Response to Reviews: *Scalable Bayesian Preference Learning for Crowds*, *Edwin Simpson and Iryna Gurevych*.

We thank all the reviewers and editors for their time in reviewing our paper and giving very helpful feedback. We have heavily revised the submission to address all of your concerns as much as possible. We provide a point-by-point response to the reviews below.

Request from ECMLPKDD 2019 Journal Track Chairs

“Moreover, please note that the call for papers for ECMLPKDD journal track explicitly states the following: "Consequently, journal versions of previously published conference papers, or survey papers will not be considered for the special issue." This is in contrast to "regular" journal submissions, which can be extended versions of previous conference papers. The earlier publication you have on this topic was presented at ACL 2018. We would thus like you to make sure that the revision puts an even stronger emphasises on the new results, to make sure that the manuscript solidly goes beyond an "extended conference paper".”

We have reworded the paper to emphasise its new contributions. In particular, we propose a new model, crowdGPPL, which tackles a different task from the ACL 2018 work, namely predicting personal preferences for individual users. This model is substantially different because it uses matrix factorisation to model correlations between different users. We have now made it clearer that the new model is the focus of the paper, and that the ACL 2018 paper is previous work.

Reviewer #1

“Fix all of the typos, especially in the introduction”

We have revised the introduction and sorted out the typos.

“- y(a,b) versus y(a > b): choose one notation”

This was a mistake -- resolved to use y(a,b).

“- what is f? Sure, a function. But is f : R^m -> R^k for some k? Or to some other codomain? Not clear.”

We have clarified that f:R^D -> R, i.e. f maps a D-dimensional vector to a scalar utility value.

“- I have no clue what you mean by (2). Do you mean \int_{\delta_a \sim N(..)} \int{\delta_b \sim N(..)} perhaps?”

The equation appeared incorrect as it was split across two lines – this is now on one line. The equation states the integral of the joint distribution over label y(a,b) and the noise terms \delta_a and \delta_b given f, integrated over \delta_a and \delta_b. That is, it shows how the noise terms are marginalised from the joint distribution over observations and noise.

"- In (4), what is K_\theta?"

The prior covariance matrix for items in the training set – this is now introduced on page 5 before the equation.

"- What is \mathcal{R}? Do you mean the real numbers \mathbb{R}?"

Yes – the typo has been fixed.

"- Sizes of the matrices F, V, and W are inconsistent with one another."

Apologies for the error – this has now been corrected on page 6.

"- In (6), what are the K's?"

We have now clarified in the text that K_{\theta} is the prior covariance of the items and L_{\eta} is the prior covariance of the users (using new notation).

"- In (7), recall that the z-hats are as in (3)."

This section has been rearranged. We now refer the reader back to Equation 3 when z-hats are used in Equations 9 and 10.

"- In (11), what are the q's? Are you postulating a separated representation for some unknown function q? Not clear."

The q() functions are the variational factors. The use of q represents an approximate distribution. In our revision, we explain this explicitly when discussing this equation (now Equation 12).

Reviewer #2:

"This paper extended the pairwise ranking model in Xi Chen's work, Pairwise Ranking Aggregation in a Crowdsourced Setting with Gaussian Process."

We apologise that the main contributions of our paper were not clear in the previous version of the paper. Our method does not extend Chen et al. (2013), i.e., the crowdBT model. In the new version of the paper, we set out more explicitly how our model and crowdBT are related (see pages 2, 3, 6, 13). Briefly:

- CrowdBT learns the accuracy of each annotator, assuming that their errors are due to random noise. We do not use the crowdBT model of annotator accuracy, so our model cannot be seen as an extension of crowdBT. Instead, we model each annotator's individual preferences, assuming that their disagreements are due to personal preferences/bias rather than noise/random errors.
- CrowdBT aggregates pairwise labels from multiple annotators to estimate the consensus. Our method also predicts the personal preferences of individual users.

- In contrast with our model, the input features of the items are not used by CrowdBT, so it cannot predict scores for new items with no pairwise labels. This is the scenario we test in the experiments on *Argument Convincingness*. Our method is aimed not just at aggregating crowdsourced data, but at learning a predictive model for test data.
- The pairwise label likelihood in our method is a Thurstone-Mosteller model, whereas CrowdBT is an extension of the Bradley-Terry model.

“The preference learning has well-studied. Particularly, the pairwise label using BT model has been published by Xi Chen's work. Recently, the model has been extended to more general PL model, such as:

Han et al., Robust Plackett-Luce model for k-ary crowdsourced preferences. Machine Learning 107(4): 675-702 (2018)
Pan et al., Stagewise learning for noisy k-ary preferences. Machine Learning 107(8-10): 1333-1361 (2018)

Those models are more general and robust to the noisy preferences in crowdsourcing settings.

In terms of robustness of noise modeling, I do not find the proposed model can be significantly better than the previous works. There is no experiments comparing with any baselines in crowdsourcing, even Chen's work. ”

We have now cited these works. They tackle a somewhat different problem, like crowdBT, as they focus on labelling noise, whereas we model individual preferences of each user in the crowd. However, it is a good suggestion to include a comparison with one of these methods for the consensus prediction task, so we now include a new baseline, crowdBT-GP, which our method outperforms. This show the benefits of modelling individual preferences in a subjective task.

“Another claim from the authors is the SVI for large-scale crowdGPPL, which is due to limitation of GP. Indeed, the online strategy has been well-studied in crowdBT, crowdPL, the previous works and the following work.
Li et al,. Hybrid-MST: A Hybrid Active Sampling Strategy for Pairwise Preference Aggregation, NIPS 2018

Therefore, I don't find any new insight of SVI here.”

The core reason for needing a scalable approach is the use of Gaussian processes to provide a Bayesian, nonlinear mapping from input features to latent utilities. This problem does not arise in methods based on crowdBT as they do not integrate input features to learn a predictive model for new users and items.

“The only new point is equation (5), but it is not clear why introducing t. I expected the authors should have more discussion on this.”

We agree that there was not enough discussion of this core equation, so have added a full explanation. We use t to model the crowd consensus, which captures the widespread appeal of an object; the other terms capture the preferences of individual users.

“There are many approximations in Sec 4.1 in order to make a tractable inference. It is not clear after so many approximations, eq (7), eq (8), eq (9), eq (11)... , the resultant objective may be very far away from the original one. The properties from GP may not be effective. There is no study to check the gap.”

We have added references to previous work that also used Gaussian likelihood approximations for classification and preference learning with GPs, and for extended Kalman filters. These works include empirical comparisons that support this type of approximation. Nonetheless, we find that our method performs well empirically, and therefore is a practical solution for large datasets.

“This paper simply applies the GP to Xi Chen's model. The technical quality is weak.”

We have answered this point above – the paper has been modified heavily to make the new contributions clearer.

“CrowdBT using GP may be interesting. However, as studied by Han's work and Pan's work, the crowdPL model is more robust than the crowdBT model, so the proposed model does not provide any improvement to crowdsourcing.
Overall, the current version of this manuscript is not sufficient for machine learning models. The authors should empirically compare other baselines in crowdsourcing setting to demonstrate the significance.”

As mentioned above, we include a new baseline, crowdBT-GP, which shows that a de-noising methods does not perform as well as a method that models user biases. Furthermore, the new method addresses subjective crowdsourcing tasks, where predicting the different preferences of each user is important – we show this in the results for the *Argument Convincingness* experiments.

Reviewer #3

“… the algorithm is complex and I would strongly advise the authors to publish the code alongside their article. Otherwise, I suspect the work could not be replicated easily.”

We fully agree that publishing the code is important for replication. We have made the code available at the link in the paper and are continuing to improve the documentation and tidy up the repo. Our collaborators are currently applying the code to a new task, so we fully intend to further improve the quality of the software. To further aid with replicating the method, we believe we have included all the equations required to implement the algorithm in the appendix as well as the main body of the paper.

“… the paper suffers many typographical errors that need to be addressed, along with the issues mentioned above, before publication.”

We apologise for the typos, and have taken care to resolve them.

“Page 1, Define, or give an example of 'sparse data' (e.g. 'when some items are not explicitly ranked in the dataset' ?)”

This is now on page 2 – “sparse, i.e., many items or users have few or no labels”.

“Page 2, line 5: gold-stanard
Page 2, line 22: thesse
Page 3, line 17: parentheses for citation of Simpson and Gurevych 2018 are wrong.
Page 3, line 19: We then we develop
Page 4, line 2: parentheses wrong around Joachim citation.”

Fixed; the text has been edited so these are no longer on the same pages/lines.

“Section 2.2: Define or give an example of 'input features' when introduced. “

Thank you for the suggestion. We now include an example on page 2 in the introduction, as the concept is important to clarify early on. We also specify the input features for the Argument Convincingness task, which gives another example.

“Page 6, equation (2): several extra commas.
Page 6, preamble to equation (3): If we have Gaussian distributions over ....
Page 6, equation (3): there should be a 2 before \sigma_{a,b}.
Page 6, preamble to equation (4): I would add 'where a_p and b_p are items'.”

Typos are now corrected.

“Page 7, line 20: you have used \sigma as output scale here as well as cross variance in equation (3). This might lead to confusion.”

Thank you for noting this – we have changed the notation so that ‘s’ is used for all output scales, with a superscript to show which variable it scales.

“Page 7, equation (5): you might want to cite precedence to this form of equation
e.g. K. Mo, E. Zhong, and Q. Yang. Cross-task crowdsourcing. In Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '13, pages 677-685, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2174-7. doi: 10.1145/2487575.2487593. URL http://doi.acm.org/10.1145/2487575.2487593.”

Thank you for the suggestion – we have included the reference in a brief discussion of the equation.

“Page 8, line 12: VB is only guaranteed to converge for conjugate distributions.”

Thank you – we have modified the text to avoid over-claiming.

“Page 8, second paragraph: you criticise GFITC for not being decentralisable. It is not clear to me that crowdGPPL is decentralisable. “

We have now included a short description of how our SVI method could be decentralised. Separate nodes can process different mini-batches inside one VB iteration (the outer loop of our algorithms), then send the sufficient statistics for the inducing points to be merged by a central node using the SVI update equations 17 and 18.

“I got really confused by Section 4.1. In equation (7), p(y|f) is a product of the cdfs. Where does the expectation come from just after the first equal sign? The second equal sign in (7) should be an approximation. In equation (7), Q is (approximately) the covariance of y conditioned on f and yet, in equation (8), Q is determined by the expectation over f. Can you please check this section and provide some further explanation.”

We agree that this section was unclear so we have rewritten it extensively. Q represents observation noise variance; it approximates a Bernoulli distribution, whose variance depends on its mean; therefore, if Q is to approximate the variance in the Bernoulli likelihood, it has to depend on f; hence we estimate Q by estimating f at each point independently of other points; given the estimate of Q, we can use it to obtain a better posterior approximation for f. We now cite previous work that also uses this method.

“Page 9, Section 4.2: Please state up front that you are proposing a mean-field variational approximation in equation (11). This will then justify why E[s] appears in equation (13). “

This is now stated explicitly when the approximation is introduced.

“Page 9, equation (14): something wrong with the parentheses.
Page 9, line 50: S should be bold.
Page 10, line 47, equation (18): final two matrices in equation for C* are in the wrong order.”

These have been corrected.

“Page 12, Section 5. Can you provide a web reference to data sources (not just papers that use them)?”

We now provide links in the footer.

“Page 13, line 22: please restate that 'C' is the number of features.”

We have added this note.

“Page 13, equation (25): please provide reference for 'Dmedian' function.”

We have now clarified this equation: for the simulations and Sushi experiments, we set length-scales to median distances – the function is just the median. For Argument Convincingness, the length-scales are increased by scaling the median function.

“Page 13, line 41: Can you provide some intuition as to how values for r, |P| and epsilon are chosen in general?”

We have now included a study of the effect of |P| (with our new notation, this is labelled P_i). For r and \epsilon, we follow recommendations from Hoffman et al. (2013), which found r=0.9 to be a good trade-off between performance and convergence time, and did not find any benefit to changing \epsilon. Our preliminary experiments drew the same conclusions. We have added some discussion to the text to explain this.

“Page 14, line 21: to to.”

Fixed.

“Page 14, Figure 1: include error bars, especially in sub-figure (c).”

The simulation and runtime plots now show error bars.

“Page 15, Table 2: Please provide a definition of 'accuracy'. Is this the proportion of pairs that are correctly ranked?”

We have now defined the metrics explicitly on page 13. Accuracy is the proportion of pairwise labels in the test set that were predicted correctly.

“Page 17, lines 30 - 35: include, in Table 3, a comprehensive set of results for rival algorithms.”

We have included the two closest approaches to crowdGPPL in the Sushi experiments. In the paper, we discuss how they relate to crowdGPPL.

“Appendix A, line 21: f to the left of the semi-colons should not have a hat over them. Appendix A, line 28: s should be s hat in log|K/s|

Please check rest of the Appendices.”

We have checked the appendix, fixed some typos and simplified the equations for the variational lower bounds.