Post-Processing Posterior Predictive P-values

Nils Lid Hjort, Fredrik A. Dahl and Gunnhildur Högnadóttir Steinbakk

University of Oslo

– *July 2004* –

ABSTRACT: This article addresses issues of model choice in Bayesian contexts, and focusses on the use of the so-called posterior predictive p-values (ppp values). 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 for purposes of comparing different combinations of prior and model. The distribution of such ppp values is however far from uniform, 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 with nonparametric model specifications, in that genuine 'measures of surprise' are put on the same canonical uniform scale. Our techniques are illustrated 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, quantification of surprise*

1. Introduction and summary

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 one might often need to consider several candidates for both prior and model. This leads to questions on how to meaningfully 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 remains relevant (also Bayesians might need to check if 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) and Marden (2000). Related to these again is the so-called Bayesian information criterion, the BIC (see e.g. Robert, 2001, Ch. 7). A quite general model-evaluation strategy employs the deviance information criterion, the DIC, of Spiegelhalter, Best, Carlin and van der Linde (2002) (see also van der Linde, 2004); this method is in widespread use since its computation is an easy by-product of the MCMC simulations that are often used to simulate from the posterior distributions. Model adequacy evaluation tools partly geared specifically towards use in hierarchical models are proposed and discussed in Gelfand and Dey (1994), Dey, Gelfand, Swartz and Vlachos (1998), O'Hagan (2003), Lu, Hodges and Carlin (2004), 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; references here include Gelfand and Ghosh (1998), Gutiérrez-Peña and Walker (2001), Claeskens and Hjort (2003), Hjort and Claeskens (2003), and Kadane and Lazar (2004).

In addition, various authors have attempted to construct 'Bayesian p-values', which can be thought of as quantifying the degree of surprise from data, given the prior and the model, sometimes also focussing on certain hypotheses. The Bayesian p-values come in many forms, and range from the prior predictive p-values of Box (1980) to the posterior predictive p-values touched on in Guttman (1967) and Rubin (1984), a tool worked out more fully by Gelman, Meng and Stern (1996) and Meng (1995). Important variations and improvements are introduced in Bayarri and Berger (2000), further discussed and analysed in Robins, van der Vaart and Ventura (2000), and in Bayarri and Castellanos (2004).

1.2. The ppp. This article focusses on one of the above-mentioned mechanisms for carrying out such evaluation and comparisons, namely the so-called posterior predictive p-value (henceforth, the 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, Meng and Stern (1996), the ppp is defined as

$$\mathrm{ppp} = \mathrm{ppp}(y^{\mathrm{obs}}) = \Pr\{D(y^{\mathrm{rep}}, \theta) \ge D(y^{\mathrm{obs}}, \theta) \,|\, \mathrm{data}\}. \tag{1.1}$$

Here y^{obs} signifies the observed data, or in some cases a suitably relevant subset of the full data-set, while y^{rep} represents a new ('future') data-set of the same type, drawn conjointly with θ from the posterior distribution $\pi(\theta \,|\, \mathrm{data})$. More concretely,

$$(\theta, y^{\mathrm{rep}}) \sim \pi(\theta \,|\, \mathrm{data}) f(y \,|\, \theta), \tag{1.2}$$

so that, in particular, y^{rep} comes from the predictive distribution $\int f(y \,|\, \theta)\pi(\theta \,|\, \mathrm{data})\, \mathrm{d}\theta$. Provided we can simulate (i) θ_js from the posterior and (ii) y_j^{rep} data-sets from the model $f(y \,|\, \theta_j)$, we may evaluate the ppp as

$$\mathrm{ppp}(y^{\mathrm{obs}}) \doteq \frac{1}{A} \sum_{j=1}^{A} I\{D(y_j^{\mathrm{rep}}, \theta_j) \ge D(y^{\mathrm{obs}}, \theta_j)\}, \tag{1.3}$$

across a high number A of simulations. It is useful to plot $D(y^{\rm obs},\theta_j)$ vs. $D(y_j^{\rm rep},\theta_j)$, whereby the ppp number is identified as the proportion of points above the diagonal.

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 upon the unknown parameters.

The ppp value of observed data $y^{\rm obs}$ has the combined intention of checking adequacy of the prior distribution $\pi(\theta)$ as well as of the model $f(y,\theta)$. In the (relatively rare) cases where we view our prior as 'the distribution Nature used when creating the world' we might prefer the prior predictive p-value (prpp), suggested by Box (1980); this is also a bona fide p-value (with null distribution uniform on $[0,1]$). The prior predictive p-values can only handle test statistics $D=D(y)$ that do not depend on the θ, however. Even in cases when the prior is taken quite literally, therefore, there is a place for ppp values.

1.3. The calibrated ppp. Loosely speaking, the ppp calculation uses the data twice; first by updating the prior to fit the data better, and then by estimating how surprising the data are, relative to the posterior parameter distribution. Thus it is not surprising that its distribution, across likely values of $y^{\rm obs}$, is not uniform; we shall in fact demonstrate various extreme aspects of non-uniformity in several situations. This makes the interpretation and comparison of ppp values a difficult and risky matter. To alleviate this problem our proposal is to post-process or calibrate the ppp value, to

$$\mathrm{cppp}(y^{\rm obs}) = \Pr\{\mathrm{ppp}(Y)\le \mathrm{ppp}(y^{\rm obs})\}, \tag{1.4}$$

where the distribution of Y is that implied by the prior and model conditions. This cppp, the perfected ppp value, will now by construction have a uniform null distribution, i.e. may be seen 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 re-scaled cppp 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, Meng and Stern (1996). There have been later warnings in the literature, however; Dey, Gelfand, Swartz and Vlachos (1998) complained that the ppp values were not able to pick out model inadequacies in a string of Bayesian logistic–normal regression setups, and 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. We will in fact demonstrate below that this is a fairly typical situation; 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, pointed to by other authors, are solved by the calibration mechanism alluded to above, as discussed more firmly in Section 4 and afterwards in our article. When properly calibrated, the cppp 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 prior predictive p-value, the ppp value can also 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. One may demonstrate mathematically that the ppp value becomes less and less dependent on the prior, for a given data model, as data information accumulates. The cppp value, however, will remain critically dependent on the actual prior used. This is at it should be; the calibration transform (1.4) is instructed to actively use and assess the implications of a given prior.

1.4. The present article. The lay-out of our article is as follows. In Section 2 we investigate the structurally simple situation where data are normal with a normal prior on 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 more flat prior, and to having a fixed prior with increasing sample size. Such calculations and analysis may also be extended to normal–normal hierarchical models. We also include a brief large-sample analysis of the ppp for general parametric models. We learn for example that the ppp value often becomes tightly concentrated around $\frac{1}{2}$ in situations with growing amounts of data for a fixed parametric model. Further properties and aspects of the ppp are gleaned from studying situations with a mixture of sub-priors, in Section 3.

Then in Section 4 we propose and develop our perfected ppp values, the cppp, and discuss their computation and interpretation. The cppp may in general be computed via 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 data set, pertaining to the modelling 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) have earlier analysed the same data, using the ppp apparatus, but via our calibrated ppp we reach somewhat different conclusions.

Our cppp values have by construction been transformed to a 'canonical scale of surprise', namely the uniform on the unit interval. Observed cppp numbers therefore enjoy a clear interpretation and can soundly be compared across several proposed or imagined combinations of prior and model. We may even apply the techniques to comparison of parametric vs. nonparametric model specifications, as illustrated in Section 8. Such comparisons are often quite difficult, conceptually and operationally, with other approaches. Our article ends in Section 9 with a list of ppp and cppp related topics and issues, some worthy of further research efforts.

The success or not of our cppp analysis depends crucially on the choice of discrepancy measure, the $D(y, \theta)$. This choice should be aided by context and aspects of the actual application, and might in particular be constructed to address those model aspects that are seen as crucial for the principal conclusions of the statistical analysis. See in this connection constructions of Dey, Gelfand, Swartz and Vlachos (1998), O'Hagan (2003) and 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 with the methodology and machinery of this article would mean choosing good and problem-relevant discrepancy measures, followed by appropriate cppp analysis.

2. The normal–normal model

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

2.1. A formula for the ppp. Assume therefore that the data $y = (y_1, \ldots, y_n)$, conditional on θ, are i.i.d. from $\mathrm{N}(\theta, \sigma^2)$, with known standard deviation σ, and let the prior be $\theta \sim \mathrm{N}(\theta_0, \sigma_0^2)$. We choose to work with

$$D(y, \theta) = n(\bar{y} - \theta)^2/\sigma^2 = \mathrm{monotone}(|\bar{y} - \theta|)$$

as discrepancy measure, where $\bar{y}$ is the mean of y_is; any monotone increasing function of $|\bar{y} - \theta|$ gives the same ppp value, as we see from (1.1). This is a pivotal quantity, with distribution being equal to a χ_1^2 for each given θ. For the following result, let $F_{1,1}(v, \kappa)$ be the cumulative distribution function of a non-central Fisher variable with degrees of freedom $(1, 1)$ and excentricity parameter κ, i.e. of a variable of the form $(X + \kappa^{1/2})^2/Y^2$, where X and Y are independent and standard normals.

Proposition. *For the situation described, the posterior predictive p-value, as a function of the observed data, may be expressed as*

$$\begin{aligned} \mathrm{ppp}(y^{\mathrm{obs}}) &= F_{1,1}\Big(\frac{1}{\rho_n}, \frac{(1-\rho_n)^2}{\rho_n} \frac{n(\bar{y}^{\mathrm{obs}} - \theta_0)^2}{\sigma^2}\Big) \\ &= F_{1,1}\Big(1 + \frac{\sigma^2}{n\sigma_0^2}, \frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \frac{(\bar{y}^{\mathrm{obs}} - \theta_0)^2}{\sigma_0^2}\Big), \end{aligned} \qquad (2.1)$$

where $\rho_n = n\sigma_0^2/(n\sigma_0^2 + \sigma^2)$.

PROOF: By well-known techniques, the posterior distribution for θ is found to be normal with mean $(1-\rho_n)\theta_0 + \rho_n \bar{y}^{\rm obs}$ and variance $\rho_n \sigma^2/n$. Conditional on θ, the $\bar{y}^{\rm rep}$ is of course a $\mathrm{N}(\theta, \sigma^2/n)$. Write now

$$(\theta \mid {\rm data}) \sim (1-\rho_n)\theta_0 + \rho_n \bar{y}^{\rm obs} + \rho_n^{1/2}(\sigma/\sqrt{n})N_0 \quad \text{and} \quad (\bar{y}^{\rm rep} \mid \theta) \sim \theta + (\sigma/\sqrt{n})N,$$

in terms of two independent standard normals N_0 and N. We learn that $D(y^{\rm rep}, \theta) = N^2$, independent of N_0, while

$$D(y^{\rm obs}, \theta) = \frac{n}{\sigma^2}\Big\{(1-\rho_n)(\theta_0 - \bar{y}^{\rm obs}) + \frac{\sigma}{\sqrt{n}}\rho_n^{1/2} N_0\Big\}^2 = \Big\{\rho_n^{1/2} N_0 - (1-\rho_n)\frac{\sqrt{n}(\bar{y}^{\rm obs} - \theta_0)}{\sigma}\Big\}^2.$$

Hence the ppp becomes

$$\mathrm{ppp} = \Pr\Big[N^2 \ge \rho_n \Big\{N_0 - \frac{1-\rho_n}{\rho_n^{1/2}}\frac{\sqrt{n}}{\sigma}(\bar{y}^{\rm obs} - \theta_0)\Big\}^2\Big],$$

from which the result follows. ■

2.2. Special cases. The formulae above offer insight into the ppp. Here are some remarks and consequences, also to be followed up in later sections.

Large n, or flat prior, or both. Suppose $\sqrt{n}\sigma_0$ is large, i.e. either n is large, or the prior is flat, or both. Then ρ_n goes to 1, and

$$\mathrm{ppp}(y^{\rm obs}) \to \Pr\{N^2 \ge N_0^2\} = \tfrac{1}{2} \quad \text{for all } y^{\rm obs}.$$

This happens irrespective of the observed $\bar{y}^{\rm obs}$. This is illustrated in Figure 2.1, for $n = 10$, $\sigma = 1$, $\theta_0 = 0$ and $\sigma_0 = 5$, giving $\rho_n = 0.996$, where the ppp is very close to $\frac{1}{2}$ over a broad range of $\bar{y}^{\rm obs}$.

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 to sample size, i.e. it is believed rather firmly that θ is close to θ_0. Then ρ_n is close to zero, and in the limit

$$\mathrm{ppp}(y^{\rm obs}) \to p^*(\bar{y}^{\rm obs}) = \Pr\{N^2 \ge n(\bar{y}^{\rm obs} - \theta_0)^2/\sigma^2\}.$$

This is the classic p-value for testing the hypothesis $\theta = \theta_0$. See Figure 2.1, for $n = 10$ and $\sigma_0 = 0.1$, with $\rho_n = 0.091$. Observe also that for given ρ_n, there is a maximum attainable value for ppp, namely $F_{1,1}(1/\rho_n)$, for $\bar{y}^{\rm obs} = \theta_0$; this is the 'maximally unsurprising' value we might have of $\bar{y}$, under the given prior. The ppp distribution is in particular not symmetric around $\frac{1}{2}$. The figure also displays the ppp as a function of $\bar{y}^{\rm obs}$ for an intermediate case of $n = 10$ and $\sigma_0 = 1$, for which $\rho_n = 0.909$. We observe that the distribution of ppp has a clear maximum value just above $\frac{1}{2}$ (actually, 0.5151), with most mass from say 0.25 to this max value.

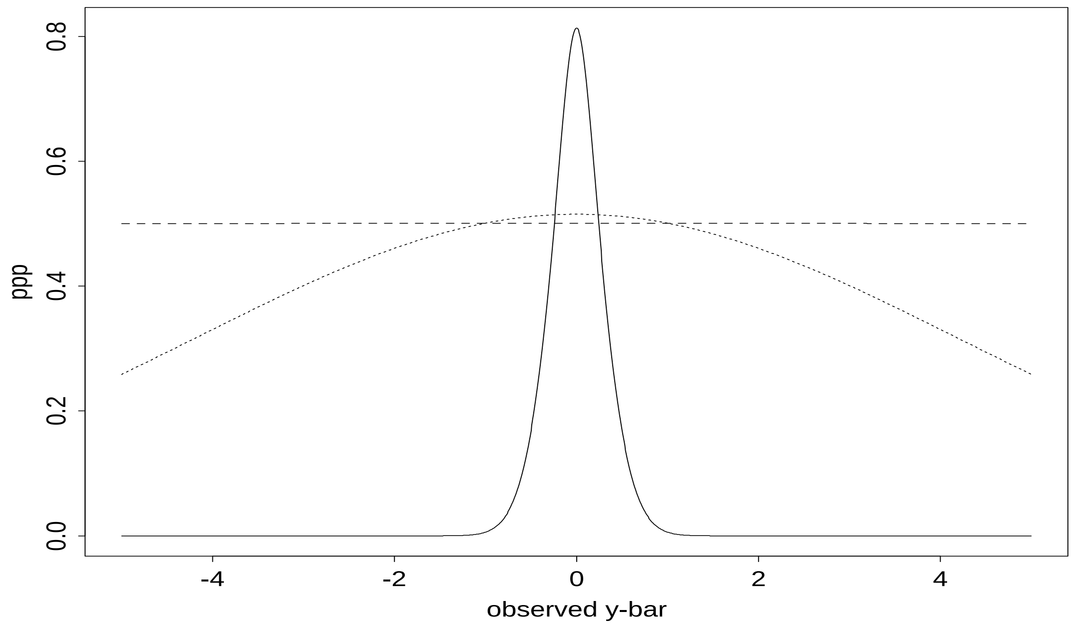

FIGURE 2.1. *The* ppp$(y^{\rm obs})$ *of formula (2.1) is displayed as a function of* $\bar y^{\rm obs}$*, for* $n=10$, $\sigma=1$, $\theta_0=0$, *for the three cases* σ_0 *equal to 0.1 (solid line), 1.0 (dotted line), 5.0 (dashed line).*

2.3. Full normal analysis. Above we took the data standard deviation to be known. More realistically, both parameters of the $\mathrm{N}(\mu,\sigma^2)$ 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 $\mathrm{GN}(\frac12 a,\frac12 b,\mu_0,c_0)$ distribution if

$$\lambda\sim\mathrm{Gam}(\tfrac12 a,\tfrac12 b)\quad\text{and}\quad(\mu\,|\,\lambda)\sim\mathrm{N}(\mu_0,(c_0\lambda)^{-1}).$$

Its density is accordingly

$$\pi(\lambda,\mu)\propto\lambda^{a/2-1}\exp(-\tfrac12 b\lambda)\lambda^{1/2}\exp\{-\tfrac12\lambda c_0(\mu-\mu_0)^2\}.$$

The likelihood for a data set $y_1,\ldots,y_n$ from the normal may be written as proportional to $\lambda^{n/2}\exp[-\frac12\lambda\{Q_0^{\rm obs}+n(\mu-\bar y)^2\}]$, where $Q_0^{\rm obs}=\sum_{i=1}^n(y_i-\bar y)^2$. The posterior density becomes

$$\begin{aligned}&\propto\lambda^{(a+n)/2-1}\lambda^{1/2}\exp[-\tfrac12\lambda\{b+Q_0^{\rm obs}+n(\mu-\bar y)^2+c_0(\mu-\mu_0)^2\}]\\&=\lambda^{(a+n)/2-1+1/2}\exp[-\tfrac12\lambda\{b+Q_0^{\rm obs}+(n^{-1}+c_0^{-1})^{-1}(\bar y-\mu_0)^2+(c_0+n)(\mu-\widetilde\mu)^2\}],\end{aligned}$$

proving that

$$\{(\lambda,\mu)\,|\,\text{data}\}\sim\mathrm{GN}\big(\tfrac12(a+n),\tfrac12(b+Q_0^{\rm obs}+(c_0^{-1}+n^{-1})^{-1}(\bar y^{\rm obs}-\mu_0)^2),\widetilde\mu,c_0+n\big),$$

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

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

$$
\begin{aligned}
D(y^{\rm obs},\theta) &= n\lambda\{\widetilde\mu+(c_0+n)^{-1/2}\lambda^{-1/2}N_0-\bar y^{\rm obs}\}^2\\
&= \frac{n}{c_0+n}\Big\{N_0+(c_0+n)^{1/2}\lambda^{1/2}\frac{c_0}{c_0+n}(\bar y^{\rm obs}-\mu_0)\Big\}^2,
\end{aligned}
$$

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

$$
\begin{aligned}
{\rm ppp}(y^{\rm obs}) &= \int_0^\infty {\rm Pr}\Big[N^2\ge\frac{n}{c_0+n}\Big\{N_0+\lambda^{1/2}\frac{c_0}{(c_0+n)^{1/2}}(\bar y^{\rm obs}-\mu_0)\Big\}^2\,|\,\lambda\Big]\,g_n(\lambda)\,{\rm d}\lambda\\
&= \int_0^\infty F_{1,1}\Big(\frac{c_0+n}{n},\frac{c_0^2\lambda}{c_0+n}(\bar y^{\rm obs}-\mu_0)^2\Big)\,g_n(\lambda)\,{\rm d}\lambda,
\end{aligned}
\tag{2.2}
$$

in terms of the gamma density $g_n(\lambda)$ with parameters $(\frac12 a_n,\frac12 b_n)$, where, from the above, $a_n=a+n$ and $b_n=b+Q_0^{\rm obs}+c_0n(\bar y^{\rm obs}-\mu_0)^2/(c_0+n)$.

The previous case of known σ corresponds to a and b going to infinity with $a/b=1/\sigma^2$ and $1/(c_0\lambda)=\sigma^2/c_0=\sigma_0^2$, i.e. $c_0=\sigma^2/\sigma_0^2$. In this case the above formula reduces to (2.1).

A simple numerical approximation to (2.2) is

$$
{\rm ppp}(y^{\rm obs})\doteq F_{1,1}\Big(1+\frac{c_0}{n},\frac{c_0^2}{c_0+n}\frac{(\bar y^{\rm obs}-\mu_0)^2}{(\widetilde\sigma^{\rm obs})^2}\Big),
$$

where $1/(\widetilde\sigma^{\rm obs})^2=a_n/b_n$ is the posterior mean of λ. For growing n the posterior variance of λ is small, and furthermore $\widetilde\sigma^2$ will be close to the empirical variance of the data.

Remark 2.1. The ppp of (2.1) converges rather rapidly to $\frac12$ as sample size increases. Some analysis reveals that $F_{1,1}(1+a/n,b/n)\doteq\frac12+\frac12(a-b)/(\pi n)$, with consequences for the (2.1) formula and approximations. ■

Remark 2.2. The formulae found in Meng (1994, Section 3) relate to the ppp value for another discrepancy measure, namely $D(y,\theta_0)=n(\bar y-\theta_0)^2/\sigma^2$, used as a test statistic for testing $\theta=\theta_0$. Our ppp measure uses the more general $D(y,\theta)$. ■

2.4. ppp behaviour for general parametric models. Calculations similar to those above can be carried out also for more complicated models, with additional efforts. In work not reported on here we have for example found explicitly computable formulae for the ppp 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 formulae for the ppp(y), however, and we would need to resort to the simulation scheme of (1.4) to compute the ppp value.

Some general phenomena may be recognised from the above analysis of the normal–normal and the other cases mentioned. Suppose that data $y = (y_1, \ldots, y_n)$ are i.i.d. from a density $g(y,\theta)$, conditional on θ. Assume also that a discrepancy measure of the type

$$D(y,\theta) = H(\sqrt{n}(\widehat\theta(y) - \theta), \theta)$$

is used, where $\widehat\theta(y)$ is the maximum likelihood estimator. We might e.g. take $H(v,\theta) = v^{\rm t} J(\theta) v$, with the Fisher information matrix $J(\theta)$; in this case $D(y,\theta)$ is close to a χ^2_p for large n, where p is the dimension of θ. In general,

$$\mathrm{ppp}(y^{\rm obs}) = \Pr\{H(\sqrt{n}(\widehat\theta(y^{\rm rep}) - \theta), \theta) \ge H(\sqrt{n}(\widehat\theta(y^{\rm obs}) - \theta), \theta) \,|\, \mathrm{data}\}.$$

Two cases of interest are as follows.

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

$$\mathrm{ppp}(y^{\rm obs}) \doteq \Pr\{H(\sqrt{n}(\widehat\theta(y^{\rm rep}) - \theta_0), \theta_0) \ge H(\sqrt{n}(\widehat\theta(y^{\rm obs}) - \theta_0), \theta_0)\} = p^*(y^{\rm obs}).$$

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

Then study the large-sample scenario where data really follow the $g(y,\theta_{\rm tr})$ model, for a suitable true parameter value, and n grows. Then, under mild regularity conditions, the distribution of $\sqrt{n}(\widehat\theta(y^{\rm rep}) - \theta)$ is close to say $V(\theta)$, which is $\mathrm{N}_p(0, J(\theta)^{-1})$, and this approximation statement holds uniformly in a neighbourhood around the $\theta_{\rm tr}$ value. In particular, $\sqrt{n}(\widehat\theta(y^{\rm rep}) - \theta) \to_d V$, which is $\mathrm{N}_p(0, J(\theta_{\rm tr})^{-1})$. Secondly, from Bernshteĭn–von Mises type theorems, see e.g. Lehmann (1983, Ch. 6), the posterior distribution of $\sqrt{n}(\theta - \widehat\theta^{\rm obs})$ is with probability 1 coming close to that of V_0, another and independent $\mathrm{N}_p(0, J(\theta_{\rm tr})^{-1})$ variable. All this implies $D(y^{\rm rep}, \theta) \to_d H(V, \theta_{\rm tr})$ and $D(y^{\rm obs}, \theta) \to_d H(V_0, \theta_{\rm tr})$, with probability 1. As long as $H(v,\theta)$ is continuous, therefore,

$$\mathrm{ppp}(y^{\rm obs}) \to \Pr\{H(V, \theta_{\rm tr}) \ge H(V^0, \theta_{\rm tr})\} = \tfrac{1}{2} \quad \text{a.s.}$$

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

3. The ppp when the prior is a mixture

Here we study the ppp 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 $y_1, \ldots, y_n$ conditional on θ are i.i.d. from $\mathrm{N}(\theta, \sigma^2)$ and that the same discrepancy measure $D(y, \theta)$ is used, but that the prior is a mixture of two different hypotheses about nature; $\theta \sim p_1 \mathrm{N}(\theta_{0,1}, \sigma_{0,1}^2) + p_2 \mathrm{N}(\theta_{0,2}, \sigma_{0,2}^2)$. Then

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

in terms of the posterior densities $\pi_j(\theta \,|\, \mathrm{data})$ 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

$$(\theta \,|\, \mathrm{data}) \sim \widetilde{p}_1(\bar{y}^{\mathrm{obs}}) \pi_1(\theta \,|\, \mathrm{data}) + \widetilde{p}_2(\bar{y}^{\mathrm{obs}}) \pi_2(\theta \,|\, \mathrm{data}),$$

where

$$\widetilde{p}_j(y^{\mathrm{obs}}) = \frac{p_j f_j(\bar{y}^{\mathrm{obs}})}{p_1 f_1(\bar{y}^{\mathrm{obs}}) + p_2 f_2(\bar{y}^{\mathrm{obs}})} \quad \text{for } j = 1, 2.$$

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

$$\mathrm{ppp}(y^{\mathrm{obs}}) = \sum_{j=1}^{2} \widetilde{p}_j(y^{\mathrm{obs}}) F_{1,1}\Big(\frac{1}{\rho_{n,j}}, \frac{(1-\rho_{n,j})^2}{\rho_{n,j}} \frac{n(\bar{y}^{\mathrm{obs}} - \theta_{0,j})^2}{\sigma^2}\Big).$$

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

There are several general ppp aspects to be learned from applying this formula in different mixture prior settings. A variety of possible shapes for the $\mathrm{ppp}(\bar{y}^{\mathrm{obs}})$ curves emerge by different combinations of the parameters. Among findings of interest are the following points.

(i) Suppose 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 $\mathrm{ppp}(y^{\mathrm{obs}})$ function is relatively unaffected by the balance parameters p_1, 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 it is the second prior component that is picked out by the data.

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

$$\lim_{p_2 \to 0} \mathrm{ppp}(\bar{y}^{\mathrm{obs}} = \theta_{0,2}) = 1 \quad \text{and} \quad \lim_{p_2 \to 0} \mathrm{prpp}(\bar{y}^{\mathrm{obs}} = \theta_{0,2}) = 0$$

in the situation with two sharp hypotheses, i.e. $\sigma_{0,j} = 0$ for $j = 1, 2$. Thus the ppp may in this respect be seen as a highly non-continuous operator, since the two $\text{ppp}(y^{\text{obs}})$ curves are highly different for $p_2 = 0$ and $p_2 = 0.001$.

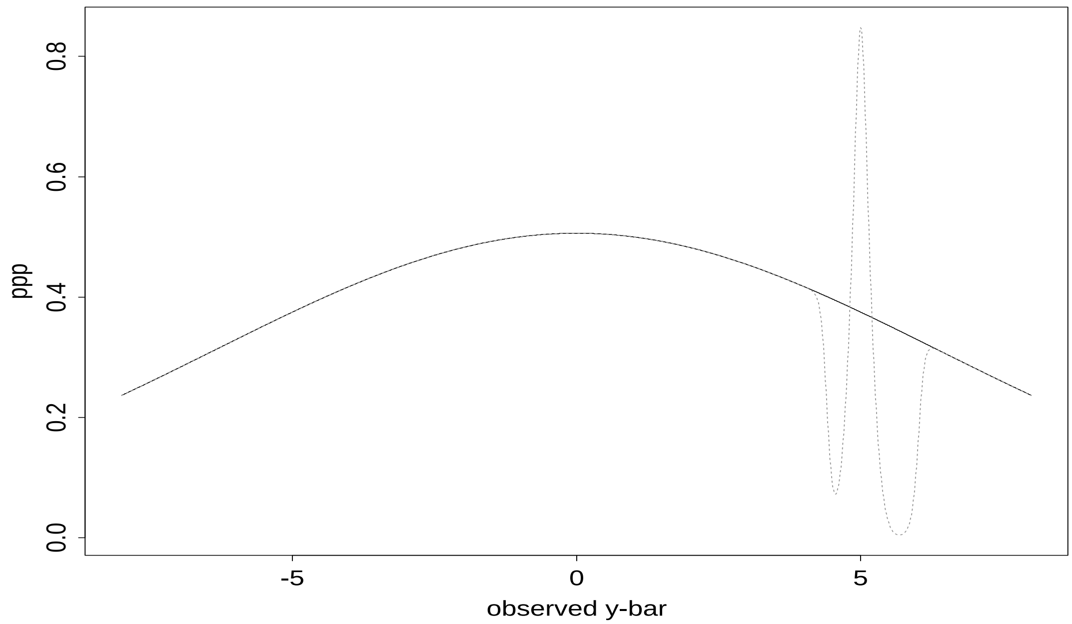

FIGURE 3.1. *The* $\text{ppp}(y^{\text{obs}})$ *as a function of* $\bar{y}^{\text{obs}}$, *for* $n = 25$ *and* $\sigma = 1$, *displayed for two priors; for the normal* $(0, 1)$ *(solid line) and for a sharp bi-mixture with* $p_1 = 0.999$ *and* $p_2 = 0.001$, *of two normals* $(0, 1)$ *and* $(5, 0.05^2)$ *(dotted line).*

(ii) Suppose $\sigma_{0,1}$ is moderate while $\sigma_{0,2}$ is small. Then when $\bar{y}^{\text{obs}}$ is close to $\theta_{0,2}$ the ppp acts as a classic p-value for testing $\theta = \theta_{0,2}$; if $\bar{y}^{\text{obs}}$ on the other hand is some distance away from $\theta_{0,2}$, then the ppp is close to what it would be for the first prior component alone. Again this behaviour is relatively independent of the p_1, p_2 parameters. It is interesting and perhaps mildly contra-intuitive that even when p_2 is very small, the ppp indicates non-surprise 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 Figure 3.1.

(iii) Let us now hold p_2 fixed (but small), while $\theta_{0,2}$ varies. We have seen that when $\theta_{0,2}$ is large, $\text{ppp}(\bar{y}^{\text{obs}} = \theta_{0,2})$ will be close to 1. Also, if $\theta_{0,2} \to 0$, we clearly get $\text{ppp}(\bar{y}^{\text{obs}} = \theta_{0,2}) \to 1$. 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 zero. This is a bit of a paradox: we have two competing models, represented 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, while if they 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 each other. This is in particular true if the different parameter values give very different distributions for the observed $y^{\rm obs}$. If this is the case, conditioning on the observation will essentially eliminate parameter values that fail to explain the data, making their prior likelihood irrelevant.

4. ppp calibration and the double-level simulation

Suppose that one computes $\text{ppp}(y^{\rm obs}) = 0.28$ from one's data set, for a given prior and model. This is a well-defined probability, as in (1.1), but to judge the significance of the 0.28 number, possibly in comparison with other ppp values for other combinations of prior and model for the same data set, we are forced to define the underlying probability scale: how rare, or how common, are values less than 0.28?

There has perhaps been a certain Pavlovian tendency in applied statistics work to interpret the ppp numbers on the uniform scale, like for classic p-values. We have already seen that the distribution of ppp is often quite non-uniform, however, and more precise information is provided below. We shall see that an observed $\text{ppp} = 0.28$ (say) may be extremely surprising in one situation, whereas the value $\text{ppp} = 0.06$ (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 is forced to consider the values of y that one did not observe. In this section we define and study the appropriate distribution of $\text{ppp}(Y)$ across values of Y, and use this to introduce our calibrated ppp value, the $\text{cppp}(y^{\rm obs})$. Its actual use is illustrated with real data in Sections 6 and 7.

4.1. The null-null distribution of the ppp. Classical Bayes analysis operates of course conditional on $y^{\rm obs}$. When one in addition wishes to test validity of aspects of model and prior, however, one needs to bring into the picture also the distribution of $y^{\rm obs}$. This leads us to define 'the null-null distribution' of $\text{ppp}(Y)$, corresponding to the distribution of $\text{ppp}(y)$ across precisely those y values that occur by the combined mechanism of the prior and the model. The null-null distribution of $U = \text{ppp}(Y)$, under perfect prior and perfect model, is

$$G(u) = \Pr\{\text{ppp}(Y) \le u\} \quad \text{where } Y \sim \int f(y,\theta)\pi(\theta)\,{\rm d}\theta. \tag{4.1}$$

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

$$\text{cppp}(y^{\rm obs}) = G(\text{ppp}(y^{\rm obs})) = \Pr\{\text{ppp}(Y) \le \text{ppp}(y^{\rm obs})\}. \tag{4.2}$$

The distribution of $\text{cppp}(Y)$, across values of Y as above, 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 the world is exactly as we have imagined, both regarding Nature's distribution of θ and our own modelling 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

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

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

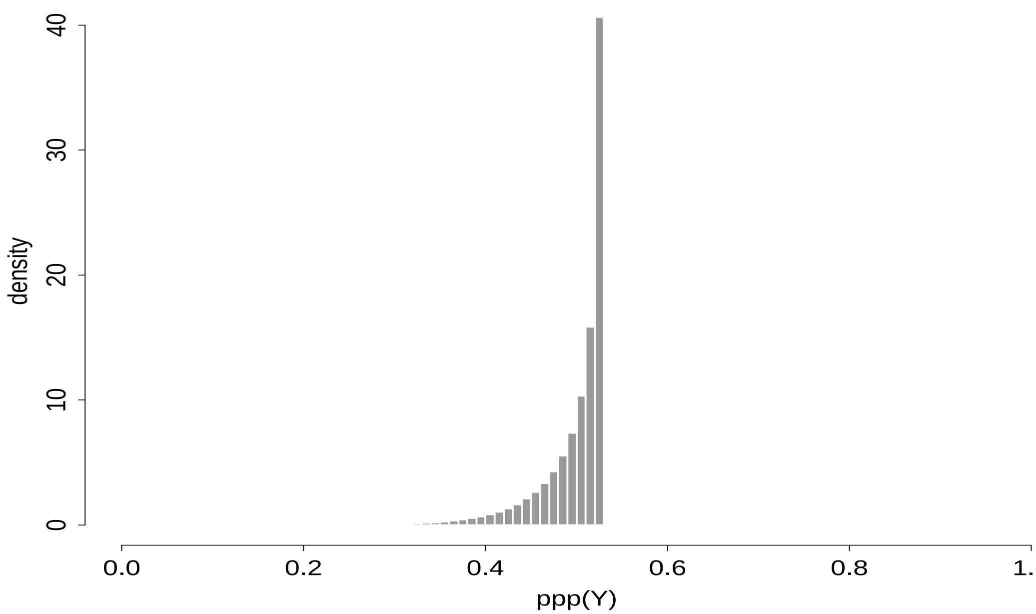

FIGURE 4.1. *Density of* ppp(Y)*, displayed as a normalised histogram with a million simulations, for* $n=5$*,* $\sigma=1$*,* $\sigma_0=1$*. The distribution has exact mean* $\frac{1}{2}$*, with sharp right cut-off point at* $F_{1,1}(1/\rho_n)$*.*

The ppp distribution can also easily be displayed via simulation of

$$U = F_{1,1}\Big(\frac{1}{\rho_n},\frac{(1-\rho_n)^2}{\rho_n}(1+n\sigma_0^2/\sigma^2)M^2\Big) = F_{1,1}\Big(\frac{1}{\rho_n},\frac{1-\rho_n}{\rho_n}M^2\Big), \qquad (4.3)$$

where $M\sim \mathrm{N}(0,1)$. To compute the cppp$(y^{\rm obs})$ we may simply check the relative frequency of such simulated Us that fall below the observed ppp$(y^{\rm obs})$. For this particular situation we may even get an explicit expression;

$$\begin{aligned}\mathrm{cppp}(y^{\mathrm{obs}}) &= \Pr\Big\{F_{1,1}\Big(\frac{1}{\rho_n}, \frac{(1-\rho_n)^2}{\rho_n}\frac{n(\bar Y-\theta_0)^2}{\sigma^2}\Big) \le F_{1,1}\Big(\frac{1}{\rho_n}, \frac{(1-\rho_n)^2}{\rho_n}\frac{n(\bar y^{\mathrm{obs}}-\theta_0)^2}{\sigma^2}\Big)\Big\} \\ &= \Pr\{n(\bar Y-\theta_0)^2/\sigma^2 \ge n(\bar y^{\mathrm{obs}}-\theta_0)^2/\sigma^2\} \\ &= \Pr\Big\{\chi_1^2 \ge \frac{n(\bar y^{\mathrm{obs}}-\theta_0)^2/\sigma^2}{1+n\sigma_0^2/\sigma^2}\Big\}.\end{aligned} \tag{4.4}$$

While the classic p-value statistic is uniform on the unit interval under the null hypothesis, the present null-null distribution is quite far from having such a form. The U variable of (4.3) 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 zero, does the ppp distribution come close to the uniform one on the unit interval (which is the limit case as $\rho_n \to 0$). The statistical distribution of $\mathrm{cppp}(Y)$ of (4.4), however, is by construction exactly a uniform on $[0,1]$, provided the prior and the model are correct. See in this connection Figure 4.1, which shows the partly extreme nature of the ppp distribution, and Figure 4.2, which gives the ppp and cppp curves, along with a third variant treated in a later subsection.

Remark 4.1. When n grows, the observed $\bar y^{\mathrm{obs}}$ will go to the true underlying mean value θ_{tr}, and from (4.4) we see that

$$\mathrm{cppp}(y^{\mathrm{obs}}) \to \Pr\{\chi_1^2 \ge \{(\theta_{\mathrm{tr}}-\theta_0)/\sigma_0\}^2\} = \Pr\{|N(0,1)| \ge |\theta_{\mathrm{tr}}-\theta_0|/\sigma_0\} \quad \text{a.s.}$$

Thus

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

emerges as a natural measure of conflict between real data distribution and the $\mathrm{N}(\theta_0,\sigma_0^2)$ prior. If $\mathrm{conf} \ge 1.96$, then $\mathrm{cppp}(y^{\mathrm{obs}})$ will for large n be below the critical value 0.05, etc. The second representation of conf uses $c_0 = \sigma^2/\sigma_0^2$, which has interpretation as prior sample size. See also Section 9.1. ■

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

$$\mathrm{cppp}(y^{\mathrm{obs}}) \doteq \frac{1}{B}\sum_{k=1}^B I\{\mathrm{ppp}(y_k) \le \mathrm{ppp}(y^{\mathrm{obs}})\}. \tag{4.5}$$

It is the perhaps chief claim of our article that while the $\mathrm{ppp}(y^{\mathrm{obs}})$ of (1.3) may have a difficult interpretation and sometimes a low information value, the $\mathrm{cppp}(y^{\mathrm{obs}})$ of (4.5) has a clear meaning and can be highly informative.

While clear in interpretation and natural qua strategy, the (4.5) operation might of course be both cumbersome and computer time costly from an operational point of view, since it in general must amount to a double simulation, with AB operations in total, following (1.3). One should therefore look for ways of simplifying the computational burden. Sometimes an explicit formula may be worked out for the $\mathrm{ppp}(y^{\mathrm{obs}})$, with considerable benefit for the cppp computations, as we also see in the next section, or one may be helped by knowing the distribution of $D(y^{\mathrm{rep}},\theta)$. In other cases one may look for variance reduction tricks or for ways of approximating the $\mathrm{ppp}(Y)$ distribution, the benefit being that one may be allowed a moderate rather than a large number B of repeated sampling. We have actually developed some methods of this kind in connection with the cppp analysis of the bird survival data in Section 7, but this will be reported on elsewhere.

4.3. Alternative what-if scenarios. We judge the above calibration to be the canonical one, transforming the ppp numbers under the proposed prior and model to the uniform scale. One may however also consider other what-if scenarios that from different perspectives lead to other potentially interesting distributions for Y, and hence to other calibrations for $\mathrm{ppp}(Y)$, in the formulation of (4.1)–(4.2). Suppose in general terms that a distribution for such Y^* data-sets is being considered, where such Y^* are drawn from a mechanism different from the canonical one given in (4.1); this other distribution could for example take the form $\int f(y,\theta)\pi^*(\theta)\,\mathrm{d}\theta$, for a prior $\pi^*(\theta)$ different from $\pi(\theta)$. This defines an alternative ppp distribution $G^*(u)=\Pr\{\mathrm{ppp}(Y^*)\le u\}$ and in its turn a differently calibrated ppp value,

$$\mathrm{cppp}^*(y^{\mathrm{obs}})=G^*(\mathrm{ppp}(y^{\mathrm{obs}}))=\Pr\{\mathrm{ppp}(Y^*)\le \mathrm{ppp}(y^{\mathrm{obs}})\}. \tag{4.6}$$

A ppp number is computed under a given model specification, say $\mathcal{M}$, and we may write $\mathrm{ppp}(y^{\mathrm{obs}},\mathcal{M})$ to indicate this. Whereas we above calibrated $\mathrm{ppp}(y^{\mathrm{obs}},\mathcal{M})$ via the same model $\mathcal{M}$, operations as described above amount to calibrating $\mathrm{ppp}(y^{\mathrm{obs}},\mathcal{M})$ using an alternative model $\mathcal{M}^*$ for data-sets Y^*. The (4.6) scheme may then be thought of in terms of

$$\begin{aligned}\mathrm{cppp}^*(y^{\mathrm{obs}})&=\mathrm{c}[\mathrm{ppp}(y^{\mathrm{obs}},\mathcal{M}),\mathcal{M}^*]\\&=\text{calibrated }\mathrm{ppp}(y^{\mathrm{obs}},\text{model }\mathcal{M}),\text{ under model }\mathcal{M}^*.\end{aligned} \tag{4.7}$$

The definition here is fully general and operational, and the cppp^* may be computed for any well-defined Y^* distribution, via double simulation if necessary. For the normal–normal model we may see more clearly the implications of the (4.6) idea; the reasoning that led to (4.4) now gives

$$\mathrm{cppp}^*(y^{\mathrm{obs}})=\Pr\{n(\bar Y^*-\theta_0)^2/\sigma^2\ge n(\bar y^{\mathrm{obs}}-\theta_0)^2/\sigma^2\},$$

in terms of the distribution of the sample average $\bar Y^*$ stemming from data-sets Y^* drawn from the intended alternative what-if distribution. This gives formulae generalising that of (4.4), for different models $\mathcal{M}^*$ for data-sets Y^*.

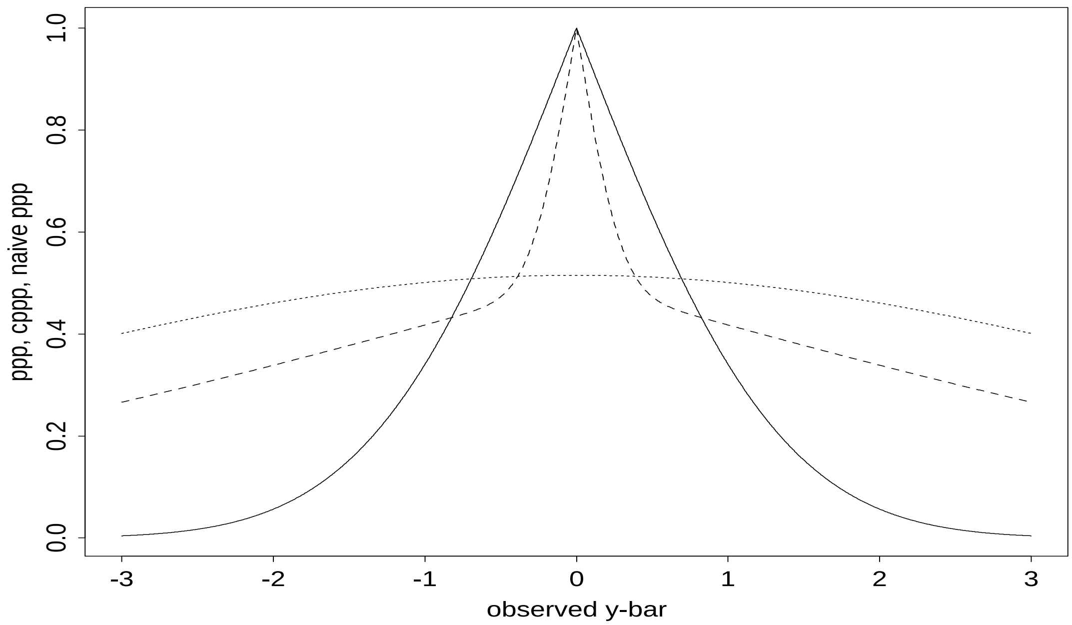

FIGURE 4.2. *The three functions* ppp *(dotted line),* cppp *(solid line),* cppp* *(dashed line) are displayed as functions of* $\bar y^{\rm obs}$*, for* $n = 10$*,* $\sigma = 1$*, with a* $\mathrm{N}(0,1)$ *prior for* θ*.*

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

For the normal–normal situation, once more, we know that $\theta\,|\,{\rm data}$ is distributed as a normal with mean $(1-\rho_n)\theta_0+\rho_n\bar y^{\rm obs}$ and variance $\rho_n\sigma^2/n$. Sampling data Y_i^* from the $\mathrm{N}(\theta,\sigma^2)$ for such θ we find that

$$\bar Y^*-\theta_0\sim \mathrm{N}(\rho_n(\bar y^{\rm obs}-\theta_0),(1+\rho_n)\sigma^2/n).$$

Using the cppp* formula above this yields

$$\begin{aligned}{\rm cppp}^*(y^{\rm obs})&={\rm Pr}\big[\{\rho_n(\bar y^{\rm obs}-\theta_0)+(1+\rho_n)^{1/2}(\sigma/\sqrt{n})N\}^2\ge(\bar y^{\rm obs}-\theta_0)^2\big]\\&={\rm Pr}\Big[\frac{\sigma^2}{n}(1+\rho_n)\Big\{\frac{\sqrt{n}}{\sigma}\frac{\rho_n}{(1+\rho_n)^{1/2}}(\bar y^{\rm obs}-\theta_0)+N\Big\}^2\ge(\bar y^{\rm obs}-\theta_0)^2\Big]\\&={\rm Pr}\Big\{\chi^2_1\Big(\frac{\rho_n^2}{1+\rho_n}z_n\Big)\ge\frac{z_n}{1+\rho_n}\Big\},\end{aligned}$$

in terms of $z_n = n(\bar{y}^{\rm obs} - \theta_0)^2/\sigma^2$. In the terminology and thinking of (4.7) this would be the formula for

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

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

Figure 4.2 illustrates the ppp, cppp and cppp* curves for $n = 10$, $\sigma = 1$, with a normal $(0, 1)$ prior for θ, as a function of $\bar{y}^{\rm obs}$. We see that the cppp* is not a satisfactory measure of surprise; we would scarcely ever be surprised, if the numbers are interpreted on the uniform scale. This is due to the double use of data when we draw Y^* from the posterior predictive. There could be other choices for Y^* distributions with more relevance and with a less over-cautious cppp* curve than here.

5. ppp and cppp analysis for general regression

Here we study the general linear regression model for data (x_i, y_i) for which $y_i = x_i^{\rm t}\beta + \varepsilon_i$, for $i = 1, \ldots, n$, where x_i is a p-dimensional covariate vector for individual i, and the ε_is are independent and normal with standard deviation σ. In standard matrix formulation, $y = X\beta + \varepsilon$, with least squares estimator $\widehat{\beta} = (X^{\rm t}X)^{-1}X^{\rm t}y$ for β; it is assumed that the $n \times p$ matrix X is of full rank. We shall 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 $D(y, \theta) = (\widehat{\beta} - \beta)^{\rm t}\Omega_n(\widehat{\beta} - \beta)/\sigma^2$, where $\Omega_n = X^{\rm t}X = \sum_{i=1}^n x_i x_i^{\rm t}$, and where we write $\theta = (\beta, \sigma)$ for the full parameter vector. It may also be expressed as

$$D(y, \theta) = \|\widehat{\mu} - \mu\|^2/\sigma^2 = \sum_{i=1}^n (\widehat{\mu}_i - \mu_i)^2/\sigma^2$$

in terms of the vector μ of means $\mu_i = x_i^{\rm t}\beta$ and their fitted values $\widehat{\mu}_i = x_i^{\rm t}\widehat{\beta}$.

One may of course include other discrepancies too for one's analysis, like

$$\max_{i \le n} |y_i - x_i^{\rm t}\beta|/\sigma \quad {\rm or} \quad \max_t \Big| n^{-1} \sum_{i=1}^n I\{(y_i - x_i^{\rm t}\beta)/\sigma \le t\} - \Phi(t) \Big|.$$

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 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 known, so we study that case first.

Assume β has a prior distribution of the form $\mathrm{N}_p(\beta_0, \sigma^2(c_0\Omega_0)^{-1})$, with a matrix Ω_0 and a scalar c_0. The parameterisation is a bit redundant, since c_0 may be taken into the Ω_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. One finds using multivariate normal theory and some matrix algebra that β given data is normal, with

$$\begin{aligned} \mathrm{E}(\beta \mid \mathrm{data}) &= \widetilde{\beta} = (c_0\Omega_0 + \Omega_n)^{-1}(c_0\Omega_0\beta_0 + \Omega_n\widehat{\beta}^{\mathrm{obs}}), \\ \mathrm{Var}(\beta \mid \mathrm{data}) &= \sigma^2(c_0\Omega_0 + \Omega_n)^{-1}. \end{aligned}$$

Also, $\widetilde{\beta} - \widehat{\beta}^{\mathrm{obs}} = (c_0\Omega_0 + \Omega_n)^{-1}c_0\Omega_0(\beta_0 - \widehat{\beta}^{\mathrm{obs}})$. Since $D(y, \theta^{\mathrm{rep}}) \sim \chi^2_p$, independent of $D(y^{\mathrm{obs}}, \theta)$, this leads to the formula

$$\mathrm{ppp}(y^{\mathrm{obs}}) = \Pr\{\chi^2_p \ge (U + f)^{\mathrm{t}}\Omega_n(U + f)\}, \tag{5.1}$$

where $U \sim \mathrm{N}_p(0, (c_0\Omega_0 + \Omega_n)^{-1})$ and $f = (c_0\Omega_0 + \Omega_n)^{-1}c_0\Omega_0(\widehat{\beta}^{\mathrm{obs}} - \beta_0)/\sigma$. Computation would most often be simplest via simulation of U vectors. We note that when c_0 is large, U goes to zero and f goes to $(\widehat{\beta}^{\mathrm{obs}} - \beta_0)/\sigma$, making $\mathrm{ppp}(y^{\mathrm{obs}}) \to p^*(y)$, the traditional p-value for testing $\beta = \beta_0$ using the $(\widehat{\beta} - \beta_0)^{\mathrm{t}}\Omega_n(\widehat{\beta} - \beta_0)/\sigma^2$ test. On the other hand, when c_0 becomes small, $f \to 0$ and $\mathrm{ppp}(y^{\mathrm{obs}}) \to \frac{1}{2}$ for each data set y^{obs}.

To calibrate the ppp we need to compute a number B of $\mathrm{ppp}(y_k)$ values from simulated data sets y_k, as per (4.5). We would draw these sets by first drawing β_k from the posterior distribution and then creating $y_k = X\beta_k + \varepsilon_k$, i.e. without changing or resampling the covariate vectors. For each of these B sets one would find the $\mathrm{ppp}(y_k)$ number via simulation of A random vectors U, i.e.

$$\mathrm{ppp}(y_k) \doteq \frac{1}{A}\sum_{j=1}^{A}\Pr\{\chi^2_p \ge (U_{k,j} - f(y_k))^{\mathrm{t}}\Omega_n(U_{k,j} - f(y_k))\}.$$

The only simple case is that of Ω_0 being proportional to Ω_n. We may in that case without loss of generality take $\Omega_0 = n^{-1}\Omega_n$, with appropriate 'prior sample size' interpretation for the flexible factor c_0 in $\mathrm{Var}\,\beta = \sigma^2(n/c_0)\Omega_n^{-1}$; the Ω_n matrix indeed grows linearly with sample size n. We then have

$$\widetilde{\beta} = \frac{c_0}{c_0 + n}\beta_0 + \frac{n}{c_0 + n}\widehat{\beta}^{\mathrm{obs}}, \quad U \sim \mathrm{N}_p\Big(0, \frac{n}{c_0 + n}\Omega_n^{-1}\Big), \quad f = \frac{c_0}{c_0 + n}\frac{\widehat{\beta}^{\mathrm{obs}} - \beta_0}{\sigma}.$$

This leads to a representation for $(U + f)^{\mathrm{t}}\Omega_n(U + f)$, in terms of $W \sim \mathrm{N}_p(0, \Omega_n^{-1})$, as

$$\begin{aligned} &\frac{n}{c_0 + n}\Big(W + \Big(\frac{c_0 + n}{n}\Big)^{1/2}\frac{c_0}{c_0 + n}\frac{\widehat{\beta}^{\mathrm{obs}} - \beta_0}{\sigma}\Big)^{\mathrm{t}}\Omega_n\Big(W + \Big(\frac{c_0 + n}{n}\Big)^{1/2}\frac{c_0}{c_0 + n}\frac{\widehat{\beta}^{\mathrm{obs}} - \beta_0}{\sigma}\Big) \\ &\quad \sim \frac{n}{c_0 + n}\chi^2_p\Big(\frac{c_0^2}{n}\frac{1}{c_0 + n}\frac{(\widehat{\beta}^{\mathrm{obs}} - \beta_0)^{\mathrm{t}}\Omega_n(\widehat{\beta}^{\mathrm{obs}} - \beta_0)}{\sigma^2}\Big). \end{aligned}$$

In consequence, under the $\Omega_0 = n^{-1}\Omega_n$ scenario,

$$\text{ppp}(y^{\text{obs}}) = F_{p,p}\Big(1 + \frac{c_0}{n}, \frac{c_0^2}{c_0+n}\frac{\kappa_n^{\text{obs}}}{\sigma^2}\Big), \tag{5.2}$$

in terms of $\kappa_n^{\text{obs}} = n^{-1}(\widehat\beta^{\text{obs}}-\beta_0)^{\rm t}\Omega_n(\widehat\beta^{\text{obs}}-\beta_0)$ and the non-central F distribution function.

This also leads to an explicit formula for the cppp, as follows:

$$\begin{aligned}\text{cppp}(y^{\text{obs}}) &= \Pr\{\text{ppp}(Y) \le \text{ppp}(y^{\text{obs}})\}\\ &= \Pr\{(\widehat\beta(Y)-\beta_0)^{\rm t}\Omega_n(\widehat\beta(Y)-\beta_0)/\sigma^2 \ge (\widehat\beta^{\text{obs}}-\beta_0)^{\rm t}\Omega_n(\widehat\beta^{\text{obs}}-\beta_0)/\sigma^2\}\\ &= \Pr\{\chi^2_p \ge \frac{c_0}{c_0+n}(\widehat\beta^{\text{obs}}-\beta_0)^{\rm t}\Omega_n(\widehat\beta^{\text{obs}}-\beta_0)/\sigma^2\}.\end{aligned}$$

We use here that under prior and model conditions, $\widehat\beta(Y)$ is normal with mean β_0 and variance $\sigma^2(1+n/c_0)\Omega_n^{-1}$.

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 β given σ. In generalisation of material of Section 2.3, agree now to say that $(\lambda,\mu) = (1/\sigma^2,\beta)$ comes from the $\text{GN}_p(\frac12 a, \frac12 b, \mu_0, c_0\Omega_0)$ distribution if

$$\lambda \sim \text{Gam}(\tfrac12 a, \tfrac12 b) \quad\text{and}\quad (\beta\,|\,\lambda) \sim \text{N}_p(\beta_0, \lambda^{-1}(c_0\Omega_0)^{-1}). \tag{5.3}$$

Thus β has prior mean β_0 and prior variance $(\text{E}\lambda^{-1})c_0^{-1}\Omega_0^{-1}$. We shall see that the posterior takes the form

$$\{(\lambda,\beta)\,|\,\text{data}\} \sim \text{GN}_p\big(\tfrac12(a+n), \tfrac12(b+Q_0^{\text{obs}}+(\widehat\beta^{\text{obs}}-\beta_0)^{\rm t}K(\widehat\beta^{\text{obs}}-\beta_0)), \widetilde\beta, c_0\Omega_0+\Omega_n\big), \tag{5.4}$$

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

$$\widetilde\beta = (c_0\Omega_0+\Omega_n)^{-1}(c_0\Omega_0\beta_0 + \Omega_n\widehat\beta^{\text{obs}}) \quad\text{and}\quad K = (c_0^{-1}\Omega_0^{-1} + \Omega_n^{-1})^{-1}.$$

Proving this is accomplished partly via a little algebraic lemma, which states that for vectors a_0, a_1 and positive definite matrices G_0, G_1, the following identity holds:

$$(x-a_0)^{\rm t}G_0(x-a_0)^{\rm t} + (x-a_1)^{\rm t}G_1(x-a_1) = (x-\widetilde a)^{\rm t}(G_0+G_1)(x-\widetilde a) + (a_1-a_0)^{\rm t}M(a_1-a_0),$$

where $\widetilde a = (G_0+G_1)^{-1}(G_0a_0+G_1a_1)$ and $M = (G_0^{-1}+G_1^{-1})^{-1}$.

To prove (5.4), start with the prior density for (λ,β), which is proportional to

$$\lambda^{a/2-1}\exp(-\tfrac12 b\lambda)\lambda^{p/2}\exp\{-\tfrac12 c_0\lambda(\beta-\beta_0)^{\rm t}\Omega_0(\beta-\beta_0)\}.$$

The likelihood for data is proportional to $\lambda^{n/2}\exp[-\frac{1}{2}\lambda\{Q_0^{\rm obs}+(\widehat\beta^{\rm obs}-\beta)^{\rm t}\Omega_n(\widehat\beta^{\rm obs}-\beta)\}]$. The posterior accordingly becomes proportional to

$$\begin{aligned}
&\lambda^{(a+n)/2-1}\lambda^{p/2}\exp[-\tfrac{1}{2}\lambda\{b+Q_0^{\rm obs}+c_0(\beta-\beta_0)^{\rm t}\Omega_0(\beta-\beta_0)\\
&\qquad\qquad+(\beta-\widehat\beta^{\rm obs})^{\rm t}\Omega_n(\beta-\widehat\beta^{\rm obs})\}]\\
&=\lambda^{(a+n)/2-1+p/2}\exp[-\tfrac{1}{2}\lambda\{b+Q_0^{\rm obs}+(\beta-\widetilde\beta)^{\rm t}(c_0\Omega_0+\Omega_n)^{-1}(\beta-\widetilde\beta)\\
&\qquad\qquad+(\widehat\beta^{\rm obs}-\beta_0)^{\rm t}K(\widehat\beta^{\rm obs}-\beta_0)\}],
\end{aligned}$$

using the algebraic identity above. This is of the required form.

This leads to an accessible scheme for computing the ppp, as follows. Under model conditions, conditional on (β,σ), one knows that $\widehat\beta$ is a ${\rm N}_p(\beta,\sigma^2\Omega_n^{-1})$, allowing $D(y^{\rm rep},\theta)$ a representation of the form $V^{\rm t}\Omega_nV$, where $V\sim{\rm N}_p(0,\Omega_n^{-1})$, i.e. $D(y^{\rm rep},\beta)$ is a χ^2_p. To work with $D(y^{\rm obs},\theta)$, we represent for the posterior distribution β for given λ as $\widetilde\beta+\lambda^{-1/2}U$, where $U\sim{\rm N}_p(0,(c_0\Omega_0+\Omega_n)^{-1})$ and independent of the χ^2_p distribution of $D(y^{\rm rep},\theta)$. Thus

$$\begin{aligned}
D(y^{\rm obs},\theta)&=\lambda(\widetilde\beta-\widehat\beta^{\rm obs}+\lambda^{-1/2}U)^{\rm t}\Omega_n(\widetilde\beta-\widehat\beta^{\rm obs}+\lambda^{-1/2}U)\\
&=(\lambda^{1/2}(c_0\Omega_0+\Omega_n)^{-1}c_0\Omega_0(\widehat\beta^{\rm obs}-\beta_0)+U)^{\rm t}\Omega_n\\
&\qquad(\lambda^{1/2}(c_0\Omega_0+\Omega_n)^{-1}c_0\Omega_0(\widehat\beta^{\rm obs}-\beta_0)+U).
\end{aligned}$$

It is now relatively easy to simulate a large number A of (λ_j,U_j) replicates, with

$$\lambda\sim{\rm Gam}(\tfrac{1}{2}a_n,\tfrac{1}{2}b_n),\quad\text{where } a_n=a+n \text{ and } b_n=b+Q_0^{\rm obs}+(\widehat\beta^{\rm obs}-\beta_0)^{\rm t}K(\widehat\beta^{\rm obs}-\beta_0).$$

This gives replicates of $D(y^{\rm obs},\theta_j)$ and the required simulation approximation

$${\rm ppp}(y^{\rm obs})\doteq\frac{1}{A}\sum_{j=1}^A{\rm Pr}\{\chi^2_p\ge D(y^{\rm obs},\theta_j)\}.\tag{5.5}$$

This is a more precise estimate than the general-recipe version used in (1.3); the (5.5) option is available here since $D(y^{\rm rep},\theta)$ has the known χ^2_p distribution, regardless of θ.

Formula (5.5) also lends itself to the computation of the cppp, through a double simulation regime that computes ${\rm ppp}(y_k)$ for many simulated data set, as per the general (4.5) strategy.

5.3. The case of proportional Ω_0 and Ω_n. Important simplifications are found for the special case where the prior variance of β is specified as being proportional to the sample variance of its least squares estimator. In that case we may again take $\Omega_0=n^{-1}\Omega_n$, as above, thereby also giving a more precise interpretation of c_0 in relation to sample size. With efforts similar to those exuded in the previous subsection one finds

$$D(y^{\rm obs},\theta)=\frac{n}{c_0+n}\chi^2_p\Big(\lambda\frac{c_0^2}{c_0+n}\frac{1}{n}(\widehat\beta^{\rm obs}-\beta_0)^{\rm t}\Omega_n(\widehat\beta^{\rm obs}-\beta_0)\Big),$$

leading to the formula

$$\begin{aligned}\mathrm{ppp}(y^{\mathrm{obs}}) &= \Pr\Big\{\chi^2_p \ge \frac{n}{c_0+n}\chi^2_p\Big(\lambda\frac{c_0^2}{c_0+n}\kappa_n^{\mathrm{obs}}\Big)\Big\}\\ &= \int_0^\infty F_{p,p}\Big(\frac{c_0+n}{n},\lambda\frac{c_0^2}{c_0+n}\kappa_n^{\mathrm{obs}}\Big)g_n(\lambda)\,\mathrm{d}\lambda,\end{aligned} \tag{5.6}$$

in terms of $\kappa_n^{\mathrm{obs}} = n^{-1}(\widehat\beta^{\mathrm{obs}}-\beta_0)^{\mathrm{t}}\Omega_n(\widehat\beta^{\mathrm{obs}}-\beta_0)$ and the posterior gamma density g_n for λ implicit in (5.4). Its parameters are now

$$a_n = a+n \quad\text{and}\quad b_n = b + Q_0^{\mathrm{obs}} + \frac{c_0 n}{c_0+n}\kappa_n^{\mathrm{obs}}.$$

The (5.6) result is a generalisation of (2.2). A sometimes satisfactory approximation to the ppp is

$$\mathrm{ppp}(y^{\mathrm{obs}}) \approx F_{p,p}\Big(\frac{c_0+n}{n}, \frac{c_0^2}{c_0+n}\frac{\kappa_n^{\mathrm{obs}}}{(\widetilde\sigma^{\mathrm{obs}})^2}\Big),$$

where $1/(\widetilde\sigma^{\mathrm{obs}})^2 = a_n/b_n$ is the posterior mean of λ. This is as in (5.2), but with the estimated $\widetilde\sigma$, and works well if n is large.

For cppp analysis it is not a difficult task to compute $\mathrm{ppp}(y_k)$ as in (5.6) for a high number B of simulated data sets, leading by recipe (4.5) to the appropriate $\mathrm{cppp}(y^{\mathrm{obs}})$ number. Special properties of the linear regression model make it possible to simplify this step, however, as we now demonstrate.

Proposition. *Suppose the prior for (σ,β) is as in (5.3), with $\Omega_0 = n^{-1}\Omega_n$, and that Y given these parameters really follows the linear regression model $\mathrm{N}_n(X\beta,\sigma^2 I_n)$. Then the ppp distribution is*

$$G(u) = \Pr\{\mathrm{ppp}(Y)\le u\} = \Pr\{F_{p,p}(1+c_0/n,(c_0/n)Z)\le u\},$$

where $Z\sim\chi^2_p$.

This makes it easy to compute the necessary $\mathrm{cppp}(y^{\mathrm{obs}}) = G(y^{\mathrm{obs}})$ as the relative frequency of $F_{p,p}(1+c_0/n,(c_0/n)Z_k)\le \mathrm{ppp}(y^{\mathrm{obs}})$, across a million copies Z_k from the χ^2_p, since the non-central $F_{p,p}$ function is implemented in software packages like R; in particular there is no need to carry out the updating of the inverse gamma parameters or to perform the numerical integration in (5.6) each time. The G distribution is close to a uniform for large c_0 (corresponding almost to a classic p-value for testing $\beta=\beta_0$), is very tightly concentrated around $\frac12$ for small c_0 (corresponding to a nearly non-informative 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. Let $H = X(X^{\mathrm{t}}X)^{-1}X = X\Omega_n^{-1}X$ be the familiar 'hat matrix'. It is symmetric and idempotent, and a standard result is that for given (σ,β),

$$X\widehat\beta = HY\sim \mathrm{N}_n(X\beta,\sigma^2 H),\quad \widehat\varepsilon = Y - X\widehat\beta = (I-H)Y\sim\mathrm{N}_n(0,\sigma^2(I-H)),$$

and that these two n-vectors are independent. Taking (5.3) into account for the distribution of β given σ, one finds that the vectors remain independent, for given σ, with

$$X\widehat\beta \sim \mathrm{N}_n(X\beta_0, \sigma^2(1+n/c_0)H), \quad \widehat\varepsilon \sim \mathrm{N}_n(0, \sigma^2(I-H)).$$

This implies $Q_0(Y) = \sigma^2\|\widehat\varepsilon\|^2 \sim \sigma^2(V^*)^{\rm t}V^*$ and

$$\kappa_n(Y) = n^{-1}(X\widehat\beta - X\beta_0)^{\rm t}(X\widehat\beta - X\beta_0) = n^{-1}\sigma^2(1+n/c_0)V^{\rm t}V,$$

in terms of $V^* \sim \mathrm{N}_n(0, I-H)$ and $V \sim \mathrm{N}_n(0, H)$, with these two being independent. Hence $\kappa_n(Y) = \sigma^2(n^{-1} + c_0^{-1})Z$ and $Q_0(Y) = \sigma^2 Z^*$, where Z and Z^* are χ^2 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 $a_n = a + n$ and $b_n = b + W_n$ where $W_n \sim \sigma^2(Z + Z^*)$, and, more importantly,

$$\mathrm{ppp}(Y) = \Pr\Big\{\chi^2_p \ge \frac{n}{c_0+n}\chi^2_p\Big(\frac{1}{\sigma^2}\frac{c_0^2}{c_0+n}\sigma^2\Big(\frac{1}{n}+\frac{1}{c_0}\Big)Z\Big)\Big\},$$

in terms of a central and a non-central χ^2_p variable that both are independent of Z. The claim follows. ■

REMARK 5.1. When sample size n is large, $n^{-1}\Omega_n$ will be close to a limiting covariance matrix Σ for the covariates, and $\widehat\beta^{\rm obs}$ will be close to the true regression coefficient vector $\beta_{\rm tr}$. Some analysis shows that $\mathrm{cppp}(y^{\rm obs}) \to \Pr\{\chi^2_p \ge \mathrm{conf}^2\}$, where

$$\mathrm{conf} = c_0^{1/2}\{(\beta_{\rm tr} - \beta_0)^{\rm t}\Sigma(\beta_{\rm tr} - \beta_0)\}^{1/2}/\sigma$$

acts as a conflict measure between prior and model, also dependent on prior sample size c_0. This generalises Remark 4.1. See also Section 9.1. ■

6. Case study: regressing speedskaters

In the annual World Sprint Championships in speedskating, competitors skate the 500 m and 1000 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, t_4 are the results (in seconds) of the skater for the four distances; see Table 1, which gives the top of the result list for the 2004 championships at Nagano. In this section we will 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 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.

1	Erben Wennemars	140.755	35.70 (9)	1.09.46 (1)	35.63 (7)	1.09.39 (1)
2	Jeremy Wotherspoon	140.970	35.25 (1)	1.10.40 (4)	35.43 (3)	1.10.18 (4)
3	Mike Ireland	141.010	35.40 (3)	1.10.52 (5)	35.46 (4)	1.09.78 (3)
4	Gerard van Velde	141.150	35.64 (7)	1.10.73 (6)	35.27 (1)	1.09.75 (2)
5	Kip Carpenter	141.795	35.36 (2)	1.10.94 (9)	35.80 (9)	1.10.33 (5)
6	Casey FitzRandolph	141.995	35.50 (6)	1.10.84 (8)	35.55 (6)	1.11.05 (11)
7	Hiroyasu Shimizu	142.030	35.40 (3)	1.11.38 (14)	35.29 (2)	1.11.30 (14)
8	Janne Hänninen	142.070	35.71 (10)	1.10.38 (3)	35.91 (11)	1.10.52 (6)
9	Dmitri Lobkov	142.520	35.49 (5)	1.11.22 (13)	35.70 (8)	1.11.44 (16)
9	Masaaki Kobayashi	142.520	35.85 (12)	1.11.49 (15)	35.51 (5)	1.10.83 (9)
. . .						

Table 1. *The best ten skaters in the 2004 World Sprint Championships, held in Nagano, Japan. The results are the combined point sum, then the 500 m and 1000 m results for Saturday, followed by the 500 m and 1000 m results for Sunday.*

6.1. Setting the prior. The natural model to consider takes

$$y_i = b_0 + b_1 x_{i,1} + \cdots + b_p x_{i,p} + \varepsilon_i = b_0 + x_i^{\rm t} b + \varepsilon_i \quad \text{for } i = 1, \ldots, n,$$

in terms of a p-dimensional covariate vector x_i for each of n individuals, where the ε_is are independent zero-mean normals with standard deviation σ. For the speedskating applications, y_i is the 1000 m Sunday result, while $x_{i,1}, x_{i,2}, x_{i,3}$ are the results of the three previous distances. Our model is intended to convey the main mechanism of achievements during the World Championships among the say 30 best skaters or the world, during races without falls or accidents, i.e. outliers are screened out and not allowed to enter the model.

Setting a clear prior for (b_0, b, σ) is not an easy exercise. It does not quite do to have (b_1, b_2, b_3) centred 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 other), is most easily quantified in terms of (i) the overall level and its variability (expected average and standard deviation for the y_is) and (ii) the correlations between covariates $x_{\cdot,j}$ and y. We therefore develop a method for bringing such prior knowledge into a proper prior for the model parameters.

It is helpful first to centre the covariates by subtracting the averages $\bar{x}_{\cdot,j}$ from the $x_{i,j}$s above. Considering this done, $\bar{y} = b_0 + \bar{\varepsilon}$, giving b_0 a clear interpretation as the expected average result. Furthermore, $n^{-1}\sum_{i=1}^n y_i x_i = S_n b + n^{-1}\sum_{i=1}^n x_i \varepsilon_i$, where $S_n = n^{-1}\sum_{i=1}^n x_i x_i^{\rm t}$ is the empirical variance matrix of the covariate vectors, with elements say $s_{j,k}$ for $j, k = 1, \ldots, p$. While the considerations that follow actually might invite also other types of priors we wish for this illustration to follow the development and recipes of Section 5, so we are to specify a prior for σ and for the regression curve parameters given σ of the form

$$\begin{pmatrix} b_0 \\ b \end{pmatrix} \mid \sigma \sim \mathrm{N}_{p+1}(\begin{pmatrix} \bar b_0 \\ \bar b \end{pmatrix}, \sigma^2 \frac{1}{c_0} \begin{pmatrix} 1 & 0 \\ 0 & S_n^{-1} \end{pmatrix}). \qquad (6.1)$$

Since $n^{-1}\sum_{i=1}^n (y_i-\bar y)^2$ is seen to have mean value $b^{\mathrm{t}} S_n b + \sigma^2$, we see that the correlations $\rho=(\rho_1,\ldots,\rho_p)^{\mathrm{t}}$ between y and the respective covariates $x_{\cdot,j}$ may be represented as

$$\rho_j = \frac{(S_n b)_j}{(b^{\mathrm{t}} S_n b + \sigma^2)^{1/2} s_j} \quad \text{for } j=1,\ldots,p,$$

where $s_j = s_{j,j}^{1/2}$ is the standard deviation of $x_{\cdot,j}$. We therefore need a mechanism for turning prior information about ρ into a prior for 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 the equation

$$\frac{S_n b}{(b^{\mathrm{t}} S_n b + \sigma^2)^{1/2}} = v = D_n^{1/2}\rho$$

for b, where $D_n = \mathrm{diag}(s_1^2,\ldots,s_p^2)$ is the matrix with empirical covariate variances down its diagonal; thus v has components $s_j\rho_j$. The solution is

$$b = \sigma \frac{S_n^{-1} v}{(1 - v^{\mathrm{t}} S_n^{-1} v)^{1/2}}.$$

To follow the intended scheme we take

$$\frac{S_n^{-1/2} v}{(1 - v^{\mathrm{t}} S_n^{-1} v)^{1/2}} = z_0/\sigma + \tau N \sim \mathrm{N}_p(z_0/\sigma, \tau^2 I_p),$$

for a suitable location z_0 and scale parameter τ, in terms of an $N \sim \mathrm{N}_p(0, I_p)$. This also means that

$$(b \mid \sigma) \sim \mathrm{N}_p(S_n^{-1/2} z_0, \sigma^2 \tau^2 S_n^{-1}), \qquad (6.2)$$

which is consistent with the conjugate set-up of (5.3) and (6.1). To fine-tune the parameters, solve for $v = (s_1\rho_1,\ldots,s_p\rho_p)$ to find

$$S_n^{-1/2} v = \frac{z_0/\sigma + \tau N}{(1 + \|z_0/\sigma + \tau N\|^2)^{1/2}}. \qquad (6.3)$$

Our strategy is as follows: (i) Put up prior guess parameters $\rho_{j,0}$ for the correlations $\rho_1,\ldots,\rho_p$. This must in particular be done in a manner which reflects $v^{\mathrm{t}} S_n^{-1} v < 1$, for

$v = D_n^{1/2}\rho$. (ii) Then, for each of a sequence of trial values of the parameter τ, solve the p equations

$$g(z_0) = \mathrm{E}\frac{z_0/\sigma_0 + \tau N}{(1 + \|z_0/\sigma_0 + \tau N\|^2)^{1/2}} = S_n^{-1/2}v_0 = S_n^{-1/2}D_n^{1/2}\rho_0 \qquad (6.4)$$

for z_0, utilising also an initial prior guess value $1/\sigma_0$ for $1/\sigma$. In practice we solve these equations using stochastic simulation to compute $g_1(z_0), \ldots, g_p(z_0)$, since no formulae are available, and a non-linear minimisation algorithm like `nlm` in the `R` software package to minimise $\sum_{j=1}^p \{g_j(z_0) - (S_n^{-1/2}v_0)_j\}^2$. (iii) For the trial value of τ, and the found z_0, one generates ρ vectors from $\rho = D_n^{-1/2}v$, where v is taken according to (6.3); in particular one may monitor the spread of the ρ_js via their standard deviations and correlations (again, there are no explicit formulae to help us, hence the simulations). (iv) The procedure stops when a value of τ is found that well reflects the uncertainty level for the ρ_js 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 Λ, for ρ, and then to choose τ to match the mean of $v^{\mathrm{t}}S^{-1}v$, which is $v_0^{\mathrm{t}}S_n^{-1}v_0 + \mathrm{Tr}(D_n^{1/2}\Lambda S_n^{-1/2})$, with the mean of $\|z_0/\sigma + \tau N\|^2/(1 + \|z_0/\sigma + \tau N\|^2)$, from (6.3). The problem with this is the difficulty of setting a good Λ matrix, in that attempts at doing this would clash with the structure imposed by (6.3). Our strategy avoids this quandary.

6.2. ppp and cppp analysis of the 2004 speedskating model. We carried out the above exercise for the World Championships 2004 data. We used prior guess values 0.6, 0.8, 0.6 for the correlations (ρ_1, ρ_2, ρ_3) of $x_{\cdot,1}, x_{\cdot,2}, x_{\cdot,3}$ with y, and $1/0.75$ for $1/\sigma_0$; these values were elicited based on discussion amongst the authors and speedskating compatriots, and are meant to be based on solid experiences; for more background, see e.g. Hjort (2002, 2003a). One had

$$S_n = \begin{pmatrix} 0.365 & 0.625 & 0.343 \\ 0.625 & 1.878 & 0.572 \\ 0.343 & 0.572 & 0.386 \end{pmatrix},$$

and we could for each trial value τ solve (6.4) for z_0, in terms of $S_n^{-1/2}D_n^{1/2}\rho_0 = (0.094, 0.756, 0.294)^{\mathrm{t}}$. Monitoring spread and correlations in the prior distribution for ρ we settled on $\tau = 0.4$ to reflect prior beliefs. This corresponds to 'prior sample size' $c_0 = 1/\tau^2 = 6.25$, to standard deviations $0.144, 0.086, 0.144$ for the three correlations, centred around $0.6, 0.8, 0.6$, and to correlations $0.477, 0.846, 0.218$ for (ρ_1, ρ_2), (ρ_1, ρ_3), and (ρ_2, ρ_3). This also leads to $S_n^{-1/2}z_0 = (-1.014, 0.948, 1.015)^{\mathrm{t}}$ in (6.2), i.e. the prior mean $\bar{b}$ in (6.1). As prior mean $\bar{b}_0$ for b_0 we use the mean of all Saturday's 1000 m results, which is 71.856.

It remains only to specify a suitable inverse gamma prior for σ^2, to keep with the (5.3) recipe. To this end we may first show that

$$b^{\mathrm{t}}S_nb = \sigma^2\frac{v^{\mathrm{t}}S_n^{-1}v}{1 - v^{\mathrm{t}}S_n^{-1}v} \quad \text{and} \quad \kappa^2 = \mathrm{Var}\, y_i = \frac{\sigma^2}{1 - v^{\mathrm{t}}S_n^{-1}v},$$

which leads to $\sigma = \kappa(1 - v^{\rm t} S_n^{-1} v)^{1/2}$ in terms of the standard deviation for the y_i results themselves, for which a prior estimate is the standard deviation of Saturday's 1000 m races, namely 1.394. With our prior estimates $(0.6, 0.8, 0.6)^{\rm t}$ for the correlations this suggests prior mean value $a/b = 1/(\sigma^*)^2 = 1/0.804^2$ for $1/\sigma^2$, and the variance is $2a/b^2 = (2/0.804^4)/a = 3.094/a$. Comparisons with the estimated variance for the standard deviation of Saturday's 1000 m races leads us finally to put $a = 4.2$ and $b = 0.804^2 a = 2.815$ for the two Gamma prior parameters.

After the hard work of setting a proper prior it is a pleasure to finally watch Sunday's 1000 m races, and an easy task to produce the ppp and cppp numbers, following the results of Section 5, and to 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. centred at about $(0.75, 0.85, 0.75)$), but is uncorrelated with the main story of this article, so we report instead on the ppp analysis. Formula (5.6) gives $\text{ppp}(y^{\rm obs}) = 0.539$, while that section's Proposition readily can be applied to yield $\text{cppp}(y^{\rm obs}) = 0.781$. We may conclude that there is absolutely no conflict between our carefully constructed prior and the data, as monitored via the canonical discrepancy function $D(y, \theta) = \sum_{i=1}^n (\widehat\mu_i - \mu_i)^2/\sigma^2$. 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 so here.

7. ppp and cppp analysis of two bird survival models

Brooks, Catchpole and Morgan (2000; henceforth BCM) analysed recapture data for the European Dipper species *(Cinclus cinclus)* using ppp values. The same data have previously been analysed 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 six years in question; see below. Here we offer a re-analysis of their data, utilising our cppp tools, and reach conclusions partly different from those of BCM regarding adequacy of suggested models.

7.1. The data, the model, and the discrepancy. The data are given in Table 1. The models considered by BCM involve up to twelve parameters: for $i = 1, \ldots, 6$, ϕ_i represents the probability that a given bird survives year $1980 + i$, while p_i represents the probability of capturing a particular bird in year $1980 + i$.

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

TABLE 1. *Recapture data for the European Dipper.*

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

$$L(\phi, p) = \text{const.}\, \Delta(\phi, p) \prod_{i=1}^{6} \prod_{j=i+1}^{6} \Big(\phi_i p_j \prod_{k=i+1}^{j-1} \phi_i (1 - p_k)\Big)^{y_{i,j}},$$

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

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

We have analysed two versions of this general model. The large model (T/T in the terminology of BCM) employs all twelve parameters (ϕ_i, p_i) while the small model (C/C in their terminology) takes all ϕ_js equal and all p_js 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. Firstly, the more focussed priors also briefly worked with by BCM do not appear to fit data particularly well, and BCM give little indication that they represent actual prior knowledge of the parameters. Secondly, they quote 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 0.075, while BCM quote the number 0.086; similarly, we find ppp equal to 0.060 where BCM give 0.069, for the small model). 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.

7.2. ppp and cppp analysis of the two models.

The large model. We have sampled the distribution of ppp along the lines indicated above, and found $\text{ppp}(y^{\text{obs}}) = 0.075$. 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 clearly unimodal. The sample standard deviation is 0.172, compared to 0.289 for a uniformly distributed p-value. More important is that the fact that only a single one of our 500 simulated ppp values was below 0.075, giving an estimated cppp value (the true surprise level) of 0.002.

The small model. In this case we find $\text{ppp}(y^{\text{obs}}) = 0.060$. 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 to the twelve parameter case. The sample standard deviation of ppp is now 0.257. With calibration we find cppp $= 0.022$ (11 out of 500 simulated ppp(Y) values were less than 0.060). We therefore conclude that the given data are much less likely under the large model than under the small one, although the nominal ppp values give the opposite relation.

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 one. We indicate how this might be put to use in two situations.

8.1. Nonparametric vs. parametric cdf. Consider independent data $y_1, \ldots, y_n$ drawn from a distribution F, where two quite different priors are under consideration for F. The first 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 $F \sim \mathrm{Dir}(aF_0)$, a Dirichlet process with centre distribution F_0 and concentration parameter a; see e.g. Hjort (2003b) for a recent review of nonparametric Bayesian statistics. For discrepancy measure we use

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

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 there to better understand the situation when n grows. In fact, the $D(y^{\mathrm{rep}}, F) = \sqrt{n}\|F_n^{\mathrm{rep}} - F\|$ is then close 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, Ch. 4). In particular, $D(y^{\mathrm{rep}}, F) \to_d \|W^0\| = \max_t |W^0(t)|$, 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 are to study the behaviour of the nonparametric and normal-parametric prior specifications, say $\mathrm{ppp}_{\mathrm{Dir}}$ and $\mathrm{ppp}_{\mathrm{N}}$. First study the nonparametric prior. Here F given data is a Dirichlet process with parameter $aF_0 + nF_n^{\mathrm{obs}}$. One may show that $\sqrt{n}(F - F_n^{\mathrm{obs}}) \to_d \widetilde{W}^0(F_{\mathrm{tr}}(\cdot))$ a.s., where $\widetilde{W}^0$ is another Brownian bridge, independent of W^0; this follows e.g. from work of Hjort and Ongaro (2004). Hence

$$\mathrm{ppp}_{\mathrm{Dir}}(y^{\mathrm{obs}}) = \Pr\{\sqrt{n}\|F_n^{\mathrm{rep}} - F\| \ge \sqrt{n}\|F_n^{\mathrm{obs}} - F\| \,|\, \mathrm{data}\} \to \Pr\{\|W^0\| \ge \|\widetilde{W}^0\|\} = \tfrac{1}{2}$$

a.s., as n grows towards infinity. Then 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 $\pi(\mu, \sigma \,|\, \mathrm{data})$. Here $\|F_n^{\mathrm{obs}} - F\|$ goes a.s. to $\|F_{\mathrm{tr}} - F_{\mathrm{appr}}\|$, essentially by the Glivenko–Cantelli theorem, where $F_{\rm appr}(t) = \Phi((t-\mu_{\rm tr})/\sigma_{\rm tr})$ is the best parametric approximant to $F_{\rm tr}(t)$, inside the normal family. It follows that $\sqrt{n}\|F_n^{\rm obs} - F\|$ goes to infinity a.s., as long as the real $F_{\rm tr}$ is not fully equal to a normal distribution function, and, in particular, $\text{ppp}_{\rm N}(y^{\rm obs}) \to 0$ a.s. as $n \to \infty$.

The results above help us understand the behaviour of the two ppp measures, for large n. We would also need to calibrate these, to reach the more interpretable cppp values given data. This can be done via double simulation, as per the general guidelines about this laid out in Section 4.2.

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

8.2. Nonparametric vs. parametric hazard rate models. Bayesian analysis of survival and event history data appear to fall into one of two separate categories, viz. the parametric and the nonparametric. Only rarely does one see any formal justification for choosing one path over the other. The cppp analysis makes it in principle easy, conceptually and operationally, to do such a comparison. The parametric model might use a Weibull hazard rate $H(t) = (\theta t)^\gamma$, with a prior on (θ, γ), where the nonparametric alternative could use a Beta process. As discrepancy measure one might use $D(y, H) = \int_0^\tau w(t)\{H_n(t) - H(t)\}^2\,{\rm d}t$, for a weight function w, involving the Nelson–Aalen estimator H_n for H.

9. Related themes and concluding remarks

We end our article with a list of comments and indication of themes, some of which might warrant further research efforts.

9.1. 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 centre 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 scheme or way of thinking not infrequently followed is to try out different precision levels, from 'reasonably precise' to 'quite non-informative', 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' elicitations indirectly are of such a form.

The procedure just described is somewhat ad hoc, of course, and may not be easy to follow or formalise in practice. The cppp mechanism offers a venue leading to a precise version of this idea, however. The proposal is to monitor the $\text{cppp}(y^{\rm obs})$ 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 $\mathrm{cppp}(y^{\mathrm{obs}}) = 0.10$, but not lower.

As a mundane illustration of this cppp-induced conservative prior, study again the normal–normal setup of Section 2 and Section 4.1. Assume 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 (4.4), the idea above amounts to selection σ_0 to have

$$\frac{n(\bar{y}^{\mathrm{obs}} - \theta_0)^2/\sigma^2}{1 + n\sigma_0^2/\sigma^2} = 1.645^2, \quad \text{or} \quad \sigma_0 = \frac{\sigma}{\sqrt{n}}\Big(\frac{z_n^{\mathrm{obs}}}{1.645^2} - 1\Big)^{1/2},$$

where $z_n^{\mathrm{obs}} = n(\bar{y}^{\mathrm{obs}} - \theta_0)^2/\sigma^2$ is the usual test statistic for testing the null hypothesis that $\theta = \theta_0$. This formula would be used when $z_n^{\mathrm{obs}} \geq 1.645^2$, i.e. when the test rejects θ_0 at significance level 0.10. In cases where the $\theta = \theta_0$ hypothesis is accepted by the data, the scheme above would allow a sharp prior at θ_0, i.e. setting σ_0 to a small value.

We stress that the idea is quite general and might be used in situations much less clear-cut than the above, e.g. in semi- and nonparametric contexts. One type of application would be to nonparametric setups involving Dirichlet or Beta processes, where the statistician knows where to centre these but is unsure about the concentration parameters.

9.2. Detection power. There is not room here for properly discussing the detection power of the cppp assessors. This clearly depends on aspects of the situation at hand, including the discrepancy function $D(y, \theta)$. For the normal–normal setup of Section 2 it is not difficult to study the distribution of $U = \mathrm{ppp}(Y)$ under various conditions different from those implied by the prior and the model; we will report elsewhere on some such findings. While 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 $D(y, \theta)$. For detecting special types of violations one might therefore need to devise corresponding special discrepancy functions.

9.3. cppp as a p-value. One way of viewing our cppp construction is that the $\mathrm{ppp}(y^{\mathrm{obs}})$ of (1.1), albeit clearly having Bayesian interpretation and inspiration, is nothing but a test statistic. Since 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.

9.4. Tail versus height. When attempting to assess the degree of surprise in an observation $U = u$, one might compute tail areas, i.e. $G(u) = \mathrm{Pr}_0\{U \leq u\}$ for the relevant null distribution, or a suitable ratio of densities $h(u)/g(u)$. This article has been concerned with the first general direction, partly because is would be difficult to find good general candidates for the $h(u)$ in question, and partly since the null densities $g(u)$ have been seen to be so extreme, cf. Figure 4.1.

9.5. General hierarchical models. In complicated hierarchical models it is easy to lose track of the different implications of a many-levelled prior. Ordinary assessment methods may not work well in such cases, see e.g. comments made in Lu, Hodges and Carlin (2004) about the otherwise general-purpose DIC measure. The cppp methods of our article may be generalised to various hierarchical models, and we believe they may be useful for screening out unfortunate combinations of prior and model there, when employed with appropriate discrepancy functions. Such might be constructed along the lines of Day, Gelfand, Swartz and Vlachos (1998).

References

Bayarri, M.J. and Berger, J.O. (2000). P values in composite null models [with discussion and a rejoinder]. *Journal of the American Statistical Association* **95**, 1127–1142.

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.* Wiley, Singapore.

Box, G.E.P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society* **A 143**, 383–430.

Claeskens, G. and Hjort, N.L. (2003). The focused information criterion [with discussion contributions and a rejoinder]. *Journal of the American Statistical Association* **98**, 900–916 and 938–945.

Brooks, S.P., Catchpole, E.A. and Morgan, B.J.T. (2000). Bayesian animal survival estimation. *Statistical Science* **15**, 357–376.

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* **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., Meng, X.-L. and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies [with discussion]. *Statistica Sinica* **6**, 733–807.

Guttiérrez-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* **B 29**, 83–100.

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 University Press, Oxford, pp. 423–453.

Hjort, N.L. (2002). Pre-Olympic practice of the Canadian anthem. (Report on the World Sprint Championships.) *SpeedSkating World* **7**, 26–29.

Hjort, N.L. (2003a). Catriona LeMay Doan: Going for Gold. (Book review.) *SpeedSkating World* **8**, 14–15.

Hjort, N.L. (2003b). Topics in nonparametric Bayesian statistics [with discussion]. In *Highly Structured Stochastic Systems* (eds. P.J. Green, N.L. Hjort and S. Richardson), Oxford University Press, Oxford, pp. 455–478.

Hjort, N.L. and Claeskens, G. (2003). Frequentist model average estimators [with discussion contributions and a rejoinder]. *Journal of the American Statistical Association* **98**, 879–899 and 938–945.

Hjort, N.L. and Ongaro, A. (2004). Exact inference for random Dirichlet means. To appear.

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.* Wiley, New York.

van der Linde, A. (2004). On the association between a random variable and an observable. *Test*, to appear.

Lu, H., Hodges, J.S. and Carlin, B.P. (2004). Measuring the complexity of generalized linear hierarchical models. Technical report.

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. *Annals of Statistics* **22**, 1142–1160.

Robins, J.M., van der Vaart, A. and Ventura, V. (2000). Asymptotic distribution of p values in composite null models [with discussion and a rejoinder]. *Journal of the American Statistical Association* **95**, 1143–1156.

Rubin, D.B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *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* **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 and a rejoinder]. *Journal of the Royal Statistical Society* **B 64**, 583–639.