Computational Statistics and Data Analysis 79 (2014) 277–291

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Comparative investigation of three Bayesian *p* values

Junni L. Zhang *

Guanghua School of Management and Center for Statistical Science, Peking University, Beijing 100871, PR China

ARTICLE INFO

Article history:
Received 7 September 2012
Received in revised form 5 May 2014
Accepted 17 May 2014
Available online 2 June 2014

Keywords:
Bayesian model checking
Posterior predictive *p* value
Sampled posterior *p* value
Calibrated posterior predictive *p* value
Hierarchical model
Causal effect

ABSTRACT

Bayesian *p* values are a popular and important class of approaches for Bayesian model checking. They are used to quantify the degree of surprise from the observed data given the specified data model and prior distribution. A systematic investigation is conducted to compare three Bayesian *p* values — the posterior predictive *p* value, the sampled posterior *p* value and the calibrated posterior predictive *p* value. Their general computation costs are compared, and several examples that incorporate both simple and complex Bayesian models are used to compare their frequency properties. It is recommended to use the sampled posterior *p* value because it is computationally least expensive and safest.

1. Introduction

In Bayesian modeling, it is assumed that the observed data $\boldsymbol{y}^{obs}$ come from a data model, $f(\boldsymbol{y}|\boldsymbol{\theta})$, for data $\boldsymbol{y}$ given parameters $\boldsymbol{\theta}$, and a prior distribution $\pi(\boldsymbol{\theta})$ is specified for the parameters. Bayesian statisticians often want to check whether a given set of model conditions, including both the data model and the prior, are adequate for the observed data, sometimes without explicit alternative models. The role of model checking versus model selection has been addressed, for example, in Bayarri and Berger (2000), Robins et al. (2000), Johnson (2004) and Hjort et al. (2006), and we will not discuss it further.

Bayesian *p* values are a popular and important class of approaches for addressing the model checking issue. Here a *p* value is regarded as being "Bayesian" when it uses the prior or posterior distribution for all occurrences of $\boldsymbol{\theta}$. Hence, for example, some *p* values discussed in Robins et al. (2000), including the plug-in *p* value which uses the MLE of $\boldsymbol{\theta}$, the conditional plug-in *p* value which uses a conditional MLE of $\boldsymbol{\theta}$, and the adjusted *p* values which use the MLE of $\boldsymbol{\theta}$, are not considered here. Bayesian *p* values involve discrepancy measures $D(\boldsymbol{y}, \boldsymbol{\theta})$, functions of the data and the parameters that are ideally "chosen to reflect aspects of the model that are relevant to the scientific purposes to which the inference will be applied" and are "commonly chosen to measure a feature of the data not directly addressed by the probability model" (Section 6.5 of Gelman et al., 2003). Bayesian *p* values are then defined as the probability that $D(\boldsymbol{y}^{rep}, \boldsymbol{\theta})$ is no smaller (or larger) than $D(\boldsymbol{y}^{obs}, \boldsymbol{\theta})$ when the parameters $\boldsymbol{\theta}$ and the replicated data $\boldsymbol{y}^{rep}$ are randomly drawn from some distribution related to a given set of model conditions, hence quantifying the degree of surprise from the observed data under the given model conditions. Note that a statistic is a special case of a discrepancy measure in that it does not depend on the parameters. In problems with missing or latent data, extensions for the discrepancy measures to depend on the complete data (including both $\boldsymbol{y}$ and the missing or latent data) can be made along the same line as in Gelman et al. (2005), and we will not pursue this issue further.

Various forms of Bayesian *p* values have been proposed in the literature, including the prior predictive *p* value (Box, 1980), the posterior predictive *p* value (Rubin, 1984; Meng, 1994; Gelman et al., 1996), the partial posterior predictive *p* value and

* Tel.: +86 10 62757922; fax: +86 10 62757922.
E-mail address: zjn@gsm.pku.edu.cn.

http://dx.doi.org/10.1016/j.csda.2014.05.012
0167-9473/

the conditional predictive p value (Bayarri and Berger, 2000, 2004), the sampled posterior p value (Johnson, 2004, 2007; Gosselin, 2011), and the calibrated posterior predictive p value (Hjort et al., 2006). We will not investigate the prior predictive p value because it is undefined under improper prior distributions and is strongly sensitive to the prior distribution. Also, we will not investigate the partial posterior predictive p value or the conditional predictive p value because they can only be easily worked out for specific discrepancy measures. We will conduct a systematic investigation of the posterior predictive p (ppp) value, the sampled posterior p (spp) value and the calibrated posterior predictive p ($cppp$) value, since these three Bayesian p values can be applied with general Bayesian models and discrepancy measures. Following Box (1980), Rubin (1984) and others, we deem it important to investigate the frequency properties of Bayesian procedures when $\boldsymbol{\theta}$ and $\boldsymbol{y}^{obs}$ are repeatedly sampled from the true model conditions in order to recommend procedures for general applied Bayesian analysis. Hence we will focus on investigating the frequency properties of the above three Bayesian p values.

In the following text, we use the notation $\Pr^{m(\mathbf{V})}(\boldsymbol{E})$ to denote the probability of event $\boldsymbol{E}$ when the random variables in $\mathbf{V}$ follow the distribution $m(\cdot)$. The ppp value is defined as

$$ppp(\boldsymbol{y}^{obs}) = \Pr^{m_{ppp}(\boldsymbol{\theta},\boldsymbol{y}^{rep})}\left[D(\boldsymbol{y}^{rep},\boldsymbol{\theta}) \geq D(\boldsymbol{y}^{obs},\boldsymbol{\theta})\right], \tag{1}$$

where $m_{ppp}(\boldsymbol{\theta},\boldsymbol{y}^{rep}) = \pi(\boldsymbol{\theta}|\boldsymbol{y}^{obs})f(\boldsymbol{y}^{rep}|\boldsymbol{\theta})$, in which $\pi(\boldsymbol{\theta}|\boldsymbol{y}^{obs})$ denotes the posterior distribution of $\boldsymbol{\theta}$. Provided that we can simulate $\boldsymbol{\theta}_j^A$ from $\pi(\boldsymbol{\theta}|\boldsymbol{y}^{obs})$ and $\boldsymbol{y}_j^{rep\ A}$ from $f(\boldsymbol{y}|\boldsymbol{\theta}_j^A)$ for $j = 1, \ldots, M_A$, the ppp value can be approximated by

$$ppp(\boldsymbol{y}^{obs}) \approx \frac{1}{M_A}\sum_{j=1}^{M_A} I[D(\boldsymbol{y}_j^{rep\ A},\boldsymbol{\theta}_j^A) \geq D(\boldsymbol{y}^{obs},\boldsymbol{\theta}_j^A)]. \tag{2}$$

The ppp value involves double use of the data, because $\boldsymbol{y}^{obs}$ is first used to obtain $\pi(\boldsymbol{\theta}|\boldsymbol{y}^{obs})$ and then used in computing the probability. Bayarri and Berger (2000), Robins et al. (2000), Bayarri and Castellanos (2007) and others found that when the data model $f(\boldsymbol{y}|\boldsymbol{\theta})$ is correctly specified and the true parameter value $\boldsymbol{\theta}^{tr}$ is fixed, the ppp value is conservative, i.e., its distribution is more concentrated around 1/2 than a uniform. In particular, Robins et al. (2000) showed theoretically that, under certain regularity conditions, this conservativeness holds asymptotically when the data consist of independent observations. On the other hand, Dey et al. (1998), Bayarri and Berger (2000), Sinharay and Stern (2003), Bayarri and Castellanos (2007) and others found that when the data model is incorrectly specified, the ppp value lacks power to detect model inadequacies. Hjort et al. (2006) noted that the distribution of ppp can take a variety of possible shapes, making its interpretation difficult and risky.

The spp value is defined as

$$spp = \Pr^{m_{spp}(\boldsymbol{\theta},\boldsymbol{y}^{rep})}\left[D(\boldsymbol{y}^{rep},\boldsymbol{\theta}) \geq D(\boldsymbol{y}^{obs},\boldsymbol{\theta})\right], \tag{3}$$

where $m_{spp}(\boldsymbol{\theta},\boldsymbol{y}^{rep}) = \delta_{\boldsymbol{\theta}^{single}}(\boldsymbol{\theta})f(\boldsymbol{y}^{rep}|\boldsymbol{\theta})$, in which $\delta(\cdot)$ is the degenerate point mass distribution and $\boldsymbol{\theta}^{single}$ is a single random parameter value drawn from $\pi(\boldsymbol{\theta}|\boldsymbol{y}^{obs})$. Provided that we can simulate $\boldsymbol{\theta}^{single}$ from $\pi(\boldsymbol{\theta}|\boldsymbol{y}^{obs})$ and $\boldsymbol{y}_k^{rep\ B}$ from $f(\boldsymbol{y}|\boldsymbol{\theta}^{single})$ for $k = 1, \ldots, M_B$, the spp value can be approximated by

$$spp \approx \frac{1}{M_B}\sum_{k=1}^{M_B} I[D(\boldsymbol{y}_k^{rep\ B},\boldsymbol{\theta}^{single}) \geq D(\boldsymbol{y}^{obs},\boldsymbol{\theta}^{single})]. \tag{4}$$

Hence, rather than comparing, as what the ppp value does, the observed data with a collection of replicated data sets where a single replicated data set is drawn given each of the multiple posterior draws of $\boldsymbol{\theta}$, the spp value compares the observed data with multiple replicated data sets drawn given a single random posterior draw of $\boldsymbol{\theta}$.

The use of a single random posterior draw of $\boldsymbol{\theta}$ in Bayesian model checking was pioneered by Johnson (2004, 2007). Let π^{tr} denote the distribution of $\boldsymbol{\theta}$ in the true data-generating process. Johnson (2007) noted that if the data model is correctly specified and $\pi = \pi^{tr}$, the marginal distribution of $\boldsymbol{\theta}^{single}$ is equal to π, and the marginal distributions of $D(\boldsymbol{y}^{obs},\boldsymbol{\theta}^{single})$ and $D(\boldsymbol{y}^{obs},\boldsymbol{\theta}^{tr})$ are identical; when D is pivotal, the null distribution of $D(\boldsymbol{y}^{obs},\boldsymbol{\theta}^{tr})$ becomes known, and hence we can compare $D(\boldsymbol{y}^{obs},\boldsymbol{\theta}^{single})$ against this reference distribution. Gosselin (2011) further noted that if the data model is correctly specified and $\pi = \pi^{tr}$, the conditional distribution of $\boldsymbol{y}^{obs}$ given $\boldsymbol{\theta}^{single}$ is equal to f, and therefore the conditional distributions of $D(\boldsymbol{y}^{obs},\boldsymbol{\theta}^{single})$ and $D(\boldsymbol{y}^{rep},\boldsymbol{\theta}^{single})$ given $\boldsymbol{\theta}^{single}$ are identical, so the spp value is uniformly distributed. Gosselin (2011) also proved that if the data model is correctly specified but $\pi \neq \pi^{tr}$, under certain regularity conditions, the distribution of spp is asymptotically uniform if the data consist of independent observations. To investigate the finite sample behavior of the spp value when the data model is correctly specified, Gosselin (2011) also used simulation scenarios with independent observations generated from several simple Bayesian models, and found that the spp value was close to being uniform when the data model is correctly specified and π was "not too informative and not too uninformative", and not too far off-centered relative to π^{tr}, and that the spp value has more power than the ppp value when the data model is incorrectly specified.

Note that given $\boldsymbol{y}^{obs}$, the ppp value is a statistic but the spp value is a random variable that depends on a random posterior sample of $\boldsymbol{\theta}$. Therefore for the same observed data set and the same statistical model, model checking using the spp values based on two different posterior samples of $\boldsymbol{\theta}$ may give different conclusions. However, it is exactly this random sampling of parameters that removes the influence of double use of the data. As (Gosselin, 2011) pointed out, "as we are working on

sampled data to fit statistical models, we should also agree to work on *sampled* parameters to criticize the model. Indeed, this double sampling allowed us to make the roles of data and parameters symmetrical [...]". Just as the fact that the sampled data may have low probability conditional on the true model conditions does not preclude their use in fitting models, the fact that the sampled parameters may have low probability conditional on the data does not preclude their use in checking models.

We would also like to note that, the roles of data and parameters in model checking being made symmetrical does not mean that two samples of $\boldsymbol{\theta}$ given two data sets of different sizes have the same credibility. If the model conditions are correctly specified, the *spp* values based on two data sets of different sizes both follow the uniform distribution, and therefore their probabilities of type I error are the same. However, if the model conditions are not correctly specified, the *spp* value based on a data set with larger size has more power than that based on a data set with smaller size. This point will be demonstrated in Section 4.3.

The *cppp* value was intended to define the underlying probability scale of the *ppp* value by comparing the *ppp* value for the observed data with other possible *ppp* values that could result from data sets generated using the assumed data model and a calibration prior $\pi^*(\boldsymbol{\theta})$, and it is thus defined as

$$cppp(\boldsymbol{y}^{obs}) = \Pr^{m_{cppp}(\boldsymbol{y}^{rep})}\left\{ppp(\boldsymbol{y}^{rep}) \le ppp(\boldsymbol{y}^{obs})\right\}, \tag{5}$$

where $m_{cppp}(\boldsymbol{y}^{rep}) = \int \pi^*(\boldsymbol{\theta}) f(\boldsymbol{y}^{rep}|\boldsymbol{\theta}) d\boldsymbol{\theta}$. The *cppp* value is generally computed by a double simulation scheme. Provided that we can simulate $\boldsymbol{\theta}_l^C$ from $\pi^*(\boldsymbol{\theta})$ and $\boldsymbol{y}_l^{rep\ C}$ from $f(\boldsymbol{y}|\boldsymbol{\theta}_l^C)$ for $l = 1, \ldots, M_C$, the *cppp* value can be approximated by

$$cppp(\boldsymbol{y}^{obs}) \approx \frac{1}{M_C}\sum_{l=1}^{M_C} I[ppp(\boldsymbol{y}_l^{rep\ C}) \le ppp(\boldsymbol{y}^{obs})], \tag{6}$$

where $ppp(\boldsymbol{y}^{obs})$ can be approximated by (2) and $ppp(\boldsymbol{y}_l^{rep\ C})$ can be approximated similarly. Hjort et al. (2006) illustrated the *cppp* analysis with independent observations through several theoretical cases and several applications, and pointed out that the *cppp* method may be generalized to various hierarchical models.

According to the definition of the *cppp* value in (5), when $\boldsymbol{y}^{rep}$ has the same distribution as $\boldsymbol{y}^{obs}$, the *cppp* value has a uniform distribution. This happens when the data model $f(\boldsymbol{y}|\boldsymbol{\theta})$ is correctly specified and the calibration prior $\pi^*(\boldsymbol{\theta})$ is equal to $\pi^{tr}(\boldsymbol{\theta})$, the true distribution of $\boldsymbol{\theta}$ with which $\boldsymbol{y}^{obs}$ is generated. The influence of double use of the data is thus removed through the right calibration of the *ppp* value. Typically, the calibration prior $\pi^*(\boldsymbol{\theta})$ is taken to be equal to $\pi(\boldsymbol{\theta})$ used in fitting the statistical model. Hjort et al. (2006) pointed out that it may be of interest to investigate some "what if" scenarios with π^* different from π. Hjort et al. (2006) illustrated this point with a survival data analysis example in which it is preferred to fit the model under a vague prior π that indicates "nearly every human being would die before age 65 and very few would celebrate their 60th birthday" in order that the data get the chance to show the way to the "true" parameter value, whereas it is preferred to calibrate the *ppp* value against a narrower prior π^* that makes practical sense or is a closer guess of π^{tr}. In such "what if" scenarios, the analysis of the goodness-of-fit of the model is conducted using different priors on the unknown parameters. Although it is unusual in a single Bayesian analysis to have different prior opinions, these are needed in order to avoid the double use of the data when using the *ppp* value and achieve the right calibration of it. In practice, however, even if the data model is specified correctly, π^{tr} can never be known; for example, it is often assumed that π^{tr} is the degenerate point mass distribution at some true unknown parameter value $\boldsymbol{\theta}^{tr}$. Therefore, π^* can never equal π^{tr} and hence the distribution of the *cppp* value is different from uniform.

The major purpose of this paper is to compare the frequency properties of the above three Bayesian p values through three demonstrative examples that incorporate a simple Bayesian model with independent observations and two other Bayesian models that go beyond independent observations. We will study both the probability of type I error when the data model is correctly specified and the power when the data model is incorrectly specified. We will also vary the relationship among π, π^{tr} and π^*.

The rest of the paper is organized as follows. In Section 2, we compare the general computational cost for the three Bayesian p values. Elements of such a comparison have appeared in Hjort et al. (2006) and Gosselin (2011), but here we will do a more quantitative comparison. In Section 3, we investigate a simple model specified in Hjort et al. (2006) where data are assumed to be iid normally distributed with known variance and a normal prior for the unknown mean, and analytically assess the null distributions of the three Bayesian p values. We also discuss the asymptotic frequency properties of the three Bayesian p values when the true unknown mean is fixed. In Section 4, we investigate a hierarchical model specified in Sinharay and Stern (2003) where data are assumed to be independently normally distributed with the same known variance and different unknown means that come from the same normal distribution. Through different simulation scenarios for the second-level model and the hyperpriors, we compare the frequency properties of the three Bayesian p values. In Section 5, we investigate a causal effect model that is more complicated and requires higher computational cost, and again compare the frequency properties of the three Bayesian p values through simulation. Section 6 then concludes with a summarization of results in this paper and a discussion connecting this paper with the relevant literature.

2. General computational cost for the three Bayesian *p* values

The posterior draws of $\boldsymbol{\theta}$ are usually obtained using a Markov Chain Monte Carlo (MCMC) algorithm. Suppose we discard the first Q_b draws for the burnin period, and keep every Q_s'th draw after burnin for posterior inference. Let $T_{\boldsymbol{\theta}}$, $T_{\boldsymbol{\theta}}^*$, $T_{\boldsymbol{y}}$ and

T_D, respectively, denote the computational cost associated with generating one draw of $\boldsymbol{\theta}$ from the posterior distribution through one full MCMC step, generating one draw of $\boldsymbol{\theta}$ from the calibration prior π^*, generating one replicated data set $\boldsymbol{y}^{rep}$ given the value of $\boldsymbol{\theta}$ and evaluating the discrepancy measure once given the value of $\boldsymbol{y}$ and $\boldsymbol{\theta}$.

For approximation of the *ppp* value, we need to generate $(Q_b + Q_sM_A)$ draws of $\boldsymbol{\theta}$ in order to obtain $\boldsymbol{\theta}_j^A$ $(j = 1, \ldots, M_A)$, generate one replicated data set $\boldsymbol{y}_j^{rep\ A}$ given each $\boldsymbol{\theta}_j^A$ $(j = 1, \ldots, M_A)$, and evaluate $D(\boldsymbol{y}_j^{rep\ A}, \boldsymbol{\theta}_j^A)$ and $D(\boldsymbol{y}^{obs}, \boldsymbol{\theta}_j^A)$ $(j = 1, \ldots, M_A)$. Hence, the associated computational cost can be derived as

$$T_{ppp} = (Q_b + Q_sM_A)T_{\boldsymbol{\theta}} + M_AT_{\boldsymbol{y}} + 2M_AT_D.$$

Suppose that in computing the *spp* value, we keep the first posterior draw after burnin as $\boldsymbol{\theta}^{single}$. We need to generate M_B replicated data sets $\boldsymbol{y}_k^{rep\ B}$ $(k = 1, \ldots, M_B)$ given $\boldsymbol{\theta}^{single}$, and evaluate $D(\boldsymbol{y}_k^{rep\ B}, \boldsymbol{\theta}^{single})$ $(k = 1, \ldots, M_B)$ and $D(\boldsymbol{y}^{obs}, \boldsymbol{\theta}^{single})$. The associated computational cost is therefore

$$T_{spp} = (Q_b + 1)T_{\boldsymbol{\theta}} + M_BT_{\boldsymbol{y}} + (M_B + 1)T_D.$$

For approximation of the *cppp* value, we need to generate $\boldsymbol{\theta}_l^C$ $(l = 1, \ldots, M_C)$ from π^*, generate one replicated data set $\boldsymbol{y}_l^{rep\ C}$ given each $\boldsymbol{\theta}_l^C$ $(l = 1, \ldots, M_C)$, and approximate the *ppp* value for each $\boldsymbol{y}_l^{rep\ C}$ $(l = 1, \ldots, M_C)$ and also for $\boldsymbol{y}^{obs}$. The associated computational cost can therefore be derived as

$$T_{cppp} = M_C(T_{\boldsymbol{\theta}}^* + T_{\boldsymbol{y}}) + (M_C + 1)[(Q_b + Q_sM_A)T_{\boldsymbol{\theta}} + M_AT_{\boldsymbol{y}} + 2M_AT_D].$$

The double-simulation scheme for the *cppp* value makes it most computationally expensive. To compare the computational costs of the *ppp* and *spp* values, note that

$$T_{ppp} - T_{spp} = (Q_sM_A - 1)T_{\boldsymbol{\theta}} + (M_A - M_B)T_{\boldsymbol{y}} + (2M_A - M_B - 1)T_D. \tag{7}$$

When $M_A = M_B$, apparently the *spp* value has much lower computational cost than the *ppp* value.

3. The normal–normal model

3.1. The null distributions of the three Bayesian p values

We now investigate the simple normal–normal model given by Hjort et al. (2006). Assume that the data $\boldsymbol{y} = (y_1, \ldots, y_n)$ are iid from $N(\theta, \sigma^2)$ where σ^2 is known, and let the prior be $\pi(\theta) \sim N(\theta_0, \sigma_0^2)$. As in Hjort et al. (2006), suppose that the discrepancy measure of interest is

$$D(\boldsymbol{y}, \theta) = \frac{n(\bar{y} - \theta)^2}{\sigma^2}, \tag{8}$$

where $\bar{y}$ is the mean of y_i's.

Let $\boldsymbol{y}^{obs} = (y_1^{obs}, \ldots, y_n^{obs})$ denote the observed data, and let $\bar{y}^{obs}$ denote the mean of y_i^{obs}'s. If the data model is correctly specified and $\pi^{tr} = \pi$, it can be easily derived that marginally $\bar{y}^{obs} \sim N(\theta_0, \sigma_0^2 + \sigma^2/n)$, and hence

$$\bar{y}^{obs} \sim \theta_0 + \sqrt{\sigma_0^2 + \sigma^2/n}\, Z_1, \tag{9}$$

where Z_1 is a standard normal random variable.

The posterior distribution of θ is also normal, and it can be written that

$$\theta|\boldsymbol{y}^{obs} \sim \frac{\sigma^2/n}{\sigma_0^2 + \sigma^2/n}\theta_0 + \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2/n}\bar{y}^{obs} + \frac{\sigma_0\sigma/\sqrt{n}}{\sqrt{\sigma_0^2 + \sigma^2/n}}Z_2, \tag{10}$$

where Z_2 is a standard normal random variable and is independent of Z_1.

Conditional on θ, the replicated data $\boldsymbol{y}^{rep} = (y_1^{rep}, \ldots, y_n^{rep})$ are iid from $N(\theta, \sigma^2)$, and it can be written that

$$\bar{y}^{rep}|\theta \sim \theta + (\sigma/\sqrt{n})Z_3, \tag{11}$$

where $\bar{y}^{rep}$ is the mean of y_i^{rep}'s, and Z_3 is a standard normal random variable and is independent of Z_1 or Z_2.

The *ppp* value

Suppose that $(\theta, \boldsymbol{y}^{rep}) \sim m_{ppp}(\theta, \boldsymbol{y}^{rep}) = \pi(\theta|\boldsymbol{y}^{obs})f(\boldsymbol{y}^{rep}|\theta)$. Plugging (9) and (10) into (8) and after some calculations, we can derive that

$$\begin{aligned} D(\boldsymbol{y}^{obs}, \theta) &\sim \left\{ \frac{\sigma_0}{\sqrt{\sigma_0^2 + \sigma^2/n}}Z_2 - \frac{\sigma/\sqrt{n}}{\sqrt{\sigma_0^2 + \sigma^2/n}}Z_1 \right\}^2 \Big| Z_1 \\ &\sim \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2/n}\chi^2_{nc,1}\left(\frac{\sigma^2/n}{\sigma_0^2}Z_1^2\right), \end{aligned} \tag{12}$$

where $\chi^2_{nc,v}(\Delta)$ denotes a non-central chi-square variable with v degrees of freedom and noncentrality parameter Δ. Plugging (11) into (8), we can derive that

$$D(\boldsymbol{y}^{rep},\theta) \sim Z_3^2 \sim \chi_1^2, \tag{13}$$

where χ^2_v denotes a chi-square variable with v degrees of freedom.

Plugging (12) and (13) into (1), we can derive that

$$ppp(\boldsymbol{y}^{obs}) = \Pr\left[\chi_1^2 \geq \frac{\sigma_0^2}{\sigma_0^2+\sigma^2/n}\chi^2_{nc,1}\left(\frac{\sigma^2/n}{\sigma_0^2}Z_1^2\right)\right]. \tag{14}$$

Note that this representation is different from that in Hjort et al. (2006). When Z_1 follows the standard normal distribution (and hence $\bar{y}^{obs}$ follows its marginal distribution), the distribution of $ppp(\boldsymbol{y}^{obs})$ is apparently not uniform.

The *spp* value

Let θ^{single} denote a single posterior draw from (10), and it is then associated with a particular value of Z_2. Suppose that $(\theta,\boldsymbol{y}^{rep}) \sim m_{spp}(\theta,\boldsymbol{y}^{rep}) = \delta_{\theta^{single}}(\theta)f(\boldsymbol{y}^{rep}|\theta)$. We can similarly derive that

$$\begin{aligned} spp &= \Pr\left[\chi_1^2 \geq \left\{\frac{\sigma_0}{\sqrt{\sigma_0^2+\sigma^2/n}}Z_2 - \frac{\sigma/\sqrt{n}}{\sqrt{\sigma_0^2+\sigma^2/n}}Z_1\right\}^2 \middle| Z_1, Z_2\right] \\ &= 1 - F_{\chi_1^2}\left[\left\{\frac{\sigma_0}{\sqrt{\sigma_0^2+\sigma^2/n}}Z_2 - \frac{\sigma/\sqrt{n}}{\sqrt{\sigma_0^2+\sigma^2/n}}Z_1\right\}^2\right], \end{aligned} \tag{15}$$

where $F_{\chi_1^2}$ is the cumulative distribution function of a χ_1^2 variable. When Z_1 and Z_2 independently follow the standard normal distribution, we can easily derive that

$$\left\{\frac{\sigma_0}{\sqrt{\sigma_0^2+\sigma^2/n}}Z_2 - \frac{\sigma/\sqrt{n}}{\sqrt{\sigma_0^2+\sigma^2/n}}Z_1\right\}^2 \sim \chi_1^2.$$

Hence *spp* in (15) follows a uniform distribution.

The *cppp* value

Suppose that $\boldsymbol{y}^{rep} \sim m_{cppp}(\boldsymbol{y}^{rep}) = \int \pi^*(\theta)f(\boldsymbol{y}^{rep}|\theta)d\theta$, and the calibration prior π^* is equal to π. For calculation of $ppp(\boldsymbol{y}^{rep})$, like Z_1 in (9), Z_2 in (10) and Z_3 in (11), let $\widetilde{Z}_1, \widetilde{Z}_2$ and $\widetilde{Z}_3$ respectively denote the standard normal variables associated with $\bar{y}^{rep}$, $\theta|\boldsymbol{y}^{rep}$ and the doubly replicated data conditional on θ. Plugging (14) and a similar expression for $ppp(\boldsymbol{y}^{rep})$ into (5), we can derive that

$$\begin{aligned} cppp(\boldsymbol{y}^{obs}) &= \Pr\left\{\Pr\left[\chi_1^2 \geq \frac{\sigma_0^2}{\sigma_0^2+\sigma^2/n}\chi^2_{nc,1}\left(\frac{\sigma^2/n}{\sigma_0^2}\widetilde{Z}_1^2\right)\right]\right. \\ &\leq \left.\Pr\left[\chi_1^2 \geq \frac{\sigma_0^2}{\sigma_0^2+\sigma^2/n}\chi^2_{nc,1}\left(\frac{\sigma^2/n}{\sigma_0^2}Z_1^2\right)\right] \middle| Z_1\right\} \\ &= \Pr\left(\widetilde{Z}_1^2 \geq Z_1^2 | Z_1\right) = 1 - F_{\chi_1^2}\left(Z_1^2\right). \end{aligned} \tag{16}$$

When Z_1 follows the standard normal distribution, the distribution of $cppp(\boldsymbol{y}^{obs})$ is uniform because $Z_1^2 \sim \chi_1^2$.

3.2. The asymptotic scenario when the true unknown mean is fixed

We now consider the large sample situation when the observed data $\boldsymbol{y}^{obs}$ are iid from a distribution with unknown fixed mean θ^{tr} and known variance σ^2, and the assumed data model, π and π^* remain the same as in Section 3.1.

Suppose that it is concluded that the model conditions are not valid for the observed data if the p value in use is below the critical value 0.05. We will investigate the frequency properties of the three Bayesian p values in terms of the probability of type I error when the true data model is normal and the power when the true data model is not normal. Notice that, when $n \to \infty$, $\bar{y}^{obs} \xrightarrow{p} \theta^{tr}$, and hence $Z_1 \xrightarrow{p} (\theta^{tr}-\theta_0)/\sigma_0$ according to (9), regardless of whether the true data model is normal or not. Since the distribution of each p value based on the particular discrepancy measure in (8) depends on $\bar{y}^{obs}$ or Z_1, the asymptotic distribution of each p value remains the same regardless of whether the true data model is normal or not, and hence the probability of type I error and the power are asymptotically equal.

The *ppp* value

When $n \to \infty$, we can easily derive that

$$\frac{\sigma_0^2}{\sigma_0^2+\sigma^2/n}\chi^2_{nc,1}\left(\frac{\sigma^2/n}{\sigma_0^2}Z_1^2\right) \xrightarrow{d} \chi^2_1.$$

Hence $ppp(\boldsymbol{y}^{obs})$ in (14) converges to 0.5 in probability. Both the probability of type I error and the power are asymptotically equal to 0.

The *spp* value

When $n \to \infty$, we can easily derive that

$$\left\{\frac{\sigma_0}{\sqrt{\sigma_0^2+\sigma^2/n}}Z_2 - \frac{\sigma/\sqrt{n}}{\sqrt{\sigma_0^2+\sigma^2/n}}Z_1\right\}^2 \xrightarrow{d} \chi^2_1.$$

Therefore asymptotically $spp(\boldsymbol{y}^{obs})$ in (15) follows a uniform distribution. Both the probability of type I error and the power are asymptotically equal to 0.05.

The *cppp* value

We can easily derive that $cppp(\boldsymbol{y}^{obs})$ in (16) converges to $1 - F_{\chi^2_1}\left[(\theta^{tr}-\theta_0)^2/\sigma_0^2\right]$ in probability. Therefore, the *cppp* value being below 0.05 is equivalent to $|\theta^{tr}-\theta_0|/\sigma_0 > 1.96$. If the calibration prior is not too off for θ^{tr} such that $|\theta^{tr}-\theta_0|/\sigma_0 \le 1.96$, then we always conclude that the model conditions are valid, and both the probability of type I error and the power are asymptotically equal to 0. If $|\theta^{tr}-\theta_0|/\sigma_0 > 1.96$ instead, then both the probability of type I error and the power are asymptotically equal to 1.

To summarize, in this example, the *spp* value reflects some tradeoff between the probability of type I error and the power, whereas the *ppp* and *cppp* values do not, making the *spp* value a safer choice than the other two.

4. A hierarchical model

4.1. The model and the discrepancy measures

We now investigate the hierarchical model specified in Sinharay and Stern (2003). Assume that the data $\boldsymbol{y} = (y_1, \ldots, y_J)$ are independent normal random variables with unknown means $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_J)$ and known variance σ^2, where ξ_j's are iid from $N(\mu, \tau^2)$. Let the hyperprior for the hyperparameters μ and τ be $\pi(\mu,\tau) = \pi(\tau)\pi(\mu|\tau)$, where $\pi(\tau)$ follows a discrete uniform distribution over equispaced points from 0.004 to ϕ_τ with distance between each two adjacent points equal to 0.004, and $\pi(\mu|\tau)$ either follows a $N(\mu_0, \phi_\mu\tau^2)$ distribution or takes the improper form $\pi(\mu|\tau) \propto 1$. Sinharay and Stern (2003) used $\phi_\tau = 40$ and the improper form of $\pi(\mu|\tau)$. Here we will investigate broader scenarios for the hyperprior.

Let $\boldsymbol{\theta} = (\mu, \tau, \boldsymbol{\xi})$ denote all the parameters. The algorithm details for generating draws from the posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ are given in Appendix A. The replicated data set $\boldsymbol{y}^{rep}$ can be easily generated given $\boldsymbol{\theta}$. As in Sinharay and Stern (2003), we use the following four discrepancy measures: (1) $\bar{y}$, the mean of y_j's; (2) s_y, the standard deviation of y_j's; (3) y_{min}, the minimum of y_j's; (4) y_{max}, the maximum of y_j's. We also use a discrepancy measure $d_y = |y_{max} - y_{med}| - |y_{min} - y_{med}|$, where y_{med} is the median of y_j's. If the data model is specified correctly, y_j's are iid draws from $N(\mu, \sigma^2+\tau^2)$ and d_y should be distributed around zero.

4.2. Four simulation scenarios

Suppose that the first level model $f(\boldsymbol{y}|\boldsymbol{\xi})$ is correctly specified. We will investigate the following four simulation scenarios for the second level model for $\boldsymbol{\xi}$ and the hyperprior.

Simulation scenario 1. $f(\boldsymbol{\xi}|\mu,\tau)$ is correctly specified. In the hyperprior π, $\phi_\tau = 40$, and $\pi(\mu|\tau)$ takes the proper form with $\mu_0 = 7.9$ and $\phi_\mu = 25$. $\pi^{tr} = \pi$. The form of the calibration prior π^* is similar to that of π, with $\phi_\tau^* \in \{20,\ 40,\ 80\}$, $\mu_0^* = 7.9$ and $\phi_\mu^* \in \{5,\ 25,\ 125\}$, covering different cases when π^* is equal to, more concentrated than and less concentrated than π^{tr}.

Simulation scenario 2. $f(\boldsymbol{\xi}|\mu,\tau)$ is incorrectly specified and ξ_j's are actually iid draws from $\exp(\mu^{tr})$ with $\mu^{tr} = 7.9$ as in Sinharay and Stern (2003). π and π^* are the same as in simulation scenario 1.

Simulation scenario 3. $f(\boldsymbol{\xi}|\mu,\tau)$ is correctly specified. As in Sinharay and Stern (2003), assume that $\pi^{tr}(\mu,\tau) = \delta_{7.9}(\mu)\delta_{7.9}(\tau)$, and that in π, $\phi_\tau = 40$ and $\pi(\mu|\tau)$ follows the improper form. π^* is the same as in simulation scenario 1.

Simulation scenario 4. $f(\boldsymbol{\xi}|\mu,\tau)$ is incorrectly specified and is the same as in simulation scenario 2. π and π^* are the same as in simulation scenario 3.

We now fix $J = 100$ and $\sigma^2 = 1$. For each simulation scenario, 200 sets of observed data are generated from the true model conditions, and the three Bayesian *p* values given each data set are approximated with $M_A = M_B = M_C = 1000$.

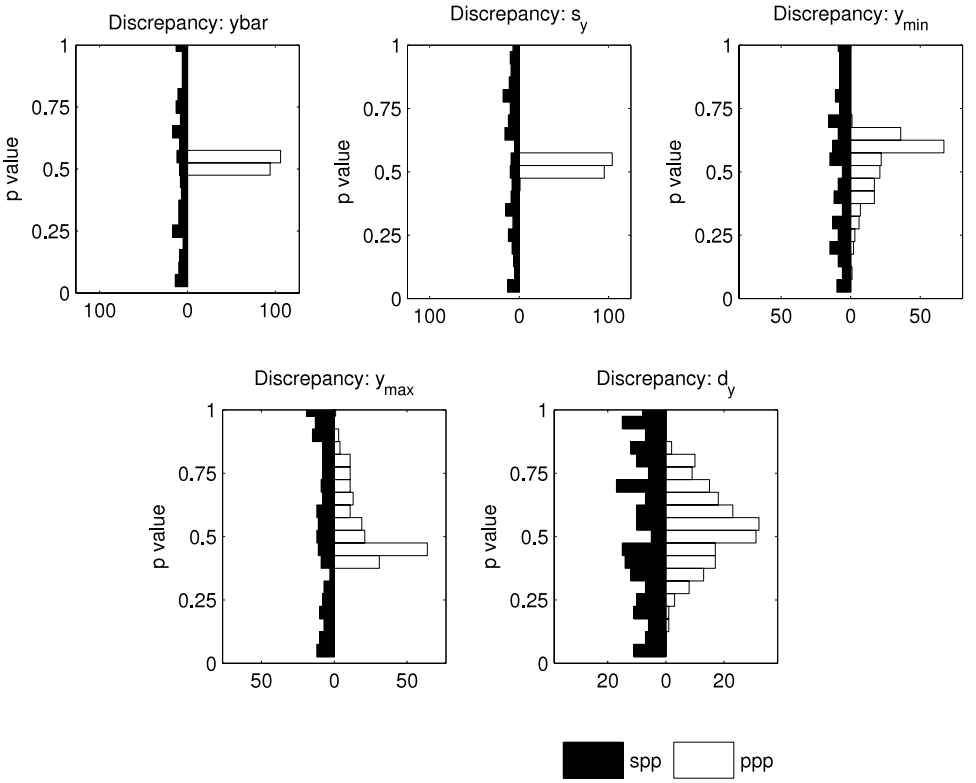

Fig. 1. Histograms of the *ppp* and *spp* values based on 200 data sets generated from the true model conditions specified in simulation scenario 3 for the hierarchical model.

Table 1
For the hierarchical model with $J = 100$ and $\sigma^2 = 1$, for five discrepancy measures and the three Bayesian p values (nine different settings for π^* in *cppp*), the entries record the probability of type I error (α) under simulation scenario 1 and the power ($1 - \beta$) under simulation scenario 2.

		ppp	*spp*	*cppp*								
				$\phi_\tau^* = 20$			$\phi_\tau^* = 40$			$\phi_\tau^* = 80$		
				ϕ_μ^*			ϕ_μ^*			ϕ_μ^*		
				5	25	125	5	25	125	5	25	125
α	$\bar{y}$	0	0.035	0.05	0.04	0.04	0.04	0.05	0.045	0.045	0.055	0.05
	s_y	0	0.055	0.035	0.04	0.06	0.05	0.04	0.05	0.085	0.055	0.035
	y_{min}	0	0.065	0.025	0.03	0.03	0.05	0.05	0.045	0.075	0.08	0.09
	y_{max}	0	0.025	0.04	0.035	0.035	0.055	0.065	0.06	0.12	0.115	0.11
	d_y	0.005	0.045	0.03	0.03	0.025	0.045	0.045	0.05	0.07	0.08	0.065
$1 - \beta$	$\bar{y}$	0	0.045	0.06	0.06	0.06	0.055	0.065	0.055	0.055	0.06	0.05
	s_y	0	0.08	0.045	0.05	0.06	0.035	0.045	0.055	0.005	0.005	0
	y_{min}	0.355	0.44	0.74	0.76	0.74	0.855	0.84	0.83	0.9	0.905	0.9
	y_{max}	0	0.07	0.005	0.015	0.02	0.11	0.14	0.2	0.59	0.62	0.64
	d_y	0.01	0.295	0.725	0.72	0.7	0.805	0.8	0.82	0.87	0.885	0.885

4.3. Simulation results and implications

Fig. 1 shows the histograms of the *ppp* and *spp* values for the 200 data sets generated from the true model conditions specified in simulation scenario 3. The *ppp* value tends to be more concentrated around 1/2 than a uniform, whereas the *spp* value is close to being uniform. Fig. 2 shows the corresponding histograms for simulation scenario 4. For the discrepancy measures $\bar{y}$, s_y and y_{max}, the *ppp* value does not show much discriminating power (i.e., being very small or very large) to indicate that the model is incorrectly specified, whereas the *spp* value has discriminating power for a small proportion of the data sets. For the discrepancy measures y_{min} and d_y, both the *ppp* and *spp* values show discriminating power for some simulated data sets, with the *spp* value showing discriminating power for larger proportion of the simulated data sets.

Supposing that it is concluded that the model conditions are not valid for the observed data if the p value in use is below 0.025 or above 0.975, Table 1 records the probability of type I error under simulation scenario 1 and the power under simulation scenario 2 for all five discrepancy measures and all three Bayesian p values. In all cases under scenario 1, the *ppp*

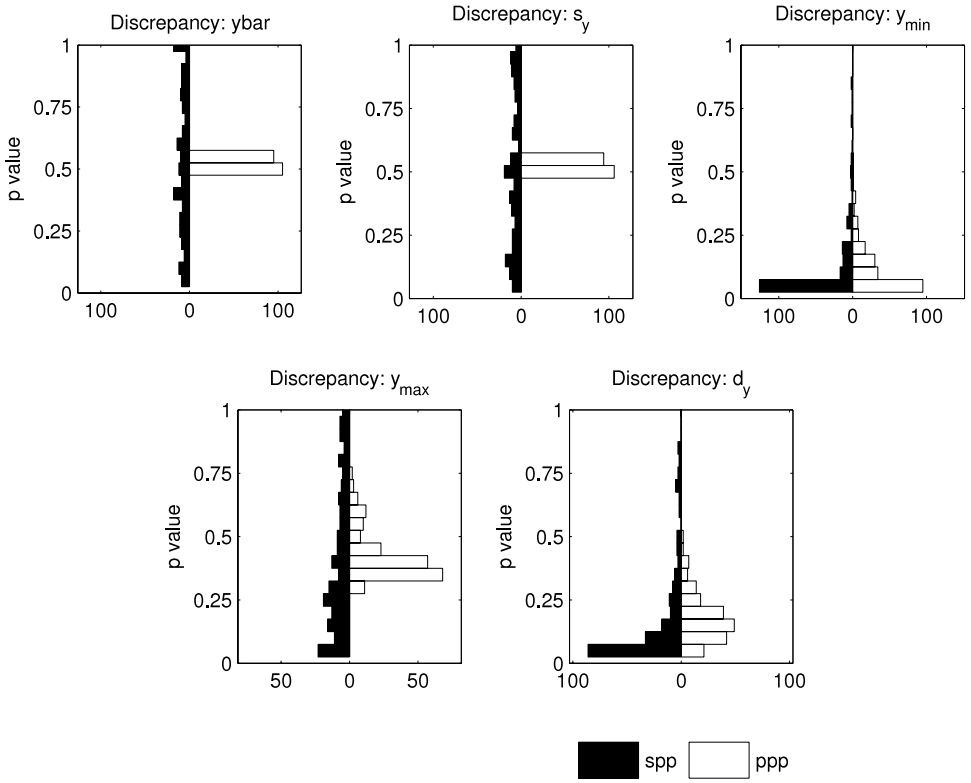

Fig. 2. Histograms of the *ppp* and *spp* values based on 200 data sets generated from the true model conditions specified in simulation scenario 4 for the hierarchical model.

Table 2
For the hierarchical model with $J = 100$ and $\sigma^2 = 1$, for five discrepancy measures and the three Bayesian p values (nine different settings for π^* in *cppp*), the entries record the probability of type I error (α) under simulation scenario 3 and the power ($1-\beta$) under simulation scenario 4.

		ppp	*spp*	*cppp*								
				$\phi_\tau^* = 20$			$\phi_\tau^* = 40$			$\phi_\tau^* = 80$		
				ϕ_μ^*			ϕ_μ^*			ϕ_μ^*		
				5	25	125	5	25	125	5	25	125
α	$\bar{y}$	0	0.05	0.05	0.06	0.055	0.045	0.055	0.05	0.05	0.055	0.055
	s_y	0	0.05	0.035	0.03	0.02	0.04	0.035	0.03	0.06	0.05	0.055
	y_{min}	0	0.06	0.005	0.005	0.005	0.015	0.025	0.02	0.115	0.09	0.115
	y_{max}	0	0.065	0.025	0.025	0.025	0.06	0.065	0.055	0.175	0.16	0.155
	d_y	0	0.03	0.015	0.015	0.01	0.06	0.065	0.04	0.105	0.105	0.11
$1-\beta$	$\bar{y}$	0	0.075	0.04	0.03	0.035	0.04	0.035	0.03	0.04	0.04	0.035
	s_y	0	0.02	0.02	0.005	0.02	0.03	0.035	0.03	0.06	0.055	0.055
	y_{min}	0.295	0.485	0.845	0.83	0.815	0.895	0.9	0.9	0.925	0.925	0.935
	y_{max}	0	0.055	0	0.005	0	0.115	0.12	0.12	0.56	0.515	0.52
	d_y	0.015	0.33	0.67	0.66	0.69	0.82	0.82	0.805	0.865	0.87	0.86

value has a probability of type I error below the nominal level of 0.05; in all cases under scenario 2, the *ppp* value has smaller power than either the *spp* value or the *cppp* value. Under scenario 1, the *spp* value has a probability of type I error around the nominal level in all cases; the *cppp* value has a probability of type I error around the nominal level when $\phi_\tau^* = 20$ or 40, but could have a probability of type I error above the nominal level when $\phi_\tau^* = 80$. Under scenario 2, the power of the *spp* value is sometimes smaller and sometimes larger than that of the *cppp* value. We also observe that the power for the *cppp* value for the same discrepancy measure y_{max} could be wildly different for different calibration priors.

Table 2 records the probability of type I error under simulation scenario 3 and the power under simulation scenario 4. The findings are rather similar to those above.

Apparently, the *ppp* value is inferior to the *spp* value. Comparison between the *spp* value and the *cppp* value, however, is a more complicated issue. When the model is incorrectly specified, the *cppp* value could have considerably higher power than the *spp* value for some discrepancy measures. However, the *cppp* value is computationally more expensive and its power could depend heavily on the calibration prior. Also, when the model is correctly specified, depending on how different π^* is

Table 3
For the hierarchical model with $J \in \{10, 50, 100\}$ and $\sigma^2 = 1$, the entries record the probability of type I error (α) under simulation scenario 3 and the power ($1-\beta$) under simulation scenario 4 for the *spp* values based on five discrepancy measures.

		$J = 10$	$J = 50$	$J = 100$
α	$\bar{y}$	0.065	0.06	0.05
	s_y	0.07	0.06	0.05
	y_{min}	0.065	0.035	0.06
	y_{max}	0.05	0.045	0.065
	d_y	0.06	0.04	0.03
$1-\beta$	$\bar{y}$	0.04	0.04	0.075
	s_y	0.035	0.06	0.02
	y_{min}	0.045	0.28	0.485
	y_{max}	0.07	0.075	0.055
	d_y	0.075	0.185	0.33

from π^{tr}, which is never known in real data analysis, the *cppp* value could have a probability of type I error larger than the nominal level. We would therefore recommend using the *spp* value as a computationally less expensive and safer choice.

In order to demonstrate how the *spp* value varies with the sample size, Table 3 compares the probability of type I error under simulation scenario 3 and the power under simulation scenario 4 for the *spp* values based on all five discrepancy measures when the sample size J takes value among 10, 50 and 100 (with the results for $J = 100$ copied from Table 2). The probability of type I error for the *spp* value is around the nominal level in all cases, whereas the power of the *spp* values based on discrepancy measures y_{min} and d_y grows larger when the sample size gets larger.

5. A causal effect model

5.1. The model and the discrepancy measures

Consider the following randomized experiment with N participants, among whom half are randomly assigned to the treatment with $Z_i = 1$, and the other half are assigned to the control with $Z_i = 0$. Suppose that those who are assigned to the control have no access to the treatment and thus can only take the control, whereas those who are assigned to the treatment may not comply to the assignment and take the control instead. This happens, for example, when the treatment represents an experimental drug and the control represents a standard drug.

We follow the potential outcome framework, also known as the Rubin Causal Model (Rubin, 1974, 1978), to formally describe the scenario. Each participant i can be potentially assigned to the treatment or the control. Let $S_i(z)$ ($z = 1$ or 0) denote the potentially observable treatment arm taken by participant i if he is assigned to treatment arm z. In our context, $S_i(0)$ must be equal to 0, whereas $S_i(1)$ can be 1 or 0. The participants can hence be classified into two principal strata (Frangakis and Rubin, 2002): compliers who would comply to the assignment, for whom $S_i(1) = 1$, and never-takers who would not take the treatment even if they are assigned to it, for whom $S_i(1) = 0$. Let $Y_i(z)$ ($z = 1$ or 0) denote the potentially observable primary outcome (e.g. change in blood pressure) for participant i if he is assigned to treatment arm z.

Since each participant is observed to be assigned to only one treatment arm Z_i, for participant i, the treatment arm actually observed to be taken is $S_i^{act} = S_i(Z_i)$, and the actually observed primary outcome is $Y_i^{act} = Y_i(Z_i)$. Note that the principal stratum is partially latent. Among participants with $Z_i = 1$, those with $S_i^{act} = 1$ are identified as compliers and those with $S_i^{act} = 0$ are identified as never-takers. However, participants with $Z_i = 0$ (and $S_i^{act} = 0$) are mixture of compliers and never-takers.

Suppose that there is a univariate covariate that follows a standard normal distribution, and let x_i denote the value of the covariate for participant i. In the data model, assume that the probability of being a complier is characterized by the following logistic model:

$$\Pr[S_i(1) = 1|x_i] = \frac{\exp(a + bx_i)}{1 + \exp(a + bx_i)}, \tag{17}$$

and that conditional on the principal stratum, $Y_i(1)$ and $Y_i(0)$ are distributed according to the following parallel regressions:

$$\begin{aligned} &Y_i(1)|S_i(1) = 1, x_i \sim N(\gamma_{c,1} + \eta x_i, \sigma^2),\\ &Y_i(0)|S_i(1) = 1, x_i \sim N(\gamma_{c,0} + \eta x_i, \sigma^2),\\ &Y_i(1)|S_i(1) = 0, x_i \sim N(\gamma_{n,1} + \eta x_i, \sigma^2)\\ \text{and } &Y_i(0)|S_i(1) = 0, x_i \sim N(\gamma_{n,0} + \eta x_i, \sigma^2). \end{aligned} \tag{18}$$

The observed data are $\{\boldsymbol{Z}, \boldsymbol{x}, \boldsymbol{S}^{act}, \boldsymbol{Y}^{act}\}$, where $\boldsymbol{Z}$ denotes the collection of Z_i, $\boldsymbol{x}$ denotes the collection of x_i, $\boldsymbol{S}^{act}$ denotes the collection of S_i^{act} and $\boldsymbol{Y}^{act}$ denotes the collection of Y_i^{act}. Let $\boldsymbol{\theta}$ denote all the parameters. Assume that the prior distributions

are $\pi(a) \sim N(0, K_0)$, $\pi(b) \sim N(0, K_0)$, $\pi(\sigma^2) \sim Inv-\chi^2(v_0, s_0^2)$ and $\pi(\gamma_{c,1}, \gamma_{c,0}, \gamma_{n,1}, \gamma_{n,0}, \eta|\sigma^2) \sim MVN(\mathbf{0}, K_0\sigma^2\boldsymbol{I}_5)$ independently, where $K_0 = 100$, $v_0 = 2$, $s_0^2 = 1.0$, $MVN(\boldsymbol{\lambda}, \boldsymbol{\kappa})$ denotes a multivariate normal distribution with mean vector $\boldsymbol{\lambda}$ and covariance matrix $\boldsymbol{\kappa}$, and $\boldsymbol{I}_5$ denotes a 5×5 identity matrix. The MCMC algorithm for generating posterior samples of $\boldsymbol{\theta}$ is detailed in Appendix B. Given $\boldsymbol{\theta}$, the replicated data are generated as follows. For participant i, Z_i and x_i are fixed as in the observed data; if $Z_i = 1$, $S_i^{act\ rep} = S_i^{rep}(1)$ and $Y_i^{act\ rep} = Y_i^{rep}(1)$ are respectively generated according to (17) and (18); if $Z_i = 0$, $S_i^{act\ rep}$ is set to be zero, and $Y_i^{act\ rep} = Y_i^{rep}(0)$ is generated according to (18).

The following six discrepancy measures are used: (1) $\bar{y}_1$, the mean of $\{Y_i^{act} : Z_i = 1\}$; (2) s_1, the standard deviation of $\{Y_i^{act} : Z_i = 1\}$; (3) $\bar{y}_0$, the mean of $\{Y_i^{act} : Z_i = 0\}$; (4) s_0, the standard deviation of $\{Y_i^{act} : Z_i = 0\}$; (5) $\bar{y}_{diff} = \bar{y}_1 - \bar{y}_0$, the average effect of treatment assignment on Y; (6) the standard error for (5), $s_{diff} = \sqrt{s_1^2/n_1 + s_0^2/n_0}$ where n_z is the number of participants satisfying $Z_i = z$ ($z = 1$ or 0).

The seventh discrepancy measure $\hat{p}_{c1}$ estimates the proportion of compliers using data on participants with $Z_i = 1$. Since among participants with $Z_i = 1$, those with $S_i^{act} = 1$ are identified as compliers, we let $\hat{p}_{c1} = n_{11}/n_1$, where n_{11} is the number of participants satisfying $Z_i = 1$ and $S_i^{act} = 1$.

The eighth discrepancy measure $\hat{p}_{c0}$ estimates the proportion of compliers using data on participants with $Z_i = 0$. Participants with $Z_i = 0$ are mixture of compliers and never-takers, and the probability of participant i being a complier can be estimated by

$$\widetilde{w}_{c:i} \equiv \Pr[S_i(1) = 1|\boldsymbol{Z}, \boldsymbol{x}, \boldsymbol{S}^{act}, \boldsymbol{Y}^{act}, \boldsymbol{\theta}] = \frac{\omega_{c:i}N_i(\gamma_{c,0} + \eta x_i, \sigma^2)}{\omega_{c:i}N_i(\gamma_{c,0} + \eta x_i, \sigma^2) + \omega_{n:i}N_i(\gamma_{n,0} + \eta x_i, \sigma^2)}, \tag{19}$$

where $\omega_{c:i} = \Pr[S_i(1) = 1|x_i]$ is given by (17), $\omega_{n:i} = 1 - \omega_{c:i}$, and $N_i(\lambda, \kappa^2)$ is the probability density function of $N(\lambda, \kappa^2)$ evaluated at Y_i^{act}. We then let $\hat{p}_{c0} = \sum_{Z_i=0} \widetilde{w}_{c:i}$.

Because the compliers can be induced by treatment assignment to change the treatment arm taken, the average treatment effect on Y for compliers is often of interest in causal inference studies. The ninth discrepancy measure $\bar{y}_{c-diff}$ estimates this effect as follows. First, the average value of $Y_i(1)$ for compliers is estimated by

$$\bar{y}_c(1) = \frac{\sum_{Z_i=1, S_i^{act}=1} Y_i^{act}}{n_{11}}.$$

Second, for each participant with $Z_i = 0$, the probability of being a complier is $w_{c:i}$, and if he is a complier, his primary outcome is expected to be $\gamma_{c,0} + \eta x_i$; therefore, the average value of $Y_i(0)$ for compliers can be estimated by

$$\bar{y}_c(0) = \frac{\sum_{Z_i=0} w_{c:i}(\gamma_{c,0} + \eta x_i)}{\sum_{Z_i=0} w_{c:i}}.$$

Third, let $\bar{y}_{c-diff} = \bar{y}_c(1) - \bar{y}_c(0)$.

Note that the eighth and ninth discrepancy measures depend not only on the data, but also on the parameters.

5.2. Two simulation scenarios

Simulation scenario 1. Assume that the data model is correctly specified, and that π^{tr} is a Dirac distribution in which the true parameter values are $a^{tr} = 0.5$, $b^{tr} = 1$, $\gamma_{c,1}^{tr} = 4$, $\gamma_{c,0}^{tr} = 2$, $\gamma_{n,1}^{tr} = \gamma_{n,0}^{tr} = 1$, $\eta^{tr} = 1$ and $\sigma^{2\ tr} = 1$. The calibration prior π^* is equal to π.

Simulation scenario 2. Assume that the data model is incorrectly specified, and the true model is similar to that specified in Section 5.1 except that $Y_i(1)$ and $Y_i(0)$ are distributed according to the following parallel regressions:

$$\begin{aligned}
&Y_i(1)|S_i(1) = 1, x_i \sim N(\gamma_{c,1} + \eta x_i + \zeta x_i^2, \sigma^2),\\
&Y_i(0)|S_i(1) = 1, x_i \sim N(\gamma_{c,0} + \eta x_i + \zeta x_i^2, \sigma^2),\\
&Y_i(1)|S_i(1) = 0, x_i \sim N(\gamma_{n,1} + \eta x_i + \zeta x_i^2, \sigma^2)\\
\text{and } &Y_i(0)|S_i(1) = 0, x_i \sim N(\gamma_{n,0} + \eta x_i + \zeta x_i^2, \sigma^2).
\end{aligned}$$

The true values of $a, b, \gamma_{c,1}, \gamma_{c,0}, \gamma_{n,1}, \gamma_{n,0}, \eta$ and σ^2 are the same as in simulation scenario 1, and the true value of ζ is $\zeta^{tr} = 1$. π and π^* remain the same as in simulation scenario 1.

For each simulation scenario, we generate 200 data sets with $N = 300$ from the true model, and approximate the three Bayesian p values given each data set with $M_A = M_B = M_C = 100$.

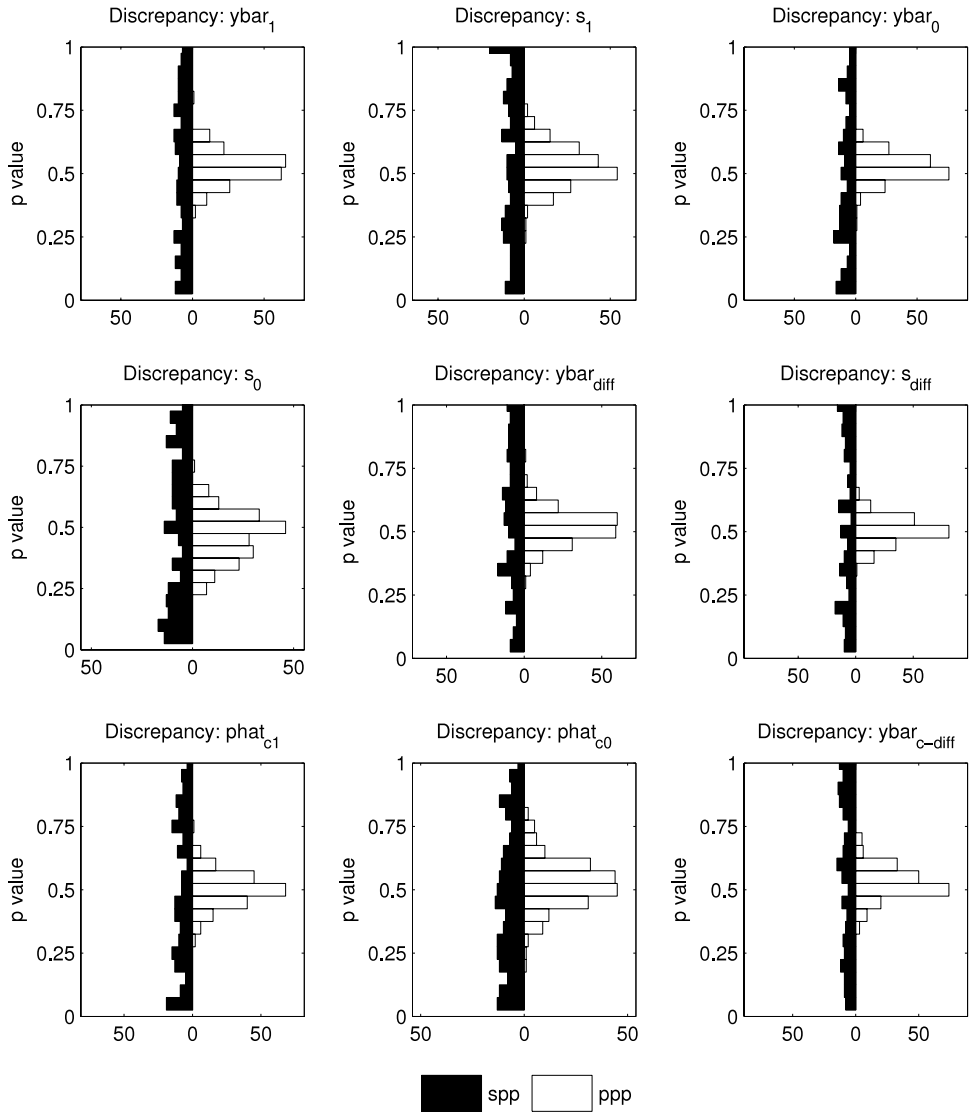

Fig. 3. Histograms of the *ppp* and *spp* values based on 200 data sets generated from the true model specified in simulation scenario 1 for the causal model.

5.3. Simulation results and implications

Figs. 3 and 4, respectively, show the histograms of the *ppp* and *spp* values for the 200 data sets generated from the true model specified in simulation scenario 1 and simulation scenario 2. Supposing that it is concluded that the model conditions are not valid for the observed data if the *p* value in use is below 0.025 or above 0.975, Table 4 records the probability of type I error under simulation scenario 1 and the power under simulation scenario 2 for all nine discrepancy measures and all three Bayesian *p* values.

The findings are similar to those in Section 4.3, except that the *cppp* value under scenario 1 has a probability of type I error smaller than the nominal level of 0.05 and that the power of the *cppp* value could be considerably less than that of the *spp* value in some cases. Note that, in this example, the generation of posterior samples of the parameters is computationally rather expensive, making the *cppp* value very computationally expensive. We would still recommend the *spp* value as a better choice.

6. Conclusion

Through an analytical example with independent observations and two simulation examples that go beyond independent observations, we have demonstrated that the *spp* value is superior to the *ppp* value in terms of both type I error and power, and that the *spp* value is safer than the *cppp* value in terms of type I error whereas the power of the *cppp* value relative to

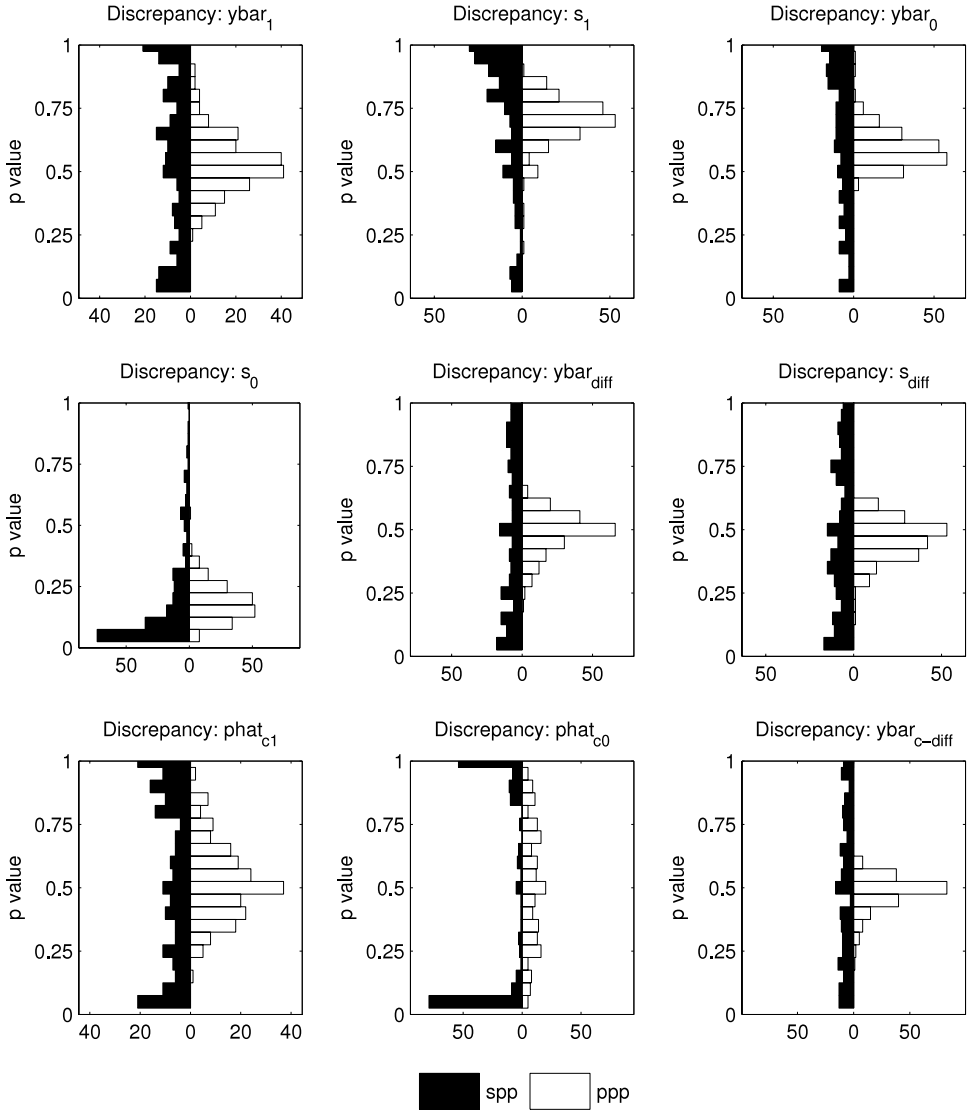

Fig. 4. Histograms of the *ppp* and *spp* values based on 200 data sets generated from the true model specified in simulation scenario 2 for the causal model.

the *spp* value largely depends on the calibration prior and the discrepancy measure. Computationally, the *spp* value is least expensive. We therefore recommend using the *spp* value.

Our results on the *ppp* value agree with those in the literature in that the *ppp* value is conservative when the data model is correctly specified and lacks power when the data model is incorrectly specified. Our results on the *spp* value extend those in the literature and suggest that even in cases with non-independent observations, the *spp* value is close to being uniform when the data model is correctly specified and has more power than the *ppp* value when the data model is incorrectly specified. We have further made contributions to the literature on investigation of the *cppp* value because the extensive study of the *cppp* value with different calibration priors and discrepancy measures has not been previously seen.

We would also like to note that, in graphical Bayesian model checking (e.g., Section 6.4 of Gelman et al., 2003), we can follow the approach adopted by the *spp* value to compare a graph of the observed data with graphs of the replicated data sets generated given a single posterior draw of the parameters.

Acknowledgement

This research is partially supported by the Key Laboratory of Mathematical Economics and Quantitative Finance (Peking University), Ministry of Education, China.

Table 4
For the causal effect model, for nine discrepancy measures and the three Bayesian p values, the entries record the probability of type I error (α) in simulation scenario 1 and power ($1-\beta$) in simulation scenario 2.

		ppp	*spp*	*cppp*
α	$\bar{y}_1$	0	0.045	0.005
	s_1	0	0.095	0
	$\bar{y}_0$	0	0.045	0
	s_0	0	0.055	0
	$\bar{y}_{diff}$	0	0.065	0
	s_{diff}	0	0.075	0
	$\hat{p}_{c1}$	0	0.045	0
	$\hat{p}_{c0}$	0	0.045	0
	$\bar{y}_{c-diff}$	0	0.055	0
$1-\beta$	$\bar{y}_1$	0	0.12	0.04
	s_1	0	0.13	0.03
	$\bar{y}_0$	0	0.115	0.02
	s_0	0	0.285	0.53
	$\bar{y}_{diff}$	0	0.085	0
	s_{diff}	0	0.07	0.015
	$\hat{p}_{c1}$	0	0.105	0.05
	$\hat{p}_{c0}$	0.01	0.57	0.155
	$\bar{y}_{c-diff}$	0	0.07	0.005

Appendix A. Simulation algorithm for the Hierarchical model

If the statistical prior $\pi(\mu|\tau)$ takes the improper form $\pi(\mu|\tau) \propto 1$, then the algorithm for simulation from the posterior distribution is as follows.

Step 1. Draw τ from the following discrete distribution over equispaced points from 0.004 to ϕ_τ with distance between each two adjacent points equal to 0.004:

$$\pi(\tau|\boldsymbol{y}) \propto V_\mu^{1/2}(\sigma^2+\tau^2)^{-J/2}\exp\left(-\frac{\sum_{j=1}^{J}(y_j-\bar{y})^2}{2(\sigma^2+\tau^2)}\right),$$

where $V_\mu = (\sigma^2+\tau^2)/J$.

Step 2. Draw μ from $\mu|\tau, \boldsymbol{y} \sim N(\bar{y}, V_\mu)$.

Step 3. For $j = 1, \ldots, J$, draw ξ_j from $\xi_j|\mu, \tau, \boldsymbol{y} \sim N(\hat{\xi}_j, V_\xi)$ where

$$\hat{\xi}_j = \frac{y_j/\sigma^2 + \mu/\tau^2}{1/\sigma^2 + 1/\tau^2} \quad \text{and} \quad V_\xi = \frac{1}{1/\sigma^2+1/\tau^2}.$$

If the statistical prior $\pi(\mu|\tau)$ takes the form $\mu|\tau \sim N(\mu_0, \phi_\mu\tau^2)$, then the simulation algorithm from the posterior distribution is as follows.

Step 1. Draw τ from the following discrete distribution over equispaced points from 0.004 to ϕ_τ with distance between each two adjacent points equal to 0.004:

$$\pi(\tau|\boldsymbol{y}) \propto \tau^{-1}\widetilde{V}_\mu^{1/2}(\sigma^2+\tau^2)^{-J/2}\exp\left(-\frac{\sum_{j=1}^{J}(y_j-\bar{y})^2}{2(\sigma^2+\tau^2)}\right)\exp\left[-\frac{1}{2}\left(\frac{\mu_0^2}{\phi_\mu\tau^2}+\frac{\bar{y}^2}{V_\mu}-\frac{\widetilde{\mu}^2}{\widetilde{V}_\mu}\right)\right],$$

where

$$\widetilde{\mu} = \frac{\mu_0/(\phi_\mu\tau^2)+\bar{y}/V_\mu}{1/(\phi_\mu\tau^2)+1/V_\mu} \quad \text{and} \quad \widetilde{V}_\mu = \frac{1}{1/(\phi_\mu\tau^2)+1/V_\mu}.$$

Step 2. Draw μ from $\mu|\tau, \boldsymbol{y} \sim N(\widetilde{\mu}, \widetilde{V}_\mu)$.

Step 3. For $j = 1, \ldots, J$, draw ξ_j from $\xi_j|\mu, \tau, \boldsymbol{y} \sim N(\hat{\xi}_j, V_\xi)$.

Note that the algorithm is not a MCMC algorithm, and all the posterior draws can be obtained in parallel.

Appendix B. Simulation algorithm for the causal effect model

Let $\boldsymbol{S}(1)$ denote the collection of $S_i(1)$ which are partially missing. Utilizing data augmentation (Tanner and Wong, 1987), the MCMC algorithm for simulation from the posterior distribution is as follows.

Step 1. Draw a and b from $f(a, b|\boldsymbol{x}, \boldsymbol{S}(1))$.

(1) Let $\mathbf{A} = (a, b)^\top$, $\mathbf{w}_i = (1, x_i)^\top$, and define

$$l(\mathbf{A}) = -\frac{\mathbf{A}^\top \mathbf{A}}{2K_0} + \sum_{S_i(1)=1} \mathbf{A}^\top \mathbf{w}_i - \sum_{i=1}^{N} \log[1 + \exp(\mathbf{A}^\top \mathbf{w}_i)].$$

Use the Newton–Raphson algorithm to find the mode $\mathbf{A}^{mode}$ for $l(\mathbf{A})$ and the curvature matrix $\boldsymbol{\Omega}$ of $l(\mathbf{A})$ at the mode.

(2) Generate $\mathbf{A}^{new}$ from a multivariate t distribution $t_4(\mathbf{A}^{mode}, \boldsymbol{\Omega})$, and update the current value of $\mathbf{A}$ to $\mathbf{A}^{new}$ with probability given by the Metropolis–Hastings ratio.

Step 2. Generate σ^2 from $f(\sigma^2|\boldsymbol{Z}, \boldsymbol{x}, \boldsymbol{Y}^{act}, \boldsymbol{S}(1))$. Define

$$\begin{aligned} R_{c,1} &= \{i:\ Z_i = 1 \text{ and } S_i(1) = 1\}, \\ R_{c,0} &= \{i:\ Z_i = 0 \text{ and } S_i(1) = 1\}, \\ R_{n,1} &= \{i:\ Z_i = 1 \text{ and } S_i(1) = 0\} \\ \text{and } R_{n,0} &= \{i:\ Z_i = 0 \text{ and } S_i(1) = 0\}. \end{aligned}$$

Let $N_{c,1}$, $N_{c,0}$, $N_{n,1}$ and $N_{n,0}$ respectively denote the number of units in $R_{c,1}$, $R_{c,0}$, $R_{n,1}$ and $R_{n,0}$. Define $\boldsymbol{\Psi} = (\gamma_{c,1}, \gamma_{c,0}, \gamma_{n,1}, \gamma_{n,0}, \eta)^\top$. Define

$$\boldsymbol{\Sigma}_{\boldsymbol{\Psi}} = \left\{ \frac{1}{K_0}\boldsymbol{I}_5 + \begin{bmatrix} N_{c,1} & 0 & 0 & 0 & \sum_{i\in R_{c,1}} x_i \\ 0 & N_{c,0} & 0 & 0 & \sum_{i\in R_{c,0}} x_i \\ 0 & 0 & N_{n,1} & 0 & \sum_{i\in R_{n,1}} x_i \\ 0 & 0 & 0 & N_{n,0} & \sum_{i\in R_{n,0}} x_i \\ \sum_{i\in R_{c,1}} x_i & \sum_{i\in R_{c,0}} x_i & \sum_{i\in R_{n,1}} x_i & \sum_{i\in R_{n,0}} x_i & \sum_{i=1}^{N} x_i^2 \end{bmatrix} \right\}^{-1}$$

and

$$\widetilde{\boldsymbol{\Psi}} = \boldsymbol{\Sigma}_{\boldsymbol{\Psi}} \left(\sum_{i\in R_{c,1}} Y_i^{act}, \sum_{i\in R_{c,0}} Y_i^{act}, \sum_{i\in R_{n,1}} Y_i^{act}, \sum_{i\in R_{n,0}} Y_i^{act}, \sum_{i=1}^{N} x_i Y_i^{act} \right)^\top.$$

Draw σ^2 from

$$\sigma^2 \sim Inv - \chi^2 \left(v_0 + N,\ \frac{v_0 s_0^2 + \sum_{i=1}^{N} (Y_i^{act})^2 - \widetilde{\boldsymbol{\Psi}}^\top \boldsymbol{\Sigma}_{\boldsymbol{\Psi}}^{-1} \widetilde{\boldsymbol{\Psi}}}{v_0 + N} \right).$$

Step 3. Generate $\boldsymbol{\Psi}$ from $MVN(\widetilde{\boldsymbol{\Psi}}, \boldsymbol{\Sigma}_{\boldsymbol{\Psi}}\sigma^2)$.

Step 4. Generate $\boldsymbol{S}(1)$ from $f(\boldsymbol{S}(1)|\boldsymbol{Z}, \boldsymbol{x}, \boldsymbol{Y}^{act}, \boldsymbol{S}^{act}, \boldsymbol{\theta})$. For participants with $Z_i = 1$, $S_i(1) = S_i^{act}$. For participants with $Z_i = 0$, $S_i(1)$ is generated from a Bernoulli distribution with probability of $S_i(1) = 1$ given by (19) in the text.

For approximation of the *ppp* value, we run the MCMC chain for a total of 4000 draws, use the first 2000 draws for burnin, and keep every 20th draws afterwards to get $M_A = 100$ draws. For approximation of the *spp* value, we use the first draw after burnin as $\boldsymbol{\theta}^{single}$.

References

Bayarri, M.J., Berger, J.O., 2000. P values in composite null models. J. Amer. Statist. Assoc. 95, 1127–1142.
Bayarri, M.J., Berger, J.O., 2004. The interplay of Bayesian and frequentist analysis. Statist. Sci. 19, 58–80.
Bayarri, M.J., Castellanos, M.E., 2007. Bayesian checking of the second levels of Hierarchical models. Statist. Sci. 22, 322–343.
Box, G.E.P., 1980. Sampling and Bayes' inference in scientific modelling and robustness. J. Roy. Statist. Soc. A 143, 383–430.
Dey, D.K., Gelfand, A.E., Swartz, T.B., Vlachos, P.K., 1998. A simulation-intensive approach for checking Hierarchical models. Test 7, 325–346.
Frangakis, C.E., Rubin, D.B., 2002. Principal stratification in causal inference. Biometrics 58, 21–29.
Gelman, A., Carlin, A.B., Stern, H.S., Rubin, D.B., 2003. Bayesian Data Analysis, Second ed. Chapman and Hall/CRC.
Gelman, A., Meng, X.L., Stern, H.S., 1996. Posterior predictive assessment of model fitness via realized discrepancies (with discussion). Statist. Sinica 6, 733–807.
Gelman, A., Van Mechelen, I., 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.
Gosselin, F., 2011. A new calibrated Bayesian internal goodness-of-fit method: sampled posterior p-values as simple and general p-values that allow double use of the data. PloS one 6.
Hjort, N.L., Dahl, F.A., Steinbakk, G.H., 2006. Post-processing posterior predictive p values. J. Amer. Statist. Assoc. 101, 1157–1174.
Johnson, V.E., 2004. A Bayesian chi(2) test for goodness-of-fit. Ann. Statist. 32, 2361–2384.
Johnson, V.E., 2007. Bayesian model assessment using pivotal quantities. Bayesian Anal. 2, 719–734.
Meng, X.L., 1994. Posterior predictive p values. Ann. Statist. 15, 1142–1160.
Robins, J.M., van der Vaart, A., Ventura, V., 2000. Asymptotic distribution of p values in composite null models (with discussion). J. Amer. Statist. Assoc. 95, 1143–1156.
Rubin, D.B., 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. J. Educ. Phychol. 66, 688–701.
Rubin, D.B., 1978. Bayesian inference for causal effects: the role of randomization. Ann. Statist. 6, 34–59.
Rubin, D.B., 1984. Bayesianly justifiable and relevant frequency calculations for the applied statistician. Ann. Statist. 12, 1151–1172.
Sinharay, S., Stern, H.S., 2003. Posterior predictive model checking in hierarchical models. J. Statist. Plann. Inference 111, 209–221.
Tanner, M.A., Wong, W.H., 1987. The calculation of posterior distributions by data augmentation (with discussion). J. Amer. Statist. Assoc. 82, 805–811.