Bayesian model criticism using uniform parametrization checks

Christian T. Covington* Jeffrey W. Miller†

Abstract

Models are often misspecified in practice, making model criticism a key part of Bayesian analysis. It is important to detect not only when a model is wrong, but which aspects are wrong, and to do so in a computationally convenient and statistically rigorous way. We introduce a novel method for model criticism based on the fact that if the parameters are drawn from the prior, and the dataset is generated according to the assumed likelihood, then a sample from the posterior will be distributed according to the prior. Thus, departures from the assumed likelihood or prior can be detected by testing whether a posterior sample could plausibly have been generated by the prior. Building upon this idea, we propose to reparametrize all random elements of the likelihood and prior in terms of independent uniform random variables, or u-values. This makes it possible to aggregate across arbitrary subsets of the u-values for data points and parameters to test for model departures using classical hypothesis tests for dependence or non-uniformity. We demonstrate empirically how this method of uniform parametrization checks (UPCs) facilitates model criticism in several examples, and we develop supporting theoretical results.

1 Introduction

Bayesian statistics proceeds by defining a model—consisting of a prior and likelihood—and drawing posterior inferences based on the assumption that this model is correct. However, if the model is not correct then the resulting inferences may be misleading. In practice, it can be difficult to know whether a model is sufficiently accurate to provide reliable inferences and, if not, which aspects of the model need improvement. The task of detecting a model's inadequacies is called "model criticism" (Box, 1980; Gelman et al., 2013; Blei, 2014).

While many methods have been proposed for Bayesian model criticism (see Section 3), posterior predictive checks (PPCs) are currently the most commonly used technique. PPCs compare the observed value of a test statistic—or more generally, a test quantity that may depend on the data and parameters—to its distribution under the posterior predictive (Guttman, 1967; Rubin, 1984; Meng, 1994; Gelman et al., 2013).

*Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, ccovington@g.harvard.edu

†Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, jwmiller@hsph.harvard.edu

However, it is well known that the "p-values" produced by PPCs are not valid, in the sense that they are not uniformly distributed under the null hypothesis that the model is correct, even asymptotically (Bayarri and Berger, 1999, 2000; Robins et al., 2000). Valid PPC p-values can be obtained by using a partial posterior or conditional predictive (Bayarri and Berger, 1999, 2000), however, these may be hard to implement in many models. Alternatively, Moran et al. (2019) and Li and Huggins (2022) propose using data splitting to obtain valid PPC p-values, however, this (i) entails a loss of information since the split-data posterior does not use the full dataset, and (ii) typically involves several posterior inference runs over various splits. Furthermore, an even more fundamental difficulty of using PPCs is that they require one to design test quantities to detect the various types of misspecification of concern in the model at hand. There are infinitely many possible test quantities, and choosing which ones to consider requires (i) confidence about the kinds of misspecification that may be present and (ii) statistical insight into what makes a good PPC test quantity, including subtle considerations of sufficiency and ancillarity (Gelman et al., 2013; Mimno et al., 2015; Bolsinova and Tijmstra, 2016). This puts a major burden on the analyst, hindering the adoption of PPCs in practice.

In this paper, we introduce a novel method for model criticism that overcomes these limitations. The method is inspired by the observation that if the parameters θ are drawn from the prior, a dataset $Y \mid \theta$ is drawn according to the assumed likelihood, and $\widetilde{\theta} \mid Y$ is drawn from the posterior, then the marginal distribution of $\widetilde{\theta}$ (integrating out θ and Y) is equal to the prior. Thus, if a posterior sample could not plausibly have been generated by the prior, then this indicates misspecification of some part of the model likelihood or prior (Gelman et al., 2013, Section 6.4). Building upon this basic idea, we propose to reparametrize all random elements in the model—including data and parameters—in terms of independent and identically distributed uniform random variables ("u-values") on the unit interval. Then, under the null that the model is correct, a posterior sample of the u-values is exactly distributed as i.i.d. uniform. This enables one to easily perform hypothesis tests probing for misspecification of various parts of the model simply by testing for departures from independence and uniformity of the u-values in various ways. It is important to emphasize that *multiple* posterior samples given the same dataset will *not* be independent, because the dataset introduces dependence across samples. Thus, to use multiple posterior samples—for example, from a Markov chain Monte Carlo (MCMC) run—we use a p-value aggregation technique for dependent p-values.

The proposed method, referred to as "uniform parametrization checks" (UPCs), shares a number of attractive features with PPCs. Like PPCs, UPCs help reveal not only whether a model is wrong, but which parts are wrong, and how to improve it. Also like PPCs, UPCs are computationally tractable and easy to implement, requiring only one posterior inference run (for example, using MCMC) that yields samples from the standard posterior given the observed data. And like PPCs, UPCs are applicable to a very wide range of models.

Additionally, UPCs have several advantages compared to PPCs. First, UPCs yield uniform p-values under the null that the model is correct, since—up to posterior inference approximations—the posterior u-values are exactly i.i.d. uniform when the model is correct. Thus, Type I error rate is correctly controlled—regardless of the dataset size—and one can even perform iterative model building in a principled way via alpha spending. Second, there are natural default choices of UPC tests that apply to any model, and it is usually intuitively clear how to design new customized tests. Third, in contrast to PPCs, which only provide model criticism based on a collection of selected statistics, UPCs provide a more comprehensive understanding of where a model is going wrong, since we know the exact joint distribution of all u-values under the null.

The paper is organized as follows. In Section 2, we introduce the UPC methodology. Section 3 briefly reviews previous work, and in Section 4, we establish theoretical properties of UPCs. In Section 5, we demonstrate the UPC method in several examples involving real and simulated data. Finally, Section 6 concludes with a brief discussion and directions for future work.

2 Methodology

Suppose Π is a prior distribution on the parameter θ and P_θ is a hypothesized distribution of the dataset Y given θ. Letting $\theta \sim \Pi$ and $Y \sim P_\theta$ given θ, this defines a joint distribution on parameters and data, (θ, Y), which we refer to as the *hypothesized model*. Assume we can write $(\theta, Y) = g(U)$ where g is a known function and $U \sim \text{Uniform}_D(0,1)$, that is, $U = (U_1, \ldots, U_D)$ and $U_1, \ldots, U_D$ i.i.d. $\sim \text{Uniform}(0,1)$. Most Bayesian models used in practice can be written in this way, including, for example, complex hierarchical models with continuous and discrete variables and identifiability constraints. We refer to $U_1, \ldots, U_D$ as the *u-values*. For simplicity, we abuse terminology slightly by referring to P_θ as the likelihood.

To perform model criticism, we view the hypothesized model as the null hypothesis. To test for departures from this hypothesis, we sample from the posterior of U, that is, from the conditional distribution $U|Y$ that arises from the joint distribution of (U, Y) defined by the assumptions that $(\theta, Y) = g(U)$ and $U \sim \text{Uniform}_D(0,1)$. Sampling from $U|Y$ can either be done by sampling $\theta|Y$ and then $U|\theta, Y$, or by directly targeting $U|Y$; either option can often be implemented by taking an algorithm for sampling from $\theta|Y$ and making minor modifications (see Section 2.6). Now, the key observation is that if Y is sampled from the hypothesized model and U is sampled from $U|Y$, then $U \sim \text{Uniform}_D(0,1)$ marginally, integrating out Y. In other words, if the model is correct, then a single posterior sample of $U = (U_1, \ldots, U_D)$ is uniformly distributed, that is, $U_1, \ldots, U_D$ i.i.d. $\sim \text{Uniform}(0,1)$. There is no approximation here – if the model is correct, then a posterior draw of U is exactly uniform, in complete generality. Consequently, if we detect that $U_1, \ldots, U_D$ are not i.i.d. $\text{Uniform}(0,1)$ under the posterior, then this implies misspecification of some

aspect of the model, that is, there is mismatch between the true distribution and some aspect of the prior Π or likelihood P_θ. This enables one to perform model criticism in a simple yet powerful way by testing for departures from uniformity or independence in various respects; see Section 2.1.

For posterior inference, we typically draw multiple posterior samples, say, $U^{(1)}, \ldots, U^{(T)} \in (0,1)^D$, where each $U^{(t)} = (U_1^{(t)}, \ldots, U_D^{(t)}) \in (0,1)^D$ is drawn from $U|Y$. A subtle but crucial point is that, while $U_1^{(t)}, \ldots, U_D^{(t)}$ are i.i.d. Uniform$(0,1)$ for any given t, it does not hold that $U_d^{(t)} \sim$ Uniform$(0,1)$ i.i.d. for all t and d. This is because the dataset Y creates dependence among the samples $U^{(1)}, \ldots, U^{(T)}$. If an independent dataset $Y^{(t)}$ were used to generate each $U^{(t)}$, then $U^{(1)}, \ldots, U^{(T)}$ would indeed be independent, but this is not the case since we only observe one dataset Y. In Section 2.3, we describe how to combine across samples.

Example 2.1. *AR(1) model. Consider an autoregression model where* $Y_1 = \sigma\varepsilon_1$ *and* $Y_i = \phi Y_{i-1} + \sigma\varepsilon_i$ *for* $i = 2, \ldots, n$, *where* $\varepsilon_1, \ldots, \varepsilon_n$ *i.i.d.* $\sim \mathcal{N}(0,1)$, *with prior* $\phi \sim$ Uniform$(-0.5, 0.5)$ *and* $\sigma \sim$ Exponential(1) *independently. Then we can write* $(\phi, \sigma, Y_1, \ldots, Y_n) = g(U_1, \ldots, U_{2+n})$, *where* $U_1, \ldots, U_{2+n}$ *i.i.d.* $\sim$ Uniform$(0,1)$, *by setting* $\phi = F_\phi^{-1}(U_1)$, $\sigma = F_\sigma^{-1}(U_2)$, $\varepsilon_i = \Phi^{-1}(U_{d_i})$, $Y_1 = \sigma\varepsilon_1$, *and* $Y_i = \phi Y_{i-1} + \sigma\varepsilon_i$ *for* $i = 2, \ldots, n$, *where* F_ϕ, F_σ, *and* Φ *denote the cumulative distribution functions (CDFs) of the* Uniform$(-0.5, 0.5)$, Exponential(1), *and* $\mathcal{N}(0,1)$ *distributions, respectively, and* $d_i = 2 + i$.

2.1 Choice of tests

In terms of model criticism, one advantage of using a uniform parametrization is that it greatly simplifies the construction of tests. In particular, since any subset of u-values is i.i.d. uniform under the null hypothesis that the model is correct, (i) the u-values can be grouped in various ways and (ii) the same tests can be applied to any model. Some natural choices of test are as follows. In each case, the test is performed on the u-values $U_1, \ldots, U_D$ from a single posterior sample; see Section 2.3 for combining across samples.

1. *Testing for extreme values.* Outliers or poor choices of prior can often be detected by looking at individual u-values. For instance, in the AR(1) example, if U_2 is very close to 1 then this indicates that the inferred value of σ is much larger than expected under the Exponential(1) prior. To test for extreme values, we use $2\min\{U_d, 1 - U_d\}$ as a p-value; note that this is Uniform$(0,1)$ under the null.

2. *Testing for non-uniformity.* Misspecification of model distributions can often be detected by testing for non-uniformity of a relevant subset of u-values. In the AR(1) example, if the empirical distribution of $U_{d_1}, \ldots, U_{d_n}$ is significantly non-uniform (where $d_i = 2+i$), then this suggests possible misspecification of either the normal distribution or the first-order linear assumption encoded in the equation $Y_i =$

$\phi Y_{i-1} + \sigma\varepsilon_i$. To test for non-uniformity, we use the Anderson–Darling test with Uniform$(0, 1)$ as the null hypothesis (Anderson and Darling, 1954), but any goodness-of-fit test could be used.

3. *Testing for internal dependence.* Structural misspecification can sometimes be detected by testing for dependence between relevant subsets of u-values. For instance, in the AR(1) example, dependence between U_{d_i} and U_{d_i+2} (for $i = 1, \ldots, n-2$) suggests that higher-order dependence is present, rather than just the first-order dependence assumed in the AR(1) model. To test for internal dependence between sets of u-values, we use Hoeffding's test (Hoeffding, 1948).

4. *Testing for external dependence.* Missing structural features can sometimes be detected by testing for dependence with external variables such as covariates. In the AR(1) example, dependence between U_{d_i} and the index i (for $i = 1, \ldots, n$) suggests an unmodeled dependence on time, such as heteroskedasticity or a trend. More generally, if x_i is a covariate, then dependence between U_{d_i} and x_i suggests the presence of unmodeled dependence on the covariate. To test for dependence (i) between two continuous variables, we use Hoeffding's test (Hoeffding, 1948); (ii) between continuous and binary variables, we use the Mann–Whitney U test (Mann and Whitney, 1947); and (iii) between continuous and non-binary discrete variables, we use the Kruskal–Wallis H test (Kruskal and Wallis, 1952).

See Section 5 for detailed illustrations, and see Section 5.4 in particular for the AR(1) example. Typically, one will want to perform a number of tests, so it is necessary to adjust for multiple testing, for instance, using the Bonferroni, Holm, or Benjamini–Hochberg procedures; see Section 2.4.

2.2 Choice of uniform parametrization

An attractive feature of UPCs is the intuitive way in which a test can connect with the type of misspecification indicated, however, the choice of parametrization plays an important role in this connection. There are many possible ways to uniformly parametrize a given model, since there is not a unique choice of function g such that $(\theta, Y) = g(U)$. To see why, note that there are many distribution-preserving operations on U, such as mapping U_d to $1 - U_d$ as a simple example. Usually, though, the model specification will suggest a natural way of defining g by following the generative process that defines the model; see Example 2.1 and more examples in Section 5. It is less clear how to choose g when there are constraints on θ or Y. For handling constraints, we generally recommend defining a u-value for each interpretable univariate quantity and then mapping these through a function that enforces the constraint, rather than defining u-values for only a subset of variables and then setting the remaining variables to satisfy the constraint. For example, in regression models, it is common to have constraints of the form $\sum_{j=1}^{J} \alpha_j = 0$, for which we suggest

defining $\tilde{\alpha}_j = F_j^{-1}(U_j)$ and $\alpha_j = \tilde{\alpha}_j - \frac{1}{J}\sum_{j=1}^{J} \tilde{\alpha}_j$, for appropriately chosen CDFs F_j. While this leads to non-identifiability of these u-values, it improves their interpretability.

2.3 Combining multiple posterior samples

So far, we have described UPC tests based on a single posterior sample of $U = (U_1, \ldots, U_D) \in (0,1)^D$ given the dataset Y. Since there is randomness in any one posterior sample, it is preferable to aggregate across many posterior samples $U^{(1)}, \ldots, U^{(T)} \in (0,1)^D$ drawn from the conditional distribution of $U|Y$. However, care must be taken to combine these in a valid way, because $U^{(1)}, \ldots, U^{(T)}$ are not marginally independent, integrating out Y. Thus, one cannot simply pool all of the u-values together to perform tests.

To understand why, consider the posterior of the ϕ parameter in the AR(1) model in Example 2.1, given a random dataset Y generated according to the hypothesized model for (θ, Y). When the number of data points n is large, the posterior of ϕ will tend to be concentrated. Likewise, the posterior of the corresponding u-value U_1 will also be concentrated. Hence, the posterior samples $U_1^{(1)}, \ldots, U_1^{(T)}$ will tend to be clustered together, and thus, for any given dataset Y, their empirical distribution will clearly not be close to uniform. This is not a contradiction: Each $U_1^{(t)}$ is marginally Uniform$(0,1)$ when integrating out Y, but the dependence among $U_1^{(1)}, \ldots, U_1^{(T)}$ induced by Y makes them tend to take similar values. See Figures 5.2 and 5.5 for illustrations of this effect in examples.

We use the following approach to combine posterior samples in a valid way for any given test. For each $t = 1, \ldots, T$, we perform the test on the posterior sample $U^{(t)} = (U_1^{(t)}, \ldots, U_D^{(t)})$ to obtain a p-value $p^{(t)}$. Under the null that the model is correct, we know $U_1^{(t)}, \ldots, U_D^{(t)}$ i.i.d. $\sim$ Uniform$(0,1)$, so a valid test will produce a uniformly distributed p-value $p^{(t)} \sim$ Uniform$(0,1)$. Then, we combine the p-values $p^{(1)}, \ldots, p^{(T)}$ using the Cauchy combination method for dependent p-values (Liu and Xie, 2020). Specifically, we compute

$$p^* = 1 - F_{\text{Cauchy}}\Big(\frac{1}{T}\sum_{t=1}^{T} \tan\big((0.5 - p^{(t)})\pi\big)\Big), \tag{2.1}$$

where F_{Cauchy} is the CDF of the Cauchy distribution. We then compare the aggregated p-value p^* to a pre-specified level α to decide whether to reject the null that the model is correct. Although p^* is not uniformly distributed on all of $(0,1)$ under the null, empirically we find that $\mathbb{P}(p^* \leqslant \alpha) \approx \alpha$ for $\alpha \in (0, 0.05)$. Thus, for practically relevant values of α, the Type I error is near the target level. Several methods for aggregating dependent p-values have been proposed—for instance, the method of Gasparin et al. (2024) yields similar results on the examples we consider—but overall we find that the Cauchy combination method tends to exhibit the greatest power while still controlling Type I error rate.

2.4 Performing multiple tests

To perform multiple different UPC tests, we apply the p-value aggregation technique in Section 2.3 for each test, yielding aggregated p-values $p_1^*, \ldots, p_M^*$. Standard multiple testing adjustment procedures can then be applied to $p_1^*, \ldots, p_M^*$. For instance, one can control family-wise error rate (FWER) with the Bonferroni or Holm procedures (Holm, 1979). One can control false discovery rate (FDR) with the procedure of Benjamini and Hochberg (1995) in the case of independent tests, which holds when the tests are computed from disjoint sets of u-values; more generally, the procedure of Benjamini and Yekutieli (2001) can be used in the case of dependent tests.

Furthermore, UPCs can be used to iteratively criticize and improve the model (Box, 1980; Blei, 2014) in a principled way that controls Type I error. Specifically, one can use *alpha spending*, in which the rejection thresholds $\alpha_1, \ldots, \alpha_M$ for a sequence of tests are pre-selected such that they satisfy $\alpha = \sum_{m=1}^{M} \alpha_m$, where α is the overall FWER that one wishes to permit. We illustrate with a logistic regression model in Section 5.3. Similarly, *alpha investing* can be used to control FDR for a sequence of tests (Foster and Stine, 2008).

2.5 Interpretation of tests

When there is an explicit generative description of the model, defining $(\theta, Y) = g(U)$ accordingly provides a natural correspondence between the entries of U and the entries of θ and Y, which aids in the interpretation of the u-values and tests. As described in Section 2.1, a rejection of the null under a test for extreme values, non-uniformity, internal dependence, or external dependence suggests a possible issue with the corresponding part of the model from which the tested u-values arise. However, the interpretation is not always straightforward, since misspecification of one part of a model can lead to departures from i.i.d. uniformity of u-values in other parts of the model.

For instance, in the AR(1) example, suppose the true distribution of $\varepsilon_1, \ldots, \varepsilon_n$ is Cauchy rather than $\mathcal{N}(0, 1)$. Then a posterior sample of the u-values corresponding to $\varepsilon_1, \ldots, \varepsilon_n$ will exhibit non-uniformity, and outliers with u-values close to 0 or 1 will likely be observed. However, σ will also tend to be wildly overestimated, causing the σ u-value to be very close to 1. Thus, there is not always a direct link between departure from i.i.d. uniformity of u-values and misspecification of the corresponding part of the model. Nonetheless, empirically we find that the strongest departures from i.i.d. uniformity tend to correspond to the aspects of the model that are misspecified.

2.6 Computation of u-values

In this section, we assume the following condition on the definition of the function g.

Condition 2.2. *$\theta = g_{\mathrm{p}}(U_{1:K})$ and $Y = g_{\mathrm{d}}(U_{K+1:D}; \theta)$ for some K and some functions g_{p} and g_{d}.*

When this holds, we refer to $U_{1:K}$ as the *parameter u-values* and $U_{K+1:D}$ as the *data u-values*. Sampling the parameter u-values from $U_{1:K}|Y$ can either be done directly with standard Bayesian techniques or by post-processing from samples of θ; see Section 2.6.1 for details. Sampling the data u-values from $U_{K+1:D} \mid U_{1:K}, Y$ requires post-processing; see Section 2.6.2. In the AR(1) model (Example 2.1), U_1 and U_2 are the parameter u-values and $U_3, \ldots, U_{2+n}$ are the data u-values.

2.6.1 Computation of parameter u-values

One method of sampling $U_{1:K}|Y$ is simply to reparametrize – that is, to view $U_{1:K} \sim \mathrm{Uniform}_K(0,1)$ as the prior and $Y|U_{1:K}$ as defining the likelihood. Letting $\mathcal{L}(\theta; Y)$ denote the likelihood function of P_θ for data Y, the posterior is then $\pi(u_{1:K} \mid Y) \propto \mathcal{L}(g_{\mathrm{p}}(u_{1:K}); Y)$ since the prior is uniform. Thus, when using algorithms based directly on the target density, such as Hamiltonian Monte Carlo as in Stan and PyMC (Carpenter et al., 2017b; Abril-Pla et al., 2023) or Langevin algorithms (Roberts and Tweedie, 1996), it is trivial to modify an MCMC sampler targeting $\theta|Y$ to instead target $U_{1:K}|Y$. For instance, in the AR(1) example, the posterior of the parameter u-values is $\pi(u_1, u_2 \mid Y) \propto \prod_{i=1}^n \mathcal{N}(Y_i \mid F_\phi^{-1}(u_1) Y_{i-1},\, F_\sigma^{-1}(u_2)^2)$, if we define $Y_0 = 0$.

Sometimes this reparametrization approach might be undesirable, for instance, if the model is not conducive to sampling based on evaluation of the posterior density or if one wishes to use an existing MCMC algorithm for sampling $\theta|Y$. In such cases, one can first sample $\theta|Y$ and then sample $U_{1:K}|\theta, Y$ in a post-processing step. Recall that $U_{1:K} \sim \mathrm{Uniform}_K(0,1)$ and, in this section, we assume $\theta = g_{\mathrm{p}}(U_{1:K})$. The simplest situation is when g_{p} is an invertible function, in which case we can deterministically transform each posterior sample of θ into a sample of the parameter u-values via $U_{1:K} = g_{\mathrm{p}}^{-1}(\theta)$. This is the case in the AR(1) model in Example 2.1, for which $U_1 = F_\phi(\phi)$ and $U_2 = F_\sigma(\sigma)$.

Often, however, $\theta = g_{\mathrm{p}}(U_{1:K})$ is not invertible, such as when there are discrete latent variables or identifiability constraints. Then one can stochastically generate the parameter u-values from the conditional distribution $U_{1:K}|\theta, Y$. This is straightforward when the non-invertibility is solely due to discreteness of one or more entries of θ, since then $U_{1:K}|\theta, Y$ is uniformly distributed subject to the constraint that $\theta = g_{\mathrm{p}}(U_{1:K})$. A common situation is that $\theta = (\theta_1, \ldots, \theta_K)$ and there are functions $g_1, \ldots, g_K$ such that $\theta_1 = g_1(U_1)$ and $\theta_k = g_k(\theta_1, \ldots, \theta_{k-1}, U_k)$ for $k = 2, \ldots, K$. Then a sample of $U_{1:K}|\theta, Y$ can be obtained by sampling $U_k \mid \theta_{1:k}$ sequentially for $k = 1, \ldots, K$. For instance, if θ_k is discrete then we draw $U_k \sim \mathrm{Uniform}(\{u \in (0,1) : \theta_k = g_k(\theta_1, \ldots, \theta_{k-1}, u)\})$; this is equivalent to the randomized probability integral transform described by Czado et al. (2009). We use this technique for the component assignment variables in the mixture model in

Section 5.3. More generally, when g_{p} is non-invertible due to more than just discreteness of latent variables, care must be taken to determine valid conditional distributions for $U_{1:K}|\theta, Y$; see Chang and Pollard (1997) for a general framework for conditioning.

2.6.2 Computation of data u-values

Computing the data u-values is very similar to computing the parameter u-values. First, if $Y = g_{\mathrm{d}}(U_{K+1:D}; \theta)$ is invertible (as a function from $U_{K+1:D}$ to Y) for all θ, then we can simply transform the data into the data u-values via $U_{K+1:D} = g_{\mathrm{d}}^{-1}(Y; \theta)$ for any given posterior sample of θ. Again, this is the case in the AR(1) model, where we have $\theta = (\phi, \sigma)$ and $U_{d_i} = \Phi\big((Y_i - \phi Y_{i-1})/\sigma\big)$ for $i = 1, \ldots, n$, where $Y_0 = 0$ and $d_i = 2 + i$.

On the other hand, if g_{d} is not invertible, then just like in the case of the parameter u-values, we sample from $U_{K+1:D} \mid \theta, Y, U_{1:K}$. Similar to before, a common situation is that for some ordering of the univariate entries of the data, say, $Y = (Y_1, \ldots, Y_n)$, there are functions $g_{K+1}, \ldots, g_{K+n}$ such that $Y_i = g_{K+i}(Y_1, \ldots, Y_{i-1}, U_{K+i}; \theta)$ for $i = 1, \ldots, n$, where $K + n = D$. Then we can sample from the joint distribution of the data u-values $U_{K+1:D} \mid \theta, Y, U_{1:K}$ by drawing $U_{K+i} \mid \theta, Y_{1:i}$ sequentially. For instance, if Y_i is a discrete random variable, then we draw $U_{K+i} \sim \mathrm{Uniform}(\{u \in (0,1) : Y_i = g_{K+i}(Y_1, \ldots, Y_{i-1}, u; \theta)\})$.

Example 2.3 (Bernoulli model)**.** *As a simple example, consider an i.i.d. Bernoulli model where θ is a discrete random variable such that $\mathbb{P}(\theta = 1/4) = \mathbb{P}(\theta = 3/4) = 0.5$, and $Y_1, \ldots, Y_n \mid \theta$ i.i.d. $\sim \mathrm{Bernoulli}(\theta)$. We can write this model as $U_1, \ldots, U_{n+1}$ i.i.d. $\sim \mathrm{Uniform}(0,1)$, $\theta = g_1(U_1) = (1/4) + (1/2)\mathbb{1}(U_1 \geqslant 0.5)$, and $Y_i = g_{i+1}(U_{i+1}; \theta) = \mathbb{1}(U_{i+1} \geqslant 1 - \theta)$ for $i = 1, \ldots, n$. Then, given data $Y = (Y_1, \ldots, Y_n)$ and a posterior sample of θ, we can sample the parameter u-value $U_1|\theta, Y$ by drawing $U_1 \sim \mathrm{Uniform}(\{u \in (0,1) : g_1(u) = \theta\})$, that is, $U_1 \sim \mathrm{Uniform}(0, 0.5)$ if $\theta = 1/4$ and $U_1 \sim \mathrm{Uniform}(0.5, 1)$ if $\theta = 3/4$. Likewise, we can sample the data u-values $U_{2:n+1}|\theta, Y, U_1$ by drawing $U_{i+1} \sim \mathrm{Uniform}(\{u \in (0,1) : g_{i+1}(u; \theta) = Y_i\})$ independently for $i = 1, \ldots, n$, that is, $U_{i+1} \sim \mathrm{Uniform}(0,\ 1 - \theta)$ if $Y_i = 0$ and $U_{i+1} \sim \mathrm{Uniform}(1 - \theta,\ 1)$ if $Y_i = 1$.*

3 Previous work

The observation that each posterior draw is marginally distributed according the prior is a simple consequence of the definition of the joint distribution of the parameters and the data. This fact has previously been used to assess the validity of posterior sampling algorithms using simulation-based calibration (SBC) (Geweke, 2004; Cook et al., 2006; Talts et al., 2018; Modrák et al., 2023). Specifically, these authors propose to generate simulated datasets from the hypothesized model, perform posterior inference for each simulated dataset, and compare the resulting approximate posteriors to the prior. Since, in this case, the datasets are known to be drawn from the model, any discrepancy between the posteriors and the prior is attributable to errors in

the posterior approximation algorithm. This is fundamentally different from our proposed method, since (i) they are not performing model criticism, (ii) they simulate datasets from the hypothesized model, and (iii) they run the posterior inference algorithm many times, once for each dataset. Talts et al. (2018) refer to the marginal distribution of θ, integrating out the data distribution, as the "data-averaged posterior".

There is an extensive literature on Bayesian model criticism, also referred to as model checking. The dominant approach is based on posterior predictive checks (PPCs), first introduced by Guttman (1967). The modern formulation of PPCs was developed by Rubin (1984), generalized by Meng (1994) to include test quantities that are functions of the data and parameters, and further developed by Gelman et al. (1996) on more complex models. Issues with the lack of uniformity of PPC p-values were demonstrated by Bayarri and Berger (1999, 2000) and Robins et al. (2000), who considered techniques for obtaining asymptotically valid PPC p-values using partial posterior or conditional predictive approaches. Valid PPC p-values can also be obtained by splitting the data, as shown by Moran et al. (2019) and Li and Huggins (2022). While we require p-values to be uniform under the null by definition, there is not universal agreement on this definition; see Gelman (2023) for other perspectives. The PPC approach has several limitations that are resolved by UPCs, as discussed in the introduction.

To our knowledge, the nearest precedent to our proposed method appears in Section 6.4 of Gelman et al. (2013) on "Graphical Posterior Predictive Checks", in which parameters of a certain hierarchical model were given $\text{Beta}(2, 2)$ priors, and Gelman et al. (2013) visually compare a posterior sample of these parameters to the $\text{Beta}(2, 2)$ prior, noting that the prior clearly does not match the posterior samples. This example was the initial seed of the idea which eventually led to our proposed methodology. Johnson (2007) and Yuan and Johnson (2012) develop a related method based on the use of pivotal discrepancy measures (PDMs), which are test quantities $T(Y, \theta)$ whose distribution is invariant to the value of θ when Y is distributed according to the hypothesized model with parameter θ; also see Gosselin (2011) and Zhang (2014), who consider sampled posterior p-values (SPPs) based on a particular class of PDMs. PDMs are used for model criticism by comparing this invariant distribution to the distribution of $T(Y^0, \tilde{\theta})$, where Y^0 is the observed data and $\tilde{\theta}$ is drawn from the posterior given Y^0. Like UPCs, PDMs only require posterior sampling given the observed data and the null distribution is known exactly, but similar to PPCs, the burden is on the analyst to design PDMs that simultaneously have the required invariance property and are useful for detecting misspecification in a given model. For continuous data, each data u-value can be viewed as a PDM of the form $T(y, \theta) = \mathbb{P}(Y_i \leqslant y_i \mid \theta)$. However, not all UPCs are PDMs, and not all PDMs are UPCs.

Johnson (2004) proposes another related idea, using what we refer to as the data u-values to construct a goodness-of-fit test based on a chi-squared test statistic; however, Johnson (2004) does not consider parameter u-values or any other types of test, and in general, Johnson's test statistic needs to be calibrated by

sampling from the posterior predictive and then from the resulting posteriors given these simulated datasets. Furthermore, the associated theory is restricted to the asymptotic setting and requires regularity conditions. Many other model checking methods have been proposed as well, for instance, based on simulation-based approaches (Dey et al., 1998; Hjort et al., 2006), cross-validation (Gelfand et al., 1992; Marshall and Spiegelhalter, 2003), and assessing within-model conflict (O'Hagan, 2003; Dahl et al., 2007), but these tend to be either computationally intensive, model-specific, or do not provide well-calibrated tests.

Model criticism methods make it possible to refine a model by identifying and correcting its inadequacies. Box (1980) proposed to iteratively perform model criticism and improvement, in a process dubbed "Box's loop" by Blei (2014). Box (1980) employed what later became known as prior predictive checks (Meng, 1994), but the same iterative refinement process can be implemented with other model checks such as PPCs (Belin and Rubin, 1995) or with our proposed UPC method, as we demonstrate in Section 5.3. Since UPCs control Type I error rate, they provide a theoretically well-founded method for implementing Box's loop while controlling the overall error rate (Section 2.4). Interestingly, Box (1980) and Gelfand et al. (1992) argue in favor of using predictive distributions for model criticism rather than the posterior, since the posterior alone cannot reveal any lack of fit. While this may be true for the usual posterior on parameters, the posterior on u-values provides additional context since (i) the u-values are on a model-based scale that captures information about goodness-of-fit, and (ii) the data u-values also provide more information than the posterior alone.

4 Theory

Let $\Theta \subseteq \mathbb{R}^L$ and $\mathcal{Y} \subseteq \mathbb{R}^n$ be measurable subsets; we use the Borel sigma-algebra on all topological spaces, unless otherwise specified. Let $g : (0,1)^D \to \Theta \times \mathcal{Y}$ be a measurable function, and define $(\boldsymbol{\theta}, Y) = g(U)$ where $U \sim \text{Uniform}_D(0,1)$, that is, $U = (U_1, \ldots, U_D)$ and $U_1, \ldots, U_D$ i.i.d. $\sim \text{Uniform}(0,1)$. Let Π denote the resulting distribution of $\boldsymbol{\theta}$ and let P_θ denote the conditional distribution of $Y \mid \boldsymbol{\theta} = \theta$. In this section, we use boldface $\boldsymbol{\theta}$ to denote the random vector and θ for particular values. All random elements are assumed to be defined on a common probability space $(\Omega, \mathcal{A}, \mathbb{P})$.

The interpretation is that Π and $(P_\theta : \theta \in \Theta)$ represent the analyst's prior and likelihood, and $(\boldsymbol{\theta}, Y)$ is jointly distributed according to this hypothesized model for the parameter θ and dataset Y. Write $\Pi_{\theta|y}$ to denote the posterior resulting from dataset y, that is, the conditional distribution of $\boldsymbol{\theta} \mid Y = y$, and let $\Pi_{u|\theta,y}$ denote the conditional distribution of $U \mid \boldsymbol{\theta} = \theta, Y = y$. The conditional distributions for $Y|\boldsymbol{\theta}$, $\boldsymbol{\theta}|Y$, and $U|\boldsymbol{\theta}, Y$ are guaranteed to exist almost everywhere due to the existence of regular conditional distributions in standard Borel spaces (Durrett, 2019). We assume, further, that these conditional distributions exist for all values of $\theta \in \Theta$ and $y \in \mathcal{Y}$, and that the maps $\theta \mapsto P_\theta(A)$, $y \mapsto \Pi_{\theta|y}(B)$, and $(\theta, y) \mapsto \Pi_{u|\theta,y}(E)$ are

measurable for all measurable $A \subseteq \mathcal{Y}$, $B \subseteq \Theta$, and $E \subset (0,1)^D$; these are very mild conditions in practice.

Now, we envision that the true distribution, which we refer to as *nature's model*, arises from some probability measures Π^0 and $(P^0_\theta : \theta \in \Theta)$ on Θ and $\mathcal{Y}$, respectively. Let $\boldsymbol{\theta}^0 \sim \Pi^0$ and $Y^0 \sim P^0_\theta$ given $\boldsymbol{\theta}^0 = \theta$, so that $(\boldsymbol{\theta}^0, Y^0)$ is jointly distributed according to nature's model; here, it is assumed that $\theta \mapsto P^0_\theta(A)$ is measurable for all measurable $A \subseteq \mathcal{Y}$. In practice, the observed dataset is generated from nature's model as Y^0 and one uses the hypothesized model to perform posterior inference using $\Pi_{\theta|y}$, setting y equal to Y_0. To represent this, we introduce a new random vector $\widetilde{\boldsymbol{\theta}}$ with conditional distribution $\Pi_{\theta|y}$ given $Y^0 = y$. Likewise, we introduce $\widetilde{U}$ with conditional distribution $\Pi_{u|\theta,y}$ given $\widetilde{\boldsymbol{\theta}} = \theta$ and $Y^0 = y$, where $\Pi_{u|\theta,y}$ is the conditional distribution under the hypothesized model. Then, we have that $(\widetilde{U}, \widetilde{\boldsymbol{\theta}}, Y^0)$ represents the true joint distribution of the observed dataset Y^0 and posterior draws of parameters $\widetilde{\boldsymbol{\theta}}$ and u-values $\widetilde{U}$ under the hypothesized model given the observed dataset.

4.1 Uniformity and independence of the u-values

This section contains the properties justifying the UPC method. Our first result, Theorem 4.1, provides the primary basis for the UPC method. All proofs are provided in Section S1.

Theorem 4.1. *Suppose* $Y^0 \stackrel{d}{=} Y$. *Then* $\widetilde{\boldsymbol{\theta}} \stackrel{d}{=} \boldsymbol{\theta}$ *and* $\widetilde{U} \stackrel{d}{=} U$, *that is,* $\widetilde{\boldsymbol{\theta}} \sim \Pi$ *and* $\widetilde{U} \sim \text{Uniform}_D(0,1)$.

Here, $\stackrel{d}{=}$ denotes equality in distribution. In other words, Theorem 4.1 says that if the marginal distribution of the dataset is the same under nature's model and the hypothesized model, then posterior draws of the parameter and u-values are distributed according to their respective priors, integrating out the dataset. In particular, if the model is correct then a draw from the posterior of the u-values is i.i.d. uniform. Thus, if we can reject the hypothesis that the u-values are i.i.d. uniform, then this implies the model is incorrect. Our next result shows that when the function $g : (0,1)^D \to \Theta \times \mathcal{Y}$ is bijective, the converse also holds.

Theorem 4.2. *If* $\widetilde{U} \stackrel{d}{=} U$ *and* g *is a bijection, then* $Y^0 \stackrel{d}{=} Y$.

Therefore, when g is bijective, $\widetilde{U} \stackrel{d}{=} U$ is a necessary and sufficient condition for the hypothesized model to be correct, at least in the sense that it matches nature's model in terms of the marginal distribution of the dataset. Thus, when performing UPCs in such cases, if our tests fail to reject the null hypothesis that $\widetilde{U} \sim \text{Uniform}_D(0,1)$ then—although we can never "accept" the null—this provides an indication that the model appears to be reasonable. The next result justifies our approach to testing for dependence with external variables X such as covariates. Although covariates are usually treated as fixed non-random quantities, here we treat them as random in order to have a formal notion of independence.

Theorem 4.3. *If* X *is a random element such that* $X \perp\!\!\!\perp Y^0$ *and* $(\widetilde{U}, \widetilde{\boldsymbol{\theta}}) \perp\!\!\!\perp X \mid Y^0$, *then* $(\widetilde{U}, \widetilde{\boldsymbol{\theta}}) \perp\!\!\!\perp X$.

The interpretation of Theorem 4.3 is that if (i) nature's model is independent of X, that is, $X \perp\!\!\!\perp Y^0$, and (ii) the hypothesized model does not use X when computing the posterior, that is, $(\widetilde{U}, \widetilde{\boldsymbol{\theta}}) \perp\!\!\!\perp X \mid Y^0$, then the u-values $\widetilde{U}$ and parameter $\widetilde{\boldsymbol{\theta}}$ are independent of X. Roughly speaking, under the null hypothesis that X is irrelevant, the u-values are independent of X. Consequently, if we observe dependence between X and the u-values, then this suggests that X or some other related variable may need to be added to the model.

4.2 Identifiability of the u-values

The u-values U are not necessarily identifiable, even if the parameter vector θ is identifiable. For instance, if there are discrete variables in the model, then it is clear that the u-values will not be uniquely determined. Another common situation is when Condition 2.2 holds and the vector of parameter u-values $U_{1:K}$ is in a higher dimensional space than the parameter vector $\theta \in \mathbb{R}^L$ (that is, $K > L$), in which case the u-values usually will not be uniquely determined; in practice this may occur when defining θ values that satisfy constraints. Likewise, the data u-values are not uniquely determined under similar circumstances.

However, if θ is identifiable and the parameter u-values $U_{1:K}$ are uniquely determined by θ (which is typically the case when all the entries of θ are continuous and there are no constraints on them), then $U_{1:K}$ is identifiable. If, further, the data u-values are uniquely determined by θ and the dataset Y, then they are identifiable as well for any given dataset. The following theorem formally states these results.

Theorem 4.4. *Assume θ is identifiable, that is, if $\theta \neq \theta'$ then $P_\theta \neq P_{\theta'}$. (i) If $\boldsymbol{\theta} = g_{\mathrm{p}}(U_{1:K})$ for some one-to-one function g_{p} and some K, then $U_{1:K}$ is identifiable, in the sense that these u-values are uniquely determined by P_θ. (ii) If $(\boldsymbol{\theta}, Y) = g(U)$ for some one-to-one function g, then all the u-values $U = U_{1:D}$ are uniquely determined by P_θ and Y.*

5 Examples

We demonstrate the UPC methodology on examples involving a univariate normal model (Section 5.1), Bernoulli trials (Section 5.2), logistic regression (Section 5.3), and an autoregression model (Section 5.4).

5.1 Normal model for Newcomb's speed of light data

We begin by considering Simon Newcomb's 66 measurements of the speed of light from 1882, a standard example for model criticism methods (Gelman et al., 2013). As seen in Figure 5.1(a), Newcomb's data look plausibly normal except for two outlying values in the left tail of the distribution. Following Gelman et al.

(2013, Section 6.3), we consider modeling the data as $Y_1, \ldots, Y_n$ i.i.d. $\sim \mathcal{N}(\mu, \sigma^2)$, where $n = 66$. We use a weakly informative Normal-InverseGamma prior: $\mu \mid \sigma^2 \sim \mathcal{N}(\mu_0, \sigma^2/\kappa_0)$ and $\sigma^2 \sim \text{InvGamma}(\alpha_0, \beta_0)$, with $\mu_0 = 0$, $\kappa_0 = 1/10$, $\alpha_0 = 2$, and $\beta_0 = 300$. We draw 500,000 samples from the posterior distribution, which is also a Normal-InverseGamma distribution by conjugacy.

Gelman et al. (2013) suggest using $T = \min\{Y_1, \ldots, Y_n\}$ as a posterior predictive check (PPC) test statistic. As shown in Figure 5.1(b), the observed value of T is not representative of the posterior predictive distribution, successfully detecting that there is an issue with the model. However, as we can see from Figure 5.1(c), the distribution of PPC p-values is not uniform under the null that the model is correct, so this is not a well-calibrated test. A deeper issue with this PPC is that this choice of T was apparently made by looking at the data. Without looking at the data, it would be difficult to know whether use the maximum instead of the minimum, or perhaps the interquartile range, or some other statistic.

To implement the UPC approach, we write $\sigma^2 = F_{\sigma^2}^{-1}(U_2)$, $\mu = \mu_0 + \sigma\kappa_0^{-1/2}\Phi^{-1}(U_1)$, and $Y_i = \mu + \sigma\Phi^{-1}(U_{d_i})$ where F_{σ^2} is the CDF of σ^2 (that is, the InvGamma(α_0, β_0) CDF), Φ is the standard normal CDF, and $d_i = 2 + i$. For each posterior sample of $(\mu, \sigma^2) \mid Y_{1:n}$, we compute the u-values by inverting these functions, that is, $\widetilde{U}_1 = \Phi((\mu - \mu_0)/(\sigma\kappa_0^{-1/2}))$, $\widetilde{U}_2 = F_{\sigma^2}(\sigma^2)$, and $\widetilde{U}_{d_i} = \Phi((Y_i - \mu)/\sigma)$.

As described in Section 2.1, we test for extreme values of μ and σ by computing p-values $p_\mu = 2\min\{\widetilde{U}_1, 1 - \widetilde{U}_1\}$ and $p_\sigma = 2\min\{\widetilde{U}_2, 1 - \widetilde{U}_2\}$, and we test for non-uniformity of the data u-values $\widetilde{U}_{d_1}, \ldots, \widetilde{U}_{d_n}$ by performing an Anderson–Darling test to obtain a p-value $p_{\text{data,unif}}$. Aggregating across posterior samples using the Cauchy combination method (Section 2.3), we obtain $p_\mu^* = 0.45$, $p_\sigma^* = 0.83$, and $p_{\text{data,unif}}^* = 1.60 \times 10^{-4}$. This suggests no issues with the priors, but indicates that the normal model may be misspecified.

To visualize what is happening, Figure 5.2 (top) shows the posterior distribution of each of these p-values given the observed data. Figure 5.2 (bottom) shows several simulated null-distributed posteriors – specifically, each curve is produced by sampling μ, σ^2 and $Y_1, \ldots, Y_n \mid \mu, \sigma^2$ from the hypothesized model, and computing the resulting posterior. The solid black lines in Figure 5.2 (bottom) show the average of the null-distributed posteriors over 100,000 simulated datasets, which we know from Theorem 4.1 is uniform in expectation. We see that the observed posteriors for p_μ and p_σ are representative of the null-distributed posteriors, but the observed posterior for $p_{\text{data,unif}}$ is much more concentrated near zero, which explains the small value of $p_{\text{data,unif}}^*$.

To examine the data u-values in more detail, Figure 5.3 (top, left/middle) shows a histogram and the empirical CDF $\hat{F}_{\text{d}}(u) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}(\widetilde{U}_{d_i} \leqslant u)$ of the data u-values for a single sample of (μ, σ^2) from the posterior given the Newcomb data. We can see that these data u-values do not appear to be uniformly distributed, particularly in comparison to the corresponding histogram and empirical CDF given a simulated

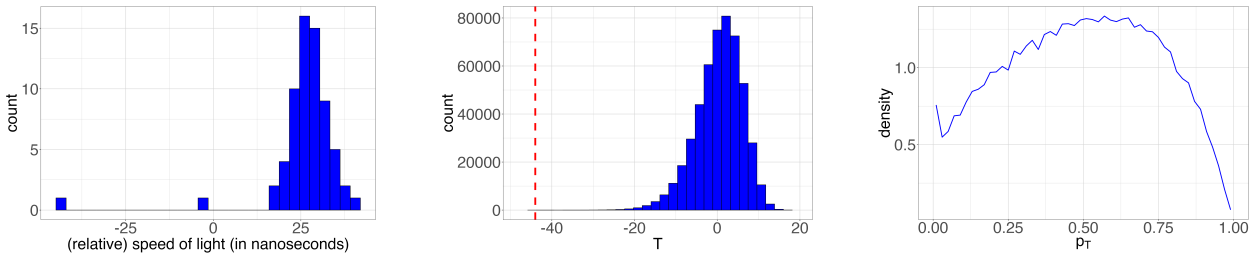


Figure 5.1: Newcomb data and results using PPCs. (a) Histogram of Newcomb's speed of light data, relative to 24,800 nanoseconds (ns). (b) Histogram of the distribution of $T = \min\{Y_1, \ldots, Y_n\}$ under the posterior predictive, along with the observed value of T on the Newcomb data (vertical red dashed line). (c) Density of the PPC p-values when sampling from the hypothesized model.

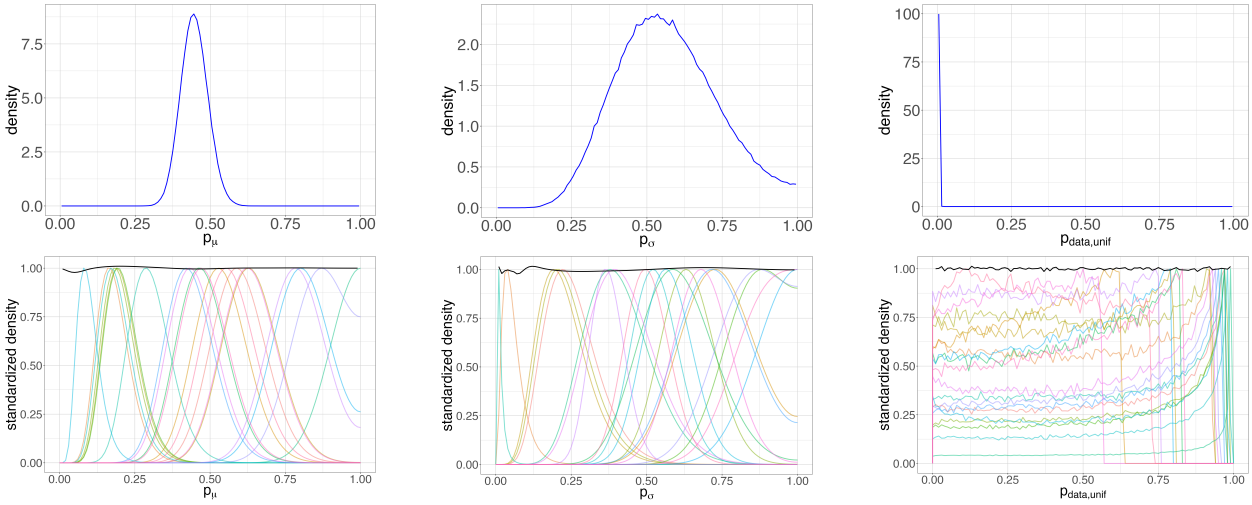


Figure 5.2: Results using UPCs on the Newcomb data. (Top) Posterior densities of p_μ, p_σ, and $p_{\text{data,unif}}$ given the Newcomb data. (Bottom) Samples of the posterior densities of p_μ, p_σ, and $p_{\text{data,unif}}$ given simulated datasets from the hypothesized model. To aid visualization, each density is standardized to have a maximum of 1, so that they are all visible in a single plot. The black lines show the average of these (unstandardized) densities over 100,000 simulated datasets.

dataset from the hypothesized model (Figure 5.3; bottom, left/middle). To more clearly see the differences between the empirical CDF $\hat{F}_{\text{d}}(u)$ and the uniform CDF $F_{\text{U}(0,1)}(u) = u$, in Figure 5.3 (right), we plot $\hat{F}_{\text{d}}(u) - u$ for multiple posterior samples given the Newcomb data (top) and simulated data from the model (bottom). For a CDF F, we refer to $F(u) - u$ as the corresponding "tilted CDF." This illustrates that the data u-values clearly do not appear to be uniformly distributed, which explains the small value of $p^*_{\text{data,unif}} = 1.60 \times 10^{-4}$.

To evaluate the effect of the prior on the UPC results, we compare three choices of prior: (1) the weakly informative prior described above, in which $\mu_0 = 0$, $\kappa_0 = 1/10$, $\alpha_0 = 2$, and $\beta_0 = 300$, (2) a data-dependent prior, in which $\mu_0 = \overline{Y}$, $\kappa_0 = n$, $\alpha_0 = n/2$, and $\beta_0 = \hat{\sigma}^2\alpha_0$, where $\overline{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$ and $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (Y_i - \overline{Y})^2$ and (3) a poorly chosen informative prior based on previous data, in which $\mu_0 = 179$, $\kappa_0 = n$, $\alpha_0 = n/2$,

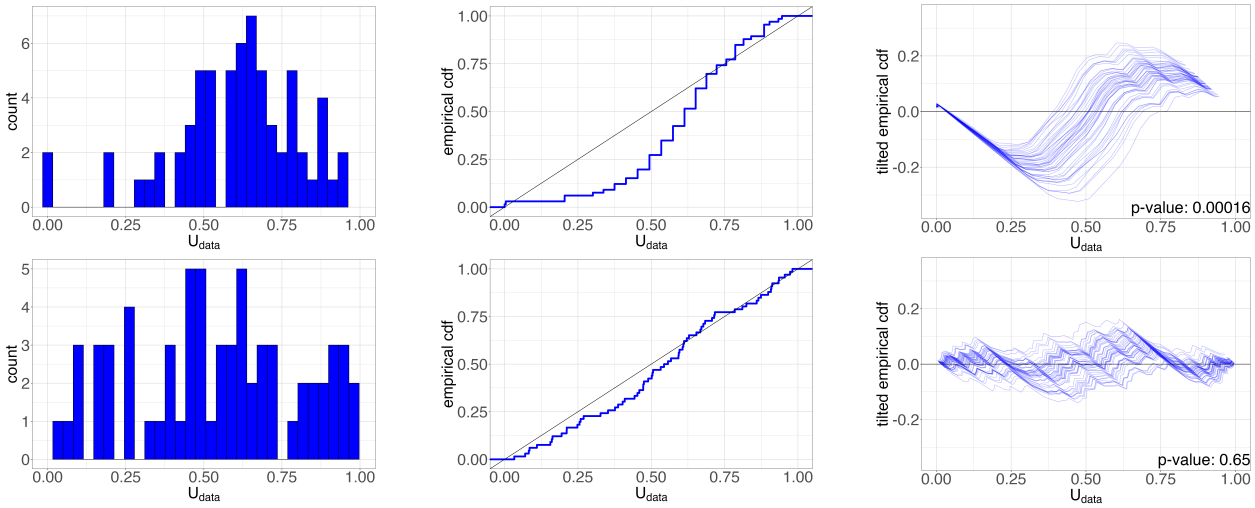


Figure 5.3: Visualizing the data u-values for the Newcomb data. (Top, left/middle) Histogram and empirical CDF $\hat{F}_{\text{d}}$ of the data u-values $U_{d_1}, \ldots, U_{d_n}$ from a single posterior sample, given the Newcomb data. (Top, right) Tilted empirical CDF of the data u-values, $\hat{F}_{\text{d}}(u) - u$, for multiple posterior samples given the Newcomb data. (Bottom) Same as the top row, but for samples of data u-values from the posterior given a simulated dataset from the hypothesized model.

Prior	**Data**	p^*_μ	p^*_σ	$p^*_{\text{data,unif}}$
Weakly informative prior	Newcomb data	0.45	0.83	1.60×10^{-4}
	Normal data	0.37	0.48	0.98
Data-dependent prior	Newcomb data	0.96	0.93	4.44×10^{-4}
	Normal data	0.93	0.95	0.50
Poorly chosen informative prior	Newcomb data	2.41×10^{-4}	3.81×10^{-10}	9.09×10^{-6}
	Normal data	2.25×10^{-4}	2.01×10^{-10}	9.09×10^{-6}

Table 5.1: Results on Newcomb's speed of light data: Aggregated p-values from UPC tests.

and $\beta_0 = 42^2 \alpha_0 \kappa_0$. Prior #3 is based on experiment of Foucault in 1862, who only twenty years prior to Newcomb estimated the speed of light to be 2.98×10^8 m/s with an error of $\pm$500,000 m/s (Froome et al., 1971). This corresponds to a value of $24{,}979 \pm 42$ ns in Newcomb's measurements, which represent the time required to travel 7.44373 km. Relative to 24,800 ns, this translates to 179 ± 42 ns.

Table 5.1 shows the aggregated p-values for the UPC results for all three choices of prior, on the Newcomb data. Table 5.1 also reports the aggregated p-values on data simulated from $\mathcal{N}(\overline{Y}, \hat{\sigma}^2)$, that is, a normal distribution with mean and variance equal to the sample mean and sample variance of the Newcomb data. Under the weakly informative and data-dependent priors, we see exactly what we would hope for, no evidence against the model in the correctly specified case (Normal data) and evidence against the uniformity of the residuals in the misspecified case (Newcomb data). Under the poorly chosen prior, we see evidence of misspecification in all three tests, on both the Newcomb data and the simulated normal data. We see this theme throughout our examples: Although UPCs do not actively distinguish between misspecification of the

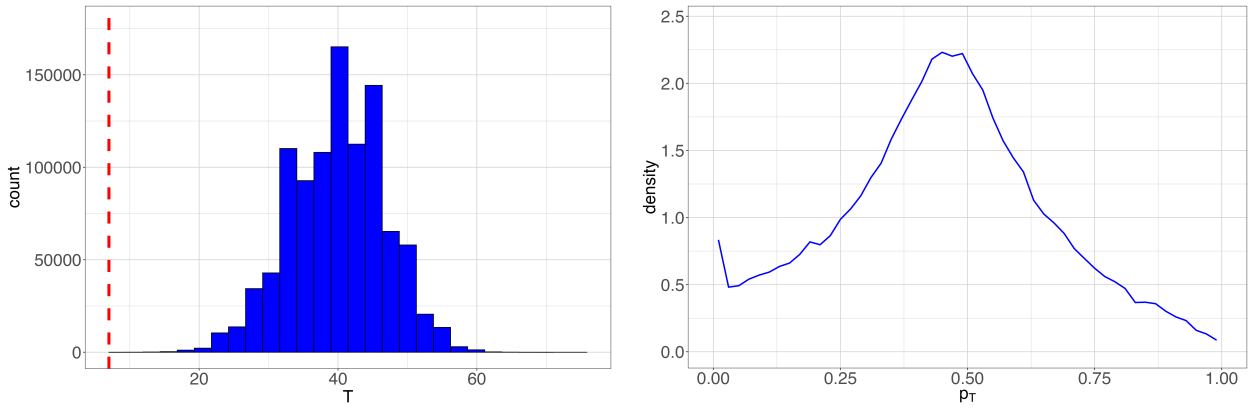


Figure 5.4: Results using PPCs on dependent Bernoulli example. (a) Histogram of the distribution of $T = \sum_{i=1}^{n-1} \mathbb{1}(Y_i \neq Y_{i+1})$ under the posterior predictive, along with the observed value of T on the data (vertical red dashed line). (b) Density of the PPC p-values when sampling from the hypothesized model.

prior and the likelihood, the choice of prior does not generally affect UPCs meaningfully unless the prior is chosen very poorly. See Section S2 for plots of results using the data-dependent prior and poorly chosen informative prior.

5.2 Dependent Bernoulli trials

Next, we consider a simulation like the dependent Bernoulli trials example presented by Gelman et al. (2013), but with a larger sample size. We generate the following simulated dataset by drawing $Y_1 \sim \text{Bernoulli}(0.5)$ and for $i = 2, \ldots, 100$, setting $Y_i = Y_{i-1}$ with probability 0.8, and drawing $Y_i \sim \text{Bernoulli}(0.5)$ otherwise:

$$\begin{aligned} &0\ 0\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 1\ 1\ 1\ 1\ 1 \\ &0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 1\ 1\ 1. \end{aligned} \tag{5.1}$$

Following Gelman et al. (2013, Section 6.3), we consider modeling these data as $Y_1, \ldots, Y_n$ i.i.d. $\sim$ Bernoulli(θ), where $n = 100$. For the prior, we consider $\theta \sim \text{Beta}(1, 1)$. We draw 10^6 samples of θ from the posterior, $\theta \mid Y_{1:n} \sim \text{Beta}(1 + \sum_i Y_i,\ 1 + n - \sum_i Y_i)$.

As a PPC test statistic, Gelman et al. (2013) suggest using the number of switches between 0 and 1, that is, $T = \sum_{i=1}^{n-1} \mathbb{1}(Y_i \neq Y_{i+1})$. Figure 5.4(a) shows that the observed value of T on the data in Equation 5.1 is extremely unlikely under the posterior predictive, so this PPC successfully detects the fact that the hypothesized model is misspecified. However, again, this is not a well-calibrated test, as we can see from the non-uniformity of the PPC p-values in Figure 5.4(b) under simulated datasets from the hypothesized model.

For the UPC approach, we reparametrize as $\theta = F_\theta^{-1}(U_1) = U_1$ and $Y_i = \mathbb{1}(U_{i+1} \geqslant 1-\theta)$ for $i = 1, \ldots, n$. For each posterior sample of θ, we sample the u-values by setting $\widetilde{U}_1 = \theta$ and drawing $\widetilde{U}_{i+1} \sim \text{Uniform}(0,\ 1-\theta)$

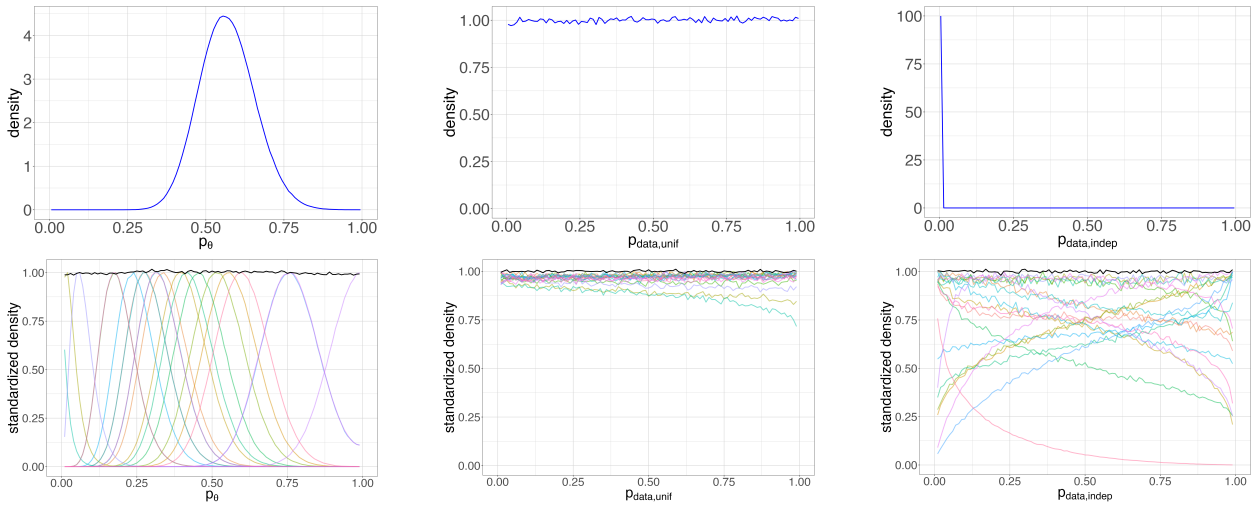


Figure 5.5: Results using UPCs on dependent Bernoulli example. (Top) Posterior densities of the p-values p_θ, $p_{\text{data,unif}}$, and $p_{\text{data,indep}}$ given the observed data in Equation 5.1. (Bottom) Several samples of the same posterior densities, given simulated datasets generated by sampling θ and $Y_1, \ldots, Y_n$ from the hypothesized model. To aid visualization, each density is standardized to have a maximum of 1, so they are all visible in a single plot. The black solid lines show the average of these posterior densities over 100,000 simulations, which is exactly uniform in expectation.

if $Y_i = 0$, or $\widetilde{U}_{i+1} \sim \text{Uniform}(1-\theta, 1)$ if $Y_i = 1$, independently for $i = 1, \ldots, n$, as described in Section 2.6.2.

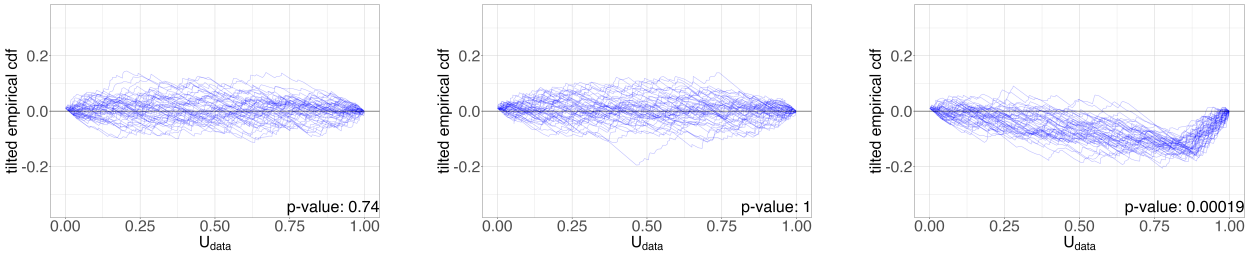
Since any binary variables must necessarily be Bernoulli, the empirical distribution of the data u-values $\widetilde{U}_2, \ldots, \widetilde{U}_{n+1}$ will be close to uniform when the inferred value of θ is close to the sample mean of the Y_i values. To verify this empirically, we test for non-uniformity of the data u-values using an Anderson–Darling test to compute $p_{\text{data,unif}}$. This yields an aggregated p-value of $p^*_{\text{data,unif}} = 0.74$, correctly indicating no issues with the Bernoulli aspect of the model.

Under the null, the u-values are not just marginally Uniform$(0, 1)$, they are also independent. To test for dependence, we compute a Hoeffding test of independence (Hoeffding, 1948) between $\tilde{U}_i$ and $\tilde{U}_{i+1}$ over $i = 2, \ldots, n$, which produces an aggregated p-value of $p^*_{\text{data,indep}} = 4.61 \times 10^{-6}$. This provides strong evidence of dependence, correctly detecting the form of misspecification present in the hypothesized model relative to the true data generating process. As before, we also test for extreme values of θ using p-value $p_\theta = 2\min\{\widetilde{U}_1, 1 - \widetilde{U}_1\}$, which yields $p^*_\theta = 0.58$, correctly indicating no issues with the prior.

Figure 5.5 (top) shows the posterior densities of p_θ, $p_{\text{data,unif}}$, and $p_{\text{data,indep}}$ given the data in Equation 5.1. Figure 5.5 (bottom) shows samples of these same posteriors under the null, that is, given simulated datasets from the hypothesized model. As in Figure 5.2, the black lines show the average of these null-distributed posteriors over 100,000 simulated datasets, which we know from the theory is uniform in expectation. We can see that the observed posterior for $p_{\text{data,indep}}$ is concentrated near zero, which visually illustrates why $p^*_{\text{data,indep}}$ is small.

prior	p^*_θ	$p^*_{\text{data,unif}}$	$p^*_{\text{data,indep}}$
Uniform prior, Beta$(1,1)$	0.58	0.74	4.61×10^{-6}
Jeffreys prior, Beta$(1/2,1/2)$	1	0.72	4.61×10^{-6}
Poorly chosen prior, Beta$(1,50)$	1.89×10^{-4}	2.08×10^{-4}	4.61×10^{-6}

Table 5.2: Dependent Bernoulli trials: Aggregated p-values from UPC tests for extreme θ values (p^*_θ), non-uniform data u-values ($p^*_{\text{data,unif}}$), and dependent data u-values ($p^*_{\text{data,indep}}$).



(a) Uniform Prior (b) Jeffreys Prior (c) Poorly Chosen Prior

Figure 5.6: Tilted CDF of the data u-values, $\hat{F}_{\text{d}}(u) - u$, for multiple posterior samples given the Bernoulli data, under each prior. The departure from uniformity is visibly clear under the poorly chosen prior.

To assess the effect of the choice of prior, we compare (1) the uniform prior as described above, $\theta \sim \text{Beta}(1,1)$, (2) the Jeffreys prior, $\theta \sim \text{Beta}(1/2,1/2)$, and (3) a poorly chosen prior, $\theta \sim \text{Beta}(1,50)$. Table 5.2 shows the aggregated p-values for each combination of prior and test. Under the two reasonable priors (uniform and Jeffreys), we see no evidence against uniformity of the parameter u-value and data u-values (as expected), and we do see evidence against the independence assumption (as desired, since the data exhibit dependence). Under the poorly chosen prior, the null of model correctness is rejected in all three tests; thus, while it is not as clear which aspect of the model is problematic under this prior, the extreme θ value suggests that improving the prior is a good place to start. See Figure S3.1 for plots of the posteriors under the Jeffreys prior and poorly chosen prior. For each prior, Figure 5.6 shows the tilted empirical CDF of the data u-values, $\hat{F}_{\text{d}}(u) - u$, for multiple posterior samples given the Bernoulli data. In contrast to the PPC approach, we simply use default choices of UPC test, without having to design test statistics.

5.3 Logistic regression model for adolescent smoking data

In this section, we apply the UPC methodology to a logistic regression example used by Gelman et al. (2013, Section 6.3) to illustrate PPCs. Each of the $n = 8{,}730$ observations comes from one of $m = 1{,}760$ individuals (`newid`) who were surveyed at up to six time points (`wave`) and asked whether or not they smoked regularly (`smkreg`). Individuals' sex (`sex`) and parental smoking status (`parsmk`) were also recorded as potential covariates. Figure 5.7 (bottom right) shows the proportion of individuals that smoked regularly versus `wave`, stratified by `sex` and `parsmk`. We treat `smkreg` as the outcome and `sex`, `parsmk`, and `wave` as covariates,

standardizing each covariate to have zero mean and unit variance across all observations.

We show how to use the UPC framework for iterative model criticism. In particular, we build up from a simple model to a more complex model, using UPCs to reject certain model assumptions in each round. At the outset, we commit to performing at most two rounds of model criticism (producing three models in total), with a cumulative Type I error rate of $\alpha = 0.2$, which we split evenly between the rounds.

Model #1. We start with a simple random effects model with weakly informative priors. The outcome $Y_{jk} \in \{0, 1\}$ represents whether individual j smokes regularly at wave k (`smkreg`), which is modeled as

$$\begin{aligned}(Y_{jk} \mid \alpha) &\sim \text{Bernoulli}\big(\text{logit}^{-1}(\alpha_j)\big),\\ \alpha_j &\sim \mathcal{N}(\mu, 5^2),\\ \mu &\sim \mathcal{N}(0, 5^2)\end{aligned}$$

for $j \in \{1, \ldots, m\}$, $k \in \{1, \ldots, 6\}$. Since some individuals were not observed at some waves, we handle missing entries by assuming they are missing completely at random; thus, they can be marginalized out of the model and contribute nothing to the likelihood. Indexing the observations $i = 1, \ldots, n$, we define $i(j, k)$ to be the index of the observation for individual j at wave k when it is nonmissing, otherwise $i(j, k)$ is undefined. We use JAGS (Plummer, 2003) for posterior inference with MCMC, drawing 100,000 posterior samples after a burn-in of 5,000 iterations. We thin the posterior samples, keeping only every 100th sample, leaving 1,000 samples. To implement the UPC approach, we write $\mu = 5\Phi^{-1}(U_1)$, $\alpha_j = \mu + 5\Phi^{-1}(U_{j+1})$, and $Y_{jk} = \mathbb{1}(U_{i(j,k)+m+1} \geqslant 1 - q_{jk})$, where $q_{jk} = \text{logit}^{-1}(\alpha_j)$. Inverting these as before, for each posterior sample, we sample the u-values by setting $\widetilde{U}_1 = \Phi(\mu/5)$, $\widetilde{U}_{j+1} = \Phi\big((\alpha_j - \mu)/5\big)$, $\widetilde{U}_{i(j,k)+m+1} \sim \text{Uniform}(0, 1 - q_{jk})$ if $Y_{jk} = 0$, and $\widetilde{U}_{i(j,k)+m+1} \sim \text{Uniform}(1 - q_{jk}, 1)$ if $Y_{jk} = 1$.

For this initial model, a primary question is whether the exclusion of all covariates is problematic. Thus, as discussed in Section 2.1, a natural first step is to test for external dependence between u-values and covariates. By Theorem 4.3, if the true distribution of the outcome is independent of a covariate, then the u-values will independent of that covariate as well. Therefore, we test for dependence between the data u-values and (1) `wave`, (2) `sex`, and (3) `parsmk`. We also test for dependence between the α u-values and (4) `sex` and (5) `parsmk`. Testing for dependence between α and `wave` is excluded since α_j is a subject-level parameter and `wave` varies within subject. We use the Mann–Whitney U test for `sex` and `parsmk`, and Hoeffding's test for `wave`; see Section 2.1.

For each posterior sample, we compute the p-values for each of these five tests, and aggregate across posterior samples using the Cauchy combination method for each test, obtaining $p^*_{\text{data,wave}} = 1.67 \times 10^{-7}$,

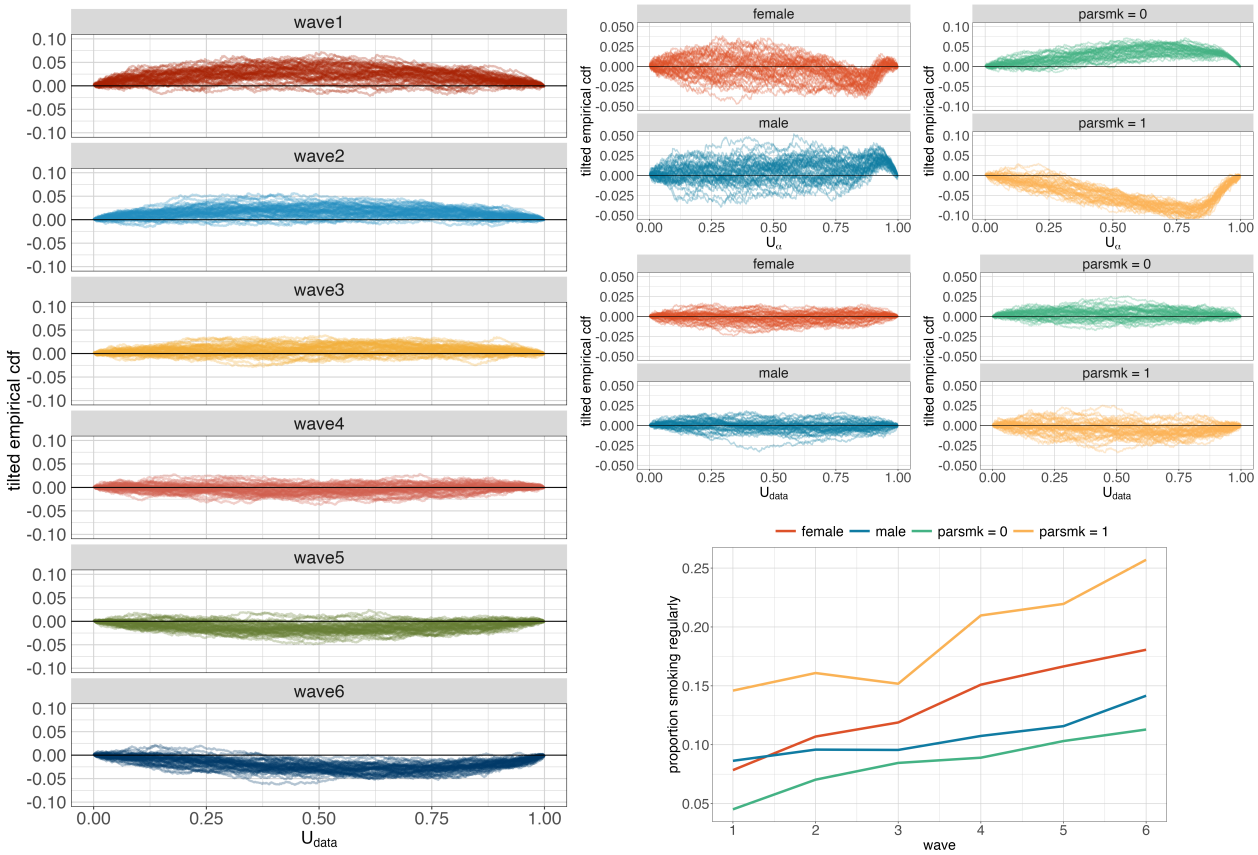


Figure 5.7: UPC results for Model #1 on the logistic regression example. (Left) Posterior samples of the tilted empirical CDFs of the data u-values, stratified by `wave`. (Top right) Posterior samples of the tilted empirical CDFs of the α u-values and the data u-values, stratified by `sex` and `parsmk`. (Bottom right) Proportion of individuals that smoke regularly as a function of `wave`, stratified by `sex` and `parsmk`.

$p^*_{\text{data,sex}} = 0.72$, $p^*_{\text{data,parsmk}} = 8.47 \times 10^{-3}$, $p^*_{\alpha,\text{sex}} = 0.68$, and $p^*_{\alpha,\text{parsmk}} = 1.81 \times 10^{-11}$. To control Type I error under multiple testing, we apply the Holm–Bonferroni correction to these five aggregated p-values, after which the p-values for `wave` and `parsmk` remain significant at the 0.1 level; see Section S4 for details.

This provides strong evidence of dependence with `wave` and `parsmk`, but not `sex`. To visually confirm that these formal results make sense, Figure 5.7 contrasts the empirical CDFs of u-values across strata defined by covariate values. For instance, Figure 5.7 (left) shows $\hat{F}^{\texttt{wave}=k}_{\text{data}}(u) - u$ for $k \in \{1, \ldots, 6\}$ for several posterior samples, where $\hat{F}^{\texttt{wave}=k}_{\text{data}}$ is the empirical CDF of the data u-values for data points at $\texttt{wave} = k$ for a given posterior sample. The clear trend as `wave` goes from 1 to 6 reflects the very small value of $p^*_{\text{data,wave}}$.

Model #2. Based on the Model #1 results, we augment the model to include `wave` and `parsmk`:

$$
\begin{aligned}
(Y_{jk} \mid \beta, \alpha) &\sim \text{Bernoulli}\big(\text{logit}^{-1}(\alpha_j + \beta_1 \texttt{wave}_{jk} + \beta_2 \texttt{parsmk}_{jk})\big), \\
\beta_1, \beta_2 &\sim \mathcal{N}(0, 5^2), \\
\alpha_j &\sim \mathcal{N}(\mu, 5^2), \\
\mu &\sim \mathcal{N}(0, 5^2)
\end{aligned}
$$

for all j, k. As before, we use JAGS to draw 100,000 posterior samples after 5,000 burn-in iterations, we thin to 1,000 samples, and we compute the u-values. To illustrate the effectiveness of the UPC method, we first verify that dependence with `wave` and `parsmk` is no longer detected. Indeed, running the same tests as before yields $p^*_{\text{data,wave}} \approx 1$, $p^*_{\text{data,sex}} = 0.55$, $p^*_{\text{data,parsmk}} \approx 1$, $p^*_{\alpha,\text{sex}} = 0.75$, and $p^*_{\alpha,\text{parsmk}} = 0.20$. Figure 5.8 provides visual confirmation of these results.

To explore possible deficiencies in Model #2, we consider two tests: (1) we test for dependence between the data u-values and $\texttt{wave} \times \texttt{parsmk}$ using Hoeffding's test, to assess whether this interaction is needed, and (2) we test for uniformity of the α u-values using Anderson–Darling, to evaluate the choice of prior on the α parameters. These tests yield aggregated p-values of $p^*_{\text{data,wave}\times\text{parsmk}} \approx 1$ and $p^*_{\alpha,\text{unif}} = 3.41 \times 10^{-7}$. This suggests that there is no need to add the interaction to the model, but that the prior on the α values is defective in some way.

To visually confirm and understand these findings, Figure 5.8 shows posterior samples of tilted empirical CDF of (left) the data u-values, stratified by $\texttt{wave} \times \texttt{parsmk}$, and (bottom right) the α u-values, as well as a density histogram of the α u-values across posterior samples. There is little variation across values of $\texttt{wave} \times \texttt{parsmk}$, which explains the large p-value for the interaction test. Meanwhile, the α u-values tend to be approximately uniform except for a medium-sized bump at around 0.9. This suggests that there may be some additional latent structure that is not accounted for in the measured covariates.

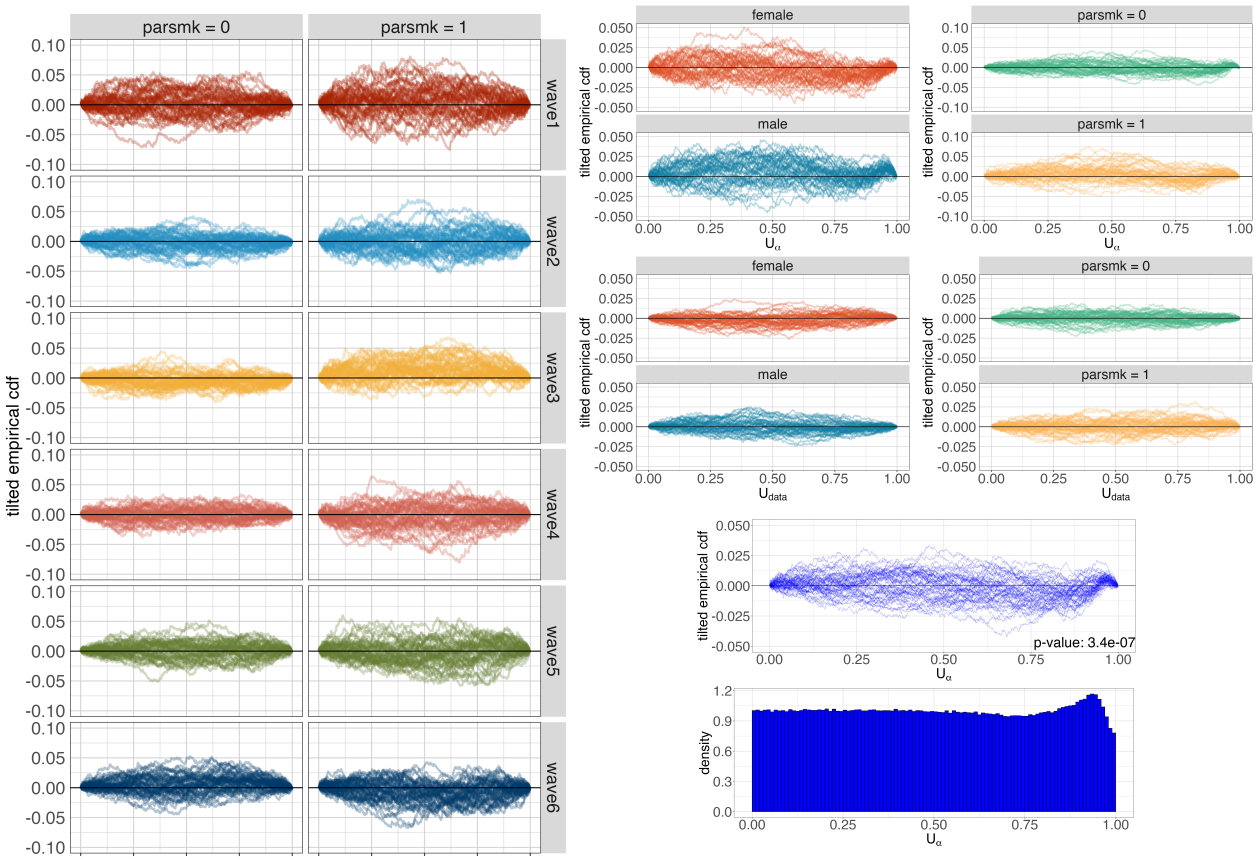


Figure 5.8: UPC results for Model #2 on the logistic regression example. (Left) Tilted empirical CDFs of the data u-values, stratified by `wave` × `parsmk`. (Top right) Same as the corresponding plots in Figure 5.7, but for Model #2. (Bottom right) Tilted empirical CDFs of the α u-values for multiple posterior samples, and a density histogram of the α u-values over all posterior samples.

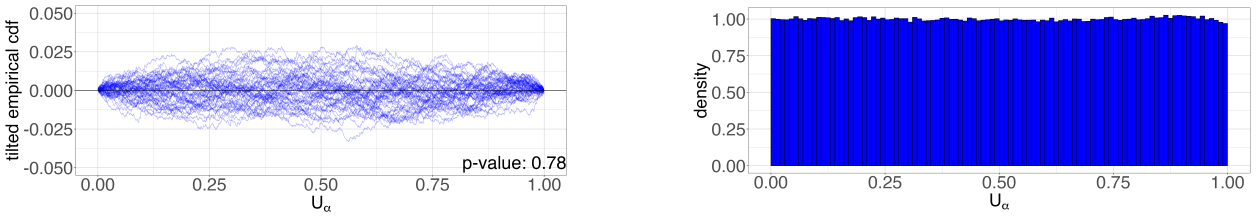


Figure 5.9: UPC results for Model #3 on the logistic regression example. (Left) Tilted empirical CDFs of the α u-values for multiple posterior samples. (Right) Histogram of the α u-values over all posterior samples.

Model #3. To account for the non-uniformity of the α u-values in Model #2, we augment the model to employ a two-component mixture model for the prior on the α values, as follows:

$$
\begin{aligned}
(Y_{jk} \mid \beta, \alpha) &\sim \text{Bernoulli}\big(\text{logit}^{-1}(\alpha_j + \beta_1 \texttt{wave}_{jk} + \beta_2 \texttt{parsmk}_{jk})\big), \\
\beta_1, \beta_2 &\sim \mathcal{N}(0, 5^2), \\
(\alpha_j \mid \mu, \tau, Z) &\sim \mathcal{N}(\mu_{Z_j}, \tau_{Z_j}^{-1}), \\
(Z_j \mid \pi) &\sim \text{Bernoulli}(\pi), \\
\pi &\sim \text{Beta}(1, 1), \\
\mu_1, \mu_2 &\sim \mathcal{N}(0, 5^2), \\
\tau_1, \tau_2 &\sim \text{Gamma}(1, 0.2)
\end{aligned}
$$

for all j, k, where $\text{Gamma}(a, b)$ denotes the gamma distribution with shape a and rate b. Again, we use JAGS to draw 100,000 posterior samples after 5,000 burn-in iterations, we thin to 1,000 samples, and we compute the u-values. Again, to demonstrate the effectiveness of UPCs, we verify that this addresses the non-uniformity of the α u-values that was seen in Model #2. Computing the same tests as done for Model #2, we obtain $p^*_{\text{data},\texttt{wave}\times\texttt{parsmk}} \approx 1$ and $p^*_{\alpha,\text{unif}} = 0.78$ for Model #3, which suggests that the inadequacy of the α prior in Model #2 has been sufficiently addressed. Figure 5.9 provides visual confirmation that the α u-values do indeed appear to be close to uniform in Model #3. In fact, for Model #3, the null of model correctness is not rejected under any of the UPC tests considered in this section, indicating that this model is at least providing a reasonable fit to these data. Additionally, Table S4.1 shows that the mixture component assignments Z_j have a clear interpretation, since nearly all individuals with $Z_j = 0$ never smoke regularly throughout the study.

5.4 Autoregressive model - Simulation study

We present a simulation study using an autoregressive model. In the other examples, we consider only the observed dataset and simulated datasets from the hypothesized model. In this example, we simulate datasets from various perturbations of the hypothesized model in order to see the effect on the UPCs. This demonstrates which UPCs are affected by different types of perturbations, and the power we have to detect each type of perturbation. Suppose the hypothesized model is an autoregressive AR(1) model:

$$\begin{aligned} Y_i &= \phi Y_{i-1} + \sigma\varepsilon_i, \\ \phi &\sim \pi(\phi), \quad \sigma \sim \pi(\sigma) \end{aligned} \tag{5.2}$$

for $i = 1, \ldots, n$, with $Y_0 = 0$ and $\varepsilon_1, \ldots, \varepsilon_n$ i.i.d. $\sim \mathcal{N}(0, 1)$. We will consider various choices for the priors $\pi(\phi)$, $\pi(\sigma)$ and for the true data generating process (DGP). In each scenario, we simulate 100,000 datasets from the true DGP with $n = 500$, and for each dataset, we use Stan (Carpenter et al., 2017a) to draw one sample from the posterior after 1,000 burn-in iterations.

To implement UPCs, we reparametrize in terms of u-values via $\phi = F_\phi^{-1}(U_1)$, $\sigma = F_\sigma^{-1}(U_2)$, and $Y_i = \phi Y_{i-1} + \sigma\Phi^{-1}(U_{d_i})$, where $d_i = i + 2$. For each posterior sample of ϕ and σ, we compute the u-values via $\widetilde{U}_1 = F_\phi(\phi)$, $\widetilde{U}_2 = F_\sigma(\sigma)$, and $\widetilde{U}_{d_i} = \Phi((Y_i - \phi Y_{i-1})/\sigma)$. Consider the following UPCs: (1) test for extreme values of ϕ and σ via the p-values $p_\phi = 2\min\{\widetilde{U}_1, 1 - \widetilde{U}_1\}$ and $p_\sigma = 2\min\{\widetilde{U}_2, 1 - \widetilde{U}_2\}$, (2) test for non-uniformity of the data u-values using Anderson–Darling, yielding $p_{\text{data,unif}}$, (3) test for dependence between the data u-values $\tilde{U}_{d_i}$ and the index i (for $i = 1, \ldots, n$) using the Hoeffding independence test, yielding $p_{\text{data,index}}$, (4) test for first-order serial correlation by using a Hoeffding independence test between $\tilde{U}_{d_i}$ and $\tilde{U}_{d_i+1}$ (for $i = 1, \ldots, n-1$), yielding $p_{\text{data,lag1}}$, and (5) test for second-order serial correlation by using a Hoeffding independence test between $\tilde{U}_{d_i}$ and $\tilde{U}_{d_i+2}$ (for $i = 1, \ldots, n-2$), yielding $p_{\text{data,lag2}}$.

Scenario #1: Correct model. As a sanity check, we begin with the case where the hypothesized model is identical to the true DGP. Let $\mathcal{TN}(\mu, \sigma^2, [a, b])$ denote the truncated normal distribution obtained by restricting the support of $\mathcal{N}(\mu, \sigma^2)$ to the interval $[a, b]$. Suppose that under both the true DGP and the hypothesized model, a dataset is generated by drawing parameters $\phi \sim \mathcal{TN}(0, 0.4^2, [-0.5, 0.5])$ and $\sigma \sim \mathcal{TN}(1.5, 0.4^2, [1, 2])$ independently, and generating $Y_1, \ldots, Y_n$ as in Equation 5.2. Figure 5.10 shows the expected posterior CDFs of p_ϕ, p_σ, $p_{\text{data,unif}}$, $p_{\text{data,index}}$, $p_{\text{data,lag1}}$, and $p_{\text{data,lag2}}$, where by "expected" we are referring to averaging over datasets drawn from the true DGP; this is called the "data-averaged posterior" by Talts et al. (2018). In Figure 5.10, we see that the p-values are all uniformly distributed, as expected based on our theory.

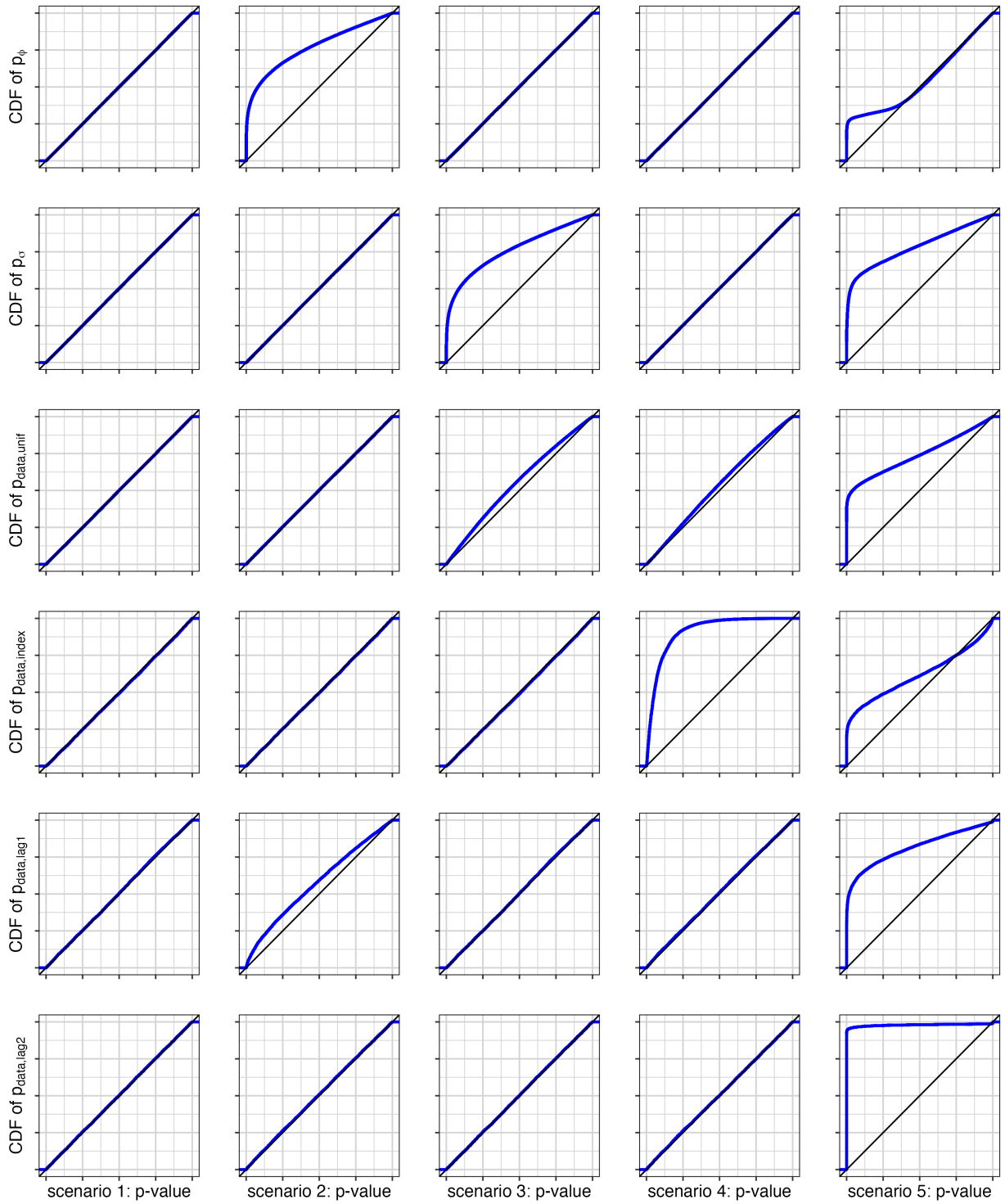


Figure 5.10: Expected CDFs of p-values of each test in the AR simulation study scenarios, where the expectation is taken with respect to datasets drawn from the true DGP: Scenario 1 (correct model), Scenario 2 (narrow prior for ϕ), Scenario 3 (narrow prior for σ), Scenario 4 (heteroskedastic errors), and Scenario 5 (second-order dependence).

Scenario #2: Narrow prior for ϕ. Next, we consider a case where the hypothesized likelihood is correct but the prior on ϕ is too concentrated. Suppose the true DGP and hypothesized model are the same as in Scenario #1, except that $\phi \sim \mathcal{TN}(0, 0.1^2, [-0.5, 0.5])$ under the hypothesized model. In Figure 5.10, we see that p_ϕ is relatively concentrated near zero. Thus, the UPCs correctly detect that the ϕ values needed to explain the data are extreme, relative to the hypothesized model prior. Meanwhile, the rest of the p-values are approximately uniform, except that $p_{\text{data,lag1}}$ exhibits a slight departure from uniformity. This is appealing since the prior on ϕ is in fact the only aspect of the model that does not match the true DGP.

Scenario #3: Narrow prior for σ. Now, we consider a case where the hypothesized prior on σ is too concentrated. Again, suppose the same true DGP and hypothesized model as in Scenario #1, except that $\sigma \sim \mathcal{TN}(1.5, 0.1^2, [1, 2])$ under the hypothesized model. Thus, in this case the prior on σ is the one that does not match the true DGP. As one might hope, the UPCs correctly detect extreme values of σ relative to the prior, and detect no other serious issues.

Scenario #4: Heteroskedastic errors. In this case, we suppose the true DGP and hypothesized model are the same as in Scenario #1, except under the true DGP the data are generated according to $Y_i = \phi Y_{i-1} + \sqrt{c_i}\sigma\varepsilon_i$ where $c_i = 1 + (2i - n - 1)/n$ for $i = 1, \ldots, n$. Meanwhile, the hypothesized model still uses Equation 5.2. Thus, the true DGP exhibits heteroskedasticity, whereas the model assumes homoskedasticity. In Figure 5.10, we see extreme non-uniformity in $p_{\text{data,index}}$, correctly reflecting that the model does not capture the index dependence in the data. Again, as desired, the UPCs detect no other serious issues.

Scenario #5: Second-order dependence. Finally, we suppose the true DGP and hypothesized model are again the same as in Scenario #1, except the true DGP follows an AR(2) structure with $Y_i = \phi Y_{i-1} + \text{sgn}(\phi)|\phi|^{1/4} Y_{i-2} + \sigma\varepsilon_i$. Thus, the true DGP exhibits second-order dependence, whereas the hypothesized model assumes only first-order dependence. In Figure 5.10, we see that $p_{\text{data,lag2}}$ is highly concentrated near zero, providing a clear indication that there is an issue pertaining to higher-order dependence. Interestingly, this perturbation of the model has a more global impact than in the other scenarios, affecting the other tests as well. None of the other p-value distributions as extreme as $p_{\text{data,lag2}}$, but all of them put significant mass on p-values very near zero.

6 Discussion

Uniform parametrization checks (UPCs) provide a general-purpose technique for Bayesian model criticism. As demonstrated in the examples, UPCs are easy-to-use and provide insight into which aspects of a model

are misspecified. As shown in the theory, UPCs are guaranteed to yield valid p-values under the null of model correctness and—in many cases—the u-values are uniform if and only if the model is correct. Compared to posterior predictive checks (PPCs), a key advantage of UPCs is that there is a default set of UPC tests that can be used on any model, rather than having to design PPC test quantities in a model-specific way.

There are several interesting directions for future work on UPCs. First, as observed in the autoregressive model example (Section 5.4), misspecification of one aspect of a model can affect UPCs pertaining to other aspects of the model. A more complete theory is needed for understanding the relationship between misspecification of each part of a model and the resulting posterior distribution of u-values of other parts. Relatedly, we have observed that some parametrizations are preferable, in that the effect of perturbing one part of a model is more isolated to the corresponding subset of u-values. Characterizing the role of model parametrization in this correspondence would make it possible to implement models in a way that is conducive to model criticism. Another important direction is to develop a standard "best practices" workflow for iterative model improvement, generalizing and formalizing the process illustrated in the logistic regression example (Section 5.3). Finally, it will be interesting to use the UPC methodology in other models and employ it in practical applications.

Acknowledgments

We would like to thank Aki Vehtari, Andrew Gelman, Yuling Yao, David Dunson, and Ryan Giordano for helpful comments. C.T.C. was supported by National Institutes of Health (NIH) Training Grant T32CA09337. J.W.M. was supported in part by the National Cancer Institute of the NIH under award number R01CA240299. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Abril-Pla, O., V. Andreani, C. Carroll, L. Dong, C. J. Fonnesbeck, M. Kochurov, R. Kumar, J. Lao, C. C. Luhmann, O. A. Martin, et al. (2023). PyMC: a modern, and comprehensive probabilistic programming framework in Python. *PeerJ Computer Science 9*, e1516.

Anderson, T. W. and D. A. Darling (1954). A test of goodness of fit. *Journal of the American Statistical Association 49*(268), 765–769.

Bayarri, M. and J. O. Berger (1999). Quantifying surprise in the data and model verification. *Bayesian Statistics 6*, 53–82.

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

Belin, T. R. and D. B. Rubin (1995). The analysis of repeated-measures data on schizophrenic reaction times using mixture models. *Statistics in Medicine 14*(8), 747–768.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological) 57*(1), 289–300.

Benjamini, Y. and D. Yekutieli (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 1165–1188.

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

Bolsinova, M. and J. Tijmstra (2016). Posterior predictive checks for conditional independence between response time and accuracy. *Journal of Educational and Behavioral Statistics 41*(2), 123–145.

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

Carpenter, B., A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell (2017a). Stan: A probabilistic programming language. *Journal of Statistical Software 76*, 1–32.

Carpenter, B., A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. A. Brubaker, J. Guo, P. Li, and A. Riddell (2017b). Stan: A probabilistic programming language. *Journal of Statistical Software 76*.

Chang, J. T. and D. Pollard (1997). Conditioning as disintegration. *Statistica Neerlandica 51*(3), 287–317.

Cook, S. R., A. Gelman, and D. B. Rubin (2006). Validation of software for Bayesian models using posterior quantiles. *Journal of Computational and Graphical Statistics 15*(3), 675–692.

Czado, C., T. Gneiting, and L. Held (2009). Predictive model assessment for count data. *Biometrics 65*(4), 1254–1261.

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

Dey, D. K., A. E. Gelfand, T. B. Swartz, and P. K. Vlachos (1998). A simulation-intensive approach for checking hierarchical models. *Test 7*(2), 325–346.

Durrett, R. (2019). *Probability: Theory and Examples*. Cambridge University Press.

Fisher, R. A. (1934). Statistical methods for research workers.

Foster, D. P. and R. A. Stine (2008). α-investing: a procedure for sequential control of expected false discoveries. *Journal of the Royal Statistical Society Series B: Statistical Methodology 70*(2), 429–444.

Froome, K., L. Essen, and R. A. Rhodes (1971). The velocity of light and radio waves. *Physics Today 24*(5), 49–49.

Gasparin, M., R. Wang, and A. Ramdas (2024). Combining exchangeable p-values. *arXiv preprint arXiv:2404.03484*.

Gelfand, A. E., D. K. Dey, and H. Chang (1992). Model determination using predictive distributions with implementation via sampling-based methods. *Bayesian Statistics 4*, 147–168.

Gelman, A. (2023, Apr). 4 different meanings of p-value (and how my thinking has changed) — statistical modeling, causal inference, and social science.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC.

Gelman, A., X.-L. Meng, and H. Stern (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 733–760.

Geweke, J. (2004). Getting it right: Joint distribution tests of posterior simulators. *Journal of the American Statistical Association 99*(467), 799–804.

Gosselin, F. (2011). A new calibrated bayesian internal goodness-of-fit method: Sampled posterior p-values as simple and general p-values that allow double use of the data. *PloS one 6*(3), e14770.

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) 29*(1), 83–100.

Harrell, Jr, F. E. (2024). *Hmisc: Harrell Miscellaneous*. R package version 5.1-2.

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

Hoeffding, W. (1948). A Non-Parametric Test of Independence. *The Annals of Mathematical Statistics 19*(4), 546–557.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 65–70.

Johnson, V. E. (2004). A Bayesian χ^2 test for goodness-of-fit. *The Annals of Statistics 32*(6), 2361–2384.

Johnson, V. E. (2007). Bayesian model assessment using pivotal quantities. *Bayesian Analysis 2*(4), 719–734.

Kruskal, W. H. and W. A. Wallis (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association 47*(260), 583–621.

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

Liu, Y. and J. Xie (2020). Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. *Journal of the American Statistical Association*.

Mann, H. B. and D. R. Whitney (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 50–60.

Marshall, E. and D. Spiegelhalter (2003). Approximate cross-validatory predictive checks in disease mapping models. *Statistics in Medicine 22*(10), 1649–1660.

Meng, X.-L. (1994). Posterior predictive p-values. *The Annals of Statistics 22*(3), 1142–1160.

Mimno, D., D. M. Blei, and B. E. Engelhardt (2015). Posterior predictive checks to quantify lack-of-fit in admixture models of latent population structure. *Proceedings of the National Academy of Sciences 112*(26), E3441–E3450.

Modrák, M., A. H. Moon, S. Kim, P. Bürkner, N. Huurre, K. Faltejsková, A. Gelman, and A. Vehtari (2023). Simulation-based calibration checking for Bayesian computation: The choice of test quantities shapes sensitivity. *Bayesian Analysis 1*(1), 1–28.

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

O'Hagan, A. (2003). HSSS model criticism. In *Highly Structured Stochastic Systems*. Oxford University Press.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling.

Roberts, G. O. and R. L. Tweedie (1996). Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli 2*, 341–363.

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

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

Talts, S., M. Betancourt, D. Simpson, A. Vehtari, and A. Gelman (2018). Validating Bayesian inference algorithms with simulation-based calibration. *arXiv preprint arXiv:1804.06788*.

Yuan, Y. and V. E. Johnson (2012). Goodness-of-fit diagnostics for Bayesian hierarchical models. *Biometrics 68*(1), 156–164.

Zhang, J. L. (2014). Comparative investigation of three bayesian p values. *Computational Statistics & Data Analysis 79*, 277–291.

Supplementary material for "Bayesian model criticism using uniform parametrization checks"

S1 Proofs

Proof of Theorem 4.1. Suppose $Y^0 \stackrel{d}{=} Y$. Since $\widetilde{\boldsymbol{\theta}} \mid Y^0$ and $\widetilde{U} \mid \widetilde{\boldsymbol{\theta}}, Y^0$ are defined according to the conditional distributions $\Pi_{\theta|y}$ and $\Pi_{u|\theta,y}$ under the hypothesized model, it follows that $(\widetilde{U}, \widetilde{\boldsymbol{\theta}}, Y^0) \stackrel{d}{=} (U, \boldsymbol{\theta}, Y)$, where $(U, \boldsymbol{\theta}, Y)$ is distributed according to the hypothesized model. In particular, $\widetilde{\boldsymbol{\theta}} \stackrel{d}{=} \boldsymbol{\theta}$ and $\widetilde{U} \stackrel{d}{=} U$. □

Proof of Theorem 4.2. Suppose $\widetilde{U} \stackrel{d}{=} U$ and g is a bijection. Recall that by definition, $\widetilde{U}$ is distributed according to $\Pi_{u|\theta,y}$ given $\widetilde{\boldsymbol{\theta}} = \theta$ and $Y^0 = y$, where $\Pi_{u|\theta,y}$ is the conditional distribution under the hypothesized model. Since $(\boldsymbol{\theta}, Y) = g(U)$ under the hypothesized model and g is a bijection, we have

$$U = g^{-1}(\boldsymbol{\theta}, Y). \tag{S1.1}$$

Since the conditional distribution of $\widetilde{U} \mid \widetilde{\boldsymbol{\theta}}, Y^0$ is defined according to the hypothesized model, we have

$$\widetilde{U} \stackrel{d}{=} g^{-1}(\widetilde{\boldsymbol{\theta}}, Y^0). \tag{S1.2}$$

Since $\widetilde{U} \stackrel{d}{=} U$ and g is a bijection, Equations S1.1 and S1.2 imply that $(\boldsymbol{\theta}, Y) \stackrel{d}{=} (\widetilde{\boldsymbol{\theta}}, Y^0)$. In particular, $Y \stackrel{d}{=} Y^0$. □

Proof of Theorem 4.3. This is a simple consequence of the assumed independence relations. We claim that for any random elements X, Y, and Z, if $X \perp\!\!\!\perp Y$ and $Z \perp\!\!\!\perp X \mid Y$ then $Z \perp\!\!\!\perp X$. To see this, observe that for any bounded, measurable real-valued functions f and g,

$$\begin{aligned}
\mathbb{E}\left(f(Z)g(X)\right) &= \mathbb{E}\left(\mathbb{E}\left(f(Z)g(X) \mid X, Y\right)\right) = \mathbb{E}\left(g(X)\,\mathbb{E}\left(f(Z) \mid X, Y\right)\right) \\
&= \mathbb{E}\left(g(X)\,\mathbb{E}\left(f(Z) \mid Y\right)\right) = \mathbb{E}\left(g(X)\right)\mathbb{E}\left(\mathbb{E}(f(Z) \mid Y)\right) \\
&= \mathbb{E}(g(X))\,\mathbb{E}(f(Z)).
\end{aligned}$$

Therefore, $Z \perp\!\!\!\perp X$. The result follows by applying this with $(\widetilde{U}, \widetilde{\boldsymbol{\theta}})$ and Y^0 playing the role of Z and Y. □

Proof of Theorem 4.4. (i) Let $u_{1:K}, u'_{1:K} \in (0,1)^K$ such that $u_{1:K} \neq u'_{1:K}$, that is, $u_k \neq u'_k$ for some k. Then, letting $\theta = g_{\mathrm{p}}(u_{1:K})$ and $\theta' = g_{\mathrm{p}}(u'_{1:K})$, we have $\theta \neq \theta'$ since g_{p} is one-to-one. Hence, $P_\theta \neq P_{\theta'}$. This establishes identifiability of $U_{1:K}$. (ii) Let $u, u' \in (0,1)^D$ such that $u \neq u'$. Letting $(\theta, y) = g(u)$

and $(\theta', y') = g(u')$, we have that either $\theta \neq \theta'$ or $y \neq y'$, since g is one-to-one. If $\theta \neq \theta'$ then $P_\theta \neq P_{\theta'}$. Otherwise, $y \neq y'$. This proves the claim. □

S2 Additional details for the normal model example

This section contains additional empirical results and implementation details for the example from Section 5.1 on the normal model for Newcomb's speed of light data.

S2.1 Additional empirical results

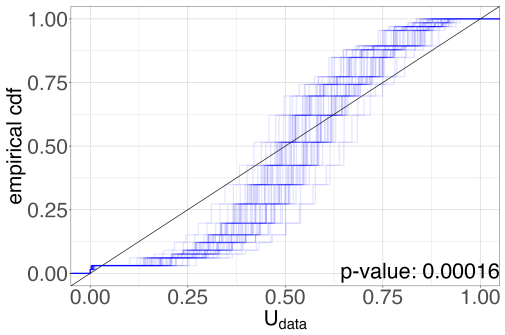
In Figure S2.1, we show the empirical CDF of the data u-values given (i) Newcomb's speed of light data and (ii) a simulated dataset from the normal distribution with mean and variance matching the sample mean and sample variance of the Newcomb data, under each of the three choices of prior considered. For all three choices of prior, the data u-values for the Newcomb data are clearly non-uniform, which correctly reflects the model misspecification. For the weakly informative and data-dependent priors, the data u-values for the normal data are approximately uniform. Meanwhile, the poorly chosen prior has such a detrimental effect that it also makes the data u-values non-uniform even under normally distributed data.

Recall that in Table 5.1, we report the aggregated p-values from the UPC tests. To visualize the distributions across which this aggregation occurs, Figure S2.2 shows the posterior densities of p_μ, p_σ, and $p_{\text{data,unif}}$ given the Newcomb data, under the data-dependent prior and the poorly chosen informative prior; see Figure 5.2 for the corresponding plots under the weakly informative prior. Under all three priors, the distribution of $p_{\text{data,unif}}$ is concentrated near zero, which leads to the small aggregated p-values $p_{\text{data,unif}^*}$ on the Newcomb data. The distributions of p_μ and p_σ are also concentrated near zero under the poorly chosen prior, which leads to the small aggregated values p_μ^* and p_σ^* for this prior.

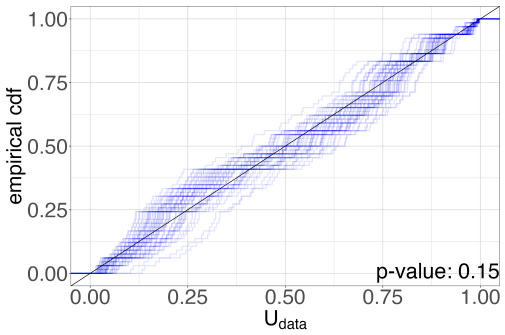
S2.2 Calculating the posterior densities of the p-values

Approximating the PPC p-values for $T(Y) = \min\{Y_1, \ldots, Y_n\}$. For any i.i.d. random variables $Y_1, \ldots, Y_n$ with common CDF F_Y, the CDF of the minimum $T(Y) = \min\{Y_1, \ldots, Y_n\}$ is given by

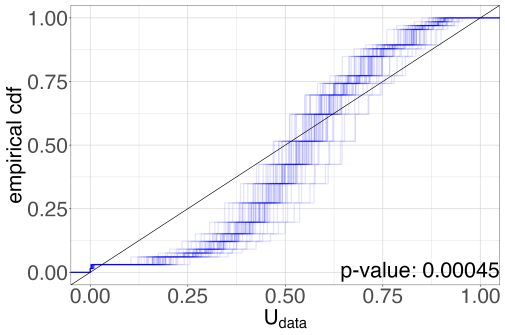
$$F_T(t) = 1 - (1 - F_Y(t))^n. \tag{S2.1}$$



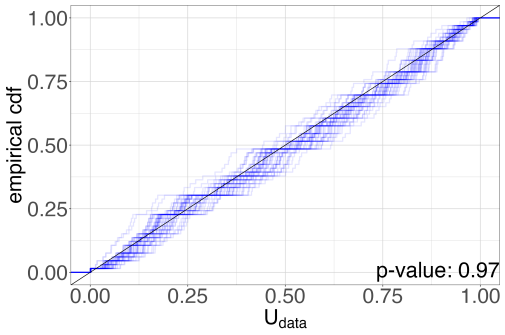
(a) Newcomb data, Weakly informative prior



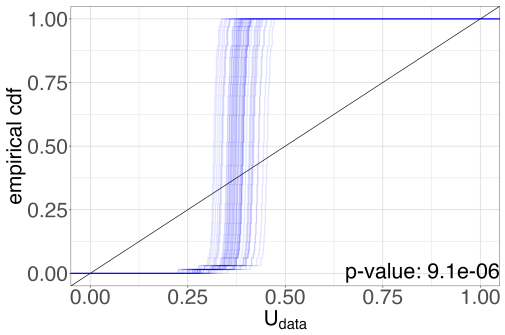
(b) Normal data, Weakly informative prior



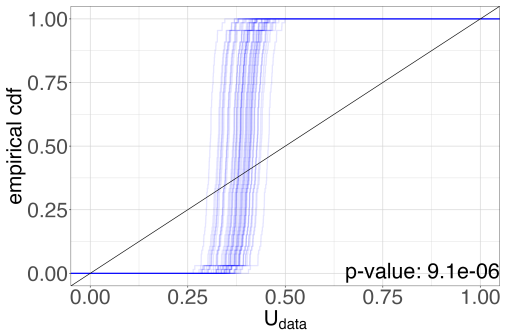
(c) Newcomb data, Data-dependent prior



(d) Normal data, Data-dependent prior



(e) Newcomb data, Poorly chosen prior



(f) Normal data, Poorly chosen prior

Figure S2.1: (Left) Empirical CDF $\hat{F}_{\mathrm{d}}$ of the data u-values $U_{d_1}, \ldots, U_{d_n}$ from multiple posterior samples given the Newcomb data, when using the (a) weakly informative prior, (c) data-dependent prior, and (e) poorly chosen informative prior. (Right) Same, but given a simulated dataset from a normal distribution with mean and variance matching the Newcomb data, when using the (b) weakly informative prior, (d) data-dependent prior, and (f) poorly chosen informative prior.

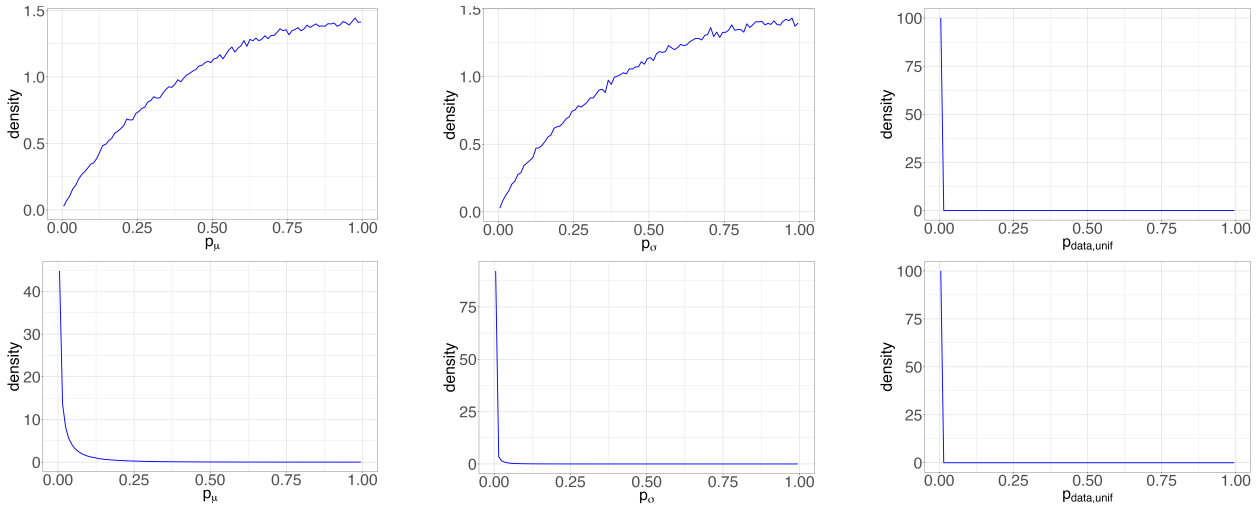


Figure S2.2: Results using UPCs on Newcomb data. Posterior densities of p_μ, p_σ, and $p_{\text{data,unif}}$ given the Newcomb data, under the data-dependent prior (top) and poorly chosen informative prior (bottom). See Figure 5.2 (top) for the corresponding plots under the weakly informative prior.

Given the dataset $y = (y_1, \ldots, y_n)$, let $Y^* = (Y_1^*, \ldots, Y_n^*)$ be distributed according to the posterior predictive distribution. Under the hypothesized normal model, the posterior predictive p-value for T given y is

$$\mathbb{P}\big(T(Y^*) \leqslant T(y) \,\big|\, y\big) = \int \mathbb{P}\big(T(Y^*) \leqslant T(y) \,\big|\, \mu, \sigma, y\big) f(\mu, \sigma \mid y)\, d\mu\, d\sigma \tag{S2.2}$$

$$\approx \frac{1}{S} \sum_{s=1}^{S} \mathbb{P}\big(T(Y^*) \leqslant T(y) \,\big|\, \mu_s, \sigma_s\big) \tag{S2.3}$$

where $(\mu_1, \sigma_1), \ldots, (\mu_S, \sigma_S)$ are samples drawn from the posterior of μ, σ given the dataset $y = (y_1, \ldots, y_n)$. By Equation S2.1,

$$\mathbb{P}\big(T(Y^*) \leqslant T(y) \,\big|\, \mu_s, \sigma_s\big) = 1 - \left(1 - \Phi\left(\frac{T(y) - \mu_s}{\sigma_s}\right)\right)^n, \tag{S2.4}$$

where Φ is the standard normal CDF. By plugging Equation S2.4 into Equation S2.2, we can obtain a better approximation to the posterior predictive p-value, compared to sampling $T(Y^*)$ directly.

Approximating the density of the UPC p-values for μ. To produce smooth plots of the densities of the p-values in Figure 5.2, we derive a partially analytic estimate of the density of p_μ given y. The following calculations are only for plotting purposes, and not needed to implement the UPC method.

To simplify the notation, we work with the precision $\lambda = 1/\sigma^2$ rather than the variance σ^2. In terms of μ and λ, our prior is $\mu \mid \lambda \sim \mathcal{N}(\mu_0, 1/(\kappa_0 \lambda))$ and $\lambda \sim \text{Gamma}(\alpha_0, \beta_0)$. Let $(\mu_1, \lambda_1), \ldots, (\mu_S, \lambda_S)$ be samples from the posterior given the dataset $y = (y_1, \ldots, y_n)$. Then we can form a Monte Carlo approximation to

the posterior density of p_μ, the UPC p-value for μ, via

$$f(p_\mu|y) = \int f(p_\mu|\lambda, y) f(\lambda|y) d\lambda \approx \frac{1}{S} \sum_{s=1}^{S} f(p_\mu|\lambda_s, y). \tag{S2.5}$$

As a function of μ and λ, the p-value for μ is

$$p_\mu = 2\min\{U_1, 1-U_1\} = 2\min\Big\{\Phi\big((\mu-\mu_0)\sqrt{\kappa_0\lambda}\big),\ 1-\Phi\big((\mu-\mu_0)\sqrt{\kappa_0\lambda}\big)\Big\}, \tag{S2.6}$$

where Φ is the standard normal CDF. By Jacobi's formula for transformation of random variables, the conditional density of p_μ is

$$f(p_\mu \mid \lambda, y) = f_{\mu|\lambda,y}\big(g_{1,\lambda}^{-1}(p_\mu) \mid \lambda, y\big)\left|\frac{d}{dp_\mu} g_{1,\lambda}^{-1}(p_\mu)\right| + f_{\mu|\lambda,y}\big(g_{2,\lambda}^{-1}(p_\mu) \mid \lambda, y\big)\left|\frac{d}{dp_\mu} g_{2,\lambda}^{-1}(p_\mu)\right|, \tag{S2.7}$$

where $f_{\mu|\lambda,y}(\cdot \mid \lambda, y)$ is the full conditional of $\mu|\lambda, y$ under the hypothesized model, specifically,

$$f_{\mu|\lambda,y}(\mu \mid \lambda, y) = \mathcal{N}\left(\mu \;\middle|\; \frac{\mu_0\kappa_0 + \sum_{i=1}^n y_i}{\kappa_0 + n},\ \frac{1}{(\kappa_0+n)\lambda}\right), \tag{S2.8}$$

and

$$\begin{aligned} g_{1,\lambda}^{-1}(p_\mu) &= \mu_0 + \Phi^{-1}(p_\mu/2)/\sqrt{\kappa_0\lambda} \\ \left|\frac{d}{dp_\mu} g_{1,\lambda}^{-1}(p_\mu)\right| &= \frac{1}{2\sqrt{\kappa_0\lambda}} \frac{1}{\varphi(\Phi^{-1}(p_\mu/2))} \\ g_{2,\lambda}^{-1}(p_\mu) &= \mu_0 + \Phi^{-1}(1-p_\mu/2)/\sqrt{\kappa_0\lambda} \\ \left|\frac{d}{dp_\mu} g_{2,\lambda}^{-1}(p_\mu)\right| &= \frac{1}{2\sqrt{\kappa_0\lambda}} \frac{1}{\varphi(\Phi^{-1}(1-p_\mu/2))} \end{aligned} \tag{S2.9}$$

where φ is the standard normal PDF.

Plugging Equations S2.6, S2.8, and S2.9 into Equation S2.7, and plugging Equation S2.7 into Equation S2.5 yields a partially analytic estimate of the density $f(p_\mu|y)$, which we find to be considerably more accurate than a purely Monte Carlo based estimate.

Approximating the density of the UPC p-values for σ. To produce smooth plots of the densities of the p-values in Figure 5.2, we derive an closed-form expression of the density of p_σ given y. The following calculations are only for plotting purposes, and are not needed to implement the UPC method.

Next, we derive a closed-form expression for the posterior density of the p-values for σ. As in the case of

μ, we reparametrize in terms of $\lambda = 1/\sigma^2$ to simplify the calculations. This reparametrization does not affect the p-value since there a strictly monotone relationship between σ and λ. By straightforward calculations, the posterior density of λ given y is

$$f_{\lambda|y}(\lambda) = \text{Gamma}(\lambda \mid \alpha_n, \beta_n) \tag{S2.10}$$

where $\alpha_n = \alpha_0 + n/2$ and

$$\beta_n = \beta_0 + \frac{1}{2}\sum_{i=1}^{n}(y_i - \bar{y})^2 + \frac{1}{2}\frac{\kappa_0 n}{\kappa_0 + n}(\bar{y} - \mu_0)^2.$$

As a function of λ, the p-value for λ is

$$p_\lambda = 2\min\{U_2, 1 - U_2\} = 2\min\Big\{F_{\text{Gamma}}(\lambda \mid \alpha_0, \beta_0),\, 1 - F_{\text{Gamma}}(\lambda \mid \alpha_0, \beta_0)\Big\} \tag{S2.11}$$

where $F_{\text{Gamma}}(x \mid \alpha_0, \beta_0) = \gamma(\alpha_0, \beta_0 x)/\Gamma(\alpha_0)$ is the CDF of a gamma distribution with shape α_0 and rate β_0. Here, Γ denotes the gamma function and γ is the lower incomplete gamma function.

When $\alpha_0 = \beta_0 = 1$, which are the only values we consider for this example, the posterior density of p_λ is

$$f(p_\lambda|y) = f_{\lambda|y}\big(-\log(1 - p_\lambda/2)\big)\frac{1}{2 - p_\lambda} + f_{\lambda|y}\big(-\log(p_\lambda/2)\big)\frac{1}{p_\lambda} \tag{S2.12}$$

by Jacobi's formula for transformation of random variables. Plugging Equations S2.10 and S2.11 into Equation S2.12 yields a closed-form expression for $f(p_\lambda|y)$.

S3 Additional details for the dependent Bernoulli trials example

This section contains additional empirical results and implementation details for the example from Section 5.2 on dependent Bernoulli trials.

S3.1 Additional empirical results

Recall that in Table 5.2, we report the aggregated p-values for the dependent Bernoulli trials example, for each of the three choices of prior (uniform, Jeffreys, and poorly chosen). To visualize how these aggregated p-values arise from the posteriors on UPC p-values, Figure S3.1 shows the posterior densities of the p-values p_θ, $p_{\text{data,unif}}$, and $p_{\text{data,indep}}$ given the dataset in Equation 5.1, under the Jeffreys and poorly chosen priors; see Figure 5.5 for the corresponding plots under the uniform prior. The posterior of $p_{\text{data,indep}}$ is concentrated near zero for all three priors, correctly indicating that the hypothesized i.i.d. Bernoulli model

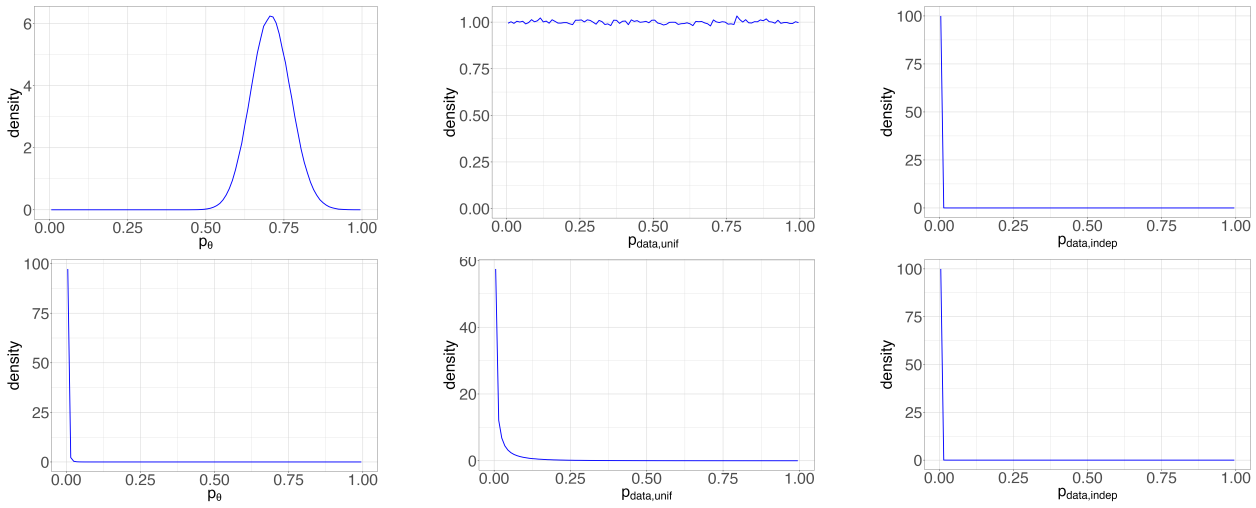


Figure S3.1: Results using UPCs on dependent Bernoulli example. Posterior densities of the p-values p_θ, $p_{\text{data,unif}}$, and $p_{\text{data,indep}}$ given the observed data in Equation 5.1, under the Jeffreys prior (top) and poorly chosen prior (bottom). See Figure 5.5 (top) for the corresponding plots under the uniform prior.

does not adequately capture the dependency in the observed data. This gives rise to the small aggregated p-values $p^*_{\text{data,indep}}$ seen in Table 5.2. Under the uniform and Jeffreys priors, the posterior of $p_{\text{data,unif}}$ is nearly uniform, reflecting the fact that the Bernoulli model must be correct marginally, and the parameter θ has been reasonably well estimated. Under the poorly chosen prior, all of the p-values are small; in the case of $p_{\text{data,unif}}$, this is because the prior too strongly biases the inferred value of θ.

S3.2 Hoeffding independence test

To calculate the test statistics for the Hoeffding independence test, we use the implementation in the Hmisc R package (Harrell, 2024). Hmisc uses a series of rules based on the exact value of the test statistic it calculates to produce a p-value. These rules produce p-values which give valid Type I error rate control under common α values and thus are reasonable to use for most practical use cases. However, they do not yield uniform p-values under the null. Thus, to obtain uniform p-values under the null, we construct an empirical null distribution by Monte Carlo simulation and we base our reported p-values on this instead. Specifically, to generate the empirical null, we (1) independently generate $U_1^{(j)}, \ldots, U_n^{(j)} \sim \text{Uniform}(0,1)$ i.i.d. and $V_1^{(j)}, \ldots, V_n^{(j)} \sim \text{Uniform}(0,1)$ i.i.d., for $j = 1, \ldots, J$ where $J = 10^5$, then (2) compute the Hoeffding test statistic value $t_j = T(U_{1:n}^{(j)}, V_{1:n}^{(j)})$ for $j = 1, \ldots, J$, and (3) store $t_1, \ldots, t_J$ as defining the empirical null distribution. Note, in particular, that the empirical null does not depend on the model or data at all – it only depends on the sample size n. Figure S3.2 shows the distribution of p-values when using the raw Hmisc procedure versus using our empirical null.

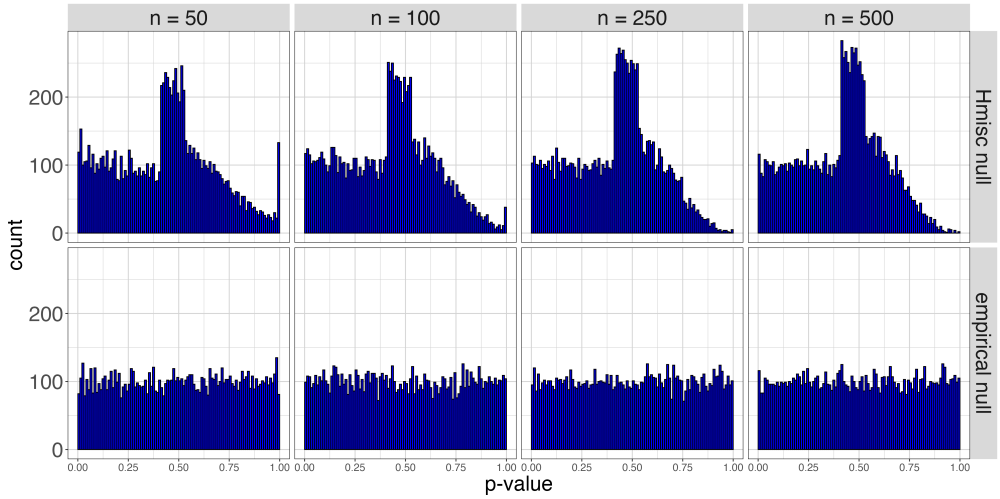


Figure S3.2: Calibration of the Hoeffding independence test: Distribution of p-values produced by the raw HMISC procedure and when using our empirical null distribution, for a range of sample sizes.

S4 Additional details for the logistic regression example

This section contains additional empirical results for the example from Section 5.3 on the logistic regression model for adolescent smoking data.

Model #1. Recall from the main text that we conduct five different tests on the u-values from Model #1. To use the corresponding p-values as part of a decision rule, we choose to control the FWER for this set of tests at the $\alpha = 0.1$ level using the Holm–Bonferroni correction. The unadjusted p-values are $p^*_{\text{data,wave}} = 1.67 \times 10^{-7}$, $p^*_{\text{data,sex}} = 0.72$, $p^*_{\text{data,parsmk}} = 8.47 \times 10^{-3}$, $p^*_{\alpha,\text{sex}} = 0.68$, and $p^*_{\alpha,\text{parsmk}} = 1.81 \times 10^{-11}$ which, after adjustment via the Holm–Bonferroni correction, become $p^{*,\text{adj}}_{\text{data,wave}} = 6.69 \times 10^{-7}$, $p^{*,\text{adj}}_{\text{data,sex}} \approx 1$, $p^{*,\text{adj}}_{\text{data,parsmk}} = 0.03$, $p^{*,\text{adj}}_{\alpha,\text{sex}} \approx 1$, and $p^{*,\text{adj}}_{\alpha,\text{parsmk}} = 9.07 \times 10^{-11}$.

Model #2. Recall that our two tests for Model #2 yield unadjusted p-values of $p^*_{\text{data,wave}\times\text{parsmk}} \approx 1$ and $p^*_{\alpha,\text{unif}} = 3.41 \times 10^{-7}$. After Holm–Bonferroni correction, the adjusted p-values are $p^{*,\text{adj}}_{\text{data,wave}\times\text{parsmk}} \approx 1$ and $p^{*,\text{adj}}_{\alpha,\text{unif}} = 6.82 \times 10^{-7}$. Note that these p-values are independent under H_0, so we could alternatively use Fisher's p-value combination method (Fisher, 1934), rather than Holm–Bonferroni.

Model #3. To interpret the latent variables Z_j in Model #3, we split the individuals into three groups: "always smokers" (those who smoke regularly at every wave of the study), "never smokers" (those who never smoke regularly), and "sometimes smokers" (those who switch between smoking regularly and not throughout the study); these groups were previously used by Gelman et al. (2013, Section 6.3) to define PPC

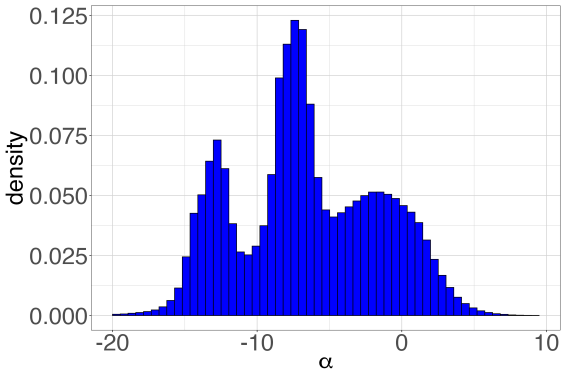


Figure S4.1: UPC results for Model #3 on the logistic regression example. Histogram of the α values over all posterior samples.

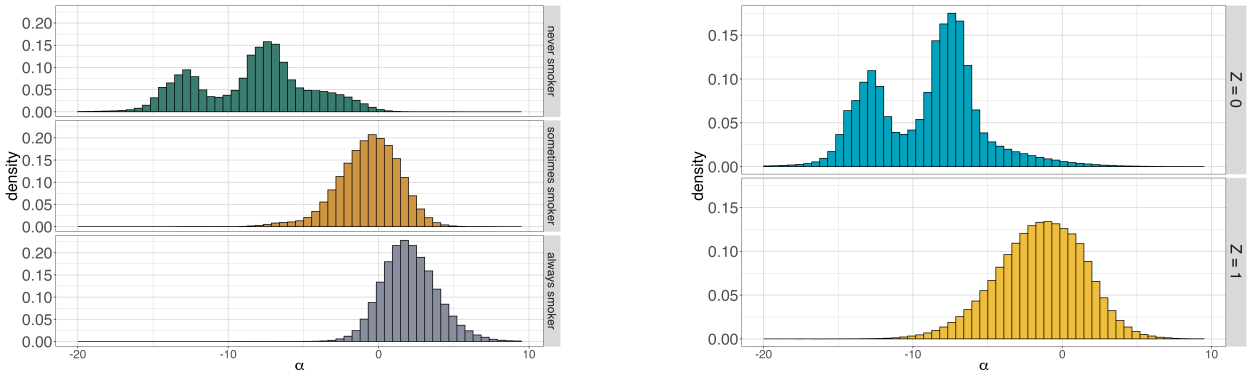


Figure S4.2: Additional results for Model #3 for logistic regression smoking example: Distribution of α values by smoking status, averaged over 1,000 posterior samples. Distribution of α values conditional on Z_j, averaged over 1,000 posterior samples.

test statistics. Table S4.1 shows the posterior mean of the fraction of individuals in each group, given $Z_j = 0$ and $Z_j = 1$, respectively. We see that the latent variable Z_j correlates strongly with the "never smoker" group, specifically, nearly all of the individuals with $Z_j = 0$ are "never smokers". Thus, the two mixture components in the prior on α represent groups with higher and lower probability of smoking, respectively, beyond the effects due to `wave`, `sex`, and `parsmk` (parent smoking status). To visualize the relationship between the α values and these groups, Figure S4.2 shows histograms of the empirical distribution of α values for individuals in the "never smoker", "sometimes smoker", and "always smoker" groups averaged over 1,000 posterior samples. Figure S4.2 also shows histograms of the empirical distribution of α values for individuals with $Z_j = 0$ and $Z_j = 1$, averaged over 1,000 posterior samples.

posterior Z_j	**never smoker**	**sometimes smoker**	**always smoker**
$Z_j = 0$	0.96	0.04	6.4×10^{-3}
$Z_j = 1$	0.41	0.45	0.14

Table S4.1: Probabilities of latent smoking status conditional on posterior Z_j.

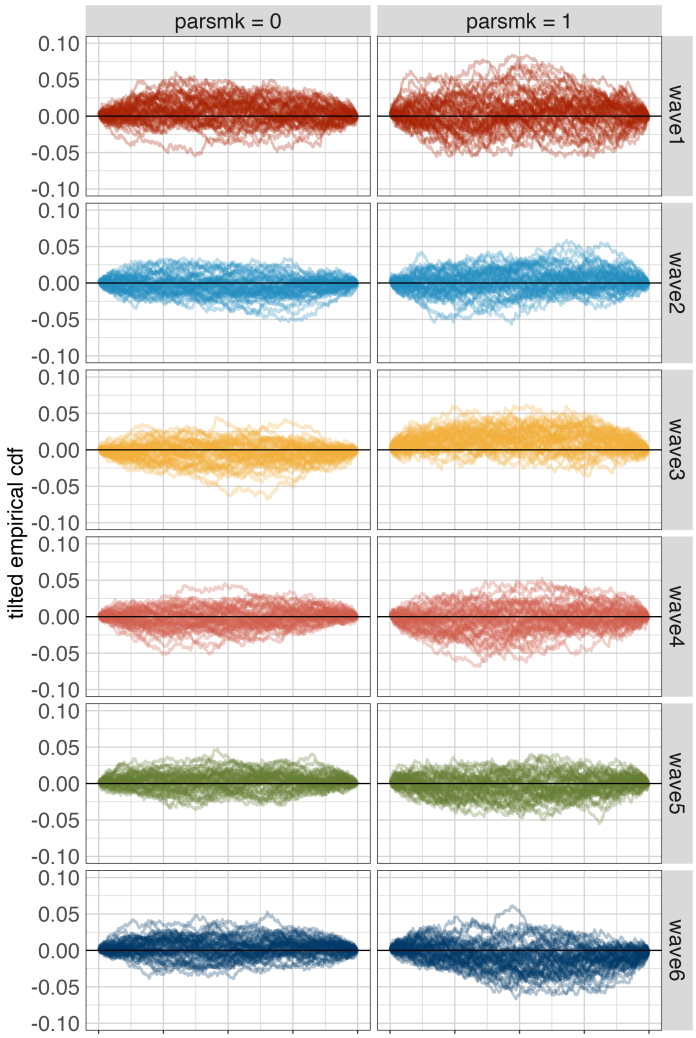


Figure S4.3: Additional results for Model #3 for the logistic regression example: Tilted CDFs for data u-values stratified by `wave` × `parsmk`. Visually, there is little variation in the data u-value distribution across values of `wave` × `parsmk`, corroborating the fact that $p^*_{\text{data,wave}\times\text{parsmk}}$ is very large.