Overcoming Calibration Problems in Pattern Labeling with Pairwise Ratings: Application to Personality Traits

Baiyu Chen[4], Sergio Escalera[1,2,3], Isabelle Guyon[3,5], Víctor Ponce-López[1,2,6], Nihar Shah[4], and Marc Oliu Simón[6]

[1] Computer Vision Center, Campus UAB, Barcelona,
[2] Dept. Mathematics and Computer Science, University of Barcelona,
[3] ChaLearn, California, USA,
[4] University of California Berkeley, California, USA,
[5] U. ParisSaclay, France,
[6] EIMT at the Open University of Catalonia, Barcelona.

Abstract. We address the problem of calibration of workers whose task is to label patterns with continuous variables, which arises for instance in labeling images of videos of humans with continuous traits. Worker bias is particularly difficult to evaluate and correct when many workers contribute just a few labels, a situation arising typically when labeling is crowd-sourced. In the scenario of labeling short videos of people facing a camera with personality traits, we evaluate the feasibility of the pairwise ranking method to alleviate bias problems. Workers are exposed to pairs of videos at a time and must order by preference. The variable levels are reconstructed by fitting a Bradley-Terry-Luce model with maximum likelihood. This method may at first sight, seem prohibitively expensive because for N videos, $p = N(N-1)/2$ pairs must be potentially processed by workers rather that N videos. However, by performing extensive simulations, we determine an empirical law for the scaling of the number of pairs needed as a function of the number of videos in order to achieve a given accuracy of score reconstruction and show that the pairwise method is affordable. We apply the method to the labeling of a large scale dataset of 10,000 videos used in the ChaLearn Apparent Personality Trait challenge.

Keywords: Calibration of labels, Label bias, Ordinal labeling, Variance Models, Bradley-Terry-Luce model, Continuous labels, Regression, Personality traits, Crowd-sourced labels.

1 Introduction

Computer vision problems often involve labeled data with continuous values (regression problems). This includes, job interview assessments [1], personality analysis [2,3], or age estimation [4], among others. To acquire continuous labeled data, it is often necessary to hire professionals that have had training on the task

of visually examining image or video patterns. For example, the data collection that motivated this study requires the labeling of 10,000 short videos with personality traits on a scale of -5 to 5. Because of the limited availability of trained professionals, one often resorts to the "wisdom of crowds" and hire a large number of untrained workers whose proposed labels are averaged to reduce variance. A typical service frequently used for crowd-sourcing labeling is Amazon Mechanical Turk[1] (AMT). In this paper, we work on the problem of obtaining accurate labeling for continuous target variables, with time and budgetary constraints.

The variance between labels obtained by crowd-sourcing stems from several factors, including the **intrinsic** variability of labeling of a single worker (who, due to fatigue and concentration may be inconsistent with his/her own assessments), and the **bias** that a worker may have (his/her propensity to over-rate or under-rate, e.g. a given personality trait). Intrinsic variability is often referred to as "random error" while "bias" is referred to as "systematic error". The problem of intrinsic variability can be alleviated by pre-selecting workers for their consistency and by shortening labeling sessions to reduce worker fatigue. The problem of **bias reduction** is the central subjet of this paper.

Reducing bias has been tackled in various ways in the literature. Beyond simple averaging, aggregation models using confusion matrices have been considered for classification problems with binary or categorical labels (e.g [5]). Aggregating continuous labels is reminiscent of Analysis of Variance (ANOVA) models and factor analysis (see, e.g. [6]) and has been generalized with the use of factor graphs [5]. Such methods are referred to in the literature as "cardinal" methods to distinguish them from "ordinal methods", which we consider in this paper.

Ordinal methods require that workers rank patterns as opposed to rating them. Typically, a pair of patterns A and B is presented to a worker and he/she is asked to judge whether $value(A) < value(B)$, for instance $extroverted(A) < extroverted(B)$. Ordinal methods are by design immune to additive biases (at least global biases, not discriminative biases, such as gender or race bias). Because of their built-in insensitivity to global biases ordinal methods are well suited when many workers contribute each only a few labels [7]. In addition, there is a large body of literature [8–13] showing evidence that ordinal feed-back is easier to provide than cardinal feed-back from untrained workers. In preliminary experiments we conducted ourselves, workers were also more engaged and less easily bored if they had to make comparisons rather than rating single items.

In the applications we consider, however, the end goal is to obtain for every pattern a cardinal rating (such as the level of friendliness). To that end, pairwise comparisons must be converted to cardinal ratings such as to obtain the desired labels. Various models have been proposed in the literature, including the Bradley-Terry-Luce (BTL) model [14], the Thurstone class of models [15], and non-parametric models based on stochastic transitivity assumptions [16]. Such methods are commonly used, for instance, to convert tournament wins in chess to ratings and in online video games such as Microsoft's Xbox [17]. In this paper, we present experiments performed with the Bradley-Terry-Luce (BTL)

[1] https://www.mturk.com/.

model [14], which provided us with satisfactory results. By performing simulations, we demonstrate the viability of the method within the time and budget constraints of our data collection.

Contribution

For a given target accuracy of cardinal rating reconstruction, we determine the practical economical feasibility of running such a data labeling and the practical computational feasibility by running extensive numerical experiments with artificial and real sample data from the problem at hand. We investigate the advantage of our proposed method from the scalability, noise resistance, and stability points of view. We derive an empirical scaling law of the number of pairs necessary to achieve a given level of accuracy of cardinal rating reconstruction from a given number of pairs. We provide a fast implementation of the method using Newton's conjugate gradient algorithm that we make publicly available on Github. We propose a novel design for the choice of pairs based on small-world graph connectivity and experimentally prove its superiority over random selection of pairs.

2 Problem Formulation

2.1 Application Setting: The Design of a Challenge

The main focus of this research is the organization of a pattern recognition challenge in the ChaLearn Looking at People (LAP) series [18–25], which is being run for ECCV 2016 [3] and ICPR 2016 . This paper provides a methodology, which we are using in our challenge on automatic personality trait analysis from video data [26]. The automatic analysis of videos to characterize human behavior has become an area of active research with a wide range of applications [1, 2, 27, 28]. Research advances in computer vision and pattern recognition have lead to methodologies that can successfully recognize consciously executed actions, or intended movements, for instance, gestures, actions, interactions with objects and other people [29]. However, much remains to be done in characterizing sub-conscious behaviors [30], which may be exploited to reveal aptitudes or competence, hidden intentions, and personality traits. Our present research focuses on a quantitative evaluation of personality traits represented by a numerical score for a number of well established psychological traits known as the "big five" [31]: Extraversion, agreableness, conscientiousness, neurotism, and openness to experience.

Personality refers to individual differences in characteristic patterns of thinking, feeling and behaving. Characterizing personality automatically from video analysis is far from being a trivial task because perceiving personality traits is difficult even to professionally trained psychologists and recruiting specialists. Additionally, quantitatively assessing personality traits is also challenging due to the subjectivity of assessors and lack of precise metrics. We are organizing a

challenge on "first impressions", in which participants will develop solutions for recognizing personality traits of subjects from a short video sequence of the person facing the camera. This work could become very relevant to training young people to present themselves better by changing their behavior in simple ways, as the first impression made is very important in many contexts, such as job interviews.

We made available a large newly collected data set sponsored by Microsoft Research of 10,000 15-second videos collected from YouTube, annotated with the "big-five" personality traits by AMT workers. See the data collection interface in Figure 1.

We budgeted 20,000 USD for labeling the 10,000 videos. We originally estimated that by paying 10 cents per rating of video pair (a conservative estimate of cost per task), we could afford rating 200,000 pairs. This paper presents the methodology we used to evaluate whether this budget would allows us to accurately estimate the cardinal ratings, which we support by numerical experiments on artificial data. Furthermore, we investigated the computational feasibility of running maximum likelihood estimation of the BTL model for such a large number of videos. Since this methodology is general, it could be used in other contexts.

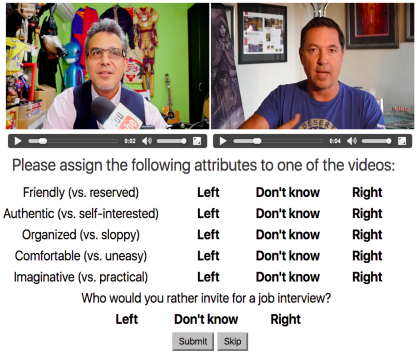

Fig. 1: Data collection interface. The AMT workers must indicate their preference for five attributes representing the "big five" personality traits.

2.2 Model Definition

Our problem is parameterized as follows. Given a collection of N videos, each video has a trait with value in $[-5, 5]$ (this range is arbitrary, other ranges can be chosen). We treat each trait separately; in what follows, we consider a single trait. We require that only p pairs will be labeled by the AMT workers out of the $P = N(N-1)/2$ possible pairs. For scaling reasons that we explain later, p is normalized by $N \log N$ to obtain parameter $\alpha = p/(N \log N)$. We

consider a model in which the ideal ranking may be corrupted by "noise", the noise representing errors made by the AMT workers (a certain parameter σ). The three parameters α, N, and σ fully characterize our experimental setting depicted in Figure 2 that we now describe.

Let $\mathbf{w}^*$ be the N dimensional vector of "true" (unknown) cardinal ratings (e.g. of videos) and $\tilde{\mathbf{w}}$ be the N dimensional vector of estimated ratings obtained from the votes of workers after applying our reconstruction method based on pairwise ratings. We consider that i is the index of a *pair* of videos $\{j, k\}$, $i = 1 : p$ and that $y_i \in \{-1, 1\}$ represents the ideal ordinal rating (+1 if $w_j^* > w_k^*$ and -1 otherwise, ignoring ties). We use the notation $\mathbf{x_i}$ to represent a special kind of indicator vector, which has value +1 at position j, −1 at position k and zero otherwise, such that $< \mathbf{x}_i, \mathbf{w}^* >= w_j^* - w_k^*$.

We formulate the problem as estimating the cardinal rating values of all videos based on p independent samples of ordinal ratings $y_i \in \{-1, 1\}$ coming from the distribution:

$$P[y_i = 1|\mathbf{x_i}, \mathbf{w}^*] = \mathbf{F}(\frac{< \mathbf{x_i}, \mathbf{w}^* >}{\sigma}),$$

where F is a known function that has value in $[0, 1]$ and σ is the noise parameter. We use Bradley-Terry-Luce model, which is a special case where F is logistic function, $F(t) = 1/(1 + exp(-t))$.

In our simulated experiments, we first draw the w_j^* cardinal ratings uniformly in [-5, 5], then we draw p pairs randomly as training data and apply noise to get the ordinal ratings y_i. As test data, we draw another set of p pairs from the remaining data.

It can be verified that the likelihood function of the BTL model is log-concave. We simply use the maximum likelihood method to estimate the cardinal rating values and get our estimation $\tilde{\mathbf{w}}$. This method should lead to a single global optimum for such a convex optimization problem.

2.3 Evaluation

To evaluate the accuracy of our cardinal rating reconstruction, we use two different scores (computed on test data):

Coefficient of Determination (R^2). We use the coefficient of determination to measure how well $\tilde{\mathbf{w}}$ reconstructs $\mathbf{w}^*$. The residual residual sum of squares is defined as $SS_{res} = \sum_i (w_i^* - \tilde{w}_i)^2$. The total sum of squares SS_{var} is defined as: $SS_{var} = \sum_i (w_i^* - \overline{w^*})^2$, where $\overline{w^*}$ denotes the average rating. The coefficient of Determination is defined as $R^2 = 1 - SS_{res}/SS_{var}$. Note that since the w_i^* are on an arbitrary scale $[-5, +5]$, we must normalize the $\tilde{w}_i$ before computing the R^2. This is achieved by finding the optimum shift and scale to maximize the R^2.

Test-accuracy. We define test Accuracy as the fraction of pairs correctly reoriented using $\tilde{\mathbf{w}}$ from the test data pairs, i.e. those pairs not used for evaluating $\tilde{\mathbf{w}}$.

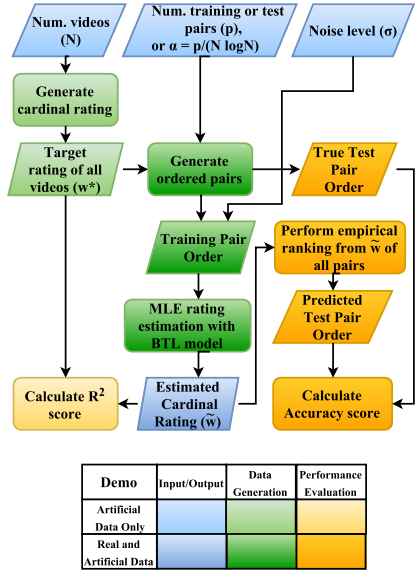

Fig. 2: Work Flow Diagram

2.4 Experiment Design

In our simulations, we follow the workflow of Figure 2. We first generate a score vector $\mathbf{w}^*$ using a uniform distribution in $[-5, 5]^N$. Once $\mathbf{w}^*$ is chosen, we select training and test pairs.

One original contribution of our paper is the choice of pairs. We propose to use a small-world graph construction method to generate the pairs [32]. Small-world graphs provide high connectivity, avoid disconnected regions in the graph, have a well distributed edges, and minimum distance between nodes [33]. An edge is selected at random from the underlying graph, and the chosen edge determines the pair of items compared. We compare the small-world strategy to draw pairs with drawing pairs at random from a uniform distribution, which according to [7] yield near-optimal results.

The ordinal rating of the pairs is generated with the BTL model using the chosen $\mathbf{w}^*$ as the underlying cardinal rating, flipping pairs according to the noise level. Finally, the maximum likelihood estimator for the BTL model is employed to estimate $\tilde{\mathbf{w}}$.

We are interested in the effect of three variables: total number of pairs available, p; total number of videos, N; noise level, σ. First we experiment on performance progress (as measured by R^2 and Accuracy on test data) for fixed values of N and σ, by varying the number of pairs p. According to [14] with no noise and error, the minimum number of pairs needed for exactly recovering of original ordering of data is $NlogN$. This prompted us to vary p as a multiple of $NlogN$. We define the parameter $\alpha = p/(N \log N)$. The results are shown in Figures 3

and 7. This allows us, for a given level of reconstruction accuracy (e.g. 0.95) or R^2 (e.g. 0.9) to determine the number of pairs needed. We then fix p and σ and observe how performance progress with N (Figures 6 and 8).

3 Results and Discussion

In this section, we examine performances in terms of test set R^2 and Accuracy for reconstructing the cardinal scores and recovering the correct pairwise ratings when noise is applied at various levels in the BTL model.

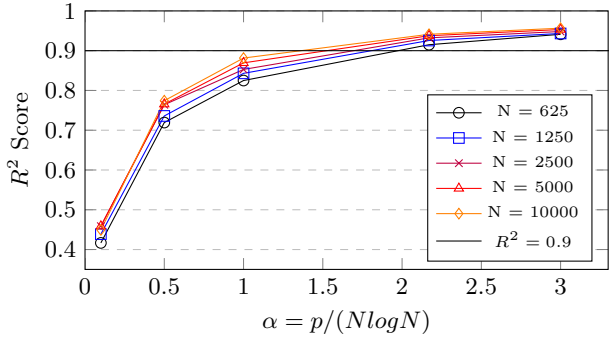

Fig. 3: Evolution of R^2 for different α with noise level $\sigma = 1$.

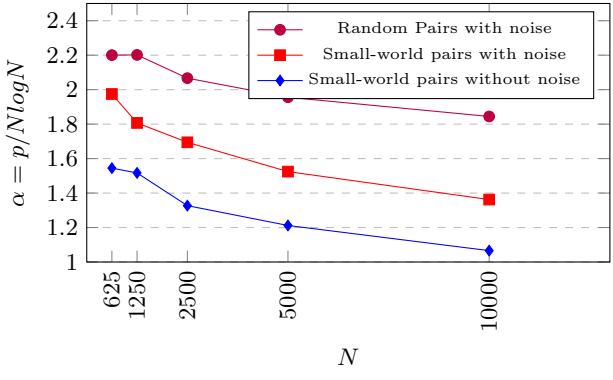

Fig. 4: Evolution of $\alpha^* : \alpha$ at $R^2 = 0.9$ for with and without noise, with $\sigma = 1$.

3.1 Number of pairs needed

We recall that one of the goals of our experiments was to figure out scaling laws for the number of pairs p as a function of N for various levels of noise. From theoretical analyses, we expected that p would scale with $NlogN$ rather than N^2. In a first set of experiments, we fixed the noise level at $\sigma = 1$. We were pleased to see in in Figures 3 and 7 that our two scores (the R^2 and Accuracy) in fact *increase* with $\alpha = p/(NlogN)$. This indicates that our presumed scaling law is, in fact, pessimistic.

To determine an empirical scaling law, we fixed a desired value of R^2 (0.9, see horizontal line in Figure 3). We then plotted the five points resulting from the intersection of the curves and the horizontal line as a function of N to obtain the red curve in Figure 4. The two other curves are shown for comparison: The blue curve is obtained without noise and the brown curve with an initialisation with the small-world heuristic. All three curves present a quasi-linear decrease of *alpha* with N with the same slope. From this we infer that $\alpha = p/(NlogN) \simeq \alpha_0 - 4 \times 10^{-5} N$. And thus we obtain the following empirical scaling law of p as a function of N:

$$p = \alpha_0 NlogN - 4 \times 10^{-5} N^2 logN.$$

In this formula, the intercept α_0 changes with the various conditions (choices of pairs and noise), but the scaling law remains the same. A similar scaling law is obtained if we use Accuracy rather than R^2 as score.

3.2 Small-world heuristic

Our experiments indicate that an increase in performance is obtained with the small-world heuristic compared to a random choice of pairs (Figure 4). This is therefore what was adopted in all other experiments.

3.3 Experiment budget

In the introduction, we indicated that our budget to pay AMT workers would cover at least $p = 200,000$ pairs. However, the efficiency of our data collection setting reduced the cost per elementary task and we ended up labeling $p = 321,684$ pairs within our budget. For our $N = 10,000$ videos, this corresponds to $\alpha = p/(NlogN) = 3.49$. We see in Figure 4 that, for $N = 10,000$ videos, in all cases examined, the α required to attain $R^2 = 0.9$ is lower than 2.17, and therefore, our budget was sufficient to obtain this level of accuracy.

Furthermore, we varied the noise level in Figures 6 and 8. In these plots, we selected a smaller value of α than what our monetary budget could afford ($\alpha = 1.56$). Even at that level, we can see that we have a sufficient number of pairs to achieve $R^2 = 0.9$ for all levels of noise considered and all values of N considered. We also achieve an accuracy near 0.95 for $N = 10,000$ for all levels of noise considered. As expected, a larger σ requires a larger number of pairs to achieve the same level of R^2 or Accuracy.

3.4 Computational time

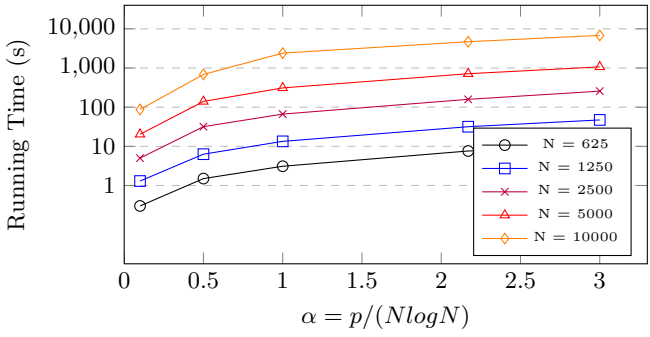

Fig. 5: Evolution of running time for different α and N with noise and $\sigma = 1$ on log scale.

One of the feasibility aspect of using ordinal ranking concerns computational time. Given that collecting and annotating data takes months of work, any computational time ranging from a few hours to a few days would be reasonnable. However, to be able to run systematic experiments, we optimized our algorithm sufficiently that any experiment we performed took less than three hours. Our implementation, which uses Newton's conjugate gradient algorithm [34], was made publicly available on Github[2]. In Figure 5 we see that the log of running time increases quite rapidly with α at the beginning and then almost linearly. We also see that the log of the running time increases linearly with N for any fixed value of α. In the case of our data collection, we were interested in $\alpha = 2.17$ (see the previous section), which corresponds to using 200,000 pairs for 10,000 videos (our original estimate). For this value of α, we were pleased to see that the calculation of the cardinal labels would take less than three hours. This comforted us on the feasibility of using this method for out particular application.

3.5 Experiments on real data

The data collection process included collecting labels from AMT workers. Each worker followed the protocol we described in Section 2 (see Figure 1). We obtained 321,684 pairs of real human votes for each trait, which were divided into 300,000 pairs for training and used the remainder 21,684 pairs for testing. This corresponds to $\alpha = 3.26$ for training.[3]

[2] `https://github.com/andrewcby/Speed-Interview`

[3] These experiments concern only cardinal label reconstruction, they have nothing to do with the pattern recognition task from the videos, for which a different split between training/validation/test sets was done for the challenge.

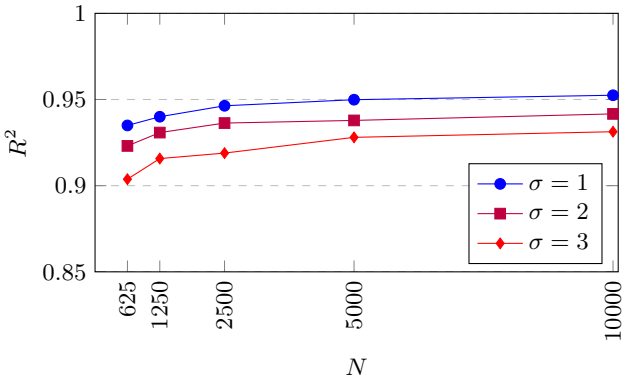

Fig. 6: Evolution of R^2 for different σ with $\alpha = 1.56$, a value that guarantees $R^2 \geq 0.9$ when $\sigma = 1$.

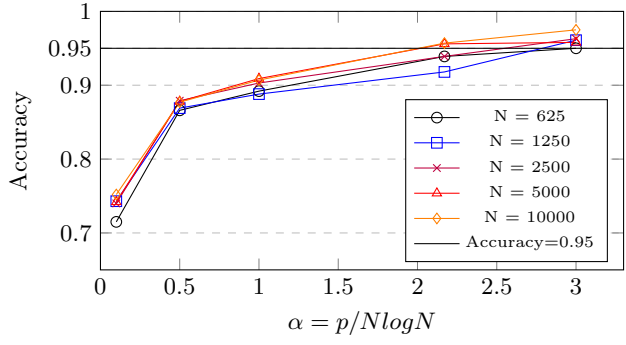

Fig. 7: Evolution of Accuracy for different α with noise with $\sigma = 1$.

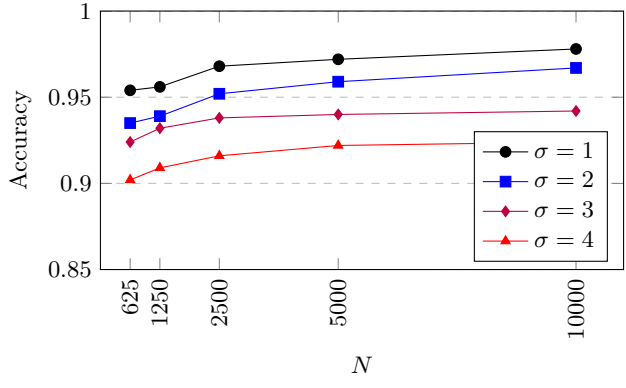

Fig. 8: Evolution of accuracy for different σ with $\alpha = 1.56$, a value that guarantees accuracy ≥ 0.9 when $\sigma = 1$.

We ran our cardinal score reconstruction algorithm on these data set and computed test accuracy. The results, shown in Table 1, give test accuracies between 0.66 and 0.73 for the various traits. Such reconstruction accuracies are significantly worse than those predicted by our simulated experiments. Looking at figure 7, the accuracies for $\alpha > 3$ are larger than 0.95.

Several factors can explain such lower accuracies of reconstruction:

1. **Use of "noisy" ground truth** estimation in real data to compute the target ranking in the accuracy calculation. The overly optimistic estimation of the accuracy in simulations stems in part from using exact ground truth, not available in real data. In real data, we compared the human ranking and the BTL model reconstructed ranking in test data. This may account for at least doubling the variance, one source of error being introduced when estimating the cardinal scores, and the other when estimating the accuracy using pair reconstruction with "noisy" real data.
2. **Departure of the real label distribution** from the uniform distribution. We carried out complementary simulations with a Gaussian distribution instead of a uniform distribution of labels (closer to a natural distribution) and observed a decrease of 6% in accuracy and a decrease of 7% in R^2.
3. **Departure of the real noise distribution from the BTL model**. We evaluated the validity of the BTL model by comparing the results to those produced with a simple baseline method introduced in [35]. This method consists in averaging the ordinal ratings for each video (counting +1 if it is rated higher than another video an -1 if it is rated lower). The performances of the BTL model are consistently better across all traits, based on the one sigma error bar calculated with 30 repeat experiments. Therefore, even though the baseline method is considerably simpler and faster, it is worth running the BTL model for the estimation of cardinal ratings. Unfortunately, there is no way to quantitatively estimate the effect of the third reason.
4. **Under-estimation of the intrinsic noise level** (random inconsistencies in rating the same video pair by the same worker). We evaluated the σ in the BTL model using bootstrap re-sampling of the video pairs. With an increasing level of σ, the results are consistently decreasing, as shown in figure 8. Therefore the parameters we chose for the simulation model proved to be optimistic and underestimated the intrinsic noise level.
5. **Sources of bias not accounted for** (we only took into account a global source of bias, not stratified sources of bias such as gender bias and racial bias. This is a voter-specific factor that we did not take into consideration when setting up the simulation. As this kind of bias is hard to measure, especially quantitatively, it can negatively influence the accuracy of the prediction.

4 Discussion and conclusion

In this paper we evaluated the viability of an ordinal rating method based on labeling pairs of videos, a method intrinsically insensitive to (global) worker bias.

Using simulations, we showed that it is in principle possible to accurately produce a cardinal rating by fitting the BTL model with maximum likelihood, using artificial data generated with this model. We calculated that it was possible

to remain within our financial budget of 200,000 pairs and incur a reasonable computational time (under 3 hours).

However, although in simulations we pushed the model to levels of noise that we thought were realistic, the performance we attained with simulations ($R^2 = 0.9$ of Accuracy$= 0.95$ on test data) turned out to be optimistic. Reconstruction of cardinal ratings from ordinal ratings on real data lead to a lower level of accuracy (in the range 69% and 73%), showing that there are still other types of noise that are not reducible by the model. Future work can focus on methods to reduce this noise.

Our financial budget and time constraints also did not allow us to conduct a comparison with direct cardinal rating. An ideal, but expensive, experiment could be to duplicate the ground truth estimation by using AMT workers to directly estimate cardinal ratings, within the same financial budget. Future work includes validating our labeling technique in this way on real data.

Acknowledgment

This work was supported in part by donations of Microsoft Research to prepare the personality trait challenge, and Spanish Projects TIN2012-38187-C03-02, TIN2013-43478-P and the European Comission Horizon 2020 granted project SEE.4C under call H2020-ICT-2015. We are grateful to Evelyne Viegas, Albert Clapés i Sintes, Hugo Jair Escalante, Ciprian Corneanu, Xavier Baró Solé, Cécile Capponi, and Stéphane Ayache for stimulating discussions. We are thankful for Prof. Alyosha Efros for his support and guidance.

Table 1: Estimation Accuracy of 10,000 videos and 321,684 pairs ($3.49 \times NlogN$).

Trait	BTL Model		Averaging ordinal ratings	
	Accuracy	STD	Accuracy	STD
Extraversion	0.692	± 0.027	0.575	± 0.095
Agreeableness	0.720	± 0.025	0.533	± 0.087
Conscientiousness	0.669	± 0.032	0.559	± 0.092
Neuroticism	0.706	± 0.022	0.549	± 0.084
Openness	0.735	± 0.021	0.542	± 0.089

References

1. Marcos-Ramiro, A., Pizarro-Perez, D., Marron-Romera, M., Nguyen, L., Gatica-Perez, D.: Body communicative cue extraction for conversational analysis. In: Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on. (April 2013) 1–8
2. Aran, O., Gatica-Perez, D.: One of a kind: Inferring personality impressions in meetings. In: Proceedings of the 15th ACM on International Conference on Multimodal Interaction. ICMI, New York, NY, USA, ACM (2013) 11–18
3. : Chalearn lap 2016: First round challenge on first impressions - dataset and results
4. Escalera, S., Gonzlez, J., Bar, X., Pardo, P., Fabian, J., Oliu, M., Escalante, H.J., Huerta, I., Guyon, I.: Chalearn looking at people 2015 new competitions: Age estimation and cultural event recognition. In: 2015 International Joint Conference on Neural Networks (IJCNN). (July 2015) 1–8
5. Venanzi, M., Guiver, J., Kazai, G., Kohli, P., Shokouhi, M.: Community-based bayesian aggregation models for crowdsourcing. In: Proceedings of the 23rd International Conference on World Wide Web. WWW '14, New York, NY, USA, ACM (2014) 155–164
6. Miller, J., Haden, P.: Statistical Analysis with The General Linear Model. (2006)
7. Shah, N., Balakrishnan, S., Bradley, J., Parekh, A., Ramchandran, K., Wainwright, M.: Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. CoRR **abs/1505.01462** (2015)
8. Whitehill, J., fan Wu, T., Bergsma, J., Movellan, J.R., Ruvolo, P.L.: Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In Bengio, Y., Schuurmans, D., Lafferty, J.D., Williams, C.K.I., Culotta, A., eds.: Advances in Neural Information Processing Systems 22. Curran Associates, Inc. (2009) 2035–2043
9. Welinder, P., Branson, S., Perona, P., Belongie, S.J.: The multidimensional wisdom of crowds. In Lafferty, J., Williams, C., Shawe-taylor, J., Zemel, R., Culotta, A., eds.: Advances in Neural Information Processing Systems 23. (2010) 2424–2432
10. Welinder, P., Perona, P.: Online crowdsourcing: rating annotators and obtaining cost-effective labels. In: In W. on Advancing Computer Vision with Humans in the Loop. (2010)
11. Raykar, V.C., Yu, S., Zhao, L.H., Valadez, G.H., Florin, C., Bogoni, L., Moy, L.: Learning from crowds. J. Mach. Learn. Res. **11** (August 2010) 1297–1322
12. Kamar, E., Hacker, S., Horvitz, E.: Combining human and machine intelligence in large-scale crowdsourcing. In: Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems - Volume 1. AAMAS '12, Richland, SC, International Foundation for Autonomous Agents and Multiagent Systems (2012) 467–474
13. Bachrach, Y., Graepel, T., Minka, T., Guiver, J.: How To Grade a Test Without Knowing the Answers — A Bayesian Graphical Model for Adaptive Crowdsourcing and Aptitude Testing. ArXiv e-prints (2012)
14. Bradley, R., Terry, M.: Rank analysis of incomplete block designs: The method of paired comparisons. Biometrika **39: 324-345.** (1952)
15. Thurstone, L.L.: A law of comparative judgment. Psychological Review **34**(4) (1927) 273
16. Shah, N.B., Balakrishnan, S., Guntuboyina, A., Wainwright, M.J.: Stochastically transitive models for pairwise comparisons: Statistical and computational issues. arXiv preprint arXiv:1510.05610 (2015)

17. Herbrich, R., Minka, T., Graepel, T.: Trueskill: A Bayesian skill rating system. Advances in Neural Information Processing Systems **19** (2007) 569
18. Escalera, S., Gonzàlez, J., Baró, X., Reyes, M., Lopés, O., Guyon, I., Athitsos, V., Escalante, H.J.: Multi-modal gesture recognition challenge 2013: Dataset and results. In: ChaLearn Multi-Modal Gesture Recognition Workshop, ICMI. (2013)
19. Escalera, S., Gonzàlez, J., Baro, X., Reyes, M., Guyon, I., Athitsos, V., Escalante, H., Argyros, A., Sminchisescu, C., Bowden, R., Sclarof, S.: Chalearn multi-modal gesture recognition 2013: grand challenge and workshop summary. ICMI (2013) 365–368
20. Escalera, S., Baro, X., Gonzàlez, J., Bautista, M., Madadi, M., Reyes, M., Ponce-López, V., Escalante, H., Shotton, J., Guyon, I.: Chalearn looking at people challenge 2014: Dataset and results. (2014)
21. Escalera, S., Gonzàlez, J., Baro, X., Pardo, P., Fabian, J., Oliu, M., Escalante, H.J., Huerta, I., Guyon, I.: Chalearn looking at people 2015 new competitions: Age estimation and cultural event recognition. In: IJCNN. (2015)
22. Baro, X., Gonzàlez, J., Fabian, J., Bautista, M., Oliu, M., Escalante, H., Guyon, I., Escalera, S.: Chalearn looking at people 2015 challenges: action spotting and cultural event recognition. In: ChaLearn LAP Workshop, CVPR. (2015)
23. Escalera, S., Fabian, J., Pardo, P., , Baró, X., Gonzàlez, J., Escalante, H., Misevic, D., Steiner, U., Guyon, I.: Chalearn looking at people 2015: Apparent age and cultural event recognition datasets and results. In: International Conference in Computer Vision, ICCVW. (2015)
24. Escalera, S., Athitsos, V., Guyon, I.: Challenges in multimodal gesture recognition. Journal on Machine Learning Research (2016)
25. Escalera, S., Gonzàlez, J., Baró, X., Shotton, J.: Special issue on multimodal human pose recovery and behavior analysis. IEEE Tans. Pattern Analysis and Machine Intelligence (2016)
26. Park, G., Schwartz, H., Eichstaedt, J., Kern, M., Stillwell, D., Kosinski, M., Ungar, L., Seligman, M.: Automatic personality assessment through social media language. Journal of Personality and Social Psychology **108** (2014) 934–952
27. Ponce-López, V., Escalera, S., Baró, X.: Multi-modal social signal analysis for predicting agreement in conversation settings. In: Proceedings of the 15th ACM on International Conference on Multimodal Interaction. ICMI, New York, NY, USA, ACM (2013) 495–502
28. Ponce-López, V., Escalera, S., Pérez, M., Janés, O., Baró, X.: Non-verbal communication analysis in victim-offender mediations. Pattern Recognition Letters **67, Part 1** (2015) 19 – 27 Cognitive Systems for Knowledge Discovery.
29. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. (June 2008) 1–8
30. Pentland, A.: Honest Signals: How They Shape Our World. The MIT Press, Massachusetts (2008)
31. Goldberg, L.: The structure of phenotypic personality traits. (1993)
32. Watts, D.J., Strogatz, S.H.: Collective dynamics of'small-world'networks. Nature **393**(6684) (1998) 409–10
33. Humphries, M., Gurney, K., Prescott, T.: The brainstem reticular formation is a small-world, not scale-free, network. Proceedings of the Royal Society of London B: Biological Sciences **273**(1585) (2006) 503–511
34. Knoll, D.A., Keyes, D.E.: Jacobian-free Newton-Krylov methods: a survey of approaches and applications. Journal of Computational Physics **193** (January 2004) 357–397

35. Shah, N.B., Wainwright, M.J.: Simple, robust and optimal ranking from pairwise comparisons. arXiv preprint arXiv:1512.08949 (2015)