BAYESIAN AGGREGATION OF CROWD JUDGMENTS FOR QUANTITATIVE FACT CHECKING

M. Lambardi di San Miniato [1] , M. Battauz[1] , R. Bellio[1] and P. Vidoni[1]

[1] Department of Economics and Statistics, University of Udine, (e-mail: [michele.lambardi, michela.battauz, ruggero.bellio, paolo.vidoni]@uniud.it)

ABSTRACT: Political fact-checking can be carried out by crowd workers, provided they are supervised by experts. We propose a Bayesian latent variable ordinal probit model for truthfulness rating data, to estimate workers' reliability, weigh in their contributions, and surrogate expert judgments. This is a notable example of aggregation function of an implicit type. This method may be used to dynamically assign workers to new tasks, as illustrated with an analysis of PolitiFact data.

KEYWORDS: Bayesian statistics, judgment aggregation, ordered probit.

1 Introduction

Fact-checking is about assessing the truthfulness of public statements to combat misinformation and improve debates. However, expert fact-checkers are few, while crowd workers are readily available but potentially biased. There is a stream of scientific research about how to surrogate expert judgements by means of workers, after some suitable calibration; see for example Roitero *et al.*, 2021. Latent traits of statements and workers are at stake, like truthfulness and political orientation. Methods from Item Response Theory (see for example Bartholomew *et al.*, 2011) can be adapted to this end. Here we adopt a Bayesian approach, which is suitable for the task.

We propose an ordinal probit model for quantitative fact-checking. An aggregation function is involved, which mimics expert judgments via the wisdom of crowds (Roitero *et al.*, 2021). The truthfulness of statements, even when encoded as an ordinal variable, is often treated as numeric. This allows to summarize ratings across workers by means of a simple average, but by using a generative model there is room for improvement. We argue that, as far as the aggregation function needs not be explicit, the Bayesian inferential approach always provides one, namely, the posterior distribution of expert judgments conditional to workers'. A different proposal, with some similarities, exists in the literature (Nguyen *et al.*, 2018).

2 Model

Let $i = 1,\dots,n$ and $j = 1,\dots,m$ be two indices to identify statements and workers, respectively. The truthfulness of statement i is rated as $Z_i = 1,\dots,k$ by a single expert and as $W_{ij} = 1,\dots,k$ by worker $j \in C_i$, with C_i a subset of workers that evaluate statement i. The aim is to predict Z_i through $(W_{ij})_{j\in C_i}$.

As typical in ordinal regression models, we think of Z_i as the observed discretization of a latent numeric variable Z_i^*, which is defined as

$$Z_i^* = \sigma_\xi\,\xi_i + \varepsilon_i\,, \quad \xi_i, \varepsilon_i \sim \mathcal{N}(0,1)\,.$$

Here, ε_i is a noise term, ξ_i is the truthfulness of the i-th statement and $\sigma_\xi > 0$ is a signal strength parameter. Analogously, we think of W_{ij} as the observed discretization of a latent numeric variable W_{ij}^*, which is defined as

$$W_{ij}^* = \alpha_j + \beta_j\xi_i + \eta_{ij}\,, \quad \alpha_j \sim \mathcal{N}(0,\sigma_\alpha^2)\,, \quad \beta_j \sim \mathcal{N}(0,\sigma_\beta^2)\,, \quad \eta_{ij} \sim \mathcal{N}(0,1)\,.$$

Here, η_{ij} is a noise term, while α_j and β_j are worker-specific parameters that affect their judging behavior. The worker-specific parameters α_j and β_j account for correlation within workers. All the terms $\xi_i, \varepsilon_i, \alpha_j, \beta_j, \eta_{ij}$ are assumed independent. Lastly, we define two sets of thresholds $(\gamma_h)_{h=0}^k$ and $(\delta_l)_{l=0}^k$ constrained as $\gamma_0 = \delta_0 = -\infty$, $\gamma_h < \gamma_{h+1}$, $\delta_l < \delta_{l+1}$, $\gamma_k = \delta_k = +\infty$, such that

$$\gamma_{h-1} < W_{ij}^* \le \gamma_h \iff W_{ij} = h\,, \quad \delta_{l-1} < Z_i^* \le \delta_l \iff Z_i = l\,.$$

Probit models are implied for Z_i and W_{ij}. As an original proposal, parameters α_j and β_j allow to represent the alignment with the experts. The model specification is then completed by assigning weakly informative priors to scale parameters and uniform priors on thresholds.

3 Example

We analyse a publicly available dataset (Roitero *et al.*, 2020), which includes expert ratings obtained from PolitiFact. Data relate to $m = 100$ workers and $n = 62$ public statements on COVID-19. The truthfulness ratings Z_i and W_{ij} have $k = 6$ levels, labeled as: "pants-on-fire", "false", "mostly-false", "half-true", "mostly-true" or "true". There were eight statements per worker and ten workers per statement, but two *gold* statements were rated by all the workers for control purposes. Gold statements have either $Z_i = 1$ or $Z_i = k$, while all

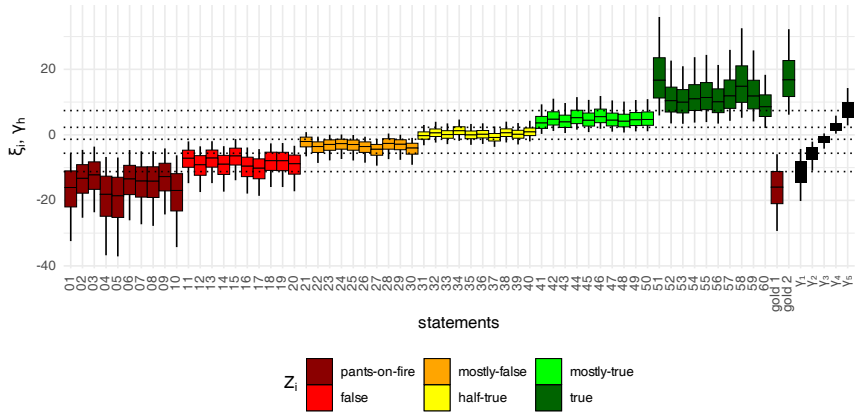

Figure 1. *Posterior percentiles (5%, 25%, 50%, 75% and 95% level) of truthfulness* ξ_i *and thresholds* γ_h.

the other statements administered to each worker cover different Z values. We estimate the model via the `R` interface to the `Stan` probabilistic programming language (Stan Development Team, 2023).

Figure 1 shows the posterior distribution of ξ_i, along with the thresholds γ_h. Were it only for the model on Z_i, the boxplots should all be similar, but the model on W_{ij} complements that information, so that there can be gradients of ξ within levels of Z.

Figure 2 summarizes the inferential results for α and β, which are affected by political orientation. Liberals tend to be more aligned with the truth (large β_j) and tend to give lower ratings (small α_j). Instead, conservatives seem more gullible (large α_j) and less aligned with the truth (small β_j). There are even two workers with negative β_j, who are detrimental on fact checking.

4 Conclusion

Our analyses support that Bayesian generative models may lead to important advances for crowd-based fact checking. Future research will focus on the usage of the model for prediction of Z given W, and on the extension to more complex settings.

Acknowledgments This work was supported by the Departmental Strategic Plan (PSD) of the University of Udine, Interdepartmental Project on Artificial

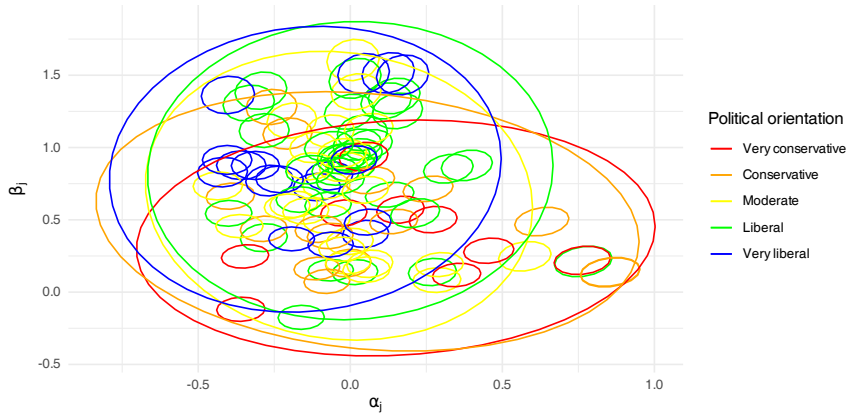

Figure 2. *Posterior normal ellipses of* α_j *and* β_j *(5%, smaller) grouped by political orientation of workers (90%, larger).*

Intelligence (2020-2025).

References

BARTHOLOMEW, D., KNOTT, M., & MOUSTAKI, I. 2011. *Latent Variable Models and Factor Analysis*. 3rd edn. Wiley.

NGUYEN, A., KHAROSEKAR, A., LEASE, M., & WALLACE, B. 2018. An interpretable joint graphical model for fact-checking from crowds. *In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence, Part II*, vol. 32. New Orleans, USA: AAAI Press.

ROITERO, K., SOPRANO, S., PORTELLI, B., SPINA, D., DELLA MEA, V., SERRA, G., MIZZARO, S., & DEMARTINI, G. 2020. The COVID-19 infodemic: Can the crowd judge recent misinformation objectively? *In: Proceedings of the 29th ACM International Conference on Information and Knowledge Management*. Virtual event, Ireland: ACM.

ROITERO, K., SOPRANO, M., PORTELLI, B., DE LUISE, M., SPINA, D., DELLA MEA, V., SERRA, G., MIZZARO, S., & DEMARTINI, G. 2021. Can the crowd judge truthfulness? A longitudinal study on recent misinformation about COVID-19. *Personal and Ubiquitous Computing*, **27**, 1–31.

STAN DEVELOPMENT TEAM. 2023. *RStan: The R interface to Stan*. R package version 2.21.8.