Bayesian correction of systematic reviewer bias

Kristopher T. Jensen

ktj21@cam.ac.uk

1 Motivation

Submitting and presenting research at large-scale conferences such as Cosyne (for computational neuroscience) and NeurIPS (for machine learning) is becoming increasingly important for the career progression of junior researchers. These conferences often rely on a large pool of reviewers giving scores to an even larger pool of paper since submission numbers in the thousands make it impossible for any single reviewer or chair to consider all submissions. However, since all papers are reviewed by different finite sets of reviewers, we are introducing additional stochasticity into the review process if the reviewers have non-zero bias. Given the relative importance of this process, it might be desirable to account for such biases in a systematic way.

First we need to define what we mean by bias. Here, we take the bias of a reviewer to be the expectation over a hypothetical infinite number of papers of the difference between their review score and a hypothetical mean score across infinite reviewers:

$$b_i^{true} = \langle r_{ij} - \langle r_{kj} \rangle_{n_{p_j} \to \infty} \rangle_{n_{r_i} \to \infty} \tag{1}$$

Where r_{ij} is the score given to paper j by reviewer i, n_{p_j} is the total number of reviewers for paper j and n_{r_i} is the total number of papers reviewed by reviewer i.

Furthermore, we will take the hypothetical expectation over infinite reviewers to be the 'true goodness' of paper j:

$$p_j^{true} = \langle r_{kj} \rangle_{n_{p_j} \to \infty} \tag{2}$$

Thus the bias of reviewer i is the expected deviation from the true paper goodness over infinitely many reviews.

2 Model

Let each paper have a ground truth goodness of p_j. We could model this explicitly as a probability distribution, but for now we will just fit a delta function (corresponding to the $\sigma_p \to 0$ limit).

In addition, let each reviewer have a bias b_i and variability σ_i. This variability can be interpreted as 'measurement noise' or an MSE loss, but it also captures variability from things like topic preferences which we cannot take into account with this approach.

We now model each review as a composition of these two processes such that

$$r_{ij} = \tilde{p}_j + \tilde{b}_i \tag{3}$$

where

$$\tilde{p}_j \sim \mathcal{N}(p_j, \sigma_j = 0) = p_j \tag{4}$$

and

$$\tilde{b}_i \sim \mathcal{N}(b_i, \sigma_i) \tag{5}$$

Since we model p as a delta function, the distribution of r_{ij} has variance $\sigma_i^2 + \sigma_j^2 = \sigma_i^2$ and we can simply write

$$r_{ij} \sim \mathcal{N}(p_j + b_i, \sigma_i) \tag{6}$$

This is illustrated in Figure 1a. This distribution has a simple log likelihood

$$\log p(r_{ij}) = -\log \sigma_i - \frac{(p_j + b_i - r_{ij})^2}{2\sigma_i^2} - 0.5 \log 2\pi \tag{7}$$

If we assume some prior over the true goodness of papers

$$p_j \sim \mathcal{N}(\mu_p, \sigma_p) \tag{8}$$

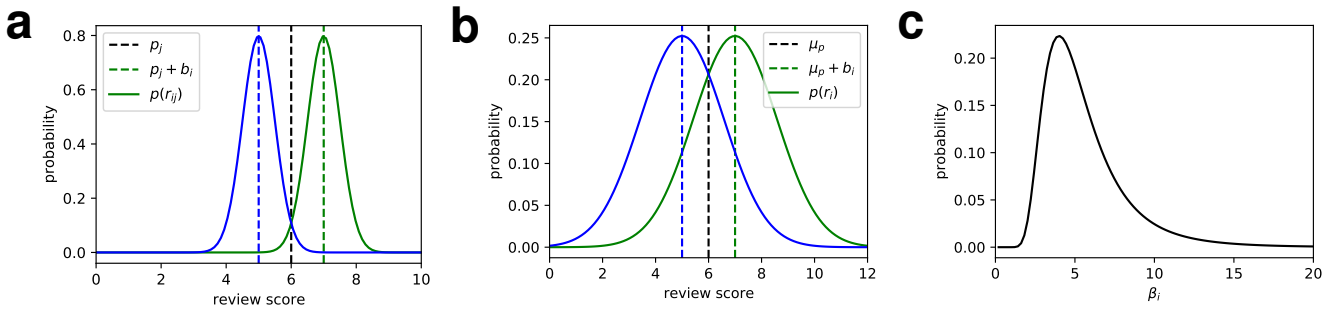

Figure 1: **Bayesian model. (a)** Example $p(r_{ij}|b_i, p_j, \sigma_i)$ for $\sigma_i = 0.5$, $p_j = 6$ and $b_i = +1$ (green) or $b_i = -1$ (blue). **(b)** Example $p(r_i|b_i, \sigma_i, \mu_p, \sigma_p)$ for $\sigma_i = 0.5$, $\sigma_p = 1.5$, $\mu_p = 6$ and $b_i = +1$ (green) or $b_i = -1$ (blue). **(c)** Prior over $\beta_i \sim$ Inv-Gamma$(\alpha = 6, \beta = 28)$.

Equation 6 leads to each reviewer having a different distribution of review scores in the hypothetical limit of infinitely many papers (Figure 1b)

$$p(r_i) = \int p(r_i|p_j)p(p_j)dp_j = \mathcal{N}(r_i|b_i + \mu_p, \sqrt{\sigma_i^2 + \sigma_p^2}) \tag{9}$$

We also include a unit Gaussian prior over the biases:

$$b_i \sim \mathcal{N}(0, 1) \tag{10}$$

This reflects our prior belief that reviewers are at least somewhat consistent, and also conveniently ensures that the mean bias is zero in accordance with Equation 1 – and this will be true independently for each group if there are disjoint groups of reviewers. In practice, it simply amounts to a ridge regression penalty on the biases.

Finally we need a prior over the precisions $\beta_i = \sigma_i^{-2}$. This is necessary to avoid overfitting a single reviewer, assuming them to be an oracle and others to be useless:

$$\beta_i \sim \text{Inv-Gamma}(\alpha_\beta, \beta_\beta) \tag{11}$$

Note that we need $\alpha_\beta + 1 > 0.5n_r$ for ALL reviewers; i.e. $2\alpha_\beta + 2$ must be larger than the largest number of papers reviewed by a single person (Figure 1c).

We now proceed to find the MAP estimate of all parameters under a uniform prior over p_j

$$p(p, b, \beta|r) \propto p(r|p, b, \beta)p(b)p(\beta) \tag{12}$$

such that we can simply minimize the loss function

$$\mathcal{L} = -\log p(\{r\}|\{p\}, \{b\}, \{\beta\}) - \log p(\{b\}) - \log p(\{\beta\}) + const. \tag{13}$$

Importantly the model has a simple log likelihood

$$\begin{aligned}
\mathcal{L} &= -\log \prod_{(i,j)} p(r_{ij}|\{b\}, \{p\}, \{\beta\}) - \log \prod_i p(b_i) - \log \prod_i p(\beta_i) + const. \\
&= -\sum_{(i,j)} \log p(r_{ij}|b_i, p_j, \beta_i) - \sum_i \log p(b_i) - \sum_i \log p(\beta_i) + const. \\
&= \sum_{(i,j)} \left[-0.5 \log \beta_i + 0.5\beta_i(p_j + b_i - r_{ij})^2\right] + \sum_i \frac{b_i^2}{2} + \sum_i (\alpha_\beta + 1) \log \beta_i + \beta_\beta \beta_i^{-1}) + const.
\end{aligned} \tag{14}$$

where the sum over (i, j) indicates all reviewer-paper pairs and $\beta_i = \sigma_i^{-2}$

Fortunately this is easy to optimize w.r.t all parameters $\{b\}, \{p\}, \{\beta\}$ since the corresponding derivatives are given by

$$\frac{\partial \mathcal{L}}{\partial b_i} = b_i + \sum_{i \in (i,j)} \beta_i \delta_{ij} \tag{15}$$

$$\frac{\partial \mathcal{L}}{\partial p_j} = \sum_{j \in (i,j)} \beta_i \delta_{ij} \tag{16}$$

$$\frac{\partial \mathcal{L}}{\partial \beta_i} = (\alpha_\beta + 1)\beta_i^{-1} - \beta_\beta \beta_i^{-2} + \sum_{i \in (i,j)} 0.5\delta_{ij}^2 - 0.5\beta_i^{-1} \tag{17}$$

Where $\delta_{ij} = p_j + b_i - r_{ij}$ and the sums run over all pairs (i, j) which include the index with respect to which we are taking a derivative. We could also treat the variance of the prior $p(b)$ as a hyperparameter to be optimized by crossvalidation. Note that if we let the prior variance go to zero, all biases go to zero and we recover the naive solution where p_j is simply the (uncertainty-weighted) mean of the reviews.

Conveniently, this likelihood allows us to compute closed form updates in a self-consistent manner

$$p_j|b, \beta = \frac{1}{\sum_{j \in (i,j)} \beta_i} \sum_{j \in (i,j)} \beta_i (r_{ij} - b_i) \tag{18}$$

$$b_i|p, \beta = \frac{\beta_i}{1 + n_{r_i}\beta_i} \sum_{i \in (i,j)} r_{ij} - p_j \tag{19}$$

$$\beta_i|p, b = \frac{1}{2\sum_{i \in (i,j)} 0.5\delta_{ij}^2} \left[0.5n_{r_i} - (\alpha_\beta + 1) + \left((0.5n_{r_i} - (\alpha_\beta + 1))^2 + 4\beta_\beta \sum_{i \in (i,j)} 0.5\delta_{ij}^2 \right)^{0.5} \right] \tag{20}$$

This corresponds to an uncertainty-weighted average goodness for each paper given the current biases, the mean regularized residual bias for each reviewer, and the regularized empirical variance for the noise term.

For comparison we consider a 'mean field' approach to debiasing where we compute

$$b_i^{MF} = \frac{1}{n_{r,i}} \sum_{i \in (i,j)} r_{ij} - \frac{1}{N} \sum_{(i,j)}, r_{ij} \tag{21}$$

i.e. we estimate the bias of each reviewer as the difference between the mean of their scores and the global mean, taking into account their interactions with other reviews only through this averaged quantity (reminiscent of e.g. the mean field approach in the Ising model). Contrast this with the correlated model where we consider direct interactions between all reviewers sharing a paper (through the p_j variables).

This allows us to compute de-biased mean field review scores

$$r_{ij}^{MF} = r_{ij} - b_i^{MF} \tag{22}$$

and mean field paper goodnesses

$$p_j^{MF} = \frac{1}{n_{p_j}} \sum_{j \in (i,j)} r_{ij}^{MF} \tag{23}$$

3 Simulations

We now simulate data for $N_r = 280$ reviewers each reviewing $n_r = 10$ submissions such that each of $N_p = 700$ submissions receive $n_p = 4$ reviews for approximate consistency with the Cosyne 2020 numbers. We assume all submissions to be drawn from a Gaussian with $\mu_p = 6$ and $\sigma_p = 2$. Following generation of all r_{ij} according to the model above, we round all scores to the nearest integer and threshold at 0 and 10 for consistency with real reviews and to avoid direct model recovery. Finally, we assume that our reviewers are ϵ-greedy such that $p = 0.04$ of the reviews are replaced with a random integer between 0 and 10.

This gives a distribution of review scores (Figure 2a) to which we can fit our Bayesian model. We find that the model does an excellent job recovering both the true bias of each reviewer (Figure 2b) and as a consequence also the true paper goodness (Figure 2c).

Of course the importance of bias correction depends on the magnitude of the biases in the population of reviewers. We therefore draw the true biases from a distribution

$$b_i \sim \mathcal{N}(0, \sigma_b) \tag{24}$$

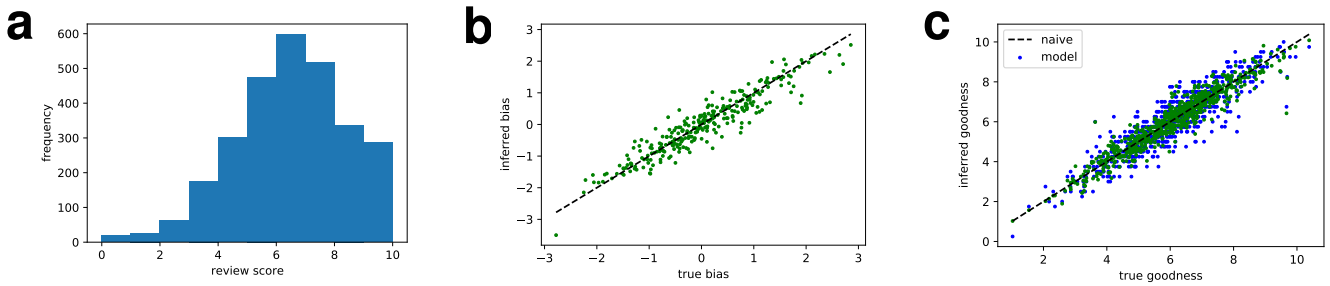

Figure 2: **Simulated example.** **(a)** Distribution of review scores for a synthetic dataset. **(b)** Bias inferred by the Bayesian model vs true bias. **(c)** Paper goodness inferred by the model (green) and simple average review scores (blue) vs true paper goodness.

and quantify the correlation between inferred paper goodness and true paper goodness for a range of σ_b. We find that the quality of our predictions are largely invariant to σ_b in stark contrast to the raw average score which deteriorates rapidly as a predictor of p^{true} as a function of σ_b (Figure 3a). This illustrates the importance of de-biasing when there are systematic biases in the reviewer population but only a finite number of reviews per paper. Like Bayesian debiasing, the mean field approach also leads to an invariance of performance to σ_b. However, while the Bayesian approach consistently outperforms naive averaging, the mean field approach is worse than raw averaging for low levels of reviewer bias and worse than Bayesian debiasing for all values of σ_b. This highlights the importance of using an approach that takes into account the correlations between individual reviewers. Note that the improvement of Bayesian debiasing over raw averaging for zero reviewer bias is due to the reviewer-specific noise model allowing us to discount the opinion of worse reviewers; this gap is closed if the reviewers are not ϵ-greedy and all have similar 'observation noise'.

Instead of correlations between true and inferred goodness, we can also quantify the actual effect in a hypothetical conference setting where only 40% of submissions are accepted. To do this, we quantify the overlap between the true top 40% papers and the top 40% of inferred paper goodnesses according to each of our models. We find that this 'succes rate' exhibits a qualitatively similar pattern to the goodness correlations and that the Bayesian bias-correction again outperforms the other models across all values of σ_b with particular importance at high biases (Figure 3b). Finally we can compare the standard deviation of the fitted bias distributions to the true values of σ_b. We find the empirical value to be a good predictor of σ_b (Figure 3c), suggesting that this quantity can be used to empirically estimate bias of the population and thus the importance of debiasing in real data.

It is perhaps at first sight surprising that the mean field approach performs so poorly at low bias levels. This deficiency occurs because the MF model does not take reviewer interactions into account to explicitly model paper goodness, such that the mean field bias is correlated with the true goodness of the assigned papers (Figure 4a). In the case where there is no bias to remove, the mean field model thus still removes the part of the true signal that is correlated with mean reviewer scores, leading to worse predictions. This effect will be smaller as reviewers review more papers

Figure 3: **Bayesian debiasing improves paper ratings.** **(a)** Mean and std of the correlation between true and inferred goodness of papers as a function of the width of the bias distribution in the population of reviewers. **(b)** Fraction of true papers in the top 40% that are correctly selected when using the inferred scores from each model as a proxy for the true scores. **(c)** The std of the inferred bias is a good estimator of the true bias distribution.

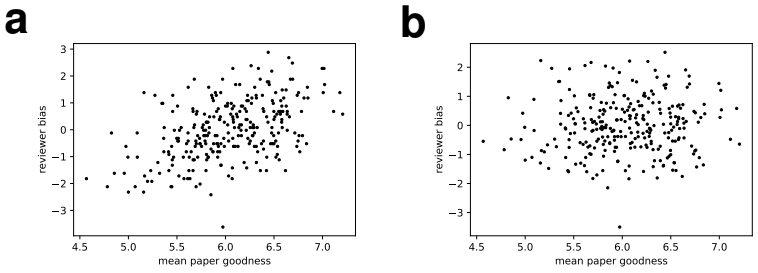

Figure 4: **Correlations between bias and paper quality. (a)** Mean field bias estimates are correlated with the true goodness of the reviewed papers (r = 0.47, p = 5e-17) showing that this approach removes true signal in addition to de-biasing. **(b)** Bayesian bias estimates are uncorrelated with the true paper goodness (r = 0.05, p = 0.37).

each, and the mean field model is exact in the limit of infinite reviews per reviewer (c.f. Equation 1) just like simple averaging is exact in the limit of infinite reviews per paper. Conversely the mean field performance decay will increase when reviewers review fewer than 10 papers each as in most ML conferences. In contrast to the mean field method, the inferred bias in the Bayesian model is uncorrelated with the true paper goodness (Figure 4b), explaining why its performance matches that of simple averaging even when all reviewers are unbiased.

We now proceed to validate the method using cross-validation to develop a metric that is not dependent on knowing the true paper goodness. This is important as we seek to quantify the effect of bias-correction in real data where we do not have access to a ground truth p_j. Firstly, to consider how well each model accounts for the data, we perform hold-one-out crossvalidation on each individual review, with predictions consisting of the sum of the corresponding paper goodness and reviewer bias for each model ($r_{ij}^{predict} = p_j^{model} + b_i^{model}$). We then compute the error $\epsilon = |r_{ij}^{pred} - r_{ij}^{true}|$ for each paper and each method. When averaging this CV error over all reviews, we find that it increases as a function of reviewer bias when taking a raw average but not when fitting either bias-correcting model (Figure 5a). Notably, these CV errors follow a very similar pattern to what we saw for correctly accepted papers and paper goodness correlation.

We therefore proceed to correlate (i) the difference in CV error for each of the Bayesian and MF models compared to the simple average estimate ($\langle \Delta\epsilon_i \rangle$) and (ii) the difference in r^2 between true and predicted p_j values for each model compared to the simple average estimate (Δr^2). We find that these are very strongly correlated, suggesting that CV error provides an excellent predictor of how well we approximate the true paper goodness without the need for access to any ground truth data (Figure 5b). Note that these analyses were done without the ϵ-greedy reviewer feature; the Bayesian and MF models are slightly worse at crossvalidation with ϵ-greedy reviewers (i.e. we risk thinking a Bayesian model is not necessary when in fact it is).

To understand why such a CV approach might be a good predictor of how well we capture the ability to predict p_j, let us write down a loss function for our predictions for a hypothetical infinite set of new reviewers.

$$\mathcal{L} = \langle (r_{ij}^{pred} - r_{ij})^2 \rangle_{n_{r_i}, n_{p_j} \to \infty} \tag{25}$$

The squared error for a given paper is minimized by the posterior mean, which by definition is the true goodness of the paper:

$$r_{ij}^{pred} = \langle r_{ij} \rangle_{n_{r_i} \to \infty} = p_j^{true}. \tag{26}$$

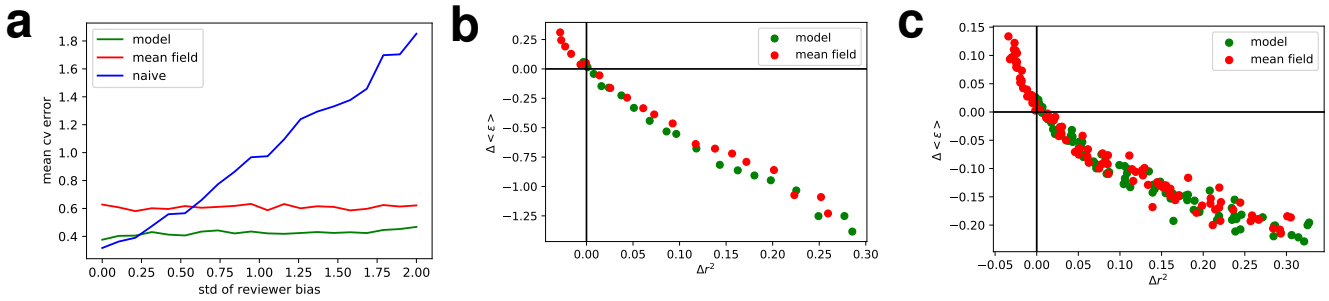

Figure 5: **Crossvalidation provides a good predictor of model quality. (a)** Cross-validation error as a function of the std of the reviewer bias for each of the models. **(b)** Difference in crossvalidation error as a function of difference in r^2 between true and inferred p_j. **(c)** As in (b), now predicting scores for held-out reviewers instead of individual reviews.

Thus the best we can do to predict a review from an unknown reviewer is to predict the true paper goodness, providing us with an unbiased way of estimating the quality of our predicted scores. Indeed, we can perform a similar analysis to the one described above, but now performing hold-one-out crossvalidation for each reviewer instead of each review. We similarly find that the CV performance in this case is highly correlated with the ability to predict paper scores (Figure 5c), although it is more noisy than the single-review predictions given the absence of a bias estimate for each new reviewer.

4 Outlook

It would be interesting to get access to an anonymized dataset with all review scores given by a set of reviewers to a set of conference submissions in order to apply the model to real data. We could then use hold-one-out crossvalidation to estimate the hypothetical effect of bias-correction for a more informed view of whether such approaches should be utilized in practice. In addition, it would provide an estimate of the distribution of biases in the reviewer population which could be interesting to analyze in its own right as it would tell us something about individual differences in how scientific work is judged.

On one hand, the effect of the debiasing proposed here in terms of the actual conference content is likely to be small since the left-out submissions are replaced by alternative submissions of only slightly lower quality. However, conferences etc. have a disproportionate effect on the careers of junior scientists, and the feeling of 'unfairness' can be quite demotivating.

In this view, rejecting a slightly worse submission can be likened to crowning the wrong winner in a very close 100m race. While it may have made no difference to the spectators who the true winner was, we do not argue that we 'might as well' give the gold medal to the second place finisher since they were close anyways – instead we spend large sums of money on high-tech equipment to closely analyze photo finishes. With this in mind, one might also expect large conference like NeurIPS or Cosyne to consider additional ways to increase the fairness of their submission selection policy.

Conversely, in cases where debiasing has only little effect on the results, it is likely to be desirable to use the raw scores rather than a Bayesian bias-correction model. This is due to the psychological effect of having 'an algorithm' determine the outcome inspiring less confidence than real humans – an effect we recently got a very prominent example of with the ofqual grade correction fiasco in the UK (although one would hope that the conference organizers are more qualified than the UK government and less inclined to use a broken model).

Here we have provided one way to address systematic biases which seems to perform well at least on our synthetic datasets. However, we have made several essentially arbitrary model choices, and there may be other methods that perform better on real data. Additionally, we are not taking into account the fact that reviewers might be confident in one field (low uncertainty) but also have to review a subset of papers they know less about (high uncertainty). In many cases, reviewers are asked to provide 'confidence scores' for each reviewed paper which can help mitigate this issue. We envisage that these confidence scores could be explicitly incorporated into the noise model but leave the specifics for future considerations.

Finally it would be desirable to develop online methods which may be of use in other contexts such as journals which can update their reviewer biases after each review. For this purpose, we envisage something resembling the TrueSkill algorithm where a single instance of a paper being reviewed by several reviewers can be treated as an all-vs-all competition in the gaming industry.