Posterior Predictive p-values in Bayesian Hierarchical models

Gunnhildur Högnadóttir Steinbakk

Department of Mathematics, University of Oslo, Norway

Geir Olve Storvik

Department of Mathematics and Centre for Ecological and Evolutionary Synthesis, University of Oslo, Norway

Summary. The question of interest is whether a particular model appears to provide an adequate fit with the observed data. In complicated hierarchical models it is easy to loose track of the different layers, and standard and existing methods may not work very well in such cases. The present work focuses on extensions of the posterior predictive p-value for models with hierarchical structure, designed for testing assumptions made on underlying processes. The posterior predictive p-values are typically non-uniform under the model assumptions. However, a post-processing of the p-values, generally computed by a double simulation scheme, makes the resulting calibrated p-values uniformly distributed under the prior and model conditions. By exploring posterior predictive p-values and their post-processed versions in models of simple structures, the effect of signal to noise levels is evaluated for testing different parts of the model. A real application illustrates how the post-processed posterior predictive p-values can be used in practise.

Keywords: Calibration of p-values, Hierarchical models, Model criticism, Posterior predictive p-values

1. Introduction

The class of models considered here is hierarchical, where the data process y and the underlying processes (x, θ) are related through a parametric model $\pi(y|x, \theta)$. Then x is in turn modelled with distribution $\pi(x|\theta)$ conditional on the hyper parameter θ. To complete our Bayesian model, we select a prior distribution $\pi(\theta)$ for the hyper parameter θ. Note that

the information about the x's and the parameters θ is available through the data process $\pi(y|x,\theta)$ and their model specification $\pi(x,\theta)$. We observe y^{obs} and want to assess whether y^{obs} is compatible with the assumed model and prior. In complicated hierarchical models it is easy to loose track of the different layers which highlights the increased need to check adequacy in such models.

The posterior predictive p-value (ppp) (Guttman 1967, Rubin 1984, Meng 1994, Gelman et al. 1996) is a method for carrying out model evaluations and comparisons which has become fairly popular in Bayesian model checking, partly in consequence of it's easy implementation by MCMC methods. The intension is to quantify the degree of surprise by observing what we actually have observed, in view of the prior and model. The ppp is defined as (Gelman et al. 1996)

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

where D indicates a particular discrepancy measure used for computing a ppp number. The replication y^{rep} follows the predictive distribution $\int \pi(y^{\text{rep}}|\theta)\pi(\theta|y^{\text{obs}})\,\mathrm{d}\theta$, with $y^{\text{rep}}|\theta$ equals $y^{\text{obs}}|\theta$ in distribution. The ppp apparatus supplies generous flexibility through choices of discrepancy functions, chosen to reflect important aspects of the statistical model. Unlike traditional test statistics, discrepancy measures are allowed to depend upon the unknown parameters.

For comparing and evaluating, we want the p-values to have the same interpretation across different combinations of models and priors. An important characteristic of a frequentistic p-value is, when considered as a random variable, a uniform distribution under the null model (at least approximatively) Recent work has shown that ppp-values are far from uniform (Robins et al. 2000, Bayarri & Berger 2000, Hjort et al. 2006), being conservative because of their double use of data. Alternative posterior predictive p-values assumed to be closer to uniformly distributed have been suggested by Bayarri & Berger (2000). Those p-values (e.g., conditional posterior predictive and partial posterior predictive) are sometimes difficult to compute. Robins et al. (2000) proposed a re-parameterisation of the test statistics (and the discrepancy measure) making the resulting ppp asymptotically uniform under the null-model. They also suggested some adjusted p-value using the non-uniform asymptotic distributions. In contrast to posterior predictive p-values, Box (1980) considered prior predictive p-values. These p-values are by definition uniformly distributed under the

prior and model assumptions. Another approach was proposed by Hjort et al. (2006) where they used a double simulation scheme for calibrating the ppp-value in order to make them uniform. In this connection, there are different ways of regarding the null-model. Either we may look at the distribution of p-values under fixed θ's (Robins et al. 2000, Bayarri & Berger 2000), or under the total model including the prior distribution (Box 1980, Meng 1994, Dey et al. 1998, Hjort et al. 2006).

Hierarchical models are used for representing complex phenomena, providing a flexible framework for modelling dependencies and interrelationships between the models units. The hierarchical structure gives great flexibility of modelling in the sense that it can be used to represent widely different phenomena, incorporating interrelationship between the model elements in a natural manner (Banerjee et al. 2004, Skrondal & Rabe-Hesketh 2004). For example, when the underlying x process is assumed Gaussian (the far most used type of model, see Diggle et al. 1998), the hierarchical model formulation allows us to specify models for dependency between outcomes through correlations in the underlying x's. Often, the data points are assumed to be independent conditional on x and θ, while the unconditional observation process is correlated induces by the underlying structure of x.

Assessing the validity of model assumptions in Bayesian hierarchical models, Dey et al. (1998) carried out a simulation based model checking, for which their work requires proper priors. Alternatively, O'Hagan (2003) describes a methods with use of conflict measures in Bayesian hierarchical models. According to Bayarri & Castellanos (2007) the method is highly sensitive to the prior, i.e it is conservative with non-informative priors. Dahl et al. (2007) proposed a data-splitting method, avoiding a double use of data, for compensating for this conservatism. Conflict p-values, generalisations of cross-validation methods, were proposed by Marshall & Spiegelhalter (2003). Bayarri & Castellanos (2007) have extended the 'partial posterior predictive' p-value to hierarchical models. Within the same paper, they discuss some model checking in hierarchical models using posterior predictive p-values with test statistics that are functions depending on data only. A simulation study of posterior predictive checking in hierarchical models is presented in Sinharay & Stern (2003). Lu et al. (2006) looked at the complexity in generalised linear hierarchical models, but their method requires exchangeable processes. Gelman et al. (2005) focused on checking qualitative aspects of the hidden process, using a generlisation of the standard ppp.

Extending the definition of (1) to hierarchical models provides different possibilities for

defining a ppp-number. The standard ppp is limited to discrepancy measures that depend upon data, otherwise the discrepancy measure on each side of the inequality in (1) becomes equal quantities and the value of ppp is always one. In hierarchical settings, it may be difficult to construct discrepancy measures only based on the observations and the hyper parameters. Depending on the context, we may test assumptions made on θ, $x|\theta$ or $y|(x,\theta)$, as model failure can occur at each stage. In many situations we are more interested in the dynamics of the x process than the observation process itself, for example in disease mapping based on count data, where the number of observed cases of a specific disease is used for making inference about the underlying mortality rate. Latent continuous responses underlying categorical variables, for example dichotomous or ordinal data describing the level of pain, demonstrates another setting where the process of x could typically be in focus. Hence, creating discrepancy measures directly on the underlying processes independently of the data, are reasonable, and often easier, if our intention is to test assumptions made on x or θ (Sinharay & Stern 2003).

Allowing for discrepancy measures depending on latent variables is no guarantee of an appropriate validation of specific model aspects. Assumptions made on unobservable parameters are usually difficult to check. The prior-specifications are often intentionally vague, since information about the model processes decreases further away from data (Gelfand 2003). Model criticism or adequacy for vague hyper prior models, over-parameterised models and multi-level models are been stated as almost impossible (Gelfand 2003). Instead, he claims that model adequacy should rather be issued in moderate dimensional models, both regarding the parameter space and hierarchical layers, with slightly informative prior. Dey et al. (1998) distinguish stage-wise model checking from checking the model adequacy with observed data since the former is a property of the model specification.

In this paper we discuss extensions of the standard ppp-value, suitable for checking hierarchical models. These extensions allow for discrepancy measures depending on the latent x process. The *extended* posterior predictive p-value (eppp) first define a complete ppp value assuming x to be known, and then calculate the posterior expectation of this complete ppp value. This approach is also suggested by Gelman et al. (2005). The posterior predictive marginalised p-value (ppp^{mrg}) on the other hand first integrate out the unknown x in the discrepancy measures giving a discrepancy measure only depending on y^{obs} and θ for which the standard definition of ppp (1) can be used. Both eppp and ppp^{mrg} are

easily computed by outputs from a MCMC routine. There is almost no need for additional programming or computation as we have simulations from the posterior distributions. As is the case for the ppp, the eppp is typically not uniformly distributed under the model assumptions. To get a reference scale for our eppp-value in order to know if the value is extreme or not, we calibrate the eppp-value under the (perfect) model conditions. These post-processed posterior predictive p-values, named ceppp and cppp$^{\text{mrg}}$, will by definition have uniform null distributions under the model and the prior. Note that the present article considers the total model situation, so within the hierarchical model, we do concern about the p-values being uniform under the combination of the data model $\pi(y|x,\theta)$ and the prior $\pi(x,\theta)$.

The outline of this article is as follows. In section 2 we present the extended versions of the ppp's to Bayesian hierarchical models and their transformations to uniformly distributed p-values. An important task for this paper is to study the level of ambitions for model adequacy in multi-level models. In section 3 properties and aspects of the p-values are studied in the context of two structurally simple models where all levels are normal, using discrepancy measures testing prior specifications. Another discrepancy measure testing the underlying process is applied to the same simple models in section 4. An important task of section 3 and 4 is to investigate whether we can criticise the model at different stages, including stages with hidden processes and hyper parameter specifications within some rather simple example models in the context of the p-values that are now to be described. The methods are then illustrated on a real data set, a subset of observations used for estimation of catch-at-age for landings of fish, in section 5. Finally, the section 6 includes summary and conclusions.

2. The ppp in Bayesian hierarchical models

We wish to use a similar definition of ppp in (1) for hierarchical models. Within such models there is some flexibility on how to treat the latent x process. Treating x as a parameter along with θ, allow us to use the same definition (1) with a discrepancy function $D(y,x,\theta)$, by which ppp becomes a probability within the sample space of $(y^{\text{rep}},x,\theta)$ with density $\pi(y^{\text{rep}}|x,\theta)\pi(\theta,x|y^{\text{obs}})$ conditional on the data (see Sinharay & Stern (2003) for an illustration). Alternatively, the x process can be considered as a set of missing variables

which we would have treated as part of y^{obs} if they where known. The hidden property of x can then be treated by using the posterior expectation of the "complete" ppp measure.

The first scheme requires the discrepancy measure D to depend on y, while the other scheme is possible to use also in cases where D only depend on x and θ. Gelman et al. (2004) state that both schemes can be used to assess model adequacy. We will concentrate on the latter scheme. Our main motivation for this is the larger flexibility of choosing discrepancy measures. Nevertheless, the measure must be carefully constructed to test specific and relevant aspects of interest in our data and model. Which part of the model you want to check will always depend on the actual application and should be directed towards those model aspects that are essential and relevant for the main conclusions of the statistical analysis.

2.1. The extended posterior predictive p-value

The intention is to quantify conflict in data in view of the multi level model. If we could observe $x = x^{\text{obs}}$, we would continue to use the ppp definition in (1) utilising the same ppp-framework as before with

$$\text{ppp}^{\text{compl}}(y^{\text{obs}}, x^{\text{obs}}) \quad = \quad \Pr\{D(y^{\text{rep}}, x^{\text{rep}}, \theta) \geq D(y^{\text{obs}}, x^{\text{obs}}, \theta) | x^{\text{obs}}, y^{\text{obs}}\}, \tag{2}$$

directly treating x as part of the observations. Here y^{rep} and x^{rep} are distributed as $\pi(y^{\text{rep}}|x^{\text{rep}}, \theta)$ and $\pi(x^{\text{rep}}|\theta)$, respectively. In the case where the x's are missing, we define

$$\begin{aligned} \text{eppp}(y^{\text{obs}}) =& \text{E}^{x|y^{\text{obs}}}[\text{ppp}^{\text{compl}}(y^{\text{obs}}, x)] \\ =& \text{E}^{y^{\text{rep}}, x^{\text{rep}}, x, \theta | y^{\text{obs}}}\big[I\{D(y^{\text{rep}}, x^{\text{rep}}, \theta) \geq D(y^{\text{obs}}, x, \theta)\}\big], \end{aligned} \tag{3}$$

where $\pi(y^{\text{rep}}, x^{\text{rep}}, x, \theta|y^{\text{obs}})$ is given by the products $\pi(y^{\text{rep}}|x^{\text{rep}}, \theta)\pi(x^{\text{rep}}|\theta)\pi(x, \theta|y^{\text{obs}})$ and the densities $\pi(y^{\text{rep}}|x^{\text{rep}}, \theta)$ and $\pi(x^{\text{rep}}|\theta)$ are equal to $\pi(y|x, \theta)$ and $\pi(x|\theta)$, respectively. We see that eppp includes an expression for ppp. By pretending that the x's are observed, the ppp in (2) gives nothing new, but is only the definition of the posterior predictive p-value treating x as a part of the observations. In general we do not observe the x's, and our natural solution is to average over the x by taking the expectation of $\text{ppp}^{\text{compl}}$ with respect to the posteriori distribution of x. In this way, we have constructed an eppp within a coherent framework that clearly distinguish the different role of $x \sim \pi(x|y^{\text{obs}})$ and $x^{\text{rep}} \sim \pi(x|\theta)$,

and consequently, allows for any $D(y, x, \theta)$ that could be a function of the x's only. Note that eppp is similar to the definition of Gelman et al. (2005).

In general, there is no analytical expression for the eppp, but we can approximate eppp by

$$\text{eppp}(y^{\text{obs}}) \approx \frac{1}{M}\sum_{j=1}^{M}\left[I\{D(y^{\text{rep},j}, x^{\text{rep},j}, \theta^j) \geq D(y^{\text{obs}}, x^j, \theta^j)\}\right], \tag{4}$$

provided we can simulate $(y^{\text{rep},j}, x^{\text{rep},j}, x^j, \theta^j)$ from $\pi(y^{\text{rep}}, x^{\text{rep}}, x, \theta|y^{\text{obs}})$ across a high number of M simulations. Due to the way we have defined eppp, the conditional joint distribution $\pi(y^{\text{rep}}, x^{\text{rep}}, x, \theta|y^{\text{obs}})$ is simplified by the products of conditional independent densities $\pi(y^{\text{rep}}|x^{\text{rep}}, \theta)\pi(x^{\text{rep}}|\theta)\pi(x, \theta|y^{\text{obs}})$. Hence, we need to sample from the above conditional distributions, where the most difficult distribution to sample from is $\pi(x, \theta|y^{\text{obs}})$, which is usually carried out when assessing inference on the parameters and the system process. Simulations from $\pi(y^{\text{rep}}|x^{\text{rep}}, \theta)$ as well as $\pi(x^{\text{rep}}|\theta)$, are simple and require little extra computation.

Similar to ppp, eppp will in general not be uniformly distributed under the model assumptions. However, the eppp can be calibrated or transformed to a uniform scale under the nature of assumed model, see section 2.3.

2.2. The marginalised discrepancy measure

In eppp we average over ppp, but instead we can average over the discrepancy measure before computing ppp. The standard ppp framework needs discrepancy measures that are functions of data, possible depending on an optional parameter vector θ. Hence, to utilise the existing methods for ppp in (1) for hierarchical models, we remake $D(y, x, \theta)$ so that it depends on y^{obs} and θ only, by computing

$$D^{\text{mrg}}(y, \theta) = \text{E}_\theta[D(y, x, \theta)|y] \tag{5}$$

for fixed θ's, and define

$$\begin{aligned}\text{ppp}^{\text{mrg}}(y) &= \text{ppp}(y; D^{\text{mrg}}) \\ &= \Pr\{D^{\text{mrg}}(y^{\text{rep}}, \theta) \geq D^{\text{mrg}}(y^{\text{obs}}, \theta)|y^{\text{obs}}\}.\end{aligned} \tag{6}$$

We simply integrate the discrepancy measure over the x-process instead of integrating x out of the ppp. Other possibilities of eliminating x, is to compute $\min_x D(x, \theta)$ or $D(\hat{x}, \theta)$

for an estimate of $\hat{x}$. Transforming the discrepancy measure to a function of y and θ, we may apply the standard ppp apparatus (1) directly. Gelman et al. (1996) proposed similar techniques for constructing a classical test statistic from a discrepancy measure in one-level models, as did Sinharay & Stern (2003) for eliminating x from $D(y, x, \theta)$. Computing D^{mrg} is often non-trivial. ppp^{mrg} is approximated through Monte Carlo simulations, similar to the approximation (4). In addition, D^{mrg} typically needs to be approximated as well, making the need for extra Monte Carlo simulations within the each iteration of the standard simulation from $\pi(x, \theta|y)$ for fixed θ's This procedure can be computer time demanding. Gelman et al. (1996) say that those kind of measures are only interesting for a theoretically point of view, due the computational complication, especially for complex model.

We could also average (5) over θ, where the resulted marginalised measure is a test statistic that do not depend on any unknown quantities. Because our main intention here has been to consider posterior predictive p-values obtained directly from standard MCMC runs, our discussion is restricted to ppp^{mrg}.

2.3. Calibration of the p-values

Both eppp and ppp^{mrg} are well defined and legitimate probabilities, but willing to use a p-value as a tool for model adequacy, you need to know if the computed p-value is surprising or not. We will in this section discuss how to interpret eppp values although an equivalent approach can be applied to ppp^{mrg}.

How can we judge the significance of an eppp number or compare it to other eppp values computed for different combinations of model and prior without a common underlying probability scale? Here we introduce the calibrated eppp where we use an appropriate distribution of $\text{eppp}(Y)$ in order to calibrate eppp to an uniform [0,1] scale.

In a non-hierarchical setting, the 'null-null distribution' was motivated by Hjort et al. (2006) as the distribution of $\text{ppp}(y)$ across precisely those y values that occur under a perfect world, that is, by the mechanism of the assumed prior and model. The null-null distribution of $\text{eppp}(Y)$ is similarly defined by

$$G(u) = \Pr\{\text{eppp}(Y) \leq u\} \tag{7}$$

where Y follows the marginal distribution $p(y) = \int\int \pi(y|x, \theta)\pi(x|\theta)\pi(\theta)\,\mathrm{d}\theta\,\mathrm{d}x$. In the same

way as Hjort et al. (2006), we define

$$\text{ceppp}(y^{\text{obs}}) = G(\text{eppp}(y^{\text{obs}})) = \Pr\{\text{eppp}(Y) \leq \text{eppp}(y^{\text{obs}})\} \tag{8}$$

as the calibrated eppp, whose distribution is by construction uniform on [0,1]. While the eppp might give limited information itself, the rescaled ceppp version becomes interpretable across different combinations of priors and models. Whereas the eppp value typically becomes less dependent on the prior for a given data model as the amount of data increases, the construction of (8) implies that ceppp remains dependent on the prior specifications. The implication of a given prior is actively used and assessed by the calibrated eppp. Obviously, we may transform ppp^{mrg} to an appropriate scale by the techniques presented here, resulting in cppp^{mrg}.

The calibration of eppp can be done under alternatives to the perfect distribution of (7), for example under a slightly more informative prior than the one used for computing the eppp-value. In particular, we may use the marginal likelihood $\pi(y;\hat{\theta})$ for calibration, where $\hat{\theta}$ is an estimate of the (true) θ (Robins et al. 2000, Hjort et al. 2006, Draper & Krnjajic 2006). It should be noted that our calibrated p-value relies on that we can sample from the prior. Obviously, we can not sample from improper priors. But also mathematically proper priors may give unreasonable outcomes. Fairly typically statisticians use Bayesian methods for computational convenience, which involve assigning an improper prior to the model and drawing samples from the posterior distribution by MCMC techniques. Non-informative priors are also used from the perspective of wanting to be objective. In this context, Hjort et al. (2006) argued the distinction should rather be between prior that makes sense to sample under and prior that produces highly unreasonable outcomes, and not between the mathematically proper and improper prior

Standard prior predictive p-values (Box 1980) can only handle test statistics that do not depend on the unknown (x, θ). Within the ceppp-machinery, we can select discrepancy measure specially designed for testing unknown variables and parameters. The $\text{ceppp}(y^{\text{obs}})$ is interpretable in this context as a prior predictive p-value using $\text{eppp}(y^{\text{obs}})$ of (3) as a test-statistic.

In general ceppp has to be approximated by the following double simulation method. Simulate (y_k, x_k, θ_k) for $k = 1, ..., B$, for a high number B, where (x_k, θ_k) is drawn from

$\pi(x, \theta)$ and y_k from $\pi(y|x_k, \theta_k)$. Then compute

$$\text{ceppp}(y^{\text{obs}}) \approx \frac{1}{B}\sum_{k=1}^{B} I\{\text{eppp}(y_k) \leq \text{eppp}(y^{\text{obs}})\}. \tag{9}$$

The double simulation follows since in addition, eppp is approximated by (4), resulting in B times M operations.

3. Discrepancy measures testing prior specifications

In this section we investigate the different p-values in two simple Gaussian models using discrepancy measures checking qualifications of the prior. These example models are described in detail in sections 3.1 and 3.2, respectively, and have hierarchical structures that effort insight in the behaviour of the p-values, helping us to understand the important mechanisms in more complicated models.

These example models might seem too simple. In complicated models we easily loose the details, so the simplicity and transparency will more clearly set out important characteristics and mechanisms of the different p-values (e.g., eppp, ceppp, ppp^{mrg} and cppp^{mrg}). We assume all variances to be known. The situation of unknown variances will complicate the formulae, making it harder to interprete the results.

While we in this section concentrate on discrepancy measures designed for testing specifications of the prior, we use smeasure inspired by the Kolmogorov-Smirnov test in section 4, intending to check the underlying process x. A key issue is that neither of these discrepancy measures do contain the observational process. Through these simple examples, we discuss the degree of complexity of testing the underlying x-process and the hyper parameter θ regarding the size of observation error and a priori knowledge. The proofs of this section propositions and corollaries are given in the appendix.

3.1. The simple normal model

Model $\mathcal{M}_1$ assumes normally distributed data y_i for $i = 1, \ldots, n$, with mean x_i and known variance τ^2, conditionally independent given the x's. The model further assumes the underlying process $x_1, ..., x_n$ conditional on θ, to be i.i.d. normal with mean θ and known variance σ^2, and finally, the hyper parameter θ is $N(\theta_0, \sigma_0^2)$. Table 1 summarises the model along with the other model $\mathcal{M}_2$ treated in section 3.2.

Table 1. The two simple hierarchical normal models ($\mathcal{M}_1$ and $\mathcal{M}_2$) used to illustrate the posterior predictive p-values in sections 3 and 4. The column furthest to the right lists the discrepancy measure used in the corresponding model.

	1st stage	2nd stage	Hyper parameters	D-measures
$\mathcal{M}_1$:	$y_i \sim N(x_i, \tau^2)$	$x_i \sim N(\theta, \sigma^2)$	$\theta \sim N(\theta_0, \sigma_0^2)$	D^{chisq}, D^{kolmo}
$\mathcal{M}_2$:	$y_i \sim N(x_i, \tau^2)$	$x_i \sim N(Z\beta, \sigma^2)$, $\beta = (\beta_1, \beta_2)'$	$\beta \sim N(\beta_0, \sigma^2 M_0^{-1})$, $\beta_0 = (\beta_{0,1}, \beta_{0,2})'$, $M_0 = \text{diag}(\kappa_1, \kappa_2)$	D^{β_2}, D^{kolmo}

Following Hjort et al. (2006), we consider the discrepancy measure

$$D^{\text{chisq}}(x, \theta) = D(y, x, \theta) = \frac{n(\bar{x} - \theta)^2}{\sigma^2}, \tag{10}$$

which is a function of x and θ only. Our purpose here is to examine the properties of this discrepancy measure in a hierarchical setting.

3.1.1. The eppp *with* D^{chisq}

For the following result, let $F_{1,1}(\nu, \kappa) = \Pr\{(U + \sqrt{\kappa})^2 \leq \nu V^2\}$ be the cumulative distribution function of a non central Fisher variable with degrees of freedom (1,1) and eccentricity parameter κ, where U and V are independent and standard normal variables.

Proposition 1 *The expected posterior predictive p-value for the simple normal model* $\mathcal{M}_1$ *can be written as a function of data expressed by*

$$\begin{aligned} \text{eppp}(y^{\text{obs}}; D^{\text{chisq}}) &= \Pr\left\{ \left(U + \frac{\sigma}{\sqrt{n\sigma_0^2 + \tau^2}} \frac{\sqrt{n}(\bar{y}^{\text{obs}} - \theta_0)}{\sqrt{n\sigma_0^2 + \sigma^2 + \tau^2}} \right)^2 \leq \left(1 + \frac{\sigma^2}{n\sigma_0^2 + \tau^2} \right) V^2 \right\} \\ &= F_{1,1}\left(1 + \frac{\sigma^2}{n\sigma_0^2 + \tau^2}, \frac{\sigma^2}{n\sigma_0^2 + \tau^2} \frac{n(\bar{y}^{\text{obs}} - \theta_0)^2}{n\sigma_0^2 + \sigma^2 + \tau^2} \right), \end{aligned} \tag{11}$$

where U *and* V *are independent standard normal variables.*

The formula above gives some insight into to eppp. Note that eppp depends on τ, σ and σ_0 only through the ratios τ/σ and σ_0/σ in formula (11), so therefore without any loss generality we can assume σ to be 1. In the following we will discuss some special cases by varying the observation error (e.g., τ) and the prior assumption (e.g., σ_0) in different combinations, so as varying the number of observations. Figure 1 shows eppp as a function of y^{obs} plotted for different values of the error variables σ_0 and τ.

Small observation error: Assume τ is close to zero, which corresponds to small observation error. Then the model is close to a non-hierarchical model and the eppp$(y^{\text{obs}}; D^{\text{chisq}})$ becomes in the limit

$$\text{ppp}^{\text{compl}}(y^{\text{obs}}, x; D^{\text{chisq}}) = F_{1,1}\left(1 + \frac{\sigma^2}{n\sigma_0^2}, \frac{\sigma^2}{n\sigma_0^2}\frac{n(\bar{x}-\theta_0)^2}{n\sigma_0^2+\sigma^2}\right),$$

with $x = y^{\text{obs}}$. This corresponds to a situation described in Hjort et al. (2006). For the combination sharp prior and small observation error (e.g., both $\sigma_0 \to 0$ and $\tau \to 0$), eppp goes to $\Pr\{n(\bar{y}^{\text{obs}} - \theta_0)/\sigma^2 \geq V^2\} = 2[1 - \Phi(\sqrt{n}|\bar{y}^{\text{obs}} - \theta_0|/\sigma)]$, which is the classic p-value for testing the hypothesis $\theta = \theta_0$ in the model $y_i = x_i \sim N(\theta, \sigma^2)$.

Large observation error: If the observation error is large, eppp$(\bar{y}^{\text{obs}}; D^{\text{chisq}})$ goes in the limit to $1/2$ irrespectively of both the observed data and the value of the σ_0. Even if the prior for θ is tight (small σ_0's), eppp is $1/2$ in its limit when τ is large (which again is similar to the behaviour of the classic p-value). The right plot of Figure 1 shows that the curve of eppp gets flatter toward $1/2$ as τ increases from 0.1 to 5. The small difference between τ equals to 1 and 0.1 indicates that, as long as the observation error τ is moderate (relative to σ), the loss of accuracy is negligible for observing x indirectly through y. Naturally, large τ makes it difficult to say anything about the latent processes and the hyper parameter.

Vague priors or large n, or both: Imagine now the prior is flat relative to the amount of information about data, in the sense that $\sqrt{n}\sigma_0$ is large (either $\sqrt{n}$ is large, or σ_0 is large, or both). Then eppp$(\bar{y}^{\text{obs}}; D^{\text{chisq}}) \to \Pr\{U^2 \geq V^2\} = 1/2$ independently of the data. Many different models may arise under vague priors, making it harder to perform model validation. A more bizarre consequence, is that we are almost certain to keep our model when the amount of data is large. In this case, the knowledge about θ is mainly based on data compared to the prior information. When σ_0 is large, there is also a priori relatively little information about θ. The dashed line in the left plot of Figure 1 visualises that eppp plotted as a function of $\bar{y}^{\text{obs}}$ is close to a $1/2$ for $\sigma_0 = 5$.

The limit considerations above (e.g., $\sigma_0 \to \infty$ or $\tau \to \infty$) hold even when taking the variability of $\bar{y}^{\text{obs}}$ into account as the distribution of $n(\bar{y}^{\text{obs}} - \theta)^2/(n\sigma_0^2 + \sigma^2 + \tau^2)$ in equation (11) is χ_1^2. The characteristic of eppp that is worth taking notice of, is that

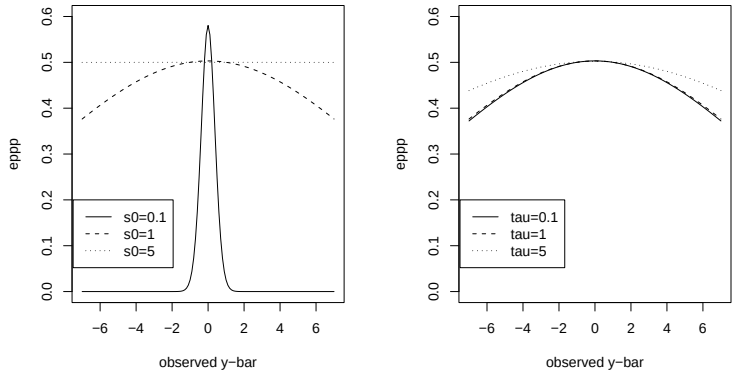

Fig. 1. The eppp$(\bar{y}^{\text{obs}}; D^{\text{chisq}})$ as a function of $\bar{y}^{\text{obs}}$ for the simple normal model $\mathcal{M}_1$ for $n = 50$, $\sigma = 1$ and $\theta_0 = 0$ varying the different levels of error processes of σ_0 (left plot) and τ (right plot). The remaining error parameter are fixed to 1.

eppp going to $1/2$ for flat priors and large observation error is indeed reasonable, while the characteristic of being close to a $1/2$ as the amount of data accumulates is not attractive.

Under the null-null model assumptions, $\sqrt{n}(\bar{y}^{\text{obs}} - \theta_0)/\sqrt{n\sigma_0^2 + \sigma^2 + \tau^2}$ is standard normal distributed. Hence, the null-null distribution of eppp$(Y; D^{\text{chisq}})$, formally defined in equation (7), is equal in distribution to

$$\text{eppp}(Y; D^{\text{chisq}}) \quad = \quad F_{1,1}\left(1 + \frac{\sigma^2}{n\sigma_0^2 + \tau^2}, \frac{\sigma^2}{n\sigma_0^2 + \tau^2} W^2\right). \tag{12}$$

The middle plot in Figure 1 shows a histogram of eppp$(Y; D^{\text{chisq}})$ for $n = 50$, $\theta_0 = 0$ and all variances equal 1 (e.g., $\sigma_0 = \sigma = \tau = 1$), illustrating that the distribution can be far from uniform. Only for both $\sqrt{n}\sigma_0$ and τ quite small does the eppp distribution come close to the uniform one on the unit interval, as eppp is the classic p-value in it's limit.

3.1.2. Marginalisation of D^{chisq}

Using $D(y, x, \theta) = D^{\text{chisq}}(x, \theta)$ in our simple normal-model with known variances, (5) gives

$$D^{\text{mrg}}(y, \theta) = \frac{\tau^2}{\tau^2 + \sigma^2} + \frac{n\sigma^2(\bar{y} - \theta)^2}{(\tau^2 + \sigma^2)^2}$$

which is a monotone increasing function of $|\bar{y} - \theta|$. Any discrepancy measure of this form will have the same ppp value. By similar techniques as in proof of Proposition 1, we get

$$\begin{aligned} \text{ppp}(y^{\text{obs}}; D^{\text{mrg}}) &= \Pr\left\{ \left(U + \frac{\sqrt{\sigma^2+\tau^2}}{\sqrt{n}\sigma_0} \frac{\sqrt{n}(\bar{y}^{\text{obs}} - \theta_0)}{\sqrt{n\sigma_0^2+\sigma^2+\tau^2}} \right)^2 \leq \left(1 + \frac{\sigma^2+\tau^2}{n\sigma_0^2}\right) V^2 \right\} \\ &= F_{1,1}\left\{ 1 + \frac{\sigma^2+\tau^2}{n\sigma_0^2}, \frac{\sigma^2+\tau^2}{n\sigma_0^2} \frac{n(\bar{y}^{\text{obs}} - \theta_0)^2}{n\sigma_0^2+\sigma^2+\tau^2} \right\}, \end{aligned}$$

for the two independent standard normal variables U and V. In particular, $\text{ppp}(y^{\text{obs}}; D^{\text{mrg}})$ is exactly the same as $\text{ppp}(y^{\text{obs}}; n(\bar{y} - \theta)^2/(\sigma^2 + \tau^2))$ which, with $y_i \sim N(\theta, \sigma^2 + \tau^2)$, was discussed in Hjort et al. (2006). We simply "ignore" the x-process and view our model as non-hierarchical. Note that in a more general non-normal model marginalising at the latent process is much more difficult.

The formula of $\text{ppp}(y^{\text{obs}}; D^{\text{mrg}})$ shows immediately that $\text{ppp}(y^{\text{obs}}; D^{\text{mrg}})$ collapses to $\text{ppp}^{\text{compl}}(y^{\text{obs}}, x; D^{\text{chisq}})$ as τ goes to zero. Similar to $\text{eppp}(y^{\text{obs}}; D^{\text{chisq}})$, $\text{ppp}(y^{\text{obs}}; D^{\text{mrg}})$ goes to a $1/2$ as the prior error increases, the data information accumulates and the observation error increases. For small $\sqrt{n}\sigma_0$, $\text{ppp}(y^{\text{obs}}; D^{\text{mrg}})$ is $\Pr\{V^2 \geq n(\bar{y}^{\text{obs}} - \theta_0)^2/(\sigma^2+\tau^2)\}$ in the limit. Again, this is the classic p-value for testing the hypothesis $\theta = \theta_0$ for $y_i \sim N(\theta, \sigma^2 + \tau^2)$. The $\text{ppp}(\bar{y}^{\text{obs}}; D^{\text{mrg}})$ and $\text{eppp}(\bar{y}^{\text{obs}}; D^{\text{chisq}})$, along with the calibrated versions $\text{ceppp}(\bar{y}^{\text{obs}}; D^{\text{chisq}})$ and $\text{cppp}^{\text{mrg}}(\bar{y}^{\text{obs}}; D^{\text{chisq}})$ treated below, are visualised as a function of $\bar{y}^{\text{obs}}$ in the left panel of Figure 2, with $\tau = \sigma = \sigma_0 = 1$, $n = 50$ and $\theta_0 = 0$. We see that eppp is more often close to $1/2$ than ppp^{mrg}, but that both the uncalibrated p-values require extreme observations before obtaining small p-values.

The (null-null) density of the $\text{ppp}(Y; D^{\text{mrg}})$, displayed through simulations of

$$\text{ppp}(Y; D^{\text{mrg}}) = F_{1,1}\left(1 + \frac{\sigma^2+\tau^2}{n\sigma_0^2}, \frac{\sigma^2+\tau^2}{n} W^2\right)$$

for $W \sim N(0,1)$, is visualised in Figure 2 together with $\text{eppp}(Y; D^{\text{chisq}})$ given in equation (12) for $\sigma_0 = 1$ and $\tau = 1$. We see that the distribution of $\text{ppp}(Y; D^{\text{mrg}})$ is more spread out than the density of $\text{eppp}(Y; D^{\text{chisq}})$, although both are far from uniform and heavily concentrated around $1/2$. As a consequence of large observation error or choosing a wide prior, the width of the density of $\text{eppp}(Y; D^{\text{chisq}})$ and $\text{ppp}(Y; D^{\text{mrg}})$ gets smaller.

3.1.3. The calibrated eppp *with* D^{chisq}

The above computations illustrate the difficulties of judging the significance of a number computed by $\text{eppp}(y^{\text{obs}})$. The following proposition gives the calibrated eppp given in

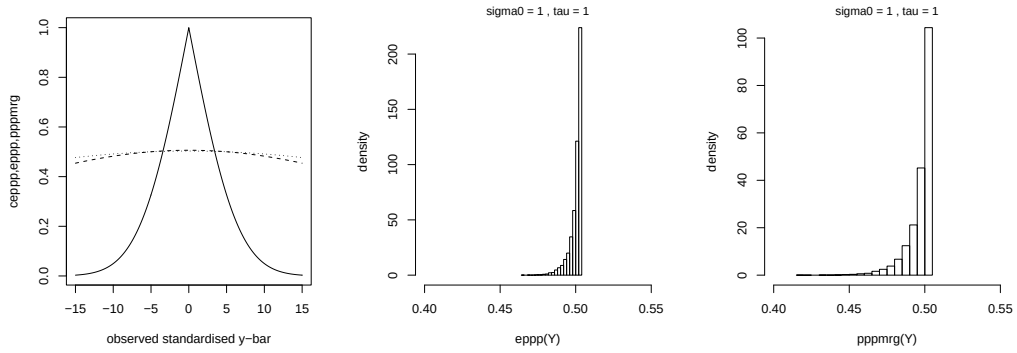

Fig. 2. Left panel: The ceppp and cppp$^{\mathrm{mrg}}$ (solid), eppp (dotted) and ppp$^{\mathrm{mrg}}$ (dashed) as a function of $\bar{y}^{\mathrm{obs}}$ for the normal model $\mathcal{M}_1$ for $\theta_0 = 0$, $\sigma_0 = \sigma = \tau = 1$ and $n = 50$. Middle: Normalised histogram of eppp$(Y; D^{\mathrm{chisq}})$. Right panel: Normalised histogram of ppp$(Y; D^{\mathrm{mrg}})$.

equation (8) for the simple normal model with the discrepancy D^{chisq}, transforming the eppp to a proper p-value. This calibrated p-value with an underlying probability scale, makes it possible to compare different combinations of prior and models computed for the same data set.

Proposition 2 *The calibrated* eppp *for the normal model $\mathcal{M}_1$ with discrepancy D^{chisq} defined in equation* (10) *can be expressed by*

$$\mathrm{ceppp}(y^{\mathrm{obs}}; D^{\mathrm{chisq}}) \quad = \quad \Pr\left\{U^2 \geq \frac{n(\bar{y}^{\mathrm{obs}} - \theta_0)^2}{n\sigma_0^2 + \sigma^2 + \tau^2}\right\}.$$

for $U \sim N(0,1)$.

When $\sigma_0 \to 0$, ceppp$(y^{\mathrm{obs}}; D^{\mathrm{chisq}})$ converges to $2[1 - \Phi(\sqrt{n}|\bar{y}^{\mathrm{obs}} - \theta_0|/\sqrt{\sigma^2 + \tau^2})]$ the p-value for the classical two-sided test of the hypothesis $H_0 : \theta = \theta_0$. In general, ceppp takes into account the extra uncertainty about θ in the model through the extra term $n\sigma_0^2$ in the denominator. From the formula it is also easy to see the effect on x only being indirectly observed through y in that the uncertainties about $x|\theta$ and $y|x$ are additive. The ceppp-formula in Proposition 2 is also useful for power computations, as we illustrate in the next paragraph. Figure 2 (left panel) illustrates ceppp$(y^{\mathrm{obs}}; D^{\mathrm{chisq}})$ curves, along with eppp$(y^{\mathrm{obs}}; D^{\mathrm{chisq}})$ and ppp$(y^{\mathrm{obs}}; D^{\mathrm{mrg}})$, as a function of y^{obs} for $n = 50$, $\theta_0 = 0$ and all variances equal 1 (e.g., $\sigma_0 = \sigma = \tau = 1$).

Since both eppp($y^{\text{obs}}; D^{\text{chisq}}$) and ppp($Y; D^{\text{mrg}}$) are decreasing functions of $(\bar{y}^{\text{obs}} - \theta_0)^2$, their calibrated versions become equivalent, indicating that we do not loose anything by using the simpler and computationally more efficient ceppp. The similarity between ceppp($y^{\text{obs}}; D^{\text{chisq}}$) and cppp($y^{\text{obs}}; D^{\text{mrg}}$) is however not a general rule.

3.1.4. The distribution of p-values under some alternatives using D^{chisq}

The eppp and model checking in general, is often intended at early stages of an analysis where no alternative models have yet been considered. The aim of this early stage exploration is to investigate if the data are likely to have arisen from the suggested model and prior. However, it is useful to study the behaviour of eppp, ppp^{mrg} and their calibrated versions when the null-null model is wrong. We do so by computing Pow(eppp) = $\Pr\{\text{eppp}(Y^{\text{alt}}) \leq \alpha\}$ where Y^{alt} is generated under an alternative model, and similar for the other p-values. In the simple normal example, the alternative model is chosen to have some violation on the location parameter θ, e.g., $\theta^{\text{alt}} \sim N(\theta_0 + \psi_0, \sigma_0^2)$ for some parameter ψ_0 on the real scale. The conditional models for the system process x and the observation process y are remained unchanged. This alternative model is motivated from the discrepancy measure D^{chisq}'s capacity to detect abnormality from the assumptions on θ.

In the following, let $q(\alpha; a)$ be the non-centrality parameter that makes the cumulative distribution $F_{1,1}(1 + a, q(\alpha; a)) = \alpha$. Inserting the marginal distribution of $\overline{Y}^{\text{alt}}$ into the formula of eppp in (11), or equivalently, imposing $(\overline{Y}^{\text{alt}} - \theta_0) = \psi_0 + \sqrt{\sigma_0^2 + \sigma^2/n + \tau^2/n} W$ by $(\bar{y}^{\text{obs}} - \theta_0)$ in (11) where W is standard normal, we get

$$\text{Pow}(\text{eppp}) = \Pr\left\{ \left(W + \frac{\sqrt{n}\psi_0}{\sqrt{n\sigma_0^2 + \sigma^2 + \tau^2}} \right)^2 \geq q\left(\alpha; \frac{\sigma^2}{n\sigma_0^2 + \tau^2} \right) \frac{n\sigma_0^2 + \tau^2}{\sigma^2} \right\}.$$

The formula follows by applying a proof similar to the proof of Proposition 2. Similarly, for ceppp and ppp^{mrg} we may express

$$\text{Pow}(\text{ppp}^{\text{mrg}}) = \Pr\left\{ \left(W + \frac{\sqrt{n}\psi_0}{\sqrt{n\sigma_0^2 + \sigma^2 + \tau^2}} \right)^2 \geq q\left(\alpha; \frac{\sigma^2 + \tau^2}{n\sigma_0^2} \right) \frac{n\sigma_0^2}{\sigma^2 + \tau^2} \right\}$$

and

$$\text{Pow}(\text{ceppp}) = \Pr\left\{ \left(W + \frac{\sqrt{n}\psi_0}{\sqrt{n\sigma_0^2 + \sigma^2 + \tau^2}} \right)^2 \geq z_{\alpha/2}^2 \right\},$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile for the standard normal distribution. The power of cppp^{mrg} is equal to the power of ceppp. The power formulae for the uncalibrated p-values, show that both Pow(eppp) and Pow(ppp^{mrg}) are zero when τ goes to infinity (because $\lim_{a\to\infty} q(\alpha, a) = \infty$). As expected, the probability of detecting wrong models is small for large observation error. But unlike eppp and ppp^{mrg}, Pow(ceppp) is close to the significance level α as the observation variance increases. Figure 3 shows the power of the tests as a function of the alternative parameter $\theta^{\text{alt}} = \theta_0 + \psi_0$ for different values of τ and σ_0. For all p-values, the chance of rejecting the model increases with the distance from $\psi_0 = 0$. The shapes of the curves are similar for each σ_0 value, with the probability of rejecting models decreasing for higher values of σ_0, illustrating that D^{chisq} is testing prior specifications. Pay particular attention to that wide priors give smaller probability of detecting wrong models no matter the size of observation error, see the right column of Figure 3. This is logical and a highly desired characteristic, since a wide prior could give arise to many possible, but different combinations of models. The power curves for eppp (dashed line) are conservative, where the probability of detecting wrong models are zero in almost all the plots, except for the sharp prior situation. In all cases, ceppp has higher power than the other variants, even much higher for large prior variance. For small σ_0, i.e. high confidence in the prior, the power of ceppp is influenced by the relative size of τ compared to σ. However, for higher σ_0 values, the effects of both σ and τ are small, giving almost identical power functions for all the given τ values. The ppp^{mrg} is doing a bit better than eppp for $\sigma_0 = 0.1$ and $\tau = 1$, otherwise they are similar.

3.2. Underlying regression process with binormal prior

For the second example model $\mathcal{M}_2$ listed in Table 1, the data y_i is again normal with mean x_i and variance τ^2 given the x's. Conditional on the prior parameters, the second stage process x_i is $N(\beta_1+\beta_2 z_i, \sigma^2)$ for z_i known covariates. The unknown hyper parameter $\beta = (\beta_1, \beta_2)$ is binormal with mean $\beta_0 = (\beta_{0,1}, \beta_{0,2})$ and covariance matrix $\sigma^2 M_0^{-1}$ with $M_0 = \text{diag}(\kappa_1, \kappa_2)$. We study the simple situation where all the (co)variances are known. In standard matrix notation, $x = Z\beta + \varepsilon$ pretending the x's are observed where $\varepsilon \sim N(0, \sigma^2 I_n)$, with the least squares estimator $\hat{\beta} = (Z'Z)^{-1}Z'x$ for the covariate matrix Z with rows $(1, z_i)$. We will further assume that the covariates are centred so that $\bar{z} = 0$.

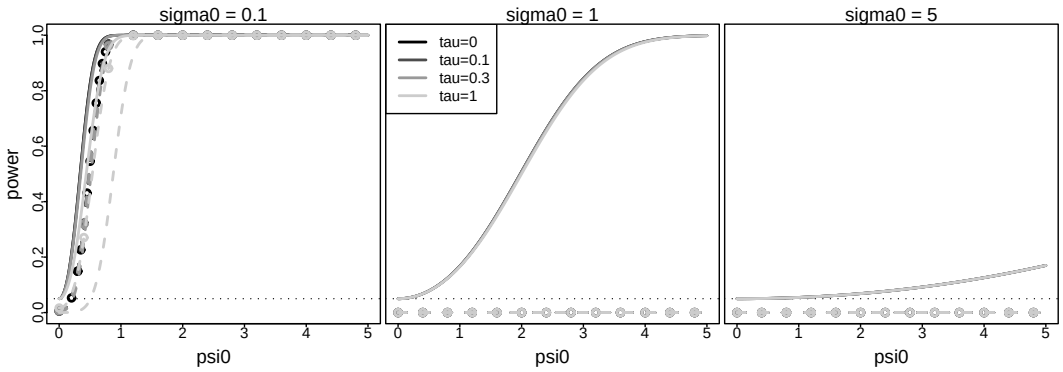

Fig. 3. Power functions where rejecting for p-values $\leq$ 0.05, with the p-values ceppp (solid), eppp (dashed) and ppp$^{\text{mrg}}$ (dashed-circle) using D^{chisq} under the alternatives $\theta^{\text{alt}} \sim N(\psi_0 + \theta_0, \sigma_0^2)$ for the simple normal model $\mathcal{M}_1$. Each plot varies with σ_0 and τ where $n = 50$, $\theta_0 = 0$ and $\sigma = 1$ are fixed. The horizontal dotted line is the significance level of 0.05. Note that the curves of eppp are over-lined by ppp$^{\text{mrg}}$ in most of the plots.

Here we study the discrepancy measure

$$D^{\beta_2}(x) = (\hat{\beta}_2(x) - \beta_{0,2})^2 \text{ where } \hat{\beta}_2(x) = \frac{z'x}{s_z^2} \text{ and } s_z^2 = \sum_i z_i^2, \tag{13}$$

i.e. $\hat{\beta}_2(x)$ is the least square estimate of β_2 if we had observed the x's.

3.2.1. The eppp *with* D^{β_2}

The following results present formulae for the ppp$^{\text{compl}}$ and eppp using the discrepancy measure (13).

Proposition 3 *The complete posterior predictive p-value for the regression model* $\mathcal{M}_2$ *can be written as*

$$\text{ppp}^{\text{compl}}(y^{\text{obs}}, x; D^{\beta_2}) = \Pr\left\{ \left[a(z) \tfrac{s_z(\hat{\beta}_2(x) - \beta_{0,2})}{\sigma} + b(z)U \right]^2 \geq \left[\tfrac{s_z(\hat{\beta}_2(x) - \beta_{0,2})}{\sigma} \right]^2 \right\} \tag{14}$$

for U *standard normal, where*

$$a(z) = \frac{s_z^2}{\kappa_2 + s_z^2} \quad \textit{and} \quad b(z) = \sqrt{\frac{\kappa_2 + 2s_z^2}{\kappa_2 + s_z^2}}. \tag{15}$$

Following from Proposition 3 is the two limit cases of interest as we now present.

Corollary 1

$$\text{ppp}^{\text{compl}}(y^{\text{obs}}, x; D^{\beta_2}) \approx \begin{cases} 2\left[1 - \Phi\left(\frac{s_z|\hat{\beta}_2(x)-\beta_{0,2}|}{\sigma}\right)\right] & \textit{for } \kappa_2 \textit{ large or } s_z \textit{ small,} \\ \frac{1}{2} + \Phi\left(-\sqrt{2}\frac{s_z|\hat{\beta}_2(x)-\beta_{0,2}|}{\sigma}\right) & \textit{for } \kappa_2 \textit{ small or } s_z \textit{ large.} \end{cases}$$

When the prior knowledge about β_2 is reasonably sharp compared with the amount of information from data, the $\text{ppp}^{\text{compl}}$ is similar to the classical p-value for testing $H_0 : \beta_2 = \beta_{0,2}$ using a traditional test statistic. On the other hand, when κ_2 is small or s_z is large, both cases corresponding to relatively little information about β_2, we obtain a large p-value. This is not a preferable result, but is in correspondence with the large n situation for model $\mathcal{M}_1$.

Proposition 4 *The expected posterior predictive p-value for the regression model $\mathcal{M}_2$ can be expressed by*

$$\text{eppp}(y^{\text{obs}}; D^{\beta_2}) = \Pr\{[a(z)V + b(z)U]^2 \geq V^2\} \tag{16}$$

for $a(z)$ and $b(z)$ given in Proposition 3, and V being normal, independent of U, with expectation and variance

$$\mu_v = \frac{\sigma^2(\kappa_2 + s_z^2)}{\sigma^2(\kappa_2 + s_z^2) + \kappa_2\tau^2} \frac{s_z(\hat{\beta}_2(y^{\text{obs}}) - \beta_{0,2})}{\sigma} \quad \textit{and} \quad \sigma_v^2 = \frac{\tau^2(\kappa_2 + s_z^2)}{\sigma^2(\kappa_2 + s_z^2) + \kappa_2\tau^2}. \tag{17}$$

Note that in this example, we may write $y = Zb + \tilde{\varepsilon}$ where $\tilde{\varepsilon}$ is $N(0, (\sigma^2 + \tau^2)I)$, which is of the same form as the regression model based on x observed. Therefore, we could in this case have used the discrepancy measure $\widetilde{D}^{\beta_2}(y) = (\hat{\beta}_2(y) - \beta_{0,2})^2$ directly and obtained

$$\text{ppp}(y^{\text{obs}}; \widetilde{D}^{\beta_2}) = \Pr\{[a(z)(\hat{\beta}_2(y^{\text{obs}}) - \beta_{0,2}) + \sqrt{\sigma^2 + \tau^2}b(z)U]^2 \geq [\hat{\beta}_2(y^{\text{obs}}) - \beta_{0,2}]^2 | y^{\text{obs}}\}.$$

In more general cases, the model for y can not be written in such a simple model of the hyper parameters alone, and the use of ppp directly will not be that easy. Both $\text{eppp}(y^{\text{obs}}; D^{\beta_2})$ and $\text{ppp}(y^{\text{obs}}; \widetilde{D}^{\beta_2})$ are decreasing functions of $|\hat{\beta}_2(y) - \beta_{0,2}|$, but will in general be slightly different.

Also in this case, we can look at the properties of $\text{eppp}(y^{\text{obs}}; D^{\beta_2})$ under different special cases.

Small observation error. When $\tau^2 \to 0$, $\text{eppp}(y^{\text{obs}}; D^{\beta_2})$ approaches $\text{ppp}^{\text{compl}}(y^{\text{obs}}, x; D^{\beta_2})$ with $x = y^{\text{obs}}$ and the properties of $\text{eppp}(y^{\text{obs}}; D^{\beta_2})$ will follow the properties of $\text{ppp}^{\text{compl}}(y^{\text{obs}}, y^{\text{obs}}; D^{\beta_2})$

Large observation error. When $\tau^2 \to \infty$, $\mu_v \to 0$ and $\sigma_v^2 \to (\kappa_2 + s_z^2)/s_z^2$, $\text{eppp}(y^{\text{obs}}; D^{\beta_2})$ becomes independent of y^{obs}. This is a reasonable result. However, $\text{eppp}(y^{\text{obs}}; D^{\beta_2})$ depends heavily on the ratio κ_2/s_z^2, ranging from 0.70 (for $\kappa_2/s_z^2 = 0$) to 0 (for $\kappa_2/s_z^2 = \infty$), showing that eppp is extremely difficult to interpret for large measurement error.

Vague prior information and/or large s_z^2. Note that since the integrand in equation (16) is finite, we may switch integral and limit sign and we obtain again that $\text{eppp}(y^{\text{obs}}; D^{\beta_2})$ is approximately 1/2.

Informative prior. We have

$$\lim_{\kappa_2 \to \infty} \text{eppp}(y^{\text{obs}}; D^{\beta_2}) = \int_u 2\left[1 - \Phi\left(\left|\frac{s_z(\hat{\beta}_2(y^{\text{obs}}) - \beta_{0,2})}{\sigma} + \sqrt{\frac{\tau^2}{\sigma^2+\tau^2}}u\right|\right)\right]\phi(u)\mathrm{d}u.$$

Note that the classical p-value in this case would be $[1-\Phi(|(\hat{\beta}_2(y^{\text{obs}})-\beta_{0,2})/(\sqrt{\sigma^2+\tau^2}/s_z)|)]$. The use of σ in eppp instead of $\sqrt{\sigma^2+\tau^2}$ would (for large $|\hat{\beta}_2(y^{\text{obs}}) - \beta_{0,2}|$ as the interesting cases) decrease the p-value. On the other hand, the integration would push the p-value towards 1/2. These two factors combined indicate that eppp in this case will be quite similar to the classical p-value.

In Figure 4 (upper part), the distribution of $\text{eppp}(y^{\text{obs}}; D^{\beta_2})$ under the true model (obtained trough simulations) is plotted for different values of κ_2 and τ. Only for high κ_2 and small τ, the distribution is close to uniform, while for other parameter sets the distribution is concentrated more around 0.5. This again indicates that interpretation of posterior predictive p-values is difficult.

3.2.2. Marginalisation of D^{β_2}

Here we study the marginalised discrepancy measure (5) for D^{β_2} given in equation (13). From the proof of Proposition 4, we get

$$\begin{aligned} D^{\text{mrg}}(y) &= \text{var}[\hat{\beta}_2(x)|y] + \text{E}[(\hat{\beta}_2(x) - \beta_{0,2})^2|y] \\ &= c_1 + c_2 \frac{(\hat{\beta}_2(y) - \beta_{0,2})^2}{(\sigma^2+\tau^2)/s_z^2}, \end{aligned}$$

for some constants c_1 and c_2 depending on $\kappa_2, \sigma^2, \tau^2$ and z. Consequently, the $\text{ppp}(y^{\text{obs}}; D^{\text{mrg}})$ is the equivalent with computing

$$\Pr\left\{\left(\frac{\hat{\beta}_2(y^{\text{rep}}) - \beta_{0,2}}{\sqrt{\sigma^2+\tau^2}/s_z}\right)^2 \geq \left(\frac{\hat{\beta}_2(y^{\text{obs}}) - \beta_{0,2}}{\sqrt{\sigma^2+\tau^2}/s_z}\right)^2 | y^{\text{obs}}\right\}. \tag{18}$$

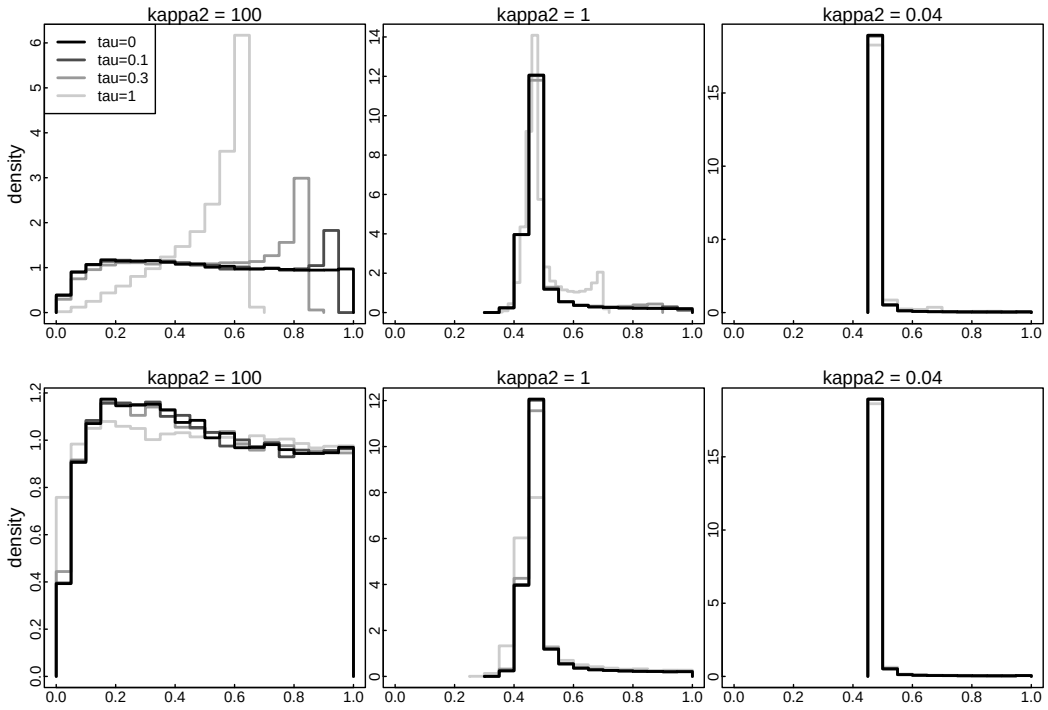

Fig. 4. Normalised histogram of $\text{eppp}(Y; D^{\beta_2})$ (upper row) and $\text{ppp}^{\text{mrg}}(Y; D^{\beta_2})$ (lower row) for the regression model $\mathcal{M}_2$ for different values τ with $\beta_{0,2} = 0$, $n = 50$, $z_i = (i - 25.5)/\sqrt{212.5})$, $\sigma = 1$ and for $\kappa_2 = 100$ (left), $\kappa_2 = 1$ (middle) $\kappa_2 = 0.04$ (right).

This leads to the following proposition.

Proposition 5 *The posterior predictive p-value for the regression model $\mathcal{M}_2$ using the marginalised discrepancy measure of D^{β_2} is given by*

$$\text{ppp}(y^{\text{obs}}, D^{\text{mrg}}) = \Pr\left\{\left[a^{\text{mrg}}(z)\tfrac{s_z(\hat{\beta}_2(y^{\text{obs}})-\beta_{0,2})}{\sqrt{\sigma^2+\tau^2}} + b^{\text{mrg}}(z)U\right]^2 \geq \left[\tfrac{s_z(\hat{\beta}_2(y^{\text{obs}})-\beta_{0,2})}{\sqrt{\sigma^2+\tau^2}}\right]^2\right\},$$

where

$$a^{\text{mrg}}(z) = \frac{s_z^2\sigma^2}{\kappa_2(\sigma^2+\tau^2)+s_z^2\sigma^2} \;\; \textit{and} \;\; b^{\text{mrg}}(z) = \sqrt{\frac{\kappa_2(\sigma^2+\tau^2)+2s_z^2\sigma^2}{\kappa_2(\sigma^2+\tau^2)+s_z^2\sigma^2}}.$$

The following corollary demonstrates that the theoretical properties of $\text{ppp}(y^{\text{obs}}; D^{\text{mrg}})$ are similar to those of $\text{eppp}(y^{\text{obs}}; D^{\beta_2})$.

Corollary 2

$$\mathrm{ppp}(y^{\mathrm{obs}};D^{\mathrm{mrg}}) \approx \begin{cases} \mathrm{ppp}^{\mathrm{compl}}(y^{\mathrm{obs}},y^{\mathrm{obs}};D^{\beta_2}) & \textit{for } \tau \textit{ small,} \\ \frac{1}{2} & \textit{for } \kappa_2 \textit{ small or } s_z \textit{ large or } \tau \textit{ large,} \\ 2\left[1-\Phi\left(\frac{s_z|\hat{\beta}_2(y^{\mathrm{obs}})-\beta_{0,2}|}{\sqrt{\sigma^2+\tau^2}}\right)\right] & \textit{for } \kappa_2 \textit{ large or } s_z \textit{ small.} \end{cases}$$

The proof of this corollary is trivial given Proposition 5, and thus, its proof is omitted. In Figure 4 (lower part), the distribution of $\mathrm{ppp}(y^{\mathrm{obs}};D^{\mathrm{mrg}})$ under the true model (obtained trough simulations) is plotted for different values of κ_2 and τ. Although $\mathrm{ppp}(y^{\mathrm{obs}};D^{\mathrm{mrg}})$ seems to have a distribution closer to the uniform distribution for a larger range of τ values, the distributions are far from uniform for many parameter sets.

3.2.3. The calibrated eppp *and* $\mathrm{ppp}^{\mathrm{mrg}}$ *with* D^{β_2}

Both $\mathrm{eppp}(y^{\mathrm{obs}};D^{\beta_2})$ and $\mathrm{ppp}(y^{\mathrm{obs}};D^{\mathrm{mrg}})$ are decreasing functions of $(\hat{\beta}_2(y^{\mathrm{obs}})-\beta_{0,2})^2$. This implies that their calibrated versions will be identical. Further, since $s_z(\hat{\beta}_2(Y)-\beta_{0,2})/\sqrt{\sigma^2+\tau^2+s_z^2\kappa_2^{-1}\sigma^2}$ is standard normal when Y is generated from the assumed model, we obtain

$$\mathrm{ceppp}(y^{\mathrm{obs}};D^{\beta_2}) = 2\left[1-\Phi\left(\frac{s_z(\hat{\beta}_2(y^{\mathrm{obs}})-\beta_{0,2})}{\sqrt{\sigma^2+\tau^2+s_z^2\kappa_2^{-1}\sigma^2}}\right)\right].$$

We see similar characteristics as for model $\mathcal{M}_1$. When $\kappa_2\to\infty$, $\mathrm{ceppp}(y^{\mathrm{obs}};D^{\beta_2})$ converges to $2[1-\Phi(s_z(\hat{\beta}_2(y^{\mathrm{obs}})-\beta_{0,2})/\sqrt{\sigma^2+\tau^2})]$, which is the p-value for the classical two-sided test of the hypothesis $H_0:\beta_2=\beta_{0,2}$. The uncertainty concerning the value of β_2 in the model is accounted for by the extra term $s_z^2\kappa_2^{-1}\sigma^2$ in the denominator. Again the effect of observation error is additive to the uncertainty of the latent variables x.

3.2.4. A simulation study of power detection using D^{β_2}

Under the assumed model we use $\beta_{0,2}=0$. Here we study eppp under the alternatives when $\beta_{0,2}\neq 0$ for which the observation process Y^{alt} is generated from the same model as before except for that the expectation is changed from $Z\beta_0$ to $Z\beta_0^{\mathrm{alt}}$ with $\beta_0^{\mathrm{alt}}=(\beta_{0,1},\beta_{0,2}^{\mathrm{alt}})$. We approximate $\Pr\{\mathrm{eppp}(Y^{\mathrm{alt}};D^{\beta_2})\leq\alpha\}$ by $\sum_b I\{\mathrm{eppp}(Y^{\mathrm{alt}};D^{\beta_2})\leq\alpha\}/B$ for a high number B of simulated Y^{alt}'s. As for the eppp, we also investigate the capability of ceppp,

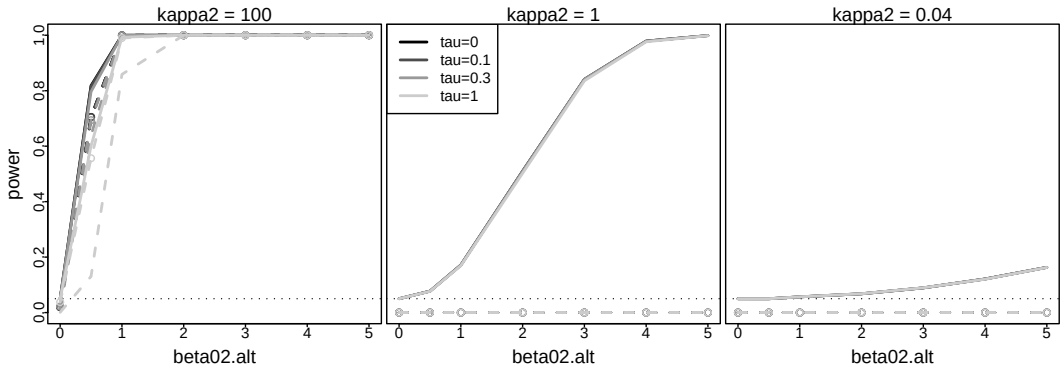

Fig. 5. Power functions where rejecting for p-values ≤ 0.05, with the p-values ceppp$(Y; D^{\beta_2})$ (solid), eppp$(Y; D^{\beta_2})$ (dashed), ppp$^{\text{mrg}}(Y; D^{\beta_2})$ (dashed-circle) for the regression model $\mathcal{M}_2$ using discrepancy measure D^{β_2} under some alternatives $\beta_{0,2}^{\text{alt}} = (0, 1, 3, 5)$. Each plot varies with κ_2 and τ, where $n = 50$ and $\sigma = 1$. The horizontal dotted line is the significance level of 0.05.

ppp$^{\text{mrg}}$ and cppp$^{\text{mrg}}$ for detecting wrong models with Y^{alt} generated from the alternative model, see results in Figure 5. The power curves of cppp$^{\text{mrg}}$ are represented by ceppp since these are identically. With $\beta_{0,2}^{\text{alt}} = 0$, the probability that the calibrated p-values detect wrong models is equal to our significance level of 5%. As for D^{chisq} in section 3.1, the probability of rejecting the null-null hypothesis under wrong models is quite small for wide priors regardless of the size of the observation error, since D^{β_2} is a discrepancy measure testing the prior specifications of β_2. Both eppp and ppp$^{\text{mrg}}$ are conservative, with eppp even more conservative for $\kappa_2 = 100$ in Figure 5. Although these uncalibrated p-values do better in the sharp prior situations, they never get close to $\alpha = 5\%$ when the null-null model is true. For the influence of observation error, we see a similar pattern as for model $\mathcal{M}_1$. For large κ_2, i.e. high confidence in the prior, the power of ceppp is influenced by the relative size of τ compared to σ. However, for low κ_2 values, corresponding to a wide prior, the effects of both σ and τ are small, giving almost identical power functions for all the given τ values.

4. Discrepancy measure testing latent assumptions

The focus in this section is to examine the adequacy of model specifications for the underlying x-process, within the same simple normal models $\mathcal{M}_1$ and $\mathcal{M}_2$ described in the

foregoing section, see summary in Table 1. Particularly, it is of interest to study how much the observation error complicates the validation of the underlying system in the context of eppp and ceppp, so is the influence of a priori knowledge. The ppp^{mrg} has shown to be slightly more uniform than eppp. Anyway, it has not proving to be superior in detecting wrong models, therefore we skip an illustration of ppp^{mrg} due to computational complications. In some situations ppp^{mrg} may however be simpler to compute than the eppp, see the entropy measures used in the multinomial model for age of fish in section 5.

The discrepancy used for checking the validity of the system process raised under the observed data is based on the Kolmogorov test statistic using standardised residuals. Model $\mathcal{M}_1$ has residuals $\varepsilon_i = \sigma^{-1}(x_i - \theta)$ for $i = 1, \ldots, n$, while for model $\mathcal{M}_2$ yields $\varepsilon_i = \sigma^{-1}(x_i - \beta_1 - \beta_2 z_i)$. We define

$$D^{\text{kolmo}}(\varepsilon) = \sup_t |F_n(t) - \Phi(t)| \tag{19}$$

for the empirical cumulative $F_n(t)$ of the residuals ε and the cumulative standard normal $\Phi(t)$. There is no analytical results, so eppp and ceppp are approximated according to equations (4) and (9) respectively. D^{kolmo} is computed by $\max_{i \leq n}(\Phi(\varepsilon_{(i)}) - (i-1)/n, i/n - \Phi(\varepsilon_{(i)}))$ for $\varepsilon_{(i)}$ the i'th largest of the ε's. Simulating the distribution of $\text{eppp}(y; D^{\text{kolmo}})$ under both models shows again that eppp is non-uniform under the model conditions, see upper row of the Figures 6 and 7, illustrating the difficulties to judge whether a computed eppp value is extreme or not.

Again, we investigate the eppp and ceppp's capability of detecting wrong models with an alternative change point model using a significance level of 5%. Simulating the power is computer intensive as it requires a triple-simulations scheme for each alternative model. We used a C-program with 500 and 50000 simulations for computing eppp and ceppp respectively, and finally 50000 replications for computing the power. For model $\mathcal{M}_1$ the alternatives are $x_i^{\text{alt}} \sim N(\theta, \sigma^2)$ for $i < 25$, and $x_i^{\text{alt}} \sim N(\theta + \psi, \sigma^2)$ for $i \geq 25$ for a real valued ψ, with the results presented in the lower row of Figure 6. Similar, results from the regression model $\mathcal{M}_2$ are given in Figure 7, with the alternatives $x_i^{\text{alt}} \sim N(\beta_1 + \beta_2 z_i, \sigma^2)$ for $i < 25$, and $x_i^{\text{alt}} \sim N(\beta_1 + (\beta_2 + \psi) z_i, \sigma^2)$ for $i \geq 25$. If the model is correct, the residuals should be identically distributed. The alternative change point model makes their distribution different. Note, however, that since the x's are predicted through the y's, they will not have the same distribution as in the usual situation when the x's are observed.

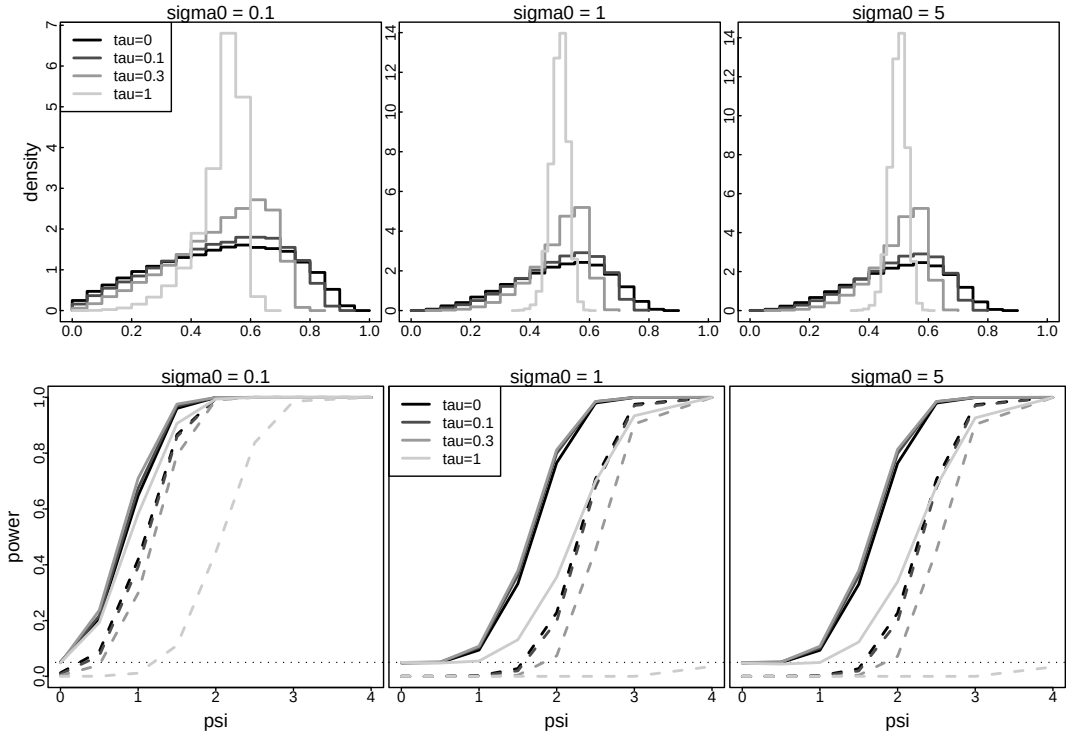

Fig. 6. Results from model $\mathcal{M}_1$ using D^{kolmo} for different values of τ and σ_0 with $\theta_0 = 0$, $\sigma = 1$ and $n = 50$. Upper panel: Normalised histogram of eppp$(Y; D^{\text{kolmo}})$. Lower row: Power functions where rejecting for p-values ≤ 0.05, for ceppp$(Y; D^{\text{kolmo}})$ (solid) and eppp$(Y; D^{\text{kolmo}})$ (dashed) where the alternative models are $x_i^{\text{alt}} \sim N(\theta, \sigma^2), i < 25$ and $x_i^{\text{alt}} \sim N(\theta + \psi, \sigma^2), i \geq 25$. The horizontal dotted line is the significance level of 0.05.

The results for model $\mathcal{M}_1$ and $\mathcal{M}_2$ follow the same trend. As expected, the calibrated p-value has more power than the uncalibrated one, similar to the results in section 3. Even when $\tau = 1$ and the power of eppp is almost zero under all alternatives, the corresponding calibrated version is able to catch lack of fit to the data, see the light grey line in the lower panel of Figures 6 and 7. The capability of detecting lack of fit to the data is better with a higly informative prior on the hyper parameters, while moderate to vague priors give similar power. This is in contrast to testing prior specifications in section 3, where decreasing the information in the prior resulted in flatter power-curves. The black curves corresponding to no observation error indicate the best we can do in a non-hierarchical situation with $x = y^{\text{obs}}$. Using D^{kolmo}, we do not observe the similar shapes across τ for a given value of

the prior deviation (e.g., σ_0 or κ_2), as was the case for the discrepancy measures constructed to test prior specifications. Especially, the distance between curves with $\tau = 1$ and the rest is notable for middle range and vague priors, see the plots with σ_0 equal 1 and 5 in Figure 6 and κ_2 equal 1 and 0.04 in Figure 7. This indicates that the effect of τ relative to σ is independent of the prior specifications.

These examples indicate that we can detect model failure regarding the underlying x by the naturally rescaled ceppp version. It should be noted that Sinharay & Stern (2003) used the measure $D(x) = |x_{\max} - x_{\text{med}}| - |x_{\min} - x_{\text{med}}|$ for testing the normal assumptions made on x, and computed the ppp^{mrg}. They indicated that detecting model failure of the underlying processes only can be detected if the violation is extreme or the observation error is small, and concluded that the data may well be described by the normal-normal model even when x's are not normally distributed, regarding this as a model robustness. However, unlike this article, their simulations schemes for computing data under the true and alternative model, are done for fixed hyper parameters, resulting in a different way of interpreting the null model.

5. A multinomial model for age of fish

Prediction of the number of fish caught in different age groups (catch at age) is of importance in the process of deciding fish quotas. The prediction is difficult due to that typically only total weight of a haul with neither age nor weight measurements available for individual fish. Samples of hauls are however selected where a limited number of fish are weighed and ages specified through otoliths. A large proportion of fish only have weights available and missing age-measurements can vary from being at random to stratifications by weight. An important ingredient in the prediction of catch at age is a description of the simultaneous (age,weight) distribution and how this change with explanatory variables.

Hirst et al. (2004, 2005) constructed a Bayesian hierarchical model for the distribution of age and weight of fish caught by commercial fisheries, applied to prediction of catch at age. For an illustration on the use of eppp values, we will here only consider age observations of cod from year 2002 (with no missing values) and the marginal model for ages. The age $A_{h,f}$ of fish f in haul h is assumed to follow a multinomial distribution with

$$\Pr(A_{h,f} = a) = \frac{\exp(x_{h,a})}{\sum_{a'} \exp(x_{h,a'})} \tag{20}$$

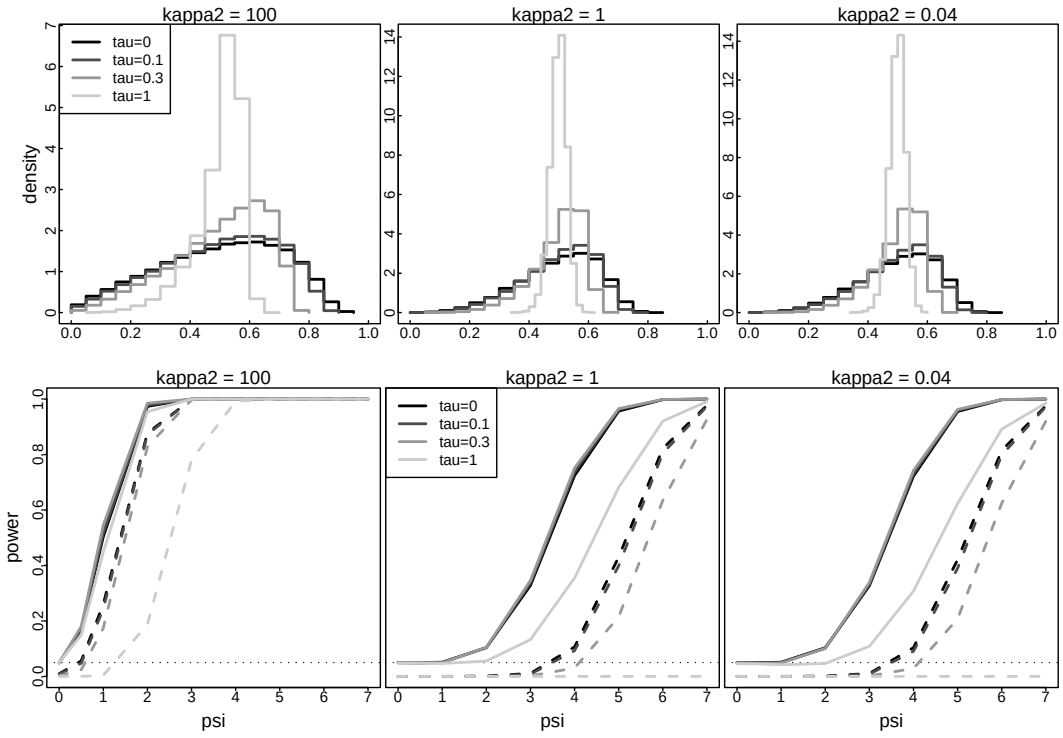

Fig. 7. Results from regression model $\mathcal{M}_2$ using D^{kolmo} for different values of τ and κ_2 with $\beta_{0,1} = 0$, $\sigma = 1$, $M_0 = I_n$ and $n = 50$. Upper panel: Normalised histogram of eppp$(Y; D^{\text{kolmo}})$. Lower row: Power functions where rejecting for p-values ≤ 0.05, for ceppp$(Y; D^{\text{kolmo}})$ (solid) and eppp$(Y; D^{\text{kolmo}})$ (dashed) where the alternative models are $x_i^{\text{alt}} \sim N(\beta_1 + \beta_2 z_i, \sigma^2)$, $i < 25$, and $x_i^{\text{alt}} \sim N(\beta_1 + (\beta_2 + \psi) z_i, \sigma^2)$, $i \geq 25$. Note that the curves of eppp are over-lined by ppp$^{\text{mrg}}$ in most of the plots.

where

$$x_{h,a} = \beta_a^{\text{const}} + \beta_{s(h),a}^{\text{seas}} + \beta_{g(h),a}^{\text{gear}} + \varepsilon_{h,a} \tag{21}$$

and where $, s(h), g(h)$ are the season and gear, respectively, for which haul h was caught. Here, $\{\beta_a^{\text{const}}, \beta_{a,s(h)}^{\text{seas}}, \beta_{a,g(h)}^{\text{gear}}\}$ are fixed effects (regression parameters corresponding to categorical covariates) while $\varepsilon_{a,h}$ is a random effect. Constraints are made on the β-parameters in order to make them identifiable (sum-constraints). We will follow Hirst et al. (2004, 2005), assuming the random process $\{\varepsilon_{h,a}\}$ to be independent Gaussian zero-mean with constant variance.

Data are also available from years $1995-2001$ and our priors are constructed from analysis of these data. In particular, all β parameters are assumed to have independent Gaussian distributions with mean and variances specified from the posterior of the $1995-2001$ data, while the variance of the random effect was assumed inverse gamma with parameters fitted from the posterior samples from previous years.

Our aim has been to check if the 2002 data is consistent with the prior and the model. In this simple situation, test statistics could be construct directly on the data and p-values could be calculated in a standard Box-type manner (Box 1980). The general aim is however to be able to perform checking in the full data/model complex where in particular missing ages complicate the problem of specifying suitable test statistics. Differences in sampling efforts between hauls complicates the matter further.

The latent x processes are well defined independent of the sampling effort and the amount of missing data. Construction of discrepancy measures on the x processes is therefore much easier. We will consider two measures. The first is the general Kolmogorov test statistic D^{kolmo} applied on the haul effects $\varepsilon_{a,h}$. The second discrepancy measure is more specifically defined towards the multinomial distribution, looking at the entropy within a haul given by

$$\text{Entr}_h = -\sum_a p_{h,a} \log(p_{h,a}),$$

which has a maximum value for all $p_{h,a}$ equal and a minimum value for $p_{h,a} = 1$ for one particular a. Different measures could be defined from the individual entropies, nevertheless, we have concentrated on maximum and minimum over all hauls in order to look for extreme situations. The marginalised discrepancy measures will be concidered for the entropy measures as those are in this case much easier to calculate. Note that for both small and large values of the entropy measures can be considered as deviations from the model.

For all discrepancy measures considered, analytical expressions for the p-values are not available and calculations has to be performed numerically. The Monte Carlo approximation given in (4) (with samples obtain through MCMC simulations, see Hirst et al. (2004, 2005) for details) has been applied to calculate the eppp values while (9) has been used for calibration.

Table 2 shows both the uncalibrated and the calibrated p-values obtained for the three discrepancy measures used. For each combination of discrepancy measure, p-value and

Table 2. eppp and ceppp for different discrepancy measures and different numbers of iterations (M) in the MCMC algorithm.

D-measure	p-value	$M = 100$	$M = 1000$	$M = 3000$	$M = 10000$
D^{kolmo}	eppp	0.470 (0.038)	0.466 (0.009)	0.466 (0.007)	0.468 (0.003)
Min Entr	eppp	0.003 (0.001)	0.003 (0.000)	0.003 (0.000)	0.003 (0.000)
Max Entr	eppp	0.319 (0.056)	0.277 (0.018)	0.280 (0.020)	0.257 (0.011)
D^{kolmo}	ceppp	0.250 (0.248)	0.011 (0.018)	0.001 (0.002)	0.003 (0.003)
Min Entr	ceppp	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
Max Entr	ceppp	0.240 (0.081)	0.177 (0.027)	0.176 (0.024)	0.144 (0.019)

number of MCMC iterations (M), 10 repetitions of eppp(y^{obs}) and ceppp(y^{obs}) were made in order to evaluate the uncertainty in calculation of the p-values. Mean and standard deviations of these to repetitions are reported

Considering D^{kolmo} first, we see a similar pattern as in the simulated examples that although the uncalibrated p-values are close to 0.5, they become much smaller when properly calibrated. We also see a typical pattern that for small numbers, the Monte Carlo variability of ceppp is large, giving relatively large p-values. When the number of iterations increases, the p-values become less variable but also more extreme, as demonstrated in Figure 8. This is the case even though the calibrated eppp values are centred around the same values for all choices of M.

The p-values based on entropy very much follow the same structure as the one based on D^{kolmo}, becoming more extreme when properly calibrated and decreasing with increasing number of MCMC-iterations. The eppp value based on the minimum entropy measure is extreme even in its uncalibrated form, but becomes even more extreme when properly calibrated. Even though the calibrated p-value is closer to zero than the uncalibrated one when regarding the maximum entropy with both p-values being closer to zero as the number of MCMC-iterations increases, their values are still being large.

The p-values in Table 2 indicate that the prior and model assumptions do not quite fit the data. The age-distribution of fish can vary considerable from year to year, and a shrinkage of the prior means towards zero improved the fit. This had however to be combined with an increase in the variance of the haul effect. Still, however, some of the p-values were quite small. A further inspection of the data revealed that within a haul h, the $\{\varepsilon_{h,a}\}$ were correlated. Changing these to follow an AR model (with the AR-coefficient equals to 0.9)

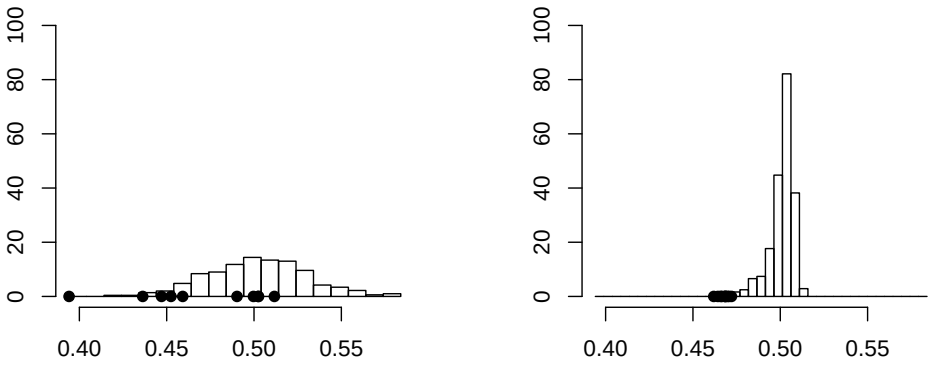

Fig. 8. Histogram of 500 simulations of eppp(Y) for the multinomial model using D^{kolmo} and with $M = 100$ (left) and $M = 10000$ (right). 10 replications of eppp(y^{obs}) are marked as black dots on the x-axis.

when simulating Y gave ceppp-values of $0.03, 0.03$ and 0.48 for D^{kolmo}, minimum entropy and maximum entropy, respectively.

Note that the model was not modified to allow for autoregressive correlations under the calculation of eppp. This means that the modification was only included in the simulations of Y. As long as we are using the same procedure for both calculating eppp(y^{obs}) and eppp(Y) this is perfectly legal, and ceppp will still be uniformly distributed under the modified model. One would however expect that using the assumed model also in the calculation of eppp would increase the power of the test.

6. Summary and conclusions

This paper has focused on posterior predictive checks in hierarchical models where we have presented two extended versions of the posterior predictive p-values (ppp), with the intention to test specifications and assumptions made on hidden processes. The extended posterior predictive p-value is computed as the posterior expectation of the ppp that we would have used if the latent process x had been observed, while the marginalised posterior predictive p-value eliminates the latent variable x by computing the posterior expectation of the discrepancy measure $D(y, x, \theta)$ for fixed θ's. Both apparatuses afford flexibility in

choosing discrepancy measures depending on hidden processes and unknown parameters, but the uniform scale for judging the significance of surprise is typically not appropriate neither for the extended nor the marginalised ppp's. We repaired this non-uniformity by calibrating the p-values into a genuine scale under model conditions, alleviated the problem of comparing across different models.

The characteristics of the uncalibrated p-values, so as the calibrated ones, were studied in the context of some structurally simple multi-level models, where an important task was to understand the consequences of observing x indirectly through y. The uncalibrated eppp and ppp^{mrg} are both conservative and have low power in detecting violations of the model, although the power increases with sharper priors. The calibrated p-values have higher power in all cases considered and are by definition uniformly distributed under the prior and model assumptions.

When using discrepancy measures related to prior assumptions, observation errors had negligible effect on the power, with the important factor being the sharpness of the prior. However, for discrepancy measures related to properties of the latent structure, the observation error seems to play a higher role, even with vague priors. We have considered only a few situations here, using known variances in the examples, but they show that by the proposed calibration techniques, posterior predictive checks improve their capability of detecting possible model violations under the model conditions.

The proposed p-values were applied to a real dataset concerning estimation of catch-at-age. Several discrepancy measures were considered, all demonstrating the conservativeness of the uncalibrated p-values. In one instance, a eppp value of 0.47 was transformed to a ceppp value of 0.003.

We stress that the Bayesians have to consider which data are likely or not to compute the level of surprise. Although we believe that the distribution of Y, implied by the model and prior in (7), is the appropriate distribution for transforming the unprocessed p-values to an uniform scale, there might be other distributions that are more relevant for other situations. Fairly commonly, statisticians use non-informative priors for computing the posterior distribution to avoid the risk of failing to capture the true parameter. In this case, we might prefer a narrower prior for the calibration, with its consequence that we in some extend can control the sampling of Y to produce reasonable values. Also, we may calibrate under proper priors while using improper priors for making inference.

Acknowledgements

We are grateful to Professor Nils Lid Hjort, for his valuable comments and suggestions during this work. We also thank Professor Alan Gelfand for contributing in discussions. The data used in Section 5 were kindly supplied by the Institute of Marine Research in Norway. This research was supported by of the Norwegian Research Council under the BeMatA program Evaluation of Hierarchical Models (project number 154911), chaired by Nils Lid Hjort. The last part of this work was performed when Storvik was a visiting Fellow at the Department of Mathematics, University of Bristol.

References

Banerjee, S., Carlin, B. P. & Gelfand, A. E. (2004), *Hierarchical Modeling and Analysis for Spatial Data*, Chapman & Hall/CRC, Florida.

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

Bayarri, M. J. & Castellanos, M. E. (2007), 'Bayesian Checking of Hierarchical Models', *Statistical Science* . to appear.

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

Dahl, F. A., Gsemyr, J. & Natvig, B. (2007), 'A robust conflict measure of inconsistencies in Bayesian hierarchical models', *Scandinavian Journal of Statistics* . To appear.

Dey, D., Gelfand, A., Swartz, T. & Vlachos, P. (1998), 'A simulation-intensive approach for checking hierarchical models', *Test* **7**, 325–346.

Diggle, P. J., Tawn, J. A. & Moyeed, R. A. (1998), 'Model-based geostatistics', *Applied Statistics* **47, part 2**, 000–000.

Draper, D. & Krnjajic, M. (2006), 'Bayesian model specification'. Submitted.

Gelfand, A. E. (2003), Some comments on model criticism, *in* P. J. Green, N. L. Hjort & S. Richardson, eds, 'Highly Structured Stochastic Systems', Oxford University Press, Oxford, pp. 449–523.

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (2004), *Bayesian Data Analysis*, Chapman & Hall, London.

Gelman, A., Mechelen, I. V., Verbeke, G., Heitjan, D. F. & Meulders, M. (2005), 'Multiple imputation for model checking: Completed-data plots with missing and latent data', *Biometrics* **61**, 74–85.

Gelman, A., Meng, X.-L. & Stern, H. (1996), 'Posterior predictive assessment of model fitness via realized discrepancies (with discussion)', *Statistics Sinica* **6**, 733–807.

Guttman, I. (1967), 'The use of the concept of a future observation in goodness-of-fit problems', *Journal of the Royal Statistical Society* **29**(B), 83–100.

Hirst, D., Aanes, S., Storvik, G., Huseby, R. B. & Tvete, I. F. (2004), 'Estimating catch-at-age from market sampling data using a bayesian hierarchical model', *Applied Statistics* **53**(1), 1–14.

Hirst, D., Storvik, G., Aldrin, M. Aanes, S. & Huseby, R. B. (2005), 'Estimating catch-at-age by combining data from different sources', *Canadian J. of Fisheries and Aquatic Sciences* **62**(6), 1377–1385.

Hjort, N. L., Dahl, F. A. & Steinbakk, G. H. (2006), 'Post-processing posterior predictive p-values', *Journal of the American Statistical Association* **101**, 1157–1174.

Lu, H., Hodges, J. S. & Carlin, B. P. (2006), 'Measuring the complexity of generalised linear hierarchical models', *Biostatistics* p. To appear.

Marshall, E. C. & Spiegelhalter, D. J. (2003), 'Approximate cross-validatory predictive checks in disease mapping models', *Statistics in Medicine* **22**, 1649–1660.

Meng, X.-L. (1994), 'Posterior predictive p-values', *Annals of Statistics* **22**, 1142–1160.

O'Hagan, A. (2003), HSSS model criticism, *in* P. J. Green, N. L. Hjort & S. Richardson, eds, 'Highly Structured Stochastic Systems', Oxford University Press, Oxford, pp. 423–444.

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

Rubin, D. B. (1984), 'Bayesian justifiable and relevant frequency calculations for the applied statistician', *Annals of Statistics* **12**, 1251–1172.

Sinharay, S. & Stern, H. (2003), 'Posterior predictive model checking in hierarchical models', *Journal of Statistical Planning and Inference* **111**, 209–221.

Skrondal, A. & Rabe-Hesketh, S. (2004), *Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models*, Chapman & Hall/CRC, Florida.

Address for correspondence: Gunnhildur Högnadóttir Steinbakk, Department of Mathematics, University of Oslo, Norway. Email:ghs@math.uio.no

Appendix

Proof of Proposition 1

We compute the eppp by integrating out $(\theta, \bar{x})'$ conditionally on the data, using the multivariate normal distribution

$$\begin{pmatrix} \theta \\ \bar{x} \\ \bar{y} \end{pmatrix} = N_3 \left[\begin{pmatrix} \theta_0 \\ \theta_0 \\ \theta_0 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \sigma_0^2 & \sigma_0^2 \\ \sigma_0^2 & \sigma_0^2 + \sigma^2/n & \sigma_0^2 + \sigma^2/n \\ \sigma_0^2 & \sigma_0^2 + \sigma^2/n & \sigma_0^2 + (\sigma^2 + \tau^2)/n \end{pmatrix} \right].$$

Hence the posterior distribution $(\theta, \bar{x})'|\bar{y}^{\text{obs}}$ is normal with mean and covariance matrix

$$\begin{aligned} \mathrm{E}[(\theta, \bar{x})'|\bar{y}] &= \frac{1}{n\sigma_0^2 + \sigma^2 + \tau^2} \begin{pmatrix} \theta_0(\sigma^2 + \tau^2) + n\sigma_0^2 \bar{y}^{\text{obs}} \\ \theta_0 \tau^2 + (n\sigma_0^2 + \sigma^2)\bar{y}^{\text{obs}} \end{pmatrix}, \\ \mathrm{var}((\theta, \bar{x})'|\bar{y}) &= \frac{1}{n\sigma_0^2 + \sigma^2 + \tau^2} \begin{pmatrix} \sigma_0^2(\sigma^2 + \tau^2) & \sigma_0^2 \tau^2 \\ \sigma_0^2 \tau^2 & \tau^2(\sigma_0^2 + \sigma^2/n) \end{pmatrix}, \end{aligned}$$

which leads us to

$$\mathrm{E}(\bar{x} - \theta|\bar{y}^{\text{obs}}) = \frac{\sigma^2(\bar{y}^{\text{obs}} - \theta_0)}{n\sigma_0^2 + \sigma^2 + \tau^2} \text{ and } \mathrm{var}(\bar{x} - \theta|\bar{y}^{\text{obs}}) = \frac{\sigma^2}{n} \frac{n\sigma_0^2 + \tau^2}{n\sigma_0^2 + \sigma^2 + \tau^2}.$$

Further, we learn that, for given y^{obs},

$$D^{\text{chisq}}(x, \theta) = \frac{n}{\sigma^2} \left(\frac{\sigma^2(\bar{y}^{\text{obs}} - \theta_0)}{n\sigma_0^2 + \sigma^2 + \tau^2} + \frac{\sigma}{\sqrt{n}} \sqrt{\frac{n\sigma_0^2 + \tau^2}{n\sigma_0^2 + \sigma^2 + \tau^2}} U \right)^2$$

for U standard normal. Since $\bar{x}^{\text{rep}} \sim \theta + (\sigma/\sqrt{n})V$ conditional on θ, for yet another standard normal V, independent of U, we get $D(x^{\text{rep}}, \theta) = V^2$. Hence, the eppp becomes

$$\begin{aligned}
\text{eppp}(\bar{y}^{\text{obs}}, D^{\text{chisq}}) &= \Pr\left\{ V^2 \geq \tfrac{n}{\sigma^2}\left(\tfrac{\sigma^2(\bar{y}^{\text{obs}}-\theta_0)}{n\sigma_0^2+\sigma^2+\tau^2} + \tfrac{\sigma}{\sqrt{n}}\sqrt{\tfrac{n\sigma_0^2+\tau^2}{n\sigma_0^2+\sigma^2+\tau^2}}U \right)^2 \right\} \\
&= \Pr\left\{ V^2 \geq \tfrac{n\sigma_0^2+\tau^2}{n\sigma_0^2+\sigma^2+\tau^2}\left(\tfrac{\sigma\sqrt{n}(\bar{y}^{\text{obs}}-\theta_0)}{\sqrt{n\sigma_0^2+\tau^2}\sqrt{n\sigma_0^2+\sigma^2+\tau^2}} + U \right)^2 \right\},
\end{aligned}$$

which claims the result.

Proof of Proposition 2

Under a perfect model and prior $\overline{Y}$ is $N(\theta_0, \sigma_0^2 + \sigma^2/n + \tau^2/n)$ and we can write $\overline{Y} = \theta_0 + \sqrt{\sigma_0^2 + \sigma^2/n + \tau^2/n}W$ for $W \sim N(0,1)$. Computing eppp with $\overline{Y}$ and $\bar{y}^{\text{obs}}$ using model $\mathcal{M}_1$ with D^{chisq} and inserting these terms in ceppp (8), we get

$$\begin{aligned}
\text{ceppp}(y^{\text{obs}}; D^{\text{chisq}}) &= \Pr\left\{ F_{1,1}\left(1 + \tfrac{\sigma^2}{n\sigma_0^2+\tau^2}, \tfrac{\sigma^2}{n\sigma_0^2+\tau^2}\tfrac{n(\overline{Y}-\theta_0)^2}{n\sigma_0^2+\sigma^2+\tau^2}\right) \right. \\
&\qquad \left. \leq F_{1,1}\left(1 + \tfrac{\sigma^2}{n\sigma_0^2+\tau^2}, \tfrac{\sigma^2}{n\sigma_0^2+\tau^2}\tfrac{n(\bar{y}^{\text{obs}}-\theta_0)^2}{n\sigma_0^2+\sigma^2+\tau^2}\right) \right\} \\
&= \Pr\{(\overline{Y}-\theta_0)^2 \geq (\bar{y}^{\text{obs}}-\theta_0)^2\} \\
&= \Pr\left\{ \tfrac{n\sigma_0^2+\sigma^2+\tau^2}{n} W^2 \geq (\bar{y}^{\text{obs}}-\theta_0)^2 \right\},
\end{aligned}$$

which claims the result. The above computation involves that $F_{1,1}(\nu;\kappa_1) \leq F_{1,1}(\nu;\kappa_2)$, or equivalently $\Pr\{(U+\sqrt{\kappa_1})^2/V^2 \leq \nu\} \leq \Pr\{(U+\sqrt{\kappa_2})^2/V^2 \leq \nu\}$, implies $\kappa_1 \geq \kappa_2$.

Proof of Proposition 3

We may write

$$\hat{\beta}_2(x^{\text{rep}}) = \tfrac{1}{s_z^2} z' x^{\text{rep}} = \tfrac{1}{s_z^2} z'[Z\beta + \varepsilon^{\text{rep}}] = \beta_2 + \tfrac{1}{s_z^2} z' \varepsilon^{\text{rep}},$$

where we have utilised the orthogonal structure of Z and $\varepsilon^{\text{rep}} \sim N(0, \sigma^2 I)$ independently of β_2. Utilising that $Z^{'}Z = \text{diag}(n, s_z^2)$, we get

$$\begin{aligned}
E[\beta|x] &= [M_0 + Z^{'}Z]^{-1}[M_0\beta_0 + Z^{'}x] = \begin{pmatrix} \frac{\kappa_1\beta_{0,1}+n\bar{x}}{\kappa_1+n} \\ \frac{\kappa_2\beta_{0,2}+s_z^2\hat{\beta}_2(x)}{\kappa_2+s_z^2} \end{pmatrix}, \\
\text{var}[\beta|x] &= \sigma^2[M_0 + Z^{'}Z]^{-1} = \sigma^2 \begin{pmatrix} \frac{1}{\kappa_1+n} & 0 \\ 0 & \frac{1}{\kappa_2+s_z^2} \end{pmatrix}.
\end{aligned}$$

Since $\beta|x$ is Gaussian, it follows that also $\hat{\beta}_2(x^{\text{rep}})|x$ is Gaussian with

$$E[\hat{\beta}_2(x^{\text{rep}})|x] = \frac{\kappa_2\beta_{0,2}+s_z^2\hat{\beta}_2(x)}{\kappa_2+s_z^2} \text{ and } \text{var}[\hat{\beta}_2(x^{\text{rep}})|x] = \frac{\sigma^2}{s_z^2}\frac{\kappa_2+2s_z^2}{\kappa_2+s_z^2}.$$

This means we can write

$$\hat{\beta}_2(x^{\text{rep}}) - \beta_{0,2} = \frac{s_z^2}{\kappa_2+s_z^2}(\hat{\beta}_2(x) - \beta_{0,2}) + \frac{\sigma}{s_z}\sqrt{\frac{\kappa_2+2s_z^2}{\kappa_2+s_z^2}}U,$$

for U a standard normal variable. Inserting this into (2), we get (14).

Proof of Corollary 1

For κ_2 large or s_z small , $a(z) \approx 0$ and $b(z) \approx 1$ and the result follows for U standard normal.

For κ_2 small or s_z large , $a(s) \approx 1$ and $b(z) \approx \sqrt{2}$. In the limit, for $\Delta = s_z(\hat{\beta}_2(x) - \beta_{0,2})/\sigma > 0$,

$$\begin{aligned}\text{ppp}^{\text{compl}}(y^{\text{obs}}, x; D^{\beta_2}) &= \Pr\left\{\left[\Delta + \sqrt{2}U\right]^2 \geq \Delta^2\right\} \\ &= \Pr\left\{\sqrt{2}U \geq 0\right\} + \Pr\left\{\sqrt{2}U \leq -2\Delta\right\} \\ &= \tfrac{1}{2} + \Phi(-\sqrt{2}\Delta),\end{aligned}$$

which by symmetry also yields for $\Delta < 0$.

Proof of Proposition 4

We have

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim N\left[\begin{pmatrix} Z\beta_0 \\ Z\beta_0 \end{pmatrix}, \begin{pmatrix} \Sigma_x & \Sigma_x \\ \Sigma_x & \Sigma_y \end{pmatrix}\right],$$

where $\Sigma_x = \sigma^2 I + \sigma^2 Z M_0^{-1} Z^{'}$ and $\Sigma_y = (\sigma^2+\tau^2)I + \sigma^2 Z M_0^{-1} Z^{'}$. Now,

$$\text{E}[\hat{\beta}_2(x)|y] = \tfrac{1}{s_z^2} z'Z\beta_0 + \tfrac{1}{s_z^2} z'\Sigma_x\Sigma_y^{-1}[y - Z\beta_0], \tag{22}$$

$$\text{var}[\hat{\beta}_2(x)|y] = \tfrac{1}{s_z^4} z'[\Sigma_x - \Sigma_x\Sigma_y^{-1}\Sigma_x]z. \tag{23}$$

Further, using that $ZM_0^{-1}Z' = \kappa_1^{-1}11^{'} + \kappa_2^{-1}zz^{'}$, we obtain

$$\Sigma_y^{-1} = \frac{1}{\sigma^2+\tau^2}[I - \frac{\sigma^2}{\kappa_1(\sigma^2+\tau^2)+n\sigma^2}11^{'} - \frac{\sigma^2}{\kappa_2(\sigma^2+\tau^2)+s_z^2\sigma^2}zz^{'}].$$

Inserting this into (22)-(23) and utilising that $z'Z = (0, s_z^2)$, we obtain

$$\begin{aligned}
\mathrm{E}\left[\tfrac{\hat{\beta}_2(x)-\beta_{0,2}}{\sigma/s_z}|y^{\mathrm{obs}}\right] &= \tfrac{\kappa_2+s_z^2}{\kappa_2+s_z^2+\kappa_2(\tau^2/\sigma^2)}\tfrac{\hat{\beta}_2(y^{\mathrm{obs}})-\beta_{0,2}}{\sigma/s_z}, \\
\mathrm{var}\left[\tfrac{\hat{\beta}_2(x)-\beta_{0,2}}{\sigma/s_z}|y^{\mathrm{obs}}\right] &= \tfrac{\tau^2(\kappa_2+s_z^2)}{\sigma^2(\kappa_2+s_z^2)+\kappa_2\tau^2}.
\end{aligned}$$

By defining $V = s_z(\hat{\beta}_2(x) - \beta_{0,2})/\sigma$, the result of the proposition is obtained. Note that U from Proposition 3 is independent of x, so V and U are independent variables.

Proof of Proposition 5

We may write $\hat{\beta}_2(y^{\mathrm{rep}}) = \beta_2 + s_z^{-2}z'\varepsilon^{\mathrm{rep}}$, which, given y^{obs}, is Gaussian with

$$\begin{aligned}
\mathrm{E}[\hat{\beta}_2(y^{\mathrm{rep}})|y^{\mathrm{obs}}] &= \tfrac{\kappa_2(\sigma^2+\tau^2)\beta_{0,2}+s_z^2\sigma^2\hat{\beta}_2(y^{\mathrm{obs}})}{\kappa_2(\sigma^2+\tau^2)+s_z^2\sigma^2}, \\
\mathrm{var}[\hat{\beta}_2(y^{\mathrm{rep}})|y^{\mathrm{obs}}] &= \tfrac{\sigma^2+\tau^2}{s_z^2}\tfrac{\kappa_2(\sigma^2+\tau^2)+2s_z^2\sigma^2}{\kappa_2(\sigma^2+\tau^2)+s_z^2\sigma^2}.
\end{aligned}$$

This means we can write

$$\tfrac{\hat{\beta}_2(y^{\mathrm{rep}})-\beta_{0,2}}{\sqrt{\sigma^2+\tau^2}/s_z} = \tfrac{s_z^2\sigma^2}{\kappa_2(\sigma^2+\tau^2)+s_z^2\sigma^2}\tfrac{\hat{\beta}_2(y^{\mathrm{obs}})-\beta_{0,2}}{\sqrt{\sigma^2+\tau^2}/s_z} + \sqrt{\tfrac{\kappa_2(\sigma^2+\tau^2)+2s_z^2\sigma^2}{\kappa_2(\sigma^2+\tau^2)+s_z^2\sigma^2}}U$$

for U standard normal, and which, inserted into (18) proves the result.