Modeling Annotator Behaviors for Crowd Labeling

Yunus Emre Kara[a,*], Gaye Genc[a], Oya Aran[b], Lale Akarun[a]

[a]*Department of Computer Engineering, Bogazici University, TR-34342 Bebek, Istanbul, Turkey*
[b]*Idiap Research Institute, Martigny, Switzerland*

Abstract

Machine learning applications can benefit greatly from vast amounts of data, provided that reliable labels are available. Mobilizing crowds to annotate the unlabeled data is a common solution. Although the labels provided by the crowd are subjective and noisy, the wisdom of crowds can be captured by a variety of techniques. Finding the mean or the median of a sample's annotations are widely used approaches for finding the consensus label of that sample. Improving consensus extraction from noisy labels is a very popular topic, the main focus being binary label data. In this paper, we focus on crowd consensus estimation of continuous labels, which is also adaptable to ordinal or binary labels. Our approach is designed to work on situations where there is no gold standard; it is only dependent on the annotations and not on the feature vectors of the instances, and does not require a training phase. For achieving a better consensus, we investigate different annotator behaviors and incorporate them into four novel Bayesian models. Moreover, we introduce a new metric to examine annotator quality, which can be used for finding good annotators to enhance consensus quality and reduce crowd labeling costs. The results show that the proposed models outperform the commonly used methods. With the use of our annotator scoring mechanism, we are able to sustain consensus quality with much fewer annotations.

*Corresponding author
Email addresses: yunus.kara@boun.edu.tr (Yunus Emre Kara), gaye.genc@boun.edu.tr (Gaye Genc), oaran@idiap.ch (Oya Aran), akarun@boun.edu.tr (Lale Akarun)

Keywords: annotation, label noise, multiple annotators, crowdsourcing, crowd labeling, annotator behavior, annotator quality, consensus

1. Introduction

In 1906, statistician Francis Galton observed a contest held in a fair; on estimating the weight of a slaughtered and dressed ox. He calculated that the median guess of 787 people was 1207 pounds which is within 0.8% of the true weight of 1198 pounds [1]. This experiment broke new ground in cognitive science; establishing the notion that opinions of a crowd on a particular subject can be represented by a probability distribution. This is what we today call the wisdom of crowds. A crowd can be any group of people, such as the students of a school, or even the general public. In daily life, when we lack knowledge about a certain concept we inquire those around us to obtain a general idea. A similar approach can also be adapted to scientific research where it is not feasible or possible to observe the phenomenon directly.

Employing the power of a crowd for a task is called crowdsourcing. Many applications in crowdsourcing exist such as fundraising, asking for people to vote their appreciation of movies and books, or dividing up and parallelizing complex tasks to be completed. The microwork concept deals with breaking up very large problem that may or may not be solved by computers. Amazon Mechanical Turk [2] and Crowdflower [3] are examples of microwork platforms where people submit lots of small tasks to be completed by other people all around the world, for a fee.

Ground truth labeling is often considered to be a menial task and consumes the valuable time of researchers acquiring datasets. For labeling tasks that do not require expert opinion, many research centers and universities prefer paying a group of people from the general population for ground truth annotation.

Assume that we have N samples and R annotators where each annotator annotates a randomized subset of N samples and every sample is annotated by a group of annotators. This is a common case for crowdsourced annotation

tasks. The aim of our work is to obtain consensus labels for each sample using these annotations.

In this paper, we focus on modeling annotator behavior and incorporating it in four new Bayesian models that we propose for the crowd labeling problem. The models we propose are designed particularly for continuous or ordinal scores, but could be applied to categorical scores as well. Our method is specifically designed for problems where there is no gold standard and we do not include a training step in our approach. We also provide a new annotator scoring mechanism, which may be used to weed out low quality annotators and reduce crowd labeling costs.

We start by addressing related work in the literature in Section 1.1 and emphasizing our contributions in Section 1.2. We investigate annotator behaviors by explaining various annotator types in Section 2. Then, we present the proposed Bayesian models in Section 3, which are used for simultaneously modeling the behaviors of annotators and finding consensus for each sample. Section 4 describes the measure we propose for scoring the competence of annotators. Since crowdsourced labeling is an expensive process, choosing good annotators is crucial for reducing the costs. That makes annotator competence scoring an important aspect of our work. In Section 5, we present the results of our experiments for evaluating our models. The experiments are performed on two crowdsourced datasets, with and without ground truth information. Finally, we conclude the work in Section 6, with possible future directions.

1.1. Related work

An annotation task completed by crowdsourcing contains vast information along with many interesting challenges. Annotators come from different backgrounds, their experiences vary, and they provide opinions over a large scale. An in-depth survey by Frenay et al. [4] focuses on defining label noise and its sources, and introduces a taxonomy on the types of label noise. Potential drawbacks and related solutions are discussed, including algorithms which are label noise-tolerant, label noise cleansing, and label noise-robust. Srivastava et al.

investigate the problem of subjective video annotation and majority opinion is shown to be the most objective annotation for a video [5]. Carpenter[6] utilizes multilevel Bayesian approaches on binary data annotations, and introduce priors on sensitivity and specificity of annotators. Singular opinions of the annotators are unreliable, but the consensus of the crowd provides a strong insight. Finding a reasonable consensus among the annotators is very important, especially in cases where the ground truth (or gold standard) does not exist. Raykar et al. estimate the gold standard and measure the competence of the annotators iteratively in a probabilistic approach [7]. Their results are challenged by Rodrigues et al. in a supervised multiclass classification problem with a simpler probabilistic model [8]. Ground truth estimation is done by annotator modeling by using the annotators' self-reported confidences in [9]. Human personality trait evaluation is also a problem where no quantifiable ground truth exists. Trait annotations collected by crowdsourcing are used in [10] for personality trait classification.

The problem of annotator reliability is a very popular subject and tackled in [11] by using Gaussian mixture models. Liu et al. approach this problem by using belief propagation and mean field methods [12]. Statistical methods are used for estimating annotator reliability and behavior [13], as well as including annotator parameters such as bias, expertise, and competence [14]. Both approaches group annotator behaviors into different 'schools of thought'. Deciding on annotator reliability is also accomplished by measuring annotator quality. Wu et al. propose a probabilistic model of active learning with multiple noisy oracles together with the oracles' labeling quality [15]. Dutta et al. also deal with annotator quality in a crowdsourcing case study where the multiple annotators provide high level categories for newspaper articles [16]. Donmez et al. introduce a new algorithm based on Interval Estimation for estimating the accuracy of multiple noisy annotators and select the best ones for active learning [17].

Annotators' varying expertise both among themselves and over different parts of the data are also factors affecting their reliability. Zhang et al. in-

vestigate annotator expertise with a combination of ML and MAP estimation [18]. An online learning algorithm weeding out unreliable annotators and asking for labels from reliable annotators for instances which have been poorly labeled has been introduced in [19]. Varying annotator expertise problems are also handled in [20] and [21] with ground truth estimation, using MAP estimation and EM approach. Whitehill et al. also study annotator expertise, taking noisy and adversarial annotators into account [22].

Detecting spammers/abusers, and biased annotators is also useful for eliminating and/or modifying specific annotations. Spectral decomposition techniques are used for moderating abusive content in [23]. Raykar et al. propose an empirical Bayesian algorithm for iteratively eliminating spammers and estimating consensus labels from good annotators [24]. Wauthier et al. present a new Bayesian model for reducing annotator bias to combine the data collection, data curation and active learning [25].

1.2. Novelty and contributions

A straightforward solution for the continuous annotation case might be taking the mean or median of annotations for each sample. For the binary case, majority voting is the first solution that comes to mind. However, a few problems arise with these approaches, such as:

- Annotator errors and outliers have a high impact on the consensus,
- Valuable information on annotator behavior and expertise is disregarded.

Investigating the behaviors of annotators and modeling their aspects would prove useful for utilizing valuable information.

The methods in the literature that we mentioned are mostly designed for binary labeled input [6, 7, 8, 14, 18, 21, 24, 26]. However, in many annotation problems, researchers request continuous or ordinal annotations and map the annotations to binary labels. An example of this is the heart wall segment level ratings where trained cardiologists are asked to rate the samples in the interval 1-5, but the input annotations are binarized as normal (1) and abnormal (2-5)

[21, 26]. Unfortunately, this binarization process results in the loss of valuable information.

Another approach is to use ordinal annotations, as if they were categories, as input to the categorical models [27, 28]. Although it is possible to employ these types of models for ordinal labels, the categorical approach falls short of preserving the ordinal and proportional relations. For continuous or ordinal annotations, it is better to employ models that make use of ordinal and proportional information.

Numerous methods also make use of features extracted from data [7, 18, 29]. In the case where feature extraction is not possible or feasible, methods such as ours can be used. Moreover, the success of data dependent methods relies heavily on the quality of extracted features. In addition, model performance across different types of problems requiring different types of features is unpredictable.

There are only a handful of works focused on ordinal or continuous annotations. Raykar et al.[7] combined sample classification with label consensus estimation. In addition, they also propose a simple data-independent model for continuous labels. Lakshminarayanan and Teh [30] focus on ordinal labels. They incorporate task difficulty to the discretization of continuous latent variables in their model. These works are pioneering elements in the continuous crowd labeling problems. However, to the best of our knowledge, our work is the first attempt to investigate the effect of diverse annotator behaviors on consensus estimation and annotator scoring mechanism for continuous crowd labeling problems.

The contributions of this study can be summarized as follows:

- We propose four new Bayesian models that model annotator behaviors for continuous or ordinal annotations to estimate the consensus scores. The proposed methods do not require any training step and are particularly designed for problems where there is no ground truth available. As a result, they are suitable to the problems where the ground truth is not available by construct, i.e. subjective annotations of human behavior. We

believe that this is the first work that incorporates numerous annotator behaviors in consensus estimation for continuous crowd labeling problems.

- We show that the consensus scores estimated by the proposed models can be converted to categorical scores using simple techniques such as thresholding. As an example, we use the binary output case and used thresholding for the binarization of continuous consensus values (i.e. model output). The experiments that we perform shows that the binarized consensus scores produced by the proposed models has higher accuracy in comparison to the state of the art techniques that are specifically designed for binary scores.

- We provide a new annotator scoring mechanism, which assigns a score to each annotator, representing the annotation quality of that annotator. This score can be used to select high quality annotators for a given task to decrease annotation cost and time. We show that the proposed annotator score successfully selects good annotators, and the consensus scores estimated using selected annotators has lower error.

- We compare the models with state-of-the art methods in the literature and report the results of our experiments on two datasets:

  - We introduce a new crowdsourced annotation dataset based on the FGNet Aging Database [31]. Although a training set with ground truth labels is not required for our methods, the existence of ground truth labels in FGNet enables us to validate our results. Obtaining reasonable consensus scores with crowd labeling tasks is especially important in problems where the ground truth does not exist (i.e. unquantifiable or subjective).

  - Our second dataset contains subjective annotations of personality impressions. Due to their highly subjective character, a ground truth for personality impressions does not exist. We produce consensus labels using our models for the personality impressions annotations

presented in [32]. The analysis of the annotator models on the personality impressions data, where there is no ground truth, is performed through the performance of the regression and classification models for predicting the personality traits trained using different consensus scores estimated by different models.

2. Annotator behaviors

Different annotator behaviors have been observed in crowdsourced tasks and discussed in several papers on analyzing crowdsourcing systems and on annotator modeling. The reasons behind these different annotator behaviors are various. While some of these behaviors are due to the level of expertise of the annotators, some may occur due to low-attention/low-concentration on the task, and some behaviors are observed due to the bad intent of the annotators. For example there are spammers [24], dishonest annotators [33] or annotators who try to game the system [34] by providing unrelated or nonsense answers. In [24], the annotators' behaviors such as biased or malicious annotators are also discussed.

We wish to understand the behavior and expertise of annotators for reaching a common annotation (consensus) for each sample. Some basic annotator types can be

- **Competent:** Low error rate
- **Spammers:** Random annotations
- **Adversaries:** Give inverted rates
- **Positively biased:** Tend to give higher rates
- **Negatively biased:** Tend to give lower rates
- **Unary annotators:** Give a single rate to all samples
- **Binary annotators:** Give rates at the opposite ends of the scale

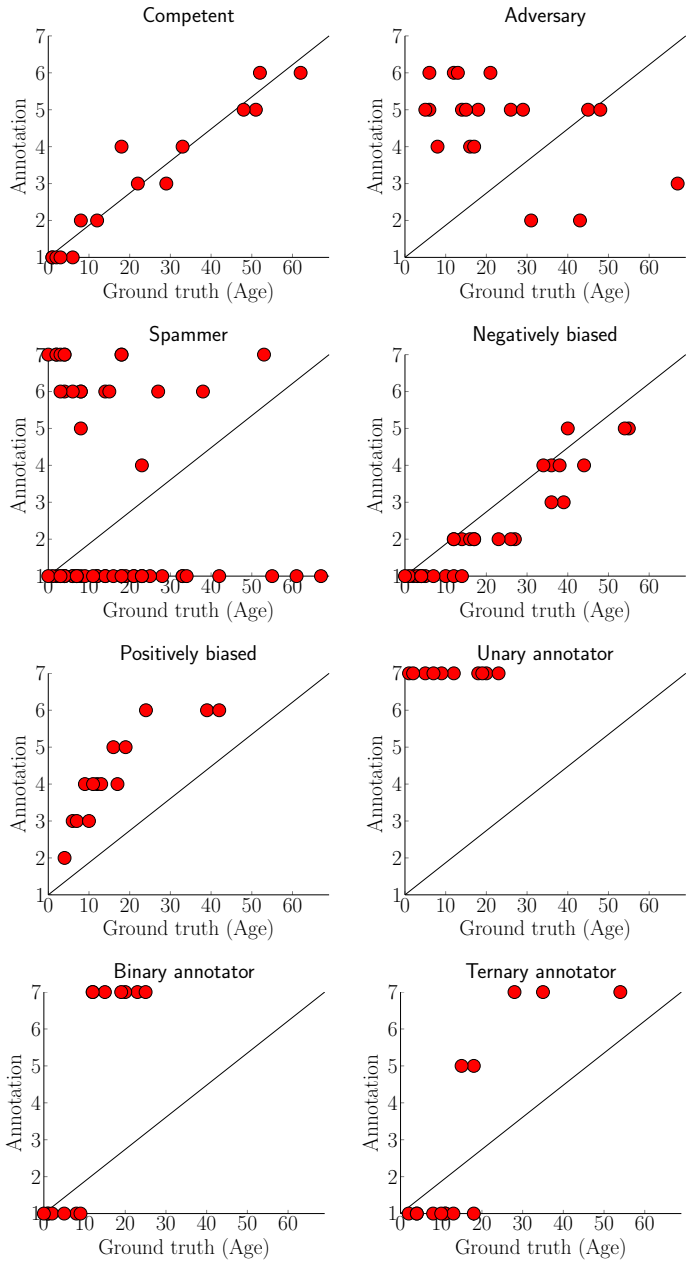


Figure 1: Real annotator examples. Each graph presents all annotations of a single annotator.

- **Ternary annotators:** Give low, mid, and high ratings

These annotator types need not be mutually exclusive; an annotator may be a combination of these types. We want to model the common behaviors of these annotator types. If we infer an annotator's behavior, we can utilize this information for our benefit. For instance, we can use competent annotators' annotations as is, we can ignore spammers, and invert the annotations of adversaries. Figure 1 shows real annotations produced by annotators of different annotator types. The dataset from which these examples have been drawn will be described in Section 5.1.1. Note that, we are not trying to classify annotator types, we incorporate the behaviors of the annotator types for designing better models.

3. Proposed Bayesian models

In our approach, we want to cover as many diverse annotator behaviors as possible. We introduce two major annotator tendencies. The first one, which we call annotator bias, explains the main behavior of positively and negatively biased annotators. Additionally, each annotator may tend to describe a similar set of samples in a wider/narrower range of rates. We call this second diversity the opinion scale.

Now, we propose four new models which handle various annotator behaviors. We assume that every sample has a single true rate (x) and an annotator tries to assign a rate (y) as a function of the unknown true rate ($\mu_\theta(x)$). The behaviors of the annotators are incorporated into our models via the annotator parameters (θ). Our models share a similar characteristic in the way that each annotation is a Gaussian random variable such that

$$\mathcal{N}(y; \mu_\theta(x), \sigma_\theta^2) \tag{1}$$

where y represents the annotation value, x is the true rate, $\mu_\theta(\cdot)$ is the annotator function, and σ_θ represents noise. x has a flat prior and the priors of the annotator parameters will be introduced with our models.

We use maximum a posteriori estimation for inferring the model parameters:

$$\mathcal{L} = \log p(Y|X,\theta) + \log p(\theta) + \log p(X) \tag{2}$$

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}}\{\mathcal{L}\} \tag{3}$$

where $\mathcal{L}$ is the log likelihood, Y are the annotations, X are the true labels, and $\theta = \{\theta_1, \ldots, \theta_N\}$ are the annotator parameters. The solution is obtained by solving

$$\frac{\partial \mathcal{L}}{\partial \theta_i} = 0, \forall \theta_i \in \theta. \tag{4}$$

The consensus rates are simultaneously inferred with the annotator parameters:

$$\frac{\partial \mathcal{L}}{\partial x_i} = 0, \forall x_i \in X. \tag{5}$$

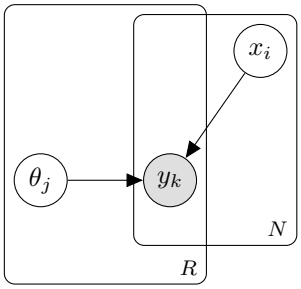


Figure 2: Bayesian network for proposed models

3.1. Model 1: Adversary handling model (M-AH)

Raykar et al.proposed a model for continuous annotation problems [7]. Their model uses features from the data in addition to the annotations. They have also adapted the same algorithm for obtaining consensus without features. Since we don't use features in our work, the latter version is more suitable for comparing with our models. This adapted version assumes that an annotator labels a sample with a rate around its true value and every annotator has a variance parameter of their own. This model does not deal with annotator behaviors.

As mentioned before, there might be some adversary annotators in crowd labeling tasks. In this model, we add adversary handling to Raykar et al.'s

model. Along with an annotator's annotation variance, we find whether the annotator is an adversary or not.

For simplicity, we assume that the annotations are zero centered in our models. For instance, if the annotators are asked to annotate between 1 and 7, we shift those annotations to the range -3 to 3. Table 1 presents model variables and parameters for all models.

Table 1: Model variables and parameters

Variable	**Description**
y_k	Value of the k^{th} annotation
i_k	Sample index of the k^{th} annotation
j_k	Annotator index of the k^{th} annotation
x_i	Consensus value of the i^{th} sample
θ_j	Parameters of the j^{th} annotator
N	Number of samples
R	Number of annotators
K	Number of annotations
N_j	Annotation count of the j^{th} annotator
Y	$\{y_{1:K}\}$
X	$\{x_{1:N}\}$
θ	$\{\theta_{1:R}\}$

We model the annotations as instances generated by a Normal distribution with the mean as the consensus x_i for that sample and variance $\sigma^2_{\theta_j} = \frac{1}{\lambda_j}$. We choose a Gamma prior for the parameter λ_j, which is a conjugate prior to the Normal distribution. It is suitable for our problem, since we want our model to fit the data well, but not too well to prevent overfitting. The prior on λ_j is

$$\lambda_j \sim \mathcal{G}\left(\lambda_j; \alpha_\lambda, \beta_\lambda\right). \tag{6}$$

We chose the hyperparameters $\alpha_\lambda = 1.2$ and $\beta_\lambda = 0.9$ since we want λ_js (which are related to noise) to be small. However, we also want them to be a bit larger

than 0, since it is evident that no annotation task is noiseless.

We want to invert the annotation if the annotator is an adversary. For the Normal distribution, inverting the mean is equivalent to inverting the value of the random variable. Thus, we set the mean parameter as $\mu_{\theta_j}(x_i) = a_j x_i$ where a_j represents the adversariness of the j^{th} annotator. If the annotator is an adversary a_j takes the value -1, if not it takes the value 1. The parameters of this model are $\theta = \{\Lambda, A\}$, where $\Lambda = \{\lambda_{1:R}\}$ and $A = \{a_{1:R}\}$. We choose a flat prior on A. Then the model is

$$\begin{aligned} p(Y, X, \theta) &= \prod_{k=1}^{K} p(y_k | x_{i_k}, \lambda_{j_k}, a_{j_k}) \prod_{j=1}^{R} p(\lambda_j) p(a_j) \prod_{i=1}^{N} p(x_i) \\ &\propto \prod_{k=1}^{K} \mathcal{N}\left(y_k; a_{j_k} x_{i_k}, \frac{1}{\lambda_j}\right) \prod_{j=1}^{R} \mathcal{G}\left(\lambda_j; \alpha_\lambda, \beta_\lambda\right) \end{aligned} \tag{7}$$

which leads to the update equations

$$x_i = \frac{\sum_{k:i_k=i} \lambda_{j_k} a_{j_k} y_k}{\sum_{k:i_k=i} \lambda_{j_k}}, \tag{8}$$

$$a_j = \operatorname{sgn}\left(\sum_{k:j_k=j} y_k x_{i_k}\right), \tag{9}$$

$$\lambda_j = \frac{N_j}{\sum_{k:j_k=j} \left(y_k - a_j x_{i_k}\right)^2}. \tag{10}$$

Note that, $a_j = \dfrac{1}{a_j}$ and $a_j^2 = 1$ for all a_j, since $a_j \in \{-1, 1\}$. The update equations are simplified using these equalities.

3.2. Model 2: Scale handling model (M-SH)

In addition to adversary handling of M-AH, we introduce opinion scale handling in this model. Some annotators tend to give rates in a wider or narrower range with respect to the ground truth. The opinion scale is represented by w. We incorporate this behavior into the model by setting the model mean as $\mu_{\theta_j}(x_i) = a_j w_j x_i$. We assume the annotators generally have a standard opinion scale, so we want to favor w being close to 1. Thus, we want to select a

distribution having 1 as its mode. As the prior for w, we select the Gamma distribution. The prior on w is

$$w_j \sim \mathcal{G}\left(w_j; \beta_w + 1, \beta_w\right) \tag{11}$$

whose hyperparameters satisfy the mode of the distribution being equal to 1. We chose $\beta_w = 4$ so that the variance of this Gamma distribution is large enough not to overconstrain w_j and small enough to favor values around 1.

The parameters of the model are $\theta = \{\Lambda, A, W\}$, where $W = \{w_{1:R}\}$. Then, we have

$$\begin{aligned} p(Y, X, \theta) &= \prod_{k=1}^{K} p(y_k | x_{i_k}, \lambda_{j_k}, a_{j_k}, w_{j_k}) \prod_{j=1}^{R} p(\lambda_j) p(a_j) p(w_j) \prod_{i=1}^{N} p(x_i) \\ &\propto \prod_{k=1}^{K} \mathcal{N}\left(y_k; a_{j_k} w_{j_k} x_{i_k}, \frac{1}{\lambda_j}\right) \prod_{j=1}^{R} \mathcal{G}\left(\lambda_j; \alpha_\lambda, \beta_\lambda\right) \\ &\quad \prod_{j=1}^{R} \mathcal{G}\left(w_j; \beta_w + 1, \beta_w\right). \end{aligned} \tag{12}$$

Then, the update equations are formulated as

$$x_i = \frac{\sum_{k:i_k=i} \lambda_{j_k} w_{j_k} a_{j_k} y_k}{\sum_{k:i_k=i} \lambda_{j_k} w_{j_k}^2}, \tag{13}$$

$$a_j = \operatorname{sgn}\left(\sum_{k:j_k=j} y_k x_{i_k}\right), \tag{14}$$

$$\lambda_j = \frac{N_j}{\sum_{k:j_k=j} \left(y_k - a_j w_j x_{i_k}\right)^2}, \tag{15}$$

$$\begin{aligned} &w_j \text{ satisfies } V_2 w_j^2 + V_1 w_j + V_0 = 0 \text{ where} \\ &\quad V_0 = -\beta_w, \\ &\quad V_1 = \beta_w - \lambda_j a_j \sum_{k:j_k=j} y_k x_{i_k}, \\ &\quad V_2 = \lambda_j \sum_{k:j_k=j} x_{i_k}^2. \end{aligned} \tag{16}$$

Out of the solutions of Equation 16, the root maximizing the posterior is selected.

3.3. Model 3: Annotation bias sensitive model (M-ABS)

In this model, we incorporate annotation bias into M-SH. This is the bias which is added after scaling and has an unscaled effect on the annotation. We incorporate this behavior into the model by setting the model mean as $\mu_{\theta_j}(x_i) = a_j(w_j x_i + t_j)$ where t_j represents either positive or negative bias. Since we model the bias as being unaffected by the opinion scale, t_j is not multiplied by w_j. Moreover, we desire the prior of negative and positive bias to be symmetrical. Thus, we find the Normal distribution suitable for our needs, resulting in the prior

$$t_j \sim \mathcal{N}\left(t_j; \mu_T, s_T^2\right). \tag{17}$$

We want the mode of the bias distribution to be at 0. We favor unbiased annotators. However, a consistent annotator with very low noise and a slight bias would be dismissed by having too much noise if the bias parameter is strictly constrained at 0. We set its standard deviation $s_T = 0.05$ to allow some positive and negative bias.

The parameters for this model are $\theta = \{\Lambda, A, W, T\}$, where $T = \{t_{1:R}\}$ and the model is defined as

$$\begin{aligned} p(Y, X, \theta) &= \prod_{k=1}^{K} p(y_k | x_{i_k}, \lambda_{j_k}, a_{j_k}, w_{j_k}, t_{j_k}) \prod_{j=1}^{R} p(\lambda_j) p(a_j) p(w_j) p(t_j) \prod_{i=1}^{N} p(x_i) \\ &\propto \prod_{k=1}^{K} \mathcal{N}\left(y_k; a_{j_k}(w_{j_k} x_{i_k} + t_{j_k}), \frac{1}{\lambda_j}\right) \prod_{j=1}^{R} \mathcal{G}\left(\lambda_j; \alpha_\lambda, \beta_\lambda\right) \\ &\quad \prod_{j=1}^{R} \mathcal{G}\left(w_j; \beta_w + 1, \beta_w\right) \prod_{j=1}^{R} \mathcal{N}\left(t_j; \mu_T, s_T^2\right) \end{aligned} \tag{18}$$

which yields the following update equations

$$x_i = \frac{\sum_{k:i_k=i} \lambda_{j_k} w_{j_k} (a_{j_k} y_k - t_{j_k})}{\sum_{k:i_k=i} \lambda_{j_k} w_{j_k}^2}, \tag{19}$$

$$a_j = \operatorname{sgn}\left(\sum_{k:j_k=j} y_k (w_j x_{i_k} + t_j)\right), \tag{20}$$

$$\lambda_j = \frac{N_j}{\sum_{k:j_k=j} \left(y_k - a_j(w_j x_{i_k} + t_j)\right)^2}, \tag{21}$$

$$t_j = \frac{a_j \sum_{k:j_k=j} y_k - w_j \sum_{k:j_k=j} x_{i_k} + \frac{\mu_T}{\lambda_j s_T^2}}{N_j + \frac{1}{\lambda_j s_T^2}}, \tag{22}$$

w_j satisfies $V_2 w_j^2 + V_1 w_j + V_0 = 0$ where

$$\begin{aligned} V_0 &= -\beta_w, \\ V_1 &= \beta_w - \lambda_j \sum_{k:j_k=j} (a_j y_k - t_j) x_{i_k}, \\ V_2 &= \lambda_j \sum_{k:j_k=j} x_{i_k}^2. \end{aligned} \tag{23}$$

Out of the solutions of Equation 23, the root maximizing the posterior is selected.

3.4. Model 4: Consensus bias sensitive model (M-CBS)

In this model, we incorporate consensus bias into M-SH. This is the bias which is affected by the annotator's scaling parameter. Since we model the bias as being affected by the opinion scale, t_j is multiplied by w_j in contrast to M-ABS. We incorporate this bias behavior into the model via setting the model mean as $\mu_{\theta_j}(x_i) = a_j w_j (x_i + t_j)$. The prior on t_j is the same as in M-ABS. In this model, we also assume that the noise introduced by an annotator is affected by their opinion scale. We achieve this effect by scaling the standard deviation of the model with the parameter w_j, resulting in the variance $\sigma_{\theta_j}^2 = \frac{w_j^2}{\lambda_j}$. The parameters are again $\theta = \{\Lambda, A, W, T\}$. Thus, we have

$$\begin{aligned} p(Y, X, \theta) &= \prod_{k=1}^{K} p(y_k | x_{i_k}, \lambda_{j_k}, a_{j_k}, w_{j_k}, t_{j_k}) \prod_{j=1}^{R} p(\lambda_j) p(a_j) p(w_j) p(t_j) \prod_{i=1}^{N} p(x_i) \\ &\propto \prod_{k=1}^{K} \mathcal{N}\left(y_k; a_{j_k} w_{j_k} (x_{i_k} + t_{j_k}), \frac{w_{j_k}^2}{\lambda_{j_k}}\right) \prod_{j=1}^{R} \mathcal{G}\left(\lambda_j; \alpha_\lambda, \beta_\lambda\right) \\ &\quad \prod_{j=1}^{R} \mathcal{G}\left(w_j; \beta_w + 1, \beta_w\right) \prod_{j=1}^{R} \mathcal{N}\left(t_j; \mu_T, s_T^2\right). \end{aligned} \tag{24}$$

The update equations are calculated as

$$x_i = \frac{\sum_{k:i_k=i} \lambda_{j_k} \left(\frac{a_{j_k} y_k}{w_{j_k}} - t_{j_k} \right)}{\sum_{k:i_k=i} \lambda_{j_k}}, \tag{25}$$

$$a_j = \text{sgn} \left(\sum_{k:j_k=j} y_k (x_{i_k} + t_j) \right), \tag{26}$$

$$\lambda_j = \frac{N_j}{\sum_{k:j_k=j} \left(\frac{y_k}{w_j} - a_j (x_i + t_j) \right)^2}, \tag{27}$$

$$t_j = \frac{\frac{a_j}{w_j} \sum_{k:j_k=j} y_k - \sum_{k:j_k=j} x_{i_k} + \frac{\mu_T}{\lambda_j s_T^2}}{N_j + \frac{1}{\lambda_j s_T^2}}, \tag{28}$$

$$\begin{aligned} &w_j \text{ satisfies } V_3 \left(\frac{1}{w_j} \right)^3 + V_2 \left(\frac{1}{w_j} \right)^2 + V_1 \left(\frac{1}{w_j} \right) + V_0 = 0 \text{ where} \\ &V_0 = -\beta_w, \\ &V_1 = \beta_w - N_j, \\ &V_2 = -\lambda_j a_j \sum_{k:j_k=j} y_k (x_{i_k} + t_j), \\ &V_3 = \lambda_j \sum_{k:j_k=j} y_k^2. \end{aligned} \tag{29}$$

Out of the solutions of Equation 29, the root maximizing the posterior is selected.

4. Annotator competence scoring

So far, we have proposed novel Bayesian models with the purpose of extracting more reliable consensus from annotations via incorporating annotator behaviors. Unfortunately, some people try to abuse the crowdsourcing system for easy money. The results are either random annotations that do not provide any solid information or ill-intentioned/absent-minded annotators marking the opposite of what they think. Naturally, one would expect to achieve a better consensus with more annotations. However, increasing annotations will also

increase costs. Due to these challenges, an annotator scoring mechanism is beneficial for both improving consensus quality and reducing annotation costs by weeding out low quality annotators. Throughout this paper we have been interested in using a group of annotators to infer the label of a sample. Using the annotator scoring mechanism to select individually good performing annotators will help us increase the crowd performance.

Now, we derive an annotator scoring function using the annotator parameters that we introduced in our models. The annotator score that we define is the sum of the joint probabilities of all possible annotations that can be produced by an annotator and the most probable originating label for those annotations given the annotator parameters. In Equation 1, we defined $\mu_\theta(\cdot)$ as the annotator function and in Table 2, we show these functions for each of our models.

Suppose that we have annotations of only a single annotator in our dataset. Although it is not the case in real annotation scenarios, let us also suppose that we are given the parameters θ of this annotator (Normally, we would infer these parameters using our models.) Given an annotation y of this annotator, we can use the inverse of the annotator function and try to obtain the originating label x. Because of σ_θ, the obtained value $\mu_\theta^{-1}(y)$ may not be equal to the originating label x. However, we can calculate the probability that the obtained value is indeed the true label as $p(x = \mu_\theta^{-1}(y)|y, \theta)$. This probability defines the accuracy of obtaining the original label of a given sample using only a single annotator. By incorporating the probability $p(y|\theta)$ of encountering the sample of interest, we obtain the joint probability of x and y conditioned on θ:

$$\begin{aligned} p(x = \mu_\theta^{-1}(y), y|\theta) &= p(x = \mu_\theta^{-1}(y)|y, \theta)p(y|\theta) \\ &= p(y|x = \mu_\theta^{-1}(y), \theta)p(x = \mu_\theta^{-1}(y)) \\ &= \mathcal{N}(y; \mu_\theta(\mu_\theta^{-1}(y)), \sigma_\theta^2)\frac{1}{2c} \\ &= \frac{1}{2c\sigma_\theta\sqrt{2\pi}} \end{aligned} \tag{30}$$

where $x \in [-c, c]$ and $p(x) = \frac{1}{2c}$ since it is flat. c is a problem specific constant for defining the annotation range. Recall that, we also shift annotations to fit

in the $[-c, c]$ range, as we explained in Section 3.1. Therefore, we have the following constraints:

$$-c \leq y \leq c,$$
$$-c \leq x = \mu_\theta^{-1}(y) \leq c.$$

For all of our models, $\mu_\theta(x)$ is monotonically increasing if and only if $a_\theta = 1$, and monotonically decreasing if and only if $a_\theta = -1$. Thus, we have

$$-c \leq \mu_\theta^{-1}(y) \leq c \implies \begin{cases} \mu_\theta(-c) \leq y \leq \mu_\theta(c), & \text{if } a_\theta = 1 \\ \mu_\theta(-c) \geq y \geq \mu_\theta(c), & \text{if } a_\theta = -1 \end{cases}$$
$$\implies a_\theta \mu_\theta(-c) \leq a_\theta y \leq a_\theta \mu_\theta(c). \tag{31}$$

By symmetry, we also have

$$-c \leq y \leq c \implies -c \leq a_\theta y \leq c. \tag{32}$$

From Inequalities 31 and 32, we have

$$\underbrace{\min\{c, \max\{a_\theta \mu_\theta(-c), -c\}\}}_{d_\theta} \leq \underbrace{a_\theta y}_{r} \leq \underbrace{\max\{-c, \min\{a_\theta \mu_\theta(c), c\}\}}_{e_\theta}. \tag{33}$$

Note that, $r = a_\theta y \implies y = \frac{r}{a_\theta} \implies y = a_\theta r$, since $a_\theta = \frac{1}{a_\theta}$, $\forall a_\theta \in \{-1, 1\}$.

We can define a path for the tuple $(x = \mu_\theta^{-1}(y), y)$ on the joint distribution as follows

$$\begin{aligned} l : [d_\theta, e_\theta] &\to \mathbb{R}^2 \\ r \mapsto (x(r), y(r)) &\implies r \mapsto (\mu_\theta^{-1}(a_\theta r), a_\theta r) \end{aligned} \tag{34}$$

We are interested in this path since it contains all possible annotations y that can be produced by an annotator, coupled with the estimations $\mu_\theta^{-1}(y)$ for the originating labels.

We define the annotator score $S(\theta)$ as the sum of the joint probabilities along

the path l:

$$\begin{aligned} S(\theta) &= \int_l p(x, y|\theta) ds \\ &= \int_{d_{\mu_\theta}}^{e_{\mu_\theta}} p(\mu_\theta^{-1}(a_\theta r), a_\theta r|\theta) \|l'(r)\| dr, \\ &= \frac{1}{2c\sigma_\theta\sqrt{2\pi}} \int_{d_{\mu_\theta}}^{e_{\mu_\theta}} \|l'(r)\| dr. \end{aligned} \tag{35}$$

Figure 3 portrays the annotator behavior deduced from the θ parameters for a selected annotator. This figure is provided for visualizing the annotator score calculation and its sub-elements. Brighter areas indicate a higher probability $p(y|x, \theta)$. For example, an annotation $y = 0$ most possibly originated from $x = -0.25$ with the probability $p(y|x, \theta) = 0.30902$. The originating label being anything other than -0.25 is still possible, but less probable. The score is the sum of these probabilities along the red line for every possible y.

Table 2 shows the derived annotator score formulas for the proposed models. Note that, d_{μ_θ} and e_{μ_θ} depend on the related model's $\mu_\theta(\cdot)$ function and their definition is given in Equation 33. It is also notable to mention that $S(\theta)$ does not depend on annotations or samples; it only depends on the parameters of the annotator.

Table 2: Annotator score formulas for the proposed models

Model	$\boldsymbol{\mu_\theta}$	$\boldsymbol{\sigma_\theta^2}$	$\boldsymbol{\|l'(r)\|}$	$\boldsymbol{S(\theta)}$
M-AH	ax	$\frac{1}{\lambda}$	$\sqrt{2}$	$\sqrt{\frac{\lambda}{\pi}}\,(e_{\mu_\theta} - d_{\mu_\theta})$
M-SH	awx	$\frac{1}{\lambda}$	$\sqrt{1 + \frac{1}{w^2}}$	$\frac{1}{w}\sqrt{\frac{\lambda(1+w^2)}{2\pi}}\,(e_{\mu_\theta} - d_{\mu_\theta})$
M-ABS	$awx + t$	$\frac{1}{\lambda}$	$\sqrt{1 + \frac{1}{w^2}}$	$\frac{1}{w}\sqrt{\frac{\lambda(1+w^2)}{2\pi}}\,(e_{\mu_\theta} - d_{\mu_\theta})$
M-CBS	$aw(x + t)$	$\frac{w^2}{\lambda}$	$\sqrt{1 + \frac{1}{w^2}}$	$\frac{1}{w^2}\sqrt{\frac{\lambda(1+w^2)}{2\pi}}\,(e_{\mu_\theta} - d_{\mu_\theta})$

In Figure 4, we demonstrate the change in annotator scores using the for-

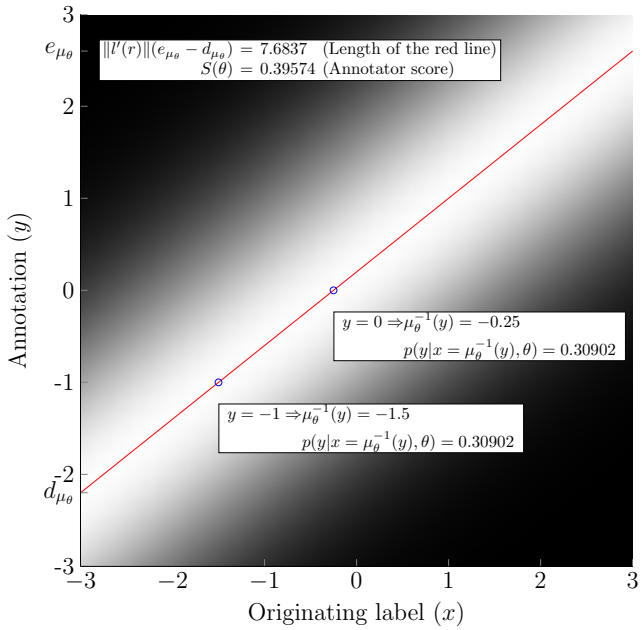


Figure 3: Score calculation for an annotator with parameters $a = 1, w = 0.8, t = 0.2, \lambda = 0.6$. The annotator is modeled using M-ABS. The intensity values depict the probability of the annotator rating a sample with respect to the ground truth. Brighter areas indicate a higher probability. This means the annotator will operate close to, but around the red line.

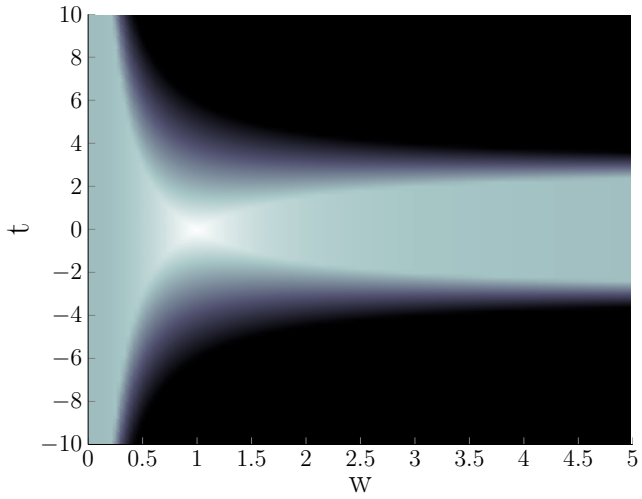


Figure 4: The change in annotator scores with respect to w and t parameters of M-CBS when the variance is fixed. Higher intensities correspond to higher annotator scores.

mulas for M-CBS with respect to w and t when the variance is fixed. When selecting our priors, we preferred w to be around 1 and t to be around 0. By examining Figure 4, we can observe that our scoring mechanism reflects our constraints successfully. When w is very small, it means that the annotator is giving rates in a narrow range providing very little to no information. If an annotator marks every sample with the same rate, it does not matter which rate they give. In this case, the effect of t diminishes and the annotator scores do not vary for different t. In the case where w is large, the annotator rates the samples whose ground truths are similar to each other in a very wide range. This is an unwanted behavior and even if the annotator is unbiased, their score will not be high since their annotations easily deviate under the smallest of changes.

5. Experimental Validation

In this section, we first evaluate the performance of our models on an annotation dataset with ground truth. We show how accurately the consensus values found by our models estimate the ground truth. Additionally, we discuss the effect of the annotators on consensus values.

In the second part, we use our models' consensus values for creating training and test scores/labels for a regression and a binary classification task and compare the performance of the trained regression and classification models, with respect to the model that is used to produce consensus scores.

5.1. Results on real data with ground truth

5.1.1. Collecting crowdsourced data with ground truth

For evaluating our models properly, we needed an annotation dataset with ground truth. We have decided to use a dataset of face images which also has the ground truth age information of the subjects in the pictures. We found the FGNet Aging Database [31] suitable for our needs. The dataset consists of a total of 1002 pictures from 82 subjects. The age range of the dataset is 0–69. Figure 5 shows some samples from this dataset and Figure 6 shows the

age histogram of the dataset. The dataset consists mostly of baby, child, and young adult photos.

Figure 5: Sample images from the FGNet Aging Database

For the annotation task, we prepared a questionnaire in which we show a facial picture and ask the annotator to rate the age of the person in the picture. The annotators are asked to rate the age from 1 to 7 where a lower rate means young and a higher rate means old. We used CrowdFlower [3] for collecting the annotation data and executed two sets of data collection. In the first set, a task for an annotator consisted of 10 annotations which means that the annotators were asked to annotate a batch of 10 images. However, if they desired they could annotate more than one batch. In the second set, a batch consisted of 15 annotations. In both sets, we set the system up to collect 5 annotations per sample. Table 3 shows annotation counts for these two sets and their joint set. The table describes the frequency of annotators' annotations. For example, there are 208 annotators in Set 1 that have provided 10 annotations and there are 292 annotators in Set 2 that have provided 15 annotations. It can be seen that not all of the annotation counts per annotator are multiples of 10 or 15. This is because the system decides to collect fewer annotations when the '5 annotations per sample' criterion is met.

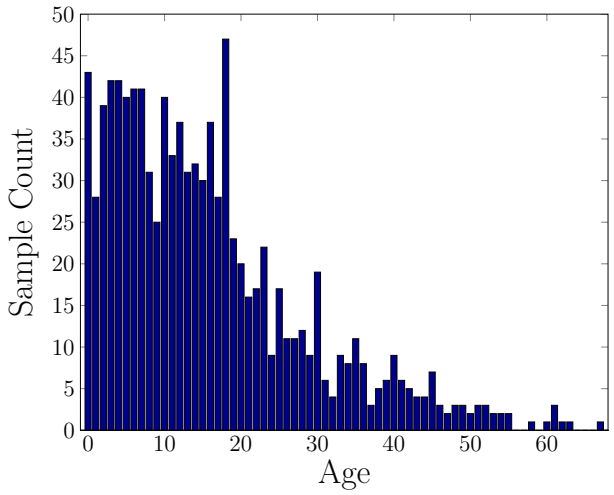


Figure 6: The FGNet Aging Database Age Histogram

Table 3: Annotator workload (the number of annotations made by an annotator)

Annotator workload	Number of annotators			Annotator workload	Number of annotators		
	Set 1	Set 2	Joint		Set 1	Set 2	Joint
1	2	4	6	29	1	1	2
6	0	1	1	30	26	12	38
7	1	0	1	31	1	0	1
9	2	0	2	33	0	1	1
10	208	0	208	36	1	0	1
11	1	0	1	40	5	0	5
14	1	0	1	42	0	1	1
15	0	292	292	43	0	1	1
16	0	1	1	45	0	1	1
19	1	0	1	50	3	0	3
20	82	0	82	59	0	1	1

5.1.2. How accurately do the models estimate ground truth?

In order to evaluate the estimation accuracy of our models, we compare the estimated consensus values against the ground truth. However, since the consensus values are in the range of 1 to 7, we need to rescale them to be compatible with the ground truth values.

The error metrics that we use in this work are Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), which are defined as follows:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|g(x_i) - z_i| \tag{36}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(g(x_i) - z_i)^2} \tag{37}$$

where z_i is the ground truth value of the i^{th} sample and $g(\cdot)$ is the linear scaling function from the consensus domain to the ground truth range. Different types of problems may require different scaling approaches. However, for the age mapping problem linear scaling is simple and intuitive. Since we map the consensus values $[1, 7]$ to the ground truth value range $[0, 69]$, the unit of error is in years of age. Note that, because of the discretization process even if the consensus values were exactly the same as the discretized ground truth labels, the error would not be zero. We call this error *the baseline error* for this dataset.

In order to compare the performance of the models among themselves, we conduct one-tailed paired-t tests with significance level $\alpha = 0.05$ for every model pair. We repeated each experiment 100 times, each time starting with randomly initialized parameters in accordance with their prior distributions. By repeating the experiments 100 times, we show that the initial parameter values (drawn from their prior distributions) do not affect the convergence of the results. The results showed us that the statistically significant order of performance is:

$$\text{Mean} < \text{Median} < \text{Raykar[7]} < \text{M-AH} < \text{M-SH} = \text{M-ABS} < \text{M-CBS}.$$

The tests between M-SH and M-ABS are inconclusive.

Table 4 shows mean errors and standard deviations for the proposed and reference models. M-CBS outperforms all other models for all sets. Simpler

Table 4: Errors on Set 1, Set 2 and the joint set. The results are presented as mean and standard deviation for 100 repetitions.

(a) Mean absolute error (baseline=2.92)

Model	**Set 1**	**Set 2**	**Joint**
Mean	9.68	8.95	8.91
Median	8.34	7.94	7.39
Raykar[7]	7.20 ± 0.048	6.94 ± 0.062	6.46 ± 0.019
M-AH	6.59 ± 0.002	6.35 ± 0.001	6.06 ± 0.000
M-SH	6.06 ± 0.112	6.04 ± 0.098	5.56 ± 0.087
M-ABS	6.07 ± 0.116	6.04 ± 0.103	5.58 ± 0.083
M-CBS	**5.91 ± 0.011**	**5.84 ± 0.006**	**5.36 ± 0.008**

(b) Root mean square error (baseline=3.40)

Model	**Set 1**	**Set 2**	**Joint**
Mean	12.10	11.50	10.90
Median	10.92	10.55	9.58
Raykar[7]	9.57 ± 0.052	9.18 ± 0.073	8.52 ± 0.020
M-AH	8.71 ± 0.003	8.49 ± 0.001	8.04 ± 0.000
M-SH	8.54 ± 0.146	8.37 ± 0.128	7.68 ± 0.100
M-ABS	8.55 ± 0.150	8.40 ± 0.134	7.70 ± 0.101
M-CBS	**8.35 ± 0.016**	**8.13 ± 0.010**	**7.50 ± 0.010**

models are prone to errors arising from outliers. Since the median model is more robust to outliers than the mean model, it performs slightly better. However, in the case of crowd labeling where lots of outliers are expected, the median model also fails to perform successfully.

The results of Set 2 are slightly better than that of Set 1. The reason for this might be that, the second set of annotators rated the samples in batches of 15 rather than 10, or they just might be more competent. Note that the proposed models do not make any assumptions on the number of samples that each annotator should annotate. However, the more annotations we gather from an annotator, the more about the annotator's behavior we can learn. One would expect a better modeling when there are more annotations from an annotator. Further examination of this phenomenon is beyond the scope of this study and is left as a future work.

The best performance is achieved in the joint set. Remember that, each sample is annotated by 5 annotators in Sets 1 and 2, which results in 10 annotations per sample in the joint set. Having more annotations per sample decreases the effect of incompetent annotators and helps to achieve better consensus values. When we investigate the samples with high error, we observe that most annotators actually do have an agreement. However, this agreement is very different from the ground truth. This is due to the fact that some samples are actually very hard to annotate where the subjects in question look much younger or older than their real age.

In Figure 7, we show the cumulative match curves (CMC) of the models. The y coordinate of a point on the CMC is the ratio of the samples that have less error than the related x coordinate. If we are interested in the consensus being in the 5–year vicinity of the ground truth, we fix the x coordinate at 5 and observe the y coordinate values of each model. 59.88% of the sample consensus values obtained with M-CBS fall within the 5–year error range of the ground truth values. When we observe the curves, Models 2, 3, and 4 perform very similarly in terms of maximum absolute age error, with M-CBS being marginally better.

Figure 8 shows the models' ground truth estimation performances of each

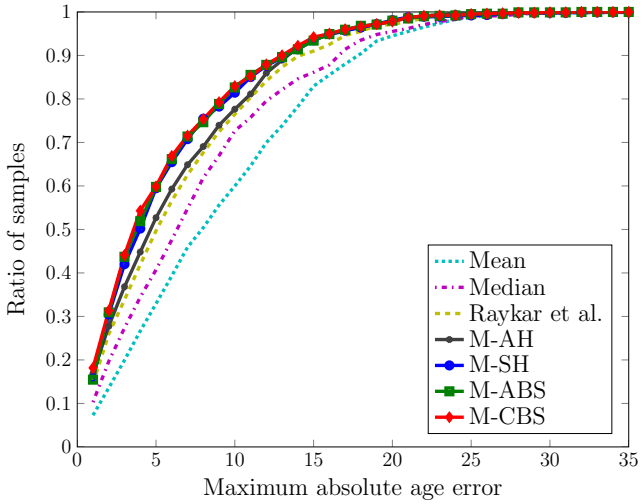


Figure 7: Cumulative match curves for the models.

sample for the joint set. As we can see, the annotations by themselves contain a huge amount of noise and do not fit to the ideal line. Using even the simplest of models allows us to reach an acceptable consensus with respect to the ground truth. We observe that the mean model has a tendency to contain more noise around the ideal line, especially in the 0–20 range. Observing Raykar et al.'s model, we see that it has characteristics belonging to both the mean and median models. This is due to the fact that the annotators are modeled after the normal distribution with the consensus being their mean. The tail sections of the normal distribution provide the outlier elimination power of the median model. The four models that we have proposed perform better as the model complexity increases.

The number of iterations until convergence are given in Table 5. As it can be observed from the table, M-CBS converges faster than M-SH and M-ABS.The reason for this is the scaling of the standard deviation with "w". This way M-CBS fits better and converges faster.

5.1.3. How beneficial is annotator scoring?

In Section 4, we discussed the importance of identifying competent annotators and proposed a scoring metric. Now, we elaborate on the annotator scores

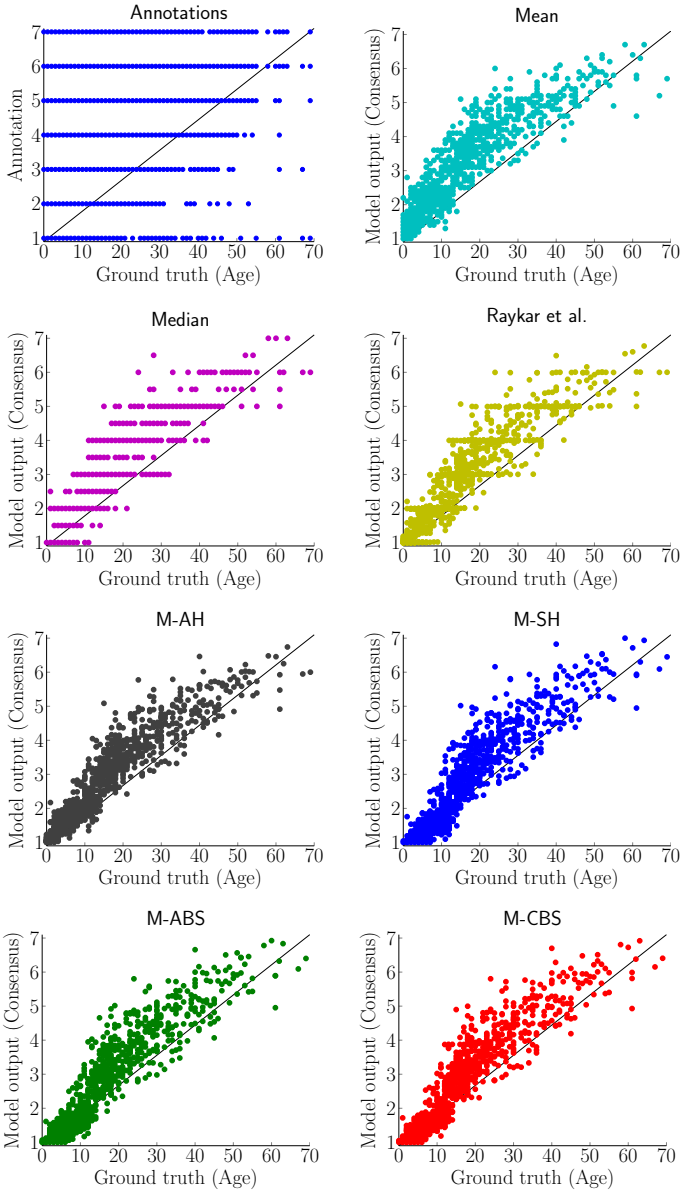


Figure 8: Ground truth estimation performance of models on joint set annotation data (The perfect fit would be on the diagonal)

Table 5: Number of iterations until convergence (100 repetitions)

	Mean	Std Dev.	Min	Max
Raykar[7]	10.02	0.556	8	12
M-AH	14.47	0.554	13	15
M-SH	45.03	4.540	16	54
M-ABS	41.03	6.649	14	54
M-CBS	12.24	2.437	8	32

calculated on real data for different models and how to make use of these scores.

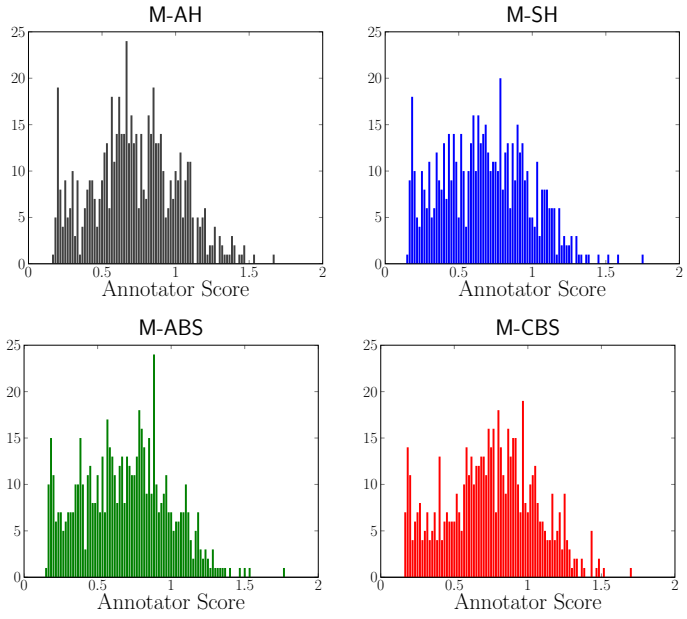


Figure 9: Annotator score histograms for the proposed models

First, we show the robustness of our scoring mechanism across different models. Figure 9 shows annotator score histograms for the models proposed in this work. It is evident that the shapes of the histograms are similar for all models. In addition, the median score improves slightly with increasing model complexity. The reason for this behavior is that, a higher complexity model

finds a higher quality consensus in which the annotators' individual opinions are represented better.

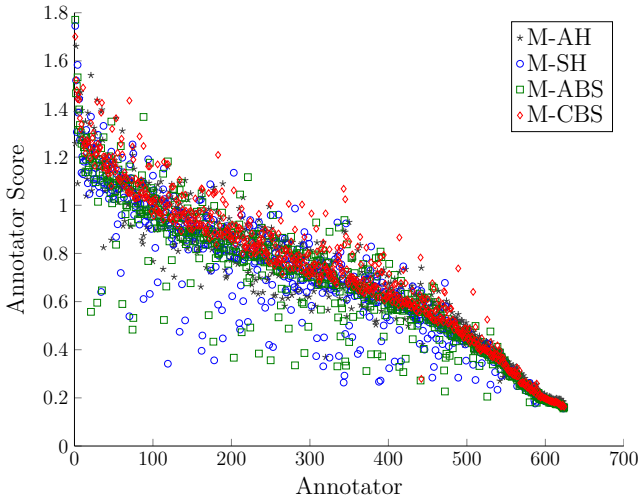
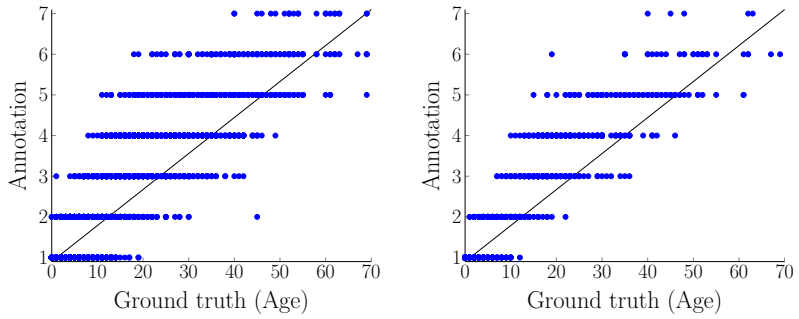


Figure 10: Annotator score comparison for the proposed models

In Figure 10, we observe the scores of every annotator for each model. For each annotator, we find the mean of the scores estimated by our proposed models. The annotators are sorted by these values for the sake of better visuality. The scoring mechanism usually agrees on similar scores for an annotator when employed with different models. In this figure, there are 2496 scores plotted, in which roughly 70 are outliers. Most of the scores follow the S–shaped trend. We also observe that for all models the scoring mechanism agrees on pointing out the most incompetent annotators, which explains the less scattered values at the tail section.

Figure 11 presents the annotations of the top scoring 50% and 10% of the annotators, respectively. We observe that the better the annotators, the better the annotations fit the ideal line. The scoring mechanism proves useful in eliminating the annotators who have given opposite or random rates to samples, as previously shown in Figure 8. It is notable to mention that, although choosing the top 10% of the annotators seems favorable, eliminating the annotators leaves some of the samples unrated, which is not desired. Thus, in the remainder of



(a) (b)

Figure 11: The annotations of (a)top 50% (b)top 10% scoring annotators.

this analysis we present our findings from the top 50% of the annotators, where each sample is rated by at least one annotator. However, if the crowd labeling task were to be continued, we would ask the top 10% to annotate more samples for solving the unrated sample problem.

Table 6: Utilizing annotator scores: Errors after using only top scoring and only bottom scoring annotators. The results are presented as mean and standard deviation for 100 repetitions.

Model	**MAE**		**RMSE**	
	Top 50%	**Bottom 50%**	**Top 50%**	**Bottom 50%**
Mean	5.56	13.49	7.61	16.29
Median	6.19	12.63	8.16	16.30
Raykar[7]	6.13 ± 0.037	12.44 ± 0.019	8.25 ± 0.044	15.34 ± 0.021
M-AH	5.65 ± 0.000	11.25 ± 0.072	7.70 ± 0.000	13.93 ± 0.066
M-SH	5.60 ± 0.075	10.06 ± 0.285	7.76 ± 0.082	13.84 ± 0.288
M-ABS	5.60 ± 0.078	10.12 ± 0.337	7.76 ± 0.085	13.86 ± 0.335
M-CBS	**5.52 ± 0.000**	**10.18 ± 0.091**	**7.65 ± 0.000**	**13.76 ± 0.094**

Table 6 shows the model errors obtained from employing top and bottom 50% scoring annotators. After separating 50% of the annotators, we re-infer the consensus values for each subset and report the related error for each model. Sets 1 and 2 have 5017 annotations each, resulting in 10034 annotations in the

joint set. Total annotation count of top 50% annotators is 5140. Although, the amount of annotations of this subset is similar to Sets 1 and 2, the model performances are almost as successful as the joint set. Since the top scoring annotators provide a better representation of the consensus, using a very simple model such as taking the mean produces very satisfactory results. For the mean model, we achieve substantially better results with approximately half of the annotations when we utilize only the top scoring annotators.

Although we strive to single out the most competent annotators, perfect annotations would result in obtaining the baseline error that we have discussed earlier. However, a little variance and annotator diversity would be preferable for beating the baseline. Since the ground truth values ($\in \{0, \ldots, 69\}$) have more precision than the annotation values ($\in \{1, \ldots, 7\}$) for this dataset, a better estimate can be obtained with increased variance in annotations.

5.1.4. Performance on binary labels

In many crowd labeling tasks, ordinal annotations are requested for binary labeled data. In these tasks, the annotators are usually asked to rate the degree of negativity or positivity of the sample. Then, continuous or ordinal valued annotations are binarized to make them compatible with methods accepting binary input. Unfortunately, this binarization process results in the loss of valuable information.

We designed our models to accept continuous and ordinal annotations. When we sought binary output labels, we used a threshold for the binarization of continuous consensus values estimated from the proposed models (i.e. model output).

We compare our binary label fitting performance with Welinder et al.'s [14] work. Their method is suitable comparison since they use a data independent approach (i.e. they don't use features) and do not have a training phase. When evaluating their work, we binarize the input annotations with a threshold of 4. For our methods, we use the annotations as they are and binarize the output consensus values with the same threshold value. The general intuition is to

Table 7: The Matthews correlation coefficient, sensitivity, specificity, and accuracy measures for binarized results. For Welinder[14] results, the input annotations are binarized, and for the other models the resulting consensus values are binarized. The results are presented as mean and standard deviation for 100 repetitions.

Model	**Input**	**MCC**	**Accuracy**	**Sensitivity**	**Specificity**
Welinder[14]	Binarized	0.427 ± 0.009	0.718 ± 0.009	0.686 ± 0.010	1.000 ± 0.002
Mean	Ordinal	0.521	0.814	0.796	0.980
Median	Ordinal	0.491	0.782	0.758	**1.000**
Raykar[7]	Ordinal	0.614 ± 0.001	0.880 ± 0.000	0.871 ± 0.000	0.961 ± 0.001
M-AH	Ordinal	0.626 ± 0.000	0.884 ± 0.000	0.874 ± 0.000	0.971 ± 0.000
M-SH	Ordinal	0.644 ± 0.007	0.896 ± 0.003	0.888 ± 0.003	0.961 ± 0.005
M-ABS	Ordinal	0.642 ± 0.008	0.895 ± 0.003	0.887 ± 0.004	0.961 ± 0.005
M-CBS	Ordinal	**0.648 ± 0.002**	**0.897 ± 0.001**	**0.890 ± 0.001**	0.961 ± 0.000

choose the median value during the binarization process. This is the reason for chosing 4 as the threshold value from the range 1–7.

In order to calculate the binary classification error, we also binarized the ground truth labels of the FGNet Aging Database to be 'young' when they are less than 35, and 'old' otherwise.

In Table 7, we present the Matthews correlation coefficient(MCC), sensitivity, specificity, and accuracy values. The Matthews correlation coefficient is a balanced statistical measure that is extracted from the confusion matrix. It can be used even if the classes are of very different sizes and symmetric in the sense of positive and negative classes. Its value is between -1 and 1 where 1 is a result of perfect prediction. It is calculated as

$$\mathrm{MCC} = \frac{\mathrm{TP} * \mathrm{TN} - \mathrm{FP} * \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \tag{38}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives and FN is the number of false negatives.

For two class problems, sensitivity and specificity values interchange when the class labels are interchanged. For these types of problems, they are only meaningful as a pair. As for the accuracy, it is strongly affected by unbalanced class sizes. Thus, out of these four statistical measures, MCC is the most suitable

measure for our problem because of its symmetry and balance.

When we analyze the results, we observe better MCC and accuracy values for M-CBS. Welinder[14] performs worse than the methods that accept continuous annotations, since it ignores lots of valuable information when binarizing the input annotations.

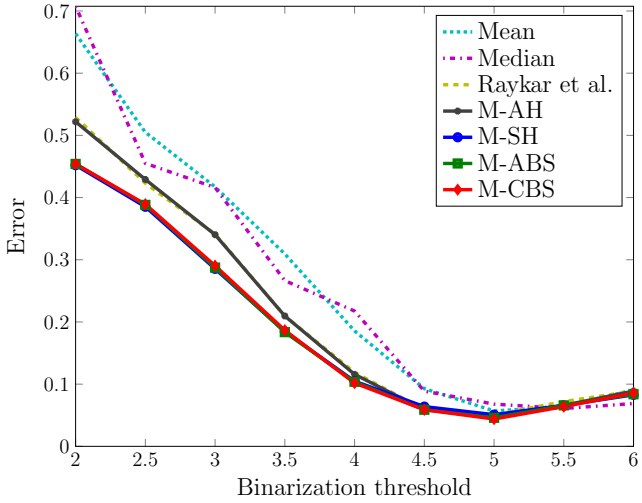


Figure 12: Change in error with respect to the change in consensus binarization threshold

In addition, we investigate the effect of different values of the consensus threshold for binarization in Figure 12. Although the value 4 would be expected to be the best threshold, we observe that 5 is a better threshold for this data. It can be deduced that the threshold selection for binarization has important effects on the final accuracy of the ground truth estimation and the best value depends on the data.

5.1.5. Discussion on global bias

With a careful look into Figures 8 and 12, one can observe that there is a positive bias in the annotations: the annotation scores are slightly above the ideal fit line. If we set the mean (μ_T) of the bias parameter(t)'s prior accordingly, we can decrease the global bias effect of the annotators. We empirically found that by setting $\mu_T = 0.7$, we would have better results. Note that, this is

only an observation of the annotations data and depends on the dataset; it is not an improvement for the models. We were able to observe this global bias, since we were in possession of the ground truth. Table 8 shows errors when the global bias is compensated for. The errors reduce drastically when this effect is removed.

Table 8: Compensating for global bias: Errors of M-ABS and M-CBS with $\mu_T = 0.7$. The results are presented as mean and standard deviation for 100 repetitions.

(a) Mean absolute error

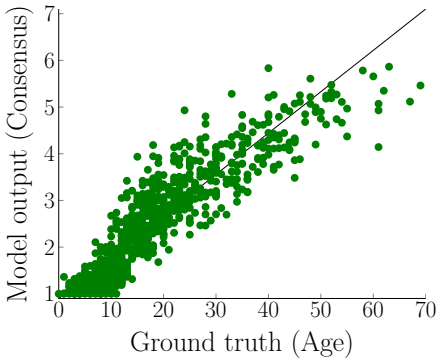
Model	**Set 1**	**Set 2**	**Joint**
M-ABS ($\mu_T = 0.7$)	4.52 ± 0.120	4.69 ± 0.102	4.24 ± 0.077
M-CBS ($\mu_T = 0.7$)	$\mathbf{4.44 \pm 0.012}$	$\mathbf{4.57 \pm 0.007}$	$\mathbf{4.14 \pm 0.010}$

(b) Root mean square error

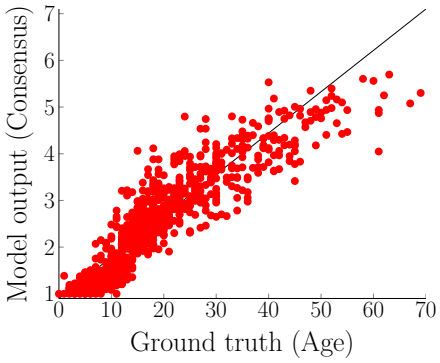
Model	**Set 1**	**Set 2**	**Joint**
M-ABS ($\mu_T = 0.7$)	6.06 ± 0.114	6.14 ± 0.119	5.45 ± 0.079
M-CBS ($\mu_T = 0.7$)	$\mathbf{5.91 \pm 0.012}$	$\mathbf{5.91 \pm 0.007}$	$\mathbf{5.33 \pm 0.009}$

In Figure 13, we observe that the models estimate the ground truth better after we take the global bias into account. In Figure 13a and Figure 13b, the estimated consensus scores are closer to the ideal fit line. In the CMC plot in Figure 13c, we see that the models perform much better after the 5-year error range. For the 10-year error range, the ratio shifts from 83% to 94%, when we compensate for the global bias with $\mu_T = 0.7$. Moreover, the binarization threshold shifts to four as one would expect (see Figure 13d).

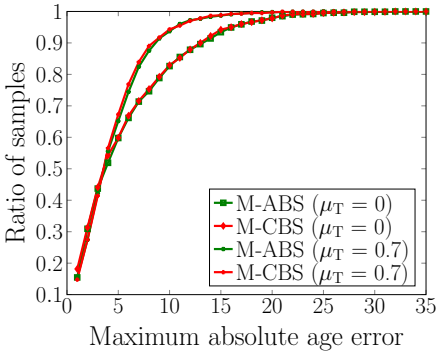
An explanation of why this global bias exists for the FGNet annotations could be related to the age range in the dataset. In the crowdsourcing phase, the annotators were not informed about the age range of the subjects in the dataset. Most of the annotators only saw young samples, since younger photos are in majority. Thus, the annotators were inclined to give higher ratings to



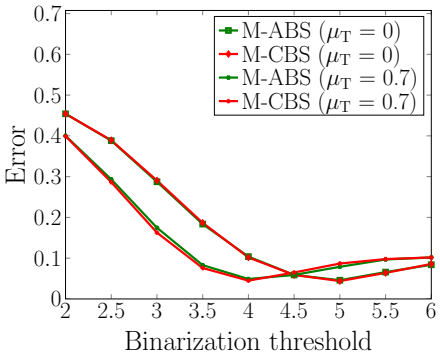
(a) M-ABS ($\mu_T = 0.7$)



(b) M-CBS ($\mu_T = 0.7$)



(c) CMC ($\mu_T = 0.7$)



(d) Change in binarization error with respect to threshold ($\mu_T = 0.7$)

Figure 13: Effect of removing global bias on the consensus scores

younger people. Since the annotators would expect the minimum age to be zero, they were more successful in annotating younger samples. Refraining from informing the annotators about the age range was intentional. Our aim is to obtain annotations where the actual score range is not exactly known by the annotators. An example for such cases is annotations for human traits, such as personality, which we investigate in the next section.

5.2. Results on real data without ground truth

In this section, we analyze the performance of the annotator models on a real dataset where there is no actual ground truth. We use the personality impressions as our domain where the annotations are highly subjective. We evaluated the performance of the annotator models on a regression and a classification task to predict the extraversion trait based on the consensus scores estimated by each model.

5.2.1. Personality Impressions Data

In parallel to the increasing existence of computers, robots, and machines equipped with various multimodal sensors in our daily lives, there is also an increasing interest in building automatic systems that are capable of inferring and predicting traits of people. One of these traits, personality, defines an individual's distinctive character as a collection of consistent behavioral and emotional traits. The Big Five model has been the widely used model, which factors personality into five different dimensions (i.e., extraversion, agreeableness, conscientiousness, emotional stability, and openness to experience). While some of those dimensions are apparent in brief observations, some others are not. For those dimensions of personality, the personality is evident in and can be predicted from people's verbal and nonverbal behavior in brief segments [10, 32, 35].

As a dataset to study personality, we used a subset from the Emergent LEAder (ELEA) corpus [36]. The ELEA AV subset consists of audio-visual recordings of 27 meetings, in which the participants perform a winter survival

Table 9: Personality annotations per annotator on the ELEA data

Ann 1	91
Ann 2	83
Ann 3	77
Ann 4	49
Ann 5	6

task with no roles assigned. The winter survival task is a simulation game where the participants in the task are the survivors of an airplane crash. They are asked to rank 12 items to take with them to survive as a group. Participants first ranked the items individually; then, as a group. The task itself is designed such that it promotes interactions among the participants in the group. The discussion and negotiation parts of the interaction present cues on the personality of the participants, making it a suitable database to study personality prediction. There are 102 participants in total in the ELEA AV subset. Each meeting lasts approximately 15 minutes and is recorded with two webcams and a microphone array. More details about the ELEA corpus can be found in [36, 37].

For each participant in the dataset, the personality impressions are obtained from external observers[32]. Ten Item Personality Inventory (TIPI) is used for measuring the Big Five personality traits of the participants [38]. The TIPI questionnaire includes two questions per trait, answered on a 7-point Likert scale. The score for each trait is also calculated on a scale of one to seven. For each participant, a one-minute segment is selected from the meeting, which corresponds to the segment that includes the participant's longest turn. Each participant was annotated by three different annotators, with a total of five annotators annotating the whole dataset. Table 9 shows the number of annotations per annotator. More details on the annotations can be found in [32].

5.2.2. Predicting personality impressions using nonverbal cues

The nonverbal cues that we display in our everyday life, particularly during our interaction with others, contains significant information regarding our personality [39]. Psychologists have long investigated the links between the nonverbal cues that we display and our personality traits and showed that several dimensions of personality is expressed through voice, face, body in the nonverbal channel [40]. In social computing literature, predicting personality using automatically extracted nonverbal cues has been addressed in several recent studies [32, 35, 10].

We use the data that is used in [32] where a large set of audio-visual nonverbal features are extracted and used in the prediction of personality. The set of features include features such as speaking turn features (speaking length, number of turns, turn duration), prosodic features (energy, pitch), visual activity features, and visual focus of attention features. More detail can be found in [32]. For the current study, we use a concatenation of all the features used in [32] when training our regression models.

We only focus on the extraversion trait for the purposes of this study. We first perform a regression task where the goal is to estimate the personality impression score. Secondly, we perform a binary classification task where the goal is to predict whether the person is high or low in extraversion. The median of the scores is used as the cut-off point for binarization.

We use linear Ridge regression for estimating the personality impression scores and report the Relative Absolute Error (RAE) on a leave-one-out cross validation setting. RAE is calculated as:

$$RAE = \frac{\sum_{i=1}^{N} |p_i - a_i|}{\sum_{i=1}^{N} |\bar{a}_i - a_i|} \tag{39}$$

where p is the score predicted by the regression model and a is the annotator consensus score as estimated by the annotator model.

For binary classification, we used the estimated scores by the regression models and labeled the samples as high and low based on the cut-off point. We report the classification accuracy.

5.2.3. Comparsion of annotator models' performance on the regression and classification tasks

We perform regression and classification to predict personality impressions using the consensus scores estimated by different annotator models. It is important to note that the consensus scores of different models could have different ranges and scales. While one model provides consensus scores in the range of 1 to 7, another model's scores could be in the range 1 to 6. As a metric which is less sensitive to such differences, we use RAE to compare the regression performances.

Table 10: Regression and classification results on extraversion prediction

	RAE	**Classification Accuracy (%)**
Mean	0.78	72.55
Median	0.82	70.59
Raykar[7]	0.88	63.73
M-AH	0.86	67.65
M-SH	0.77	74.51
M-ABS	0.77	75.49
M-CBS	0.77	73.53

The results are given in Table 10. We see that the lowest errors are obtained with consensus scores estimated by M-CBS, followed by M-ABS and M-SH. When it comes to the classification accuracy, the observations are different and not directly inline with the regression errors. The highest accuracy is achieved by M-ABS, followed by M-SH and M-CBS. The reasoning behind this observation could be related to the binarization of the scores. The errors of the regression models for the samples that are close to the cut-off point directly affect the classification accuracy. Even if a regression model has a low RAE, if the errors are concentrated around the cut-off point, a lower classification accuracy could be observed.

6. Conclusions

In this paper, we proposed four Bayesian models for obtaining consensus in continuous valued crowd labeling tasks by taking annotator behaviors into account. We also introduced a novel metric for measuring annotator quality. We acquired annotation data on a dataset with known ground truth for evaluating the performance of the proposed models. In addition, we adapted our methods to work with binary labeled data and reported their performance.

We observed various annotator behaviors and successfully compensated for this versatility with the use of scale and bias parameters. The error rates show that our methods perform better in estimating the consensus score than widely used methods. We also showed that it is possible to select competent annotators using our metric and keep the consensus error rate the same while reducing labeling costs by 50%. On a personality impressions dataset, where there is no ground truth to compare the estimated consensus scores, we have observed that the consensus scores obtained with the proposed models lead to lower regression errors in comparison to the widely used methods.

We have made several important observations in the course of this work. First of all, the samples that are hard to rate result in misleading most of the annotators where the consensus value does not agree with the ground truth. In crowdsourced efforts, this problem is inevitable. Another observation is that, the annotators may tend to be biased as a whole due to the nature of the labeling problem. Informing the annotators about the opposite ends of the scale that occur in the dataset is important for alleviating the global bias problem, where possible.

Acknowledgement

This study is partially supported by the Swiss National Science Foundation through the Ambizione fellowship project (PZ00P2_136811) and the State Planning Organization (DPT) of the Republic of Turkey under the project TAM with the project number 2007K120610.

References

[1] F. Galton, Vox populi (the wisdom of crowds), Nature 75 (1907) 450–451.

[2] Amazon Mechanical Turk - https://www.mturk.com.

[3] Crowdflower - http://www.crowdflower.com.

[4] B. Frenay, M. Verleysen, Classification in the presence of label noise: A survey, Neural Networks and Learning Systems, IEEE Transactions on 25 (5) (2014) 845–869. doi:10.1109/TNNLS.2013.2292894.

[5] G. Srivastava, J. A. Yoder, J. Park, A. C. Kak, Using objective ground-truth labels created by multiple annotators for improved video classification: A comparative study, Computer Vision and Image Understanding 117 (10) (2013) 1384 – 1399. doi:http://dx.doi.org/10.1016/j.cviu.2013.06.009.

[6] B. Carpenter, Multilevel bayesian models of categorical data annotation, Unpublished manuscript.

[7] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, L. Moy, Learning from crowds, Journal of Machine Learning Research 99 (2010) 1297–1322.

[8] F. Rodrigues, F. Pereira, B. Ribeiro, Learning from multiple annotators: Distinguishing good from random labelers, Pattern Recognition Letters 34 (12) (2013) 1428 – 1436. doi:http://dx.doi.org/10.1016/j.patrec.2013.05.012.

[9] G. Chittaranjan, O. Aran, D. Gatica-Perez, Exploiting observers' judgements for nonverbal group interaction analysis, Face and Gesture 2011 (2011) 734–739doi:10.1109/FG.2011.5771339.

[10] J.-I. Biel, O. Aran, D. Gatica-Perez, You are known by how you vlog: Personality impressions and nonverbal behavior in youtube., in: ICWSM, The AAAI Press, 2011.

[11] K. Audhkhasi, S. Narayanan, A globally-variant locally-constant model for fusion of labels from multiple diverse experts without using reference labels, Pattern Analysis and Machine Intelligence, IEEE Transactions on 35 (4) (2013) 769–783. doi:10.1109/TPAMI.2012.139.

[12] Q. Liu, J. Peng, A. Ihler, Variational inference for crowdsourcing, Advances in Neural Information Processing Systems 25 (2012) 701–709.

[13] Y. Tian, J. Zhu, Learning from Crowds in the Presence of Schools of Thought, Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (2012) 226–234.

[14] P. Welinder, S. Branson, S. Belongie, P. Perona, The multidimensional wisdom of crowds, Advances in Neural Information Processing Systems 23 (2010) 2424–2432.

[15] W. Wu, Y. Liu, M. Guo, C. Wang, X. Liu, A probabilistic model of active learning with multiple noisy oracles, Neurocomputing 118 (2013) 253 – 262. doi:http://dx.doi.org/10.1016/j.neucom.2013.02.034.

[16] H. Dutta, W. Chan, Using community structure detection to rank annotators when ground truth is subjective, NIPS Workshop on Human Computation for Science and Computational Sustainability (2012) 1–4.

[17] P. Donmez, J. G. Carbonell, J. Schneider, Efficiently learning the accuracy of labeling sources for selective sampling, Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09 (2009) 259doi:10.1145/1557019.1557053.

[18] P. Zhang, Z. Obradovic, Integration of multiple annotators by aggregating experts and filtering novices, 2012 IEEE International Conference on Bioinformatics and Biomedicine (2012) 1–6doi:10.1109/BIBM.2012.6392657.

[19] P. Welinder, P. Perona, Online crowdsourcing: Rating annotators and obtaining cost-effective labels, 2010 IEEE Computer Society Conference

on Computer Vision and Pattern Recognition - Workshops (2010) 25–32doi:10.1109/CVPRW.2010.5543189.

[20] P. Zhang, Z. Obradovic, Learning from inconsistent and unreliable annotators by a Gaussian mixture model and Bayesian information criterion, Machine Learning and Knowledge Discovery in Databases (2011) 553–568.

[21] Y. Yan, R. Rosales, G. Fung, M. Schmidt, G. Hermosillo, L. Bogoni, L. Moy, J. Dy, Modeling annotator expertise: Learning when everybody knows a bit of something, International Conference on Artificial Intelligence and Statistics 9 (2010) 932–939.

[22] J. Whitehill, P. Ruvolo, T. Wu, Whose vote should count more: Optimal integration of labels from labelers of unknown expertise, Advances in Neural Information Processing Systems 22 (1) (2009) 2035–2043.

[23] A. Ghosh, S. Kale, P. McAfee, Who moderates the moderators?: crowdsourcing abuse detection in user-generated content, Proceedings of the 12th ACM conference on Electronic commerce (2011) 167–176.

[24] V. Raykar, S. Yu, Eliminating spammers and ranking annotators for crowdsourced labeling tasks, Journal of Machine Learning Research 13 (2012) 491–518.

[25] F. Wauthier, M. Jordan, Bayesian bias mitigation for crowdsourcing, Proc. of NIPS (2011) 1–9.

[26] J. Bi, X. Wang, Min-Max Optimization for Multiple Instance Learning from Multiple Data Annotators, KDD13 August.

[27] A. P. Dawid, A. M. Skene, Maximum likelihood estimation of observer error-rates using the em algorithm, Applied statistics (1979) 20–28.

[28] V. C. Raykar, S. Yu, Annotation models for crowdsourced ordinal data, NIPS workshop on Computational Social Science and the Wisdom of Crowds.

[29] H. Kajino, Y. Tsuboi, I. Sato, H. Kashima, Learning from Crowds and Experts, Proceedings of the 4th Human Computation Workshop (HCOMP) (2012) 107–113.

[30] B. Lakshminarayanan, Y. Teh, Inferring ground truth from multi-annotator ordinal data: a probabilistic approach, arXiv preprint arXiv:1305.0015 (2013) 1–19arXiv:arXiv:1305.0015v1.

[31] The FGNet Aging Database.
URL http://www-prima.inrialpes.fr/FGnet/html/benchmarks.html

[32] O. Aran, D. Gatica-Perez, One of a kind: Inferring personality impressions in meetings, in: ICMI, 2013.

[33] D. Zhu, B. Carterette, An analysis of assessor behavior in crowdsourced preference judgments, in: SIGIR 2010 workshop on crowdsourcing for search evaluation, 2010, pp. 17–20.

[34] A. Kittur, E. H. Chi, B. Suh, Crowdsourcing user studies with mechanical turk, in: Proceedings of the SIGCHI conference on human factors in computing systems, ACM, 2008, pp. 453–456.

[35] O. Aran, D. Gatica-Perez, Cross-domain personality prediction: From video blogs to small group meetings, in: ICMI, 2013.

[36] D. Sanchez-Cortes, O. Aran, M. Mast, D. Gatica-Perez, A nonverbal behavior approach to identify emergent leaders in small groups, Multimedia, IEEE Transactions on 14 (3) (2012) 816–832. doi:10.1109/TMM.2011.2181941.

[37] D. Sanchez-Cortes, O. Aran, D. Gatica-Perez, An audio visual corpus for emergent leader analysis, in: Icm-mlmi'11: workshop on multimodal corpora for machine learning: taking stock and road mapping the future, 2011.

[38] S. D. Gosling, P. J. Rentfrow, W. B. Swann, A very brief measure of the big-five personality domains, Journal of Research in Personality 37 (2003) 504–528.

[39] M. L. Knapp, J. A. Hall, Nonverbal Communication in Human Interaction, Wadsworth, Cengage Learning, 2008.

[40] R. Gifford, The SAGE Handbook of Nonverbal Communication, SAGE Publications, Inc., 2006, Ch. Personality and Nonverbal Behavior: A Complex Conundrum, pp. 159–181.