Mitigating Publication Bias Using Bayesian Stacking

Thomas A. Gibson

November 30, 2021

1 Introduction

Results from a meta-analysis may be skewed and unreliable in the presence of publication bias, where the publication or non-publication of a study depends on the statistical significance or magnitude of its results (Rothstein et al., 2006). Statistical methods for publication bias have been designed for sensitivity analysis, testing for the presence/magnitude of publication bias, and calculating bias-corrected parameter estimates. Most methods are either based on the funnel plot or selection models.

Methods based on the funnel plot (Light and Pillemer, 1984) – a scatterplot of effect sizes against their standard errors – use the plot's asymmetry to test or correct for bias. The methods assume that publication bias operates primarily on smaller studies and that larger studies are much less affected. A popular non-parametric test for publication bias (Begg and Mazumdar, 1994) uses Kendall's tau to measure the rank correlation between standardized observed effect sizes and the effect sizes' standard errors. Egger's test (Egger et al., 1997) fits a linear regression of observed standard normal deviates against the observed precision of estimates, with the null hypothesis being that the regression intercept is zero. Other regression methods

(Macaskill et al., 2001; Rücker et al., 2008; Thompson and Sharp, 1999; Peters et al., 2006) are similar and use regression weights or transformations to improve upon Egger's test in the presence of heterogeneity or for dichotomous outcomes (Jin et al., 2015). Lin and Chu (2018) develops a measure for the severity of publication bias based on the skewness of standardized deviates. The trim-and-fill method (Duval and Tweedie, 2000) estimates the number of missing studies and their effect sizes using funnel plot asymmetry and gives an adjusted pooled effect estimate. The authors recommend using it as a sensitivity analysis based on the potential number of missing studies, with general guidelines given in Shi and Lin (2019). The trim-and-fill method is the only funnel plot-based method that offers an adjusted mean estimate, and it is not recommended if there is heterogeneity present (Jin et al., 2015).

A second class of methods is based on *selection models*, first described in Hedges (1984). Some models explicitly model the probability of publication for individual studies as a function of their p-values (Hedges, 1992; Givens et al., 1997; Vevea and Hedges, 1995) or as a function of both the effect size and standard error (Copas, 1999; Copas and Shi, 2000, 2001). Earlier selection models were recommended for bias-corrected effect size estimates, and were later recommended only for sensitivity analyses because of identifiability issues in smaller meta-analyses (Vevea and Woods, 2005; Jin et al., 2015). Sensitivity analyses use a grid representing varying levels of publication bias and estimate the mean effect under each assumed scenario. If results do not change much under severe publication bias they are considered robust, and if results do change under mild publication bias they are considered sensitive to publication bias. Bayesian implementations of the Copas selection model (Mavridis et al., 2013; Bai et al., 2020) have again allowed for estimation of mean effect sizes.

Recent approaches to mitigating publication bias have used Bayesian model averaging (BMA) to consider a set of potential selection functions. Guan and Vandekerckhove (2016) considers four different selection functions, including a no-bias model, an extreme-bias model where results with p-values $p > \alpha$ are never published, a 1-step function where results with $p > \alpha$ are published with some probability $\pi < 1$, and a model inspired by Givens et al. (1997) where the probability of publication decreases exponentially with p. The authors only implement the models in a fixed-effects framework. Maier et al. (2020) considers a set of 12 models, using a $2 \times 2 \times 2$ factorial design with fixed/random effects, a true null/alternative hypothesis, and the presence/absence of publication bias. The authors consider two-step and three-step selection functions based on p-values when publication bias is assumed, where the probability of publication changes at $p = 0.05$ (two-step) or at both $p = 0.05$ and $p = 0.10$ (three-step).

BMA effectively assumes that one of the considered models is the "true" model, which we call the $\mathcal{M} - \textit{closed}$ setting. BMA does not perform as well under the $\mathcal{M} - \textit{complete}$ or $\mathcal{M} - \textit{open}$ settings, where the true data generating mechanism is too complex to implement or to put into a probabilistic framework (Clyde and Iversen, 2013). Multiple issues arise for BMA in these settings, including a) the need to specify prior model probabilities, which makes little sense when we know the true model is not in our list, and b) the model weights from BMA will converge to the 1 for the model "closest" to the true model in terms of Kullback-Leibler divergence, and 0 for all others. Bayesian stacking (Yao et al., 2018, 2021) is a related method that outperforms BMA and solves the above issues by calculating optimal weights using expected log-predictive densities and leave-one-out cross validation.

We propose using Bayesian stacking to mitigate publication bias by incorporating multiple different selection models. Copas and Shi (2001) recommends a sensitivity analysis over a grid of possible patterns of publication bias, each of which returns a bias-adjusted mean effect size estimate. A stacked estimate of the mean effect size can be obtained through a weighted average of estimates with the weights coming from Bayesian stacking. Assumed patterns of publication bias that do not fit the data will be given little weight. We propose stacking over multiple types of models, including an exponential decay model (Givens et al., 1997), step functions (Hedges, 1992; Vevea and Hedges, 1995), and Bayesian Copas selection models Mavridis et al. (2013); Bai et al. (2020).

2 Methods

2.1 Selection models based on p-values

The first selection models for publication bias assumed the selection process was based on p-values. Here we consider a subset of p-value-based models that split the unit interval $[0, 1]$ into K sub-intervals, where studies with p-values that fall into different intervals have different probabilities of publication. The data model is that of a standard random effects meta-analysis: let $y_1, \ldots, y_S$ be observed effect sizes with associated standard errors s_i. We model the observed effects as

$$y_i|\theta_i \sim \mathrm{N}(\theta_i, s_i^2) \tag{2.1}$$

$$\theta_i|\theta_0, \tau^2 \sim \mathrm{N}(\theta_0, \tau^2) \tag{2.2}$$

where θ_i represent random study effects normally distributed around global mean θ_0 and with variance τ^2. The stepped weight function $w(p)$ represents the probability that a study with p-value p is observed. Dividing the unit interval into K sub-intervals with descending endpoints a_k in which the weight function $w(p)$ is constant, we have that

$$w(p_i) = \begin{cases} \omega_1 & \text{if } a_1 < p_i < 1 \\ \\ \omega_j & \text{if } a_j < p_i < a_{j-1} \\ \\ \omega_K & \text{if } 0 < p_i < a_{K-1} \end{cases} \tag{2.3}$$

where $a_0 = 1$ and $a_K = 0$. Say $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_K)$. We then have that the likelihood contribution for each observation y_i given weight function $w(\cdot)$ is

$$f(y_i|\theta_0, \tau^2, \boldsymbol{\omega}) = \frac{\text{Normal}(y_i|\theta_0, \tau^2 + s_i^2) \times w(p_i)}{\int \text{Normal}(x|\theta_0, \tau^2 + s_i^2) \times w(1 - \Phi(x/2))dx}. \tag{2.4}$$

Maier et al. (2020) place a "cumulative-Dirichlet" prior distribution on the weights $\boldsymbol{\omega}$, which effectively means placing a symmetric Dirichlet prior on an auxiliary parameter $\widetilde{\boldsymbol{\omega}} \in (0,1)^K$ and taking the cumulative sum

$$\begin{aligned} \widetilde{\boldsymbol{\omega}} &\sim \text{Dirichlet}(\text{rep}(1, K)) \\ \boldsymbol{\omega} &= \text{cumulative-sum}(\widetilde{\boldsymbol{\omega}}). \end{aligned} \tag{2.5}$$

This restricts $\boldsymbol{\omega}$ so that the K intervals in (2.3) have increasing probability of publication with decreasing p-values, and $\omega_K = 1$. Restricting $\omega_K = 1$ means each other ω_j represents the probability of publication for a study in interval j relative to the probability of publication in the lowest interval. We consider a range of possible structures for the weight function $w(p)$ by varying both the number of intervals K and the choice of cut-points a_j.

2.2 Copas selection model

Say we have S studies in the meta-analysis, indexed by $i = 1, \ldots, S$. Each study provides an estimated effect y_i, such as a log-odds ratio, and a standard deviation

s_i associated with the effect. We model observed effects as

$$y_i = \theta_i + \sigma_i \epsilon_i \tag{2.6}$$

$$z_i = \gamma_0 + \frac{\gamma_1}{s_i} + \delta_i \tag{2.7}$$

$$\theta_i \sim \mathrm{N}(\theta, \tau^2), \tag{2.8}$$

where θ_i is study i's true effect and σ_i^2 is study i's sampling variance. The latent factor z_i models the publication process, where study i is selected (published) only if $z_i > 0$. The parameter γ_0 controls the probability of publication for a study with infinite standard deviation, and γ_1 defines the relationship between the observed standard deviation s_i and the probability of publication. Usually γ_1 is assumed to be positive, so that studies with smaller standard errors are more likely to be published. The random effects (ϵ_i, δ_i) are modeled as bivariate normal

$$\begin{pmatrix} \epsilon_i \\ \delta_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right) \tag{2.9}$$

where $\mathrm{corr}(\epsilon_i, \delta_i) = \rho$ measures how the probability of selection changes with observed effect sizes. We generally expect ρ to be positive, so that studies with larger effects are more likely to be published. Copas and Shi (2000) rewrite model (2.6) - (2.9) as

$$y_i = \theta + (\tau^2 + \sigma_i^2)^{1/2} \epsilon_i^* \tag{2.10}$$

$$z_i = \gamma_0 + \frac{\gamma_1}{s_i} + \delta_i \tag{2.11}$$

$$\begin{pmatrix} \epsilon_i^* \\ \delta_i \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \tilde{\rho}_i \\ \tilde{\rho}_i & 1 \end{pmatrix} \right) \tag{2.12}$$

$$\tilde{\rho}_i = \frac{\sigma_i}{(\tau^2 + \sigma_i^2)^{1/2}} \rho, \tag{2.13}$$

which leads to a simple form for the log-likelihood

$$\begin{aligned} L(\theta, \tau^2, \rho, \gamma_0, \gamma_1) &= \sum_{i=1}^{S} \log[p(y_i|z_i > 0, s_i)] \\ &= \sum_{i=1}^{S} \log\left[\frac{p(z_i > 0|y_i, s_i) f(y_i)}{p(z_i > 0|s_i)}\right] \\ &= \sum_{i=1}^{S} -\frac{1}{2}\log(\tau^2 + \sigma_i^2) - \frac{(y_i - \theta)^2}{2(\tau^2 + \sigma_i^2)} - \log\Phi(u_i) + \log\Phi(v_i) \end{aligned} \tag{2.14}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ represent the standard normal probability density and cumulative density functions, respectively, $u_i = \gamma_0 + \frac{\gamma_1}{s_i}$, and

$$v_i = \frac{u_i + \tilde{\rho}_i \frac{y_i - \theta}{\sqrt{\tau^2 + \sigma_i^2}}}{\sqrt{(1 - \tilde{\rho}_i^2)}}. \tag{2.15}$$

For a given pair (γ_0, γ_1), one can obtain stable estimates $\hat{\theta}$, $\hat{\tau}^2$, $\hat{\rho}$ using simple maximum likelihood optimization.

We consider two Bayesian adaptations of the Copas model (Mavridis et al., 2013; Bai et al., 2020), which put prior distributions on all parameters including γ_0 and γ_1. Bai et al. (2020) places uniform priors on γ_0 and γ_1 as

$$\begin{aligned} \gamma_0 &\sim \text{Uniform}(-2, 2) \\ \gamma_1 &\sim \text{Uniform}(0, s_{max}), \end{aligned} \tag{2.16}$$

where s_{max} is the largest observed standard error in the sample of S studies. They reason that this range of values allows for selection probabilities between 2.5% and 99.7% by restricting most of the mass for latent variables z_i to the range (-2, 3). Mavridis et al. (2013) instead places priors on the lower and upper bounds for the probability of publication, P_{low} and P_{high}, as

$$\begin{aligned} P_{low} &\sim \text{Uniform}(L_1, L_2) \\ P_{high} &\sim \text{Uniform}(U_1, U_2), \end{aligned} \tag{2.17}$$

where (L_1, L_2) and (U_1, U_2) represent plausible ranges for the probability of publication for the studies with the largest and smallest standard errors, respectively. They then transform (P_{low}, P_{high}) to (γ_0, γ_1) with a 1-to-1 transformation. Priors (2.16) are meant to be default prior distributions, while (2.17) may require more problem-specific tuning, and the two priors lead to surprisingly different posterior distributions for the mean parameter θ.

2.3 Bayesian stacking

We use $\mathcal{M} - open$ to refer to the setting in which our list of candidate models does not include the true data generating mechanism (Bernardo and Smith, 2009). Bayesian stacking (Yao et al., 2018) is an alternative to BMA that has superior performance in the $\mathcal{M}$-open setting. If we have K candidate models, the goal is to find the set of optimal weights $w \in \mathcal{S}_1^K$, $\mathcal{S}_1^K = \{w \in [0,1]^K : \sum_{k=1}^K w_k = 1\}$, that maximizes a score S comparing weighted predictive distributions $p_k(\tilde{y}|y, M_k)$ to the true distribution $p_t(\tilde{y}|y)$. Yao et al. (2018) replace $p(\tilde{y}|y)$ with observed values y_i and replace $p(\tilde{y}|y, M_k)$ with its corresponding leave-one-out (LOO) predictive distribution $\hat{p}_{k,-i}(y_i) = \int p(Y_i|\theta_k, M_k)p(\theta_k|y_{-i}, M_k)d\theta_k$, where M_k is model $k, k = 1, \ldots, K$, and θ_k are the parameters in model k. The authors recommend the logarithmic scoring rule, which reduces the stacking problem to solving for weights w in

$$\max_{w \in \mathcal{S}_1^K} \frac{1}{n} \sum_{i=1}^{n} \log \sum_{k=1}^{K} w_k p(y_i|y_{-i}, M_k) \tag{2.18}$$

via optimization. Instead of refitting each model n times to obtain LOO distributions $p_k(y_i|y_{-i}M_k)$, Yao et al. (2018) use Pareto smoothed importance sampling (PSIS) (Vehtari et al., 2017) to obtain approximations.

After stacking weights w have been optimized, the stacked posterior distribution of θ can be obtained by taking $w_k \times T$ samples from each model k's posterior distribution $p(\theta|y)$ for a total of T posterior samples.

Implementation of Bayesian stacking can be done using the R package 'loo'.

2.4 Stacking selection models for publication bias

It would be naive to think that either the stepped selection functions in Section 2.1 or the Copas models in Section 2.2 represent the true data generating mechanism for publication bias. As Bayesian stacking is designed to perform well in the event that our list of models does not contain the true model, we propose stacking over a variety of stepped selection functions and Copas models to obtain a more robust posterior distribution for the mean parameter θ.

We propose implementing both Bayesian Copas selection models (Mavridis et al., 2013; Bai et al., 2020) and a set of stepped selection models with varying numbers of steps. As a default, we use a 1-step function with the step at $p = 0.05$, a two-step function with steps at $p = (0.05, 0.10)$, and a three-step function with steps $p = (0.05, 0.10, 0.20)$. We fit the Copas models in JAGS (Plummer et al., 2003) as there is a straightforward Gibbs sampling routine, and fit the stepped selection models in Stan (Gelman et al., 2015) because of the ability to easily code the custom probability distribution (2.4).

References

Bai, R., Lin, L., Boland, M. R., and Chen, Y. (2020). A robust bayesian copas selection model for quantifying and correcting publication bias. *arXiv preprint arXiv:2005.02930* .

Begg, C. B. and Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics* pages 1088–1101.

Bernardo, J. M. and Smith, A. F. (2009). *Bayesian theory*, volume 405. John Wiley & Sons.

Clyde, M. and Iversen, E. S. (2013). Bayesian model averaging in the m-open framework. *Bayesian theory and applications* **14,** 483–498.

Copas, J. (1999). What works?: selectivity models and meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **162,** 95–109.

Copas, J. and Shi, J. Q. (2000). Meta-analysis, funnel plots and sensitivity analysis. *Biostatistics* **1,** 247–262.

Copas, J. and Shi, J. Q. (2001). A sensitivity analysis for publication bias in systematic reviews. *Statistical methods in medical research* **10,** 251–265.

Duval, S. and Tweedie, R. (2000). Trim and fill: a simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics* **56,** 455–463.

Egger, M., Smith, G. D., Schneider, M., and Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *Bmj* **315,** 629–634.

Gelman, A., Lee, D., and Guo, J. (2015). Stan: A probabilistic programming language for bayesian inference and optimization. *Journal of Educational and Behavioral Statistics* **40,** 530–543.

Givens, G. H., Smith, D., and Tweedie, R. (1997). Publication bias in meta-analysis: a bayesian data-augmentation approach to account for issues exemplified in the passive smoking debate. *Statistical Science* **12,** 221–250.

Guan, M. and Vandekerckhove, J. (2016). A bayesian approach to mitigation of publication bias. *Psychonomic bulletin & review* **23,** 74–86.

Hedges, L. V. (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics* **9,** 61–85.

Hedges, L. V. (1992). Modeling publication selection effects in meta-analysis. *Statistical Science* **7,** 246–255.

Jin, Z.-C., Zhou, X.-H., and He, J. (2015). Statistical methods for dealing with publication bias in meta-analysis. *Statistics in medicine* **34,** 343–360.

Light, R. J. and Pillemer, D. B. (1984). Summing up: the science of reviewing research.

Lin, L. and Chu, H. (2018). Quantifying publication bias in meta-analysis. *Biometrics* **74,** 785–794.

Macaskill, P., Walter, S. D., and Irwig, L. (2001). A comparison of methods to detect publication bias in meta-analysis. *Statistics in medicine* **20,** 641–654.

Maier, M., Bartoš, F., and Wagenmakers, E.-J. (2020). Robust Bayesian meta-analysis: Addressing publication bias with model-averaging.

Mavridis, D., Sutton, A., Cipriani, A., and Salanti, G. (2013). A fully bayesian application of the copas selection model for publication bias extended to network meta-analysis. *Statistics in medicine* **32,** 51–66.

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., and Rushton, L. (2006). Comparison of two methods to detect publication bias in meta-analysis. *Jama* **295,** 676–680.

Plummer, M. et al. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*, volume 124, pages 1–10. Vienna, Austria.

Rothstein, H. R., Sutton, A. J., and Borenstein, M. (2006). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. John Wiley & Sons.

Rücker, G., Schwarzer, G., and Carpenter, J. (2008). Arcsine test for publication bias in meta-analyses with binary outcomes. *Statistics in Medicine* **27,** 746–763.

Shi, L. and Lin, L. (2019). The trim-and-fill method for publication bias: practical guidelines and recommendations based on a large database of meta-analyses. *Medicine* **98,**.

Thompson, S. G. and Sharp, S. J. (1999). Explaining heterogeneity in meta-analysis: a comparison of methods. *Statistics in medicine* **18,** 2693–2708.

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing* **27,** 1413–1432.

Vevea, J. L. and Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika* **60,** 419–435.

Vevea, J. L. and Woods, C. M. (2005). Publication bias in research synthesis: sensitivity analysis using a priori weight functions. *Psychological methods* **10,** 428.

Yao, Y., Pirš, G., Vehtari, A., and Gelman, A. (2021). Bayesian hierarchical stacking. *arXiv preprint arXiv:2101.08954* .

Yao, Y., Vehtari, A., Simpson, D., and Gelman, A. (2018). Using stacking to average bayesian predictive distributions (with discussion). *Bayesian Analysis* **13,** 917–1007.