The Dogmatic Prior of the Bayesian Mixture Model Inflates False Discovery Rates in Psychological Science

Ulrich Schimmack

University of Toronto Mississauga

Author Note

Ulrich Schimmack, Department of Psychology, University of Toronto Mississauga, I thank Jerry Brunner for invaluable discussions that informed the development of these ideas and for his statistical advice on Bayesian statistics. A previous version of this article has been posted as a blog post on https://replicationindex.wordpress.com/ Correspondence concerning this article should be addressed to ulrich.schimmack@utoronto.ca

Abstract

Gronau et al. (2017) developed a Bayesian Mixture Model to estimate the false discovery rate (FDR) in experimental psychology. Their results suggested that many published results are false positives. I show that their model imposed a dogmatic prior on the heterogeneity of power for true hypotheses and that the dogmatic prior leads to an inflated FDR estimate. A model without the dogmatic prior improved fit and produced a lower FDR estimate. I also show that a frequentist mixture model, z-curve, fits the data even better and produces an even lower FDR estimate. Even a z-curve with zero false positives fit the data well and better than the BMM. The main conclusion is that Gronau et al.'s article suggested that it is possible to provide precise estimates of the FDR in psychology and that the FDR is high. Ironically, this claim is itself a false discovery that is based on a model with false assumptions.

Keywords: Bayesian Mixture Model, False Discovery Rate, Z-Curve, Power, Meta-Psychology

The Dogmatic Prior of the Bayesian Mixture Model Inflates False Discovery Rates in Psychological Science

In the 1990s, the prevalent view among psychologists was that false positive results are rare (Cohen, 1994). This changed in 2011, when Bem (2011) published an article with 9 experiments that showed evidence for extrasensory perception. Rather than demonstrating a groundbreaking discovery, this article revealed that questionable research practices can provide strong evidence for false hypotheses (Francis, 2012; John et al., 2012; Schimmack, 2012). It now seemed possible that many other discoveries in psychology were false positives (Simmons et al., 2011), but the actual rate of false discoveries in psychological science remains unknown.

Gronau et al. (2017) introduced the Bayesian Mixture Model (BMM) to provide an empirical estimate of the false discovery rate in psychology. The model is explained very well in the original article, and I will only summarize the key features of the model. The data for the model are test statistics of published results. These test statistics are converted into p-values, which are then converted into z-scores using the inverse normal distribution; z = -qnorm(p).

The BMM assumes that the observed z-scores are a mixture of two normal distributions. One normal distribution represents the standard normal distribution centered over 0. These z-scores reflect false positive results, where the null-hypothesis is true and the p-value was below .05, which is assumed to be the significance criterion for all tests. The second normal distribution reflects true positives, where a true hypothesis was tested and a significant result was obtained. Non-significant results are ignored. The mean and the standard deviation of this distribution are parameter estimates of the model that reflect the average power and the variability in power, respectively. The third parameter reflects the relative proportion of true and false hypotheses. This is the parameter of interest. The proportion of false positives is known as

the false discovery rate (FDR, Soric, 1989). The prior for the FDR is a uniform distribution ranging from 0 to 1. This prior reflects a state of absolute uncertainty about the FDR in psychological research. More important, it does not impose any restrictions on the posterior distribution of the FDR. Estimates can range from 0 to 1, and the influence of the prior diminishes as more data become available. I call this an undogmatic prior. In contrast, the BMM imposes a dogmatic prior on the standard deviation of the normal distribution for true hypotheses. This parameter can only range from 0 to 1. This makes it a dogmatic prior that restricts heterogeneity in power. This prior is extremely problematic because a standard deviation of 1 is expected even if all studies have the same power. As the sampling error of z-scores is 1, sampling error alone would produce a standard deviation of 1. Heterogeneity in power would produce values greater than 1 that the dogmatic prior does not allow. Thus, the BMM essentially imposes the constraint that observed z-scores are a mixture of false positives and studies of true positives with the same power. A straightforward prediction is that the BMM with the dogmatic prior should have poor fit to data when there is heterogeneity in power, and that a model without a dogmatic prior would fit the data better.

Example 1

Gronau et al. use data from Wetzel et al. (2011) to estimate the FDR for cognitive psychology. The dataset consists of all t-tests that were published in 2007 in the journals *Psychonomic Bulletin & Review* and *Journal of Experimental Psychology: Learning, Memory, and Cognition*. The BMM 95% credibility interval for the FDR ranged from 34% to 46%. The other model parameters were not reported.

Jerry Brunner (2018) wrote an R function that provides all parameter estimates of the BMM, and demonstrated that it reproduces the same parameter estimates. I used this function to

obtain all parameter estimates and to modify the prior on the standard deviation of true hypotheses. I also wrote an R function that computes the Root Mean Square Error (RMSE) for the observed density distribution of z-scores and the density distribution predicted by the BMM (Schimmack, 2018). The function also provides figures for visual comparisons of the observed and predicted density distributions.

The parameter estimates for the false discovery rate was 45%. This reproduces the results reported by Gronau et al. The parameter estimates for the standard deviation was 1, indicating that the estimate is squished against the boundary value set by the dogmatic prior. The estimate for the mean was 4.092. This value corresponds to an estimate of pnorm(1.96,4.092) = 98% power. Thus, the BMM implicitly assumes that experimental psychologists test 45% false hypotheses and 55% true hypothesis with a fixed power of 98%. Visual inspection of the model fit shows that the model has poor fit to the data. It overestimates z-scores close to the threshold value of 1.96 (p = .05, two-tailed) and it underestimates z-scores with moderate levels of observed power (z-scores between 2.5 to 4). The RMSE value for this model is .12.

Next, I fitted the model without the dogmatic prior by increasing the range of the prior distribution to a maximum value of 5. The parameter estimates for this model were 11.5% for the FDR, a mean of 0.20, and a standard deviation of 2.84. Visual inspection of model fit shows that this model fit the data better. This is confirmed by a lower RMSE value of .05. The most important finding is that the estimate of the false discovery rate decreased from 45% to 11.5%. Even this estimate seems to high because the model without the dogmatic prior still overestimates just significant z-scores that stem from the distribution of false positives.

A Frequentist Alternative

An alternative to the Bayesian Mixture Model is Z-curve (Brunner & Schimmack, 2018). Z-curve is a frequentist mixture model. The most important difference between BMM and Z-curve is the way true positives are modeled. BMM – without the dogmatic prior - models heterogeneity in power with the standard deviation of a normal distribution. This imposes restriction on the distribution of power. In contrast, z-curve uses multiple standard normal distribution. The use of standard normal is based on the fact that sampling error alone produces a standard deviation of 1. The use of multiple standard normal makes it possible to allow for differences in power. Each standard normal represents a fixed value of power. The model can fit actual data by giving different weights to each standard normal. The model is estimated by minimizing RMSE for the discrepancy between the observed and predicted density of z-scores. Figure 3 shows the fit of z-curve to the actual data. The RMSE of this model is .032 is lower than for either one of the two BMM, indicating that neither the dogmatic prior, not the assumption of a normal distribution of power are consistent with the actual data. Z-curve does not provide a point estimate of the FDR because it is impossible to distinguish false positives from true positives with very low power (Brunner & Schimmack). However, z-curve can provide an estimate of the maximum false discovery rate, if power of true hypothesis has to meet a minimum criterion. I chose a value of z = 1 (17% power) as the minimum value for a true hypothesis. This resulted in an estimated FDR of 6.2%. I then fitted z-curve with a fixed value of 0 for the weight assigned to the false positives. Model fit was unchanged, RMSE = .033. Thus, it is impossible to reject the hypothesis that the FDR for Wetzel's t-tests is 0. If the definition of FDRs is altered to include true positive with low power, the maximum FDR is still below 10%.

Conclusion

The main conclusion is that Gronau et al.'s Bayesian Mixture Model dramatically inflates the risk of false positive results because it imposed a dogmatic prior on the variability in z-scores for true hypothesis. Fixing the standard deviation at 1 implicitly assumed that there is no heterogeneity in power. This is an implausible assumption. Not surprisingly, the model had poor fit to the data and model fit improved when the prior was changed to an undogmatic one. Thus, Gronau et al.'s model should not be used to estimate false discovery rates and the published results should be ignored. Ideally, the authors would realize their mistake and retract the article. Modifying the model is also not an option. Even the BMM without the dogmatic prior did not fit the data well and still overestimated FDR. Further modifications might be needed, but z-curve already provides an attractive alternative that produces a good fit to the data. However, the goal of this article is not to introduce an alternative method of estimating the FDR. Whether estimating the FDR is even possible or desirable is a thorny question that requires a much longer discussion than I can provide in this short commentary (cf. Soric, 1989). The main point is to demonstrate that the BMM gives the illusion that it is possible to estimate the false discovery rate in psychology and that the false discovery rate is high. Ironically, this is a false discovery that is based on a model with false assumption.

References

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. Journal of Personality and Social Psychology, 100, 407–425. doi:10.1037/a0021524

Brunner, J. (2018). R-functions for Bayesian Mixture Model. ("http://www.utstat.toronto.edu/~brunner/Rfunctions/AltGroneau.txt")

Brunner, J. & Schimmack, U. (2018). Estimating population mean power under conditions of heterogeneity and selection for significance. Under Review.

Cohen, J. (1994). The earth is round (p < .05). American Psychologist, 49(12), 997-1003. http://dx.doi.org/10.1037/0003-066X.49.12.997

Francis, G. (2012). Too good to be true: Publication bias in two prominent studies from experimental psychology. *Psychonomic Bulletin & Review, 19*, 151–156

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*, 524–532. doi:10.1177/0956797611430953

Gronau, Q. F., Duizer, M., Bakker, M., & Wagenmakers, E.-J. (2017). Bayesian mixture modeling of significant p values: A meta-analytic method to estimate the degree of contamination from H_0. *Journal of Experimental Psychology: General, 146*(9), 1223-1233.http://dx.doi.org/10.1037/xge0000324

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods, 17(4)*, 551-566. http://dx.doi.org/10.1037/a0029487

Schimmack, U. (2018). Bayesian Mixture Model Visualization. Unpublished R-Code.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22,* 1359–1366. doi:10.1177/0956797611417632

Sorić, B. (1989) Statistical “Discoveries” and Effect-Size Estimation, Journal of the American Statistical Association, 84:406, 608-610, DOI: 10.1080/01621459.1989.10478811

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science, 6,* 291–298.

Figures title:

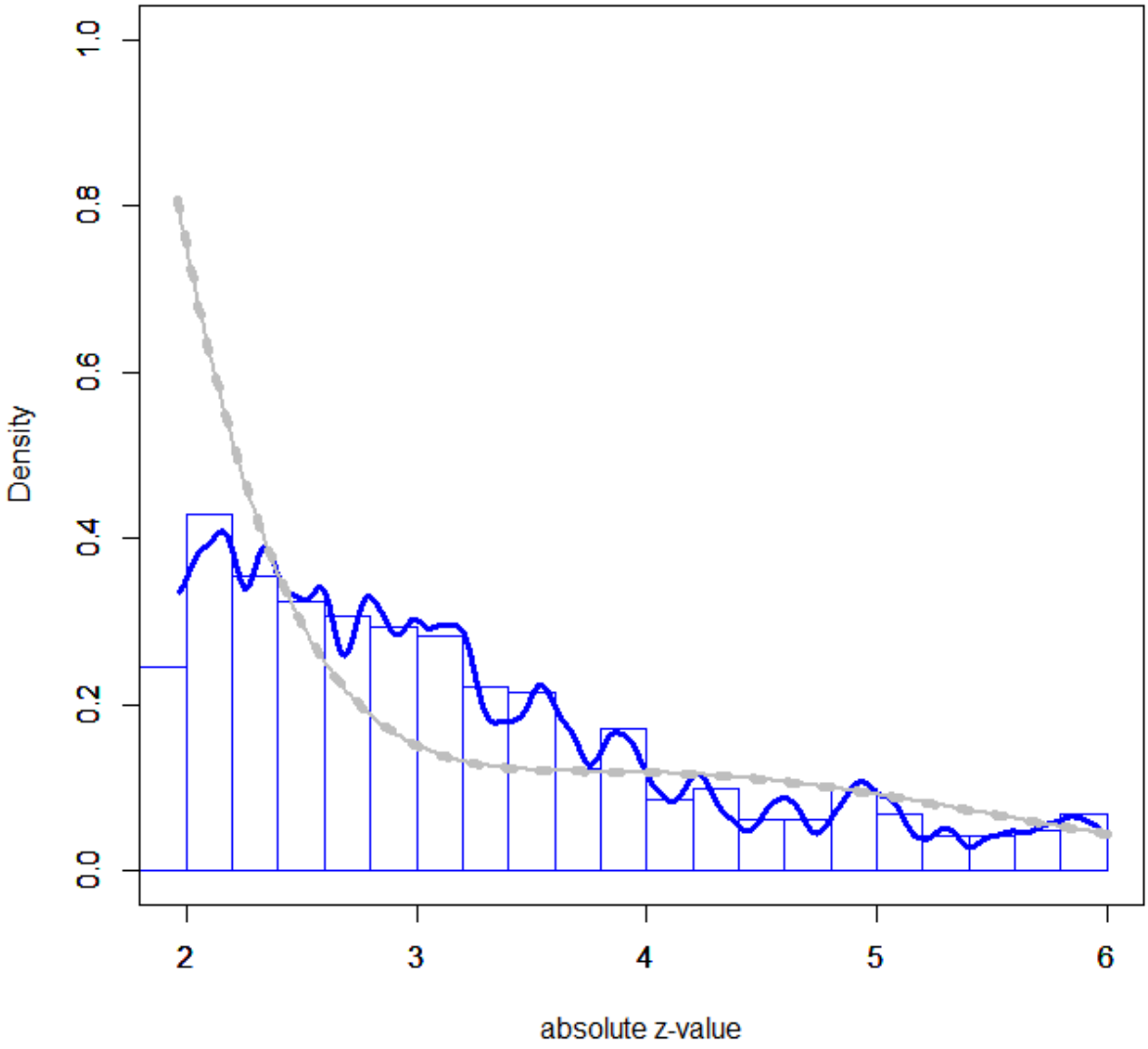

Figure 1. Fit of Bayesian Mixture Model with Dogmatic Prior (RMSE = .130)

Blue Line = Observed Density Distribution; Grey Line = Predicted Density Distribution

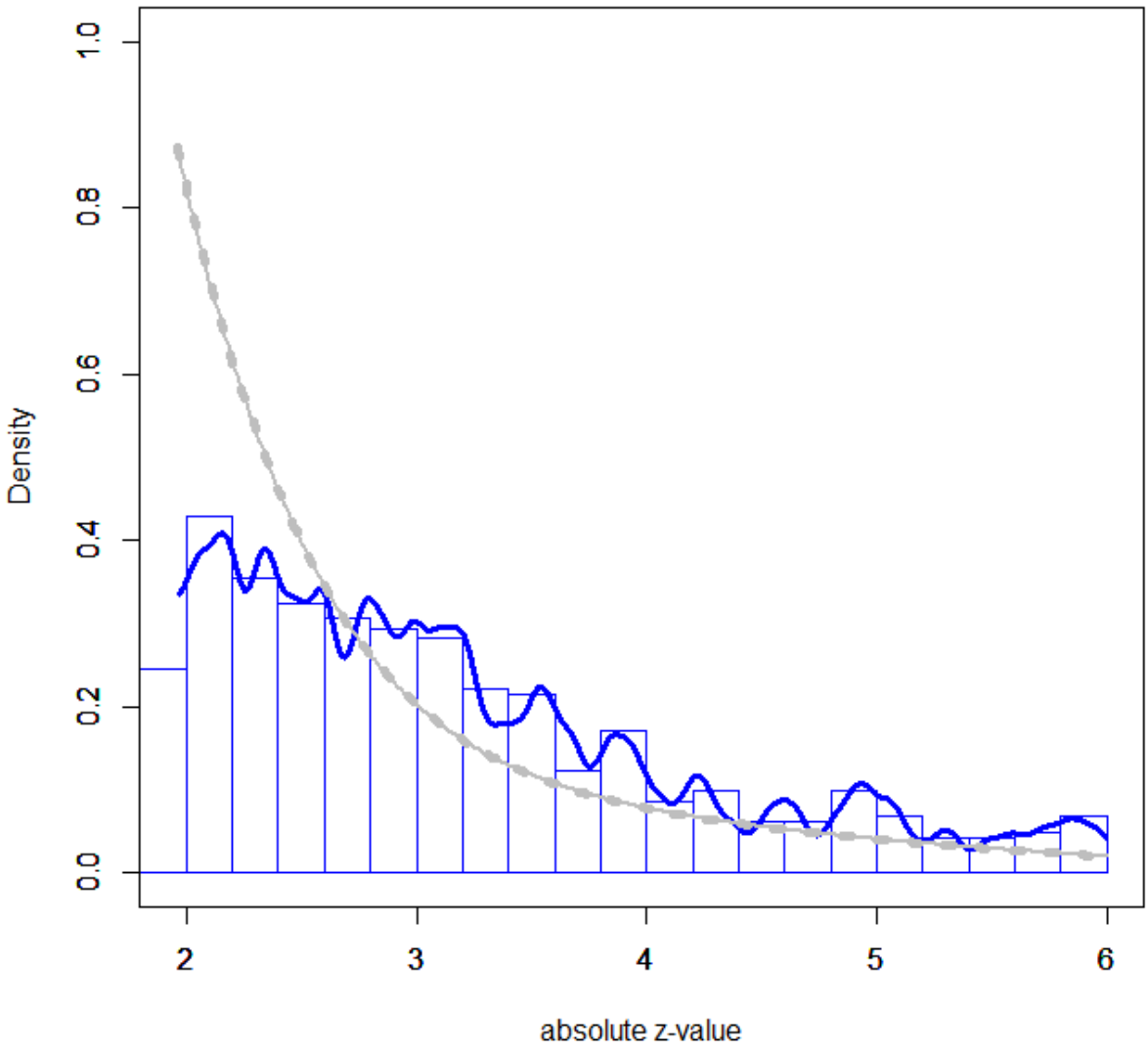

Figure 2. Fit of Bayesian Mixture Model without Dogmatic Prior (RMSE = .054)

Blue Line = Observed Density Distribution; Grey Line = Predicted Density Distribution

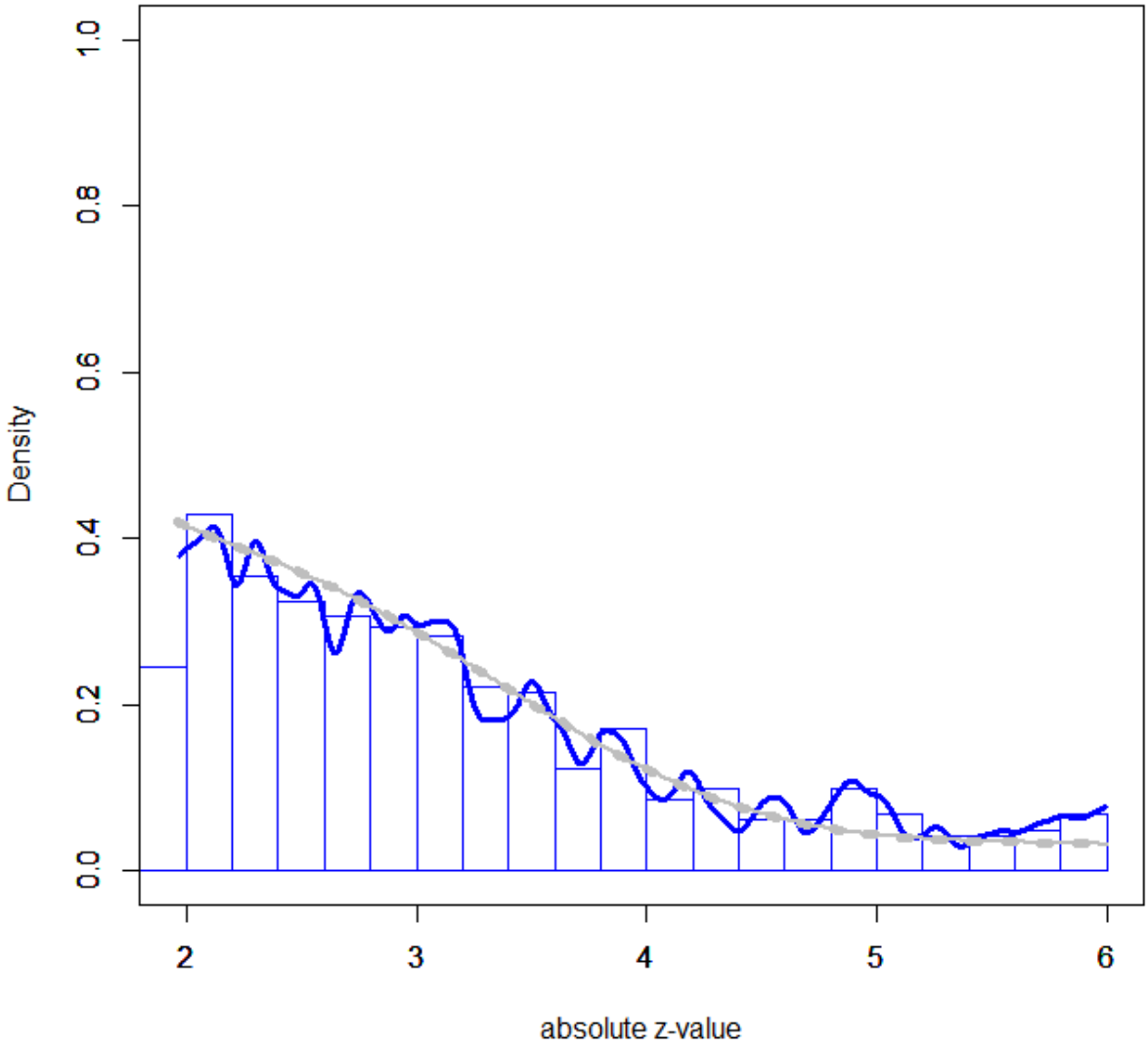

Figure 3. Fit of Z-Curve (RMSE = .033)

Blue Line = Observed Density Distribution; Grey Line = Predicted Density Distribution