Mitigating Publication Bias Using Bayesian Stacking

Thomas A. Gibson

December 23, 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 – inspect the plot's asymmetry to test or correct for bias. Say we have S studies indexed by $i = 1, \ldots, S$ and that y_i and $v_i = s_i^2$ are their estimated effect sizes and sampling variances. A popular non-parametric test for publication bias (Begg and Mazumdar, 1994) measures the rank correlation between standardized observed effect sizes y_i^* and the effect sizes' variances v_i, where

$$y_i^* = (y_i - \overline{y})/(v_i^*)^{1/2}$$
$$\overline{y} = \Big(\sum_j (v_i^{-1})/y_i\Big)/\Big(\sum_j v_i^{-1}\Big)$$
$$v_i^* = v_i - \Big(\sum_j v_i^{-1}\Big)^{-1}.$$

Begg and Mazumdar (1994) measure the correlation between pairs (y_i^*, v_i) with Kendall's tau, where a symmetric funnel plot would have correlation near zero. Egger's test (Egger et al., 1997) fits a linear regression of observed standard normal deviates $\text{SND}_i = y_i/s_i$ against the inverse standard errors $1/s_i$, i.e. $\text{SND}_i = \alpha + \beta \times (1/s_i)$ with the null hypothesis $H_0 : \alpha = 0$. Other regression methods (Macaskill et al., 2001; Rücker et al., 2008; Thompson and Sharp, 1999; Peters et al., 2006) are similar to Egger's test 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) calculates an adjusted mean effect estimate in a series of steps, by 1) estimating the number of missing studies k_0, 2) "trimming" the smaller studies that are causing funnel plot asymmetry, 3) estimating the true mean effect with the remaining studies, and 4) replacing trimmed studies and their missing counterparts and re-estimating the mean effect and its variance. However, Duval and Tweedie (2000) recommend using trim-and-fill as a sensitivity analysis based on the *potential* number of missing studies, with general guidelines for sensitivity analysis 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 in study effects (Jin et al., 2015).

A second class of methods is based on *selection models*, first described in Hedges

(1984). Let Y be a random variable representing effect sizes for all studies in a population and let Θ be the parameters determining the sampling density $f(y;\Theta)$. Selection models assume a biased sampling scheme where the probability of a study being observed (published) is represented by a weight function $w(y;\lambda)$, where λ, a scalar or vector parameter, determines how certain studies may be more or less likely to be published. The weighted density for observed effect y_i is then

$$f^*(y_i;\Theta,\lambda) = \frac{f(y_i;\Theta)w(y_i;\lambda)}{\int f(y;\Theta)w(y;\lambda)dy} \tag{1.1}$$

and the likelihood function for all observed studies is

$$L(\Theta,\lambda) = \prod_{i=1}^{S} f^*(y_i;\Theta,\lambda). \tag{1.2}$$

Some models explicitly model the probability of publication for individual studies as a function of their p-values (Iyengar and Greenhouse, 1988; 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). Hedges (1992) introduced stepped weight functions by dividing the unit interval [0, 1] into K segments with $K-1$ cut points, where studies that have p-values in different segments have different probabilities of publication. We refer to stepped selection functions by the number of cut points, i.e. a 1-step selection function might have a single cut point at $p = 0.05$, or a 2-step function might have cuts at $p = 0.05, 0.10$. 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 an assumption of severe publication bias they are robust, and if results do change under an assumption of mild publication bias they are sensitive. 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 averaged meta-analysis (BMA-MA) to consider a set of potential selection functions. Guan and Vandekerckhove (2016) considers four different selection functions based on p-values, including a no-bias model, an extreme-bias model where studies with p-values $p > \alpha$ are never published, a 1-step function where studies with $p > \alpha$ are published with some probability $\pi < 1$, and a model where the probability of publication decreases exponentially with p. Guan and Vandekerckhove (2016) only implement the models in a fixed-effects framework. Maier et al. (2020) evaluates 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. Maier et al. (2020) fit two-step and three-step selection functions based on p-values when publication bias is assumed, where the probability of publication changes at $p = 0.05$ (one-step) or at both $p = 0.05$ and $p = 0.10$ (two-step).

Bayesian model averaging (BMA) effectively assumes that one of the considered models is the "true" model, which is called the $\mathcal{M}$-closed setting (Bernardo and Smith, 2009). BMA does not perform as well under the $\mathcal{M}$-complete or $\mathcal{M}$-open settings, where the true data generating mechanism is too complex to implement or to put into a probabilistic framework (Bernardo and Smith, 2009; Le and Clarke, 2017). 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 1 for the model "closest" to the true model in terms of Kullback-Leibler divergence, and 0 for all others (Clyde and Iversen, 2013). Bayesian stacking of predictive distributions (Yao et al., 2018, 2021) is a method that outperforms Bayesian model averaging in the $\mathcal{M}$-complete and $\mathcal{M}$-open settings and avoids issues (a) and (b) above. If we have K candidate models $M_1, \ldots, M_K$, Yao et al. (2018) solve for model weights $w = (w_1, \ldots, w_K)$ under the constraint $w \in \mathcal{S}_1^K$ where $\mathcal{S}_1^K = \{w \in [0,1]^K : \sum_{k=1}^K w_k = 1\}$. They do this by solving

$$(\hat{w}_1, \ldots, \hat{w}_K) = \max_{w \in \mathcal{S}_1^K} \frac{1}{S} \sum_{i=1}^{n} \log \sum_{k=1}^{K} w_k p(y_i | y_{-i}, M_k) \tag{1.3}$$

where $p(y_i|y_{-i}, M_k)$ is the leave-one-out (LOO) posterior predictive density with the ith data point left out evaluated at y_i. It would be computationally costly to refit each model M_k S times, so LOO densities $p(y_i|y_{-i}, M_k)$ are approximated using Pareto-smoothed importance sampling (Vehtari et al., 2017). We go through Bayesian stacking more thoroughly in Section 2.3.

Given that the true data generating mechanism for publication bias is likely much too complex to be specified in a simple selection model, we propose using Bayesian stacking to mitigate publication bias by fitting multiple Bayesian selection models and stacking over them. Assumed patterns of publication bias that poorly predict the observed data with LOO cross validation will be given little weight. We propose stacking over multiple types of models, including step functions (Vevea and Hedges, 1995) and Bayesian Copas selection models Mavridis et al. (2013); Bai et al. (2020). Section 2 describes relevant selection models for publication bias, Bayesian stacking, and how to implement Bayesian stacking of selection models. We then describe and summarize a simulation study in Section 3. The purpose of the simulation is to compare a stacked estimate of the mean effect size to estimates from individual selection models when the true data generator is not one of the fitted selection models. We apply the stacked model to a real dataset on the effectiveness of antidepressants in Section 4

2 Methods

We are doing a meta-analysis and have S studies indexed by $i = 1, \ldots, S$, and each study provides an estimated effect y_i and an associated standard error s_i. Assuming estimates y_i are normally distributed, we calculate study i's 2-sided p-value as $p_i = 2 \times \big(1 - \Phi(|y_i|/s_i)\big)$ where $\Phi(\cdot)$ is the standard normal cumulative distribution function. The data model for each selection method is

$$y_i = \theta_i + s_i\epsilon_i \tag{2.1}$$

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

where θ_i represent random study effects normally distributed around global mean θ with variance τ^2, and $\epsilon_i \sim \mathrm{N}(0, 1)$. For stepped weight functions we combine s_i^2 and τ^2 into a single residual so that

$$y_i|\theta, \tau^2 \sim \mathrm{N}(\theta, s_i^2 + \tau^2). \tag{2.3}$$

2.1 Selection models based on p-values

We define a stepped weight function $w(\cdot)$ similar to Vevea and Hedges (1995) and Vevea and Woods (2005), where $w(p)$ 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

$$w(p_i) = \begin{cases} \omega_1 & \text{if } a_1 < p_i < 1 \\ \omega_j & \text{if } a_j < p_i < a_{j-1} \\ \omega_K & \text{if } 0 < p_i < a_{K-1} \end{cases} \tag{2.4}$$

where $a_0 = 1$ and $a_K = 0$. Say $\boldsymbol{\omega} = (\omega_1, \dots, \omega_K)$. We then have that the likelihood contribution for each observation y_i given θ, τ^2 and weight function $w(\cdot)$ is

$$f(y_i|\theta, \tau^2, \boldsymbol{\omega}) = \frac{\phi(y_i; \theta, \tau^2 + s_i^2) \times w(p_i)}{\int \phi(x; \theta, \tau^2 + s_i^2) \times w(1 - \Phi(x/2))dx}, \tag{2.5}$$

where $\phi(x; a, b)$ represents a normal probability density function with mean a and variance b. 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

$$\begin{aligned} \widetilde{\boldsymbol{\omega}} &\sim \text{Dirichlet}(\text{rep}(1, K)) \\ \boldsymbol{\omega} &= \text{cumulative-sum}(\widetilde{\boldsymbol{\omega}}). \end{aligned} \tag{2.6}$$

This restricts $\boldsymbol{\omega}$ so that the K intervals in (2.4) have increasing probability of publication with decreasing p-values, and $\omega_K = 1$. The symmetric Dirichlet prior on $\widetilde{\boldsymbol{\omega}}$ leads to prior means $(\frac{1}{K}, \frac{2}{K}, \dots, 1)$ for $\boldsymbol{\omega}$. 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

The Copas selection model (Copas, 1999; Copas and Shi, 2000, 2001) models the selection mechanism for publication as a function of study effect y_i and associated standard error s_i. We introduce a latent variable z_i modeled as

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

where z_i models the publication process such that study i is selected (published) only if $z_i > 0$. The parameter γ_0 controls the baseline probability of publication, 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

$$\begin{pmatrix} \epsilon_i \\ \delta_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right) \tag{2.8}$$

where $\text{corr}(\epsilon_i, \delta_i) = \rho$ measures how the probability of selection changes with observed effect sizes.

Interpretation of the parameters $(\gamma_0, \gamma_1, \rho)$ can be difficult. The marginal probability that a study with standard error s_i is published is

$$P(z_i > 0|s_i) = \Phi(\gamma_0 + \frac{\gamma_1}{s_i}).$$

Thus, if γ_0 is large and positive then all studies are published with high probability regardless of the value of s_i. We restrict γ_1 to be positive under the assumption that larger studies are more likely to be published for various reasons (e.g. more funding, quality of writing, etc.). Larger values of γ_1 lead to larger differences in publication probabilities for studies with different standard errors. The correlation parameter ρ is the main driver in how unadjusted estimates of θ are biased from the truth. If $\rho = 0$, then the selection process does not depend on observed effect sizes and unadjusted estimates are unbiased. Positive values of ρ indicate that observed effects y_i influence the selection process such that larger values of y_i (relative to their true mean θ_i) are being selected for, while negative values of ρ would show selection favoring larger negative values of y_i.

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. Mavridis et al. (2013) instead places priors on the lower and upper bounds for the probability of publication, P_{low} and P_{high}, as

$$\begin{aligned} P_{\text{low}} &\sim \text{Uniform}(L_1, L_2) \\ P_{\text{high}} &\sim \text{Uniform}(U_1, U_2), \end{aligned} \tag{2.9}$$

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 $(P_{\text{low}}, P_{\text{high}})$ to (γ_0, γ_1) with a 1-to-1 transformation using

$$\begin{aligned} \gamma_0 + \frac{\gamma_1}{s_{\max}} &= \Phi^{-1}(P_{\text{low}}) \\ \gamma_0 + \frac{\gamma_1}{s_{\min}} &= \Phi^{-1}(P_{\text{high}}) \end{aligned} \tag{2.10}$$

where $s_{\min}$ and $s_{\max}$ are the smallest and largest observed standard errors in the sample of S studies. Bai et al. (2020) place priors directly on γ_0 and γ_1 as

$$\begin{aligned} \gamma_0 &\sim \text{Uniform}(-2, 2) \\ \gamma_1 &\sim \text{Uniform}(0, s_{\max}). \end{aligned} \tag{2.11}$$

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 $(-2, 3)$. Priors (2.11) are meant to be default prior distributions, while (2.9) 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 Bayesian model averaging that has superior performance in the $\mathcal{M}$-open setting. If we have K candidate models M_k indexed by $k = 1, \ldots, K$, 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 the predictive distributions $p_k(\tilde{y}|y, M_k)$ to the true distribution $p_T(\tilde{y}|y)$. Formally, Yao et al. (2018) define the stacking problem as

$$\max_{w \in \mathcal{S}_1^K} \Big(\sum_{k=1}^{K} w_k p(\cdot|y, M_k), p_T(\cdot|y) \Big) \tag{2.12}$$

where the first argument P in $S(P, Q)$ is a weighted sum of model-specific posterior predictive distributions. Equation (2.12) is intractable as written, so Yao et al. (2018) replace the "true" predictive distribution $p_T(\tilde{y}|y)$ with observed values y_i, and replace model k's predictive distribution $p(\tilde{y}|y, M_k)$ with its corresponding leave-one-out (LOO) predictive distributions $\hat{p}_{k,-i}(y_i) = \int p(y_i|\theta_k, M_k)p(\theta_k|y_{-i}, M_k)d\theta_k$, where θ_k are the parameters in model k and subscript $-i$ denotes the data y with observation i left out. The authors recommend the logarithmic scoring rule, which reduces the stacking problem to solving for weights w with

$$(\hat{w}_1, \ldots, \hat{w}_K) = \arg\max_{w \in \mathcal{S}_1^K} \frac{1}{n} \sum_{i=1}^{n} \log \sum_{k=1}^{K} w_k \hat{p}_{k,-i}(y_i) \tag{2.13}$$

via optimization. After solving for stacking weights $\hat{w}$, the *stacked posterior distribution* of a common parameter θ can be obtained by taking $\hat{w}_k \times T$ samples from each model k's posterior distribution $p(\theta|y)$ for a total of T posterior samples.

2.3.1 Pareto smoothed importance sampling

It would be computationally costly to refit each model n times to obtain LOO distributions $p_k(y_i|y_{-i}, M_k)$. Instead, Yao et al. (2018) use Pareto smoothed importance sampling (PSIS) (Vehtari et al., 2017) to obtain approximations. If we have draws $\theta_k^{(t)}$ from the full posterior $p(\theta_k|y, M_k)$, we can calculate importance ratios as

$$r_{i,k}^{(t)} = \frac{1}{p(y_i|\theta_k^{(t)}, M_k)} \propto \frac{p(\theta_k^{(t)}|y_{-i}, M_k)}{p(\theta_k^{(t)}|y, M_k)} \tag{2.14}$$

and use $r_{i,k}^{(t)}$ to generate model k's importance sampling leave-one-out (IS-LOO) distribution

$$p_k(\tilde{y}_i|y_{-i}) \approx \frac{\sum_{t=1}^{T} r_{i,k}^{(t)}\, p(\tilde{y}_i|\theta_k^{(t)}, M_k)}{\sum_{t=1}^{T} r_{i,k}^{(t)}}, \tag{2.15}$$

which we evaluate at the left out data point y_i to obtain the IS-LOO predictive density $p_k(y_i|y_{-i})$.

This approximation may be unstable because the importance ratios $r_{i,k}^{(t)}$ can have high or infinite variance. Vehtari et al. (2017) mitigate this issue by fitting a generalized Pareto distribution to the upper tail (top 20%) of importance ratios, which returns smoothed importance weights $w_{i,k}^{(t)}$ to be used in place of the original importance ratios $r_{i,k}^{(t)}$ in (2.15). The Pareto-smoothed importance sampling leave-one-out (PSIS-LOO) expected log-predictive density (elpd) for point i in model k can be calculated as

$$\text{elpd}_{i,k} = \log\left(\frac{\sum_{t=1}^{T} w_{i,k}^{(t)}\, p(y_i|\theta_k^{(t)}, M_k)}{\sum_{t=1}^{T} w_{i,k}^{(t)}}\right). \tag{2.16}$$

We use the R package 'loo' (Vehtari et al., 2020) to implement PSIS-LOO, which requires only a $T \times n$ matrix of posterior pointwise log-likelihood samples. The loo package also solves for stacking weights w when given an $n \times K$ matrix of pointwise elpd estimates with one column for each model.

The estimated shape parameter $\hat{k}$ from the generalized Pareto distribution can be used as a diagnostic for the reliability of PSIS-LOO approximations, where $\hat{k} > 0.7$ signals a potentially unreliable approximation. Vehtari et al. (2017) recommend sampling directly from $p_{k,-i}(y_i)$ (exact LOO) for problematic data points if the number of problematic y_i is small. Exact LOO requires fully refitting model k with the leave-one-out dataset y_{-i} and calculating the log-likelihood for the held out data point y_i. The exact LOO elpd for point i can then be combined with PSIS-LOO estimates to solve for stacking weights w.

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 model list does not contain the true model, we propose stacking over both Copas models (Mavridis et al., 2013; Bai et al., 2020) and a variety of stepped selection functions to obtain a more robust posterior distribution for the mean parameter θ.

To fit the Copas models we rewrite model (**??**) - (**??**) as

$$\begin{pmatrix} y_i \\ z_i \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \theta \\ u_i \end{pmatrix}, \begin{pmatrix} \tau^2 + s_i^2 & \rho s_i \\ \rho s_i & 1 \end{pmatrix}\right) \mathbb{1}_{z_i>0}, \tag{2.17}$$

which shows that we can first sample z_i from a truncated normal $z_i \sim \mathrm{N}(u_i, 1)\mathbb{1}_{z_i>0}$, and then sample $y_i|z_i \sim \mathrm{N}\big(\mathrm{E}[y_i|z_i], \mathrm{Var}[y_i|z_i]\big)$ where $\mathrm{E}[y_i|z_i] = \theta + \rho s_i(z_i - u_i)$ and $\mathrm{Var}[y_i|z_i] = \tau^2 + s_i^2(1-\rho^2)$ (Mavridis et al., 2013). We also need to extract the log-likelihood for each observation i in order to stack models. The model construction (2.17) leads to a simple form for the log-likelihood

$$\begin{aligned} L(\theta, \tau^2, \rho, \gamma_0, \gamma_1) &= \sum_{i=1}^{S} \log[p(y_i|z_i > 0, s_i)] \\ &= \sum_{i=1}^{S} \log\left[\frac{p(z_i > 0|y_i, s_i) f(y_i)}{p(z_i > 0|s_i)}\right] \\ &= \sum_{i=1}^{S} \log\big(\phi(y_i; \theta, \tau^2)\big) - \log\Phi(u_i) + \log\Phi(v_i) \end{aligned} \tag{2.18}$$

where $\phi(y_i; \theta, \tau^2)$ represents the density at the point y_i of a normal distribution with mean θ and variance τ^2, $\Phi(\cdot)$ represents the standard normal cumulative density function, and

$$
\begin{aligned}
u_i &= \gamma_0 + \frac{\gamma_1}{s_i} \\
v_i &= \frac{u_i + \tilde{\rho}_i \frac{y_i - \theta}{\sqrt{\tau^2 + s_i^2}}}{\sqrt{(1 - \tilde{\rho}_i^2)}} \\
\tilde{\rho}_i &= \frac{s_i}{(\tau^2 + s_i^2)^{1/2}} \rho.
\end{aligned}
$$

While Bai et al. (2020) use default prior distributions (2.11) for the parameters γ_0 and γ_1, Mavridis et al. (2013) instead advise researchers to use expert elicitation or historical data to specify (L_1, L_2) and (U_1, U_2) in (2.9) as plausible bounds for the lower and upper probabilities of publication. To avoid the need for strictly informative prior values, we specify $(L_1, L_2) = (0, 0.5)$ and $(U_1, U_2) = (0.5, 1)$, meaning we believe the lower bound for publication probability is between 0-50%, and the upper bound is between 50-100%. Mavridis et al. (2013) gives τ a half-normal prior $\tau \sim \mathrm{N}(0, 10^2)\mathbb{1}_{\tau>0}$, while Bai et al. (2020) uses a half-Cauchy prior $\tau \sim \mathrm{Cauchy}(0, 1)\mathbb{1}_{\tau>0}$. We fit the two Copas models in JAGS (Plummer et al., 2003).

As a default for the stepped models described in Section 2.1, we fit a 1-step model with the step at $p = 0.05$, a two-step model with steps at $p = (0.01, 0.10)$, and a two-step model with steps $p = (0.05, 0.20)$. We fit the stepped selection models in Stan (Gelman et al., 2015) because of the ability to easily code the custom probability distribution (2.5), which also makes extraction of the log-likelihood matrix simple. For a given dataset we fit both Copas models and the three stepped models for a total of $K = 5$ models to stack over.

3 Simulation

The aim of the simulation is to assess how well the stacked model described in Section 2.4 *stacks up* against the individual models <please help me find a way to keep the pun in here>. We simulate data for meta-analyses using a 2×3 factorial design with $\theta = (0, 0.4)$ and number of studies per meta analysis $S = (30, 60, 100)$ for 6 total scenarios. For each simulation scenario $j = 1, \ldots, 6$ we set $\tau = 0.2$ and study-specific standard errors s_i, $i = 1, \ldots, S$, are distributed as Uniform(0.2, 0.8). Study-specific true mean effects θ_i are distributed as $\theta_i \sim \mathrm{N}(\theta, \tau^2)$. We then sample $y_i \sim \mathrm{N}(\theta_i, s_i^2)$ as observed study effects. Each study i is assigned a probability of publication α_i based on one of multiple selection functions which are chosen deliberately such that none of the stepped selection models or Copas models can individually capture the true selection mechanism. Whether or not each study i is included in the meta-analysis is then determined by an indicator $B_i \sim \mathrm{Bernoulli}(\alpha_i)$.

Model performance is evaluated using bias, 95% interval coverage, 95% interval length, and root-mean squared error (RMSE).

3.1 Selection functions

The first selection function $f_1(p)$ decreases with one-sided p-values, i.e. $p_i = 1 - \Phi(y_i/s_i)$. It has three steps at $p = 0.005, 0.2, 0.5$, with exponential decay between the three steps. It is constant at $f_1(p) = 1$ for $p \in [0, 0.005)$. It then decreases exponentially as $f_1(p) = \exp(-2p)$ for $p \in [0.005, 0.2)$. There is then a step at $p = 0.2$, and $f_1(p) = \exp(-4p)$ for $p \in [0.2, 0.5)$, and is constant at $f_1(p) = 0.1$ for $p \in [0.5, 1]$. See Figure [reference figure] for a graphical depiction.

3.2 Simulation results

We simulated $K = 1000$ iterations for each scenario j.

4 Numerical example

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.

Iyengar, S. and Greenhouse, J. B. (1988). Selection models and the file drawer problem. *Statistical Science* pages 109–117.

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.

Le, T. and Clarke, B. (2017). A bayes interpretation of stacking for m-complete and m-open settings. *Bayesian Analysis* **12,** 807–829.

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., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P.-C., Paananen, T., and Gelman, A. (2020). loo: Efficient leave-one-out cross-validation and waic for bayesian models. R package version 2.4.1.

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.

List of Figures

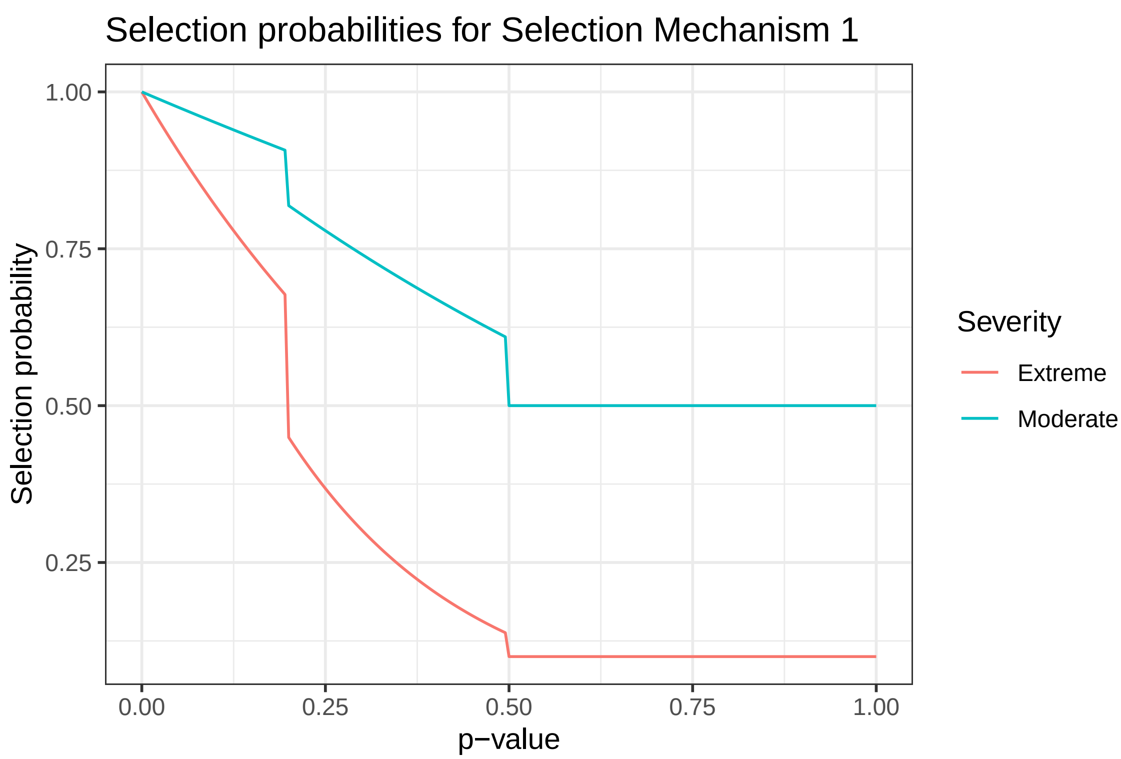

Figure 1: Selection mechanism 1: declining selection probabilities with increasing p-values. There are steps at $p = .2$ and $p = .5$ and exponential decay between cut-points. Selection probability is constant after $p = 0.5$.

List of Tables

Model	Bias	Avg. SD	95% Cover	95% Length	RMSE
Bai	0.130	0.175	0.925	0.692	0.190
Mavridis	0.066	0.184	0.978	0.728	0.164
one.step	0.169	0.125	0.808	0.491	0.194
stacked	0.060	0.167	0.951	0.664	0.160
standard	0.234	0.115	0.439	0.452	0.254
three.step	0.089	0.147	0.983	0.581	0.137
two.step	0.121	0.131	0.944	0.516	0.155

Table 1: Simulation results for selection function 1 with 30 base studies

Model	Bias	Avg. SD	95% Cover	95% Length	RMSE
Bai	0.093	0.129	0.895	0.508	0.148
Mavridis	0.031	0.122	0.969	0.483	0.117
one.step	0.172	0.084	0.468	0.329	0.185
stacked	0.036	0.117	0.937	0.462	0.119
standard	0.220	0.076	0.130	0.297	0.230
three.step	0.099	0.101	0.870	0.396	0.131
two.step	0.130	0.090	0.743	0.353	0.149

Table 2: Simulation results for selection function 1 with 60 base studies