A Bayesian Model for Calibrating Reviewer Scores

Hong Ge
University of Cambridge

Max Welling
University of Amsterdam

Zoubin Ghahramani
University of Cambridge

Abstract

A typical technical conference involves each submission being reviewed by several reviewers of different expertise, preference of technique, and attention to detail. Furthermore, some reviewers have a tendency to give high scores, some reviewers low. These are detrimental for the overall quality of the reviewing. Can we estimate the reviewer bias scientifically? Here we first review a method for calibrating review scores used by NIPS (2006-2012), and present a revised Bayesian model that is used in NIPS 2013, 2014. We also investigate the potential of improving the calibration performance by incorporating extra information like review confidence factors.

1 Reviewer Bias Modelling: Platt-Burges Model

We introduce some notations first. Let us index papers with i, and reviewers with j. r_{ij} is the score reviewer j gave to paper i, g_i is the unobserved true score for paper i (call it the "goodness" of paper i), b_j is the bias of reviewer j. Furthermore, let C_{ij} denotes reviewer j's confidence for her/his review of paper i, and $C_{ij} = 0$ if reviewer j didn't review paper i. Then the Platt-Burges model can be written down as

$$\begin{aligned} r_{ij} &= g_i + b_j + \epsilon_{ij}, \\ \epsilon_{ij} &\sim \mathcal{N}(0, \sigma). \end{aligned} \tag{1}$$

To solve the above equation, we minimise the regularised least-squares:

$$L = \frac{1}{2}\Sigma_i \Sigma_{j \in R_i} (r_{ij} - b_j - g_i)^2 + \frac{1}{2}\lambda \Sigma_j b_j^2 \tag{2}$$

where λ is the regularisation parameter, and R_i is the set of reviewers for paper i.

The Platt-Burges model estimates the per-reviewer biases, and per-paper quality scores. It has been used in several NIPS conferences to identify overly positive or negative reviewers and to adjust decisions accordingly. In the following section we first provide a Bayesian reformulation of the Platt-Burges model, and use it as a starting point for deriving more sophisticated models.

2 A Bayesian Re-formulation of the Platt-Burges Model

To begin, we place the following priors on g_i and b_j, with hyper-parameters of the prior motivated by the data:

$$g_i \sim N(5, 2), \tag{3}$$
$$b_j \sim N(0, c). \tag{4}$$

Here the prior on b_j has a regularisation effect, i.e. a small variance c pulls biases towards zero, which is similar with the effect of regularisation parameter $1/\lambda$ in the Platt-Burges model. The probability (density) distribution of review score r_{ij} is

$$\begin{aligned} r_{ij} &\sim N(g_i + b_j, 1/(dA_{ij})), \\ d &\sim \text{Gamma}(a_d, b_d) \end{aligned} \tag{5}$$

where A_{ij} is a sparse matrix with 0's and 1's indicating that reviewer i scored paper j. We refer to this Bayesian re-formulation as the Bayes-Platt-Burges model. Bayes-Platt-Burges is more general than Platt-Burges due to the introduction of variance parameter d (this possible extension was noted by Platt and Burges). [1]. In the case of $d = 1$, Bayes-Platt-Burges degenerates to Platt-Burges. In general, we can re-parametrise d to incorporate extra factors that has influence on review score. This is discussed in more details in the coming section.

[1] To quote, "Note that this formulation also permits the use of the reviewer confidences: takes the form of a precision of a Gaussian. Instead of 1 for all papers, we can set it to be the confidence that reviewer has in his review of paper . This makes the innate goodness the weighted average of the corrected reviews."

2.1 Modelling the Effect of Review Confidence

In the NIPS review score data, each review is accompanied with a review confidence. This is a useful indicator variable C_{ij} reflecting reviewer j's confidence/-expertise on paper i. Will this information help in calibrating the scores? To investigate this effect, we propose and evaluate several variants of the Bayes-Platt-Burges model that takes care of reviewer confidence when calibrating scores.

2.1.1 Bayes-Platt-Burges+ Model

We obtain the first Bayes-Platt-Burges model variant by replacing the term A_{ij} with $1/dC_{ij}$ in Equation 5:

$$r_{ij} \sim N(g_i + b_j, 1/(dC_{ij})). \tag{6}$$

Here C_{ij} denotes the confidence of reviewer j for her/his score of paper i.

2.1.2 Bayes-Platt-Burges++simple Model

The second model variant is obtained by replacing the term A_{ij} with C_j

$$r_{ij} \sim N(g_i + b_j, 1/C_j), \tag{7}$$

where C_j is reviewer specific, and $C_j \sim \text{Gamma}(a, b)$.

2.1.3 Bayes-Platt-Burges++ Model

The third model variant is obtained by replacing replacing the term A_{ij} with $1/C_jC_{ij}$:

$$r_{ij} \sim N(g_i + b_j, 1/(C_jC_{ij})), \tag{8}$$

where $C_j \sim \text{Gamma}(a, b)$.

2.1.4 Bayes-Platt-Burges+e Model

The last model variant is

$$r_{ij} \sim N(g_i + b_j, (e + C_{ij})^\alpha/d), \tag{9}$$

where $e \sim \text{Gamma}(1, 1)$, and $\alpha \sim \text{Unif}(-1, 1)$. This model generalises Bayes-Platt-Burges and Bayes-Platt-Burges+, and has the potential of selecting the "right" model automatically from data:

$$\frac{(e + C_{ij})^\alpha}{d} = \tag{10}$$
$$\begin{cases} 1/d, & \text{if } \alpha = 0 \Rightarrow \text{Bayes-Platt-Burges} \\ \frac{1}{dC_{ij}}, & \text{if } \alpha = -1, e = 0 \Rightarrow \text{Bayes-Platt-Burges+} \\ \alpha > 0 & \Rightarrow \text{Confidence unreliable} \end{cases}$$

In general, we can write all above models in the following form

$$r_{ij} = g_i + b_j + k_{ij}, \tag{11}$$

where $k_{ij} \sim N(0, \tau_{ij})$. Here the term g_i is the unobserved true "value" of paper i, b_j is the overall bias of reviewer j, k_{ij} is a noise term caused by reviewer j's confidence for her/his review of paper i. The more confident a reviewer is about score r_{ij}, the smaller the k_{ij} is.

3 Inference

During NIPS 2013, we used Gibbs sampling to perform inference in our models. More specifically, we worked out the Gibbs-conditionals for each g_i, b_j and per-reviewer confidence bias C_j (if exists, otherwise slice sampling is used). In 2014, we re-implemented the model using the STAN language, which provides an excellent automated inference engine based on the NUTS sampler.

4 Models Comparison

In order to empirically compare the proposed models and the Platt-Burges model, we split the NIPS scores into two data sets: train data set (about 90%), test data set (about 10%). The test data $R_{test} = \{r^t_{ij}\}$ is selected randomly from the score matrix with the following criteria: 1) paper i must have at least 2 reviews; 2) reviewer j must at least scored 4 papers. We consider two different metrics for accessing model's power: predictive log likelihood (PLL) and root mean squared error (RMSE).

4.1 Predictive Log Likelihood

The predictive log likelihood is computed in the following way:

$$\begin{aligned} \ell_{\text{pred}} &= \frac{1}{|R_{\text{test}}|} \log \prod_{r_{ij} \in R_{\text{test}}} P(r_{ij}|R_{\text{train}}) \\ &= \frac{1}{|R_{\text{test}}|} \log \prod_{r_{ij} \in R_{\text{test}}} \int P(r_{ij}|g, b)) P(g, b|R_{\text{train}}) dgdb \end{aligned} \tag{12}$$

where $|R_{\text{test}}|$ denotes to the number of scores in the test set, and

$$P(g, b|R_{\text{train}}) = \frac{P(g)P(b)P(R_{\text{train}}|g, b)}{\int P(g)P(b)P(R_{\text{train}}|g, b)dgdb}. \tag{13}$$

The priors over g and b are given by Equation 3 and 4.

4.2 Root Mean Squared Error

The RMSE can be computed in the follow way:

$$\text{RMSE} = \sqrt{\frac{1}{|R_{\text{test}}|} \sum_{r_{ij} \in R_{\text{test}}} (r_{ij} - \bar{g}_i - \bar{b}_j)^2} \qquad (14)$$

where $\bar{g}_i$ and $\bar{b}_j$ are mean goodness of paper i and mean bias of reviewer j, i.e., $\bar{g}_i = \frac{1}{S}\sum_s g_i^{(s)}$, $\bar{b}_j = \frac{1}{S}\sum_s b_j^{(s)}$, and s index samples from the posterior distribution defined by Equation 13. Alternatively, RMSE can be computed while MCMC is running, i.e.

$$\text{RMSE}^* = \sqrt{\frac{1}{S \times |R_{\text{test}}|} \sum_{s=1}^{S} \sum_{r_{ij} \in R_{\text{test}}} (r_{ij} - g_i^{(s)} - b_i^{(s)})^2} \qquad (15)$$

where S is the total number of samples from the posterior, and $|R_{test}|$ is the total number of scores in the test set.

The difference between RMSE and RMSE* is subtle. Technically, RMSE should be computed in the way defined by Equation 15. However, if one want to make any prediction for test data, an intuitive yet simple way is to compute the sum of paper i's mean goodness and reviewer j's mean bias, i.e. $r_{ij} = \bar{g}_i + \bar{b}_j$. This is exactly how prediction is computed in Equation 14. In practise, RMSE and RMSE* emit very similar numbers. In this article, all results are reported using RMSE.

4.3 Probability of Accepting a Paper

The probability that paper should be accepted is computed in the following way:

Initialisation: $P(\text{all papers}) = 0$, S = total number of samples after burin;
for *each sample (indexed by s)* **do**
  Rank all paper according to $g_i^{(s)}$;
  if *rank(i) in top 370* **then**
    $P(i) = P(i) + 1$
  end
end
return P/S;

5 Results

5.1 Predictive Performance

Table 1 shows the predictive performance of different models, in terms of two evaluation metrics: negative predictive log likelihood (NPLL) and root mean squared error (RMSE). Smaller numbers are better.

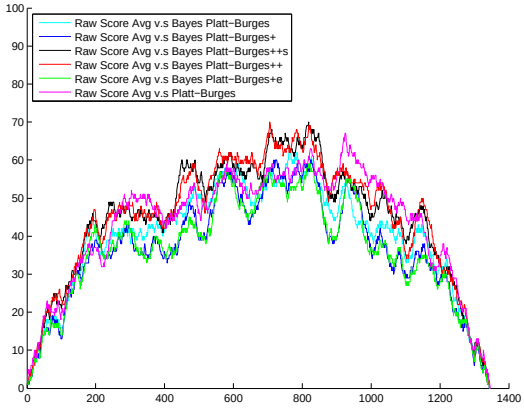

Figure 1: Comparison of paper ranking using calibrated scores returned by different models. The x axis index paper submissions, and the y axis shows ranking differences between each proposed model and the base line (average raw score): lower plots means smaller ranking difference compared to the baseline.

In terms of RMSE, we can see all Bayesian models performs much better compared with Platt-Burges. Models that take review confidences into account generally performs worse that those that do not. We can also see from the results Bayes-PlaBur+ beats all other models. This is a bit surprising because we would expect a model with more flexibly (e.g. Bayes-PlaBur+s and Bayes-PlaBur++ have an additional per-reviewer confidence bias C_j parameter, see Equation 6) would win. However this is not the case here due to the limitation of our data: each reviewer only has 2–6 review confidences. It turns out it performs better by pooling all these confidence numbers together and learn a shared confidence d (see Equation 5).

5.2 Paper Ranking using Calibrated Scores

Papers are ranked using calibrated score. Each calibration model leads to a different ranking result. To further analyse properties of different review calibration models, we compare the pair-wise distances of different rankings. More specifically, for each model, we compute the ranking distance against a baseline model (average raw review score) in the following way:

$$f_{ab}(k) = \big|\{\text{top k papers in the model a}\} \backslash \{\text{top k papers in model b}\}\big|, \qquad (16)$$

where $A \backslash B$ denotes set difference[2], and $|C|$ denotes the number of elements in set C.

[2]That is, $A \backslash B = \{x : x \in A \text{ and } x \notin B\}$.

	PlaBur	B-PlaBur	B-PlaBur+	B-PlaBur+s	B-PlaBur++	B-PlaBur+e
NPLL-Train	-	1.4581	1.5788	1.4131	1.4401	1.5769
NPLL-Test	-	2.8280	1.9743	3.3083	3.3089	1.9735
RMSE-Train	0.9114	0.9573	1.0529	0.9865	0.9941	1.0497
RMSE-Test	1.8534	1.6493	1.6042	1.6329	1.6443	1.6008

Table 1: Root mean square error (RMSE) and negative predictive log likelihood (NPLL) for different models. Smaller numbers are better. The testing data is about 10% entries randomly selected from NIPS scores.

As shown in Figure 1, paper ranking returned by Bayes-PlattBurges+ is most close to that of raw average score. This surprisingly coincides with our previous observation: Bayes-Platt-Burges+ outperforms all other models in terms of predictive performance. This hints calibrated score from a good model is close to the average raw score. This makes sense because if we have many reviews for each paper, then law of large numbers guarantees the mean is near the unobserved tree score. We also observe that almost all Bayesian models are closer to the average raw score compared to Platt-Burges.

5.3 Calibrated Scores from Different Models

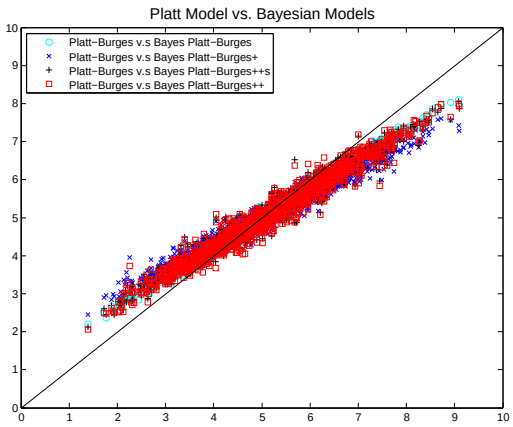

Figure 2: Comparison of corrected scores returned by different models. Bayesian models are plotted against the baseline: Platt-Burges. Here all models are run using all NIPS scores.

Figure 2 shows the calibrated score of different models. In summary, Bayesian models are slightly more conservative: calibrated scores are shifted towards the the mean score (i.e. 5) compared with Platt-Burges. If we look at this plot together with other results (ranking + predictive performance), we will see in general, the more conservative a model is, the better predictive performance it has, or the closer its ranking is to the average raw score. This observation is also consistent with paper ranking, that is, more conservative model has smaller ranking distance compared to the average raw score.

Acknowledgements

We thank John Platt and Chris Burges for sharing their NIPS model for removing reviewer bias. We also thank David Mackay for some comments on this draft.

References

Platt, B. and Burges, C. (2012). Regularized least squares to remove reviewer bias.

A Stan implementation

```
data {
  int<lower=1> I; # number of papers;
  int<lower=1> N; # total number of reviews
  int<lower=1> J; # number of reviewers
  vector<lower=0> [N] scores;      # review scores vector
  vector<lower=0> [N] confidence; # review confidences
  int reviewerID[N]; # reviewer ID for each review
  int paperID[N];     # paper ID for each review
}
parameters {
  vector [I] truescore;
  vector [J] reviewbias;
  real<lower=0> c;
  real<lower=0> d;
}

model{
  c ~ gamma(1,1);
  d ~ gamma(1,1);

  truescore ~ normal(5, 2);
  reviewbias ~ normal(0, 1/c);
  for(i in 1:N){
    scores[i] ~ normal(truescore[paperID[i]]
                       + reviewbias[reviewerID[i]], 1/d);
  }
}
```

B Examples: Posterior Distribution of Paper Goodness and Review Bias

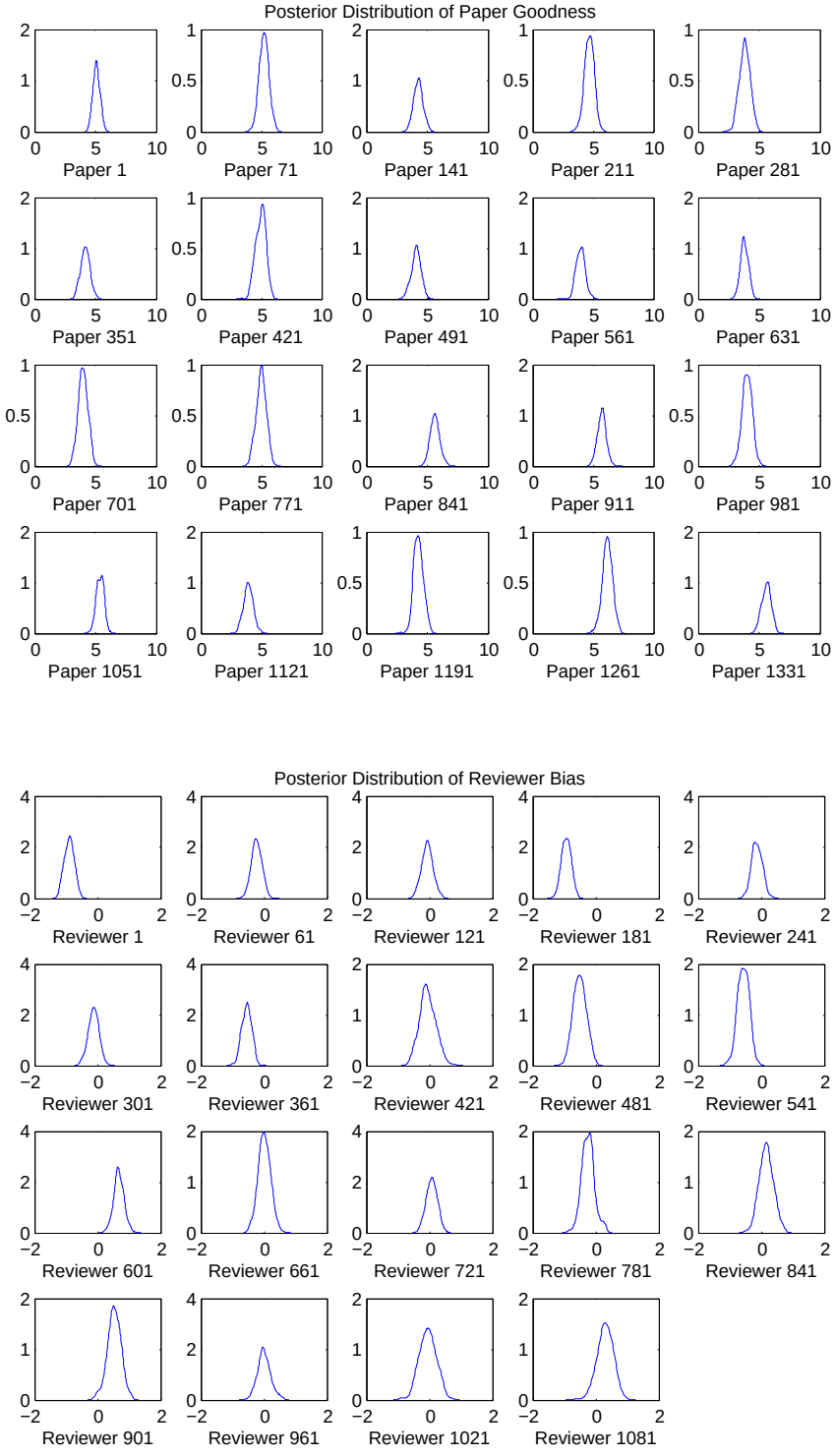

Figure 3: Posterior distribution from Bayes-Platt-Burges. Note full NIPS (2013) review data is used here.

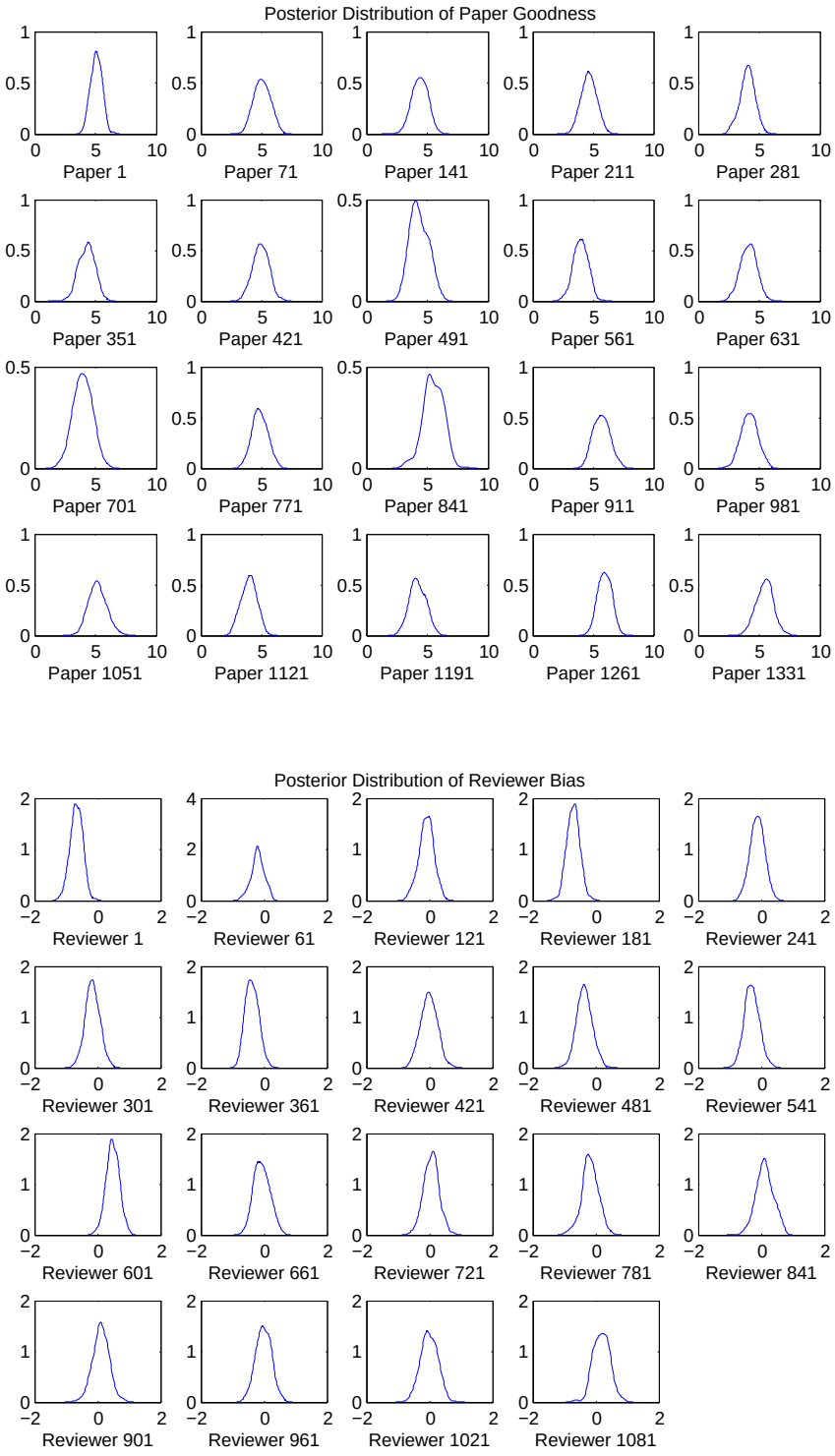

Figure 4: Posterior distribution from Bayes-Platt-Burges+. Note full NIPS (2013) review data is used here.