Bayesian Mixture Modeling of Significant P Values: A Meta-Analytic Method to Estimate the Degree of Contamination from $\mathcal{H}_0$: Supplemental Material

Quentin Frederik Gronau[1], Monique Duizer[1], Marjan Bakker[2], & Eric-Jan Wagenmakers[1]

[1]University of Amsterdam

[2]Tilburg University

I. How to Use the Online Web Application

In this section, we describe how researchers can use our online application (`https://qfgronau.shinyapps.io/bmmsp/`) to analyze any set of significant p values. To use the app, one needs to save the set of significant p values that should be analyzed in a text file as a single column without a column name. Figure 1 shows a screenshot of the input panel that allows the user to load the data. By clicking on the "Browse" button underneath "Select a Data File", the user can select the text file that contains the significant p values for the analysis. If the default "Plot Type" setting (i.e., "Observed P Value Distribution") has not been altered and the "Plot" tab is selected (default), the user immediately obtains a histogram of the distribution of the observed significant p values (Figure 2).

Before fitting the model, the user can specify how many MCMC iterations should be used (via the "Iterations" panel) and how many of them should be cut-off as a "burn-in" (via the "Number of Burn-In Samples" panel). This number should be selected in such a way that the traceplot of the chains shows nicely intermixing chains that do not appear to move up or down systematically. Furthermore, the user can select how much the samples should be "thinned" ("Thinning" panel) where the number x corresponds to keeping each x-th sample. This can be helpful in case the samples appear to be highly autocorrelated as indicated by chains that look "sticky". The user can also change the prior standard deviation for the μ parameter of the normal distribution (corresponding to the probit-transformed p values originating from $\mathcal{H}_1$) via the "Prior SD muH1" panel; this allows one to investigate how much the results change when using a prior with a different width.

Model fitting can be started by clicking on the "Fit Model" button. Note that fitting may take a considerable amount of time depending on the number of p values to be analyzed and the number of iterations requested.

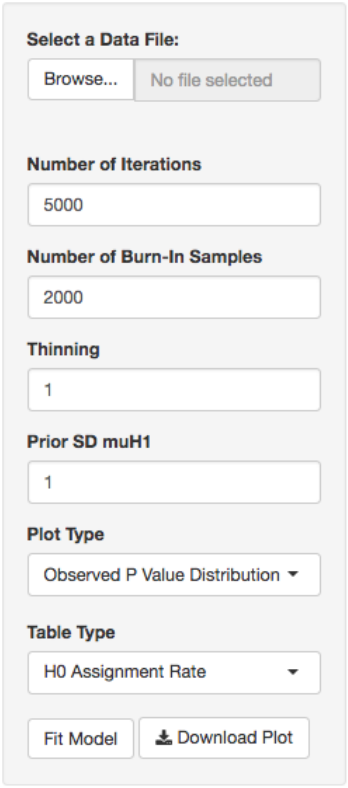

Figure 1. Input panel of the online application. Figure available at `http://tinyurl.com/z6vpfkk` under CC license `https://creativecommons.org/licenses/by/2.0/`.

Once model fitting has stopped, the user has a number of options in order to inspect the results. When the "Plot" tab is selected, the "Plot Type" panel allows the user to select between a histogram of the observed distribution of significant p values ("Observed P Value Distribution"), a traceplot of the MCMC chains for the $\mathcal{H}_0$ assignment rate parameter ("Traceplot H0 Assignment Rate"; note that this app uses three chains for fitting the model), a Q-Q plot which visualizes how well the model fits by comparing the observed p value distribution to a distribution of p values that have been generated from the model ("Q-Q Plot"), a histogram of the posterior distribution of the $\mathcal{H}_0$ assignment rate parameter ("H0 Assignment Rate"), a plot that displays for each observed p value the estimated probability that it comes from the null hypothesis ("Individual Assignment Probs."), and a five panel plot which combines these plots ("5-Panel Plot", see Figure 3). Each of these plots can be downloaded as an eps file by clicking on the "Download Plot" button.

When the "Table" tab is selected, the "Table Type" panel allows the user to choose between a table that displays the mean, median, and 95% highest density interval of the posterior distribution of the $\mathcal{H}_0$ assignment rate parameter ("H0 Assignment Rate") and a

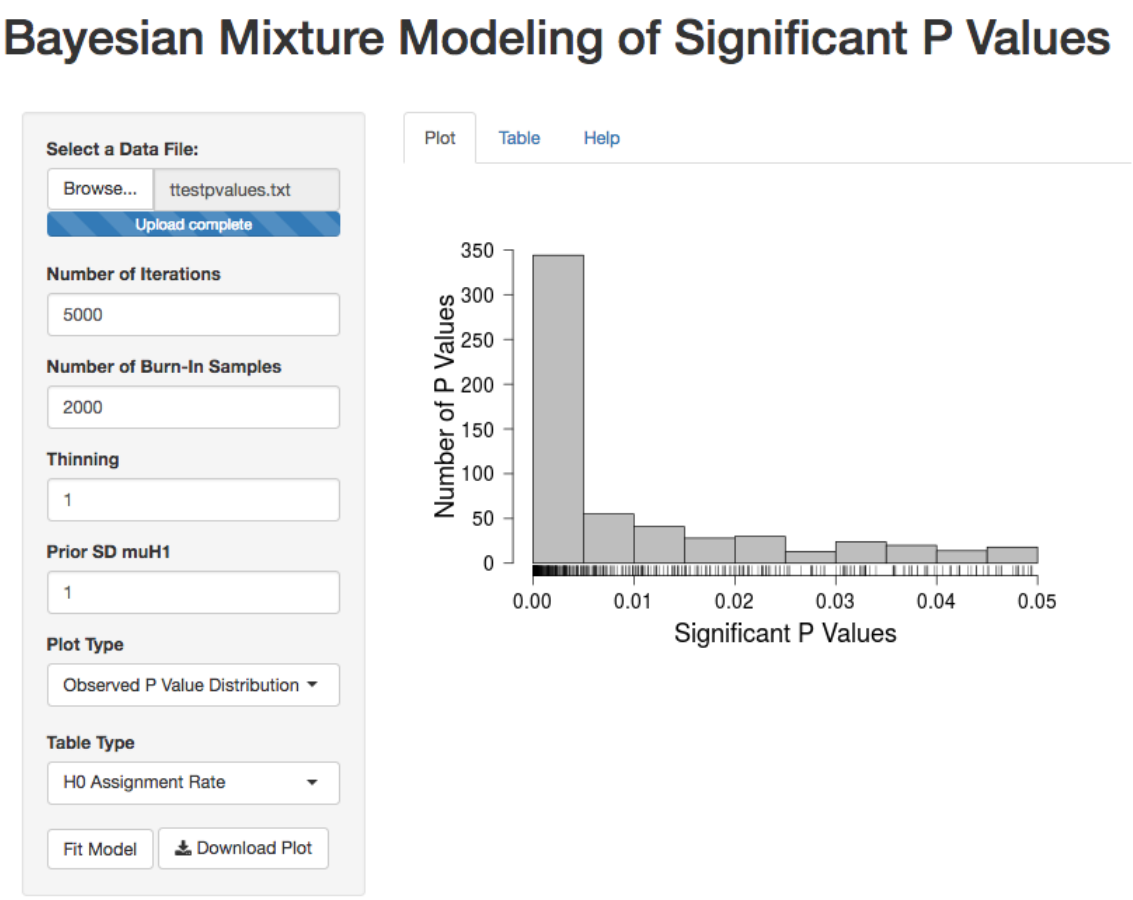

Figure 2. Plot of the observed p value distribution. Figure available at `http://tinyurl.com/h4sg26u` under CC license `https://creativecommons.org/licenses/by/2.0/`.

table that displays for each observed p value the estimated probability that it comes from the null hypothesis ("Individual Assignment Probs.", see Figure 4).

II. Simulation Studies

We conducted a number of simulation studies to investigate the effects of different effect sizes, different power of the individual studies, extreme $\mathcal{H}_0$ contamination rates ϕ, and p-hacking. For the first four simulations studies, we generated 5,000 p values from independent samples t-tests. For the fifth simulation study that investigated model performance across repeated applications, the number of p values was 200 for each replicate.

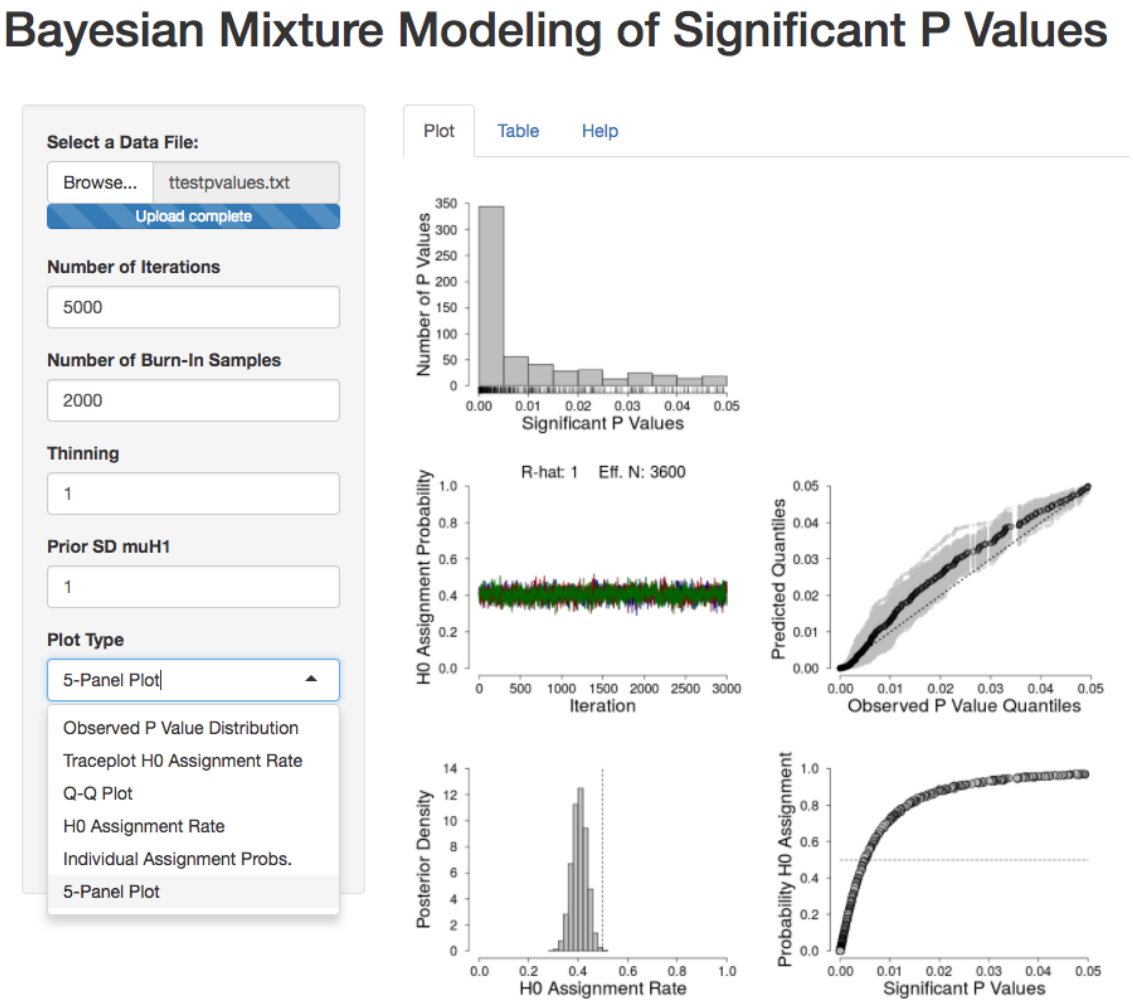

Figure 3. Five-panel plot of the results. Figure available at `http://tinyurl.com/zvw59rd` under CC license `https://creativecommons.org/licenses/by/2.0/`.

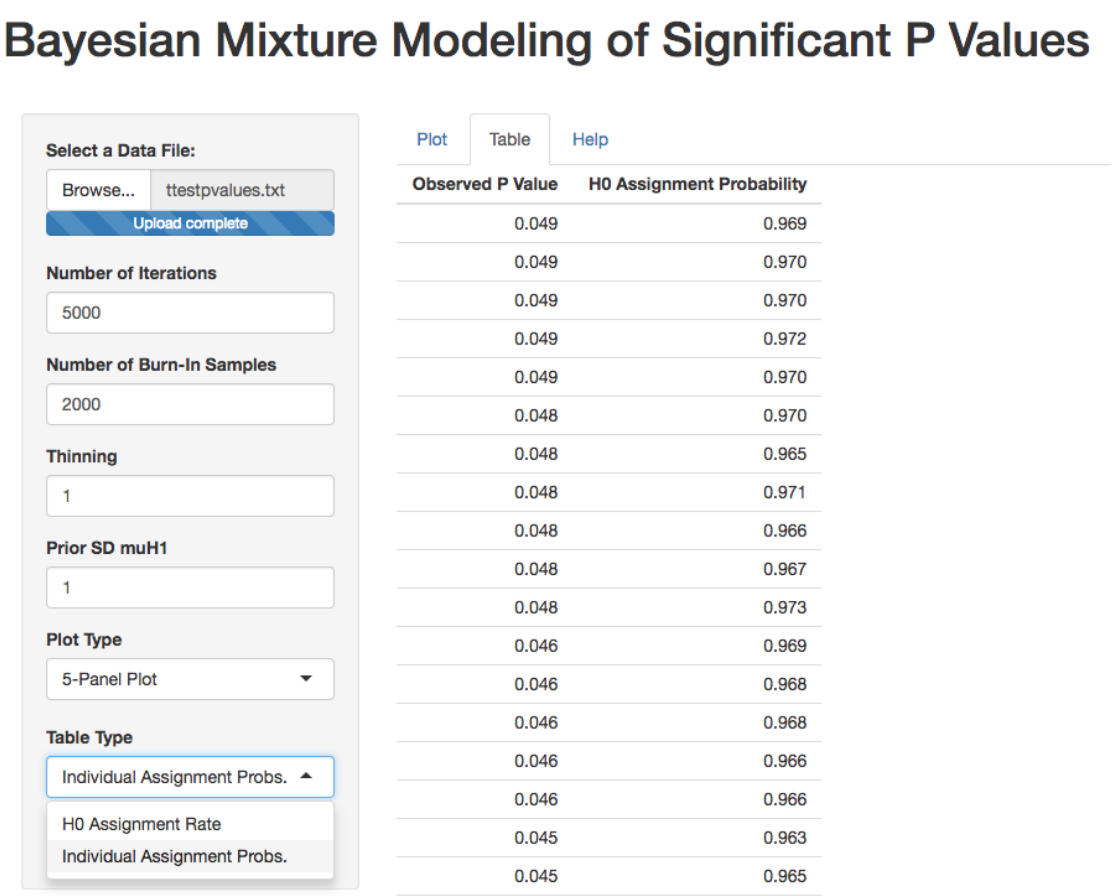

Figure 4. Table with the estimated probabilities that a specific p value stems from the null hypothesis. Figure available at `http://tinyurl.com/grwluy8` under CC license `https://creativecommons.org/licenses/by/2.0/`.

Simulation Study 1: Effect Size

The first simulation study investigated different effect sizes. We generated three sets of p values with true contamination rate $\phi = 0.50$. That means, for all three data sets, $2,500$ of the $5,000$ significant p values were generated with an effect size of zero. For the first data set, we additionally generated $2,500$ p values with effect sizes that were drawn from a truncated normal distribution which was centered on 0.15 with standard deviation 0.05 and truncation points 0.05 and 0.25, i.e., $\delta_{\mathcal{H}_1} \sim N(0.15, 0.05^2)_{T(0.05,0.25)}$. For the second data set, we additionally generated $2,500$ p values with effect sizes that were drawn from a truncated normal distribution which was centered on 0.30 with standard deviation 0.05 and truncation points 0.20 and 0.40, i.e., $\delta_{\mathcal{H}_1} \sim N(0.30, 0.05^2)_{T(0.20,0.40)}$. Finally, for the third data set, we additionally generated $2,500$ p values with effect sizes that were drawn from a truncated normal distribution which was centered on 0.45 with standard deviation

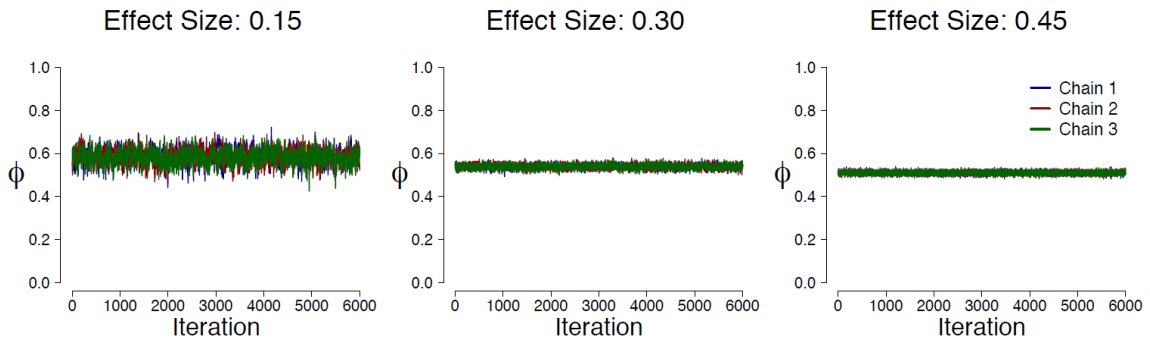

Figure 5. Simulation study 1: chains for the $\mathcal{H}_0$ assignment rate ϕ. The left panel corresponds to the data set for which the p values from $\mathcal{H}_1$ were generated with mean effect size 0.15, the middle panel to the data set for which the p values from $\mathcal{H}_1$ were generated with mean effect size 0.30, and the right panel to the data set for which the p values from $\mathcal{H}_1$ were generated with mean effect size 0.45. The true contamination rate is $\phi = 0.50$. Figure available at `http://tinyurl.com/hx3hv2y` under CC license `https://creativecommons.org/licenses/by/2.0/`.

0.05 and truncation points 0.35 and 0.55, i.e., $\delta_{\mathcal{H}_1} \sim N(0.45, 0.05^2)_{T(0.35,0.55)}$. All studies were simulated with 250 participants in each of the two t-test conditions.

We fitted the Bayesian mixture model and obtained three Markov chain Monte Carlo chains each consisting of $6,000$ posterior samples after cutting off a sufficient number of samples as burn-in for each data set (determined by visual inspection of the chains). Furthermore, for the first data set we only kept every fifth sample to decrease the amount of autocorrelation. The chains of the $\mathcal{H}_0$ assignment rate ϕ for the different effect sizes are displayed in Figure 5.

Figure 6 summarizes the results of simulation study 1. Each column corresponds to the results for one data set: the first column shows the results for the data set for which the effect size distribution under $\mathcal{H}_1$ was centered on 0.15, the second column the results for the data set for which the effect size distribution under $\mathcal{H}_1$ was centered on 0.30, and the third column the results for which the effect size distribution under $\mathcal{H}_1$ was centered on 0.45.

The first row in Figure 6 displays the posterior distributions for the $\mathcal{H}_0$ assignment rate ϕ: for effect sizes of 0.45, the model adequately recovers the true $\mathcal{H}_0$ contamination rate of 0.50. For effect size 0.30, the $\mathcal{H}_0$ assignment rate is slightly overestimated, and for effect size 0.15, this overestimation is somewhat more pronounced. However, in the latter case, the posterior distribution is also wider, indicating a higher amount of uncertainty. Nevertheless, for both effect size 0.30 and 0.15, the posterior distribution still covers the true value of the $\mathcal{H}_0$ assignment rate. Intuitively, it is expected that as the effect size becomes smaller and smaller, it will be more difficult to estimate the $\mathcal{H}_0$ assignment rate adequately, since the mixture component corresponding to $\mathcal{H}_0$ and the one corresponding

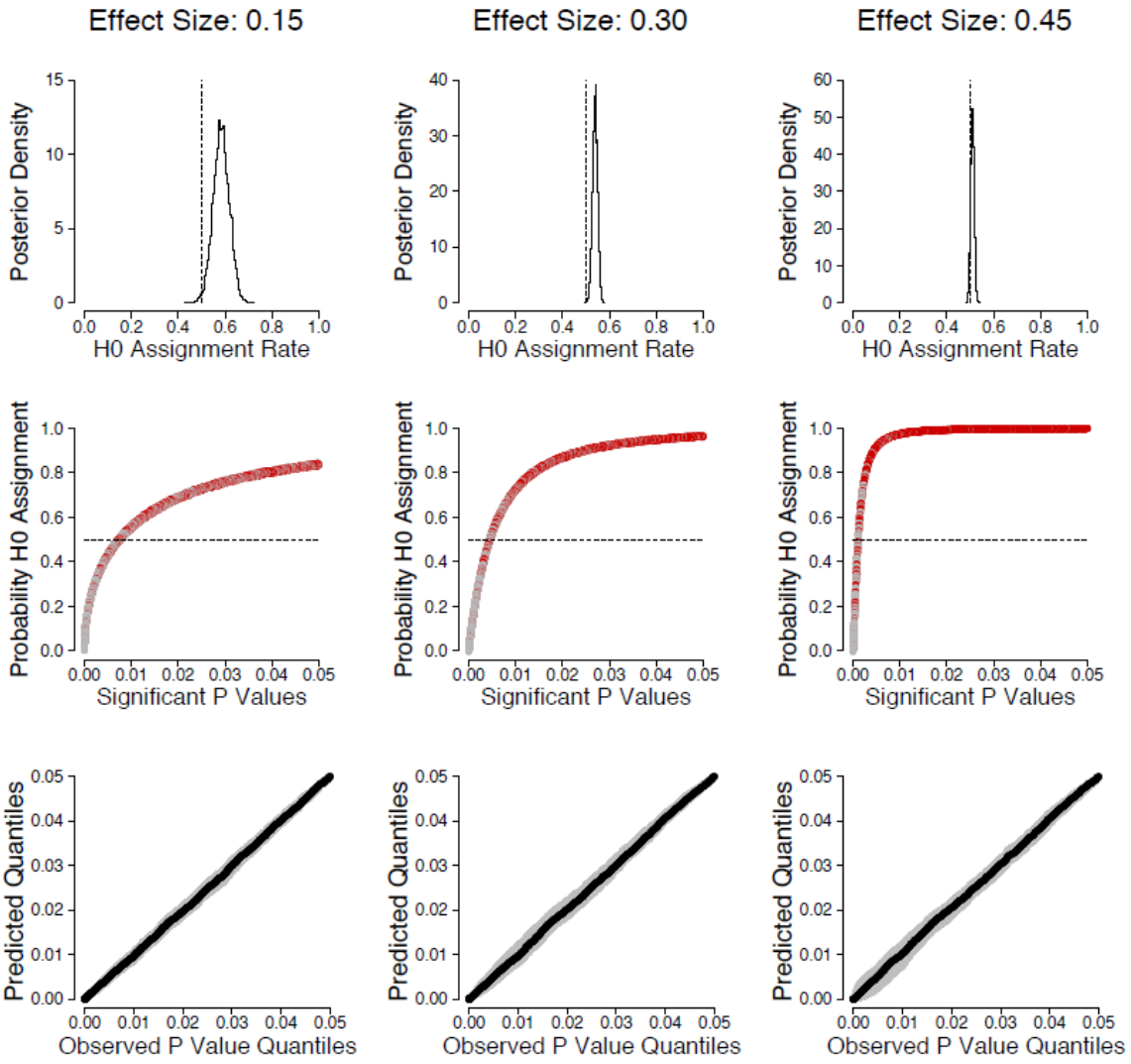

Figure 6. Simulation study 1 results. The first column corresponds to a data set for which the 2,500 p values from $\mathcal{H}_1$ were generated with mean effect size 0.15, the second column to a data set for which the 2,500 p values from $\mathcal{H}_1$ were generated with mean effect size 0.30, and the third column to a data set for which the 2,500 p values from $\mathcal{H}_1$ were generated with mean effect size 0.45. All data sets were generated with a 0.50 $\mathcal{H}_0$ contamination rate. The first row depicts the posterior distributions for the $\mathcal{H}_0$ assignment rate, the second row the individual $\mathcal{H}_0$ assignment probabilities (p values from $\mathcal{H}_0$ are colored red, those from $\mathcal{H}_1$ are colored gray), and the third row displays Q-Q plots for a comparison between observed and generated posterior predictive p values. Figure available at `http://tinyurl.com/hnvfhxx` under CC license `https://creativecommons.org/licenses/by/2.0/`.

to $\mathcal{H}_1$ become more similar and hence harder to discriminate.

The second row of Figure 6 shows the $\mathcal{H}_0$ assignment probabilities for the individual p values. P values originating from $\mathcal{H}_0$ are colored red, p values from $\mathcal{H}_1$ are colored gray. As the effect size increases, the model discriminates between p values from $\mathcal{H}_0$ and $\mathcal{H}_1$ more and more accurately.

The third row of Figure 6 displays Q-Q plots for a comparison between the distribution of observed p values and the distribution of generated posterior predictive p values which allow obtaining an impression about how well the model can account for the observed data patterns. The black dots visualize the fit that is obtained by comparing the observed p value distribution to a predicted distribution that is averaged across posterior samples. The gray dots indicate the uncertainty in the Q-Q plot by displaying the comparison of the observed p value distribution with predicted p value distributions of equal size as the observed one that are each based on one draw from the joint posterior distribution. If the distributions were exactly the same, the Q-Q plots would be straight lines with slopes equal to one. The Q-Q plots suggest that the model is able to accurately describe all three data sets.

Simulation Study 2: Power

The second simulation study investigated the effect of different power of the individual studies. We generated three sets of p values with true contamination rate $\phi = 0.50$. That means, $2,500$ of the $5,000$ significant p values were generated with an effect size of zero (with 250 participants in each t-test condition). For all three data sets, we generated $2,500$ additional p values with effect sizes that were drawn from a truncated normal distribution which was centered on 0.40 with standard deviation 0.05 and truncation points 0.30 and 0.50, i.e., $\delta_{\mathcal{H}_1} \sim N(0.40, 0.05^2)_{T(0.30,0.50)}$. The number of participants of the studies from $\mathcal{H}_1$ was chosen in such a manner that for the first data set each study had power equal to 0.33, for the second data set each study had power equal to 0.50, and for the third data set each study had power equal to 0.80.

We fitted the Bayesian mixture model and obtained three Markov chain Monte Carlo chains each consisting of $6,000$ posterior samples after cutting off a sufficient number of samples as burn-in for each data set (determined by visual inspection of the chains). For the first data set (i.e., power of 0.33) we only kept every fifth sample and for the second data set (i.e., power of 0.50) we only kept every third sample to decrease the amount of autocorrelation. The chains of the $\mathcal{H}_0$ assignment rate ϕ are displayed in Figure 7.

Figure 8 summarizes the results of simulation study 2. For the first data set (power 0.33), the contamination rate appears to be estimated relatively accurately, but there is is still quite some uncertainty as indicated by a relatively wide posterior distribution. For the second data set (power 0.50), the $\mathcal{H}_0$ contamination rate is slightly overestimated,

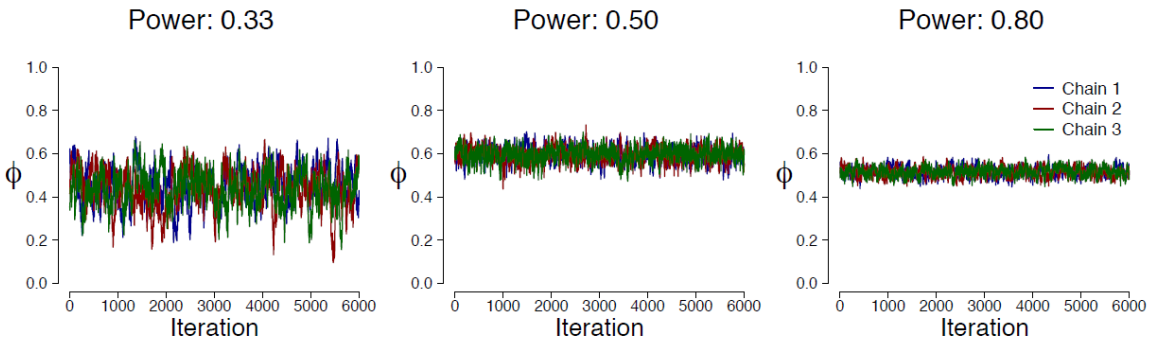

Figure 7. Simulation study 2: chains for the $\mathcal{H}_0$ assignment rate ϕ. The left panel corresponds to the data set for which the p values from $\mathcal{H}_1$ were generated from studies with power 0.33, the middle panel to the data set for which the p values from $\mathcal{H}_1$ were generated from studies with power 0.50, and the right panel to the data set for which the p values from $\mathcal{H}_1$ were generated from studies with power 0.80. The true contamination rate is $\phi = 0.50$. Figure available at `http://tinyurl.com/h234jn3` under CC license `https://creativecommons.org/licenses/by/2.0/`.

however, the posterior distribution still covers the true value of 0.50. For the third data set (power 0.80), the $\mathcal{H}_0$ contamination rate is estimated very accurately. As expected, the posterior variance decreases as power increases. This is also reflected in the $\mathcal{H}_0$ assignment probabilities for the individual p values (second row of Figure 8): as power increases, the discrimination becomes more pronounced. The Q-Q plots again suggest a good fit of the model (third row of Figure 8).

Simulation Study 3: Extreme $\mathcal{H}_0$ Contamination Rates

The third simulation study investigated the effect of extreme $\mathcal{H}_0$ contamination rates. We generated two sets of p values with true contamination rate $\phi = 0.10$ and $\phi = 0.90$. That means, for the first data set 500 and for the second data set $4,500$ of the $5,000$ significant p values were generated with an effect size of zero (with 250 participants in each t-test condition). The remaining p values were generated with effect sizes that were drawn from a truncated normal distribution which was centered on 0.40 with standard deviation 0.05 and truncation points 0.30 and 0.50, i.e., $\delta_{\mathcal{H}_1} \sim N(0.40, 0.05^2)_{T(0.30,0.50)}$ (with 250 participants in each t-test condition).

We fitted the Bayesian mixture model and obtained three Markov chain Monte Carlo chains each consisting of $6,000$ posterior samples after cutting off a sufficient number of samples as burn-in for each data set (determined by visual inspection of the chains). We only kept every third sample to decrease the amount of autocorrelation. The chains of the $\mathcal{H}_0$ assignment rate ϕ are displayed in Figure 9.

Figure 10 summarizes the results of simulation study 3. The first row of Figure 10 shows that for both data sets the $\mathcal{H}_0$ assignment rate is estimated quite accurately. The

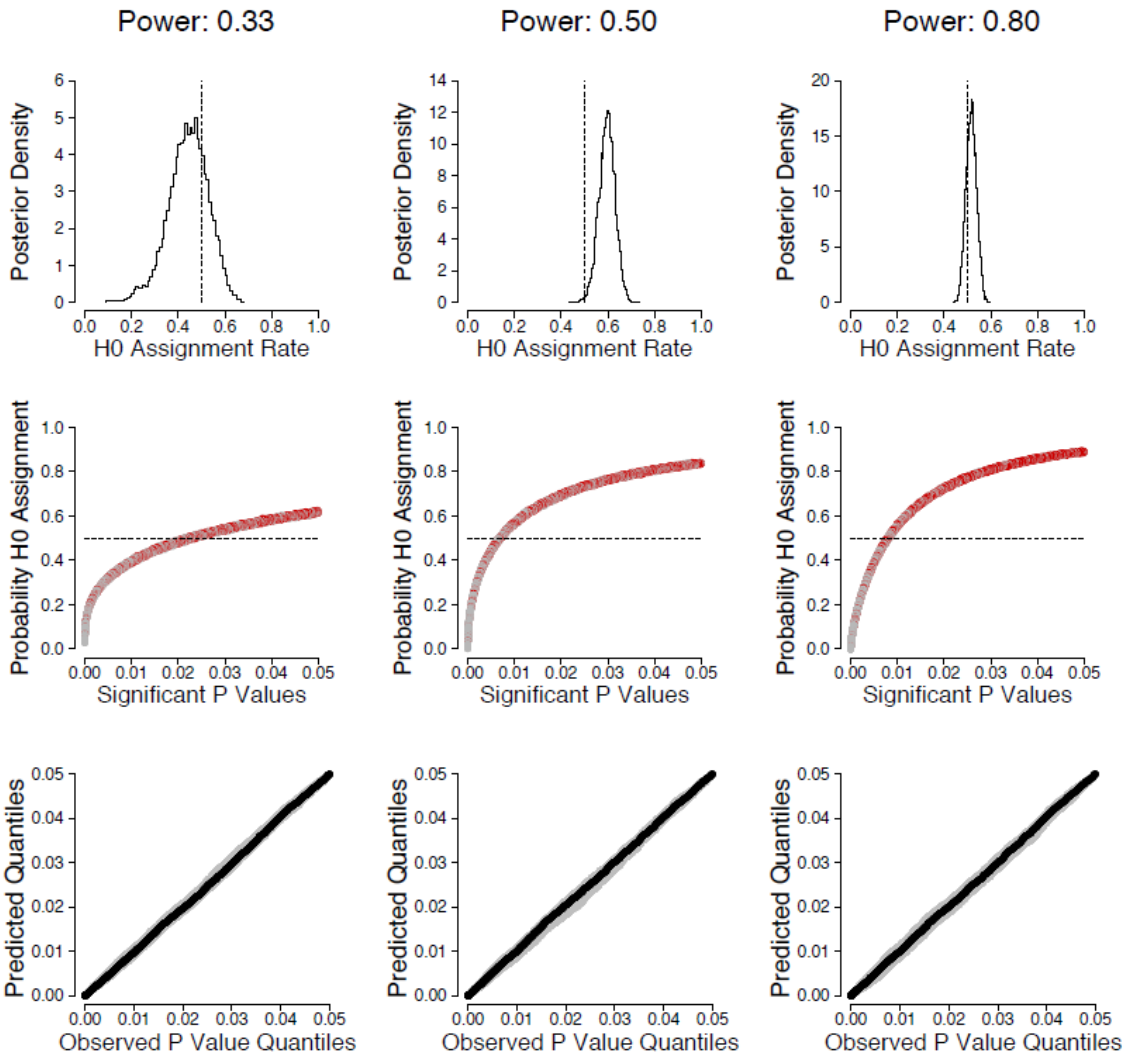

Figure 8. Simulation study 2 results. The first column corresponds to the data set for which the p values from $\mathcal{H}_1$ were generated from studies with power 0.33, the second column to the data set for which the p values from $\mathcal{H}_1$ were generated from studies with power 0.50, and the third column to the data set for which the p values from $\mathcal{H}_1$ were generated from studies with power 0.80. All data sets were generated with a 0.50 $\mathcal{H}_0$ contamination rate and a mean effect size of 0.40 for $\mathcal{H}_1$. The first row depicts the posterior distributions for the $\mathcal{H}_0$ assignment rate, the second row the individual $\mathcal{H}_0$ assignment probabilities (p values from $\mathcal{H}_0$ are colored red, those from $\mathcal{H}_1$ are colored gray), and the third row displays Q-Q plots for a comparison between observed and generated posterior predictive p values. Figure available at `http://tinyurl.com/jhljvhb` under CC license `https://creativecommons.org/licenses/by/2.0/`.

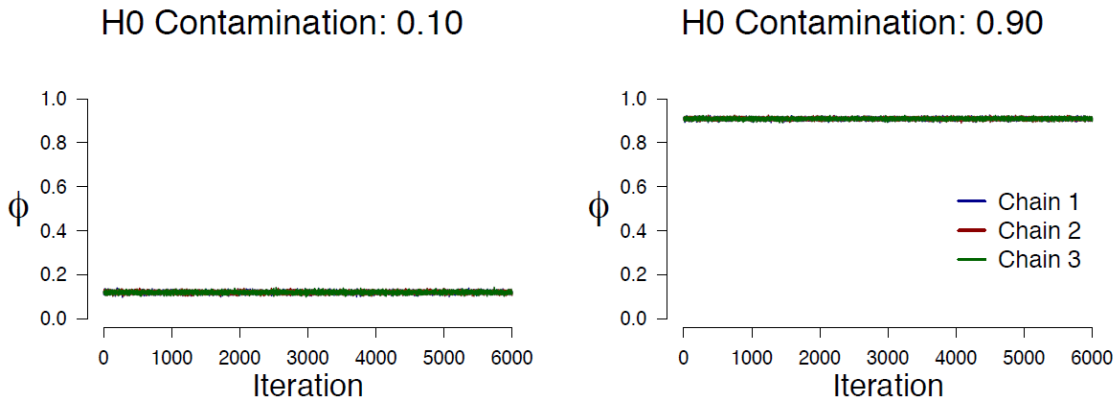

Figure 9. Simulation study 3: chains for the $\mathcal{H}_0$ assignment rate ϕ. The left panel corresponds to the data set with a 0.10 $\mathcal{H}_0$ contamination rate, the right panel to the data set with a 0.90 $\mathcal{H}_0$ contamination rate. Figure available at `http://tinyurl.com/z9b67xk` under CC license `https://creativecommons.org/licenses/by/2.0/`.

second row of Figure 10 illustrates that the discrimination between p values from $\mathcal{H}_0$ and $\mathcal{H}_1$ is good for $\phi = 0.10$ and for $\phi = 0.90$, it is excellent. Finally, the Q-Q plots suggest a good fit of the model for both data sets although the Q-Q plot for the first data set (i.e., $\phi = 0.10$) does not look as ideal as the one for the second data set (i.e., $\phi = 0.90$).

Simulation Study 4: P-Hacking

The fourth simulation study investigated the effect of p-hacking. For this simulation study, we implemented a form of p-hacking known as *data peeking* or *optional stopping*. This means that after a number of participants have been collected, the critical statistical test is conducted and if it is not significant, more participants are collected and the test is performed again without correcting for multiple testing. In theory, even when the true effect size is exactly zero, applying this method multiple times guarantees the researcher to eventually obtain a significant result (e.g., Wagenmakers, 2007). However, in practice, the researcher may not have the time and money to continue collecting participants until significance is reached.

For this simulation study we implemented p-hacking as follows: an initial 40 participants in each t-test condition were collected and the t-test was performed. If this test was not significant an additional five participants were collected in each condition and the t-test was conducted again. This procedure was repeated until a significant p value was obtained or the number of participants in each condition exceeded 250. In this way, we generated $5,000$ significant p values with a true effect size of zero (i.e., all p values were generated from $\mathcal{H}_0$). That means, the true $\mathcal{H}_0$ contamination rate was $\phi = 1$.

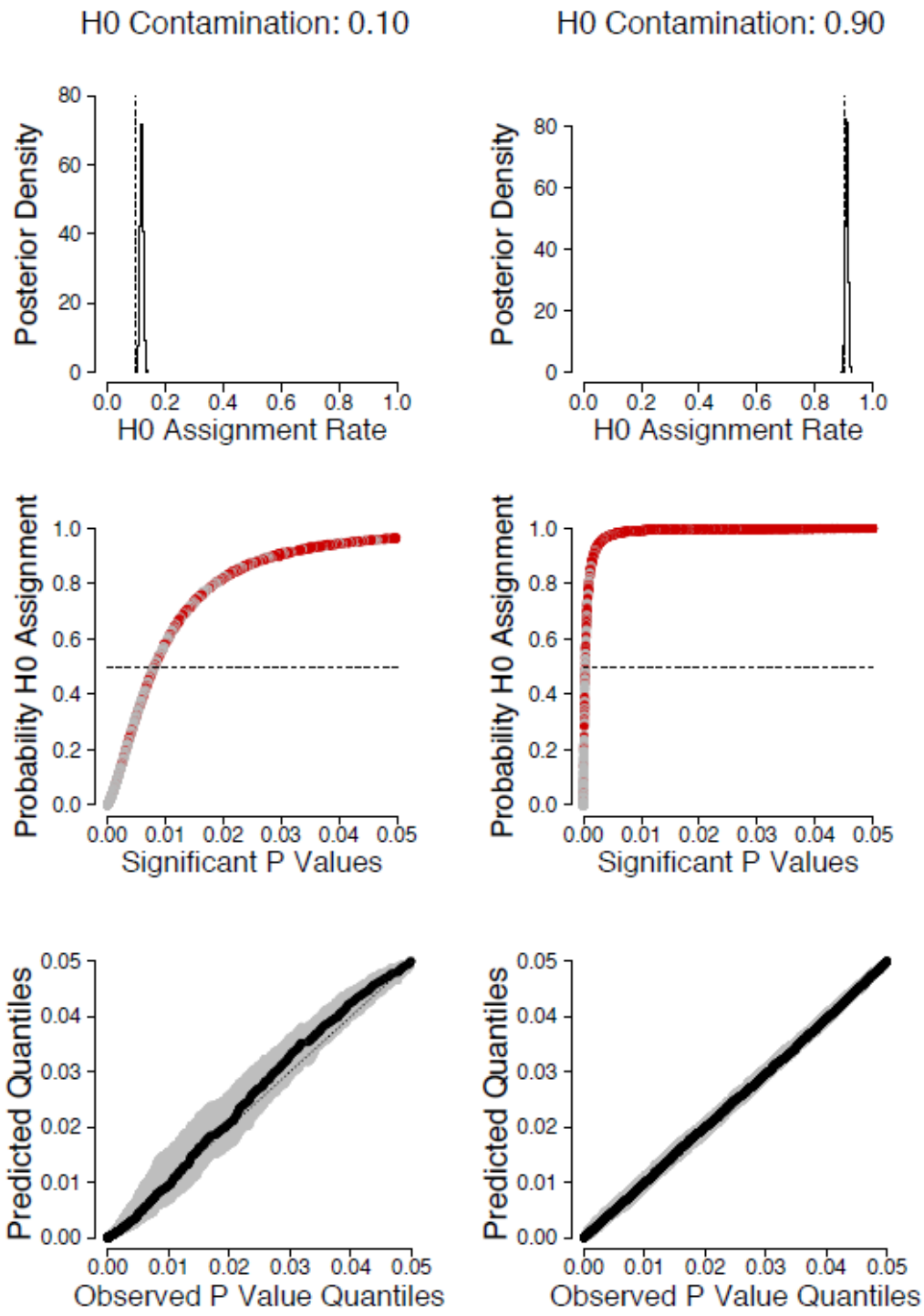

Figure 10. Simulation study 3 results. The first column corresponds to the data set with a 0.10 $\mathcal{H}_0$ contamination rate, the second column to the data set with a 0.90 $\mathcal{H}_0$ contamination rate. Both data sets were generated with a mean effect size of 0.40 for $\mathcal{H}_1$. The first row depicts the posterior distributions for the $\mathcal{H}_0$ assignment rate, the second row the individual $\mathcal{H}_0$ assignment probabilities (p values from $\mathcal{H}_0$ are colored red, those from $\mathcal{H}_1$ are colored gray), and the third row displays Q-Q plots for a comparison between observed and generated posterior predictive p values. Figure available at `http://tinyurl.com/jqybnsr` under CC license `https://creativecommons.org/licenses/by/2.0/`.

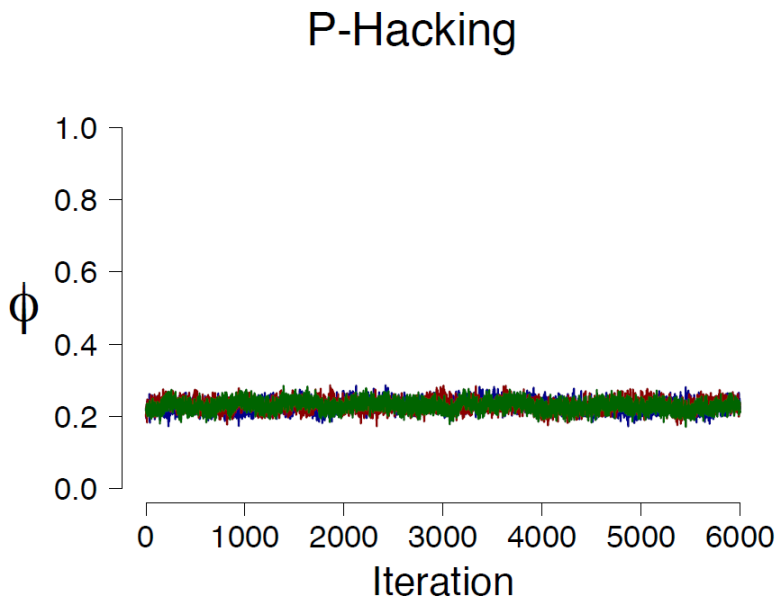

Figure 11. Simulation study 4: chains for the $\mathcal{H}_0$ assignment rate ϕ for the p-hacking simulation study. The true contamination rate is $\phi = 1$. Figure available at `http://tinyurl.com/gt5o79d` under CC license `https://creativecommons.org/licenses/by/2.0/`.

We fitted the Bayesian mixture model and obtained three Markov chain Monte Carlo chains each consisting of $6,000$ posterior samples after cutting off a sufficient number of samples as burn-in (determined by visual inspection of the chains). We only kept every fifth sample to decrease the amount of autocorrelation. The chains of the $\mathcal{H}_0$ assignment rate ϕ are displayed in Figure 11. The chains mix well, suggesting convergence to the posterior distribution.

Figure 12 summarizes the results of simulation study 4. The upper-left panel shows the distribution of significant p values which confirms that this form of p-hacking yields p value distributions that are left-skewed. It could be argued that in such a case, one should not apply a statistical test at all, because just by inspecting the p-curve it is immediately obvious that the results raise suspicion. The upper-right panel displays the posterior distribution of the contamination rate ϕ. The true contamination rate of $\phi = 1$ is dramatically underestimated as most posterior mass is close to .2. The lower-left panel shows the individual $\mathcal{H}_0$ assignment probabilities. Unexpectedly, the probability of being assigned to the null hypothesis decreases as the p values increase. When looking at the p value distribution, we observe that p values in the range $p_i \in (0, .015)$ look relatively uniformly distributed. These p values are correctly assigned to the null hypothesis with high confidence (lower-left panel of Figure 12). However, the p values larger than .15 that do not look uniformly

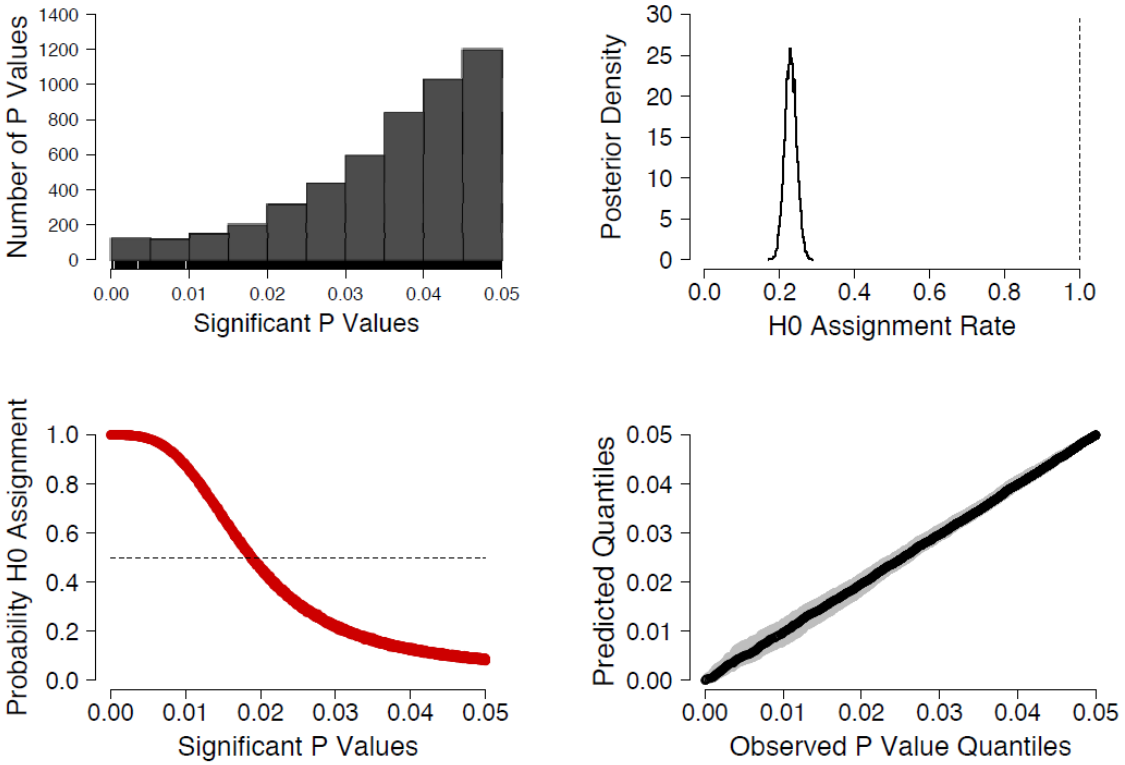

Figure 12. Simulation study 4 results. Upper-left panel: distribution of observed p values; upper-right panel: posterior distribution of the $\mathcal{H}_0$ assignment rate; lower-left panel: individual $\mathcal{H}_0$ assignment probabilities; lower-right panel: Q-Q plot for comparing the observed p value distribution to the posterior predictive distribution. The true contamination rate is $\phi = 1$. Figure available at `http://tinyurl.com/jyu4pny` under CC license `https://creativecommons.org/licenses/by/2.0/`.

distributed but show left-skew are incorrectly estimated to be more likely under $\mathcal{H}_1$ than $\mathcal{H}_0$. We recommend that researchers who obtain $\mathcal{H}_0$ assignment probabilities that are decreasing for larger p values do not interpret the model results, since this shape is highly counterintuitive and raises suspicion about whether it is appropriate to apply the model. The Q-Q plot (lower-right panel) indicates that the model predicts a p value distribution that is very similar to the observed one. In sum, in the presence of extensive p-hacking, we observe an underestimation of the $\mathcal{H}_0$ contamination rate ϕ.

Simulation Study 5: Model Performance across Repeated Applications

The fifth simulation study investigated model performance across repeated applications for sets of p values for which the p values from $\mathcal{H}_1$ were generated with power 0.33, 0.50, and 0.80. For each of these three different power values, we generated 50 sets of 200

p values with a true contamination rate of $\phi = 0.50$. Figure 13 displays the results of simulation study 5.

The top panel of Figure 13 shows the 95% highest density interval of the posterior distribution for the $\mathcal{H}_0$ contamination rate ϕ for each of the 50 simulated sets of p values for which the p values from $\mathcal{H}_1$ were generated from tests with power 0.33. Almost all of the 95% highest density intervals contain the true value of 0.50. However, these intervals are relatively wide which indicates that there is quite some uncertainty about the true value of the contamination rate ϕ. The middle panel displays the results for the 50 sets of p values for which the p values from $\mathcal{H}_1$ were generated from tests with power 0.50. Again, almost all intervals contain the true value of 0.50. Similar to the results for power 0.33, the posterior distributions are relatively wide indicating uncertainty about the true value of ϕ. The middle panel displays the results for the 50 sets of p values for which the p values from $\mathcal{H}_1$ were generated from tests with power 0.80. Once more, almost all intervals contain the true value of 0.50. However, for power 0.80, the posterior distributions are much narrower than for power 0.33 and power 0.50 indicating more certainty about the true value of ϕ.

To sum up, for the simulated data sets at hand, the Bayesian mixture model appears to do relatively well in terms of covering the true value of the contamination rate when looking at the 95% highest density interval. However, with small power, the posterior distributions are still quite wide indicating a substantial amount of uncertainty about the true value of the contamination rate ϕ.

III. Prior Sensitivity Analysis for Example 1: 587 T-Test P Values

Here we present the small prior sensitivity analysis exploring the effect of different prior choices for the μ parameter for Example 1, the 587 t-test p values, that has been mentioned in the main manuscript. For the prior sensitivity analysis, we set the prior standard deviation for the μ parameter to one half and two instead of one. Figure 14 displays the results for a prior standard deviation of one half and Figure 15 shows the results for a prior standard deviation of two. The plots highlight that for this example, the results appear to be quite robust to the prior choice for the μ parameter.

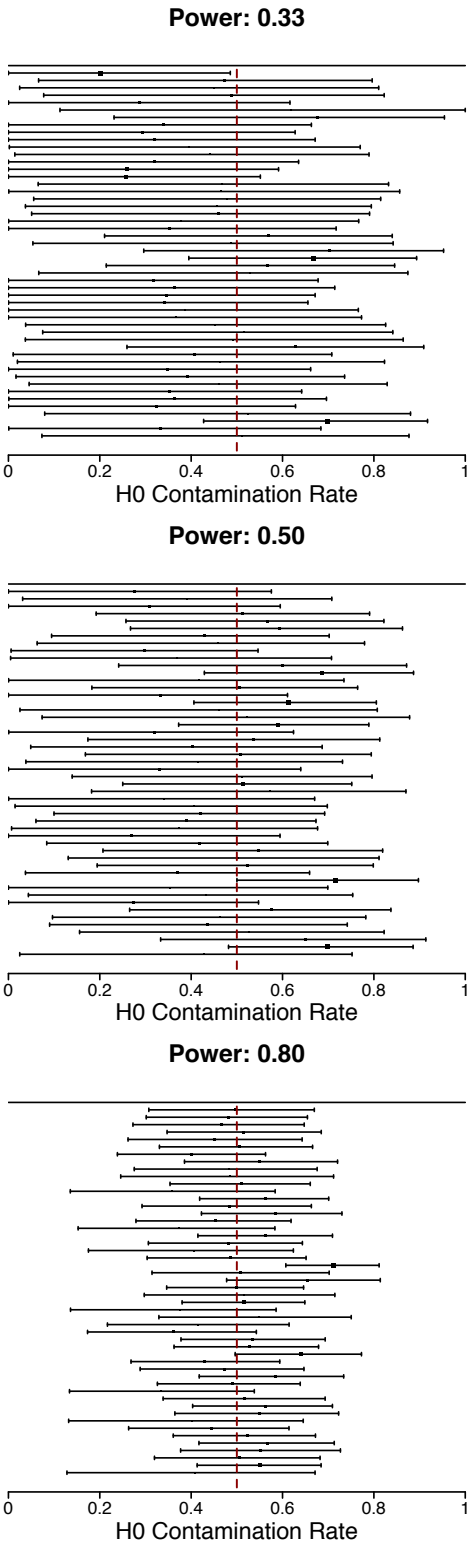

Figure 13. Simulation study 5: model performance across repeated applications. The top panel shows the 95% highest density interval of the posterior distribution for the $\mathcal{H}_0$ contamination rate ϕ for 50 simulated sets of p values of size 200 where the p values that stem from $\mathcal{H}_1$ were generated from independent samples t-tests with power 0.33. The middle panel shows the results for sets of p values for which the p values from $\mathcal{H}_1$ were generated from tests with power 0.50, the third panel the results for sets of p values for which the p values from $\mathcal{H}_1$ were generated from tests with power 0.80. The true contamination rate is $\phi = 0.50$. Figure available at `http://tinyurl.com/zty5qvp` under CC license `https://creativecommons.org/licenses/by/2.0/`.

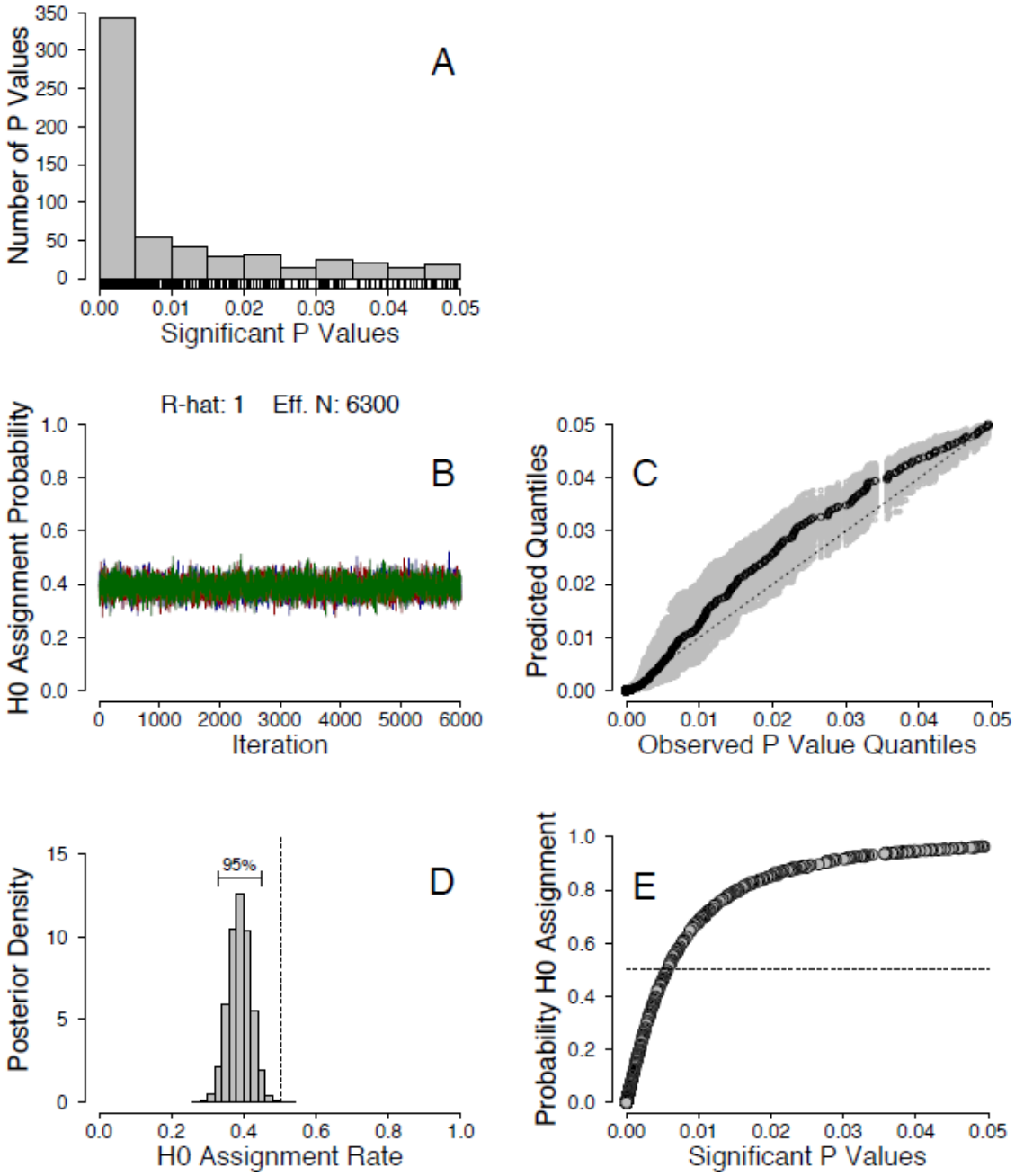

Figure 14. Application of the Bayesian mixture model to Example 1: 587 t-test p values (prior standard deviation for μ set to 0.5). Panel A: distribution of observed p values; panel B: traceplot of the MCMC chains for the $\mathcal{H}_0$ assignment rate; panel C: Q-Q plot for comparing the observed p value distribution to the posterior predictive distribution; panel D: posterior distribution of the $\mathcal{H}_0$ assignment rate; panel E: individual $\mathcal{H}_0$ assignment probabilities. Figure available at `http://tinyurl.com/zoy5eho` under CC license `https://creativecommons.org/licenses/by/2.0/`.

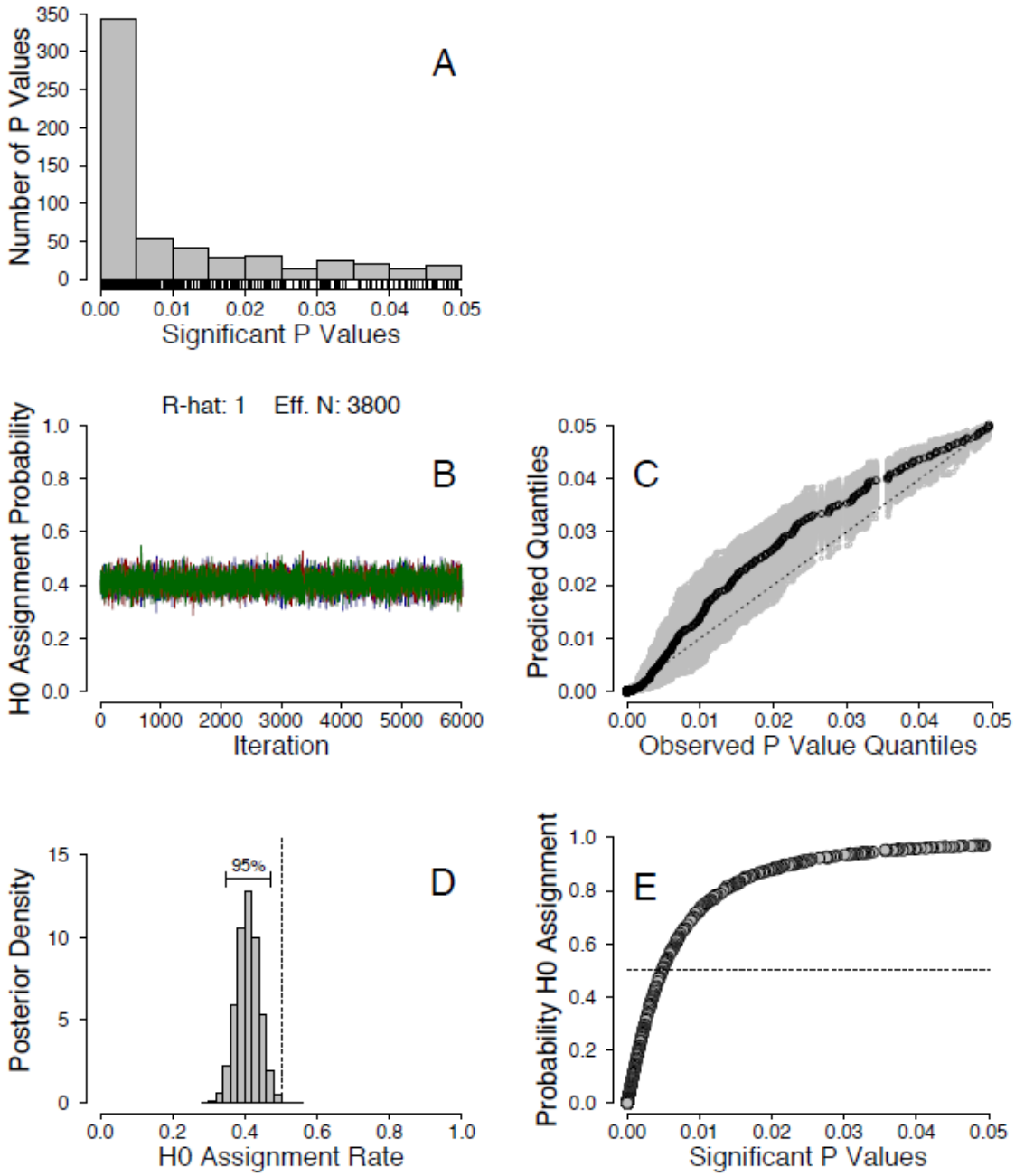

Figure 15. Application of the Bayesian mixture model to Example 1: 587 t-test p values (prior standard deviation for μ set to 2). Panel A: distribution of observed p values; panel B: traceplot of the MCMC chains for the $\mathcal{H}_0$ assignment rate; panel C: Q-Q plot for comparing the observed p value distribution to the posterior predictive distribution; panel D: posterior distribution of the $\mathcal{H}_0$ assignment rate; panel E: individual $\mathcal{H}_0$ assignment probabilities. Figure available at `http://tinyurl.com/jlssh2k` under CC license `https://creativecommons.org/licenses/by/2.0/`.

References

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*, 779–804.