Holdout Predictive Checks for Bayesian Model Criticism

Gemma E. Moran[1], David M. Blei[2], and Rajesh Ranganath[3]

[1]*Department of Statistics; Rutgers University*
[2]*Department of Computer Science and Department of Statistics; Columbia University*
[3]*Department of Computer Science and Center for Data Science; New York University*

Abstract

Bayesian modeling helps applied researchers articulate assumptions about their data and develop models tailored for specific applications. Thanks to good methods for approximate posterior inference, researchers can now easily build, use, and revise complicated Bayesian models for large and rich data. These capabilities, however, bring into focus the problem of model criticism. Researchers need tools to diagnose the fitness of their models, to understand where they fall short, and to guide their revision. In this paper we develop a new method for Bayesian model criticism, the holdout predictive check (HPC). HPCs are built on posterior predictive checks (PPCs), a seminal method that checks a model by assessing the posterior predictive distribution on the observed data. However, PPCs use the data twice—both to calculate the posterior predictive and to evaluate it—which can lead to uncalibrated p-values. HPCs, in contrast, compare the posterior predictive distribution to a draw from the population distribution, a heldout dataset. This method blends Bayesian modeling with frequentist assessment. Unlike the PPC, we prove that the HPC is properly calibrated. Empirically, we study HPCs on classical regression, a hierarchical model of text data, and factor analysis.

1 Introduction

Thanks to good algorithms for approximate Bayesian inference and good software for general Bayesian modeling, statisticians can now explore and develop a large variety of Bayesian models for a given problem. This new ease with which we can model data has turned the practice of Bayesian modeling into an iterative cycle (Blei, 2014; Gelman et al., 1995, 2020). More complex models are constructed from simpler ones, and we can evaluate earlier model "drafts" to guide the structure of subsequent revisions. But this iterative process for model-building brings into sharp focus the problem of *model criticism*. How do we navigate the space of models we can use for a problem? How do we decide when a model needs to change? In this paper, we develop the holdout predictive check (HPC), a new method of model criticism.

One of the key tools for Bayesian model criticism is the posterior predictive check (PPC) (Guttman,

1967; Rubin, 1984). In a PPC, we locate the observed data within the posterior predictive distribution, a reference distribution that is determined by the model under consideration. The spirit of a PPC is to formalize the following heuristic: "If my model is good, then its posterior predictive distribution will generate data that looks like my observations (filtered through a diagnostic function)."

Consider an observed dataset $\boldsymbol{y}^{\text{obs}}$ and a model with latent variables $\boldsymbol{\theta}$,

$$\text{p}(\boldsymbol{\theta}, \boldsymbol{y}^{\text{obs}}) = \text{p}(\boldsymbol{\theta})\text{p}(\boldsymbol{y}^{\text{obs}} \mid \boldsymbol{\theta}). \tag{1}$$

The prior is $\text{p}(\boldsymbol{\theta})$; the likelihood is $\text{p}(\boldsymbol{y}^{\text{obs}} \mid \boldsymbol{\theta})$. The model and data combine to form the posterior $\text{p}(\boldsymbol{\theta} \mid \boldsymbol{y}^{\text{obs}})$, which helps form the posterior predictive distribution,

$$\text{p}(\boldsymbol{y}^{\text{rep}} \mid \boldsymbol{y}^{\text{obs}}) = \int \text{p}(\boldsymbol{y}^{\text{rep}} \mid \boldsymbol{\theta})\text{p}(\boldsymbol{\theta} \mid \boldsymbol{y}^{\text{obs}}) \, \text{d}\boldsymbol{\theta}. \tag{2}$$

Here the variable $\boldsymbol{y}^{\text{rep}}$ denotes *replicated data*.

To implement a PPC, we first choose a *diagnostic statistic* $d(\boldsymbol{y})$, a way to measure discrepancy between a dataset $\boldsymbol{y}$ and the model. An example diagnostic is the average squared distance to the posterior mean, $d(\boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \text{E}\,[Y \mid \boldsymbol{y}])^2$. The PPC then locates $d(\boldsymbol{y}^{\text{obs}})$ in the *reference distribution* of $d(\boldsymbol{y}^{\text{rep}})$, the posterior predictive from Equation 2. One common approach is with a posterior predictive p-value,

$$p_{\text{ppc}} = \text{p}(d(\boldsymbol{y}^{\text{rep}}) \geq d(\boldsymbol{y}^{\text{obs}}) \mid \boldsymbol{y}^{\text{obs}}). \tag{3}$$

Bayarri and Morales (2003) discusses other ways to measure surprise, and Gelman et al. (1996); Gelman (2004) discuss visual approaches to checking the model.

But there is a crucial issue with the PPC—it uses the data twice. The data are first used to construct the reference distribution of the diagnostic, the posterior predictive distribution of Equation 2. The data are then used again in the observed diagnostic $d(\boldsymbol{y}^{\text{obs}})$, which is located within the reference, such as with the p-value of Equation 3. Essentially, a PPC checks how "close" $d(\boldsymbol{y}^{\text{obs}})$ is to $d(\boldsymbol{y}^{\text{rep}})$, a quantity that also depends on $\boldsymbol{y}^{\text{obs}}$. The issue is that they can be close regardless of whether the model is correct.

This paper introduces a solution to this "double use of the data" problem. The holdout predictive check (HPC) is a new method for Bayesian model criticism, one that combines the Bayesian PPC with the frequentist idea of the population distribution. The premise of the HPC is this: "If my model is good, then data drawn from the posterior predictive distribution will look like a draw from *the population distribution*."

Along with the data, model, and posterior predictive, consider new data drawn from the true population distribution $\boldsymbol{y}^{\text{new}} \sim F$. The HPC locates $d(\boldsymbol{y}^{\text{new}})$ in the posterior predictive distribution of $d(\boldsymbol{y}^{\text{rep}})$. As a p-value, a HPC is

$$p_{\text{hpc}} = \text{p}(d(\boldsymbol{y}^{\text{rep}}) \geq d(\boldsymbol{y}^{\text{new}}) \mid \boldsymbol{y}^{\text{obs}}, \boldsymbol{y}^{\text{new}}), \tag{4}$$

where $\boldsymbol{y}^{\text{rep}} \sim \text{p}(\boldsymbol{y}^{\text{rep}} \mid \boldsymbol{y}^{\text{obs}})$. This quantity no longer uses the data twice.

To implement a HPC, we split the data into $\boldsymbol{y} = (\boldsymbol{y}^{\text{obs}}, \boldsymbol{y}^{\text{new}})$, assuming $\boldsymbol{y}^{\text{new}}$ is an independent draw from the population distribution. We then calculate Equation 4. As with the PPC, the modeler can use other measures of surprise or visual checks. Figure 1 displays a schematic which relates the HPC to both the prior predictive check (Box, 1980) and the PPC (Guttman, 1967; Rubin, 1984).

Why should a modeler avoid the double use of the data? Why prefer the HPC of Equation 4 to the PPC of Equation 3? We can understand the consequences of this practice by examining some of the frequentist properties of a predictive check. Suppose we repeatedly sample an observed dataset, and then check a model; note this is a frequentist situation. We define a model to be "correct" when its posterior predictive distribution is equal to (or approaches) the distribution of the observations; this is the conceit for both a PPC and a HPC.

Now consider the sampling distribution of the PPC p-value. There are two consequences of its double use of the data. First, this p-value may result in an incorrect model not being rejected; in the language of testing, it can suffer from low *power*. Second, this p-value may not control the type I error; it may not be *calibrated*. We will show below that the HPC p-value does not suffer from these issues. Theoretically, under certain assumptions, we prove that it is calibrated. In the infinite data limit, and in a specific setup, the HPC rejects an incorrect model with probability one, while calibrated versions of the PPC (Robins et al., 2000; Hjort et al., 2006) fail to reject the incorrect model. Empirically, we study the HPC in several modeling situations. We find that it is better calibrated than the PPC and that it has higher power—a HPC more easily detects an incorrect model than either a PPC or calibrated PPCs (Robins et al., 2000; Hjort et al., 2006).

The paper is organized as follows. Section 1.1 discusses the historical development of Bayesian model evaluation and how HPCs fit in. Section 2 develops the HPC and provides an illustrative example. Section 3 proves that the HPC p-value is calibrated. Section 4 illustrates how the "double use of the data" affects the power of PPCs, unlike the HPC. Further, calibrating PPCs does not resolve this power issue. Section 5 demonstrates the HPC empirically on a regression model, a hierarchical document model, and models for factor analysis. Section 6 concludes the paper.

1.1 Related work

This work builds on predictive checks, which are part of the larger literature on Bayesian model criticism. Predictive checks locate the observations in a model-based distribution of data, the reference distribution. A brief history: Inspired by the earlier ideas of Geisser (1975), Box (1980) used the prior predictive distribution as the reference. This is a prior predictive check, which is useful for checking the conflict between prior and likelihood (Evans and Moshonov, 2006). Later, Rubin (1984) mimicked Box's framework, but replaced the prior predictive with the posterior predictive; this strategy is both more practical for diagnosing models and one that is, in Rubin's language, "Bayesianly justifiable." Guttman (1967) proposed the same approach. Finally, Gelman et al. (1996) showed how to develop diagnostic functions of the data—termed realized discrepancy functions—that depend on both a data set and the latent variables.

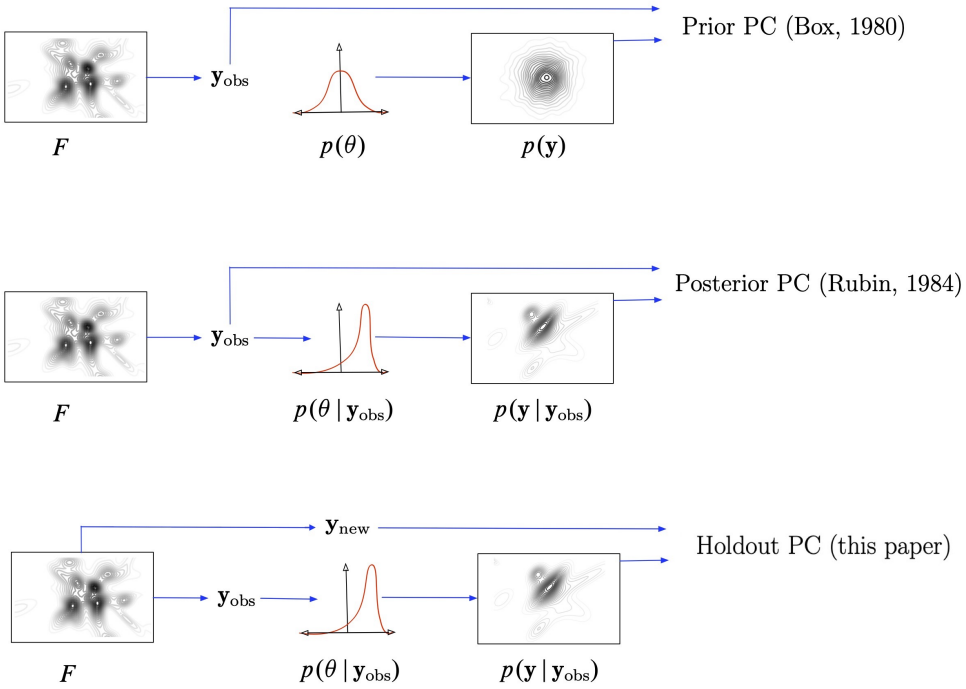


Figure 1: A schematic diagram that relates Box (1980), Rubin (1984), and this paper. This diagram posits the data come from an unknown population distribution. Box (1980) uses the data as reference in the marginal distribution induced by the model (top); Rubin (1984) uses the data as reference in the posterior predictive distribution induced by the model (middle); our method uses a draw from the population as a reference in the posterior predictive distribution (bottom).

To resolve the "double use" problem, researchers have proposed a range of strategies. One strand of research proposes to calibrate a posterior predictive check (PPC) p-value post-hoc by using its empirical distribution (Robins et al., 2000; Hjort et al., 2006). However, calibration by itself does not necessarily improve the power of the check, as we will show in Sections 4 and 5.

Alternatively, Bayarri and Berger (2000) proposes the partial predictive check. The partial predictive check calculates a predictive reference distribution that does not depend on the diagnostic, thereby removing the correlation between the diagnostic and the reference. Consequently, the partial predictive check is calibrated (Bayarri and Berger, 2000; Robins et al., 2000). Finally, Johnson (2007) proposed to use pivotal diagnostics to ensure the diagnostic and reference distribution are uncorrelated.

The holdout predictive check (HPC) can be viewed as a prior predictive check where the prior is updated based on a subset of the data. Note that this idea of using data-dependent priors was also applied to the idea of intrinsic Bayes factors (Berger and Pericchi, 1996). The intrinsic Bayes factor first calculates a posterior using only a subset of the data. This posterior is then treated as a prior, and the Bayes factor is calculated using the remaining data. In a similar way, the HPC first calculates the posterior given the observed data, and then treats this posterior as a prior in a prior predictive check of the new data.

The HPC also has close connections to the partial predictive check (Bayarri and Berger, 2000). Specifically, when the partial PC diagnostic uses only a subset of the data, the partial PC can coincide with the HPC (see Appendix D for an example). In certain cases, a partial posterior check which computes the diagnostic with only a subset of the data may be more efficient than a HPC. This efficiency is because a partial predictive check may only remove a sufficient statistic of this subset of the data, instead of the entire subset as in the HPC. A drawback of the partial predictive check, however, is that it can be difficult to calculate, and requires re-calculation for each diagnostic function. Meanwhile, a HPC is simple to implement, and the inferred posterior can be used to check many different diagnostic functions. A further limitation of the partial predictive check is that it can revert to the prior predictive check when the diagnostic contains the sufficient statistics of the model (see Appendix D for an example).

Another closely related work is Gelfand et al. (1992), which develops cross-validated checks. A cross-validated check iteratively holds out each data point, conditioning on the remaining data, and compares samples from the corresponding posterior predictive distribution to the held-out point. Similar strategies are discussed in Draper (1996); Marshall and Spiegelhalter (2003); Larsen and Lu (2007). In its relation to these other data-split checks, this paper provides a theoretical understanding and empirical evaluation for this class of methods.

In an independent and concurrent paper, Li and Huggins (2022) empirically shows that a previous definition of the HPC (POP-PC, see Appendix B for details) is not calibrated, and proposes the split predictive check (SPC). In revising this paper, we proved the POP-PC is not calibrated, which we have resolved by updating it to the HPC in Equation 5. This HPC is

the same as the single SPC of Li and Huggins (2022), and the proof here that it is calibrated (Theorem 1) is also similar to the proof of calibration in Li and Huggins (2022) for the SPC (their Theorem 3.1(1) for the case where the model is true). (The proofs are similar because both build on the work of Robins et al. (2000).)

Li and Huggins (2022) and this paper present similar methods, but complementary perspectives. Li and Huggins (2022) shows the SPC has asymptotic power of 1 under moderate-to-major model misspecification. It also proposes the divided SPC, which considers SPC p-values for multiple different splits of the data. In this paper, we examine the "double use of the data" problem of the PPC in greater detail. Specifically, we illustrate that post-hoc empirical calibration procedures for the PPC, while producing uniform p-values under the null hypothesis, do not resolve the "double use of the data" problem in terms of detecting model misspecification. Finally, we consider different diagnostics from Li and Huggins (2022) - the χ^2 diagnostic and latent Dirichlet allocation log-likelihood - and we provide empirical evidence that they are calibrated; these diagnostics are not covered by the calibration theory for the HPC, nor the SPC.

Finally, the HPC relates to metrics which assess generalization error, though it is also distinct from these ideas. Assuming a stationary data generating process, the HPC asks: does my model produce data that "looks like" future independent draws from this process? This binary outcome is in contrast to continuous metrics, such as the widely applicable information criterion (WAIC, Watanabe, 2009, 2010) or the deviance information criterion (DIC, Spiegelhalter et al., 2002). While the WAIC and DIC can be used to select from multiple models, they do not assess the adequacy of a single model, which is of interest in this work.

2 Holdout Predictive Checks

The holdout predictive check (HPC) checks a Bayesian model by considering a true population distribution: "If my model is good then data drawn from the posterior predictive distribution will look like a draw from the true population (filtered through a diagnostic function)."

The ingredients of a HPC are observed data $\boldsymbol{y}^{\text{obs}}$, replicated data $\boldsymbol{y}^{\text{rep}}$ from the posterior predictive distribution (Equation 2), and new data $\boldsymbol{y}^{\text{new}}$, drawn from the true population distribution F. As for a posterior predictive check (PPC), each check involves a diagnostic statistic $d(\boldsymbol{y})$, which measures misfit between $\boldsymbol{y}$ and the model. The HPC uses new data $\boldsymbol{y}^{\text{new}}$ to check if a draw from the population F is close to the posterior predictive distribution $\text{p}(\boldsymbol{y}^{\text{rep}} \mid \boldsymbol{y}^{\text{obs}})$, in terms of the diagnostic. If so, then the posterior predictive captures the data well, and the model passes the check.

Definition 1 (Holdout predictive check, HPC) *Consider observed data $\boldsymbol{y}^{\text{obs}}$, its posterior predictive distribution $\text{p}(\boldsymbol{y}^{\text{rep}} \mid \boldsymbol{y}^{\text{obs}})$, and a diagnostic statistic $d(\boldsymbol{y})$. Suppose we have $\boldsymbol{y}^{\text{new}}$ drawn from the population distribution of the data. As a p-value, the holdout predictive*

check is:

$$p_{\text{hpc}} = \text{p}(d(\boldsymbol{y}^{\text{rep}}) \geq d(\boldsymbol{y}^{\text{new}}) \mid \boldsymbol{y}^{\text{obs}}, \boldsymbol{y}^{\text{new}}), \tag{5}$$

where $\boldsymbol{y}^{\text{rep}} \sim \text{p}(\boldsymbol{y}^{\text{rep}} \mid \boldsymbol{y}^{\text{obs}})$.

To implement a HPC, we split the data into $\boldsymbol{y} = (\boldsymbol{y}^{\text{obs}}, \boldsymbol{y}^{\text{new}})$ and calculate Equation 5.

The diagnostic is a function of the data that measures model misfit. For example, one diagnostic is the conditional negative log-likelihood,

$$d(\boldsymbol{y}) \triangleq -\frac{1}{n} \sum_{i=1}^{n} \log \text{p}(y_i \mid \boldsymbol{y}^{\text{obs}}). \tag{6}$$

In the context of a PPC, this diagnostic is discussed in Lewis and Raftery (1996).

Algorithm 1: Holdout predictive check

---

input: data $\boldsymbol{y} = \{\boldsymbol{y}^{\text{obs}}, \boldsymbol{y}^{\text{new}}\}$, diagnostic $d(\cdot)$, # replicates R
output: holdout predictive check p-value
for $r = 1, \ldots, R$ **do**
  draw samples from the posterior $\theta_r \sim \text{p}(\theta|\boldsymbol{y}^{\text{obs}})$;
  draw posterior predictive data $\boldsymbol{y}_r^{\text{rep}} \sim \text{p}(\boldsymbol{y}^{\text{rep}}|\theta)$;

compute the empirical HPC p-value

$$p_{\text{hpc}} = \frac{1}{R} \sum_{r=1}^{R} 1 \left[d(\boldsymbol{y}_r^{\text{rep}}, \theta_r) > d(\boldsymbol{y}^{\text{new}}, \theta_r) \right];$$

return p_{hpc}

Meng (1994); Gelman et al. (1996) discuss realized diagnostics $d(\boldsymbol{y}, \theta)$, those that also depend on the latent variables. A realized diagnostic measures the strength of the connection between latent variables and a data set. An example is the negative log-likelihood

$$d(\boldsymbol{y}, \theta) \triangleq -\log \text{p}(\boldsymbol{y} \mid \theta). \tag{7}$$

Consider a joint distribution of latent variables and data,

$$\text{p}(\theta, \boldsymbol{y}) = \text{p}(\theta)\text{p}(\boldsymbol{y} \mid \theta) \tag{8}$$

and a realized diagnostic $d(\theta, \boldsymbol{y})$. The HPC is

$$p_{\text{hpc}} = \text{p}(d(\theta, \boldsymbol{y}^{\text{rep}}) > d(\theta, \boldsymbol{y}^{\text{new}}) \mid \boldsymbol{y}^{\text{obs}}, \boldsymbol{y}^{\text{new}}). \tag{9}$$

This probability is under a distribution that draws the latent variable from the posterior and the replicated data from the likelihood given the latent variable,

$$\text{p}(\theta, \boldsymbol{y}^{\text{rep}} \mid \boldsymbol{y}^{\text{obs}}) = \text{p}(\theta \mid \boldsymbol{y}^{\text{obs}})\text{p}(\boldsymbol{y}^{\text{rep}} \mid \theta). \tag{10}$$

The HPC procedure is detailed in Algorithm 1.

2.1 Example

To illustrate the HPC, we now apply a ridge regression model to synthetic data. The model is:

$$\boldsymbol{\theta} \sim \text{Normal}(0, c\boldsymbol{I}_p), \tag{11}$$

$$\sigma^2 \sim \text{Inverse-Gamma}(1/2, 1/2), \tag{12}$$

$$y_i|\boldsymbol{\theta}, \sigma^2 \overset{iid}{\sim} \text{Normal}(\boldsymbol{\theta}^\top \boldsymbol{x}_i, \sigma^2), \quad i = 1, \ldots, n. \tag{13}$$

We take $n = 50$ and $p = 100$. The covariates, $\boldsymbol{x}_i$, are drawn as uniform random variables on $[0, 1]$. Meanwhile, the true coefficients, θ, have five entries equal to 3.5 and the remaining 95 entries drawn from a standard normal distribution. Consequently, the coefficients θ have many entries close to zero, and a few large entries. The true $\sigma^2 = 1$ (note that σ^2 is treated as unknown during inference, however).

What is the impact of the prior variance parameter c on the adequacy of the model? If c is too large, the estimated coefficients will overfit to the data and not generalize well on new data. Here, we show that a HPC can detect this poor model fit while a PPC cannot. Note that we are not advocating to use the HPC to choose c; we are merely emphasizing that in the well-studied setting of ridge-regression, the HPC can detect poor model fit while the PPC cannot.

To check the model, we use the χ^2 diagnostic function:

$$d(\boldsymbol{y}, \boldsymbol{\theta}, \sigma^2) = \frac{1}{N} \sum_{i=1}^{N} \frac{(y_i - \text{E}[y_i|\boldsymbol{\theta}, \sigma^2])^2}{\text{Var}(y_i|\boldsymbol{\theta}, \sigma^2)}, \tag{14}$$

which is a sum of standardized residuals. We conduct posterior inference using a Gibbs sampler.

Figure 2 shows the HPC and PPC p-values for different values of c. Also plotted is the mean squared error:

$$\text{MSE} = \mathbb{E}\left[(\theta_0 - \theta)^2\right] \tag{15}$$

where the expectation is with respect to the posterior $\text{p}(\theta|\boldsymbol{y}, \boldsymbol{X})$ and θ_0 is the true value of the coefficents.

When c is small, there is more regularization of the coefficient estimates. This regularization prevents overfitting and results in small mean squared error. For these small values of c, both the PPC and HPC both correctly retain the model. When c is large, however, there is less regularization of the coefficients. Consequently, the model overfits to the data and the MSE is large. For these large values of c, the HPC correctly rejects the model. The PPC, however, does not reject the model; the PPC does not detect overfitting, unlike the HPC.

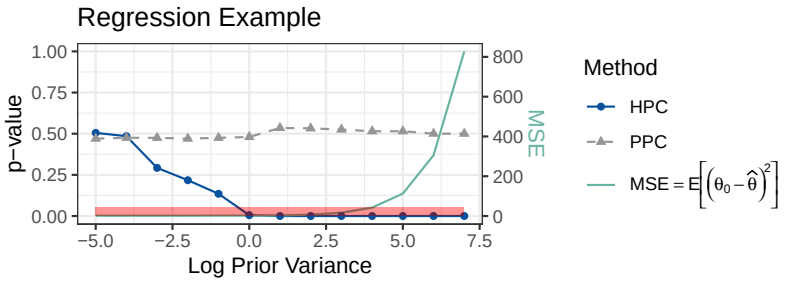


Figure 2: When the log prior variance c is small, both the HPC and PPC correctly retain the model; when c is large, the HPC correctly rejects the model, while the PPC incorrectly retains the model.

3 The asymptotic distribution of HPC p-values

In this section, we study the holdout predictive check (HPC) p-value asymptotic sampling distribution. If a p-value is uniformly distributed when the model is correct, the p-value is said to be *calibrated*. For a class of diagnostic functions, we prove that the HPC p-value is calibrated (Theorem 1).

Calibration is a frequentist property, not a Bayesian one. Although the HPC checks Bayesian models, it is still important to determine if its p-values are calibrated. Calibration helps us interpret a p-value: If the distribution of p-values is uniform, a p-value of 0.4 would not be surprising, while if the distribution was concentrated around 0.5, the value 0.4 would be surprising. For a calibrated check, if we decide to reject a model when its p-value is less than α, then the correct model will only fail such a check with that probability.

Before stating our HPC calibration result, we introduce some notation. The data are $\boldsymbol{y} = (y_1, \ldots, y_n)$ where y_i are mutually independent random variables with density $f(y; \theta)$ where $\theta \in \Theta \subset \mathbb{R}^p$.

We prove the HPC is calibrated for the same class of diagnostic functions $d(\boldsymbol{y})$ that are considered by Robins et al. (2000). Specifically, we consider diagnostic functions $d(\boldsymbol{y})$ that are asymptotically normal with asymptotic mean $\nu(\theta)$ and asymptotic variance $\sigma^2(\theta)$ when the model is correct (i.e. the density of $\boldsymbol{y}$ is $f(\boldsymbol{y}; \theta)$):

$$n^{1/2}\left[\frac{d(\boldsymbol{y}) - \nu(\theta)}{\sigma(\theta)}\right] \rightsquigarrow N(0, 1), \tag{16}$$

where $\rightsquigarrow$ denotes convergence in distribution.

Theorem 1 proves that the HPC p-values are asymptotically uniform when the model is correct. Unlike the posterior predictive check (PPC), the HPC p-values are calibrated. Note our result relies on standard regularity conditions that are detailed in Appendix A.1.

Theorem 1 *Assume Equation 16 holds and assume the regularity conditions detailed in Appendix A.1. Under the distribution $f(\boldsymbol{y}; \theta_0)$, the HPC p-value can be written as:*

$$p_{\text{hpc}}(\boldsymbol{y}) = 1 - \Phi(Q) + o_P(1), \tag{17}$$

where $o_P(1)$ denotes a random variable converging to zero in probability, Φ is the standard normal cdf, and $Q \sim N(0,1)$. Consequently, the HPC p-values are calibrated.

The proof of Theorem 1 is in Appendix A.1.

Theorem 1 proves that HPC p-values are calibrated for realized diagnostics that are asymptotically normal. In experiments (Section 5) we consider diagnostics that are not asymptotically normal, including the χ^2 diagnostic and the log-likelihood of latent Dirichlet allocation (Blei et al., 2003). With these diagnostics, we show empirically the HPC still has good calibration properties.

3.1 Data splitting for the HPC

We acknowledge that splitting the data may reduce power to detect model misspecification. This reduction in power may be avoided by using a PPC with a diagnostic that is independent of the model parameters. However, it can be difficult to verify such independence, and so the HPC allows practitioners more flexibility in diagnostic choice.

A separate aspect of data splitting is that the value of p_{hpc} will depend on the particular split of the data. However, Theorem 1 provides an asymptotic guarantee that p_{hpc} is uniformly distributed over any split of the data. Further, in finite sample settings, our experiments show that the HPC is approximately uniform.

Finally, we note that the HPC data splitting requirement limits us to diagnostic functions that can be calculated on a split of the data–we are unable to use diagnostics that depend jointly on all observations.

4 A comparison of predictive checks

In the previous section, we studied the holdout predictive check (HPC) p-value distribution when the model is correct. In this section, we consider a specific model and diagnostic function to illustrate the properties of the HPC p-value when the model is incorrect. Under this setup, we show the HPC has high power–it detects this model misfit with probability one in the infinite data limit, while the posterior predictive check (PPC) and calibrated versions of the PPC (Robins et al., 2000; Hjort et al., 2006) cannot detect this model misfit; that is, calibration by itself cannot improve the power of a test.

The setup is: we have data $\boldsymbol{y}^{\text{obs}} = \{y_i\}_{i=1}^n$ for which we posit a Gaussian model:

$$y_i \sim N(\mu, \sigma^2), \quad i = 1, \ldots, n \tag{18}$$

with unknown mean $\mu \in \mathbb{R}$ and known σ^2.

We want to check the suitability of this Gaussian model with a Bayesian predictive check. Unlike in a frequentist hypothesis test, the parameters of the model are not set to prespecified values. That is, in Equation 18 we are not checking a specific value of the mean μ, but the appropriateness of the Gaussian model.

In this section, we consider a specific alternative data distribution that would not fall in the Gaussian model (Equation 18), so we can explicitly compare the power of predictive checks. This data distribution is a Cauchy distribution with location $x_0 \in \mathbb{R}$ and scale $\gamma = 1$:

$$y_i \sim \text{Cauchy}(x_0, 1), \quad i = 1, \ldots, n. \tag{19}$$

Specifically, we analyze how the PPC, calibrated PPCs and the HPC perform when checking the Gaussian model (Equation 18) if the underlying data is actually Cauchy (Equation 19).

We use a mean diagnostic function: $d(\boldsymbol{y}^{\text{obs}}) = \overline{\boldsymbol{y}}^{\text{obs}}$. We chose this diagnostic function as it results in an analytic form for the PPC. This analytic form allows us to demonstrate clearly how the PPC fails to detect model misfit. While there are other choices of diagnostic function for which the PPC will detect model misfit in this scenario (e.g. the maximum), our main point is that the HPC will avoid the failure modes of the PPC no matter the choice of diagnostic function, allowing for more flexibility in diagnotic choice for practictioners.[1]

4.1 Posterior predictive checks

We first consider the distribution of the PPC p-value with the mean diagnostic which checks the Gaussian model in Equation 18. We show that this particular PPC p-value is degenerate at 0.5 regardless of whether the data is Gaussian or Cauchy. This degeneracy of the PPC p-value is problematic for both calibration and power. Specifically, the PPC p-value is not uniform and thus not calibrated. Further, the PPC has low power - it cannot detect model misfit when the data is actually Cauchy.

For a prior on μ, we take $\mu \sim N(\mu_0, \sigma_0^2)$ with fixed hyperparameters $\mu_0 \in \mathbb{R}, \sigma_0^2 \in \mathbb{R}^+$.

Concretely, the posterior predictive p-value with diagnostic $d(\boldsymbol{y}^{\text{obs}}) = \overline{\boldsymbol{y}}^{\text{obs}}$ is:

$$p_{\text{ppc}} = \text{p}(d(\boldsymbol{y}^{\text{rep}}) > d(\boldsymbol{y}^{\text{obs}}) | \boldsymbol{y}^{\text{obs}}) = 1 - \Phi\left(\frac{\overline{\boldsymbol{y}}^{\text{obs}} - \rho_n \overline{\boldsymbol{y}}^{\text{obs}} - (1-\rho_n)\mu_0}{\sqrt{(1+\rho_n)\sigma^2/n}}\right), \tag{20}$$

where $\rho_n = \sigma_0^2/(\sigma_0^2 + \sigma^2/n)$ (for details, see Appendix C).

Ideally, the PPC would check some aspect of model misfit. However, the PPC in Equation 20 is only checking how "close" the posterior mean $\mathbb{E}[\mu|\boldsymbol{y}^{\text{obs}}] = \rho_n \overline{\boldsymbol{y}}^{\text{obs}} + (1-\rho_n)\mu_0$ is to the MLE $\overline{\boldsymbol{y}}^{\text{obs}}$. Whether the posterior mean is close to the MLE is a property of the model, unrelated to the fit of the model to the data. That is, the posterior mean can be close to the MLE, whether or not the underlying data is actually Gaussian.

To further illustrate this problem with the PPC, consider the asymptotic distribution of p_{ppc}. In Equation 20, the numerator is $\overline{\boldsymbol{y}}^{\text{obs}} - \rho_n \overline{\boldsymbol{y}}^{\text{obs}} - (1-\rho_n)\mu_0 = O(1/n)$ and the denominator is $O(1/\sqrt{n})$. Consequently, the integrand goes to zero as $O(1/\sqrt{n})$ and the p_{ppc} converges to 0.5. What is important to note is that p_{ppc} converges to 0.5 regardless of whether the data is Gaussian or Cauchy.

[1]Others have also used the mean diagnostic to check a normal model (see, for example, the '8 Schools' example in Section 6.5, Gelman et al. (2013)).

4.2 Empirically calibrated PPCs can have low power

To fix the calibration of PPC p-values, a number of post-processing strategies have been proposed (Robins et al., 2000; Hjort et al., 2006). In this section, we show that such post-hoc calibration techniques do not improve the power of the PPC to detect model misfit.

Essentially, these calibration techniques cannot detect model misfit because for any random variable, we can calibrate it by using its cdf to transform it to a uniform distribution. If the original random variable does not detect model misfit, this transformation will not provide additional power to detect model misfit.

We make the above point more concrete by again considering Equation 18 and analyzing the following two post-hoc calibration methods:

- Robins et al. (2000) proposed to locate the observed PPC p-value in the empirical distribution of the PPC p-values. The empirical distribution of the PPC p-values is calculated by treating draws from the posterior predictive as replicates from the true model and then calculating their PPC p-values.
- Hjort et al. (2006) also propose to locate the observed PPC p-value in an empirical reference distribution. Unlike Robins et al. (2000), however, Hjort et al. (2006) use draws from the prior predictive distribution to calculate an empirical reference distribution (instead of draws from the posterior predictive).

First, we show that for the mean diagnostic, the Robins et al. (2000) empirically calibrated PPC p-value has the same distribution regardless of whether the data is Gaussian or Cauchy. Consequently, the check cannot detect model misfit.

The Robins et al. (2000) empirically calibrated PPC p-value is

$$\mathrm{p}\left(\mathrm{ppc}(\boldsymbol{y}^{\mathrm{rep}}) > \mathrm{ppc}(\boldsymbol{y}^{\mathrm{obs}}) | \boldsymbol{y}^{\mathrm{obs}}\right). \tag{21}$$

To compute this empirically calibrated p-value, we first need to determine the empirical distribution of $\mathrm{ppc}(\boldsymbol{y}^{\mathrm{rep}})$–these are the PPC values when $\boldsymbol{y}^{\mathrm{rep}}$ is treated as an observed dataset. We use the notation $\boldsymbol{y}^{\mathrm{rep,rep}}$ to denote a draw from the posterior predictive distribution conditioned on $\boldsymbol{y}^{\mathrm{rep}}$; that is, the posterior predictive where $\boldsymbol{y}^{\mathrm{rep}}$ is "observed data". More specifically, $\mathrm{ppc}(\boldsymbol{y}^{\mathrm{rep}})$ is calculated as

$$\mathrm{ppc}(\boldsymbol{y}^{\mathrm{rep}}) = \frac{1}{R}\sum_{r=1}^{R} \mathbb{1}(\overline{\boldsymbol{y}}_r^{\mathrm{rep,rep}} > \overline{\boldsymbol{y}}^{\mathrm{rep}}), \quad \boldsymbol{y}_r^{\mathrm{rep,rep}} \sim \mathrm{p}(\boldsymbol{y}^{\mathrm{rep,rep}} | \boldsymbol{y}^{\mathrm{rep}}), \tag{22}$$

with R draws from the posterior predictive distribution. Now, suppose $\rho_n \to 1$. Then, $\sqrt{n}(\overline{\boldsymbol{y}}^{\mathrm{rep,rep}} - \overline{\boldsymbol{y}}^{\mathrm{rep}}) | \overline{\boldsymbol{y}}^{\mathrm{rep}} \sim N(0, 2\sigma^2)$. In this case, the probability that $\overline{\boldsymbol{y}}^{\mathrm{rep,rep}}$ is greater than its mean is 0.5 and so the sum of indicators is a binomial random variable:

$$\sum_{r=1}^{R} \mathbb{1}(\overline{\boldsymbol{y}}_r^{\mathrm{rep,rep}} > \overline{\boldsymbol{y}}^{\mathrm{rep}}) \sim \mathrm{Binomial}(R, 0.5). \tag{23}$$

With the normal approximation to the binomial distribution,

$$\mathrm{p}\left(\mathrm{ppc}(\boldsymbol{y}^{\mathrm{rep}}) > \mathrm{ppc}(\boldsymbol{y}^{\mathrm{obs}})|\boldsymbol{y}^{\mathrm{obs}}\right) = 1 - \Phi\left(2\sqrt{R}\left[\mathrm{ppc}(\boldsymbol{y}^{\mathrm{obs}}) - 0.5\right]\right). \tag{24}$$

To determine the distribution of the empirically calibrated PPC, consider the distribution of $\mathrm{ppc}(\boldsymbol{y}^{\mathrm{obs}})$. Similarly to Equation 23, for large R,

$$2\sqrt{R}[\mathrm{ppc}(\boldsymbol{y}^{\mathrm{obs}}) - 0.5] \sim N(0, 1), \tag{25}$$

again using the normal approximation to the binomial distribution. Then, the empirically calibrated PPC is uniform:

$$\mathrm{p}\left(\mathrm{ppc}(\boldsymbol{y}^{\mathrm{rep}}) > \mathrm{ppc}(\boldsymbol{y}^{\mathrm{obs}})|\boldsymbol{y}^{\mathrm{obs}}\right) \sim U[0, 1]. \tag{26}$$

The key point is that the calibrated PPC is uniform regardless of whether the data is Gaussian or Cauchy. This is because for both data distributions, we have $\sqrt{n}(\overline{\boldsymbol{y}}^{\mathrm{rep,rep}} - \overline{\boldsymbol{y}}^{\mathrm{rep}})|\overline{\boldsymbol{y}}^{\mathrm{rep}} \sim N(0, 2\sigma^2)$ and $\sqrt{n}(\overline{\boldsymbol{y}}^{\mathrm{rep}} - \overline{\boldsymbol{y}}^{\mathrm{obs}})|\overline{\boldsymbol{y}}^{\mathrm{obs}} \sim N(0, 2\sigma^2)$. In other words, here, the empirically calibrated PPC does not depend on the underlying distribution of $\overline{\boldsymbol{y}}^{\mathrm{obs}}$ and so it cannot detect model misfit.

We now show that Hjort et al. (2006)'s calibrated p-value (cppp) also has the same distribution regardless of whether the data is Gaussian or Cauchy. Consequently, the check cannot detect model misfit.

The cppp is:

$$\mathrm{p}(\mathrm{ppc}(\boldsymbol{y}) > \mathrm{ppc}(\boldsymbol{y}^{\mathrm{obs}})|\boldsymbol{y}^{\mathrm{obs}}), \quad \text{where} \quad \boldsymbol{y} \sim \mathrm{p}(\boldsymbol{y}; \theta),\ \theta \sim \mathrm{p}(\theta), \tag{27}$$

and $\mathrm{ppc}(\cdot)$ is defined in Equation 22.

Similarly to Equation 23,

$$R \cdot \mathrm{ppc}(\boldsymbol{y}) = \sum_{r=1}^{R} \mathbb{1}(\overline{\boldsymbol{y}}_r^{\mathrm{rep}} > \overline{\boldsymbol{y}}) \sim \mathrm{Binomial}(R, 0.5), \tag{28}$$

where $\boldsymbol{y}_r^{\mathrm{rep}} \sim \mathrm{p}(\boldsymbol{y}^{\mathrm{rep}}|\boldsymbol{y})$.

Following a similar argument to the empirically calibrated PPC of Robins et al. (2000),

$$\mathrm{p}(\mathrm{ppc}(\boldsymbol{y}) > \mathrm{ppc}(\boldsymbol{y}^{\mathrm{obs}})|\boldsymbol{y}^{\mathrm{obs}}) \sim U[0, 1]. \tag{29}$$

Again, this will hold for regardless of whether the data is Gaussian or Cauchy and so the cppp will fail to detect model misfit.

In Section 5, we consider a linear regression example and demonstrate that these post-processing techniques yield calibrated PPC p-values, but that these p-values fail to detect model misspecification. In contrast, the HPC is calibrated and does detect model misspecification.

4.3 Comparison with holdout predictive checks

We have reviewed the "double use of the data" problem with PPCs, which can result in both uncalibrated p-values and minimal power to detect model misfit. These p-values can be empirically calibrated, but the calibration procedures do not necessarily improve the power of the test. In contrast, the HPC can detect model misspecification in situations where a PPC or calibrated PPC cannot.

Consider the example in Equation 18. When the data is actually Gaussian, the HPC p-value is:

$$\mathrm{p}(d(\boldsymbol{y}^{\mathrm{rep}}) > d(\boldsymbol{y}^{\mathrm{new}})|\boldsymbol{y}^{\mathrm{obs}}, \boldsymbol{y}^{\mathrm{new}}) = 1 - \Phi\left(\frac{\overline{\boldsymbol{y}}^{\mathrm{new}} - \rho_n \overline{\boldsymbol{y}}^{\mathrm{obs}} - (1-\rho_n)\mu_0}{\sqrt{(1+\rho_n)\sigma^2/n}}\right). \tag{30}$$

As $n \to \infty$, we have

$$\sqrt{n}[\overline{\boldsymbol{y}}^{\mathrm{new}} - \rho_n \overline{\boldsymbol{y}}^{\mathrm{obs}} - (1-\rho_n)\mu_0] \sim N(0, 2\sigma^2). \tag{31}$$

Then,

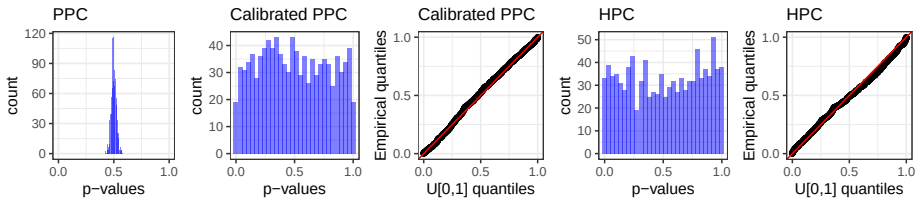
$$\mathrm{p}(d(\boldsymbol{y}^{\mathrm{rep}}) > d(\boldsymbol{y}^{\mathrm{new}})|\boldsymbol{y}^{\mathrm{obs}}, \boldsymbol{y}^{\mathrm{new}}) = 1 - \Phi(Z), \tag{32}$$

where $Z \sim N(0,1)$, and so the HPC p-value is uniform and calibrated. (This is an example of the more general calibration result proved in Theorem 1).

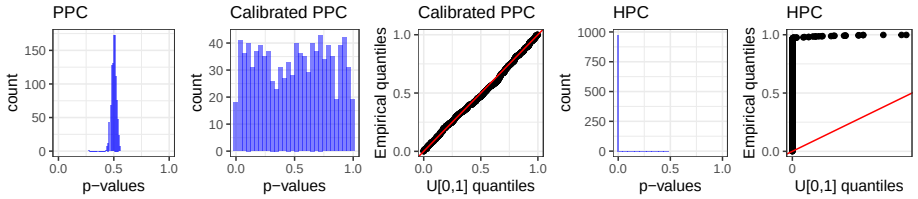
Now suppose the data is actually Cauchy (Equation 19). Then, we have $\overline{\boldsymbol{y}}^{\mathrm{new}} - \overline{\boldsymbol{y}}^{\mathrm{obs}} \sim \mathrm{Cauchy}(0,2)$. We prove that the HPC has asymptotic power of one–that is, HPC will reject the Gaussian model if the data is actually Cauchy with probability one. Let z_α denote the α-quantile of a standard Gaussian distribution. Then for the two-sided HPC, the power at rejection level α is:

$$\begin{aligned}
\text{Power} &= \mathrm{p}\left(\Phi\left(\frac{\overline{\boldsymbol{y}}^{\mathrm{new}} - \overline{\boldsymbol{y}}^{\mathrm{obs}}}{\sqrt{2\sigma^2/n}}\right) \le \alpha/2\right) + \mathrm{p}\left(\Phi\left(\frac{\overline{\boldsymbol{y}}^{\mathrm{new}} - \overline{\boldsymbol{y}}^{\mathrm{obs}}}{\sqrt{2\sigma^2/n}}\right) \ge 1 - \alpha/2\right) && (33)\\
&= \mathrm{p}\left(\overline{\boldsymbol{y}}^{\mathrm{new}} - \overline{\boldsymbol{y}}^{\mathrm{obs}} \ge z_{1-\alpha/2}\sqrt{\frac{2\sigma^2}{n}}\right) + \mathrm{p}\left(\overline{\boldsymbol{y}}^{\mathrm{new}} - \overline{\boldsymbol{y}}^{\mathrm{obs}} \le z_{\alpha/2}\sqrt{\frac{2\sigma^2}{n}}\right) && (34)\\
&\overset{n\to\infty}{\to} 1. && (35)
\end{aligned}$$

To further illustrate these points, the empirical distributions of the PPC, calibrated PPC and HPC for the simple mean example are displayed in Figure 3. We see the PPC is concentrated around 0.5 both when the data is Gaussian and when the data is Cauchy. Moreover, the calibrated PPC is uniform for both Gaussian- and Cauchy-distributed data. In contrast, the HPC is uniform when the data is Gaussian, and concentrated around 0 when the data is Cauchy. That is, the HPC detects model misspecification while the PPC or calibrated PPC do not.



(a) Distribution of p-values when the data is Gaussian (Equation 18).



(b) Distribution of p-values when the data is Cauchy (Equation 19).

Figure 3: In the Gaussian mean example, the HPC detects model misspecification (i.e. when the data is actually Cauchy) while the PPC and calibrated PPC do not.

4.4 Choosing diagnostic functions

We showed that the PPC for checking a Gaussian model with a mean diagnostic is degenerate, even under the alternative hypothesis when the true data distribution is Cauchy.

In general, when does this degeneracy occur for PPC p-values? The PPC becomes degenerate when the diagnostic is perfectly correlated with the model parameters (Robins et al., 2000). More specifically, let $\theta(\boldsymbol{y}^{\text{obs}})$ be the posterior mean of the model parameters, conditioned on the observed data. Then under the conditions of Theorem 1:

- if the diagnostic $d(\boldsymbol{y}^{\text{obs}})$ is perfectly correlated with $\theta(\boldsymbol{y}^{\text{obs}})$, then the PPC will converge to 0.5 under both the null and alternative hypothesis;
- if the diagnostic is independent of $\theta(\boldsymbol{y}^{\text{obs}})$, the PPC p-values are uniformly distributed under the null hypothesis and thus calibrated (Robins et al., 2000).

However, it is often difficult to determine the degree of association between $\theta(\boldsymbol{y}^{\text{obs}})$ and $d(\boldsymbol{y}^{\text{obs}})$, and thus difficult to assess whether the PPC will detect model misfit. In contrast, we proved that the HPC is calibrated for all normally-distributed diagnostic functions (Theorem 1). This result allows the HPC much more flexibility in the choice of diagnostic than the PPC.

The choice of diagnostic function also has implications for the power of the test. For the Gaussian/Cauchy example studied in this section, the mean diagnostic can detect model misfit because under H_0, $\overline{\boldsymbol{y}^{\text{new}}} - \overline{\boldsymbol{y}^{\text{obs}}} \sim N(0, 2)$, while under H_1, $\overline{\boldsymbol{y}^{\text{new}}} - \overline{\boldsymbol{y}^{\text{obs}}} \sim \text{Cauchy}(0, 2)$– interestingly, it is the variance in the mean diagnostic that allows us to detect model misfit.

Other choices of diagnostic functions will also change the power of the test. As a trivial

example, if we had chosen $d(\boldsymbol{y}) = c$ for some constant c, the distribution of $d(\boldsymbol{y})$ would be the same under both H_0 and H_1 and so we cannot detect model misfit.

As a final example of how the diagnotic function may affect power, suppose we take $d(\boldsymbol{y}) = \max \boldsymbol{y}$. When will we have the same distribution of the test-statistic under both H_0 and H_1? This may occur if the models in H_0 and H_1 are such that their maxima converge to the same distribution. This is possible given the extreme value theorem, which tells us that for a variety of distributions, the distribution of their maxima is either Gumbel, Fréchet or Weibull.

5 Empirical Study

We study holdout predictive checks on a regression model, a hierarchical model of documents, and a factor model.

- In the regression model study:

  (i) We show empirically that the holdout predictive check (HPC) detects overfitting when there is insuficient regularization. In contrast, the posterior predictive check (PPC), as well as the calibration suggestions of Robins et al. (2000) and Hjort et al. (2006), do not detect overfitting when there is insufficient regularization.

  (i) We show empirically that the HPC p-values are approximately uniform when the model has sufficient regularization (i.e. the HPC is calibrated). In contrast, the PPC p-values are not calibrated.

- For the hierarchical model of documents:

  (i) On synthetic data, we show empirically that the HPC p-values are approximately uniform when the data actually comes from the model (i.e. the HPC is calibrated).

  (i) On a collection of documents from the *New York Times*, we show that the HPC detects model misfit due to overfitting, while the PPC does not.

- In the factor model study:

  (i) When the model is correct, we show empirically that the HPC p-values show much greater variability than the PPC p-values, which are centered around 0.5.

  (i) When the model is incorrect, we show that the HPC correctly reject the model at a much higher rate than the PPC.

5.1 Bayesian Ridge Regression

In this section, we return to the regression example in Section 2.1. We now compare the HPC with a variety of methods for Bayesian model criticism for a range of different values of the prior variance c. Again, we set $n = 50$ and $p = 100$.

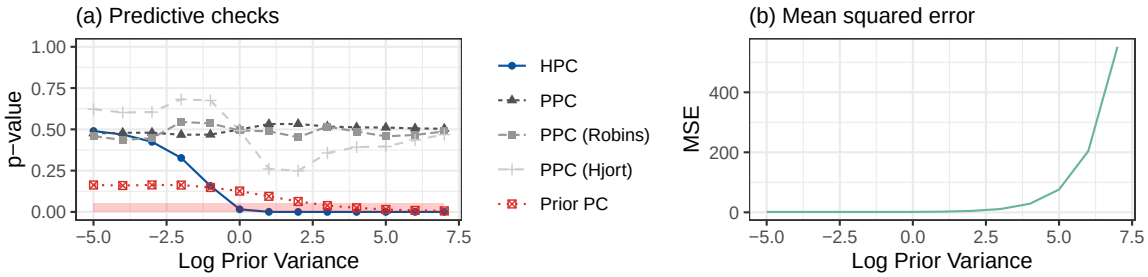


Figure 4: Bayesian ridge regression: HPC detects overfitting when there is little regularization (large c), while the PPC (including calibrated versions) does not. (a) average p-values over different c from PPC, calibrated PPCs (Robins et al., 2000; Hjort et al., 2006), HPC and prior PC. (b) mean squared error of fitted model (Equation 13) over different values of c.

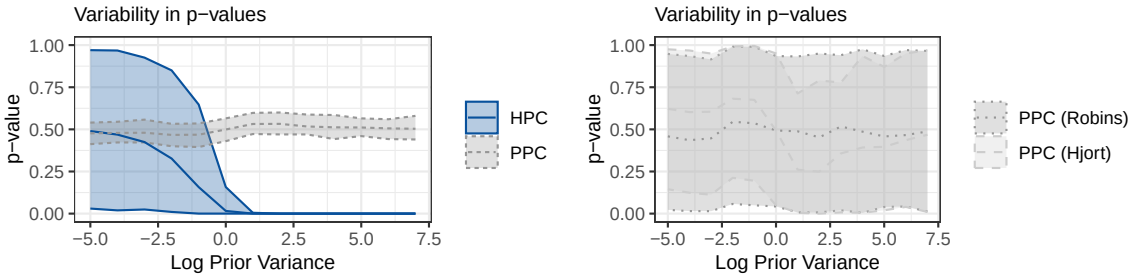
In the large c scenario, we expect that a posterior predictive check will not be able to detect the overfitting of the data. Moreover, as discussed in Section 4, we expect the calibration strategies of Robins et al. (2000) and Hjort et al. (2006) will also not detect this overfitting. In contrast, we expect that a holdout predictive check will be able to detect overfitting, and reject the model with large prior variance values. Recall that we are not using the HPC to choose the value of the regularization parameter c; we are simply emphasizing that when there is clear overfitting, the HPC is able to detect this poor model fit, while the PPC cannot.

We also consider prior predictive checks (Box, 1980).

As anticipated, the PPC p-values are constant for all values of the regularization parameter, c; that is, the PPC cannot detect model misfit for large values of c (Figure 4(b)). Moreover, the two calibrated PPCs (Robins et al., 2000; Hjort et al., 2006) also cannot detect this model misfit. Meanwhile, the HPC retains the model for small values of c. For large values of c, the HPC rejects the model as it overfits to the observed data. The prior predictive check exhibits similar behavior to the HPC, but it does not begin to reject the model until larger values of c.

We next consider the variability of the p-values from the predictive checks. As anticipated by the theoretical results of Robins et al. (2000), the PPC p-value is tightly concentrated around 0.5 for all values of c (Figure 5b). Meanwhile, the empirical calibration procedures of Robins et al. (2000); Hjort et al. (2006) give p-values with greater variability around 0.5, as expected from the calibration process; however, they still do not reject the model for large c (Figure 5b). In contrast, the HPC p-values show greater variability around 0.5 for small values of c and then concentrate around 0 for large values of c (Figure 5a).

The empirical distribution of the HPC p-values are displayed in Figure 6. The p-values are approximately uniform for small values of c. This provides empirical evidence for the calibration of HPC p-values with the χ^2-diagnostic. For large values of c, the p-values concentrate around 0; that is, the HPC has high power to detect model misfit. Additional histograms and QQ-plots of the p-value distributions for all methods are in Appendix E.



(a) HPC vs. PPC

(b) Calibrated PPCs (Robins et al., 2000; Hjort et al., 2006)

Figure 5: Bayesian ridge regression: (a) When the model is not overfit, the HPC p-values are much more variable than PPC p-values, which concentrate around 0.5. When the model is overfit, HPC p-values detect this overfitting while PPC p-values do not. (b) Calibrated PPC p-values are much more variable than the PPC; however, the calibrated PPC does not detect overfitting for large c. For both (a) and (b), lower and upper lines show 0.025 and 0.975 quantiles of p-values over $K = 100$ replicates for different values of the regularization parameter c; middle line shows mean.

In theory, the HPC is calibrated for asymptotically normal diagnostic statistics (Theorem 1). In this regression example, the diagnostic statistic is χ^2 distributed, not normal. Although our theory does not extend to this diagnostic, we showed empirically that the HPC has good calibration properties.

5.2 Topic Modeling

We next study HPCs on latent Dirichlet allocation (LDA) (Blei et al., 2003), a hierarchical model of documents. We first apply LDA to synthetic data and show empirically the HPC has good calibration properties. We then apply LDA to a collection of documents from the *New York Times* and show that the HPC detects model misfit due to overfitting.

LDA models documents as mixtures over latent topics, where each topic is a distribution over words. The kth topic is denoted by $\beta_k = \{\beta_{kv}\}_{v=1}^{V}$, where $V \in \mathbb{N}$ is the number of unique words in the corpus, and the number of topics ranges from $k = 1, \ldots, K$. The topics are drawn as:

$$\beta_k \sim \text{Dirichlet}(\alpha \mathbf{1}_V), \quad k = 1, \ldots, K, \tag{36}$$

where $\alpha \in \mathbb{R}^+$ is a hyperparameter.

For the dth document, the generative process is:

1. Draw the topic proportions $\theta_d \sim \text{Dirichlet}(\alpha \mathbf{1}_K)$
2. For words $n = 1, \ldots, N_d$:
   (a) Draw a topic $z_{dn} \sim \text{Multinomial}(\theta_d)$

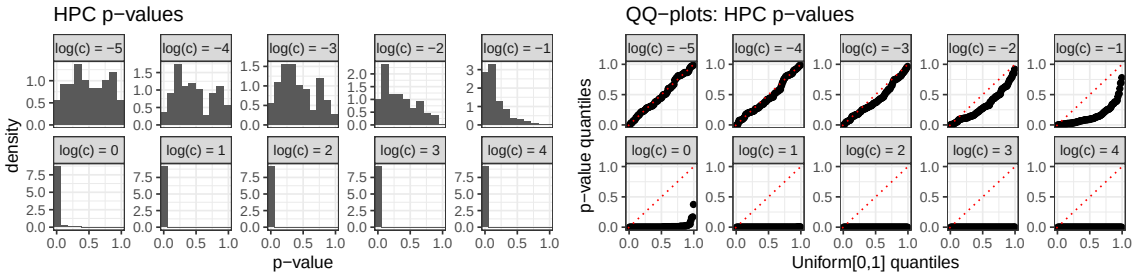


Figure 6: Bayesian ridge regression: HPC p-values are approximately uniform when there is no overfitting (small c); then, the HPC detects when there is overfitting (large c). Left: Histograms of HPC p-values. Right: QQ-plots comparing the quantiles of the HPC p-values with the quantiles of a uniform[0,1] random variable. Plots are over $K = 100$ replications of the data.

(b) Draw a word conditioned on the topic $w_{dn} \sim \text{Multinomial}(\beta_{z_{dn}})$

We set the Dirichlet hyperparameter to $\alpha = 0.1$ on both the topics and the document proportions.

For the model check diagnostic, we will use the log-likelihood. For document d, the log-likelihood is:

$$\ell(\boldsymbol{w}_d; \theta_d, \boldsymbol{\beta}) = \sum_{v=1}^{V} \left[\sum_{n=1}^{N_d} \mathbb{1}(w_{dn} = v) \right] \log \left(\sum_{k=1}^{K} \beta_{kv} \theta_{dk} \right). \tag{37}$$

To calculate the PPC p-value, we first calculate the posterior expectation of the global topics $\widehat{\boldsymbol{\beta}} = \mathbb{E}[\boldsymbol{\beta}|\widetilde{\boldsymbol{w}}]$ where $\widetilde{\boldsymbol{w}}$ is a heldout set of documents. Then, we estimate the local per-document parameters, $\{\theta_d\}_{d=1}^{D}$, and draw replicates from the posterior predictive distribution. The per-document PPC is:

$$\text{ppc}(\boldsymbol{w}_d^{\text{obs}}) = \frac{1}{R} \sum_{r=1}^{R} \mathbb{1}[\ell(\boldsymbol{w}_{d,r}^{\text{rep}}, \theta_{d,r}, \widehat{\boldsymbol{\beta}}) > \ell(\boldsymbol{w}_d^{\text{obs}}, \theta_{d,r}, \widehat{\boldsymbol{\beta}})], \tag{38}$$

$$\text{where} \quad (\boldsymbol{w}_{d,r}^{\text{rep}}, \theta_{d,r}) \sim p(\boldsymbol{w}_d^{\text{rep}}|\theta_d, \widehat{\boldsymbol{\beta}}) p(\theta_d|\boldsymbol{w}_d^{\text{obs}}, \widehat{\boldsymbol{\beta}}). \tag{39}$$

Note that the topic vector $\boldsymbol{\beta}$ is fixed as the estimate from the heldout corpus; the above PPC is checking the fit of the model based on the document topic proportions θ_d.

To calculate the HPC p-value, we split each document in half: $\boldsymbol{w}_d = \{\boldsymbol{w}_d^{\text{obs}}, \boldsymbol{w}_d^{\text{new}}\}_{d=1}^{D}$. Here, we split the data at the document level because for each document, we need to infer the local variable, θ_d. Half of the document is used to infer this per-document variable, while the other half is used to check the model:

$$\text{hpc}(\boldsymbol{w}_d^{\text{obs}}, \boldsymbol{w}_d^{\text{new}}) = \frac{1}{R} \sum_{r=1}^{R} \mathbb{1}[\ell(\boldsymbol{w}_{d,r}^{\text{rep}}, \theta_{d,r}, \widehat{\boldsymbol{\beta}}) > \ell(\boldsymbol{w}_d^{\text{new}}, \theta_{d,r}, \widehat{\boldsymbol{\beta}})], \tag{40}$$

$$\text{where} \quad (\boldsymbol{w}_{d,r}^{\text{rep}}, \theta_{d,r}) \sim p(\boldsymbol{w}_d^{\text{rep}}|\theta_d, \widehat{\boldsymbol{\beta}}) p(\theta_d|\boldsymbol{w}_d^{\text{obs}}, \widehat{\boldsymbol{\beta}}). \tag{41}$$

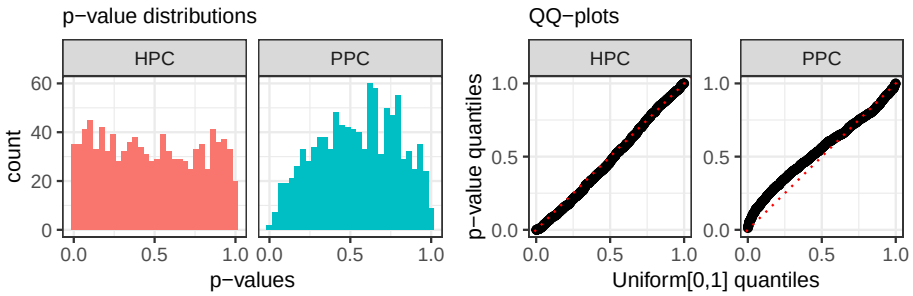


Figure 7: Simulated LDA data: when the model is true, HPC p-values are approximately uniform while PPC p-values are not. Left: Histograms of HPC and PPC p-values. Right: QQ-plots comparing the quantiles of the HPC and PPC p-values with the quantiles of a uniform[0,1] random variable.

5.3 Synthetic data

We investigate the distribution of HPC and PPC p-values when the data is drawn from the LDA generative process with $K = 10$ topics and $V = 2500$ vocabulary of unique words. The number of words (tokens) in each document is drawn as $N_d \sim \text{Poisson}(\xi)$ where $\xi = 300$.

We draw $100,000$ documents and infer the topics β using stochastic variational inference (Hoffman et al., 2013) with minibatches of size 100. Given the topics β, we then compare the PPC and HPC on a heldout set of documents of size $D = 1,000$.

For each document, we calculate the PPC and HPC p-values using the log-likelihood diagnostic with $R = 500$ replicates from the posterior predictive distribution. The HPC p-values are approximately uniformly distributed while the PPC p-values are left-skewed (Figure 7). This provides empirical evidence that the HPC is calibrated for the LDA log-likelihood diagnostic.

5.4 New York Times

We consider a corpus of 100,000 news documents with a vocabulary size of 5,000 unique words from *the New York Times*. To infer the topics β, we implement stochastic variational inference (Hoffman et al., 2013) with minibatches of size 100. Given the topics β, we then compare the PPC and HPC on a set of documents of size 1,000.

For each document, we calculate the PPC and HPC p-values using the log-likelihood diagnostic averaged over documents, where for each document we draw $R = 500$ replicates from the posterior predictive distribution.

As the number of topics increases to $K = 1,000$, the PPC p-value remains close to 0.5, and the negative log-likelihood is monotonically decreasing (Figure 8). At $K = 1,000$, the number of topics is equal to the size of the vocabulary; at this stage, the model assigns each word to its own topic, memorizing the data. The HPC p-value detects this overfitting, and begins to reject the model past $K = 200$ topics (Figure 8 left).

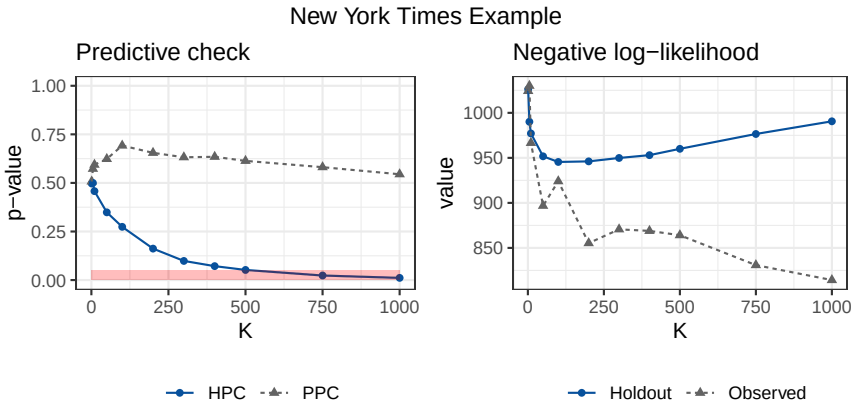


Figure 8: On the New York Times data, the HPC detects model overfitting, rejecting the model when the number of topics is large. In contrast, the PPC never rejects the model. Left: HPC and PPC p-values for different values of topics K. Right: Negative log-likelihood on observed and holdout data over different values of topics K.

Again, we are not using the HPC to choose the number of topics K; we are simply emphasizing that when there is clear overfitting, the HPC is able to detect this poor model fit, while the PPC cannot.

Note that when the number of topics is small, the HPC does not reject the model. A similar phenomenon was noted in Moran et al. (2022) for a Gaussian mixture model when the number of mixtures components is fewer than the truth. When the number of mixture components is too few, the entropy of the posterior predictive distribution increases to fit the data. The draws from this posterior distribution can be extreme relative to the true model, making it difficult to distinguish model misfit.

5.5 Factor Analysis

In this section, we study the HPC for factor analysis, using examples similar to Moran et al. (2022). We fit probabilistic principal component analysis (PPCA, Tipping and Bishop, 1999) to (i) data drawn from a well-specified linear model and (ii) data drawn from a nonlinear model for which PPCA is not well-specified. We show empirically that:

- PPC p-values are not calibrated and have low power to detect the poor fit of the nonlinear model;
- HPC p-values, while not exactly calibrated, avoid the degeneracy of the PPC p-values when the model is correct. When the model is incorrect, the HPC has much higher power than the PPC.

Data generating process

The observed data is $\boldsymbol{x}_i \in \mathbb{R}^G$, $i = 1, \ldots, N$. We assume that $\boldsymbol{x}_i$ has some low dimensional

representation $z_i \in \mathbb{R}^K$ with

$$\boldsymbol{x}_i = f(\boldsymbol{z}_i) + \boldsymbol{\varepsilon}_i \tag{42}$$

for some function $f : \mathbb{R}^K \to \mathbb{R}^G$ and noise term $\boldsymbol{\varepsilon}_i \in \mathbb{R}^G$. We will consider two cases: (i) f is set to a linear function and (ii) f is set to a nonlinear function.

Model

Probabilistic PCA (Tipping and Bishop, 1999) assumes that f is a linear mapping from the low-dimensional latent representation to the observed data,

$$\boldsymbol{x}_i = \boldsymbol{W}\boldsymbol{z}_i + \boldsymbol{\varepsilon}_i, \quad \boldsymbol{\varepsilon}_i \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}).$$

The latent variables are assigned a normal prior, $\boldsymbol{z}_i \sim N(0, \boldsymbol{I})$. We fit $\boldsymbol{W}$ and representations $\boldsymbol{z}_i$ using the EM algorithm.

According to Equation 42, when f is linear, PPCA should adequately fit the data. In contrast, when f is nonlinear, PPCA does not have the requisite complexity to fit the data.

Diagnostic function

To check the model, we use the likelihood diagnostic. For PPCA, this is:

$$d_{\text{ppca}}(\boldsymbol{x}; \boldsymbol{x}_{\text{obs}}) = \sum_{i=1}^{N} \frac{1}{2\widehat{\sigma^2}} \|\boldsymbol{x}_i - \widehat{\boldsymbol{W}}\mathbb{E}[\boldsymbol{z}|\boldsymbol{x}_i, \widehat{\boldsymbol{W}}]\|^2, \quad \text{where} \quad (\widehat{\boldsymbol{W}}, \widehat{\sigma}^2) = \arg\max_{\boldsymbol{W}, \sigma^2} \log p(\boldsymbol{x}_{\text{obs}}|\boldsymbol{W}, \sigma^2) \tag{43}$$

$$\text{and} \quad \mathbb{E}[\boldsymbol{z}_i|\boldsymbol{x}_i, \widehat{\boldsymbol{W}}, \widehat{\sigma}^2] = \boldsymbol{M}^{-1}\widehat{\boldsymbol{W}}^T\boldsymbol{x}_i, \quad \boldsymbol{M} = \widehat{\boldsymbol{W}}^T\widehat{\boldsymbol{W}} + \widehat{\sigma}^2\boldsymbol{I}. \tag{44}$$

Then, the empirical PPC and HPC are:

$$p_{\text{ppc}} = \frac{1}{R}\sum_{r=1}^{R} \mathbb{I}[d(\boldsymbol{x}_{\text{rep}}; \boldsymbol{x}_{\text{obs}}) > d(\boldsymbol{x}_{\text{obs}}; \boldsymbol{x}_{\text{obs}})]; \tag{45}$$

$$p_{\text{hpc}} = \frac{1}{R}\sum_{r=1}^{R} \mathbb{I}[d(\boldsymbol{x}_{\text{rep}}; \boldsymbol{x}_{\text{obs}}) > d(\boldsymbol{x}_{\text{new}}; \boldsymbol{x}_{\text{obs}})]. \tag{46}$$

5.5.1 Linear data generating process

We first consider the setting where f is linear. We set the number of samples to $N = 1000$, the number of observed features to $G = 11$ and the latent dimension as $K = 6$. The data is generated as

$$\boldsymbol{x}_i = \boldsymbol{W}\boldsymbol{z}_i + \boldsymbol{\varepsilon}_i,$$

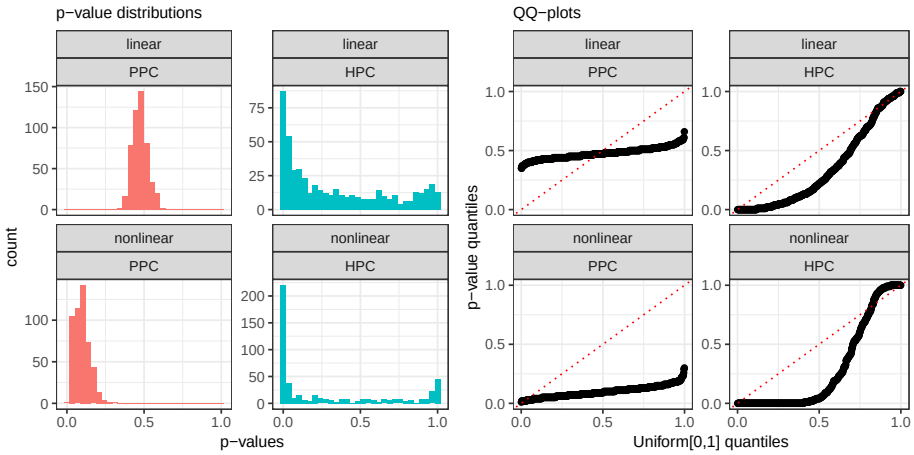


Figure 9: Factor analysis example. Top row: when the model is correctly specified (linear), the HPC avoids the degeneracy of the PPC around 0.5. Bottom row: when the model is incorrect (nonlinear), the HPC has higher power than the PPC to reject the model.

where $\boldsymbol{z}_i \sim N(0, \boldsymbol{I})$, $\boldsymbol{\varepsilon}_i \sim N(0, \sigma^2 \boldsymbol{I})$ with true $\sigma^2 = 1$. (Note however that σ^2 is treated as unknown in the inference stage). The matrix $\boldsymbol{W}$ is the matrix,

$$\boldsymbol{W}^\top = \begin{pmatrix} 5 & 5 & 5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 5 & 5 & 5 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 5 & 5 & 5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 5 & 5 & 5 \end{pmatrix}.$$

In this setting, the PPC is centered around 0.5 (Figure 9, top row). Meanwhile, the HPC, while not perfectly calibrated, avoids this degeneracy issue of the PPC.

5.5.2 Nonlinear data generating process

We next consider the setting where the true mapping f from the factors to the observed data is nonlinear. Here, we expect PPCA to be a poor fit for the nonlinear model. We set the number of samples to $N = 1000$, the number of observed features to $G = 6$ and the latent dimension to $K = 2$. The data is generated from

$$\boldsymbol{x}_i = (z_{i1},\ 2z_{i1},\ 3z_{i1}^2,\ 4z_{i2},\ 5z_{i2},\ 6\sin(\pi/2 \cdot z_{i2}))^\top + \boldsymbol{\epsilon}_i, \tag{47}$$

where $\boldsymbol{z}_i \sim N(0, \boldsymbol{I})$, $\boldsymbol{\varepsilon}_i \sim N(0, \sigma^2 \boldsymbol{I})$ with true $\sigma^2 = 1$. That is, the first three columns of $\boldsymbol{x}$ are related to the first factor and the next three columns are related to the second factor.

In this setting, the PPC p-values have a median of 0.09 and reject the model only 12.6% of the time (Figure 9, bottom row). Meanwhile, the HPC rejects the model 63.2% of the time (Figure 9, bottom row). That is, the HPC empirically has much higher power than the PPC to detect the inadequacy of PPCA for modeling the nonlinear data generating process.

6 Discussion

We developed holdout predictive checks (HPCs), a diagnostic tool that brings together Bayesian methods for model checking with frequentist estimation of goodness of fit. The HPC assesses a Bayesian model by comparing samples from the posterior predictive to a sample from the population distribution, which in practice is a holdout dataset. We proved that the HPC is calibrated for a class of asymptotically normal diagnostic functions and empirically show its calibration for other diagnostics. We revisited the "double use of the data" issue of posterior predictive checks (PPCs) and highlighted that while post-hoc procedures can be used to calibrate the PPC, post-hoc calibration does not provide power to detect model misfit. Finally, we demonstrated the utility of HPCs with Bayesian linear regression models, on probabilistic topic models of documents, and on factor analysis.

There are several areas for further research. For hierarchical models of grouped data, Marshall and Spiegelhalter (2003) define mixed predictive checks, where the reference distribution combines the prior for the group-specific latent variables with the posterior for latent variables shared across groups. This approach mitigates the issues of a PPC, but there are no guarantees; a mixed predictive check can still be uncalibrated or have low power if the influence of the posterior becomes too large. To avoid these issues, Bayarri and Castellanos (2007) extends the checks of Bayarri and Berger (2000) to hierarchical models. How to extend the HPC to these situations is one avenue of further research.

Generalizing beyond hierarchical models, researchers have studied how to check individual components of a probabilistic model, i.e., individual nodes in a directed graphical model. O'Hagan (2003) proposes some of the earlier ideas along these lines, though in a way that uses the data twice and leads to an uncalibrated check. To correct this lack of calibration, his predictive checks for individual components of a model have been extended, often by using data splitting (Marshall and Spiegelhalter, 2007; Bayarri and Castellanos, 2007; Dahl et al., 2007; Gåsemyr and Natvig, 2009; Presanis et al., 2013). Again the HPC proposed here could be extended to these settings.

Acknowledgements

This research was supported by NSF IIS 2127869, ONR N00014-17-1-2131, ONR N00014-15-1-2209, Simons Foundation, Sloan Foundation, Open Philanthropy, NIH/NHLBI Award R01HL148248, NSF Award 1922658, NSF CAREER Award 2145542 and the Eric and Wendy Schmidt Center.

We thank Jiawei Li and Jonathan Huggins for pointing out an error in an earlier version of this manuscript.

References

Bayarri, M. and Berger, J. (2000). P-values for composite null models. *Journal of the American Statistical Association*, 95(452):1127–1142.

Bayarri, M. and Castellanos, M. (2007). Bayesian checking of the second levels of hierarchical models. *Statistical Science*, 22:322–343.

Bayarri, M. and Morales, J. (2003). Bayesian measures of surprise for outlier detection. *Journal of Statistical Planning and Inference*, 111:3–22.

Berger, J. and Pericchi, L. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, 91(433):109–122.

Blei, D. (2014). Build, compute, critique, repeat: Data analysis with latent variable models. *Annual Review of Statistics and Its Application*, 1:203–232.

Blei, D., Ng, A., and Jordan, M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Box, G. (1980). Sampling and Bayes' inference in scientific modeling and robustness. *Journal of the Royal Statistical Society, Series A*, 143(4):383–430.

Dahl, F., Gåsemyr, J., and Natvig, B. (2007). A robust conflict measure of inconsistencies in bayesian hierarchical models. *Scandinavian Journal of Statistics*, 34(4):816–828.

Draper, D. (1996). Comment: Utility,sensitivity analysis, and cross-validation in Bayesian model-checking. *Statistica Sinica*, 6(760–767.).

Evans, M. and Moshonov, H. (2006). Checking for prior-data conflict. *Bayesian Analysis*, 1(4):893–914.

Gåsemyr, J. and Natvig, B. (2009). Extensions of a conflict measure of inconsistencies in Bayesian hierarchical models. *Scandinavian Journal of Statistics*, 36(4):822–838.

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association*, 70(350):320–328.

Gelfand, A., Dey, D., and Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. *Bayesian Statistics*, 4.

Gelman, A. (2004). Exploratory data analysis for complex models. *Journal of Computational and Graphical Statistics*, 13(4):755–779.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2013). *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (1995). *Bayesian Data Analysis*. Chapman & Hall, London.

Gelman, A., Meng, X., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6:733–807.

Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., and Modrák, M. (2020). Bayesian workflow. *arXiv preprint arXiv:2011.01808*.

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

Hjort, N., Dahl, F., and Steinbakk, G. (2006). Post-processing posterior predictive p values. *Journal of the American Statistical Association*, 101(475):1157–1174.

Hoffman, M., Blei, D., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *Journal of Machine Learning Research*, 14(1303–1347).

Johnson, V. E. (2007). Bayesian model assessment using pivotal quantities. *Bayesian Analysis*, 2(4):719–733.

Larsen, M. and Lu, L. (2007). Comment: Bayesian checking of the second level of hierarchical models: Cross-validated posterior predictive checks using discrepancy measures. *Statistical Science*, 22:359–362.

Lewis, S. and Raftery, A. (1996). Comment: Posterior predictive assessment for data subsets in hierarchical models via MCMC. *Statistica Sinica*, 6:779–786.

Li, J. and Huggins, J. H. (2022). Calibrated model criticism using split predictive checks. *arXiv preprint arXiv:2203.15897*.

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

Marshall, E. and Spiegelhalter, D. (2007). Identifying outliers in Bayesian hierarchical models: a simulation-based approach. *Bayesian Analysis*, 2(2):409–444.

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

Moran, G. E., Cunningham, J. P., and Blei, D. M. (2022). The posterior predictive null. *Bayesian Analysis*, 1(1):1–27.

O'Hagan, A. (2003). HSSS model criticism. In *Highly Structured Stochastic Systems*, volume 27 of *Oxford Statist. Sci. Ser.*, pages 423–453. Oxford Univ. Press, Oxford.

Presanis, A., Ohlssen, D., Spiegelhalter, D., and De Angelis, D. (2013). Conflict diagnostics in directed acyclic graphs, with applications in Bayesian evidence synthesis. *Statistical Science*, 28(3):376–397.

Ranganath, R. and Blei, D. M. (2019). Population predictive checks. *arXiv preprint arXiv:1908.00882v1*.

Robins, J., van der Vaart, A., and Ventura, V. (2000). Asymptotic distribution of p-values in composite null models. *Journal of the American Statistical Association*, 95(452):1143–1156.

Rubin, D. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, 12(4):1151–1172.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 64(4):583–639.

Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622.

Watanabe, S. (2009). *Algebraic geometry and statistical learning theory*, volume 25. Cambridge University Press.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(12).

A Proofs

A.1 Proof of calibration of the holdout predictive check

In this section, we prove Theorem 1. Recall that we assume the diagnostic $d(\boldsymbol{y})$ is asymptotically normal with asymptotic mean $\nu(\theta)$ and asymptotic variance $\sigma^2(\theta)$, under the null hypothesis that the density of $\boldsymbol{y}$ is $f(\boldsymbol{y};\theta)$:

$$n^{1/2}\left[\frac{d(\boldsymbol{y})-\nu(\theta)}{\sigma(\theta)}\right] \rightsquigarrow N(0,1), \tag{48}$$

where $\rightsquigarrow$ denotes convergence in distribution.

Notation. We let $\mathrm{p}(\theta|\boldsymbol{y})$ denote the posterior distribution of θ. We use $\|\cdot\|$ to denote the total variation distance between two distributions P and Q. The Fisher information is

$$I(\theta) = \lim_{n\to\infty} n^{-1}\mathbb{E}_\theta\left[-\partial^2 \log f(\boldsymbol{y};\theta)/\partial\theta\partial\theta'\right]. \tag{49}$$

We use the notation $y_n = O_p(a_n)$ as $n\to\infty$ to mean y_n/a_n is stochastically bounded: for any $\varepsilon > 0$, there exists finite $M > 0$ and $N > 0$ such that

$$\mathrm{p}(|y_n/a_n| > M) < \varepsilon, \quad \forall n > N. \tag{50}$$

We use $\phi(\theta;\mu,\Sigma)$ to denote the density of a $N(\mu,\Sigma)$ random variable.

Regularity conditions.

1. The asymptotic mean ν is continuously differentiable in a neighborhood of $(0, \theta)$, with partial derivatives converging to limit:

$$\dot{\nu}(\theta_0) = \lim_{n\to\infty} \partial \nu_n(0,\theta)/\partial\theta\big|_{\theta=\theta_0}. \tag{51}$$

2. For some p-vector-valued function $\theta(\boldsymbol{y})$ on the sample space, we assume

$$\|\mathrm{p}(\cdot|\boldsymbol{y}) - N(\theta(\boldsymbol{y}), I^{-1}(\theta_0)/n)\| \overset{P_{\theta_0}}{\to} 0 \tag{52}$$

and

$$n^{1/2}(\theta(\boldsymbol{y}) - \theta_0) = O_{P_{\theta_0}}(1). \tag{53}$$

Proof of Theorem 1. Our proof technique follows that of Robins et al. (2000) for their proof of Theorem 3.

The HPC p-value is:

$$p(\boldsymbol{y}^{\text{obs}}, \boldsymbol{y}^{\text{new}}) = \int_\Theta \mathrm{p}(d(\boldsymbol{y}^{\text{rep}}) > d(\boldsymbol{y}^{\text{new}})|\boldsymbol{y}^{\text{new}}, \theta)\mathrm{p}(\theta|\boldsymbol{y}^{\text{obs}})d\theta \tag{54}$$

Consider:

$$d(\boldsymbol{y}^{\text{rep}}) > d(\boldsymbol{y}^{\text{new}}) \tag{55}$$

Then,
$$\sqrt{n}[d(\boldsymbol{y}^{\text{rep}}) - \nu(\theta)] > \sqrt{n}[d(\boldsymbol{y}^{\text{new}}) - \nu(\theta_0) - (\nu(\theta) - \nu(\theta_0))]. \tag{56}$$

Consider the LHS of Equation 56. By Equation 48, we have that, conditional on θ,

$$\sqrt{n}[d(\boldsymbol{y}^{\text{rep}}) - \nu(\theta)] \sim N(0, \sigma^2(\theta_0)).$$

Consider now the RHS of Equation 56. The first term $\sqrt{n}[d(\boldsymbol{y}^{\text{new}}) - \nu(\theta_0)]$ is fixed, conditional on $\boldsymbol{y}^{\text{new}}$. For the second term, we use a Taylor expansion to obtain:

$$\nu(\theta) - \nu(\theta_0) \approx \dot{\nu}(\theta_0)(\theta - \theta_0). \tag{57}$$

Then by Equation 52, given $\boldsymbol{y}^{\text{obs}}$, the second term is approximately distributed as

$$N(\dot{\nu}(\theta_0)^T n^{1/2}(\theta(\boldsymbol{y}^{\text{obs}}) - \theta_0), \dot{\nu}(\theta_0)^T I^{-1}(\theta_0)\dot{\nu}(\theta_0)). \tag{58}$$

(Note this step using Equation 52 helps resolve the potentially difficult integral over $\mathrm{p}(\theta|\boldsymbol{y}^{\text{obs}})$.)

Then, the conditional probability, given $\boldsymbol{y}^{\text{obs}}$ and $\boldsymbol{y}^{\text{new}}$, of $d(\boldsymbol{y}^{\text{rep}}) > d(\boldsymbol{y}^{\text{new}})$ is approximately

$$\mathrm{p}(W > 0|\boldsymbol{y}^{\text{obs}}, \boldsymbol{y}^{\text{new}}), \tag{59}$$

where W is a Gaussian random variable with

$$\mathbb{E}[W|\boldsymbol{y}^{\text{obs}}, \boldsymbol{y}^{\text{new}}] = n^{1/2}\dot{\nu}(\theta_0)^T(\theta(\boldsymbol{y}^{\text{obs}}) - \theta_0) - n^{1/2}[d(\boldsymbol{y}^{\text{new}}) - \nu(\theta_0)], \tag{60}$$

$$\text{Var}(W|\boldsymbol{y}^{\text{obs}}, \boldsymbol{y}^{\text{new}}) = \sigma^2(\theta_0) + \dot{\nu}(\theta_0)^T I^{-1}(\theta_0)\dot{\nu}(\theta_0). \tag{61}$$

Then, the HPC p-value is

$$p(\boldsymbol{y}^{\text{obs}}, \boldsymbol{y}^{\text{new}}) = 1 - \Phi\left(\frac{n^{1/2}\dot{\nu}(\theta_0)^T(\theta(\boldsymbol{y}^{\text{obs}}) - \theta_0) - n^{1/2}[d(\boldsymbol{y}^{\text{new}}) - \nu(\theta_0)]}{\sqrt{\sigma^2(\theta_0) + \dot{\nu}(\theta)^T I^{-1}(\theta_0)\dot{\nu}(\theta)}}\right) + o_{P_0}(1), \tag{62}$$

where we have used Equation 52.

Now, we consider the distribution of the HPC p-value over the distributions of $\boldsymbol{y}^{\text{obs}}$ and $\boldsymbol{y}^{\text{new}}$.

By Equation 53, we have that the posterior mean converges to the true θ_0:

$$\sqrt{n}\dot{\nu}(\theta_0)^T(\theta(\boldsymbol{y}^{\text{obs}}) - \theta_0) \rightsquigarrow N(0, \dot{\nu}(\theta_0)^T I^{-1}(\theta_0)\dot{\nu}(\theta_0)). \tag{63}$$

Independently, $n^{1/2}[d(\boldsymbol{y}^{\text{new}}) - \nu(\theta_0)]$ converges to a $N(0, \sigma^2(\theta_0))$ distribution.

Hence, the HPC p-value is:

$$p(\boldsymbol{y}^{\text{obs}}, \boldsymbol{y}^{\text{new}}) = 1 - \Phi(Q) + o_{P_0}(1), \tag{64}$$

where $Q \sim N(0, 1)$.

This concludes the proof.

B Note on previous version of this paper

The previous version of this paper, Ranganath and Blei (2019), proposed the population predictive check (population predictive check (POP-PC)). The POP-PC treated $\boldsymbol{y}^{\text{new}}$ as random; in the current holdout predictive check, $\boldsymbol{y}^{\text{new}}$ is fixed. Treating $\boldsymbol{y}^{\text{new}}$ as fixed results in a calibrated check, while treating $\boldsymbol{y}^{\text{new}}$ as random does not. For completeness, we include the previous definition of the POP-PC below.

Definition 2 (Population predictive check (Version 1, 2019)) *Consider observed data $\boldsymbol{y}^{\text{obs}}$, its posterior predictive distribution $\text{p}(\boldsymbol{y}^{\text{rep}} \mid \boldsymbol{y}^{\text{obs}})$, and a diagnostic statistic $d(\boldsymbol{y})$. Suppose we have $\boldsymbol{y}^{\text{new}}$ drawn from the population distribution of the data. The population predictive check as a p-value is:*

$$p_{\text{pop-v1}} = \text{p}(d(\boldsymbol{y}^{\text{rep}}) \geq d(\boldsymbol{y}^{\text{new}}) \mid \boldsymbol{y}^{\text{obs}}), \quad \boldsymbol{y}^{\text{new}} \sim F \tag{65}$$

where $\boldsymbol{y}^{\text{rep}} \sim \text{p}(\boldsymbol{y}^{\text{rep}} \mid \boldsymbol{y}^{\text{obs}})$.

To illustrate the issue with treating $\boldsymbol{y}^{\text{new}}$ as random, we again consider the mean example in Section 4. Suppose we observe data $\boldsymbol{y}^{\text{obs}} = \{y_i\}_{i=1}^n$ drawn from a Gaussian with known variance parameter:

$$y_i \sim N(\mu, \sigma^2), \tag{66}$$

for some $\mu \in \mathbb{R}$ and fixed σ^2. For a prior on μ, we take $\mu \sim N(\mu_0, \sigma_0^2)$ with fixed hyperparameters $\mu_0 \in \mathbb{R}, \sigma_0^2 \in \mathbb{R}^+$.

In this case, the posterior predictive distribution is

$$\overline{\boldsymbol{y}}^{\text{rep}} | \overline{\boldsymbol{y}}^{\text{obs}} \sim N\left(\rho_n \overline{\boldsymbol{y}}^{\text{obs}} + (1-\rho_n)\mu_0, (1+\rho_n)\,\frac{\sigma^2}{n}\right). \tag{67}$$

As $n \to \infty$, $\overline{\boldsymbol{y}}^{\text{rep}} | \overline{\boldsymbol{y}}^{\text{obs}}$ is centered around $\overline{\boldsymbol{y}}^{\text{obs}}$. This is different from the distribution of $\boldsymbol{y}^{\text{new}}$, which is the population distribution: $\overline{\boldsymbol{y}}^{\text{new}} \sim N(\mu, \sigma^2)$ (see Figure 10). Consequently, the POP-PC is not calibrated. We demonstrate this lack of calibration theoretically below.

Consider the distribution of POP-PC when Equation 66 holds:

$$\text{p}(d(\boldsymbol{y}^{\text{rep}}) > d(\boldsymbol{y}^{\text{new}}) | \boldsymbol{y}^{\text{obs}}) = \text{p}(D > 0 | \boldsymbol{y}^{\text{obs}}), \tag{68}$$

where $D = d(\boldsymbol{y}^{\text{rep}}) - d(\boldsymbol{y}^{\text{new}})$ is a Gaussian random variable with

$$\text{E}[D | \boldsymbol{y}^{\text{obs}}] = \rho_n \overline{\boldsymbol{y}}^{\text{obs}} + (1-\rho_n)\mu_0 - \mu, \quad Var(D) = \left(2 + \frac{\rho_n}{n}\right)\frac{\sigma^2}{n}. \tag{69}$$

The POP-PC is then:

$$\text{p}(D > 0 | \boldsymbol{y}^{\text{obs}}) = 1 - \Phi\left(\frac{\mu - \rho_n \overline{\boldsymbol{y}}^{\text{obs}} - (1-\rho_n)\mu_0}{\sqrt{(2+\rho_n)\sigma^2/n}}\right) \tag{70}$$

Now, if the observed data has the same distribution as the new data, we have $\overline{\boldsymbol{y}}^{\text{obs}} \sim N(\mu, \sigma^2/n)$. Then, for large n, the POP-PC p-value (Version 1, 2019) is:

$$\text{p}(D > 0 | \boldsymbol{y}^{\text{obs}}) \to 1 - \Phi\left(Z/\sqrt{3}\right), \quad \text{where } Z \sim N(0,1). \tag{71}$$

Consequently, the POP-PC is not calibrated.

The updated definition of the HPC in Definition 1 does not have this calibration issue. Intuitively, this is because a fixed $\boldsymbol{y}^{\text{new}}$ is on average as far from the true mean as the fixed $\boldsymbol{y}^{\text{obs}}$.

For the more general case of asymptotically normal diagnostic functions, we also prove the POP-PC is not uniformly distributed.

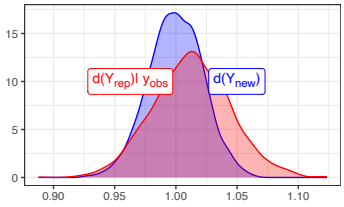


Figure 10: The distribution of $d(\boldsymbol{y}^{\text{new}}) = \overline{y}^{\text{new}}$ is not equal to the distribution of $d(\boldsymbol{y}^{\text{rep}}) = \overline{y}^{\text{rep}}$. (Data drawn as $y_i \sim N(1,1)$, $i = 1, \ldots, 2000$).

Theorem 2 *We assume Equation 16 holds, in addition to regularity conditions detailed in Appendix A.1. Under the distribution $f(\boldsymbol{y}; \theta_0)$, the* POP-PC *p-value can be written as:*

$$p_{\text{pop}}(\boldsymbol{y}) = 1 - \Phi(Q) + o_P(1), \text{ where } Q \sim N\left(0, \frac{\dot{\nu}_\theta(\theta_0)^T I^{-1}(\theta_0)\dot{\nu}_\theta(\theta_0)}{2\sigma^2(\theta_0) + \dot{\nu}_\theta(\theta_0)^T I^{-1}(\theta_0)\dot{\nu}_\theta(\theta_0)}\right), \quad (72)$$

where $o_P(1)$ denotes a random variable converging to zero in probability, Φ is the standard normal cdf, and $Q \sim N(0,1)$.

Consequently, the POP-PC *is not calibrated.*

Proof of Theorem 2. The POP-PC p-value is:

$$p(\boldsymbol{y}^{\text{obs}}) = \int_\Theta \text{p}(d(\boldsymbol{y}^{\text{rep}}) > d(\boldsymbol{y}^{\text{new}})|\theta)\text{p}(\theta|\boldsymbol{y}^{\text{obs}})\text{p}(\boldsymbol{y}^{\text{new}}|\theta_0)d\theta \quad (73)$$

Consider:

$$d(\boldsymbol{y}^{\text{rep}}) > d(\boldsymbol{y}^{\text{new}}) \quad (74)$$

Then,
$$\sqrt{n}[d(\boldsymbol{y}^{\text{rep}}) - \nu(\theta)] > \sqrt{n}[d(\boldsymbol{y}^{\text{new}}) - \nu(\theta_0) - (\nu(\theta) - \nu(\theta_0))]. \quad (75)$$

Consider the LHS of Equation 75. By Equation 48, we have

$$\sqrt{n}[d(\boldsymbol{y}^{\text{rep}}) - \nu(\theta)] \sim N(0, \sigma^2(\theta_0)).$$

Consider now the RHS of Equation 75. The first term is distributed as:

$$\sqrt{n}[d(\boldsymbol{y}^{\text{new}}) - \nu(\theta_0)] \sim N(0, \sigma^2(\theta_0)). \quad (76)$$

For the second term, we use a Taylor expansion to obtain:

$$\nu(\theta) - \nu(\theta_0) \approx \dot{\nu}(\theta_0)(\theta - \theta_0). \quad (77)$$

Then by Equation 52, given $\boldsymbol{y}^{\text{obs}}$, the second term is approximately distributed as

$$N(\dot{\nu}(\theta_0)^T n^{1/2}(\theta(\boldsymbol{y}^{\text{obs}}) - \theta_0), \dot{\nu}(\theta_0)^T I^{-1}(\theta_0)\dot{\nu}(\theta_0)). \quad (78)$$

(Note this step using Equation 52 helps resolve the potentially difficult integral over $\mathrm{p}(\theta|\boldsymbol{y}^{\mathrm{obs}})$.)

Then, the conditional probability, given $\boldsymbol{y}^{\mathrm{obs}}$, of $d(\boldsymbol{y}^{\mathrm{rep}}) > d(\boldsymbol{y}^{\mathrm{new}})$ is approximately

$$\mathrm{p}(W > 0|\boldsymbol{y}^{\mathrm{obs}}), \tag{79}$$

where W is a Gaussian random variable with

$$\mathbb{E}[W|\boldsymbol{y}^{\mathrm{obs}}] = n^{1/2}\dot{\nu}(\theta_0)^T(\theta(\boldsymbol{y}^{\mathrm{obs}}) - \theta_0), \tag{80}$$

$$\mathrm{Var}(W|\boldsymbol{y}^{\mathrm{obs}}) = 2\sigma^2(\theta_0) + \dot{\nu}(\theta_0)^T I^{-1}(\theta_0)\dot{\nu}(\theta_0). \tag{81}$$

Then, the POP-PC p-value is

$$p(\boldsymbol{y}^{\mathrm{obs}}) = 1 - \Phi\left(\frac{n^{1/2}\dot{\nu}(\theta_0)^T(\theta(\boldsymbol{y}^{\mathrm{obs}}) - \theta_0)}{\sqrt{2\sigma^2(\theta_0) + \dot{\nu}(\theta)^T I^{-1}(\theta_0)\dot{\nu}(\theta)}}\right) + o_{P_0}(1), \tag{82}$$

where we have used Equation 52.

Now, we consider the distribution of the POP-PC p-value over the distribution of $\boldsymbol{y}^{\mathrm{obs}}$.

By Equation 53, we have that the posterior mean converges to the true θ_0:

$$\sqrt{n}\dot{\nu}(\theta_0)^T(\theta(\boldsymbol{y}^{\mathrm{obs}}) - \theta_0) \rightsquigarrow N(0, \dot{\nu}(\theta_0)^T I^{-1}(\theta_0)\dot{\nu}(\theta_0)). \tag{83}$$

Hence, the POP-PC p-value is:

$$p(\boldsymbol{y}^{\mathrm{obs}}) = 1 - \Phi(Q) + o_{P_0}(1), \tag{84}$$

where Q is a Gaussian random variable with mean 0 and variance

$$\mathrm{Var}(Q) = \frac{\dot{\nu}(\theta_0)^T I^{-1}(\theta_0)\dot{\nu}(\theta_0)}{2\sigma^2(\theta_0) + \dot{\nu}(\theta_0)^T I^{-1}(\theta)\dot{\nu}(\theta_0)}. \tag{85}$$

This concludes the proof.

C Details for Section 4

Suppose we observe data $\boldsymbol{y} = \{y_i\}_{i=1}^n$, $y_i \sim N(\mu, \sigma^2)$ where σ^2 is known. The prior is taken to be $\mu \sim N(\mu_0, \sigma_0^2)$.

We consider the posterior predictive p-value with diagnostic $d(\boldsymbol{y}) = \overline{\boldsymbol{y}}$. The posterior predictive distribution of $\overline{\boldsymbol{y}}^{\mathrm{rep}}$ is:

$$\overline{\boldsymbol{y}}^{\mathrm{rep}}|\overline{\boldsymbol{y}}^{\mathrm{obs}} = \int \mathrm{p}(\overline{\boldsymbol{y}}^{\mathrm{rep}}|\theta)\mathrm{p}(\theta|\overline{\boldsymbol{y}}^{\mathrm{obs}})d\theta \tag{86}$$

$$\sim N\left(\rho_n\overline{\boldsymbol{y}}^{\mathrm{obs}} + (1-\rho_n)\mu_0, (1+\rho_n)\frac{\sigma^2}{n}\right) \tag{87}$$

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

Then, the posterior predictive p-value is:

$$\mathrm{p}(d(\boldsymbol{y}_{rep}) > d(\boldsymbol{y}^{\text{obs}})|\boldsymbol{y}^{\text{obs}})) = 1 - \Phi\left(\frac{\overline{\boldsymbol{y}}^{\text{obs}} - \rho_n\overline{\boldsymbol{y}}^{\text{obs}} - (1-\rho_n)\mu_0}{\sqrt{(1+\rho_n)\sigma^2/n}}\right). \tag{88}$$

D Discussion of the partial predictive check

An alternative method to obtain calibrated p-values is the partial predictive check of Bayarri and Berger (2000). The partial PC achieves calibration by calculating a conditional posterior predictive that is independent of the diagnostic. However, when the diagnostic includes the sufficient statistics of the model, the partial predictive check essentially becomes a prior predictive check. To illustrate this point, consider the partial predictive check for the test in Section 4. The partial predictive check uses the following predictive distribution:

$$\mathrm{p}(\boldsymbol{y}^{\text{rep}}|\mu)\mathrm{p}(\mu|\boldsymbol{y}^{\text{obs}}\backslash d(\boldsymbol{y}^{\text{obs}})), \quad \text{where} \quad \mathrm{p}(\mu|\boldsymbol{y}^{\text{obs}}\backslash d(\boldsymbol{y}^{\text{obs}})) \propto \frac{\mathrm{p}(\boldsymbol{y}^{\text{obs}}|\mu)\mathrm{p}(\mu)}{\mathrm{p}(d(\boldsymbol{y}^{\text{obs}})|\mu)}. \tag{89}$$

In this example, $\mathrm{p}(\mu|\boldsymbol{y}^{\text{obs}}\backslash\overline{\boldsymbol{y}}^{\text{obs}})$ is simply the prior on μ, as we are removing the influence of the sufficient statistic, $\overline{\boldsymbol{y}}^{\text{obs}}$. Then, the distribution of the partial posterior predictive is

$$\overline{\boldsymbol{y}}^{\text{rep}}|\boldsymbol{y}^{\text{obs}}\backslash d(\boldsymbol{y}^{\text{obs}}) \sim N\left(\mu_0, (\sigma_0^2 + \sigma^2)/n\right). \tag{90}$$

The partial predictive p-value is then:

$$\mathrm{p}(d(\boldsymbol{y}^{\text{rep}}) > d(\boldsymbol{y}^{\text{obs}})|\boldsymbol{y}^{\text{obs}}\backslash\overline{\boldsymbol{y}}^{\text{obs}}) = 1 - \Phi\left(\frac{\overline{\boldsymbol{y}}^{\text{obs}} - \mu_0}{\sqrt{(\sigma_0^2 + \sigma^2)/n}}\right). \tag{91}$$

This is the prior predictive p-value. That is, if the diagnostic is the only sufficient statistic, the partial predictive p-value coincides with the prior predictive p-value, which is not calibrated. This does not contradict Robins et al. (2000); Bayarri and Berger (2000), however, who prove the partial predictive check is calibrated under certain assumptions. One of these assumptions is that the parameters of the predictive distribution converge to the MLE, which is not the case here.

However, the partial posterior check is similar to the HPC when the partial diagnostic is defined to use a subset of the data. In the above example, we could choose $d(\boldsymbol{y}^{\text{obs}}) = 1/(n/2)\sum_{i=1}^{n/2} y_i^{\text{obs}} =: \overline{\boldsymbol{y}}_{1:n/2}^{\text{obs}}$ In this case, the conditional posterior is:

$$\mathrm{p}(\mu|\boldsymbol{y}^{\text{obs}}\backslash d(\boldsymbol{y}^{\text{obs}})) \propto \frac{\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left[y_i^{\text{obs}} - \mu\right]^2\right\}}{\exp\left\{-\frac{n}{4}\left[\overline{\boldsymbol{y}}_{1:n/2} - \mu\right]^2\right\}}\mathrm{p}(\mu) \tag{92}$$

$$\propto \prod_{i=n/2+1}^{n} \mathrm{p}(y_i^{\text{obs}}|\mu)\mathrm{p}(\mu). \tag{93}$$

For this diagnostic, the partial posterior check is equivalent to the HPC.

In certain cases, the partial posterior check with a data-split diagnostic may be more efficient in its use of the data than HPC. This is because the HPC will always use a subset of the data for posterior inference. A partial predictive check meanwhile may only remove a sufficient statistic of this subset of the data. Despite this, the partial predictive check can be difficult to calculate, and requires re-calculation for each diagnostic function. Meanwhile, a HPC is simple to implement, and the inferred posterior can be used to check many different diagnostic functions.

E Additional details for Section 5

In this section, we provide additional plots for the regression study in Section 5.1.

- Figure 11 shows the empirical distribution of PPC p-values. The p-values are concentrated around 0.5 for all values of c.
- Figure 12 shows the empirical distribution of PPC p-values calibrated following Robins et al. (2000). While the p-values are uniform, they show a similar distribution for all values of c and so are unable to detect model misfit.
- Figure 13 shows the empirical distribution of cppp p-values (Hjort et al., 2006). The cppp p-values do not detect model misfit for large c.
- Figure 6 shows the empirical distribution of the HPC p-values. For small values of c, the HPC p-values are approximately uniform. For large values of c, the HPC p-values concentrate around 0 (i.e. the HPC will always reject the model when there is insufficient regularization).

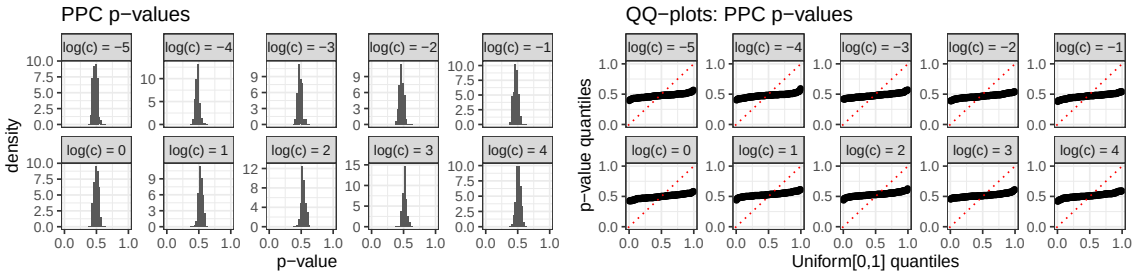


Figure 11: PPC p-values are concentrated around 0.5 for all values of the regularization parameter c. Left: Histograms of PPC p-values. Right: QQ-plots comparing the quantiles of the PPC p-values with the quantiles of a uniform[0,1] random variable.

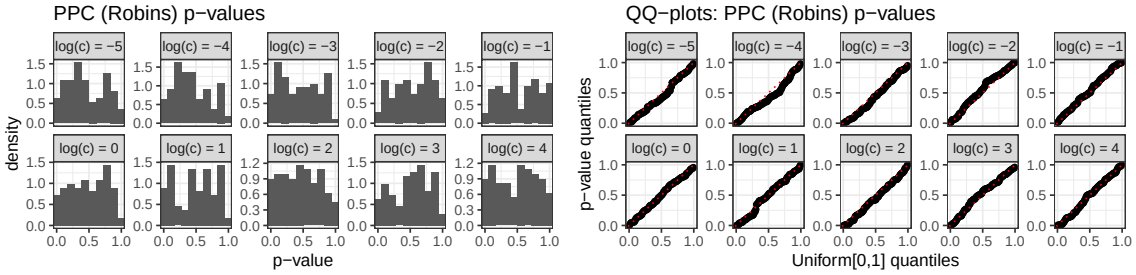


Figure 12: Calibrated PPC p-values (Robins et al., 2000) are approximately uniformly distributed but they do not detect overfitting for large values of the regularization parameter c. Left: Histograms of calibrated PPC p-values. Right: QQ-plots comparing the quantiles of the calibrated PPC p-values with the quantiles of a uniform[0,1] random variable. Plots are over $K = 100$ replications of the data.

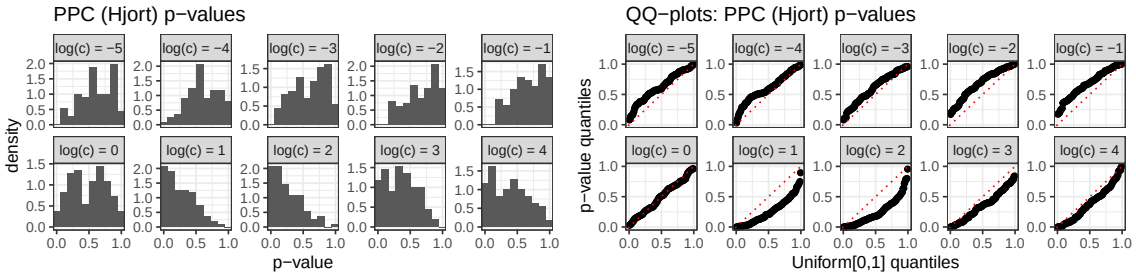


Figure 13: Calibrated cppp p-values (Hjort et al., 2006) do not detect overfitting for large values of the regularization parameter c. Left: Histograms of cppp p-values. Right: QQ-plots comparing the quantiles of the cppp p-values with the quantiles of a uniform[0,1] random variable. Plots are over $K = 100$ replications of the data.