0\boldsymbol{\Omega}_0 + \boldsymbol{\Omega}_n)^{-1}c_0\boldsymbol{\Omega}_0(\hat{\boldsymbol{\beta}}^{\text{obs}} - \boldsymbol{\beta}_0) + \mathbf{U}). \end{aligned}$$

It is now relatively easy to simulate a large number A of $(\lambda_j, \mathbf{U}_j)$ replicates, with

$$\lambda \sim \text{Gam}\left(\frac{1}{2}a_n, \frac{1}{2}b_n\right),$$

where

$$\begin{aligned} a_n &= a_0 + n \quad \text{and} \\ b_n &= b_0 + Q_0^{\text{obs}} + (\hat{\boldsymbol{\beta}}^{\text{obs}} - \boldsymbol{\beta}_0)^t \mathbf{K} (\hat{\boldsymbol{\beta}}^{\text{obs}} - \boldsymbol{\beta}_0). \end{aligned}$$

This gives replicates of $D(\mathbf{y}^{\text{obs}}, \boldsymbol{\theta}_j)$ and the required simulation approximation,

$$\text{ppp}(\mathbf{y}^{\text{obs}}) \doteq \frac{1}{A} \sum_{j=1}^A \Pr\{\chi_p^2 \geq D(\mathbf{y}^{\text{obs}}, \boldsymbol{\theta}_j)\}. \quad (21)$$

This is a more precise estimate than the general-recipe version used in (3); option (21) is available here because $D(\mathbf{y}^{\text{rep}}, \boldsymbol{\theta})$ has the known χ_p^2 distribution, regardless of $\boldsymbol{\theta}$.

Formula (21) also lends itself to the computation of the cPPP, through a double-simulation regime that computes $\text{ppp}(\mathbf{y}_k)$ for many simulated dataset, in accordance with the general (13) strategy.

5.3 The Case of Proportional $\boldsymbol{\Omega}_0$ and $\boldsymbol{\Omega}_n$

Important simplifications are found for the special case where the prior variance of $\boldsymbol{\beta}$ is specified as being proportional to the sample variance of its least squares estimator. In that case we may again take $\boldsymbol{\Omega}_0 = n^{-1}\boldsymbol{\Omega}_n$, as before, thereby also giving a more precise interpretation of c_0 in relation to sample size. With efforts similar to those exuded in the previous section, we find that

$$D(\mathbf{y}^{\text{obs}}, \boldsymbol{\theta}) = \frac{n}{c_0 + n} \chi_p^2 \left(\lambda \frac{c_0^2}{c_0 + n} \frac{1}{n} (\hat{\boldsymbol{\beta}}^{\text{obs}} - \boldsymbol{\beta}_0)^t \boldsymbol{\Omega}_n (\hat{\boldsymbol{\beta}}^{\text{obs}} - \boldsymbol{\beta}_0) \right),$$

leading to the formula

$$\begin{aligned} \text{ppp}(\mathbf{y}^{\text{obs}}) &= \Pr \left\{ \chi_p^2 \geq \frac{n}{c_0 + n} \chi_p^2 \left(\lambda \frac{c_0^2}{c_0 + n} \kappa_n^{\text{obs}} \right) \right\} \\ &= \int_0^\infty F_{p,p} \left(\frac{c_0 + n}{n}, \lambda \frac{c_0^2}{c_0 + n} \kappa_n^{\text{obs}} \right) g_n(\lambda) d\lambda, \end{aligned} \quad (22)$$

in terms of $\kappa_n^{\text{obs}} = n^{-1}(\hat{\boldsymbol{\beta}}^{\text{obs}} - \boldsymbol{\beta}_0)^t \boldsymbol{\Omega}_n (\hat{\boldsymbol{\beta}}^{\text{obs}} - \boldsymbol{\beta}_0)$ and the posterior gamma density g_n for λ implicit in (20). Its parameters are now $a_n = a_0 + n$ and $b_n = b_0 + Q_0^{\text{obs}} + c_0 n \kappa_n^{\text{obs}} / (c_0 + n)$. The result in (22) is a generalization of (20).

For cPPP analysis, it is not a difficult task to compute $\text{ppp}(\mathbf{y}_k)$ as in (22) for a high number B of simulated datasets, leading, by (13), to the appropriate cPPP($\mathbf{y}^{\text{obs}}$) number. Special properties of the linear regression model make it possible to simplify this step, as we now demonstrate.

Proposition 2. Suppose that the prior for $(\sigma, \boldsymbol{\beta})$ is as in (19), with $\boldsymbol{\Omega}_0 = n^{-1}\boldsymbol{\Omega}_n$, and that $\mathbf{Y}$ given these parameters really follows the linear regression model $N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$. Letting $Z \sim \chi_p^2$, the ppp distribution is

$$G(u) = \Pr\{\text{ppp}(\mathbf{Y}) \leq u\} = \Pr \left\{ F_{p,p} \left(1 + \frac{c_0}{n}, \frac{c_0}{n} Z \right) \leq u \right\}.$$

This makes it easy to compute the necessary cPPP($\mathbf{y}^{\text{obs}}$) = $G(\text{ppp}(\mathbf{y}^{\text{obs}}))$ as the relative frequency of $F_{p,p}(1 + c_0/n, (c_0/n)Z_k) \leq \text{ppp}(\mathbf{y}^{\text{obs}})$ events, across 1 million copies Z_k from the χ_p^2 , because the noncentral $F_{p,p}$ function is implemented in a software package like R; in particular, there is no need to update the inverse gamma parameters or to perform the numerical integration in (22) each time. The G distribution is close to a uniform for large c_0 (corresponding almost to a classic p value for testing $\boldsymbol{\beta} = \boldsymbol{\beta}_0$), and is very tightly concentrated around $\frac{1}{2}$

for small c_0 (corresponding to a nearly noninformative prior), with a sharp upper bound at $F_{p,p}(1 + c_0/n, 0)$. The case study presented in the following section has $c_0 = 6.25$, an intermediate case.

Proof of Proposition 2. Let $\mathbf{H} = \mathbf{X}(\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t = \mathbf{X} \boldsymbol{\Omega}_n^{-1} \mathbf{X}^t$ be the familiar hat matrix. It is symmetric and idempotent, and a standard result is that, for given $(\sigma, \boldsymbol{\beta})$,

$$\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{H}\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{H}),$$

$$\hat{\boldsymbol{\epsilon}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} = (\mathbf{I} - \mathbf{H})\mathbf{Y} \sim N_n(\mathbf{0}, \sigma^2(\mathbf{I} - \mathbf{H})),$$

and these two n -vectors are independent. Taking (19) into account for the distribution of $\boldsymbol{\beta}$ given σ , we find that the vectors remain independent, for given σ , with

$$\mathbf{X}\hat{\boldsymbol{\beta}} \sim N_n(\mathbf{X}\boldsymbol{\beta}_0, \sigma^2(1 + n/c_0)\mathbf{H}) \quad \text{and}$$

$$\hat{\boldsymbol{\epsilon}} \sim N_n(\mathbf{0}, \sigma^2(\mathbf{I} - \mathbf{H})).$$

This implies that $Q_0(\mathbf{Y}) = \|\hat{\boldsymbol{\epsilon}}\|^2 \sim \sigma^2(\mathbf{V}^*)^t \mathbf{V}^*$ and

$$\kappa_n(\mathbf{Y}) = n^{-1}(\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}_0)^t (\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}_0) = n^{-1}\sigma^2(1 + n/c_0)\mathbf{V}^t \mathbf{V},$$

in terms of $\mathbf{V}^* \sim N_n(\mathbf{0}, \mathbf{I} - \mathbf{H})$ and $\mathbf{V} \sim N_n(\mathbf{0}, \mathbf{H})$, with these two being independent. Hence $\kappa_n(\mathbf{Y}) = \sigma^2(n^{-1} + c_0^{-1})Z$ and $Q_0(\mathbf{Y}) = \sigma^2 Z^*$, where Z and Z^* are chi-squared distributed and independent with p and $n - p$ degrees of freedom. This simplifies the updated parameters of the gamma distribution for $\lambda = 1/\sigma^2$, to $a_n = a_0 + n$ and $b_n = b_0 + W_n$, where $W_n \sim \sigma^2(Z + Z^*)$ and, more importantly, through (22),

$$\text{ppp}(\mathbf{Y}) = \Pr \left\{ \chi_p^2 \geq \frac{n}{c_0 + n} \chi_p^2 \left(\frac{1}{\sigma^2} \frac{c_0^2}{c_0 + n} \sigma^2 \left(\frac{1}{n} + \frac{1}{c_0} \right) Z \right) \right\},$$

in terms of a central and a noncentral χ_p^2 variable that are both independent of Z . The simplifying point is that the σ^2 factors cancel out. The claim follows.

Remark 5. When sample size n is large, $n^{-1}\boldsymbol{\Omega}_n$ will be close to a limiting covariance matrix $\boldsymbol{\Sigma}$ for the covariates and $\hat{\boldsymbol{\beta}}^{\text{obs}}$ will be close to the true regression coefficient vector $\boldsymbol{\beta}_{\text{tr}}$. Some analysis shows that $\text{cPPP}(\mathbf{y}^{\text{obs}}) \rightarrow \Pr\{\chi_p^2 \geq \text{conf}^2\}$, where

$$\text{conf} = \frac{c_0^{1/2} \{(\boldsymbol{\beta}_{\text{tr}} - \boldsymbol{\beta}_0)^t \boldsymbol{\Sigma} (\boldsymbol{\beta}_{\text{tr}} - \boldsymbol{\beta}_0)\}^{1/2}}{\sigma}$$

acts as a conflict measure between prior and model, also dependent on prior sample size c_0 . This generalizes Remark 4; see also Section 10.1.

6. CASE STUDY: REGRESSING SPEEDSKATERS

In the annual World Sprint Championships in speedskating, competitors skate the 500-m and 1,000-m distances on Saturday, and then the same distances on Sunday. The champion is the skater with the lowest combined score, defined as $t_1 + t_2/2 + t_3 + t_4/2$, where t_1, t_2, t_3 , and t_4 are the results (in seconds) of the skater for the four distances. Table 1 gives the top of the result list for the 2005 championships, held at the Olympic Oval in Salt Lake City (with 33 participants from 12 countries). In this section we illustrate the general methods of Section 5 for the regression model that pertains to predicting and understanding the result of the fourth distance in terms of times achieved for the previous three distances. In this case

Table 1. The Best 10 Skaters in the 2005 World Sprint Championships, Held in Salt Lake City, January 22–23

1. Erben Wennemars	137.310	34.75 (4)	1:08.30 (4)	34.68 (3)	1:07.46 (1)
2. Jeremy Wotherspoon	137.820	34.65 (3)	1:08.40 (5)	34.67 (1)	1:08.60 (8)
3. Joey Cheek	137.975	34.88 (8)	1:08.59 (7)	34.70 (4)	1:08.20 (4)
4. Masaaki Kobayashi	138.050	34.76 (5)	1:08.28 (3)	34.90 (7)	1:08.50 (5)
5. Dmitri Lobkov	138.100	34.63 (2)	1:08.78 (9)	34.67 (1)	1:08.82 (10)
6. Jan Bos	138.470	35.25 (14)	1:08.00 (1)	35.26 (11)	1:07.92 (3)
7. Shani Davis	138.715	35.43 (20)	1:08.04 (2)	35.43 (17)	1:07.67 (2)
8. Casey FitzRandolph	138.770	34.99 (9)	1:08.84 (11)	35.11 (10)	1:08.50 (5)
9. Joji Kato	139.000	34.80 (6)	1:09.54 (17)	34.89 (6)	1:09.08 (14)
10. Hiroyasu Shimizu	139.230	35.12 (10)	1:09.49 (16)	34.82 (5)	1:09.09 (15)

NOTE: The results are the combined point sum, then the 500-m and 1,000-m results for Saturday, followed by the 500-m and 1,000-m results for Sunday (with numbers in parentheses indicating ranking).

there would be substantial prior knowledge (along with excitement and speculations) among skaters and the millions of television viewers relating to the parameters and the fruitfulness of the model, making Bayesian analysis a relevant and interesting enterprise.

6.1 Setting the Prior

The natural model to consider takes

$$y_i = b_0 + b_1 x_{i,1} + \cdots + b_p x_{i,p} + \varepsilon_i \\ = b_0 + \mathbf{x}_i^t \mathbf{b} + \varepsilon_i \quad \text{for } i = 1, \dots, n,$$

in terms of a p -dimensional covariate vector $\mathbf{x}_i$ for each of n individuals, where the ε_i 's are independent mean-0 normals with standard deviation σ . For the speedskating applications, y_i is the 1,000-m Sunday result, whereas $x_{i,1}$, $x_{i,2}$, and $x_{i,3}$ are the results of the three previous distances. Our model is intended to convey the main mechanism of achievements during the World Championships among, say, the 25–30 best skaters of the world, during races without falls or accidents. Such “negative outliers” are therefore screened out and not allowed to enter the model, whereas, on the other hand, unusually splendid results, like a new world record, are certainly allowed. We test this model with ppp scrutiny associated with discrepancy measure (13).

Setting a clear prior for (b_0, b, σ) is not an easy exercise. It is not quite sufficient to have (b_1, b_2, b_3) centered at $(0, 1, 0)$, even though $x_{i,2}$ is a reasonable guess for y_i ; this viewpoint does not take into account that y_i also tends to be positively associated with both $x_{i,1}$ and $x_{i,3}$. One option here would be to use last year's tables of results to provide that competition's posterior distribution of parameters as a prior for this year's model, perhaps scaled down in precision so as to not be too informative. To make the exercise more realistic, however, having likely future applications of our theory in mind, we adopt the attitude that we should construct our prior from subjective (but substantiated) beliefs about the parameters. We argue that the prior knowledge in this situation (and, we suggest, in many others) is most easily quantified in terms of the overall level and its variability (expected average and standard deviation for the y_i 's) and the correlations between covariates $x_{i,j}$ and the outcome y . We therefore develop a method for bringing such prior knowledge into a proper prior for the model parameters.

It is helpful to first center the covariates by subtracting the averages $\bar{x}_{\cdot,j}$ from the foregoing $x_{i,j}$'s. Considering this done, $\bar{y} = b_0 + \bar{\varepsilon}$, giving b_0 a clear interpretation as the expected average result. Furthermore, $n^{-1} \sum_{i=1}^n y_i \mathbf{x}_i = \mathbf{S}_n \mathbf{b} + n^{-1} \sum_{i=1}^n \mathbf{x}_i \varepsilon_i$, where $\mathbf{S}_n = n^{-1} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^t$ is the empirical variance matrix of

the covariate vectors with elements, say, $s_{j,k}$ for $j, k = 1, \dots, p$. Although the considerations that follow actually might also invite other types of priors, we wish for this illustration to follow the development and recipes of Section 5, so we specify a prior for σ and for the regression curve parameters given σ of the form

$$\begin{pmatrix} b_0 \\ \mathbf{b} \end{pmatrix} | \sigma \sim N_{p+1} \left(\begin{pmatrix} \bar{b}_0 \\ \bar{\mathbf{b}} \end{pmatrix}, \sigma^2 \frac{1}{c_0} \begin{pmatrix} 1 & 0 \\ 0 & \mathbf{S}_n^{-1} \end{pmatrix} \right). \quad (23)$$

Because $n^{-1} \sum_{i=1}^n (y_i - \bar{y})^2$ is seen to have mean value $\mathbf{b}^t \mathbf{S}_n \mathbf{b} + \sigma^2(1 - 1/n)$, which we approximate with $\mathbf{b}^t \mathbf{S}_n \mathbf{b} + \sigma^2$, we see that the correlations $\boldsymbol{\rho} = (\rho_1, \dots, \rho_p)^t$ between $\mathbf{y}$ and the respective covariates $x_{i,j}$ may be represented as

$$\rho_j = \frac{(\mathbf{S}_n \mathbf{b})_j}{(\mathbf{b}^t \mathbf{S}_n \mathbf{b} + \sigma^2)^{1/2} s_j} \quad \text{for } j = 1, \dots, p, \quad (24)$$

where $s_j = s_{j,j}^{1/2}$ is the standard deviation of $x_{i,j}$. Therefore, we need a mechanism for turning prior information about $\boldsymbol{\rho}$ into a prior for $\mathbf{b}$. We assume that the list of covariate vectors is available to aid us in fine-tuning the prior, as with the speedskating data.

To work with this, we solve (24) for $\mathbf{b}$. They may be expressed as $\mathbf{S}_n \mathbf{b} / (\mathbf{b}^t \mathbf{S}_n \mathbf{b} + \sigma^2)^{1/2} = \mathbf{v} = \mathbf{D}_n^{1/2} \boldsymbol{\rho}$, where $\mathbf{D}_n = \text{diag}(s_1^2, \dots, s_p^2)$ is the matrix with empirical covariate variances down its diagonal; thus $\mathbf{v}$ has components $s_j \rho_j$. The solution is $\mathbf{b} = \sigma \mathbf{S}_n^{-1} \mathbf{v} / (1 - \mathbf{v}^t \mathbf{S}_n^{-1} \mathbf{v})^{1/2}$. To follow the intended scheme, we take

$$\mathbf{S}_n^{-1/2} \mathbf{v} / (1 - \mathbf{v}^t \mathbf{S}_n^{-1} \mathbf{v})^{1/2} = \mathbf{z}_0 / \sigma + \tau \mathbf{N} \sim N_p(\mathbf{z}_0 / \sigma, \tau^2 \mathbf{I}_p)$$

for a suitable location $\mathbf{z}_0$ and scale parameter τ , in terms of an $\mathbf{N} \sim N_p(\mathbf{0}, \mathbf{I}_p)$. This also means that

$$(\mathbf{b} | \sigma) \sim N_p(\mathbf{S}_n^{-1/2} \mathbf{z}_0, \sigma^2 \tau^2 \mathbf{S}_n^{-1}), \quad (25)$$

which is consistent with the conjugate setup of (19) and (23). To fine-tune the parameters, solve for $\mathbf{v} = (s_1 \rho_1, \dots, s_p \rho_p)^t$ to find

$$\mathbf{S}_n^{-1/2} \mathbf{v} = \frac{\mathbf{z}_0 / \sigma + \tau \mathbf{N}}{(1 + \|\mathbf{z}_0 / \sigma + \tau \mathbf{N}\|^2)^{1/2}}. \quad (26)$$

Our strategy is as follows:

1. Put up prior guess parameters $\rho_{j,0}$ for the correlations $\rho_1, \dots, \rho_p$. This must be done in a manner that reflects $\mathbf{v}^t \mathbf{S}_n^{-1} \mathbf{v} < 1$, for $\mathbf{v} = \mathbf{D}_n^{1/2} \boldsymbol{\rho}$.

- For each of a sequence of trial values of the parameter τ , solve the p equations

$$g(\mathbf{z}_0) = E \frac{\mathbf{z}_0/\sigma_0 + \tau \mathbf{N}}{(1 + \|\mathbf{z}_0/\sigma_0 + \tau \mathbf{N}\|^2)^{1/2}} \\ = \mathbf{S}_n^{-1/2} \mathbf{v}_0 = \mathbf{S}_n^{-1/2} \mathbf{D}_n^{1/2} \boldsymbol{\rho}_0 \quad (27)$$

for $\mathbf{z}_0$, also utilizing an initial prior guess value $1/\sigma_0$ for $1/\sigma$. In practice we solve these equations using stochastic simulation to compute $g_1(\mathbf{z}_0), \dots, g_p(\mathbf{z}_0)$, because no formulas are available, and a nonlinear minimization algorithm like *nlm* in the R software package to minimize $\sum_{j=1}^p \{g_j(\mathbf{z}_0) - (\mathbf{S}_n^{-1/2} \mathbf{v}_0)_j\}^2$.

- For the trial value of τ , and the found $\mathbf{z}_0$, generate $\boldsymbol{\rho}$ vectors from $\boldsymbol{\rho} = \mathbf{D}_n^{-1/2} \mathbf{v}$, where $\mathbf{v}$ is taken according to (26); in particular, we may monitor the spread of the ρ_j 's through their standard deviations and correlations (again, there are no explicit formulas to help us, hence the simulations).
- The procedure stops when a value of τ is found that well reflects the uncertainty level for the ρ_j 's around the prior guesses $\rho_{j,0}$.

The procedure would perhaps have looked more precise and elegant were it possible to produce a clear prior covariance matrix, say $\mathbf{A}$, for $\boldsymbol{\rho}$, and then choose τ to match the mean of $\mathbf{v}^t \mathbf{S}_n^{-1} \mathbf{v}$, which is $\mathbf{v}_0^t \mathbf{S}_n^{-1} \mathbf{v}_0 + \text{Tr}(\mathbf{D}_n^{1/2} \mathbf{A} \mathbf{S}_n^{-1/2})$, with the mean of $\|\mathbf{z}_0/\sigma + \tau \mathbf{N}\|^2 / (1 + \|\mathbf{z}_0/\sigma + \tau \mathbf{N}\|^2)$, from (26). The problem with this is the difficulty of setting a good $\mathbf{A}$ matrix, in that attempts at doing this would clash with the structure imposed by (26). Our strategy avoids this quandary.

6.2 ppp and cppp Analysis of the 2005 Speedskating Model

We carried out the foregoing exercise for the 2005 World Championships data. A total of six skaters were eliminated from the analysis, after falls, disqualifications, and injuries, causing some skaters to not finish all four distances. This left 27 healthy skaters in the data matrix. We used prior guess values .6, .8, and .6 for the correlations (ρ_1, ρ_2, ρ_3) of $x_{\cdot,1}, x_{\cdot,2}$, and $x_{\cdot,3}$ with $\mathbf{y}$, and $1/.75$ for $1/\sigma_0$; these values were elicited based on discussion among the authors and speedskating compatriots and are meant to be based on solid experience (for more background, see, e.g., Hjort 2003a, 2005). We had

$$\mathbf{S}_n = \begin{pmatrix} .259 & .252 & .196 \\ .252 & 1.209 & .223 \\ .196 & .223 & .239 \end{pmatrix},$$

and for each trial value τ we could solve (27) for $\mathbf{z}_0$, in terms of $\mathbf{S}_n^{-1/2} \mathbf{D}_n^{1/2} \boldsymbol{\rho}_0 = (.094, .756, .294)^t$. Monitoring spread and correlations in the prior distribution for $\boldsymbol{\rho}$, we settled on $\tau = .4$ to reflect prior beliefs. This corresponds to prior sample size $c_0 = 1/\tau^2 = 6.25$, to standard deviations .132, .084, and .133 for the three correlations, centered around .6, .8, and .6. The (ρ_1, ρ_2, ρ_3) distribution is far from normal in its constrained space determined by $\boldsymbol{\rho}^t \mathbf{D}_n^{1/2} \mathbf{S}_n^{-1} \mathbf{D}_n^{1/2} \boldsymbol{\rho} < 1$. The correlation correlations are $-.444, .634$, and $-.521$ for (ρ_1, ρ_2) , (ρ_1, ρ_3) , and (ρ_2, ρ_3) . We also have $\mathbf{S}_n^{-1/2} \mathbf{z}_0 = (.398, 1.018, .856)^t$ in (25), that is, the prior mean $\bar{\mathbf{b}}$ in (23). As a prior mean $\bar{b}_0$, for b_0 , we use the mean of all Saturday 1,000-m results, which is 1:09.46.

It only remains to specify suitable parameters $(\frac{1}{2}a_0, \frac{1}{2}b_0)$ for the gamma prior for $\lambda = 1/\sigma^2$, to keep with the (19) recipe. Toward this end, we may first show that

$$\mathbf{b}^t \mathbf{S}_n \mathbf{b} = \sigma^2 \frac{\mathbf{v}^t \mathbf{S}_n^{-1} \mathbf{v}}{1 - \mathbf{v}^t \mathbf{S}_n^{-1} \mathbf{v}} \quad \text{and}$$

$$\kappa^2 = \text{var } y_i = \frac{\sigma^2}{1 - \mathbf{v}^t \mathbf{S}_n^{-1} \mathbf{v}},$$

which leads to the identity $\sigma = \kappa(1 - \mathbf{v}^t \mathbf{S}_n^{-1} \mathbf{v})^{1/2}$ in terms of the standard deviation κ for the y_i results themselves. But for κ we have quite relevant prior information after watching Saturday's 1,000-m results, with standard deviation $\hat{\kappa} = 1.121$. Therefore, we select $(\frac{1}{2}a_0, \frac{1}{2}b_0)$ such that, first,

$$E \frac{1}{\sigma^2} = \frac{a_0}{b_0} \quad \text{matches} \\ \frac{1}{(\sigma^*)^2} = \frac{1}{\hat{\kappa}^2} \frac{1}{1 - \mathbf{v}^t \mathbf{S}_n^{-1} \mathbf{v}} = \frac{1}{.580^2},$$

and second, such that $\text{var}(1/\sigma^2) = 2a_0/b_0^2$ matches the estimated variance,

$$\text{var} \left(\frac{1}{\hat{\kappa}^2} \frac{1}{1 - \mathbf{v}^t \mathbf{S}_n^{-1} \mathbf{v}} \right) = \frac{1}{\kappa^4} \frac{2m^2}{(m-2)^2(m-4)} \frac{1}{(1 - \mathbf{v}^t \mathbf{S}_n^{-1} \mathbf{v})^2} \\ \doteq \frac{1}{.580^4} \frac{2m^2}{(m-2)^2(m-4)}.$$

Here we use $\hat{\kappa}^2 \sim \kappa^2 \chi_m^2/m$ with $m = n - 1$. The procedure yields $(a_0, b_0) = (18.746, 6.307)$.

After the hard work of setting a proper prior, it is a pleasure to finally watch Sunday's 1,000-m races to easily produce the ppp and cppp numbers, following the results of Section 5, and update our prior for the correlation and regression parameters. The correlation uncertainty updating is of interest for the actual application [the posterior is, e.g., centered at about (.45, .79, .42)] but is uncorrelated with the main focus of this article, so we report instead on the ppp analysis. Formula (22) gives $\text{ppp}(\mathbf{y}^{\text{obs}}) = .444$. Applying Proposition 2, we can check, through 1 million simulations, that the $\text{ppp}(\mathbf{Y})$ distribution in this case has 90% of its probability mass on [.404, .563], for example, so .444 is closer to an actual surprise than what an ordinary (but here naive) p value interpretation would suggest. We find that $\text{cppp}(\mathbf{y}^{\text{obs}}) = .135$. We may conclude that there is no conflict between our carefully constructed prior and the data, as monitored through the canonical discrepancy function $D(\mathbf{y}, \boldsymbol{\theta}) = \sum_{i=1}^n (\hat{\mu}_i - \mu_i)^2 / \sigma^2$. It might be of interest to try other discrepancy functions, checking adequacy of other aspects of prior and model, like asymmetry, but we abstain from doing that here.

7. ppp AND cppp ANALYSIS OF TWO BIRD SURVIVAL MODELS

Brooks, Catchpole, and Morgan (2000) (henceforth BCM) analyzed recapture data for the European dipper species (*Cinclus cinclus*) using ppp values. The same data were previously analyzed by Lebreton, Burnham, Clobert, and Anderson (1992). They are in the form of a triangular 6×6 array in conjunction with a vector giving the number of released individuals

for the 6 years in question. Here we offer a reanalysis of their data, using our cppp tools, and reach conclusions that differ somewhat from those of BCM regarding the adequacy of suggested models.

7.1 The Data, the Model, and the Discrepancy

The data are given in Table 2. The models considered by BCM involve up to 12 parameters: for $i = 1, \dots, 6$, ϕ_i represents the probability that a given bird survives year $1980 + i$, whereas p_i represents the probability of capturing a particular bird in year $1980 + i$.

The multiplicative multinomial model assumption can for such recapture data be shown to lead to a likelihood of the form

$$L(\boldsymbol{\phi}, \mathbf{p}) = \text{const } \Delta(\boldsymbol{\phi}, \mathbf{p}) \prod_{i=1}^6 \prod_{j=i+1}^6 \left\{ \phi_i p_j \prod_{k=i+1}^{j-1} \phi_k (1 - p_k) \right\}^{y_{i,j}},$$

where Δ is the contribution to the likelihood for the never-captured birds, and where “empty products” are equal to 1; see BCM for details. The discrepancy measure used in BCM is

$$D(\mathbf{y}, (\boldsymbol{\phi}, \mathbf{p})) = \sum_{i,j} (y_{i,j}^{1/2} - e_{i,j}^{1/2})^2, \quad \text{where } e_{i,j} = E_{\text{model}} y_{i,j}.$$

We have analyzed two versions of this general model. The large model (T/T in the terminology of BCM) uses all 12 parameters (ϕ_i, p_i) while the small model (C/C in their terminology) takes all ϕ_j 's equal and all p_j 's equal, thus having two parameters.

For both cases we choose to use independent and uniform priors for the parameters in question. We do this for two reasons. First, the more focused priors also briefly worked with by BCM do not appear to fit data particularly well, and BCM gave little indication that they represented actual prior knowledge of the parameters. Second, BCM quoted the posterior means and standard deviations only for the uniform prior. It turns out that our computed ppp values are slightly, but statistically significantly, different from those quoted by BCM. For the full model with uniform prior we find a ppp value of .075, whereas BCM quoted .086; similarly, for the small model we find ppp equal to .060, where BCM gave .069. By verifying the quoted posterior means and standard deviations, we have eliminated the MCMC simulation as a possible reason for these differences. Although we have not been able to reproduce the exact ppp values of BCM, the differences are small enough to be ignored in the present setting of general analysis.

Table 2. Recapture Data for the European Dipper

Release year	Released	Recaptured					
		1982	1983	1984	1985	1986	1987
1981	22	11	2	0	0	0	0
1982	60		24	1	0	0	0
1983	78			34	2	0	0
1984	80				45	1	2
1985	88					51	0
1986	98						52

7.2 ppp and cppp Analysis of the Two Models

The Large Model. We sampled the distribution of ppp along the lines indicated earlier and found $\text{ppp}(\mathbf{y}^{\text{obs}}) = .075$. From plots (not shown here) of the sampled cumulative distribution function and estimated density for the ppp, based on 500 simulated values, we see that the distribution function is clearly S-shaped and the density is clearly unimodal. The sample standard deviation is .172, compared with .289 for a uniformly distributed p value. More important is the fact that only 1 of our 500 simulated ppp values was $<.075$, giving an estimated cppp value (the true surprise level) of .002.

The Small Model. In this case, we found $\text{ppp}(\mathbf{y}^{\text{obs}}) = .060$. Again we studied the sample cumulative distribution function and estimated density for the ppp. We found here that the ppp distribution is closer to a uniform distribution compared with the 12-parameter case. The sample standard deviation of ppp is now .257. With calibration, we find $\text{cppp} = .022$ [11 out of 500 simulated $\text{ppp}(\mathbf{Y})$ values were $<.060$].

We therefore conclude that the given data are much less likely under the large model than under the small model, although the nominal ppp values led BCM to the opposite conclusion.

8. COMPARING NONPARAMETRIC WITH PARAMETRIC MODELS

Our cppp can be interpreted as a “quantification of surprise,” as monitored by the discrepancy function. It is a strength of this approach that it puts all surprises on the same footing, so to speak, namely the uniform scale on the unit interval. In particular, there are no conceptual difficulties with comparing the cppp from a parametric model specification with that of a nonparametric model specification. We indicate how this might work with in two situations.

8.1 Nonparametric versus Parametric cdf

Consider independent data $y_1, \dots, y_n$ drawn from a distribution F , where two quite different priors are under consideration for F . The first of these is to take $F(t) = \Phi((t - \mu)/\sigma)$ normal, with a prior $\pi(\mu, \sigma)$ on its parameters. The second does not bound F to any parametric description and takes $F \sim \text{Dir}(aF_0)$, a Dirichlet process with center distribution F_0 and concentration parameter a (see, e.g., Hjort 2003b for a review of nonparametric Bayesian statistics). For discrepancy measure, we use

$$D(\mathbf{y}, F) = \sqrt{n} \|F_n - F\| = \sqrt{n} \max_t |F_n(t) - F(t)|,$$

in terms of the empirical distribution function F_n of the data. The $\sqrt{n}$ factor is of no consequence for the actually computed ppp and cppp numbers, but is given to better understand the situation when n grows. In fact, the $D(\mathbf{y}^{\text{rep}}, F) = \sqrt{n} \|F_n^{\text{rep}} - F\|$ is then close to being the maximum absolute value of $W^0(F(t))$, where W^0 is a Brownian bridge, by classic empirical process theory (see, e.g., Billingsley 1968, chap. 4). In particular, $D(\mathbf{y}^{\text{rep}}, F) \rightarrow_d \|W^0\| = \max_t |W^0(t)|$, both when F is fixed and when F is selected by some posterior mechanism.

Let us now assume that the y_i data really follow some continuous F_{tr} distribution. We aim at studying the behavior of the nonparametric and normal-parametric prior specifications, say

ppp_{Dir} and ppp_N. First, we study the nonparametric prior. Here F given data is a Dirichlet process with parameter $aF_0 + nF_n^{\text{obs}}$. We may show that $\sqrt{n}(F - F_n^{\text{obs}}) \rightarrow_d \tilde{W}^0(F_{\text{tr}}(\cdot))$ a.s., where $\tilde{W}^0$ is another Brownian bridge, independent of W^0 ; this follows from work of Hjort and Ongaro (2005). Hence

$$\begin{aligned} \text{ppp}_{\text{Dir}}(\mathbf{y}^{\text{obs}}) &= \Pr\{\sqrt{n}\|F_n^{\text{rep}} - F\| \geq \sqrt{n}\|F_n^{\text{obs}} - F\| \mid \text{data}\} \\ &\rightarrow \Pr\{\|W^0\| \geq \|\tilde{W}^0\|\} = \frac{1}{2} \end{aligned}$$

a.s., as n grows toward infinity. Then we consider the normal-parametric prior, for which F given data is a random normal distribution function, with (μ, σ) drawn from the appropriate posterior density $\pi(\mu, \sigma \mid \text{data})$. Here $\|F_n^{\text{obs}} - F\|$ goes a.s. to $\|F_{\text{tr}} - F_{\text{appr}}\|$, essentially by the Glivenko–Cantelli theorem, where $F_{\text{appr}}(t) = \Phi((t - \mu_{\text{tr}})/\sigma_{\text{tr}})$ is the best parametric approximant to $F_{\text{tr}}(t)$ inside the normal family. It follows that $\sqrt{n}\|F_n^{\text{obs}} - F\|$ goes to infinity a.s. as long as the real F_{tr} is not fully equal to a normal distribution function and, in particular, $\text{ppp}_{\text{N}}(\mathbf{y}^{\text{obs}}) \rightarrow 0$ a.s. as $n \rightarrow \infty$.

The foregoing results help us understand the behavior of the two ppp measures for large n . We would also need to calibrate these measures to reach the more interpretable ccpp values given data. We could do this through double simulation, in accordance with the general guidelines on this laid out in Section 4.2.

In some situations interest focuses more on some interest parameter $\kappa(F)$ than on the full distribution F , say, the interquartile range or the skewness. In such cases we might prefer working with discrepancy measures of the type $D(\mathbf{y}, F) = n\{\kappa(F_n) - \kappa(F)\}^2$. Here the parametric model could win, even when it is not fully correct as such. Again, ppp and ccpp analysis may be carried out.

8.2 Nonparametric versus Parametric Hazard Rate Models

Bayesian analysis of survival and event history data appear to fall into one of two separate categories, the parametric and the nonparametric. Only rarely do we see any formal justification for choosing one path over the other. In principle, the ccpp analysis makes it easy, conceptually and operationally, to do such a comparison. The parametric model might use a Weibull cumulative hazard rate $H(t) = (\theta t)^\gamma$ with a prior on (θ, γ) , whereas the nonparametric alternative could use a beta process. As discrepancy measure, we might use $D(\mathbf{y}, H) = \int_0^\tau w(t)\{H_n(t) - H(t)\}^2 dt$ for a weight function w , involving the Nelson–Aalen estimator H_n for H over a time observation window $[0, \tau]$.

9. ppp AND ccpp: FACTS AND FICTION

Various issues related to the rapidly evolving area of Bayesian model evaluation border on the controversial, and there are certainly conflicting views on the use and limitations of ppp measures and various other simulation-based model and model-prior diagnostics. Here we attempt to clarify our own views on some of these issues, in the light of our findings and applications.

9.1 The ppp Needs a Scale

A ppp number computed from data is a well-defined probability, but it has a difficult interpretation, due to the double

use of data aspect of its construction. In particular, an observed ppp = .28, say, cannot be taken as support for or criticism against the model, and neither can it easily be compared with other ppp values. In some cases .28 can signal extreme surprise, in the sense of being a highly unlikely value under model conditions. This may also be true for a value of, say, .52, even when the mean value of the ppp distribution is .50. Thus an appropriate scale is called for to compute our surprise level, and this scale must take into account which datasets are considered likely or not. We consider the ccpp of (4) a fairly canonical calibration, transforming the ppp to the uniform scale. Other transformations are sometimes relevant, as discussed in Section 4.3.

For an illustration, we return to the speed of light data described in the Example of Section 1, where the discrepancy $D = |y_{(61)} - \mu| - |\mu - y_{(6)}|$ is used for assessing normality in the middle of the distribution. We now examine the real surprise level of ppp scores under some natural priors, namely the inverse gamma–normal class defined in (6) with parameters $(a_0, b_0, \mu_0, c_0) = (c/25, c, 30, c)$. This corresponds to prior guess 30 for μ and prior mean $a_0/b_0 = 1/25$ for $1/\sigma^2$, whereas c , in a reasonable sense, is the prior sample size, in view of the updating mechanism for GN priors laid out in Section 2.3. In particular, the vague prior case corresponds to $c \rightarrow 0$. The following table gives ppp and ccpp values for different c values, using the double-simulation method to compute the ccpp numbers, with $A = B = 2,000$ in (3) and (13):

c	ppp	ccpp	c	ppp	ccpp
.5	.216	.067	10	.280	.146
1	.218	.069	20	.333	.227
2	.223	.083	30	.386	.335
3	.228	.089	40	.422	.396
4	.238	.093	50	.470	.466
5	.249	.107	100	.588	.619

This table indicates that $\text{ppp}(\mathbf{y}^{\text{obs}}) = .208$ may be judged to be on the statistical borderline of surprise after all (despite the conclusion in Gelman et al. 2004, p. 164). This is also consistent with results from a couple of frequentist checks of the type $T = (y_{(61)} - y_{(6)})/\hat{\sigma}$ with any robust scale estimate; normality is rejected at level .03 if $\hat{\sigma}$ is the mean-mad estimator $n^{-1} \sum_{i=1}^n |y_i - \text{med}(\mathbf{y})|$, for example.

9.2 One Discrepancy's ccpp Is not Another Discrepancy's ppp

One reaction we have met is that the ccpp for a discrepancy $D(\mathbf{y}, \theta)$ should be equivalent to the original ppp for another discrepancy $D^*(\mathbf{y}, \theta)$. But we argue that this is typically not the case. In this sense the ccpp is therefore a new invention.

The ccpp is by construction distributed as a $U(0, 1)$. Hence the desired ppp*, corresponding to D^* , must also be $U(0, 1)$. Define $h(\mathbf{z}, \theta) = \Pr\{D^*(\mathbf{y}, \theta) \geq D^*(\mathbf{z}, \theta)\}$, where $\mathbf{y} \sim f(\mathbf{y}|\theta)$. It follows from theorem 1 of Meng (1994) that $\text{ppp}^* \sim U(0, 1)$ only if $h(\mathbf{y}^{\text{obs}}, \theta)$ is constant under the posterior distribution of θ , for almost all $\mathbf{y}^{\text{obs}}$. This can happen only if the posterior distribution of θ does not affect the distribution of $\mathbf{y}|\theta$ in ways detected by D^* . In particular, defining D^* as a function of $\mathbf{y}$ only is not sufficient, because $h(\mathbf{y}^{\text{obs}}, \theta)$ will still depend on its second parameter through the distribution of $\mathbf{y}|\theta$. Essentially,

D^* can only monitor aspects of (y, θ) for which the prior distribution of θ is not random. On the other hand, the desired D^* must be able to detect the same discrepancies as D to achieve $\text{ppp}^* = \text{cPPP}$. We conclude that such a D^* is unlikely to exist except in cases where the randomness of the prior distribution is irrelevant for the discrepancy D .

9.3 Some Priors Are More Proper Than Others

There are different schools of thought among scientists that apply Bayesian methods; Good (1983) reminds us that there are at least 46,656 different types of Bayesian thinking. Purists may even imagine that the parameters of Nature were drawn according to the prior distribution when the world was created. At the other extreme, many statisticians use Bayesian methods as a means of estimating parameters through posterior means of MCMC samples, partly because this appears to be (at this stage) more computationally convenient than frequentistic estimation. The latter group typically uses improper noninformative priors, partly out of convenience and partly to be more objective, letting the data show the way. The purists, on the other hand, insist on proper priors.

From a computational standpoint, the propriety of the prior has important implications. Obviously, we cannot compute prior predictive p values, for instance, if we cannot sample from the prior. This also applies to our calibrated posterior predictive p values. We would, however, argue that the important distinction should not be between mathematically proper or improper prior distributions, but rather between reasonable and unreasonable ones. By a “reasonable prior” we mean a distribution under which it makes sense to sample. As an example of a mathematically proper, but highly unreasonable, prior, we might turn to the hazard rate analysis example of Gelman et al. (1996). To carry out Bayes analysis for the survival data in question, we need to model the probabilities θ_i of a person not surviving a year $35 + i$, given that the person survived year $34 + i$, for $i = 1, \dots, 30$. Gelman et al.’s prior is a uniform on $[0, 1]^{30}$, constrained by the assumption that the sequence $\theta_1, \dots, \theta_{30}$ is increasing and convex. Without looking at the specific data of this application, we know that θ_{30} is well below .5, most likely below .1. But under the chosen prior, $\Pr\{\theta_{30} > .5\} > 1 - 10^{-9}$; even more diabolically, under Gelman et al.’s prior, nearly every human being would die before age 65, and very few would celebrate their 60th birthday.

Therefore, it makes no sense to sample under this prior. Accordingly, our calibration method is not applicable for this application, and the problem of defining the true surprise of a computed ppp value has no obvious solution. This is a case where we might want to apply our “what-if” analysis of Section 4.3. We may keep the unreasonable prior of Gelman et al. as our fitting prior, which means using it for computing the posterior probability distribution and also for the ppp value. When calibrating this ppp value, we would use a much more informative prior, however, for example, by restricting the uniform prior on increasing convex vectors in $[0, 1]^{30}$ by $\theta_{30} \in [.03, .07]$, if we consider this a likely range. One may ask why, then we do not use this improved prior throughout the analysis. First, we may not wish to include this subjective piece of information in our otherwise objective analysis of the data. Applying a subjective prior in the process of testing the sanity of the model is

quite different from using it as a basis for posterior inference. Second, there may be a risk that our sampling prior is too narrow, failing to capture the “true” parameter value. Therefore, we may prefer to fit the model under a prior that is vague, so that the data get the chance to show the way. Similarly, we may want the sampling prior used for calibrating the ppp value to be on the narrow side, to ensure that we do not calibrate our ppp value against nonsense.

10. RELATED THEMES AND CONCLUDING REMARKS

We end this article with some comments and discussion of themes, some of which might warrant further research efforts.

10.1 Calibrating the Prior Through the Calibrated ppp

Eliciting and fine-tuning priors remains, of course, a difficult task even for experienced Bayesian statisticians. There are often situations where the statistician may translate prior knowledge into a reasonably secure center point, say a prior guess equal to θ_0 , but where setting the appropriate precision level is far from clear-cut. In such situations a not uncommon scheme or way of thinking is to try out different precision levels, from reasonably precise to quite noninformative, and make do with a level of spread that balances the two desiderata of not trusting the prior guess too much and at the same time not tolerating a serious conflict or clash between prior and data. Some prior sample size elicitation are indirectly of such a form.

The procedure just described is somewhat ad hoc, of course, and may not be easy to follow or formalize in practice. The cPPP mechanism offers a venue leading to a precise version of this idea, however. The proposal is to monitor the $\text{cPPP}(y^{\text{obs}})$ as a function of the spread or precision parameter involved, and in the end use the most conservative prior that still does not clash with data, in the sense of having, for example, $\text{cPPP}(y^{\text{obs}}) = .10$ but not lower. As a mundane illustration of this cPPP-induced conservative prior, again consider the normal–normal setup of Sections 2 and 4.1. Assume that the statistician has selected a secure prior guess parameter θ_0 for θ but is not yet certain about the choice of the prior’s spread parameter σ_0 . From (12), the foregoing idea amounts to selecting σ_0 to have

$$\frac{n(\bar{y}^{\text{obs}} - \theta_0)^2 / \sigma^2}{1 + n\sigma_0^2 / \sigma^2} = 1.645^2$$

or

$$\sigma_0 = \frac{\sigma}{\sqrt{n}} \left(\frac{z_n^{\text{obs}}}{1.645^2} - 1 \right)^{1/2},$$

where $z_n^{\text{obs}} = n(\bar{y}^{\text{obs}} - \theta_0)^2 / \sigma^2$ is the usual test statistic for testing the null hypothesis that $\theta = \theta_0$. This formula would be used when $z_n^{\text{obs}} \geq 1.645^2$, that is, when the test rejects θ_0 at significance level .10. In cases where the $\theta = \theta_0$ hypothesis is accepted by the data, the foregoing scheme would allow a sharp prior at θ_0 , that is, setting σ_0 to a small value.

We stress that this idea is quite general and might be used in much less clear-cut situations, such as in semiparametric and nonparametric contexts. One type of application would be to nonparametric setups involving Dirichlet or beta processes, where the statistician knows where to center these but is unsure about the concentration parameters.

10.2 Detection Power

A proper discussion of the detection power of the cppp assessors is beyond the scope of this article. This clearly depends on aspects of the situation at hand, including the discrepancy function $D(\mathbf{y}, \boldsymbol{\theta})$. For the normal-normal setup of Section 2, it is not difficult to study the distribution of $U = \text{ppp}(\mathbf{Y})$ under various conditions different than those implied by the prior and the model; we will report on some such findings elsewhere. Although the power may be satisfactory for some sets of alternative combinations of prior and model, there will remain types of alternatives that are difficult to detect with any given $D(\mathbf{y}, \boldsymbol{\theta})$. Detecting special types of violations may require devising corresponding special discrepancy functions.

10.3 cppp as a p Value

One way of viewing our cppp construction is that the $\text{ppp}(\mathbf{y}^{\text{obs}})$ of (1), albeit clearly having a Bayesian interpretation and inspiration, is nothing but a test statistic. Because statisticians compute it, they wish directly or indirectly to assess its significance, which amounts to comparing it to its null distribution. This is in effect what the cppp operation does.

10.4 Tail versus Height

When attempting to assess the degree of surprise in an observation $U = u$, we might compute tail areas for the relevant null distribution or a suitable ratio of densities $h(u)/g(u)$. We have focused on the first general direction, partly because it would be difficult to find good general candidates for the $h(u)$ in question and partly because the null densities $g(u)$ have been seen to be so extreme; see Figure 4.

10.5 General Hierarchical Models

In complicated hierarchical models, it is easy to lose track of the different implications of a many-level prior. Ordinary assessment methods may not work well in such cases (see, e.g., Lu et al. 2006). The cppp methods of this article may be generalized to various hierarchical models and may be useful for screening out unfortunate combinations of prior and model when used with appropriate discrepancy functions.

[Received July 2004. Revised September 2005.]

REFERENCES

- Bayarri, M. J., and Berger, J. O. (2000), "P Values in Composite Null Models" (with discussion), *Journal of the American Statistical Association*, 95, 1127–1142.
- (2004), "The Interplay of Bayesian and Frequentist Analysis," *Statistical Science*, 19, 58–80.
- Bayarri, M. J., and Castellanos, M. E. (2004), "Bayesian Checking of Hierarchical Models," technical report, University of Valencia.
- Billingsley, P. (1968), *Convergence of Probability Measures*, Singapore: Wiley.
- Box, G. E. P. (1980), "Sampling and Bayes' Inference in Scientific Modelling and Robustness," *Journal of the Royal Statistical Society, Ser. A*, 143, 383–430.
- Brooks, S. P., Catchpole, E. A., and Morgan, B. J. T. (2000), "Bayesian Animal Survival Estimation," *Statistical Science*, 15, 357–376.
- Claeskens, G., and Hjort, N. L. (2003), "The Focused Information Criterion" (with discussion), *Journal of the American Statistical Association*, 98, 900–916 and 938–945.
- Clyde, M., and George, E. (2004), "Model Uncertainty," *Statistical Science*, 19, 81–94.
- Dey, D. K., Gelfand, A. E., Swartz, T. B., and Vlachos, P. K. (1998), "A Simulation-Intensive Approach for Checking Hierarchical Models," *Test*, 7, 325–346.
- Gelfand, A. E., and Dey, D. K. (1994), "Bayesian Model Choice: Asymptotics and Exact Calculations," *Journal of the Royal Statistical Society, Ser. B*, 56, 501–514.
- Gelfand, A. E., and Ghosh, S. K. (1998), "Model Choice: A Minimum Posterior Predictive Approach," *Biometrika*, 85, 1–11.
- Gelfand, A. E., and Wang, F. (2002), "A Simulation Based Approach to Sample Size Determination Under a Given Model and for Separating Models," *Statistical Science*, 17, 193–208.
- Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004), *Bayesian Data Analysis* (2nd ed.), Boca Raton, FL: Chapman & Hall/CRC.
- Gelman, A., Meng, X.-L., and Stern, H. (1996), "Posterior Predictive Assessment of Model Fitness via Realized Discrepancies" (with discussion), *Statistica Sinica*, 6, 733–807.
- Good, I. J. (1983), *Good Thinking: The Foundations of Probability and Its Applications*, Minneapolis: University of Minnesota Press.
- Gutiérrez-Peña, E., and Walker, S. G. (2001), "A Bayesian Predictive Approach to Model Selection," *Journal of Statistical Planning and Inference*, 93, 259–276.
- Guttman, I. (1967), "The Use of the Concept of a Future Observation in Goodness-of-Fit Problems," *Journal of the Royal Statistical Society, Ser. B*, 29, 83–100.
- Hjort, N. L. (2003a), "Catriona LeMay Doan: Going for Gold" [book review], *SpeedSkating World*, 8, 14–15.
- (2003b), "Topics in Nonparametric Bayesian Statistics" (with discussion), in *Highly Structured Stochastic Systems*, eds. P. J. Green, N. L. Hjort, and S. Richardson, Oxford, U.K.: Oxford University Press, pp. 455–478.
- (2005), "The Bolshoi of Speedskating," *SpeedSkating World*, 10, 14–17.
- Hjort, N. L., and Claeskens, G. (2003), "Frequentist Model Average Estimators" (with discussion), *Journal of the American Statistical Association*, 98, 879–899 and 938–945.
- Hjort, N. L., and Ongaro, A. (2005), "Exact Inference for Random Dirichlet Means," *Statistical Inference for Stochastic Processes*, 8, 227–254.
- Kadane, J. B., and Lazar, N. A. (2004), "Methods and Criteria for Model Selection," *Journal of the American Statistical Association*, 99, 279–290.
- Kass, R. E., and Raftery, A. E. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795.
- Lebreton, J.-D., Burnham, K. P., Clobert, J., and Anderson, D. R. (1992), "Modeling Survival and Testing Biological Hypotheses Using Marked Animals: A Unified Approach With Case Studies," *Ecological Monographs*, 62, 67–118.
- Lehmann, E. L. (1983), *Theory of Point Estimation*, New York: Wiley.
- Lu, H., Hodges, J. S., and Carlin, B. P. (2006), "Measuring the Complexity of Generalized Linear Hierarchical Models," *Biostatistics*, to appear.
- Marden, J. I. (2000), "Hypothesis Testing: From p Values to Bayes Factors," *Journal of the American Statistical Association*, 95, 1316–1320.
- Meng, X.-L. (1994), "Posterior Predictive p Values," *The Annals of Statistics*, 22, 1142–1160.
- O'Hagan, T. (2003), "HSSS Model Criticism" (with discussion), in *Highly Structured Stochastic Systems*, eds. P. J. Green, N. L. Hjort, and S. Richardson, Oxford, U.K.: Oxford University Press, pp. 423–453.
- Robert, C. P. (2001), *The Bayesian Choice* (2nd ed.), New York: Springer-Verlag.
- Robins, J. M., van der Vaart, A., and Ventura, V. (2000), "Asymptotic Distribution of p Values in Composite Null Models" (with discussion), *Journal of the American Statistical Association*, 95, 1143–1156.
- Rubin, D. B. (1984), "Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician," *The Annals of Statistics*, 12, 1151–1172.
- Sinharay, S., and Stern, H. S. (2003), "Posterior Predictive Model Checking in Hierarchical Models," *Journal of Statistical Planning and Inference*, 111, 209–221.
- Smith, A. F. M., and Spiegelhalter, D. J. (1980), "Bayes Factors and Choice Criteria for Linear Models," *Journal of the Royal Statistical Society, Ser. B*, 42, 213–220.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002), "Bayesian Measures of Model Complexity and Fit" (with discussion), *Journal of the Royal Statistical Society, Ser. B*, 64, 583–639.
- van der Linde, A. (2004), "On the Association Between a Random Variable and an Observable," *Test*, 13, 85–111.