Overcoming Calibration Problems in Pattern Labeling with Pairwise Ratings

Baiyu Chen
Sergio Escalera
Isabelle Guyon
Victor Ponce-Lopez
Nihar Shah
Marc Oliu Simon

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2017-194
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2017/EECS-2017-194.html

December 5, 2017

Overcoming Calibration Problems in Pattern Labeling with Pairwise Ratings

by Baiyu Chen, Sergio Escalera, Isabelle Guyon, Vıctor Ponce-Lopez, Nihar Shah, and Marc Oliu Simon

Research Project

Submitted to the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, in partial satisfaction of the requirements for the degree of **Master of Science, Plan II**.

Approval for the Report and Comprehensive Examination:

Committee:

A. Efros

Professor Efros
Research Advisor

11/28/17

(Date)

J. Guyon

Professor Guyon
Second Reader

Dec 1, 2017

(Date)

Overcoming Calibration Problems in Pattern Labeling with Pairwise Ratings

Baiyu Chen, Sergio Escalera, Isabelle Guyon, Víctor Ponce-López, Nihar Shah, and Marc Oliu Simón

***Abstract*—We address the problem of calibration of workers whose task is to label patterns with continuous variables. An example would be labeling pictures of people with apparent age. 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 patterns at a time and must just rank them with respect to the (presumed) variable level. 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 very affordable.**

***Index Terms*—Calibration of labels, Label bias, Ordinal labeling, Variance Models, ley-Terry-Luce model, Continuous labels, Regression, Personality traits, Crowd-sourced labels.**

I. Introduction

Computer vision problems often involve labeled data with continuous values (regression problems). This includes, job interview assessments [19], personality analysis [1], or age estimation [12], 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). 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 [30]). Aggregating continuous labels is reminiscent of Analysis of Variance (ANOVA) models and factor analysis (see, e.g. [20]) and has been generalized with the use of factor graphs [30]. 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 $age(A) < age(B)$. 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 [26]. In addition, there is a large body of literature [2], [16], [25], [32]–[34] 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 (the age, the level of friendliness, etc.). 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 [4], the Thurstone class of models [29], and non-parametric models based on stochastic transitivity assumptions [27]. 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 [14]. In this paper, we present experiments performed with the Bradley-Terry-Luce (BTL) model [4], which has given us satisfaction. 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

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

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. The paper will be complemented by the results of the actual data labeling at the time of submission of the final camera-ready version of the paper (if accepted), on Sep. 05, 2016.

II. Problem Formulation

A. 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 [3], [5]–[11], which we are proposing for ICPR 2016. This paper provides a methodology, which we are using in the design of our upcoming challenge on automatic personality trait analysis from video data [21]. The automatic analysis of videos to characterize human behavior has become an area of active research with a wide range of applications [1], [19], [23], [24]. 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 [18]. However, much remains to be done in characterizing sub-conscious behaviors [22], 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" [13]: 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 subjets 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 will make available a large newly collected data set sponsored by Microsoft Research of at least 10,000 15-second videos collected from YouTube, annotated with the "big-five" personality traits by AMT workers. See the data collection webpage in Figure 1.

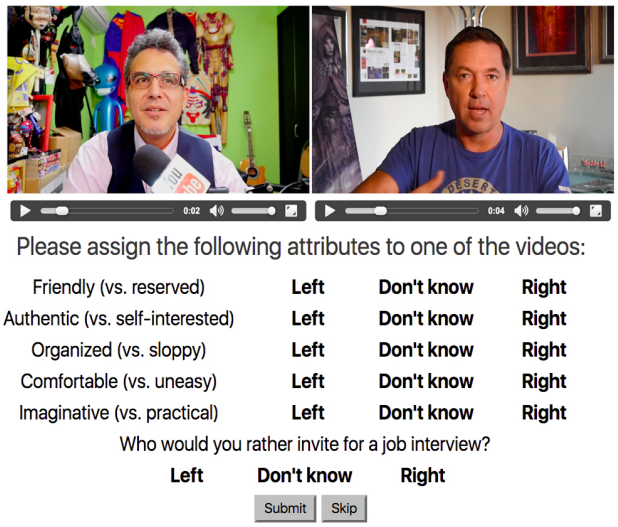

Fig. 1: Data collection webpage. The AMT workers must indicate their preference for five attributes representing the "big five" personality traits.

We budgeted 20,000 Dollars for labeling the 10,000 videos. By paying 10 cents per rating of video pair, we can afford rating 200,000 pairs. We investigate in this paper whether this budget allows us to accurately estimate the cardinal ratings. Furthermore, we investigate the computational feasibility of running maximum likelihood estimation of the BTL model for such a large number of videos.

B. 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.

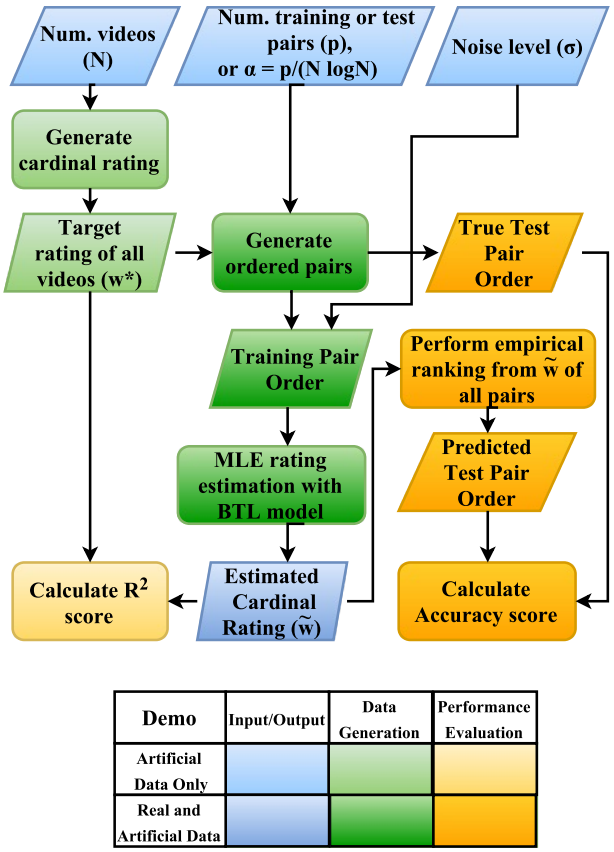

Fig. 2: Work Flow Diagram

C. 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 re-oriented using $\tilde{\mathbf{w}}$ from the test data pairs, i.e. those pairs not used for evaluating $\tilde{\mathbf{w}}$.

D. 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 [31]. Small-world graphs provide high connectivity, avoid disconnected regions in the graph, have a well distributed edges, and minimum distance between nodes [15]. 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 [26] 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 [4] 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).

III. Results and Discussion

In this section, we examine performances in terns 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.

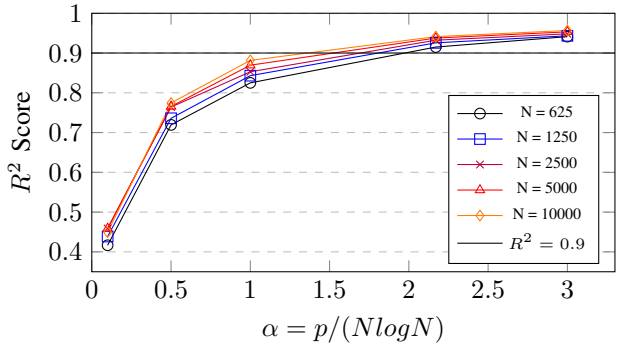

Fig. 3: Evolution of R^2 for different α with noise with $\sigma = 1$.

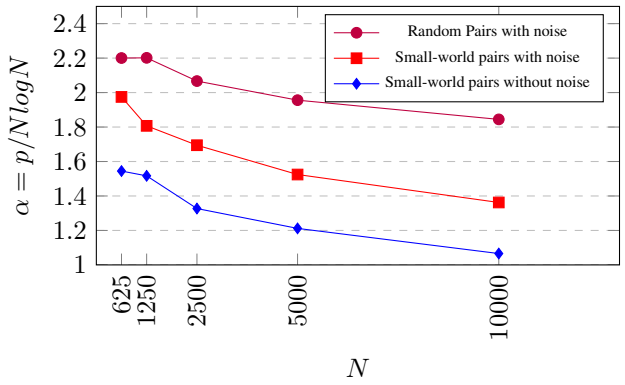

Fig. 4: Evolution of $\alpha^* : \alpha$ at $R^2 = 0.9$ for with and without noise, with $\sigma = 1$.

A. 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.

B. 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.

C. Experiment budget

In the introduction, we indicated that our budget to pay AMT workers would cover $p = 200,000$ pairs. This corresponds for $N = 10,000$ videos to $\alpha = p/(NlogN) = 2.17$. 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 is 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.

D. Computational time

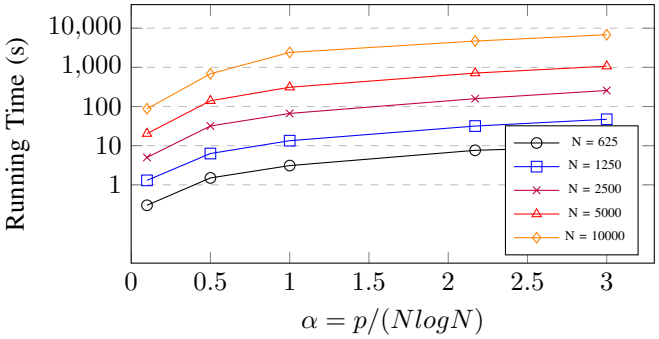

Fig. 5: Evolution of running time for different α and N with noise and $\sigma = 1$ on log scale.

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 [17], 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

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

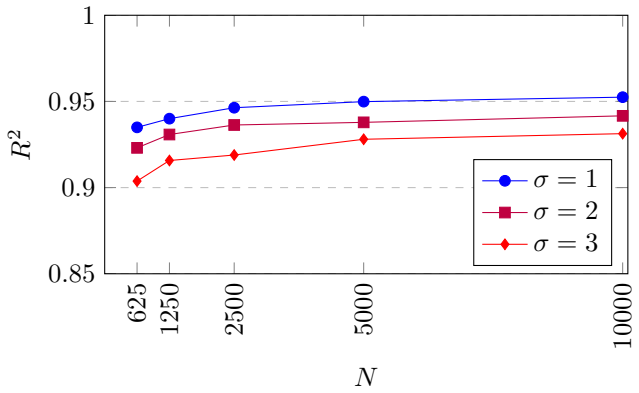

Fig. 6: Evolution of R^2 for different σ with $\alpha = 1.56$, a value that guarantees $R^2 \geq 0.9$ when $\sigma = 1$.

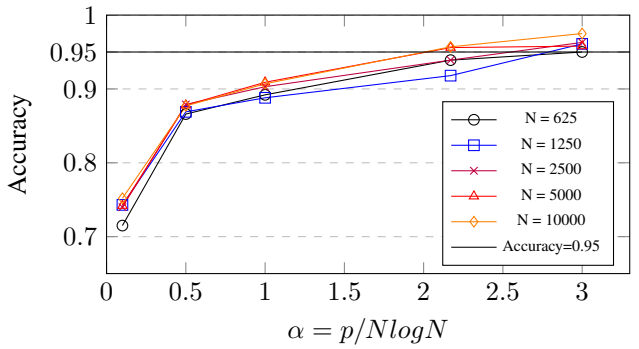

Fig. 7: Evolution of Accuracy for different α with noise with $\sigma = 1$.

of α. In the case of our data collection, we are interested in $\alpha = 2.17$ (see the previous section), which corresponds to using 200,000 pairs for 10,000 videos. For this value of α, we are pleased to see that the calculation of the cardinal labels will take less than three hours.

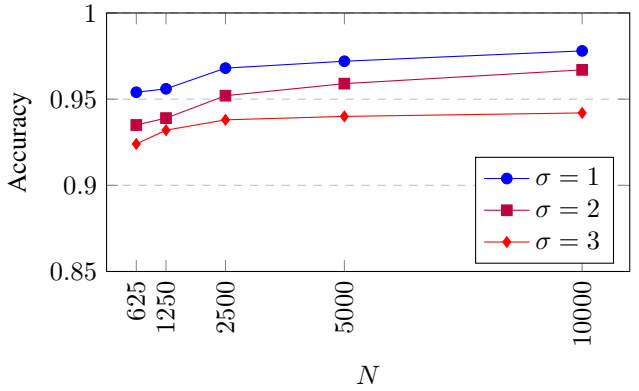

Fig. 8: Evolution of accuracy for different σ with $\alpha = 1.56$, a value that guarantees accuracy ≥ 0.9 when $\sigma = 1$.

E. Experiments on real data

The data collection process included collecting labels from AMT workers. Each worker followed the protocol we described in Section II (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 $\alpha = 0.6$.

We ran our model on this data set with respect to test accuracy. The results are shown in Table I. The results are consistently between 66% and 73% for different traits. The accuracy is not quite as good as that predicted by our simulated experiments. Looking at figure 7, the accuracy from artificial data with $\alpha = 0.6$ is just below 0.9. This may be explained by the fact that we accounted only for the global systematic bias in our model and did not take into account other types of biases (like gender bias and racial bias) as well as intrinsic noise due to the difficulty of extracting personality traits from short videos (the same viewer can inconsistently label the same video). At the same time,

For comparison, we also produced a cardinal rating with a simple baseline method introduced in [28]. 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 BLT model for the estimation of cardinal ratings.

TABLE I: Estimation Accuracy of 300000 videos and 10000 pairs ($0.6 \times NlogN$).

Trait	BTL Model		Averaging ordinal ratings	
	Accuracy	STD	Accuracy	STD
Friendly	0.692	$\pm$ 0.003	0.575	$\pm$ 0.013
Authentic	0.720	$\pm$ 0.007	0.533	$\pm$ 0.021
Organized	0.669	$\pm$ 0.004	0.559	$\pm$ 0.017
Comfortable	0.706	$\pm$ 0.004	0.549	$\pm$ 0.014
Imaginative	0.735	$\pm$ 0.003	0.542	$\pm$ 0.020

IV. Conclusion

In this paper we evaluated the viability of an ordinal rating method based on labeling pairs of videos. We showed that it is possible to accurately produce a cardinal rating by fitting the BTL model with maximum likelihood, using artificial data generated with this model. By simulation, we pushed the model to levels of noise realistic in the real world and showed that we attain $R^2 = 0.9$ of Accuracy$= 0.95$ (on test data), while remaining within our budget of 200,000 pairs and a reasonnable computational time (under 3 hours). Our method is therefore a viable way to label data without being sensitive to (global) worker bias. Experiments on real sample data lead to a lower level of accuracy (in the range 66% and 73%), showing that other types of noise are not reducible by the model.

ACKNOWLEDGEMENT

This work was supported in part by donations of Microsoft Research to prepare the personality trait challenge, and Spanish Projects TIN2012-38187-C03-02 and TIN2013-43478-P. 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.

REFERENCES

[1] O. Aran and D. Gatica-Perez. One of a kind: Inferring personality impressions in meetings. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, ICMI, pages 11–18, New York, NY, USA, 2013. ACM.

[2] Y. Bachrach, T. Graepel, T. Minka, and J. Guiver. How To Grade a Test Without Knowing the Answers — A Bayesian Graphical Model for Adaptive Crowdsourcing and Aptitude Testing. *ArXiv e-prints*, 2012.

[3] X. Baro, J. Gonzàlez, J. Fabian, M.A. Bautista, M. Oliu, H.J. Escalante, I. Guyon, and S. Escalera. Chalearn looking at people 2015 challenges: action spotting and cultural event recognition. In *ChaLearn LAP Workshop, CVPR*, 2015.

[4] R.A. Bradley and M.E. Terry. Rank analysis of incomplete block designs: The method of paired comparisons. *Biometrika*, 39: 324-345., 1952.

[5] S. Escalera, V. Athitsos, and I. Guyon. Challenges in multimodal gesture recognition. *Journal on Machine Learning Research*, 2016.

[6] S. Escalera, X. Baro, J. Gonzàlez, M. Bautista, M. Madadi, M. Reyes, V. Ponce-López, H. Escalante, J. Shotton, and I. Guyon. Chalearn looking at people challenge 2014: Dataset and results. *ChaLearn LAP Workshop, ECCV*, 2014.

[7] S. Escalera, J. Fabian, P. Pardo, , X. Baró, J. Gonzàlez, H.J. Escalante, D. Misevic, U. Steiner, and I. Guyon. Chalearn looking at people 2015: Apparent age and cultural event recognition datasets and results. In *International Conference in Computer Vision, ICCVW*, 2015.

[8] S. Escalera, J. Gonzàlez, X. Baro, P. Pardo, J. Fabian, M. Oliu, H.-J. Escalante, I. Huerta, and I. Guyon. Chalearn looking at people 2015 new competitions: Age estimation and cultural event recognition. In *IJCNN*, 2015.

[9] S. Escalera, J. Gonzàlez, X. Baro, M. Reyes, I. Guyon, V. Athitsos, HJ Escalante, A. Argyros, C. Sminchisescu, R. Bowden, and S. Sclarof. Chalearn multi-modal gesture recognition 2013: grand challenge and workshop summary. *ICMI*, pages 365–368, 2013.

[10] S. Escalera, J. Gonzàlez, X. Baró, M. Reyes, O. Lopés, I. Guyon, V. Athitsos, and H.-J. Escalante. Multi-modal gesture recognition challenge 2013: Dataset and results. In *ChaLearn Multi-Modal Gesture Recognition Workshop, ICMI*, 2013.

[11] S. Escalera, J. Gonzàlez, X. Baró, and J. Shotton. Special issue on multimodal human pose recovery and behavior analysis. *IEEE Tans. Pattern Analysis and Machine Intelligence*, 2016.

[12] S. Escalera, J. Gonzlez, X. Bar, P. Pardo, J. Fabian, M. Oliu, H. J. Escalante, I. Huerta, and I. Guyon. Chalearn looking at people 2015 new competitions: Age estimation and cultural event recognition. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, July 2015.

[13] L.R. Goldberg. The structure of phenotypic personality traits., 1993.

[14] Ralf Herbrich, Tom Minka, and Thore Graepel. Trueskill: A Bayesian skill rating system. *Advances in Neural Information Processing Systems*, 19:569, 2007.

[15] M.D Humphries, K Gurney, and T.J Prescott. The brainstem reticular formation is a small-world, not scale-free, network. *Proceedings of the Royal Society of London B: Biological Sciences*, 273(1585):503–511, 2006.

[16] Ece Kamar, Severin Hacker, and Eric Horvitz. 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, pages 467–474, Richland, SC, 2012. International Foundation for Autonomous Agents and Multiagent Systems.

[17] D. A. Knoll and D. E. Keyes. Jacobian-free Newton-Krylov methods: a survey of approaches and applications. *Journal of Computational Physics*, 193:357–397, January 2004.

[18] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, June 2008.

[19] A. Marcos-Ramiro, D. Pizarro-Perez, M. Marron-Romera, L. Nguyen, and D. Gatica-Perez. Body communicative cue extraction for conversational analysis. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–8, April 2013.

[20] Jeff Miller and Patricia Haden. *Statistical Analysis with The General Linear Model*. 2006.

[21] G. Park, H.A. Schwartz, J.C. Eichstaedt, M.L. Kern, D.J. Stillwell, M. Kosinski, L.H. Ungar, and M.EP. Seligman. Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, 108:934–952, 2014.

[22] A.S. Pentland. *Honest Signals: How They Shape Our World*. The MIT Press, Massachusetts, 2008.

[23] V. Ponce-López, S. Escalera, and X. Baró. Multi-modal social signal analysis for predicting agreement in conversation settings. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, ICMI, pages 495–502, New York, NY, USA, 2013. ACM.

[24] V. Ponce-López, S. Escalera, M. Pérez, O. Janés, and X. Baró. Non-verbal communication analysis in victim-offender mediations. *Pattern Recognition Letters*, 67, Part 1:19 – 27, 2015. Cognitive Systems for Knowledge Discovery.

[25] Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning from crowds. *J. Mach. Learn. Res.*, 11:1297–1322, August 2010.

[26] N.B. Shah, S. Balakrishnan, J.K. Bradley, A. Parekh, K. Ramchandran, and M.J. Wainwright. Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. *CoRR*, abs/1505.01462, 2015.

[27] Nihar B Shah, Sivaraman Balakrishnan, Adityanand Guntuboyina, and Martin J Wainwright. Stochastically transitive models for pairwise comparisons: Statistical and computational issues. *arXiv preprint arXiv:1510.05610*, 2015.

[28] Nihar B Shah and Martin J Wainwright. Simple, robust and optimal ranking from pairwise comparisons. *arXiv preprint arXiv:1512.08949*, 2015.

[29] Louis L Thurstone. A law of comparative judgment. *Psychological Review*, 34(4):273, 1927.

[30] M. Venanzi, J. Guiver, G. Kazai, P. Kohli, and M. Shokouhi. Community-based bayesian aggregation models for crowdsourcing. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 155–164, New York, NY, USA, 2014. ACM.

[31] D. J. Watts and S. H. Strogatz. Collective dynamics of'small-world'networks. *Nature*, 393(6684):409–10, 1998.

[32] Peter Welinder, Steve Branson, Pietro Perona, and Serge J. Belongie. The multidimensional wisdom of crowds. In J. Lafferty, C. Williams, J. Shawe-taylor, R.s. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 2424–2432. 2010.

[33] Peter Welinder and Pietro Perona. Online crowdsourcing: rating annotators and obtaining cost-effective labels. In *In W. on Advancing Computer Vision with Humans in the Loop*, 2010.

[34] Jacob Whitehill, Ting fan Wu, Jacob Bergsma, Javier R. Movellan, and Paul L. Ruvolo. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 2035–2043. Curran Associates, Inc., 2009.