Bayesian model criticism using uniform parametrization checks

Jeffrey W. Miller

Harvard T.H. Chan School of Public Health
Department of Biostatistics

SDS Seminar || University of Texas, Austin || Nov 7, 2025

Preprint: Christian T. Covington and J.W.M., 2025,
https://arxiv.org/abs/2503.18261

Christian Covington

Outline

Outline

Motivation

- 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, 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.
- We would like to do this in a computationally convenient and statistically rigorous way.

The standard approach: Posterior predictive checks

- Posterior predictive checks (PPCs) are the most commonly used technique for Bayesian model criticism.
- PPCs compare the observed value of a test quantity to its distribution under the posterior predictive.
  - Guttman (1967), Rubin (1984), Meng (1994), Gelman et al. (2013)
- However, it is well known that the "p-values" produced by PPCs are not uniformly distributed.
  - Bayarri and Berger (1999, 2000), Robins et al. (2000)

The standard approach: Posterior predictive checks

- Asymptotically uniform PPC p-values can be obtained using a partial posterior or conditional predictive, but this tends to be model specific.
  - Bayarri and Berger (1999, 2000), Robins et al. (2000)

- Data splitting also yields uniform PPC p-values, but this entails a loss of information and often involves multiple posterior inference runs.
  - Moran et al. (2019), Li and Huggins (2022)

- A more fundamental limitation is the need to design test quantities for the model at hand. Constructing good PPC test quantities requires:
  1) confidence about the kinds of misspecification that may be present and
  2) statistical insight into what makes a good PPC test quantity.

- Many other model criticism methods have been proposed as well, but they tend to be either (a) not well-calibrated, (b) model specific, or (c) computationally intensive.

Outline

Uniform parametrization checks (UPCs): Setup

- Consider a hypothesized model consisting of
  - prior: $\theta \sim \Pi$, and
  - likelihood: $Y \sim P_\theta$ given θ, where Y is the dataset.

- This defines a joint distribution on parameters and dataset, (θ, Y).

- 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 $U_d \overset{\text{iid}}{\sim} \text{U}(0, 1)$.

- 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*.

Uniform parametrization checks (UPCs): Main idea

- To perform model criticism, we view the hypothesized model as the null hypothesis.

- To test for departures, we sample from the posterior of U.
  - That is, we sample from $U|Y$, where the joint distribution of (U, Y) is induced by $(\theta, Y) = g(U)$ and $U \sim \text{Uniform}_D(0, 1)$.

- Key fact: If Y is drawn from the hypothesized model and U is drawn 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, $U_d \overset{\text{iid}}{\sim} \text{U}(0, 1)$.

  - This is a trivial consequence of conditional probabilities.

- There is no approximation here – if the model is correct, then a posterior draw of U is exactly uniform.

Uniform parametrization checks (UPCs): Main idea

- Thus, if we detect that $U_1, \ldots, U_D$ are not i.i.d. $\mathrm{U}(0,1)$ under the posterior, this implies misspecification of some aspect of the model.
- This enables model criticism in a simple yet powerful way by testing for departures from uniformity or independence in various respects.
- Specifically, we can probe different aspects of the model for possible misspecification simply by applying classical hypothesis tests for dependence or non-uniformity to subsets of $U_1, \ldots, U_D$.

Example: Autoregressive model

- Consider a 1st order autoregression model such that, for $i = 1, \ldots, n$,

$$Y_i = \phi Y_{i-1} + \sigma \varepsilon_i$$

where $\varepsilon_1, \ldots, \varepsilon_n$ i.i.d. $\sim \mathcal{N}(0, 1)$ and $Y_0 = 0$.

- For the prior, take $\phi \sim \mathrm{U}(-0.5, 0.5)$ and $\sigma \sim \mathrm{Exp}(1)$ 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. $\sim \mathrm{U}(0, 1)$, by setting

$$\begin{aligned}
\phi &= F_\phi^{-1}(U_1), \\
\sigma &= F_\sigma^{-1}(U_2), \\
\varepsilon_i &= \Phi^{-1}(U_{d_i}), \\
Y_i &= \phi Y_{i-1} + \sigma \varepsilon_i,
\end{aligned}$$

where F_ϕ, F_σ, and Φ denote the CDFs of ϕ, σ, and $\mathcal{N}(0, 1)$, respectively, and $d_i = 2 + i$.

Computing the u-values

- Sampling $U \mid Y$ is usually very straightforward if we have a way of sampling from $\theta \mid Y$.

- Given a posterior sample of $\theta \mid Y$, we need to sample $U \mid \theta, Y$.
  - Recall that $(\theta, Y) = g(U)$ and $U \sim \text{Uniform}_D(0, 1)$.

- If g is invertible, then this is simply done by setting $U = g^{-1}(\theta, Y)$.

- For instance, in the autoregressive example, we compute

$$\begin{aligned} U_1 &= F_\phi(\phi), \\ U_2 &= F_\sigma(\sigma), \\ U_{d_i} &= \Phi(\varepsilon_i), \text{ where } d_i = 2 + i \text{ and} \\ \varepsilon_i &= (Y_i - \phi Y_{i-1})/\sigma. \end{aligned}$$

- If g is non-invertible, e.g., when there are discrete variables, then we need to randomly sample $U \mid \theta, Y$, but this is usually trivial.

Default choices of UPC test

- Since any subset of u-values is i.i.d. uniform under the null,
  1) the u-values can be grouped in various ways and
  2) the same tests can be applied to any model.

- For illustration, consider the autoregressive model.

- *Testing for extreme values:*
  - Is U_2 extremely close to 0 or 1? Suggests prior on σ may be bad.
  - Is U_{d_i} extremely close to 0 or 1? Suggests Y_i may be an outlier.

- *Testing for non-uniformity:*
  - Are $U_{d_1}, \ldots, U_{d_n}$ non-uniform? Normal assumption may be incorrect.

- *Testing for internal dependence:*
  - Are U_{d_i} and U_{d_i+2} dependent? 1st order assumption may be incorrect.

- *Testing for external dependence:*
  - Is U_{d_i} dependent on covariate x_i? Suggests heteroskedasticity or trend.

Features of UPCs versus PPCs

- Similarities between UPCs and PPCs:
  - Both help reveal which aspects of a model are misspecified.
  - Both are applicable to a wide range of models.
  - Both are computationally tractable and easy to implement.
- Advantages of UPCs relative to PPCs:
  - Exactly uniform p-values under the null that the model is correct.
  - Natural default choices of tests.
  - Intuitively clear how to design new customized tests.

Combining multiple posterior samples

- Since there is randomness in any one posterior sample, it is preferable to aggregate across many posterior samples $U^{(1)}, \ldots, U^{(T)} \in (0,1)^D$.
- However, $U^{(1)}, \ldots, U^{(T)}$ are not independent, since they all depend on the same dataset Y.
- Thus, we cannot simply pool all of the u-values together to perform tests, due to the dependency.

Combining multiple posterior samples

- For a given test, we combine posterior samples as follows:
  - Apply the test to each posterior sample $U^{(t)} = (U_1^{(t)}, \ldots, U_D^{(t)})$ to obtain a p-value $p^{(t)}$.
  - Combine $p^{(1)}, \ldots, p^{(T)}$ using the Cauchy combination method (Liu and Xie, 2020), specifically,

$$p^* = 1 - F_{\text{Cauchy}}\left(\frac{1}{T}\sum_{t=1}^{T} F_{\text{Cauchy}}^{-1}(1 - p^{(t)})\right)$$

    where F_{Cauchy} is the CDF of the Cauchy distribution.
  - Compare p^* to a pre-specified level α to decide whether to reject the null of model correctness.

- Under quite general dependence structures, p^* approximately controls Type I error for $\alpha \in (0, 0.05)$ and exhibits good power.

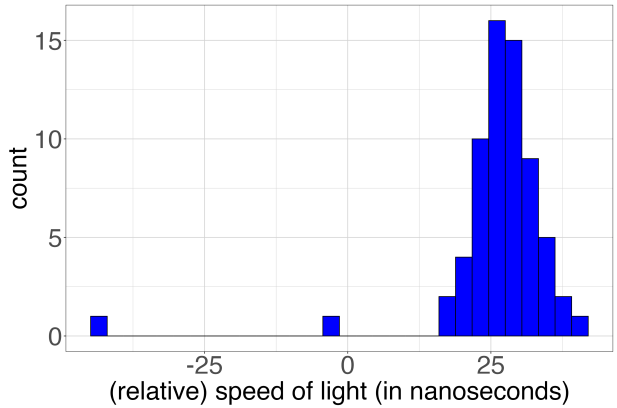
Outline

Related work

- Posterior predictive checks (PPCs)
  - Guttman (1967), Rubin (1984), Meng (1994), Gelman et al. (1996)
- Simulation-based calibration (SBC)
  - Used to check the validity of posterior inference algorithms.
  - Relies on the same key fact as UPCs: If the model is correct, then a single posterior draw is distributed according to the prior.
  - Geweke (2004), Cook et al. (2006), Talts et al. (2018), Modrák et al. (2024)
- Pivotal discrepancy measures (PDMs)
  - Model criticism using test quantities $T(Y, \theta)$ whose distribution is invariant to the value of θ when $Y \sim P_\theta$.
  - When g^{-1} exists, UPCs based on the data u-values can be viewed as PDMs. However, not all UPCs are PDMs, and not all PDMs are UPCs.
  - Johnson (2007), Yuan & Johnson (2012), Gosselin (2011), Zhang (2014)
- Many other model criticism methods in the literature
  - Dey et al. (1998), Hjort et al. (2006), Gelfand et al. (1992), Marshall & Spiegelhalter (2003), O'Hagan (2003), Dahl et al. (2007), to mention a few.

Outline

Normal model for Newcomb's speed of light data

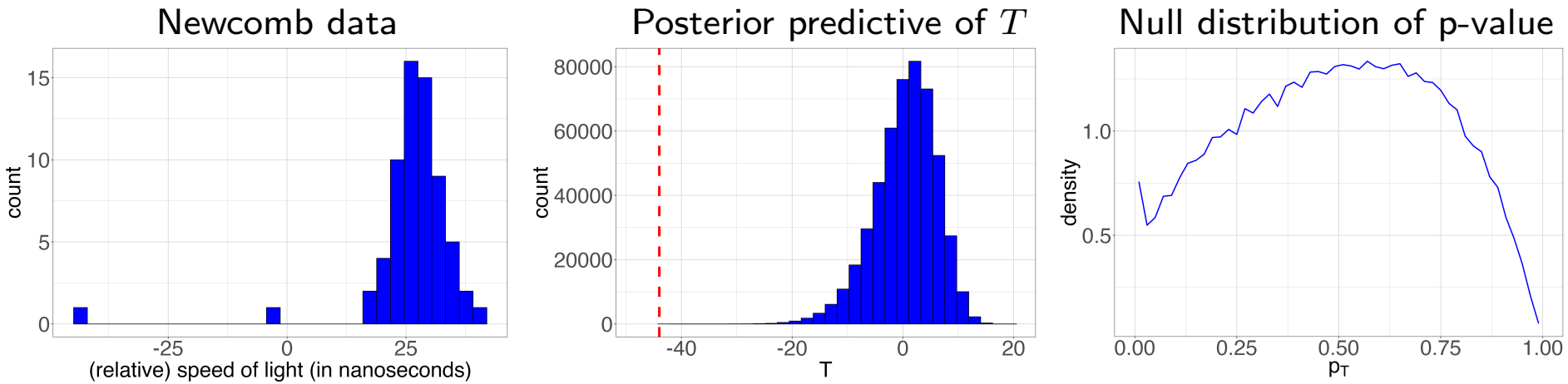


- Simon Newcomb's 66 measurements of the speed of light from 1882.
- Common example for model criticism methods (Gelman et al., 2013).
- Hypothesized model:
  - $Y_1, \ldots, Y_n$ i.i.d. $\sim \mathcal{N}(\mu, \sigma^2)$.
  - Weakly informative Normal-InverseGamma prior:

$$\mu \mid \sigma^2 \sim \mathcal{N}(\mu_0, \sigma^2/\kappa_0)$$
$$\sigma^2 \sim \text{InvGamma}(\alpha_0, \beta_0)$$

with $\mu_0 = 0$, $\kappa_0 = 1/10$, $\alpha_0 = 2$, and $\beta_0 = 300$.

Normal model for Newcomb data: PPC results



- Gelman et al. (2013) suggest using $T = \min\{Y_1, \ldots, Y_n\}$ as a PPC test statistic. Observed value of T is red dotted line in middle plot.
- This PPC successfully detects an issue with the model.
- However, the PPC p-value is not uniform under the null of model correctness. $\Rightarrow$ Loss of power; hard to interpret PPC p-value.
- Also, it seems that this choice of T was made by looking at the data.

Normal model for Newcomb data: UPC approach

- For the UPC approach, we write

$$\sigma^2 = F_{\sigma^2}^{-1}(U_2)$$
$$\mu = \mu_0 + \sigma\kappa_0^{-1/2}\Phi^{-1}(U_1)$$
$$Y_i = \mu + \sigma\Phi^{-1}(U_{d_i})$$

  where $d_i = 2 + i$.

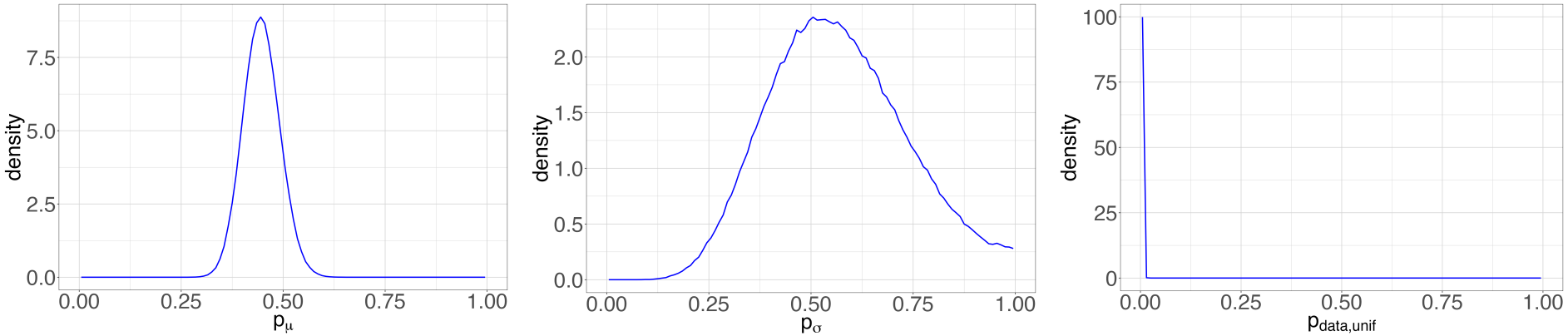
- We compute p-values for three default UPC tests:
  - *Test for extreme values of μ:* $p_\mu = 2\min\{U_1, 1 - U_1\}$.
  - *Test for extreme values of σ:* $p_\sigma = 2\min\{U_2, 1 - U_2\}$.
  - *Test for non-uniformity of the data u-values:*
    $p_{\text{data,unif}}$ = p-value from Anderson–Darling test on $U_{d_1}, \ldots, U_{d_n}$.

- Aggregating posterior samples via Cauchy combination method:

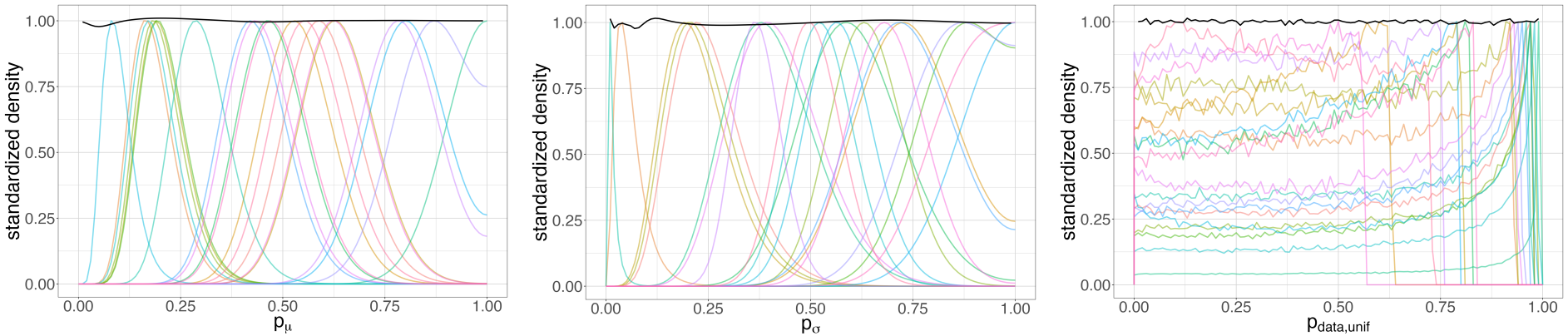
$$p_\mu^* = 0.45 \qquad p_\sigma^* = 0.83 \qquad p_{\text{data,unif}}^* = 1.60 \times 10^{-4}.$$

Normal model for Newcomb data: Visualizing the UPCs

Posteriors of the UPC p-values given the Newcomb data



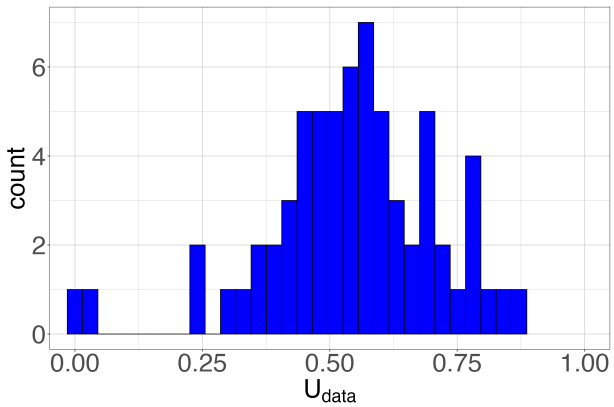
Posteriors of the UPC p-values given null-distributed data from the model

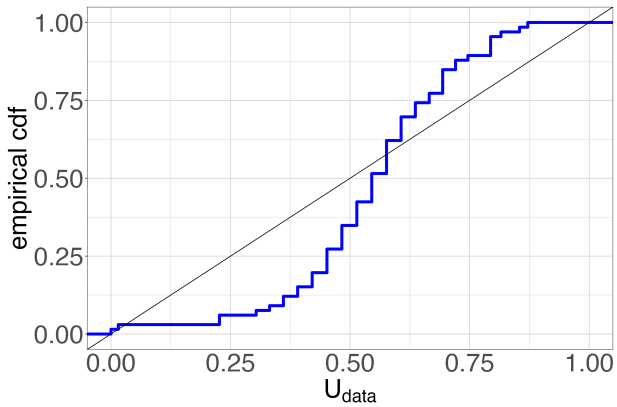


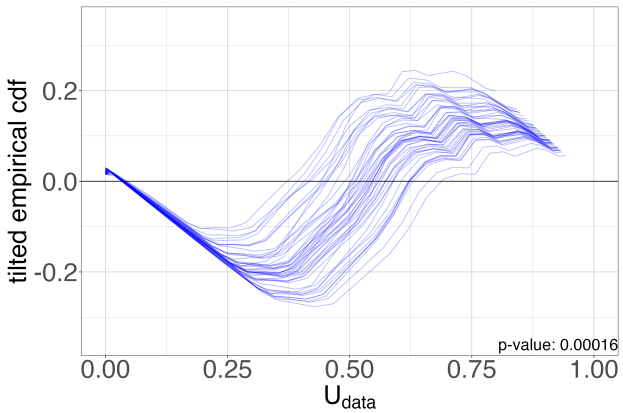
- UPCs correctly indicate no issues with the priors.
- UPCs correctly indicate a problem with the normal outcome aspect.

Normal model for Newcomb data: Visualizing the UPCs

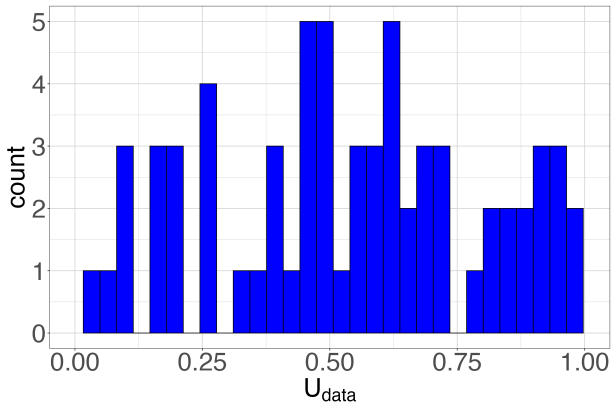
Visualizing the data u-values for the Newcomb data

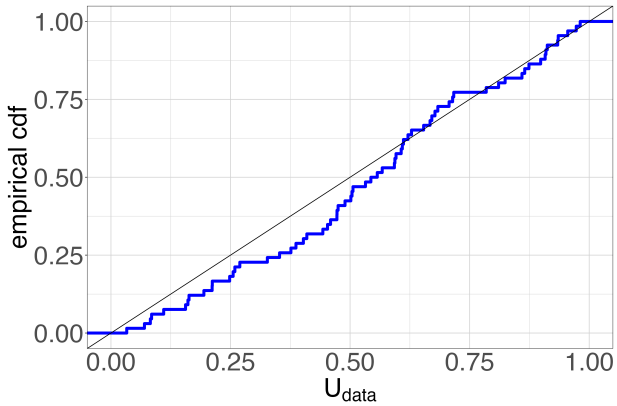


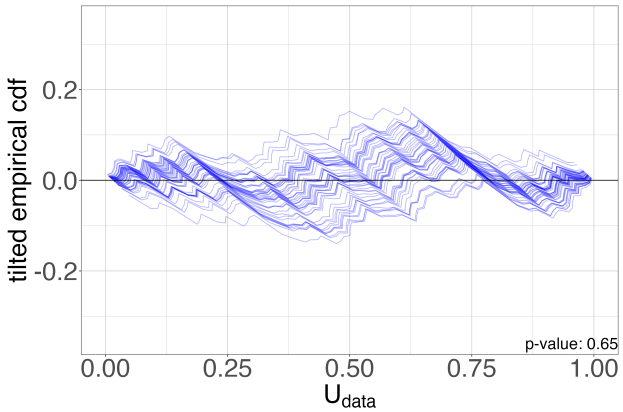




Visualizing the data u-values for null-distributed data from the model







Dependent Bernoulli trials

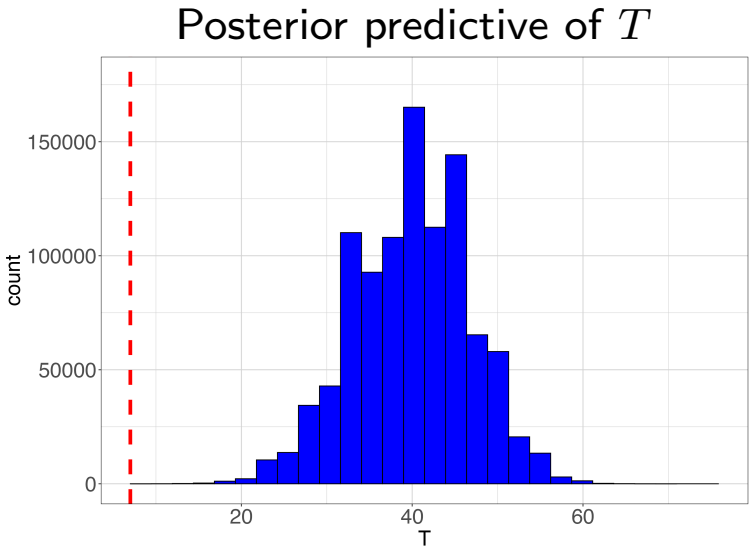
Observed data:

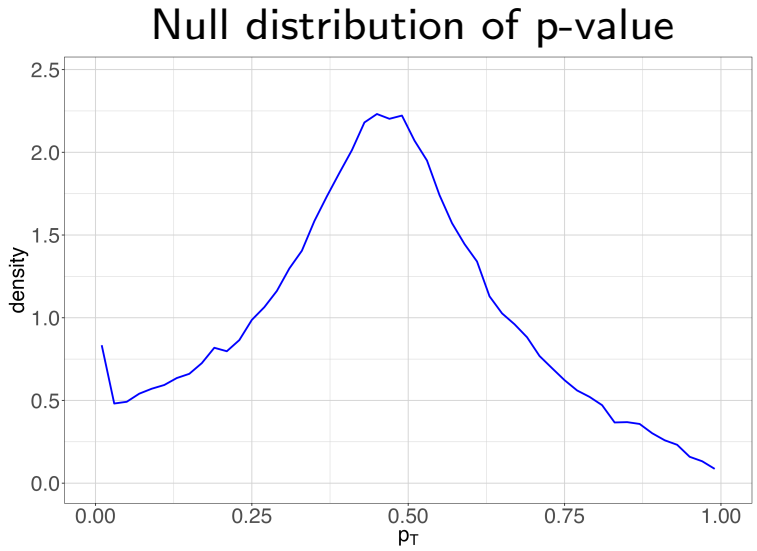
0 0 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1.

- Consider the simulated binary data above, inspired by the example of Gelman et al. (2013, Section 6.3).

- Hypothesized model:
  - $Y_1, \ldots, Y_n$ i.i.d. $\sim \text{Bernoulli}(\theta)$.
  - Uniform prior: $\theta \sim \text{Beta}(1, 1)$.

Dependent Bernoulli trials: PPC results





- Gelman et al. (2013) suggest using the number of switches, $T = \sum_{i=1}^{n-1} \mathbb{1}(Y_i \neq Y_{i+1})$, as a PPC test statistic.
- This PPC successfully detects an issue with the model.
- However, again, it is not uniform under the null of model correctness.

Dependent Bernoulli trials: UPC approach

- To implement the UPC approach, we write

$$\theta = F_\theta^{-1}(U_1) = U_1,$$
$$Y_i = \mathbb{1}(U_{i+1} \geq 1 - \theta).$$

- This mapping from U to (θ, Y) is not invertible due to the discreteness of Y_i.

- Thus, it is necessary to randomly sample $U \mid \theta, Y$, rather than deterministically compute it from θ and Y.

- For each posterior sample of $\theta \mid Y$, we sample $U \mid \theta, Y$ via

$$U_1 = \theta,$$
$$U_{i+1} \sim \begin{cases} \mathrm{U}(0,\, 1-\theta) & \text{if } Y_i = 0 \\ \mathrm{U}(1-\theta,\, 1) & \text{if } Y_i = 1 \end{cases}$$

  independently for $i = 1, \ldots, n$.

Dependent Bernoulli trials: UPC approach

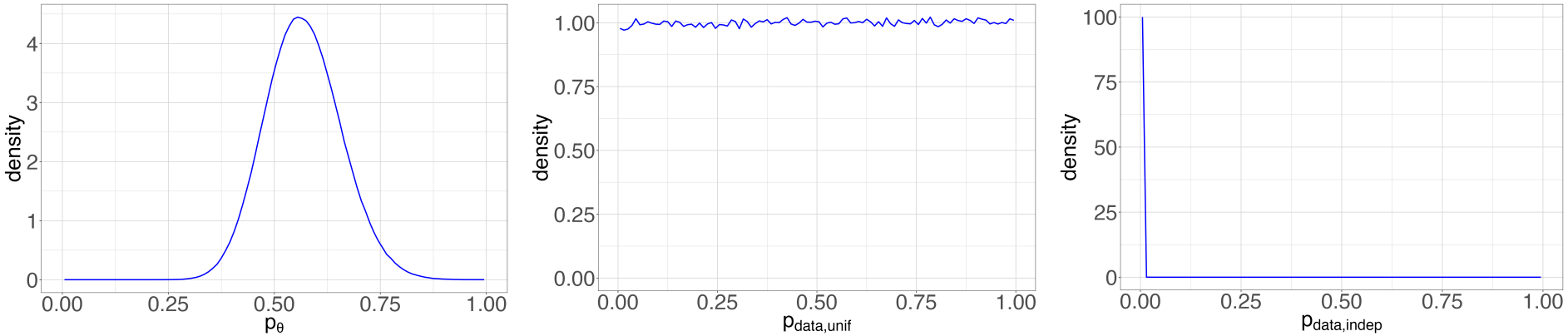
- We compute p-values for three default UPC tests:
  - *Test for extreme values of θ:* $p_\theta = 2\min\{U_1, 1 - U_1\}$.
  - *Test for non-uniformity of the data u-values:*
    $p_{\text{data,unif}}$ = p-value from Anderson–Darling test on $U_2, \ldots, U_{n+1}$.
  - *Test for dependence of successive data u-values:*
    $p_{\text{data,indep}}$ = p-value from Hoeffding's test on (U_i, U_{i+1}) for $i = 2, \ldots, n+1$.
- Aggregating posterior samples via Cauchy combination method:

$$p^*_\theta = 0.58$$
$$p^*_{\text{data,unif}} = 0.74$$
$$p^*_{\text{data,indep}} = 4.61 \times 10^{-6}.$$

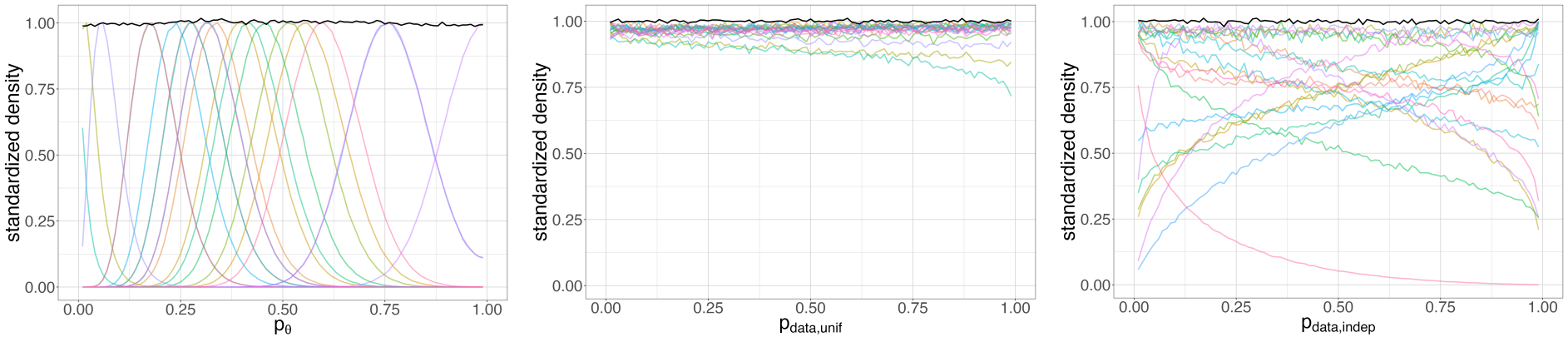
- Thus, the UPCs correctly find no issues with the prior or the Bernoulli aspect of the model, and correctly detect the serial dependence.

Dependent Bernoulli trials: Visualizing the UPCs

Posteriors of the UPC p-values given the observed Bernoulli data



Posteriors of the UPC p-values given null-distributed data from the model

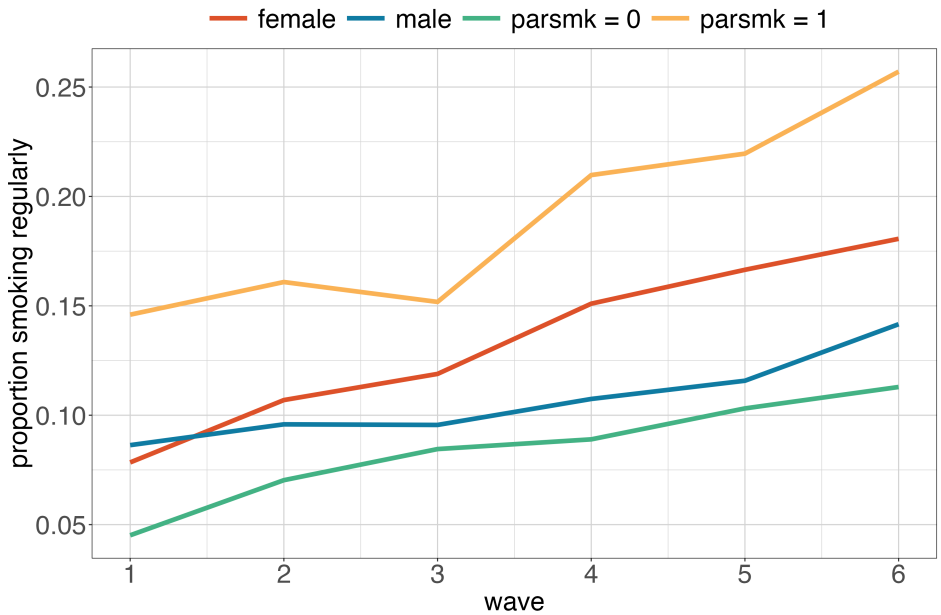


Logistic regression model for adolescent smoking data

- We illustrate performing iterative model improvement with UPCs, using a logistic regression example from Gelman et al. (2013).

- Data: $n = 8{,}730$ observations from $m = 1{,}760$ individuals who were surveyed at up to six time points (`wave`) and asked whether or not they smoked regularly (`smkreg`).

- Sex and parental smoking status (`parsmk`) were also recorded.



Logistic regression example: Model #1

- Let $Y_{jk} \in \{0, 1\}$ denote whether individual j smokes regularly at wave k.

- Model #1: Random effects model with weakly informative priors.

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

  for individual $j \in \{1, \ldots, m\}$ at wave $k \in \{1, \ldots, 6\}$.

Logistic regression example: Model #1 results

- Is it problematic that none of the covariates were used in this model?
- To address this, we *test for external dependence* between u-values and covariates.

- Test for dependence between data u-values and `wave`, `sex`, `parsmk`:

$$p^*_{\text{data},\texttt{wave}} = 1.67 \times 10^{-7}$$
$$p^*_{\text{data},\texttt{sex}} = 0.72$$
$$p^*_{\text{data},\texttt{parsmk}} = 8.47 \times 10^{-3}.$$

- Test for dependence between α u-values and `sex`, `parsmk`:

$$p^*_{\alpha,\texttt{sex}} = 0.68$$
$$p^*_{\alpha,\texttt{parsmk}} = 1.81 \times 10^{-11}.$$

(We exclude `wave` since α is subject-level and `wave` is observation-level.)

- This is evidence of dependence with `wave` and `parsmk`, but not `sex`.

Logistic regression example: Model #2

- Model #2: Augment to include `wave` and `parsmk` as covariates.

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

- We first verify that dependence with `wave` and `parsmk` is no longer detected. Indeed, running the same tests as before yields

$$
\begin{aligned}
p^*_{\text{data},\texttt{wave}} &\approx 1 \\
p^*_{\text{data},\texttt{sex}} &= 0.55 \\
p^*_{\text{data},\texttt{parsmk}} &\approx 1 \\
p^*_{\alpha,\texttt{sex}} &= 0.75 \\
p^*_{\alpha,\texttt{parsmk}} &= 0.20.
\end{aligned}
$$

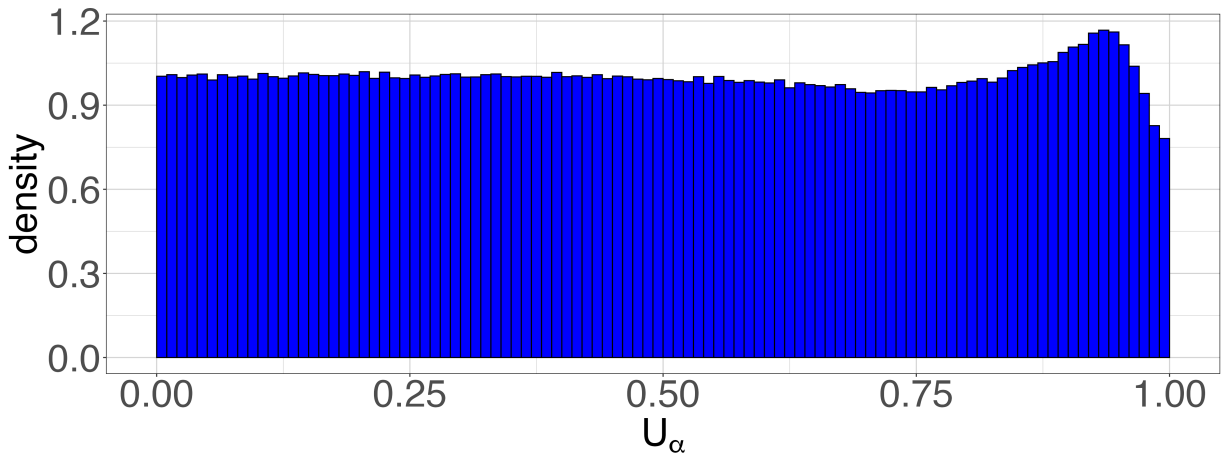
Logistic regression example: Model #2 results

- To explore possible deficiencies in Model #2, we
  1) test for dependence between the data u-values and `wave` $\times$ `parsmk`, to assess whether this interaction is needed,
  2) test for non-uniformity of the α u-values, to evaluate the choice of prior on the α's.

- These tests yield aggregated p-values of

$$p^*_{\text{data},\texttt{wave}\times\texttt{parsmk}} \approx 1$$
$$p^*_{\alpha,\text{unif}} = 3.41 \times 10^{-7}.$$

- Histogram of the α u-values over all posterior samples:



Logistic regression example: Model #3

- Model #3: Augment to use a two-component mixture prior on α's.

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

- Running the same tests as before yields

$$
\begin{aligned}
p^*_{\text{data},\texttt{wave}\times\texttt{parsmk}} &\approx 1 \\
p^*_{\alpha,\text{unif}} &= 0.78.
\end{aligned}
$$

- In fact, the null of model correctness is not rejected for any of the tests considered in this example.

Outline

Choice of parametrization

- There is not a unique choice of function g. That is, different choices of g can produce the same distribution on (θ, Y).
- Usually, the model specification will suggest a natural way of choosing g by following the generative process that defines the model.
- That said, our preliminary results do suggest that some choices of g are preferable to others.
- Specifically, some parametrizations yield a more direct correspondence between UPCs and types of misspecification.
- Understanding this better is an important direction for future work.

Summary

Features of UPCs

- General-purpose technique for Bayesian model criticism.
- Provide insight into which aspects of a model are misspecified.
- Easy to implement and computationally efficient.
- Exactly uniform p-values under the null of model correctness.
- Good power by aggregating p-values across posterior samples.
- Default tests that can be used with any model.

Bayesian model criticism using uniform parametrization checks

Jeffrey W. Miller

Harvard T.H. Chan School of Public Health
Department of Biostatistics

SDS Seminar || University of Texas, Austin || Nov 7, 2025

Preprint: Christian T. Covington and J.W.M., 2025,
https://arxiv.org/abs/2503.18261

Christian Covington