Bayesian System Inference On Shallow Pools

Rodger Benham[1], Alistair Moffat[2], and J. Shane Culpepper[1]

[1] RMIT University, Melbourne, Australia
[2] The University of Melbourne, Melbourne, Australia

Abstract. IR test collections make use of human annotated judgments. However, new systems that surface unjudged documents high in their result lists might undermine the reliability of statistical comparisons of system effectiveness, eroding the collection's value. Here we explore a Bayesian inference-based analysis in a "high uncertainty" evaluation scenario, using data from the first round of the TREC COVID 2020 Track. Our approach constrains statistical modeling and generates credible replicates derived from the judged runs' scores, comparing the relative discriminatory capacity of RBP scores by their system parameters modeled hierarchically over different response distributions. The resultant models directly compute risk measures as a posterior predictive distribution summary statistic; and also offer enhanced sensitivity.

1 Introduction

TREC COVID [20] is the first IR evaluation track to use the *residual collection scoring* pooling methodology described by Salton and Buckley [17]. The track judged multiple rounds of runs, with shallow judgments made available after each round, to allow tuning of systems in subsequent rounds. Several participants raised concerns about the generalizability of the first round judgment set, after the RBP $\phi = 0.5$ [13] residuals were found to be unacceptably high for systems not included in the judgment pool. Voorhees [19] investigated the effect that further judgments had on the system orderings between the complete set and the first round set, finding that a small portion of systems had significant changes – the worst being `RMITBFuseM2` which rose 33 ranks on P@5. Shallow judgments are also used for the MS MARCO [14] runs, a collection with so many topics that deep judgment coverage would be very costly. When system scores are uncertain, practitioners might decide to only evaluate pooled systems.

In general, when attempting to ascertain whether a ranker outperforms one or many others, a statistical test is employed to mitigate against sampling error. Sakai [15] notes that the most popular statistical test at present is the Student t-test. However, it (and all other frequentist tests) assumes that the sample of scores are one of many repeated samples from a population of score differences. Hence, using a t-test, even if the systems were both pooled, might produce *overconfident* confidence intervals, as an entire population of unseen topics are inferred against based on scores derived from low-fidelity judgments. Conversely, Bayesian inference allows the predicted score replicates to be conditioned on the measured pooled system-topic scores only.

In this paper, we adapt models initially described by Carterette [3] to infer graded RBP $\phi = 0.8$ scores over multiple systems hierarchically [1], and analyze the relative power of the resulting models using the pooled TREC COVID first round submissions and judgments, finding increased sensitivity. Other recent work [1] has also investigated Bayesian "risk" overlays which penalize systems for relative effectiveness loss against a baseline by a linear scalar r. We explore a similar summary statistic using the posterior predictive distribution (PPD).

2 Related Work

Carterette [2] was the first to use Bayesian inference as an alternative to frequentist statistical testing for IR effectiveness scores. Carterette [3] then empirically evaluated the outcomes of these models on the TREC-8, Robust04, and TREC Web 2012 track datasets. Sakai [16] shows that Bayesian Markov Chain Monte Carlo (MCMC) simulation can also be used to generate complementary information about the effect size of different systems, by calculating Glass' Δ and expected *a posteriori* (EAP) values for one-to-one system comparisons.

In early work on risk measures, Collins-Thompson [4] explored methods to measure the risk of query drift in query expansion. Similarly, Wang et al. [21] defined URisk as a learning-to-rank objective function. Dinçer et al. [6] then extended the URisk measure to be an inferential risk measure using the t-distribution, calling the result TRisk. Dinçer et al. [7] noted that in this one-to-one risk evaluation setting, experimental system comparisons will be biased to the baseline ranking; prompting the development of ZRisk and GeoRisk [5].

Benham et al. [1] recently combined Bayesian inference and risk-adjusted score overlays at the system-topic level on multiple systems. However, they did not compare the relative system effectiveness inferences over statistical models that consider system-topic-rank gain scores in the way that was proposed by Carterette [2]. That gap is targeted in this work.

3 Statistical Models

Our primary goal is to understand how increasingly sophisticated models affect assessment as to which ranker is the most effective. Bayesian inference techniques effectively reverse-engineer the parameters required to generate the underlying score observations in a parametric way, conditioned on a set of priors. Those parameters can be inferentially evaluated directly using a hierarchical model, such as a system effect parameter, to infer which system(s) are better [12]. We use the `brms` front-end to the `Stan` statistical programming language, in the `R` programming language to specify the models.[3] In our simulations, we use the default weakly-informative priors in `brms`, which are auto-scaled with MCMC to be credible fits against the observed score values. Benham et al. [1] explain the process of generating Bayesian inferences in greater detail.

[3] Code to reproduce available at: `https://github.com/rmit-ir/bayesian-shallow`

Using the pooled runs submitted to 2020 TREC COVID Track, we compare statistical outcomes when treating observed RBP score values, assuming either Gaussian or Zero-One Inflated Beta (ZOiB) distributions. Additionally, we model the RBP gain values directly on a per-document basis (cutting each system-topic ranking to the pooling depth of 7 documents), similar to Carterette [3], and compare against a Gaussian approach. The Gaussian method is a useful reference point, as it is similar in response distribution to t-distributed values [2]. Note that it is the differences in per-topic effectiveness scores between two systems that are studentized – beyond those score pairs for multiple system comparisons, many pairs of tests are run and corrected for. Therefore this exercise cannot guarantee that one approach gives inferences that are more "truthful" than others, as such a proof does not exist. The bottom 25% of pooled systems were discarded, to avoid comparisons being performed against erroneous runs.

Linear Model. The first model, Gaussian, simplistically assumes that the underlying distribution of RBP values is normally distributed, and is a function of a system and topic effect[4]:

$$
\begin{aligned}
y_{ij} &\sim N(\hat{\alpha}_i + \hat{\beta}_j, \sigma_y^2) & \sigma_{\{y,\alpha,\beta,\alpha_i,\beta_j\}} &\sim t(3, 0, 2.5) \\
\hat{\alpha}_i &= \omega_{\alpha,\alpha_i}\mu_\alpha + (1 - \omega_{\alpha,\alpha_i})\alpha_i & \mu_\alpha &\sim N(0, \sigma_\alpha^2);\ \alpha_i \sim N(0, \sigma_{\alpha_i}^2) \\
\hat{\beta}_j &= \omega_{\beta,\beta_j}\mu_\beta + (1 - \omega_{\beta,\beta_j})\beta_j & \mu_\beta &\sim N(0, \sigma_\beta^2);\ \beta_j \sim N(0, \sigma_{\beta_j}^2)\,,
\end{aligned}
$$

where y_{ij} is an RBP effectiveness score parameterized by topic j and system i. The topic and system effects, β_j and α_i respectively, are moderated by *partial pooling* in the corresponding $\hat{\beta}_j$ and $\hat{\alpha}_i$ [11], where $\omega_{Y,y}$ is the pooling factor that measures the simulated strength of the population Y versus the observed group effect y (topics for example, β is the topic population parameter averaged from all other topics in the model, and β_j is the specific topic effect for the y_{ij} observation, for example, topic 3)

$$\omega_{Y,y} = 1 - \frac{\sigma_Y^2}{\sigma_Y^2 + \sigma_y^2}\,.$$

The parameters provided to the standard deviation three-parameter Student t-distribution prior and hyperpriors correspond to the non-informative defaults in `brms` for the Gaussian family. The above approach is related to the Model 2 specified by Carterette [3], with marginally more informative priors than the Jeffreys prior ($\sigma \sim \log(1/\sigma)$).

ZOiB Model. Inspection of the PPD of the Gaussian model (top of Figure 1a) indicates that the MCMC simulation converges towards a distribution that describes some characteristics of the underlying effectiveness data. However, as Gaussian values are in the range $(-\infty, \infty)$, the replicate effectiveness scores are frequently invalid. A Beta distribution can be used to model a rate in the range

[4] This amends Benham et al. [1, Eqn. 3], which omitted the partial pooling notation.

$(0, 1)$, and a ZOiB distribution extends that range to $[0, 1]$.[5] We thus model RBP scores with the ZOiB parameters

$$y_{ij} \sim \begin{cases} \pi_0 & \text{if } y_{ij} = 0 \\ (1-\pi_0)(1-\pi_1)\beta(\mu_{ij}\phi, (1-\mu_{ij})\phi) & \text{if } 0 < y_{ij} < 1 \\ \pi_1 & \text{if } y_{ij} = 1 \end{cases}$$

$$\begin{aligned} \text{logit}\, \mu_{ij} &\sim N(\hat{\alpha_i} + \hat{\beta_j}, \sigma_y^2) & \sigma_{\{y,\alpha,\beta,\alpha_i,\beta_j\}} &\sim t(3, 0, 2.5) \\ & & \pi_0, \pi_1 &\sim \beta(1, 1) \\ & & \phi &\sim \gamma(0.01, 0.01) \\ \hat{\alpha}_i &= \omega_{\alpha,\alpha_i}\mu_\alpha + (1-\omega_{\alpha,\alpha_i})\alpha_i & \mu_\alpha &\sim N(0, \sigma_\alpha^2);\ \alpha_i \sim N(0, \sigma_{\alpha_i}^2) \\ \hat{\beta}_j &= \omega_{\beta,\beta_j}\mu_\beta + (1-\omega_{\beta,\beta_j})\beta_j & \mu_\beta &\sim N(0, \sigma_\beta^2);\ \beta_j \sim N(0, \sigma_{\beta_j}^2)\,, \end{aligned}$$

where ϕ is the precision parameter of the Beta distribution β to be modeled with a Gamma distribution (another `brms` default), π_0 and π_1 are the Bernoulli probabilities that a score will be zero or one, and μ_{ij} is logit transformed to link the linear parameterization (described in Gaussian) to the Beta distribution.

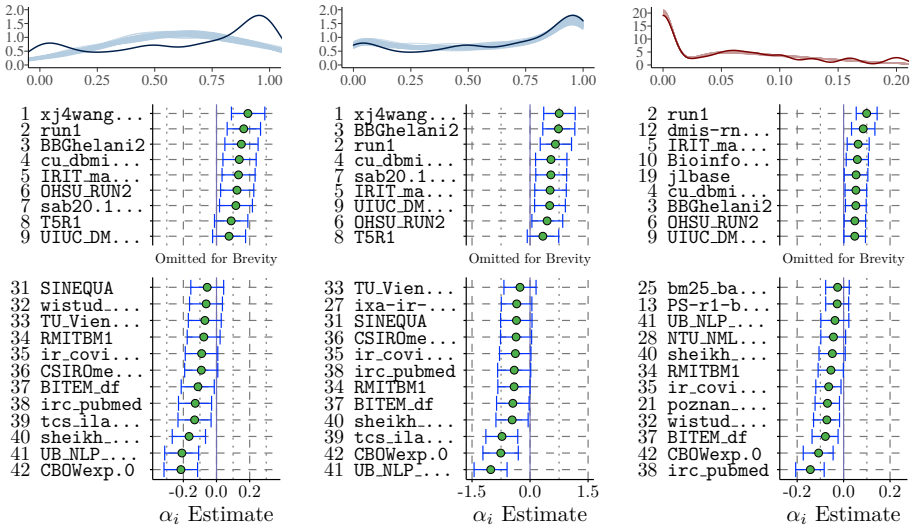
ZOiB-Rank. The ZOiB model can be extended to model y_{ijk} per-position RBP gain scores by including k as a rank parameter, modeled as a population effect. ZOiB-Rank is therefore a small modification: logit $\mu_{ijk} \sim N(\hat{\alpha_i} + \hat{\beta_j} + k, \sigma_y^2)$. (Carterette [3] used the very similar Quasi-Binomial distribution to model RBP gain scores, a response family that is not available in `brms`.) Of interest is comparing the properties of the system effect inferences of this gain-based approach against traditional RBP scores.

Posterior Predictive Risk. The URisk overlay with a challenger system against a champion computes the value:

$$URisk_r = -(1/n) \cdot \left[\sum Wins - r \cdot \sum Losses \right] . \tag{1}$$

Benham et al. [1] inferentially evaluate risk-adjusted scores using a Bayesian approach, with increasing r resulting in increased uncertainty according to their system effects. That uncertainty stems from attempting to predict instances where an experimental system would outperform the baseline (also known as the model selection problem). Here, we note that risk measures are essentially a summary statistic. As we can predict scores from experimental and baseline systems in a joint statistical model that has already been implicitly corrected for multiple comparisons in the Bayesian way (via hierarchical modeling [9], noting that the technique and any other correction approach is not flawless [10]), the PPD of what is judged to be the best fitting measure can be used to analyze the spread of the URisk values [8]. That is, for each draw from the posterior $\theta_i \sim p(\theta \mid data)$, the set of point parameter estimates from that draw θ_i is used to form *a posteriori* replicate scores supplied to URisk: $data'_i \sim p(data \mid \theta_i)$ [12].

[5] `https://rdrr.io/cran/brms/man/brmsfamily.html`, accessed October 29, 2020.



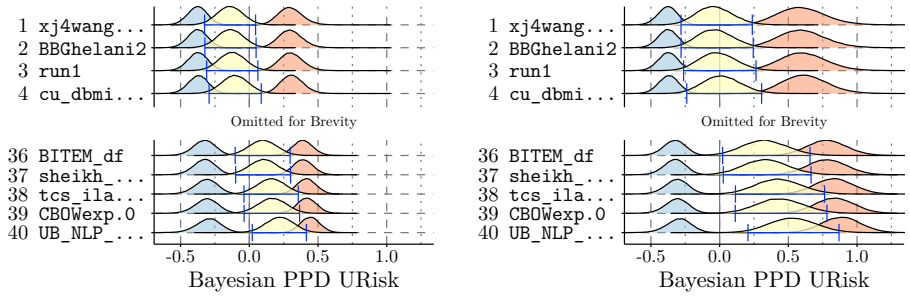
(a) Gaussian (b) ZOiB (c) ZOiB-Rank

Fig. 1: RBP with $\phi = 0.8$: Bayesian analysis of system effects for three different models, with 95% credible intervals. The top graphs are described in the text. Numbers to the left of each system corresponds to the ordering the Gaussian model invoked as a reference.

4 Analysis

Figure 1 plots the parametric inferences of the system effect for 95% credible intervals for the three models. The density plot above each column contains RBP topic scores amalgamated over all systems (blue solid line) or RBP gain scores combining all system-topic-rank scores (red solid line). Faint lines plotted behind these distributions are draws from the PPD which graphically indicates model fit – lines closer to the original distribution are preferable.

As can be seen, the two ZOiB distributed models have a better fit than the Gaussian model. The best system can be distinguished from 17 other (poor) systems, and the worst system from 23 (good) systems with the Gaussian model; with the corresponding numbers being 17 and 29 for the ZOiB model, and 20 and 31 using the ZOiB-Rank model. For ZOiB-Rank, the 12^{th} best system from the Gaussian model (`dmis-rnd1-run3`) moved up to 2^{nd} place with ZOiB-Rank, and the run `xj4wang_run1` moved from 1^{st} to 10^{th}. These shifts occur because ZOiB-Rank preferences systems more likely to report an RBP gain at any observed rank, rather than top-heavy systems that may return fewer relevant outcomes at the $\phi = 0.8$ expected viewing depth of 5 documents. Given that ZOiB visually fits the score distribution better than the Gaussian counterpart and does not draw unexpected predictions as in the ZOiB-Rank approach, the ZOiB model provides the most accurate description of system ranking dominance of the three tested, on the first round TREC COVID dataset.



(a) $r = 1$ (b) $r = 2$

Fig. 2: RBP $\phi = 0.8$ EAP risk, URisk with two risk values r against the `bm25_baseline` run, 95% credible intervals, and with wins (blue), losses (orange), and run aggregates (yellow) plotted.

Using the ZOiB model, Figure 2 compares risk-free ($r = 1$) against risk-sensitive ($r = 2$) evaluation using EAP values. In Figure 2a only `UB_NLP_RUN_1` (truncated) is able to be discriminated from the `bm25_baseline` run as the interval excludes zero, which is consistent with the extended parameter inference plot (without omitted systems) in Figure 1b. In Figure 2b, more challengers are statistically separable, while still being constrained to the observed outcomes in the pooled set. This EAP approach is therefore an improvement over the Benham et al. [1] approach, as it does not subsume the increased variance from the other challenger systems into the champion baseline system – providing more discriminative inferences in terms of the original URisk units.

5 Conclusion

Using the first round of the TREC COVID track, we modeled RBP scores via three separate distributions inspired by Carterette [3], and observed outcomes for many-to-many inferential system comparisons using a Bayesian hierarchical model. We found that the ZOiB method worked well for the corpus and smooth evaluation metrics considered, noting that further work is required to ascertain its applicability to other datasets (indeed, Urbano and Nagler [18] show that a one-size-fits-all model is rarely preferable). We also modeled risk inferentially using the PPD, which is more discriminative than modeling risk scores directly.

We posit that Bayesian hierarchical modeling may complement traditional IR statistical tests, and particularly recommend their use when there are fidelity concerns about the judgments used to form the evaluation scores. While these Bayesian methods are also amenable to more generalizing collection-based comparisons, they are not without limitations: they are orders of magnitude slower than traditional IR tests; and, in our observations to date, tend to require at least five systems to simulate the system parameters without divergent iterations.

Acknowledgments. This work was partially supported by Australian Research Council Grant DP190101113. The first author was supported by an RMIT VCPS.

Bibliography

[1] Benham, R., Carterette, B., Culpepper, J.S., Moffat, A.: Bayesian inferential risk evaluation on multiple IR systems. In: Proc. SIGIR. pp. 339–348 (2020)

[2] Carterette, B.: Model-based inference about IR systems. In: Proc. ICTIR. pp. 101–112 (2011)

[3] Carterette, B.: Bayesian inference for information retrieval evaluation. In: Proc. ICTIR. pp. 31–40 (2015)

[4] Collins-Thompson, K.: Reducing the risk of query expansion via robust constrained optimization. In: Proc. CIKM. pp. 837–846 (2009)

[5] Dinçer, B.T., Macdonald, C., Ounis, I.: Risk-sensitive evaluation and learning to rank using multiple baselines. In: Proc. SIGIR. pp. 483–492 (2016)

[6] Dinçer, B.T., Macdonald, C., Ounis, I.: Hypothesis testing for the risk-sensitive evaluation of retrieval systems. In: Proc. SIGIR. pp. 23–32 (2014)

[7] Dinçer, B.T., Ounis, I., Macdonald, C.: Tackling biased baselines in the risk-sensitive evaluation of retrieval systems. In: Proc. ECIR. pp. 26–38 (2014)

[8] Gelman, A.: Two simple examples for understanding posterior p-values whose distributions are far from uniform. Electron. J. Statist. 7, 2595–2602 (2013)

[9] Gelman, A., Hill, J., Yajima, M.: Why we (usually) don't have to worry about multiple comparisons. J. Res. Int. Educ. 5(2), 189–211 (2012)

[10] Gelman, A., Loken, E.: The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. Department of Statistics, Columbia University (2013)

[11] Gelman, A., Pardoe, I.: Bayesian measures of explained variance and pooling in multilevel (hierarchical) models. Technometrics 48(2), 241–251 (2006)

[12] Lambert, B.: A student's guide to Bayesian statistics. Sage (2018)

[13] Moffat, A., Zobel, J.: Rank-biased precision for measurement of retrieval effectiveness. ACM Trans. Inf. Sys. 27(1), 2.1–2.27 (2008)

[14] Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., Deng, L.: MS MARCO: A human-generated machine reading comprehension dataset. In: Proc. NIPS. pp. 96–105 (2016)

[15] Sakai, T.: Statistical significance, power, and sample sizes: A systematic review of SIGIR and TOIS, 2006–2015. In: Proc. SIGIR. pp. 5–14 (2016)

[16] Sakai, T.: The probability that your hypothesis is correct, credible intervals, and effect sizes for IR evaluation. In: Proc. SIGIR. pp. 25–34 (2017)

[17] Salton, G., Buckley, C.: Improving retrieval performance by relevance feedback. J. Am. Soc. Inf. Sci. 41(4), 288–297 (1990)

[18] Urbano, J., Nagler, T.: Stochastic simulation of test collections: Evaluation scores. In: Proc. SIGIR. pp. 695–704 (2018)

[19] Voorhees, E.: Effect on system rankings of further extending pools for TREC-COVID round 1 submissions. `https://ir.nist.gov/covidSubmit/papers/rnd1runs_j0.5-2.0.pdf` (2020)

[20] Voorhees, E., Alam, T., Bedrick, S., Demner-Fushman, D., Hersh, W.R., Lo, K., Roberts, K., Soboroff, I., Wang, L.L.: TREC-COVID: Constructing a pandemic information retrieval test collection. SIGIR Forum 54(1), 1–12 (2020)

[21] Wang, L., Bennett, P.N., Collins-Thompson, K.: Robust ranking models via risk-sensitive optimization. In: Proc. SIGIR. pp. 761–770 (2012)