CROWD-LABELING FOR CONTINUOUS-VALUED ANNOTATIONS

by

Yunus Emre Kara

B.S., Mathematics, Boğaziçi University, 2008

M.S., Computer Engineering, Boğaziçi University, 2011

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

Graduate Program in Computer Engineering
Boğaziçi University
2018

In loving memory of İsmail Arı, dear friend and colleague

ACKNOWLEDGEMENTS

First, I would like to thank my spouse, best friend, and colleague Gaye Genç for being my partner and foremost supporter in life, the universe and everything, always...

I would like to offer my deepest gratitude to my advisor Lale Akarun. Without her guidance and patience, this thesis would not be complete. I would like to thank all those who gave insights that helped me to complete this thesis. First, I would like to thank Oya Aran for all her contributions and comments; she made it possible for me to work in this field. I also thank Yusuf Gören for his never ending friendship and the exceptional support he gave whenever I needed; Çetin Meriçli for being a 'dayı' and a great proofreader; and Burak Gürel for his mathematical support. I thank my thesis progress committee members Ali Taylan Cemgil and Murat Şensoy for their comments and technical perspective. I also thank the defense jury members Ümit Işlak and Sinan Yıldırım.

I would also like to thank my friends and colleagues in the CMPE department for providing a great and fun working environment, whose names can not fit these pages. However, I would like to especially thank those who made it possible for me to bear the final burden – Furkan Kıraç for being a great friend and colleague; Alp Kındıroğlu for his friendship, support, and lovely personality; Ozan Özen for being a popstar as he was meant to be; Tekin Meriçli for our intellectual discussions; Birkan Yılmaz for his fast problem solving skills and being a great research partner; Itır Karaç for being a big sister to me. I would also like to thank all my friends and colleagues in PILAB and NETLAB, especially Barış Kurt, Barış Emmim Demiröz, Recep Doğa Siyli, Orhan Sönmez, Necati Cihan Camgöz, Oğulcan Özdemir, and Serhan Daniş. I would like to thank the faculty members Tuna Tuğcu, Cem Ersoy, and Albert Ali Salah for always supporting me.

Last but not the least, I would like to offer my deepest gratitude to my grandmother Emine Alver, my parents Nevin Kara and İsmail Hakkı Kara, my sister Efser

Kara, and my other parents Nurtop Genç and Hüseyin Genç. Without their support and encouragement, I could not bear the burden.

This work is supported by the Turkish Ministry of Development under the TAM Project, number 2007K120610.

ABSTRACT

CROWD-LABELING FOR CONTINUOUS-VALUED ANNOTATIONS

As machine learning gained immense popularity across a wide variety of domains in the last decade, it has become more important than ever to have fast and inexpensive ways to annotate vast amounts of data. With the emergence of crowdsourcing services, the research direction has gravitated toward putting 'the wisdom of crowds' to use. We call the process of crowdsourcing based label collection *crowd-labeling*. In this thesis, we focus on crowd consensus estimation of continuous-valued labels. Unfortunately, spammers and inattentive annotators pose a threat to the quality and trustworthiness of the consensus. Thus, we develop Bayesian models taking different annotator behaviors into account and introduce two crowd-labeled datasets for evaluating our models. High quality consensus estimation requires a meticulous choice of the candidate annotator and the sample in need of a new annotation. Due to time and budget limitations, it is beneficial to make this choice while collecting the annotations. To this end, we propose an *active crowd-labeling* approach for actively estimating consensus from continuous-valued crowd annotations. Our method is based on annotator models with unknown parameters, and Bayesian inference is employed to reach a consensus in the form of ordinal, binary, or continuous values. We introduce ranking functions for choosing the candidate annotator and sample pair for requesting an annotation. In addition, we propose a penalizing method for preventing annotator domination, investigate the explore-exploit trade-off for incorporating new annotators into the system, and study the effects of inducing a stopping criterion based on consensus quality. Experimental results on the benchmark datasets suggest that our method provides a budget and time-sensitive solution to the crowd-labeling problem. Finally, we introduce a multivariate model incorporating cross attribute correlations in multivariate annotations and present preliminary observations.

ÖZET

SÜREKLİ DEĞERLİ İŞARETLEMELER İÇİN KİTLE ETİKETLEME

Hızlı ve ucuz veri işaretleme, makine öğrenmesinin son on yılda birçok alanda aşırı rağbet görmesiyle birlikte daha da önemli bir hale geldi. Kitle kaynak servislerinin çıkışı, araştırma yönünü 'kitlelerin bilgeliğini' kullanmaya itti. Kitle kaynak temelli etiket toplama işlemini *kitle etiketleme* olarak adlandırıyoruz. Bu tezde, sürekli değerli etiketler için kitle oydaşım kestirimi üzerine odaklanıyoruz. Maalesef, kötü niyetli veya dikkatsiz işaretçiler, oydaşım etiketinin kalitesine ve güvenilirliğine kötü etki etmektedir. Bundan ötürü, değişik işaretçi davranışlarını dikkate alan Bayesçi modeller geliştiriyoruz ve modellerimizi değerlendirmek için iki yeni kitle işaretli veri kümesi tanıtıyoruz. Kaliteli oydaşım etiketi kestirimi, işaretçi ve işaretlenecek örnek seçiminin akıllı bir şekilde yapılmasını gerektirir. Zaman ve bütçe kısıtlarından dolayı, bu seçimleri işaret toplama sırasında yapmak önemlidir. Bu nedenle, sürekli değerli kitle işaretlerinden aktif bir şekilde etiket kestirimi yapan bir *aktif kitle etiketleme* yaklaşımı öneriyoruz. Yöntemimiz, bilinmeyen parametreleri olan işaretçi modellerine dayalıdır ve sıralı, ikili veya sürekli değerli etiketlere ulaşabilmek için Bayesçi çıkarım kullanır. İşaret istemek için işaretçi ve işaretlenecek örnek ikilisini seçmede kullanılan sıralama fonksiyonları tanıtıyoruz. Ek olarak, işaretçi baskınlığını engellemek için cezalandırma yöntemi öneriyoruz, sisteme yeni işaretçiler eklemek için keşfetme ve kullanma dengesini araştırıyoruz ve oydaşım etiketi kalitesine göre aktif işaretlemeyi durdurma kriteri koymanın etkilerini inceliyoruz. Kıstas veri kümelerindeki deneysel sonuçlar, yöntemimizin kitle etiketleme problemine bütçeye ve zamana duyarlı bir çözüm sağladığını göstermektedir. Son olarak, çok değişkenli işaretlemelerdeki nitelikler arası bağıntıları dikkate alan çok değişkenli bir model tanıtıyoruz ve hakkındaki ilk gözlemlerimizi sunuyoruz.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

a_j	Adverseness parameter of the j^{th} annotator
b_j	Bias parameter of the j^{th} annotator
$\boldsymbol{b_j}$	Rightmost column of $\boldsymbol{\Phi_j}$
c	Annotation upper value for zero centered annotation range
C	Number of annotator categories for the multivariate model
d_j	Lower limit of the annotator score path integral of the j^{th} annotator
e_j	Upper limit of the annotator score path integral of the j^{th} annotator
i_k	Sample index of the k^{th} annotation
$\mathcal{I}$	Set of all samples
j_k	Annotator index of the k^{th} annotation
$\mathcal{J}$	Set of all annotators
$\mathcal{J}'$	Set of currently active annotators
$\mathcal{J}^1$	Set of annotators that have at least one annotation
$\mathcal{K}$	Set of current annotations
$\mathcal{K}_i$	Set of sample i's annotations
$\mathcal{K}^j$	Set of annotator j's annotations
K	Number of annotations
$\boldsymbol{M_c}$	Location matrix prior parameter for the matrix normal random variable $\boldsymbol{\Phi_j}$
n_0	Degree of freedom prior parameter for the Wishart random matrix $\boldsymbol{\Lambda_j}$
N_j	Annotation count of the j^{th} annotator
N	Number of samples
$\boldsymbol{p}$	Probability vector prior parameter for the categorical random vector $\boldsymbol{z_j}$
$\boldsymbol{p}_c$	c^{th} element of the vector $\boldsymbol{p}$
R	Number of annotators

s_B	Standard deviation hyperparameter for the annotator bias parameter
S_A	Annotator competence score function
$S_A^{\mathcal{K}}$	Annotator selector function based only on annotator's workload
$S_A^{\mathcal{R}}$	Random annotator selector function
S_A^{φ}	Dominance suppression based annotator competence score function with dominance suppression coefficient φ
S_S	Sample consensus quality score function
$\boldsymbol{V_0}$	Among column scale matrix prior parameter for the matrix normal random variable $\boldsymbol{\Phi_j}$
w_j	Opinion scale parameter of the j^{th} annotator
$\boldsymbol{W_0}$	Scale matrix prior parameter for the Wishart random matrix $\boldsymbol{\Lambda_j}$
x_i	Consensus value of the i^{th} sample
$\boldsymbol{x_i}$	Consensus value of the i^{th} sample for the multivariate model
X	Set of sample consensus values
y_k	Value of the k^{th} annotation
$\boldsymbol{y_k}$	Value of the k^{th} annotation for the multivariate model
Y	Set of annotation values
$\boldsymbol{z_j}$	1-of-C binary vector for annotator category
$\boldsymbol{z_j}_c$	c^{th} element of the vector $\boldsymbol{z_j}$
Z	Set of $\boldsymbol{z_j}$ for all annotators

α_λ	Shape hyperparameter for the annotator precision parameter
β_λ	Rate hyperparameter for the annotator precision parameter
β_w	Rate hyperparameter for the annotator opinion scale parameter
δ	Target sample consensus posterior variance
$\mathcal{E}$	Exploration rate parameter
θ_j	Parameters of the j^{th} annotator
θ	Set of parameters of all annotators
λ_j	Precision parameter of the j^{th} annotator

Λ	Set of $\boldsymbol{\Lambda_j}$ for all annotators
$\boldsymbol{\Lambda_j}$	Random variable precision parameter for the multivariate model
μ_B	Location hyperparameter for the annotator bias parameter
τ	Target lower limit on sample score
φ	Dominance suppression coefficient
Φ	Set of $\boldsymbol{\Phi_j}$ for all annotators
$\boldsymbol{\Phi_j}$	Random variable scale and bias parameter for the multivariate model
$\boldsymbol{\chi_i}$	Consensus value of the i^{th} sample with 1 concatenated at the end
$\boldsymbol{\Omega_j}$	Left d columns of $\boldsymbol{\Phi_j}$

LIST OF ACRONYMS/ABBREVIATIONS

ACL	Active crowd labeling
CMC	Cumulative match curve
CPU	Central processing unit
ELEA	Emergent LEAder corpus
EM	Exceptation Maximization
FN	False negative
FP	False positive
M-ABS	Annotation bias sensitive model
M-AH	Adversary handling model
M-CBS	Consensus bias sensitive model
M-SH	Scale handling model
MAE	Mean absolute error
MAP	Maximum a posteriori
MCC	Matthews correlation coefficient
ML	Maximum likelihood
NLP	Natural language processing
O-CBS	Online M-CBS
O-CBS+	Online M-CBS from scratch
RAE	Relative absolute error
RMSE	Root mean squared error
TIPI	Ten item personality inventory
TN	True negative
TP	True positive

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 a 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 task givers submit lots of small tasks such as dataset labeling to be completed by annotators all around the world, for a fee.

In the machine learning domain, labeled datasets are valuable commodities. Computing resources have increased exponentially for two decades, driving machine learning toward big data applications. The introduction of the ImageNet database [4], a large crowd-labeled dataset, and the success of deep neural network methods have further pushed the research direction toward the use of large datasets. When used in a supervised manner, deep neural networks heavily rely on the availability of vast amounts of training data, with ground truth labels. Researchers in deep learning most often depend on crowd labeling to supply these labels. This popularity has resulted in the introduction of many large crowd-labeled datasets such as the recently introduced Open Images dataset [5].

Providing ground truth labels, which indicate the correct labels, for large datasets often proves to be excessively time consuming. Thus, researchers tend to outsource the labeling process, especially for the aforementioned large datasets. However, employing expert labelers is expensive. Crowdsourcing the labeling process is a cost-effective and fast method to solve this problem, especially when expertise is not necessarily required.

Crowd-labeling is the process of collecting annotations from crowds and using them for estimating consensus values to be used as labels. In crowd-labeling, each person annotates a randomized subset of samples and every sample is annotated by a subset of all annotators. If we reorganize annotations into a matrix with annotators as rows and samples as columns, the resulting matrix would often be sparse. This is a common case for crowdsourced annotation tasks. The aim of crowd-labeling is to obtain consensus labels for each sample using this sparse set of annotations.

Many crowd-labeling problems aim to obtain continuous or ordinal labels, such as the position of an object, age of a person, or air temperature. Surprisingly, active crowd-labeling for continuous-valued annotations is a rather sidelined open issue. Related literature on active crowd-labeling mainly focuses on binary annotation problems due to several reasons. First of all, formulating the active crowd-labeling problem in a binary setting is often more tractable with provable mathematical guarantees. Due to the nature of the continuous domain, providing mathematical guarantees in active crowd-labeling solutions proves to be hard, if not impossible. This has pushed researchers to work with well-studied algorithms by binarizing existing continuous or ordinal annotations. Additionally, presenting the annotation tasks in the form of yes/no or positive/negative reduces task intricacy for the annotators. Although working with binary annotations has several advantages, valuable information is often lost during binarization. Moreover, binary active crowd-labeling approaches are simply impractical when continuous labels are sought. In Section 1.1.1, we give a literature review on crowd-labeling.

A problem that is commonly encountered during crowd-labeling is reduced quality of consensuses arising from inattentive annotators and spammers. Although there are numerous methods in the literature that deal with the low quality annotations provided by said annotators, most are effective only after the annotation process is completed. At this point, valuable time and money are already spent on low quality annotations. The classical use of crowd-labeling is analogous to a careless shopper who buys excessively without proper planning and ends up throwing away their purchase when the product is of low quality or unneeded. In contrast, imagine that the researcher is a meticulous shopper with limited time and money. The most important questions on their mind would be: What am I in need of purchasing and which vendor should I purchase it from? Applying this reasoning to the crowd-labeling problem calls for a smarter solution and active learning is the remedy to this problem. The general idea of active learning can be applied to the crowd-labeling problem in terms of choosing which annotation to incorporate into the annotation pool. In this thesis, the process of smart annotation collection using crowdsourcing is called *active crowd-labeling*. We give an extensive review on active crowd-labeling literature in Section 1.1.2.

In this thesis, we focus on attaining high consensus quality from continuous-valued annotations while reducing the cost of the annotation process. We achieve this goal by modeling annotator behaviors and making use of active crowd-labeling. The crowd-labeling method we propose is based on annotator modeling and consensus estimation by Bayesian inference, which is used for producing ordinal and binary labels in addition to continuous labels. One advantage of the method is that it is unsupervised: the gold standard label is not needed for any sample. The proposed method only uses crowd or expert annotations for estimating consensus values and does not depend on the features extracted from the data to be labeled. For the active crowd-labeling part, we introduce an effective mechanism that decides which sample needs a new annotation and who should annotate it. In addition, we introduce a multivariate model for incorporating correlations across different attributes.

In the remainder of this chapter, we discuss the related work in this domain (Section 1.1), followed by the novelty and contributions of this thesis (Section 1.2). In

Chapter 2, we introduce the datasets on which we evaluate our methods. In Chapter 3, we give the definition of the crowd-labeling problem and investigate annotator behaviors by explaining various annotator types. In Chapter 4, we focus on the problem of passive crowd-labeling. We propose four novel Bayesian models which are used for simultaneously modeling the behaviors of annotators and finding the consensus for each sample. Chapter 5 describes our methodology for dealing with the problem of active crowd-labeling. Since crowdsourced labeling is an expensive process, choosing good annotators and samples that would benefit from new annotations is crucial for reducing the costs. Chapter 6 deals with how to use active crowd-labeling to improve existing consensus in crowd-labeling problems. In Chapter 7, we elaborate on how to conduct smart label collection from scratch and compare our methods with existing methods in the literature. In Chapter 8, we introduce a multivariate annotation model, give a variational Bayes solution, and present some preliminary experiments. Finally, we conclude the thesis in Chapter 9, 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.* [6] 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.

1.1.1. Crowd-Labeling

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.

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, annotator errors and outliers have a high impact on the consensuses obtained with these approaches. Moreover, these simple methods disregard valuable information on annotator behavior and expertise. Investigating and modelling annotator behaviors would prove useful for utilizing valuable information. Numerous methods also make use of features extracted from data [7–9]. Although the extracted features provide additional information, the success of data dependent methods relies heavily on the quality of the features. In addition, model performance across different types of problems requiring different types of features is unpredictable.

Crowd-labeling is a well-studied area for binary annotations [7,8,10–15]; nevertheless it is rather sidelined for continuous-valued annotations. For many annotation tasks with continuous-valued attributes, researchers either acquire the annotations in binary form or they binarize the continuous/ordinal-valued annotations after acquisition. 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) [13, 15]. Although working with binary annotations streamlines the label estimation process, valuable information may be lost during binarization. Carpenter [10] utilizes multilevel Bayesian approaches on binary data annotations, and introduce priors on sensitivity and specificity of annotators. Ground truth estimation is done by annotator modeling by using the annotators' self-reported confidences in [16]. Human personality trait evaluation is also a problem where no quantifiable ground truth exists. Trait annotations collected by crowdsourcing are used in [17] for personality trait classification.

Considering ordinal annotations as if they were categories, as input to the categorical models, is another simplification used in the literature [18, 19]. Rodrigues *et al.* [11] challenged the results of Raykar *et al.* [7] in a supervised multiclass classification problem with a simpler probabilistic model. Srivastava *et al.* investigate the problem of subjective video annotation and the majority opinion is shown to be the most objective annotation for a video [20]. 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.

There are only a handful of works focusing on ordinal or continuous annotations. Raykar *et al.* estimate the gold standard and measure the competence of the annotators iteratively in a probabilistic approach [7]. They mainly focus on the estimation of consensus by making use of features extracted from the sample data. Their method is also adapted to work without the sample features. The focus of Lakshminarayanan and Teh is on ordinal labels where task difficulty is incorporated to the discretization of continuous latent variables [21]. Peng *et al.* propose a domain-specific approach to the protein folding annotation problem by maximizing the log-likelihood of an exponential family mixture model of annotation similarities [22]. Ok *et al.* model the continuous crowd-labeling problem as a bipartite graph and use a belief propagation based Bayesian iterative algorithm when the annotator noise levels are known [23]. For the case where the annotator noise levels are unknown, they employ a non-Bayesian iterative algorithm with marginal performance loss.

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. In this thesis, we focus on estimating the crowd consensus to be used as sample labels from continuous-valued annotations by employing active crowd-labeling.

1.1.2. Active Crowd-Labeling

Active learning aims to concurrently reduce the training cost and increase the performance of machine learning algorithms by smartly selecting the instances to be included during the learning process. The concept of active learning is a well-suited approach to the crowd-labeling domain where an immense number of annotations need to be acquired, costing both money and time. Settles surveys and organizes active learning methods, practical considerations, and the relation of active learning to other research areas in detail [24]. Fu *et al.* [25] survey the active learning domain from the perspective of instance selection, where active learning methods are categorized into two main groups: those that assume independent and identically distributed instances and those that consider instance correlations.

1.1.2.1. Annotator Selection Strategies. The quality of the annotators varies largely in crowd-labeling problems. Not only do the annotators' expertise vary, but also some of them may attempt to exploit the system for profit. Donmez *et al.* use the interval estimation learning method for selecting the best annotators by incorporating the exploration-exploitation trade-off [26]. Raykar and Yu introduce an annotator ranking metric for detecting spammers [27]. Their metric works on binary, categorical, and ordinal labeling tasks. Fang *et al.* try to tackle the problem of data scarcity in crowd-labeling by using knowledge transfer from abundant unlabeled data [28]. They report that the approach helps to estimate annotator expertise better and improves performance. Li *et al.* propose a crowd targeting framework for selecting the best possible group of annotators for a specific task on binary and categorical data [29]. They introduce information gain as a measure of annotator competence and use EM based top-down and bottom-up approaches for selecting the best annotators. Jagabathula *et al.* propose a soft penalty scheme for the case of non-malicious annotators for binary labeled data [30]. For each sample, they count the number of times a given annotator agrees with other annotators and calculate the reciprocal of the harmonic mean of such quantities over all samples the given annotator has annotated. A hard penalty scheme is proposed for handling sophisticated adversaries. They use optimal semi-matchings with a quadratic cost function. Zhang *et al.* combine a reverse auction model with annotator quality and sample difficulty for conducting crowd-labeling under a budget constraint [31].

The problem of annotator reliability is a very popular subject and tackled in [32] by using Gaussian mixture models. Liu *et al.* approach this problem by using belief propagation and mean field methods [33]. Statistical methods are used for estimating annotator reliability and behavior [34], as well as including annotator parameters such as bias, expertise, and competence [12]. 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 [35]. Dutta *et al.* also deal with annotator quality in a crowdsourcing case study where multiple annotators provide high level categories for newspaper articles [36].

Annotators' varying expertise both among themselves and over different parts of the data are also factors affecting their reliability. Zhang *et al.* investigate annotator expertise with a combination of ML and MAP estimation [8]. 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 [37]. Varying annotator expertise problems are also handled in [38] and [13] with ground truth estimation, using MAP estimation and EM approach. Whitehill *et al.* study annotator expertise, taking noisy and adversarial annotators into account [39].

Detecting spammers/abusers, and biased annotators is useful for eliminating and/or modifying specific annotations. Spectral decomposition techniques are used for moderating abusive content in [40]. Raykar *et al.* propose an empirical Bayesian algorithm for iteratively eliminating spammers and estimating consensus labels from good annotators [14]. Wauthier *et al.* present a new Bayesian model for reducing annotator bias to combine the data collection, data curation and active learning [41].

1.1.2.2. Sample Selection Strategies. The problem of selecting the most suitable sample has attracted the interest of researchers. The selection criteria can depend on various factors such as informativeness or uncertainty. Donmez and Carbonell study the binary active learning problem by proposing a new sampling strategy [42]. They focus on selecting a suitable sample to include in an unsupervised learning scenario, where the annotator is considered to be infallible. Sheng *et al.* use noise-introduced benchmark datasets for sample selection strategies on binary classification problems [43]. Gao *et al.* propose an online profit estimation method that weeds out samples which do not need further annotations [44]. Lin *et al.* introduce variants of uncertainty sampling and propose impact sampling to select the most informative sample suited for the classifier [45]. Their method decides whether to obtain a new annotation for a readily annotated sample or to introduce a new sample to the crowd-labeled dataset. Khetan and Oh tackle the problem of binary active crowd-labeling by expending the annotation budget on difficult tasks [46]. They classify high and low confidence tasks in each annotation step and increase the budget allocation for more difficult tasks.

1.1.2.3. Joint Annotator and Sample Selection.

Some of the works in the literature deal with choosing the sample that needs to be annotated along with the most suitable annotator. Donmez and Carbonell [47] extend their earlier work [42] by considering multiple imperfect annotators and jointly select the optimum annotator-sample pair under a budget constraint. Hsueh *et al.* study the annotation selection problem by focusing on annotator noise, class label ambiguity, and the informativeness of a new annotation with regard to the classifier [48]. Tran-Thanh *et al.* investigate the trade-off between budget constraint and annotation quality [49, 50]. Nguyen *et al.* use a decision theoretic approach for choosing between acquiring labels from crowds and domain experts [51]. Their method selects a sample and annotator tuple to acquire an annotation. During this process, they account for the active sampling bias and estimate annotator accuracy.

1.1.2.4. Binary Annotation Problems.

Current literature on active crowd-labeling is mainly focused on binary annotation problems [26, 28, 31, 37, 42–45, 47–57]. We briefly survey the main tenets below.

Raykar and Agrawal model the crowdsourced labeling task sequentially with an epsilon-greedy exploration in a Markov Decision Process [53]. They use a utility function that considers label accuracy, cost and time. Li *et al.* deal with the budget allocation problem in crowd-labeling by using a Markov Decision Process in a sequential labeling scheme [58]. They propose a trade-off between label quality and quantity. Karger *et al.* define the crowd-labeling problem as a bipartite graph and show results supported by simulated binary data [59,60]. Their method is inspired by low-rank matrix approximation and belief propagation. Zhuang and Young verify and investigate the existence of in-batch annotation bias by using a factor graph based batch annotation model on binary data [55]. Ho *et al.* formulate the setting as a linear programming problem and work with the dual of the relaxed version. Their method requires the use of gold standard labels for assessing annotator quality and uses weighted majority voting for inferring the consensus [61]. Ho *et al.* treat the payment problem for crowdsourcing markets as a multi-armed bandit problem, where each arm represents the contract between a task and an annotator [57]. They propose a method called 'Agnostic Zooming' for selecting the most beneficial contract and study dynamic task pricing. This work focuses on annotator-sample pairing and deals with binary problems with the task giver's utility function as the main objective.

1.1.2.5. Categorical Annotation Problems. A relatively smaller portion of the existing work in the active crowd-labeling literature concentrates on categorical annotations [37,52,54,56,62–65]. These methods may also be adapted for binary annotations by considering only two categories. Yan *et al.* use uncertainty sampling for sample selection, along with learning annotator expertise on binary and categorical data [52]. Mozafari *et al.* propose two active learning algorithms based on sample uncertainty and a classifier's expected error [54]. The methods are tested on a variety of datasets. Zhu *et al.* [56] propose an online variant of the Dawid and Skene algorithm [18] that is motivated by online EM variants and stochastic approximation methods. Kamar *et al.* use the Galaxy Zoo dataset for the celestial object classification problem [62–64]. Galaxy Zoo is a crowdsourced effort mainly for the classification of different types of galaxies. They use Bayesian structure learning to incorporate the human and machine knowledge into the classification task in [62] and they tackle the problem of exploration-exploitation trade-off in worker hiring strategy by modeling the decision-making process as a Markov decision process in [63]. In [64], they focus on the problem of rectifying task-related bias of annotators and show that active learning with expert annotators can be used for alleviating bias. Venanzi *et al.* use a time-sensitive Bayesian aggregation method to estimate the labeling duration and annotator profile in crowdsourcing systems [65]. They detect bots, spammers or lazy annotators from the duration of their labeling process (either too short or too long). The study is carried out for categorical data.

1.1.2.6. Natural Language Processing Annotation Problems. In addition to categorical or numerical problems, natural language processing (NLP) is a popular area where crowd-labeling is preferred [66–68]. Ambati *et al.* use crowdsourcing for collecting translations from non-expert annotators [67]. They use inter-annotator agreement for translation reliability computation. Laws *et al.* use majority voting in the active learning scheme for crowd-labeling in the NLP domain [68]. The results show that active learning is to be preferred to random annotation selection.

1.1.2.7. Ordinal and Continuous Annotation Problems. Active crowd-labeling for continuous or ordinal valued annotations is a mostly unexplored research area. Marcus *et al.* make use of gold standard labels to identify low-quality or spammer annotators by a counting approach that combines several binary tasks into an ordinal task [69]. They also identify and avoid coordinated attacks from malicious annotators (*i.e.* Sybil attacks). Guo *et al.* deal with the problem of ordering objects in a set by aggregating pairwise comparison of said objects [70]. They devise a maximum likelihood formulation for finding the correct order of objects and show that this problem is NP-hard for their setting where all annotator accuracies are the same. However, their approach to active labeling focuses on the one-shot utilization of the additional budget. Welinder and Perona tackle the active crowd-labeling problem for continuous-valued annotations, by including the label uncertainty and annotator ability measurement in an EM based approach [37]. Their method detects and excludes spammers during the annotation process and is suitable for both binary and categorical data.

1.2. Novelty and Contributions

Annotating large datasets is a time consuming task and is usually expensive when the annotation task is outsourced to experts. The wisdom of crowds is an efficacious approach for this task in terms of budget and time constraints. However, inattentive annotators and spammers reduce the quality of annotations. Since desired sample labels are consensuses obtained from these annotations, it is beneficial to observe and understand the behavior of the annotators early on in the annotation process and take the necessary steps for improving the quality of consensus.

Our contributions in this study can be summarized as follows. First, we propose four new Bayesian models that model annotator behaviors for continuous or ordinal annotations to estimate the consensus scores [71]. 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.

Second, we show that the consensus scores estimated by the proposed models can be converted to categorical scores using simple techniques such as thresholding [71]. 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 show 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.

Third, we provide a new annotator scoring mechanism, which assigns a score to each annotator, representing the annotation quality of that annotator [71]. 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.

Fourth, we propose two active crowd-labeling methods [72] which produce continuous or ordinal valued consensus labels that can be further converted to binary or categorical labels by quantization, if necessary. The first method, O-CBS, focuses on improving the existing consensuses established from a set of previously collected annotations by selecting a sample-annotator pair for the next annotation. The second active crowd-labeling method, O-CBS+, is an extension of O-CBS. O-CBS+ eliminates the requirement of a readily available annotation set and is able to infer consensuses from scratch by means of annotator exploration/exploitation. Both methods target computational feasibility through a two-tier approach, where choosing a sample with low consensus quality is followed by choosing a high-quality annotator to annotate it. The two-tier approach makes both methods highly scalable and tractable. The proposed methods are data-independent, require no gold standard data to learn annotators, and are specifically designed for problems where the ground truth is not available or easily quantifiable.

Fifth, based on the variance of the sample's consensus posterior, we provide a novel formulation to estimate sample consensus quality, which corresponds to the total precision of the annotators that annotated the sample [72]. This scoring mechanism prevents budget exhaustion on confusing samples and provides a balanced sample selection.

Sixth, we address annotator selection problem in active crowd-labeling by introducing a family of annotator competence scoring functions that prevent annotator domination [72]. The dominance suppression mechanism that we introduce prevents ill-intentioned annotators from dominating the system and utilizes high-quality annotators in a balanced manner. We investigate the effects of both sample and annotator selection functions with extensive experiments on nine real-world datasets, three of which are introduced in the course of this thesis (Age Annotations dataset, Head Pose Annotations *Pan* and *Tilt* datasets).

Seventh, we study the effects of both a budget induced and a sample consensus quality induced stopping criteria with comparative experiments on all datasets [72]. The results show that O-CBS+ is an effective and budget-friendly (as low as one fifth of the original budget) active crowd-labeling method with high accuracy. Moreover, t-test results prove that it measures up to or surpasses contender algorithms.

Finally, we introduce a preliminary multivariate crowd-labeling model and solve it using a variational Bayes approach. Our model takes cross dimensional correlations in the datasets into account. Early results show that the proposed multivariate model performs on par with or better than univariate benchmark models and has significant potential for improvement.

2. DATASETS

In this chapter, we give the details of the datasets that we used to evaluate our work. First, we introduce two datasets: The Age Annotations dataset, and the Head Pose Annotations dataset (*tilt* and *pan* attributes). Second, we give the details of the Affective Text Analysis dataset (*anger*, *disgust*, *fear*, *joy*, *sadness*, and *surprise* attributes) [66]. Last, we describe the ELEA Personality Impressions Data [73]. Throughout this thesis, each attribute is also referred as a distinct dataset for our univariate models. Table 2.1 summarizes the datasets used in this thesis.

Most commonly, annotation schemes use binary symbols: Annotators choose between two options: These can be categorical options; such as male/female; or categorical classification of a continuous variable such as age in the form of young/old. The age of a person is actually a continuous variable; quantifying the time passed since birth. There are many ways of annotating this variable: One may choose a binary classification; such as young/old; or an ordinal representation, such as $0, 1, \ldots, 7$; each digit representing an age bracket. In this thesis, we regard the annotations as continuous variables to accommodate all possibilities, by adopting the most general representation. The continuous variables can be converted back to their ordinal or binary form by simple thresholding techniques when needed. This way, we are able to handle continuous, ordinal, or binary annotations.

For all datasets, annotations are linearly mapped to the range $[-3, 3]$ before processing. The results for the Head Pose Annotations datasets and the Age Annotations dataset are given in mean absolute degree and age error, respectively. Therefore, their inference results, which are in the range $[-3, 3]$, are linearly mapped to their related ground truth ranges (*i.e.* $[-90, 90]$ degrees and ages 0 through 69.)

As we mention in Section 1.1, the work of [37] is the only approach besides this work that estimates continuous-valued labels by means of active crowd-labeling. Thus, on the Head Pose Annotations and the Age Annotations datasets, we compare our results with the work of [37]. We also provide binarized comparisons with the work of [53] on the six Affective Text Analysis datasets.

Table 2.1. Annotation datasets used in this work. For evaluating our work, we introduce Age Annotations and Head Pose Annotations *tilt* and *pan* datasets. Additionally, we use six Affective Text Analysis [66] and ELEA Personality Impressions Data [73].

Dataset	Annotations	Samples	Annotators	Ground Truth Range	Annotation Range
Age Annotations (introduced in this work)	10020	1002	619	$\{0,\ldots,69\}$	$\{1,\ldots,7\}$
Head Pose Annotations: *tilt*, *pan* (introduced in this work)	5399	555	189	$\{-90,\ldots,90\}$	$\{1,\ldots,7\}$
Affective Text Analysis: *anger*, *disgust*, *fear*, *joy*, *sadness*, *surprise* [66]	1000	100	38	$\{0,\ldots,100\}$	$\{0,\ldots,100\}$
ELEA Personality Impressions Data: Big five personality traits [73]	306	102	5	$\{1,\ldots,7\}$	$\{1,\ldots,7\}$

2.1. Age Annotations Dataset

For evaluating our models, annotation datasets with ground truth values are invaluable. 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 [74] 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 2.1 shows some samples from this dataset and Figure 2.2 shows the age histogram of the dataset. The dataset consists mostly of baby, child, and young adult photos.

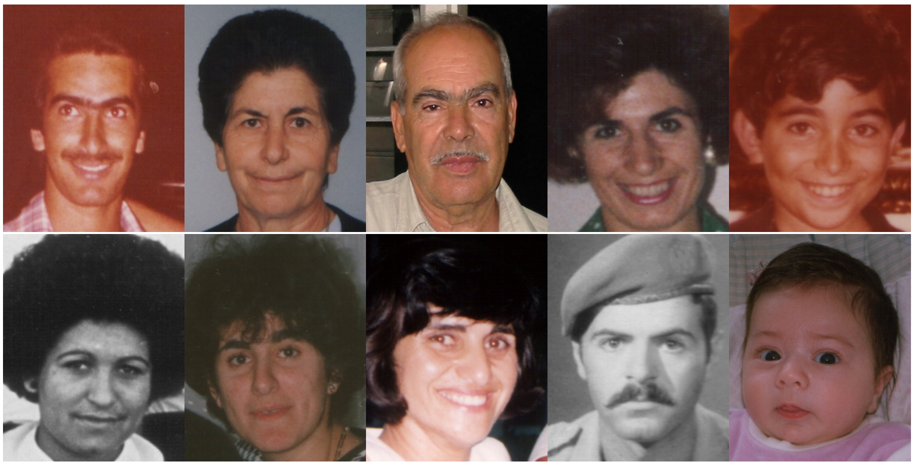


Figure 2.1. Sample images from the FGNet Aging Database

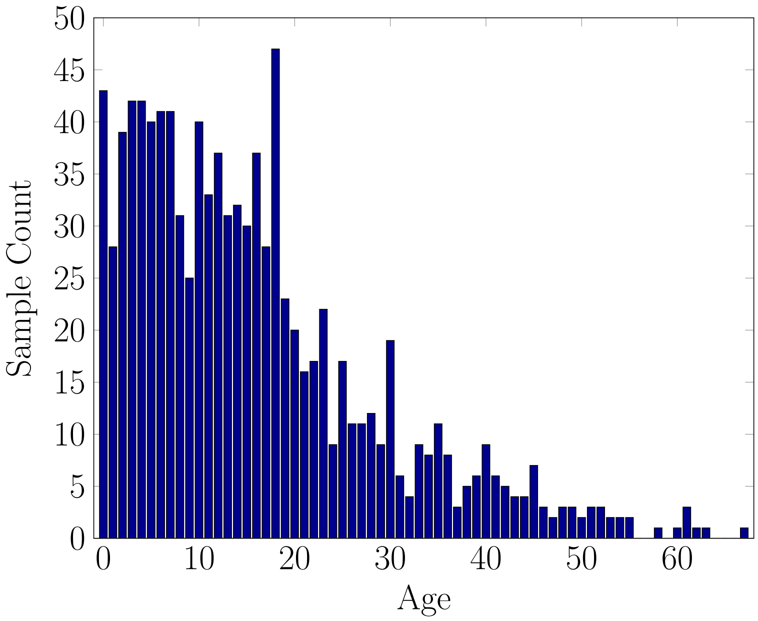


Figure 2.2. The FGNet Aging Database Age Histogram

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 2.2 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.

Table 2.2. Annotator workload for the Age Annotations Dataset (the number of annotations made by an annotator)

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

2.2. Head Pose Annotations Dataset

In addition to the Age Annotations Dataset, we also collected annotations for the Head Pose Image Database [75, 76]. The dataset has both *pan* and *tilt* ground truth values for each of the 2790 photos. The *tilt* values in the dataset are -90, -60, -30, -15, 0, +15, +30, +60, +90 degrees and the *pan* values are -90, -75, -60, -45, -30, -15, 0, +15, +30, +45, +60, +75, +90 degrees. The *pan-tilt* pairs used in the dataset result in 93 unique head pose configurations. There are two series of photos in which 15 subjects portrayed all of these configurations. Figure 2.3 shows sample images of the dataset. The ground truth values and the images were acquired using the setup shown in Figure 2.4.

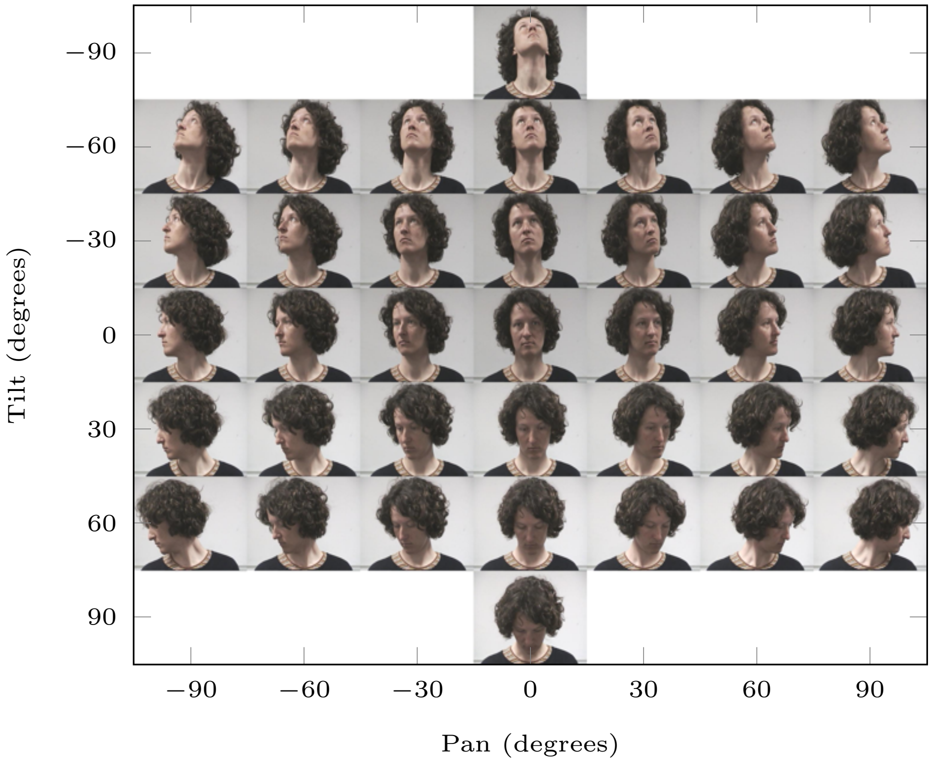
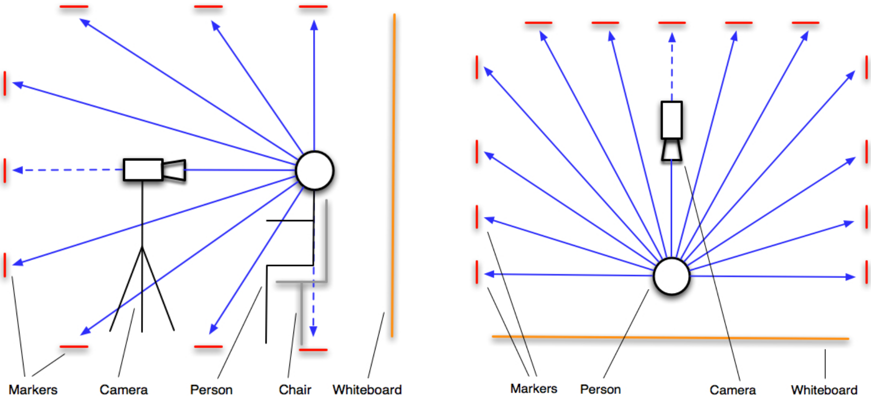


Figure 2.3. 37 distinct head poses of a person, which are chosen for the annotation tasks in the Head Pose Annotations datasets. The head pose images are taken from the Head Pose Image Database [76].



(a) Side View

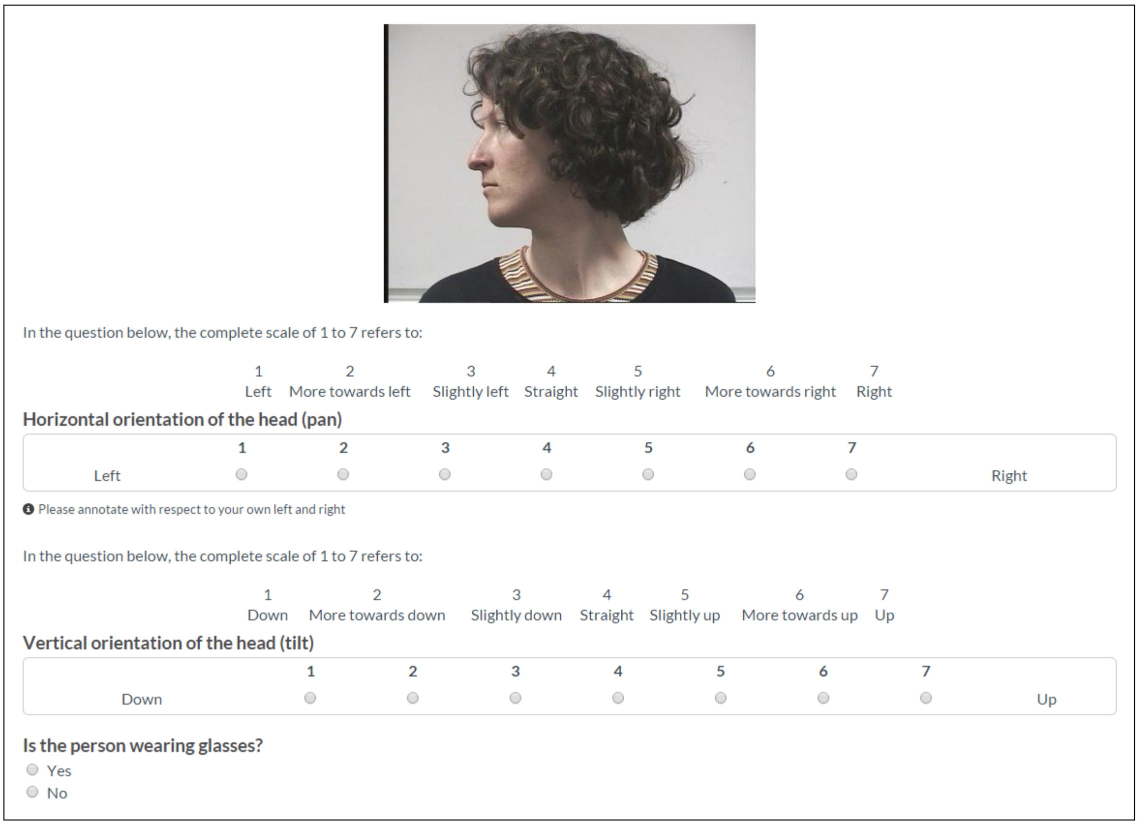
(b) Top View

Figure 2.4. Image acquisition setup of the Head Pose Image Database [75]

Due to budgetary constraints, we submitted a subset of these images to CrowdFlower [3] for annotation. We chose only one photo series for each subject. 6 subjects in the dataset wear glasses in one of their photo series. If available, we chose the photo series with glasses, otherwise the first series was used. We tried to choose a balanced combination of images with and without glasses. For *pan* and *tilt* values, we chose the photos with -90, -60, -30, 0, +30, +60, +90 degrees in both dimensions. A total of 555 photos were annotated. For each photo, we asked the participants to annotate three questions:

(i) Horizontal Orientation (*pan*): Left(1)-Right(7) (annotators' own left and right)
(ii) Vertical Orientation (*tilt*): Up(1) - Down(7)
(iii) Whether the person is wearing glasses or not.

Figure 2.5 shows a sample of what the annotators see when they are working on our head orientation tagging task.



In the question below, the complete scale of 1 to 7 refers to:

1	2	3	4	5	6	7
Left	More towards left	Slightly left	Straight	Slightly right	More towards right	Right

Horizontal orientation of the head (pan)

	1	2	3	4	5	6	7	
Left	○	○	○	○	○	○	○	Right

Please annotate with respect to your own left and right

In the question below, the complete scale of 1 to 7 refers to:

1	2	3	4	5	6	7
Down	More towards down	Slightly down	Straight	Slightly up	More towards up	Up

Vertical orientation of the head (tilt)

	1	2	3	4	5	6	7	
Down	○	○	○	○	○	○	○	Up

Is the person wearing glasses?

○ Yes
○ No

Figure 2.5. Head Pose Annotations Dataset sample question

In Table 2.3, we present the annotation frequency of the samples. Out of 555 samples, 475 have 9 annotations, with other samples having as few as 7 and as many as 17 annotations.

Table 2.3. Number of annotations per sample for the Head Pose Annotations Datasets

Sample annotation count	**7**	**8**	**9**	**15**	**16**	**17**
Number of samples	10	10	475	6	34	20

Table 2.4 shows the annotation frequency of the annotators, which we call annotator workload. A total of 189 annotators participated in the annotation tasks. Most common annotator workloads are multiples of 10 since many annotators completed the batch tasks assigned to them. For example, 61 annotators annotated 10 samples and 2 annotators annotated 100 samples.

Table 2.4. Annotator workloads for the Head Pose Annotations Datasets (the number of annotations made by an annotator)

Annotator workload	**5**	**10**	**17**	**20**	**24**	**30**	**39**	**40**	**45**	**50**	**55**	**60**	**70**	**75**	**80**	**84**	**90**	**100**
Number of annotators	1	61	1	45	1	26	1	15	2	13	1	7	5	1	4	1	2	2

2.3. Affective Text Analysis Datasets

Another group of datasets that we used for evaluating our methods are the six Affective Text Analysis datasets [66]. Each of these datasets has 1000 annotations on 100 short news headlines, drawn from various news sources [77], regarding positive and negative emotions. The task is to annotate a headline for each emotion, namely *anger*, *disgust*, *fear*, *joy*, *sadness*, and *surprise*. The annotators were asked to provide annotations in the interval of 0 to 100 for each emotion. 10 annotations per task were collected from 38 annotators using Amazon Mechanical Turk. The provided ground truth values are the averages of expert opinions.

Annotating emotions is a highly subjective task. There is no quantitative metric with which to measure the intensity of an emotion. Thus, the best possible approach is to consult experts and accept combinations of their opinions as the ground truth labels. However, comparing estimated labels obtained from crowd annotations with these ground truth values only establishes how well the crowd can estimate the average opinion of experts. Thus, it is very likely that high quality crowd opinions may be dismissed as subpar since they differ from the ground truth produced by only a few experts.

It is more common to express one's emotions in a state of existent/non-existent instead of on a scale of 0 to 100. Similarly, it is not easy for the annotator to annotate the emotion on such a fine scale. Therefore, a more practical approach is to compare the crowd's opinions against the experts' after binarization.

In light of these issues, we compare the binarized estimated labels with the binarized ground truth values for the six Affective Text Analysis datasets, as has been done in previous works that use this data [53]. Although we binarize the estimated output labels, we use the input annotations from the crowd as they are. By not binarizing the input annotations, we prevent the loss of valuable information, which may prove crucial for borderline decisions. Therefore, the results for Affective Text Analysis datasets are given as accuracies.

2.4. ELEA 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, 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 [17, 73, 78].

As a dataset to study personality, we used a subset from the Emergent LEAder (ELEA) corpus [79]. The ELEA AV subset consists of audio-visual recordings of 27 meetings, in which the participants perform a winter survival task with no roles as-

Table 2.5. Personality annotations per annotator on the ELEA data

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

signed. 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 [79, 80].

For each participant in the dataset, the personality impressions are obtained from external observers [73]. Ten Item Personality Inventory (TIPI) is used for measuring the Big Five personality traits of the participants [81]. 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 2.5 shows the number of annotations per annotator. More details on the annotations can be found in [73].

3. THE CROWD-LABELING PROBLEM DEFINITION

Assume that we consult R annotators to annotate a dataset of N samples. Due to budgetary and annotator availability constraints, it is not always possible to collect annotations for all samples from all annotators. Most often, each annotator annotates a few out of N samples and every sample is annotated by a small group of annotators. Out of all annotator-sample pairings ($N \times R$ possible annotations), we end up with $K \ll N \times R$ annotations. This is a common case for crowdsourced annotation tasks. The aim of our work is to choose these K annotations wisely and to infer high-quality consensus labels for all samples.

In Chapter 4, we work on how to infer high-quality consensus labels from a readily collected set of annotations. We call this problem passive crowd-labeling since we have no control over the annotation collection process. Active crowd-labeling is the case when we have control over the annotation collection process. In Chapters 5 to 7, we work on how to choose beneficial annotator-sample pairing during the course of active crowd-labeling.

Table 3.1. Variables pertaining to the crowd-labeling problem

Variable	**Description**
N	Number of samples
R	Number of annotators
K	Number of annotations
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
N_j	Annotation count of the j^{th} annotator
Y	$\{y_{1:K}\}$
X	$\{x_{1:N}\}$

In this thesis, collected annotation values are denoted as y_k where $k \in \{1, \ldots, K\}$ represents the annotation index. Additionally, the sample and the annotator indices of the annotation k are denoted as i_k and j_k, respectively. The sought consensus value of the i^{th} sample is denoted as x_i. Annotation count of the j^{th} annotator is represented with N_j. Table 3.1 summarizes these variables.

In the first section of this chapter, we elaborate on frequently encountered annotator behaviors which are the main motivation behind our methodology. In the following section, we give the definitions of some probability distributions that are used throughout this thesis.

3.1. 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 [14], dishonest annotators [82] or annotators who try to game the system by providing unrelated or nonsense answers [83]. In [14], annotator behaviors such as bias or maliciousness 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: Give annotations with low error rate
- Spammers: Give random annotations
- Adversaries: Give inverted rates
- Positively biased: Tend to give higher rates
- Negatively biased: Tend to give lower rates
- Unary annotators: Give the same rating to all samples

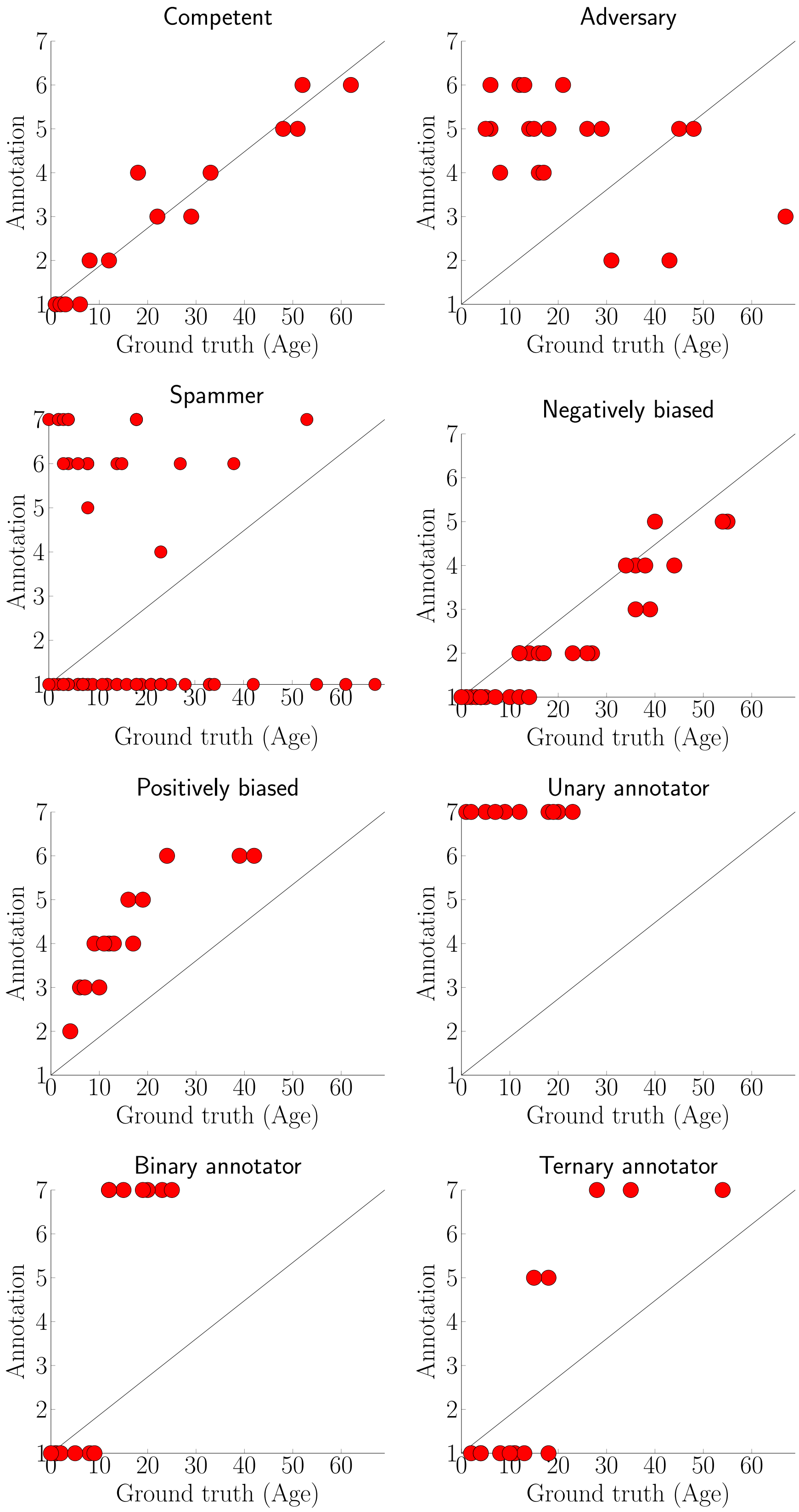


Figure 3.1. Real annotator examples from the Age Annotations Dataset. Each graph presents all annotations of a single annotator.

- Binary annotators: Give rates at the opposite ends of the scale
- 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 3.1 shows real annotations from the Age Annotations Dataset, produced by different types of annotators. Note that we do not try to classify annotator types, but we incorporate the behaviors of the annotator types for designing better models.

3.2. Probability Distributions

Definition 3.1 (Bernoulli Distribution). *Probability mass function of a Bernoulli random variable $k \in \{0,1\}$ with probability parameter $p \in (0,1)$ is given by*

$$\mathcal{B}\left(k;p\right) = p^k(1-p)^{1-k}.$$

Definition 3.2 (Gamma Distribution). *Probability density function of the Gamma distribution is*

$$\mathcal{G}\left(x;\alpha,\beta\right) = \frac{\beta^\alpha}{\Gamma\left(\alpha\right)} x^{\alpha-1} \exp\left(-x\beta\right)$$

where $x \geq 0$ and $\alpha, \beta > 0$, $\Gamma\left(\cdot\right)$ is the gamma function. Its mode is $\frac{\alpha-1}{\beta}$ when $\alpha \geq 1$.

Definition 3.3 (Normal Distribution). *Probability density function of a normally distributed variable x with mean μ and variance σ^2 is given by*

$$\mathcal{N}\left(x;\mu,\sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

and its cumulative distribution function is

$$\Phi\left(\frac{x-\mu}{\sigma}\right)=\frac{1}{2}\left(1+\operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right).$$

Definition 3.4 (Truncated Normal Distribution)**.** *Truncated normal distribution is the distribution of a bounded and normally distributed random variable. Probability density function of a truncated normally distributed variable x with parameters $\mu \in \mathbb{R}$, $\sigma > 0$, lower bound a, and upper bound b is given by*

$$\mathcal{N}_{trunc}\left(x;\mu,\sigma^2,a,b\right)=\frac{\mathcal{N}(x;\mu,\sigma^2)}{\Phi\left(\frac{b-\mu}{\sigma}\right)-\Phi\left(\frac{a-\mu}{\sigma}\right)}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

Remark. *When $a=-\infty$, $\Phi\left(\frac{a-\mu}{\sigma}\right)=0$; and when $b=\infty$, $\Phi\left(\frac{b-\mu}{\sigma}\right)=1$. Thus,*

$$\mathcal{N}_{trunc}\left(x;\mu,\sigma^2,-\infty,\infty\right)=\mathcal{N}\left(x;\mu,\sigma^2\right).$$

Now, we introduce a new distribution called the Generalized Positively Truncated Normal Distribution. To the best of our knowledge, this distribution does not appear in the statistics literature.

Definition 3.5 (Generalized Positively Truncated Normal Distribution)**.** *Probability density function of a generalized positively truncated normally distributed random variable $x > 0$ with parameters $\mu \in \mathbb{R}$, $\sigma > 0$, and $\alpha \geq 0$ is given by*

$$\mathcal{GPTN}\left(x;\mu,\sigma^2,\alpha\right)=\frac{1}{Z_{\mu,\sigma}(\alpha)}x^{\alpha}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}+\frac{\mu^2}{2\sigma^2}\right)$$

The normalization constant $Z_{\mu,\sigma}(\alpha)$ is

$$\begin{aligned}Z_{\mu,\sigma}(\alpha)=(\sigma\sqrt{2})^{\alpha+1}\Bigg(&\frac{1}{2}\Gamma\left(\frac{\alpha+1}{2}\right){}_1F_1\left(\frac{\alpha+1}{2};\frac{1}{2};\frac{\mu^2}{2\sigma^2}\right)\\&+\frac{\mu}{\sigma\sqrt{2}}\Gamma\left(\frac{\alpha}{2}+1\right){}_1F_1\left(\frac{\alpha}{2}+1;\frac{3}{2};\frac{\mu^2}{2\sigma^2}\right)\Bigg)\end{aligned}$$

where $\Gamma(\cdot)$ *is the gamma function, and* ${}_1F_1(\cdot;\cdot;\cdot)$ *is the confluent hypergeometric function of the first kind [84].*

Remark. *When* $\alpha = 0$*, the distribution reduces to the positively truncated normal distribution, i.e.* $\mathcal{GPTN}(x;\mu,\sigma^2,0) = \mathcal{N}_{trunc}(x;\mu,\sigma^2,0,\infty)$.

Theorem 3.1 (Moments of the Generalized Positively Truncated Normal Distribution)**.** *The* n^{th} *moment of the Generalized Positively Truncated Normal Distribution with parameters* α*,* μ*, and* σ *is*

$$E[x^n] = \frac{Z_{\mu,\sigma}(\alpha+n)}{Z_{\mu,\sigma}(\alpha)}.$$

Remark. *The mean of the distribution is* $\dfrac{Z_{\mu,\sigma}(\alpha+1)}{Z_{\mu,\sigma}(\alpha)}$ *and the variance is* $\dfrac{Z_{\mu,\sigma}(\alpha+2)}{Z_{\mu,\sigma}(\alpha)} - \left(\dfrac{Z_{\mu,\sigma}(\alpha+1)}{Z_{\mu,\sigma}(\alpha)}\right)^2$.

Proof.

$$\begin{aligned}
E[x^n] &= \int_0^\infty x^n \mathcal{GPTN}\left(x;\mu,\sigma^2,\alpha\right) dx \\
&= \int_0^\infty x^n \frac{1}{Z_{\mu,\sigma}(\alpha)} x^\alpha \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} + \frac{\mu^2}{2\sigma^2}\right) dx \\
&= \int_0^\infty \frac{1}{Z_{\mu,\sigma}(\alpha)} x^{\alpha+n} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} + \frac{\mu^2}{2\sigma^2}\right) dx \\
&= \frac{Z_{\mu,\sigma}(\alpha+n)}{Z_{\mu,\sigma}(\alpha)} \int_0^\infty \frac{1}{Z_{\mu,\sigma}(\alpha+n)} x^{\alpha+n} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} + \frac{\mu^2}{2\sigma^2}\right) dx \\
&= \frac{Z_{\mu,\sigma}(\alpha+n)}{Z_{\mu,\sigma}(\alpha)} \underbrace{\int_0^\infty \mathcal{GPTN}\left(x;\mu,\sigma^2,\alpha+n\right) dx}_{1} \\
&= \frac{Z_{\mu,\sigma}(\alpha+n)}{Z_{\mu,\sigma}(\alpha)}
\end{aligned}$$

□

Theorem 3.2 (Mode of the Generalized Positively Truncated Normal Distribution). *The mode of the distribution is*

$$\frac{\mu}{2} + \sqrt{\left(\frac{\mu}{2}\right)^2 + \alpha\sigma^2}.$$

Proof.

$$\begin{aligned}
\log(f) &= \alpha \log x - \frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x + Constant \\
0 = \frac{d\log(f)}{dx} &= \frac{\alpha}{x} - \frac{1}{\sigma^2}x + \frac{\mu}{\sigma^2} \\
0 &= -\frac{1}{\sigma^2}x^2 + \frac{\mu}{\sigma^2}x + \alpha \\
0 &= \frac{1}{\sigma^2}x^2 - \frac{\mu}{\sigma^2}x - \alpha \\
x &= \frac{\frac{\mu}{\sigma^2} \pm \sqrt{\left(\frac{\mu}{\sigma^2}\right)^2 + \frac{4\alpha}{\sigma^2}}}{\frac{2}{\sigma^2}} \\
&= \frac{\mu}{2} \pm \underbrace{\sqrt{\left(\frac{\mu}{2}\right)^2 + \alpha\sigma^2}}_{\geq |\mu/2| \text{ (since } \alpha \geq 0)}
\end{aligned}$$

Therefore for all μ, the only positive root of the equation is

$$x = \frac{\mu}{2} + \sqrt{\left(\frac{\mu}{2}\right)^2 + \alpha\sigma^2}.$$

□

Definition 3.6 (Multivariate Normal Distribution). *Probability density function of a normally distributed vector* $\boldsymbol{x} \in \mathbb{R}^d$ *with the mean vector* $\boldsymbol{\mu} \in \mathbb{R}^d$ *and the covariance matrix* $\boldsymbol{\Sigma}$ *is given by*

$$\mathcal{N}_d\left(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = |2\pi\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right)$$

where $|\cdot|$ *denotes the matrix determinant.*

Definition 3.7 (Matrix Normal Distribution)**.** *Probability density function of a normally distributed random matrix* $\boldsymbol{X}\,(n\times p)$ *with the location matrix* $\boldsymbol{M}\,(n\times p)$ *and the scale matrices* $\boldsymbol{U}\,(n\times n)$ *and* $\boldsymbol{V}\,(p\times p)$ *is given by*

$$\mathcal{MN}_{n,p}\left(\boldsymbol{X};\boldsymbol{M},\boldsymbol{U},\boldsymbol{V}\right)=\frac{\exp\left(-\frac{1}{2}\operatorname{Tr}\left(\boldsymbol{V}^{-1}(\boldsymbol{X}-\boldsymbol{M})^{\intercal}\boldsymbol{U}^{-1}(\boldsymbol{X}-\boldsymbol{M})\right)\right)}{(2\pi)^{np/2}|\boldsymbol{V}|^{n/2}|\boldsymbol{U}|^{p/2}}$$

where $\operatorname{Tr}(\cdot)$ *and* $|\cdot|$ *denote the trace and the matrix determinant, respectively.*

Remark. *The following statement defines the relation between the multivariate normal and the matrix normal distributions [85]:*

$$\boldsymbol{X}\sim\mathcal{MN}_{n,p}\left(\boldsymbol{X};\boldsymbol{M},\boldsymbol{U},\boldsymbol{V}\right)\iff\operatorname{vec}(\boldsymbol{X})\sim\mathcal{N}_{np}\left(\operatorname{vec}(\boldsymbol{X});\operatorname{vec}(\boldsymbol{M}),\boldsymbol{V}\otimes\boldsymbol{U}\right)$$

where $\operatorname{vec}(\cdot)$ *denotes the vectorization and* $\otimes$ *denotes the Kronecker product.*

Remark. *The following properties are shown to hold for the matrix normal distribution [85]*

$$\mathbb{E}\left[\boldsymbol{X}\right]=\boldsymbol{M}\tag{3.1}$$

$$\mathbb{E}\left[(\boldsymbol{X}-\boldsymbol{M})(\boldsymbol{X}-\boldsymbol{M})^{T}\right]=\boldsymbol{U}\operatorname{Tr}(\boldsymbol{V})\tag{3.2}$$

$$\mathbb{E}\left[(\boldsymbol{X}-\boldsymbol{M})^{T}(\boldsymbol{X}-\boldsymbol{M})\right]=\boldsymbol{V}\operatorname{Tr}(\boldsymbol{U})\tag{3.3}$$

$$\mathbb{E}\left[\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{T}\right]=\boldsymbol{U}\operatorname{Tr}(\boldsymbol{A}^{\intercal}\boldsymbol{V})+\boldsymbol{M}\boldsymbol{A}\boldsymbol{M}^{\intercal}\tag{3.4}$$

$$\mathbb{E}\left[\boldsymbol{X}^{\intercal}\boldsymbol{A}\boldsymbol{X}\right]=\boldsymbol{V}\operatorname{Tr}(\boldsymbol{U}\boldsymbol{A}^{\intercal})+\boldsymbol{M}^{\intercal}\boldsymbol{A}\boldsymbol{M}\tag{3.5}$$

$$\mathbb{E}\left[\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}\right]=\boldsymbol{U}\boldsymbol{A}^{\intercal}\boldsymbol{V}+\boldsymbol{M}\boldsymbol{A}\boldsymbol{M}\tag{3.6}$$

where $\operatorname{Tr}(\cdot)$ *denotes trace.*

Definition 3.8 (Wishart Distribution)**.** *Probability density function of the Wishart distribution is*

$$\mathcal{W}_{d}\left(\boldsymbol{X};\boldsymbol{\Psi},\nu\right)=\frac{|\boldsymbol{X}|^{\frac{\nu-d-1}{2}}}{2^{\frac{\nu d}{2}}|\boldsymbol{\Psi}|^{\frac{\nu}{2}}\boldsymbol{\Gamma}_{d}\left(\frac{\nu}{2}\right)}\exp\left(-\frac{1}{2}\operatorname{Tr}(\boldsymbol{\Psi}^{-1}\boldsymbol{X})\right)$$

where $\boldsymbol{X}$ *and* $\boldsymbol{\Psi}$ *are* $d{\times}d$ *positive definite matrices,* $\boldsymbol{\Gamma}_d\left(\cdot\right)$ *is the multivariate gamma function, and* $|\cdot|$ *denotes the matrix determinant. The expected value of* $\boldsymbol{X}$ *is* $\mathbb{E}\left[\boldsymbol{X}\right]{=}\nu\boldsymbol{\Psi}$.

Definition 3.9 (Categorical Distribution)**.** *Let* $\boldsymbol{z}$ *be a categorically distributed random 1-of-C binary vector with elements* $\boldsymbol{z}_c$ *and* $\boldsymbol{p}$ *be a vector where* $\boldsymbol{p}_c \in [0,1], \forall c$ *and* $\sum_{c=1}^{C} \boldsymbol{p}_c = 1$. *Then, the distribution of* $\boldsymbol{z}$ *is*

$$\mathcal{C}\left(\boldsymbol{z};\boldsymbol{p}\right) = \prod_{c=1}^{C} \boldsymbol{p}_c^{\boldsymbol{z}_c}.$$

The expected value of observing category c *is* $\mathbb{E}\left[\boldsymbol{z}_c\right] = \boldsymbol{p}_c$. *Note that since only one element of the vector* $\boldsymbol{z}$ *is 1,* $\sum_{c=1}^{C} \boldsymbol{z}_c = 1$.

4. PASSIVE CROWD-LABELING

In our approach, we want to cover as many diverse annotator behaviors as possible. We introduce three major annotator tendencies. The first one is the adverseness of the annotator, which describes whether or not the annotator provides inverted annotations. The second 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 third behavior the diversity of the opinion scale.

We assume that each 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

$$\mathcal{N}\left(y; \mu_\theta(x), \sigma_\theta^2\right) \tag{4.1}$$

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. Figure 4.1 shows a common Bayesian network for the proposed models.

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

$$\mathcal{L} = \log p(Y|X,\theta) + \log p(\theta) + \log p(X) \tag{4.2}$$

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}}\{\mathcal{L}\} \tag{4.3}$$

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

$$\frac{\partial \mathcal{L}}{\partial \theta_j} = 0, \forall \theta_j \in \theta. \tag{4.4}$$

The consensus rates are simultaneously inferred with the annotator parameters:

$$\frac{\partial \mathcal{L}}{\partial x_i} = 0, \forall x_i \in X. \tag{4.5}$$

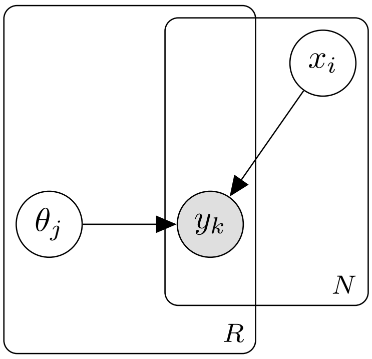


Figure 4.1. Bayesian network for the proposed models

Now, we propose four novel models which handle various annotator behaviors for univariate labels. In these models, x and y are scalars. Table 4.1 summarizes the annotator parameters, their domains, default values, and priors that appear in all four models. Table 4.2 shows a summary of the update equations for the consensus values and the parameters of the proposed models.

Table 4.1. Summary of annotator parameters

Parameter	Name	Domain	Default	Prior
Annotator Precision	λ_j	$\mathbb{R}_{>0}$	N/A	$\mathcal{G}(\lambda_j; \alpha_\lambda, \beta_\lambda)$
Adverseness	a_j	$\{-1, +1\}$	1	Flat
Opinion Scale	w_j	$\mathbb{R}_{>0}$	1	$\mathcal{G}(w_j; \beta_w + 1, \beta_w)$
Annotator Bias	b_j	$\mathbb{R}$	0	$\mathcal{N}(b_j; \mu_B, s_B^2)$

4.1. M-AH: Adversary Handling Model

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 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 centered around zero 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.

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. This 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

$$\lambda_j \sim \mathcal{G}\left(\lambda_j; \alpha_\lambda, \beta_\lambda\right). \tag{4.6}$$

We choose the hyperparameters as $\alpha_\lambda = 1.2$ 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 adverseness 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:

$$\begin{aligned} p(Y, X, \theta) &= \prod_{k=1}^{K} p(y_k | x_{i_k}, \lambda_{j_k}, a_{j_k}) \prod_{j=1}^{R} p(\lambda_j) p(a_j) \prod_{i=1}^{N} p(x_i) \\ &\propto \prod_{k=1}^{K} \mathcal{N}\left(y_k; a_{j_k} x_{i_k}, \frac{1}{\lambda_j}\right) \prod_{j=1}^{R} \mathcal{G}\left(\lambda_j; \alpha_\lambda, \beta_\lambda\right). \end{aligned} \tag{4.7}$$

Parameters of distinct annotators are independent of each other when X is given. Therefore, we are able to produce the update equations of each annotator's parameters seperately. That is, for finding the update equations of a specific annotator's parameters, we are only interested in the samples that are annotated by the said annotator. For calculating the update equations of x_i, λ_j, and a_j, we state and make use of Theorems 4.1 to 4.3.

4.1.1. Update Equation for the Consensus Value x in M-AH

Theorem 4.1 (Posterior distribution of x). *Let the distribution of y_k be*

$$\mathcal{N}\left(y_k; a_{j_k}(w_{j_k} x_{i_k} + b_{j_k}), \frac{1}{\lambda_{j_k}}\right).$$

Then, the posterior distribution of x_i is

$$x_i | \{y_k, \theta_{j_k} : k \in \mathcal{K}_i\} \sim \mathcal{N}\left(x_i; \frac{\sum\limits_{k:i_k=i} \lambda_{j_k} w_{j_k} (a_{j_k} y_k - b_{j_k})}{\sum\limits_{k:i_k=i} w_{j_k}^2 \lambda_{j_k}}, \left(\sum_{k:i_k=i} w_{j_k}^2 \lambda_{j_k}\right)^{-1} \right)$$

where $\theta_j = \{a_j, w_j, b_j, \lambda_j\}$ is the set of parameters of annotator j and $\mathcal{K}_i = \{k \in \mathcal{K} : i_k = i\}$ is the set of annotations of sample i.

Proof. See Appendix A.1. □

Using Theorem 4.1 for the setting in Equation 4.7, we find the posterior distribution of x_i as

$$\mathcal{N}\left(x_i; \frac{\sum\limits_{k:i_k=i} \lambda_{j_k} a_{j_k} y_k}{\sum\limits_{k:i_k=i} \lambda_{j_k}}, \left(\sum_{k:i_k=i} \lambda_{j_k} \right)^{-1} \right). \tag{4.8}$$

Then, the mode of this distribution is the update equation of x_i, which is

$$x_i = \frac{\sum\limits_{k:i_k=i} \lambda_{j_k} a_{j_k} y_k}{\sum\limits_{k:i_k=i} \lambda_{j_k}}. \tag{4.9}$$

4.1.2. Update Equation for the Precision Parameter λ in M-AH

Theorem 4.2 (Posterior distribution of λ)**.** *Let $x_k, y_k \in \mathbb{R}, \forall k \in \{1, \ldots, K\}$, $w > 0$, and $\lambda > 0$. If the distribution of y_k is $\mathcal{N}(y_k; wx_k, w^2\lambda^{-1})$, then the posterior distribution of λ is*

$$\mathcal{G}\left(\lambda; \frac{K}{2} + 1, \frac{1}{2} \sum_{k=1}^{K} \left(\frac{y_k}{w} - x_k \right)^2 \right).$$

Moreover, if the prior distribution of λ is $\mathcal{G}(\lambda; \alpha_\lambda, \beta_\lambda)$, then the posterior is

$$\mathcal{G}\left(\lambda; \frac{K}{2} + \alpha_\lambda, \beta_\lambda + \frac{1}{2} \sum_{k=1}^{K} \left(\frac{y_k}{w} - x_k \right)^2 \right).$$

Proof. See Appendix A.2. □

Considering the setting in Equation 4.7, we use Theorem 4.2 to find the posterior of λ_j . Then, we obtain the posterior as

$$\mathcal{G}\left(\lambda_j; \frac{N_j}{2} + \alpha_\lambda, \beta_\lambda + \frac{1}{2}\sum_{k:j_k=j}\left(\frac{y_k}{a_j} - x_{i_k}\right)^2\right). \tag{4.10}$$

The mode of this distribution is the update equation of λ_j, which is

$$\lambda_j = \frac{2(\alpha_\lambda - 1) + N_j}{2\beta_\lambda + \sum\limits_{k:j_k=j}(y_k - a_j x_{i_k})^2}. \tag{4.11}$$

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.

4.1.3. Update Equation for the Adverseness Parameter a in M-AH

Theorem 4.3 (Posterior distribution of a)**.** *Suppose that the values* $x_k, y_k \in \mathbb{R}, \forall k \in \{1, \ldots, K\}$ *and* $\lambda > 0$ *are given. Let* $c \sim \mathcal{B}(c; p)$ *and the distribution of* y_k *be* $y_k \sim \mathcal{N}(y_k; ax_k, \lambda^{-1})$ *where* $a = 2c - 1$*. Then the posterior distribution of* c *is*

$$\mathcal{B}\left(c; \left[1 + \exp\left(-2\lambda\sum_{k=1}^{K} y_k x_k\right)\right]^{-1}\right).$$

Moreover, the value a^* *that maximizes this distribution is given by*

$$a^* = \operatorname{sgn}\left(\sum_{k=1}^{K} y_k x_k\right).$$

Proof. See Appendix A.3. □

Using Theorem 4.3 for the setting in Equation 4.7, we find the update equation of a_j as

$$a_j = \operatorname{sgn}\left(\sum_{k:j_k=j} y_k x_{i_k}\right). \tag{4.12}$$

4.2. M-SH: Scale Handling Model

In addition to adversary handling of M-AH, we introduce opinion scale handling in M-SH. 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 that 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

$$w_j \sim \mathcal{G}\left(w_j; \beta_w + 1, \beta_w\right) \tag{4.13}$$

whose hyperparameters satisfy the mode of the distribution being equal to 1. We choose $\beta_w = 4$ 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

$$\begin{aligned} p(Y, X, \theta) =& \prod_{k=1}^{K} p(y_k | x_{i_k}, \lambda_{j_k}, a_{j_k}, w_{j_k}) \prod_{j=1}^{R} p(\lambda_j) p(a_j) p(w_j) \prod_{i=1}^{N} p(x_i) \\ \propto& \prod_{k=1}^{K} \mathcal{N}\left(y_k; a_{j_k} w_{j_k} x_{i_k}, \frac{1}{\lambda_j}\right) \prod_{j=1}^{R} \mathcal{G}\left(\lambda_j; \alpha_\lambda, \beta_\lambda\right) \\ & \prod_{j=1}^{R} \mathcal{G}\left(w_j; \beta_w + 1, \beta_w\right). \end{aligned} \tag{4.14}$$

For calculating the update equations of x_i, λ_j, a_j, and w_j, we make use of Theorems 4.1 to 4.4.

4.2.1. Update Equation for the Consensus Value x in M-SH

Using Theorem 4.1 for the setting in Equation 4.14, we find the posterior distribution of x_i as

$$\mathcal{N}\left(x_i; \frac{\sum\limits_{k:i_k=i} \lambda_{j_k} w_{j_k} a_{j_k} y_k}{\sum\limits_{k:i_k=i} {w_{j_k}}^2 \lambda_{j_k}}, \left(\sum_{k:i_k=i} {w_{j_k}}^2 \lambda_{j_k}\right)^{-1}\right). \tag{4.15}$$

Then, the mode of this distribution is the update equation of x_i, which is

$$x_i = \frac{\sum\limits_{k:i_k=i} \lambda_{j_k} w_{j_k} a_{j_k} y_k}{\sum\limits_{k:i_k=i} \lambda_{j_k} w_{j_k}^2}. \tag{4.16}$$

4.2.2. Update Equation for the Precision Parameter λ in M-SH

By using Theorem 4.2 for the setting in Equation 4.14, we find the posterior of λ_j as

$$\mathcal{G}\left(\lambda_j; \frac{N_j}{2} + \alpha_\lambda, \beta_\lambda + \frac{1}{2} \sum_{k:j_k=j} (y_k - a_j w_j x_{i_k})^2\right). \tag{4.17}$$

The mode of this distribution is the update equation of λ_j, which is

$$\lambda_j = \frac{2(\alpha_\lambda - 1) + N_j}{2\beta_\lambda + \sum\limits_{k:j_k=j} (y_k - a_j w_j x_{i_k})^2}. \tag{4.18}$$

4.2.3. Update Equation for the Adverseness Parameter a in M-SH

Using Theorem 4.3 for the setting in Equation 4.14, we find the update equation of a_j as

$$a_j = \operatorname{sgn}\left(\sum_{k:j_k=j} y_k x_{i_k}\right). \tag{4.19}$$

4.2.4. Update Equation for the Opinion Scale Parameter w in M-SH

Theorem 4.4 (Posterior distribution of w)**.** *Let $x_k, y_k \in \mathbb{R}, \forall k \in \{1, \ldots, K\}$, $w > 0$, and $\lambda > 0$. Let the distribution of y_k be $y_k \sim \mathcal{N}(y_k; wx_k, \lambda^{-1})$. Then, the posterior distribution of w is*

$$\mathcal{N}_{trunc}\left(w; \frac{\sum_{k=1}^{K} y_k x_k}{\sum_{k=1}^{K} x_k^2}, \left(\lambda \sum_{k=1}^{K} x_k^2\right)^{-1}, 0, \infty\right).$$

Moreover, if $w \sim \mathcal{G}(w; \alpha_w, \beta_w)$, then the posterior distribution of w becomes

$$\mathcal{GPTN}\left(x; \frac{\lambda \sum_{k=1}^{K} y_k x_k - \beta_w}{\lambda \sum_{k=1}^{K} x_k^2}, \left(\lambda \sum_{k=1}^{K} x_k^2\right)^{-1}, \alpha_w - 1\right).$$

Proof. See Appendix A.4. □

Applying Theorem 4.4 to the setting in Equation 4.14, we find the posterior of w_j as

$$\mathcal{GPTN}\left(w_j; \frac{\lambda_j a_j \sum\limits_{k:j_k=j} y_k x_{i_k} - \beta_w}{\lambda_j \sum\limits_{k:j_k=j} x_{i_k}^2}, \left(\lambda_j \sum_{k:j_k=j} x_{i_k}^2\right)^{-1}, \beta_w\right).$$

Then, we find the mode of this distribution using Theorem 3.2, which gives the update equation of w_j as

$$w_j = \frac{a_j \sum\limits_{k:j_k=j} y_k x_{i_k} - \frac{\beta_w}{\lambda_j}}{2 \sum\limits_{k:j_k=j} x_{i_k}^2} + \sqrt{\left(\frac{a_j \sum\limits_{k:j_k=j} y_k x_{i_k} - \frac{\beta_w}{\lambda_j}}{2 \sum\limits_{k:j_k=j} x_{i_k}^2}\right)^2 + \frac{\beta_w}{\lambda_j \sum\limits_{k:j_k=j} x_{i_k}^2}}. \tag{4.20}$$

4.3. M-ABS: Annotation Bias Sensitive Model

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 $\mu_{\theta_j}(x_i) = a_j(w_j x_i + b_j)$ where b_j represents either positive or negative bias. Since we model the bias as being unaffected by the opinion scale, b_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

$$b_j \sim \mathcal{N}\left(b_j; \mu_B, s_B^2\right). \tag{4.21}$$

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 $s_B = 0.05$ to allow some positive and negative bias.

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

$$\begin{aligned} p(Y, X, \theta) &= \prod_{k=1}^{K} p(y_k | x_{i_k}, \lambda_{j_k}, a_{j_k}, w_{j_k}, b_{j_k}) \prod_{j=1}^{R} p(\lambda_j) p(a_j) p(w_j) p(b_j) \prod_{i=1}^{N} p(x_i) \\ &\propto \prod_{k=1}^{K} \mathcal{N}\left(y_k; a_{j_k}(w_{j_k} x_{i_k} + b_{j_k}), \frac{1}{\lambda_j} \right) \prod_{j=1}^{R} \mathcal{G}\left(\lambda_j; \alpha_\lambda, \beta_\lambda\right) \\ &\prod_{j=1}^{R} \mathcal{G}\left(w_j; \beta_w + 1, \beta_w\right) \prod_{j=1}^{R} \mathcal{N}\left(b_j; \mu_B, s_B^2\right). \end{aligned} \tag{4.22}$$

For calculating the update equations of x_i, λ_j, a_j, w_j, and b_j, we make use of Theorems 4.1 to 4.5.

4.3.1. Update Equation for the Consensus Value x in M-ABS

Using Theorem 4.1 for the setting in Equation 4.22, we find the posterior distribution of x_i as

$$\mathcal{N}\left(x_i; \frac{\sum\limits_{k:i_k=i} \lambda_{j_k} w_{j_k} \left(a_{j_k} y_k - b_{j_k}\right)}{\sum\limits_{k:i_k=i} {w_{j_k}}^2 \lambda_{j_k}}, \left(\sum_{k:i_k=i} {w_{j_k}}^2 \lambda_{j_k} \right)^{-1} \right). \tag{4.23}$$

Then, the mode of this distribution is the update equation of x_i, which is

$$x_i = \frac{\sum\limits_{k:i_k=i} \lambda_{j_k} w_{j_k} (a_{j_k} y_k - b_{j_k})}{\sum\limits_{k:i_k=i} \lambda_{j_k} w_{j_k}^2}. \tag{4.24}$$

4.3.2. Update Equation for the Precision Parameter λ in M-ABS

By using Theorem 4.2 for the setting in Equation 4.22, we find the posterior of λ_j as

$$\mathcal{G}\left(\lambda_j; \frac{N_j}{2} + \alpha_\lambda, \beta_\lambda + \frac{1}{2}\sum_{k:j_k=j}(y_k - a_j(w_j x_{i_k} + b_j))^2\right). \tag{4.25}$$

The mode of this distribution is the update equation of λ_j, which is

$$\lambda_j = \frac{2(\alpha_\lambda - 1) + N_j}{2\beta_\lambda + \sum\limits_{k:j_k=j}(y_k - a_j(w_j x_{i_k} + b_j))^2}. \tag{4.26}$$

4.3.3. Update Equation for the Adverseness Parameter a in M-ABS

Using Theorem 4.3 for the setting in Equation 4.22, we find the update equation of a_j as

$$a_j = \operatorname{sgn}\left(\sum_{k:j_k=j} y_k(w_j x_{i_k} + b_j)\right). \tag{4.27}$$

4.3.4. Update Equation for the Opinion Scale Parameter w in M-ABS

Applying Theorem 4.4 to the setting in Equation 4.22, we find the posterior of w_j as

$$\mathcal{GPTN}\left(w_j; \frac{\lambda_j \sum\limits_{k:j_k=j}(a_j y_k - b_j)x_{i_k} - \beta_w}{\lambda_j \sum\limits_{k:j_k=j} x_{i_k}^2}, \left(\lambda_j \sum_{k:j_k=j} x_{i_k}^2\right)^{-1}, \beta_w\right).$$

Then, we find the mode of this distribution using Theorem 3.2, which gives the update equation of w_j as

$$w_j = \frac{\sum\limits_{k:j_k=j}(a_j y_k - b_j)x_{i_k} - \dfrac{\beta_w}{\lambda_j}}{2\sum\limits_{k:j_k=j} x_{i_k}^2} + \sqrt{\left(\frac{\sum\limits_{k:j_k=j}(a_j y_k - b_j)x_{i_k} - \dfrac{\beta_w}{\lambda_j}}{2\sum\limits_{k:j_k=j} x_{i_k}^2}\right)^2 + \frac{\beta_w}{\lambda_j \sum\limits_{k:j_k=j} x_{i_k}^2}}. \tag{4.28}$$

4.3.5. Update Equation for the Bias Parameter b in M-ABS

Theorem 4.5 (Posterior distribution of b). *Let $y_k \in \mathbb{R}, \forall k \in \{1, \ldots, K\}$, $b \in \mathbb{R}$, $w > 0$, and $\lambda > 0$. If the distribution of y_k is $y_k \sim \mathcal{N}(y_k; wb, w^2\lambda^{-1})$, then the posterior distribution of b is*

$$\mathcal{N}\left(b; \frac{1}{wK}\sum_{k=1}^{K} y_k, (K\lambda)^{-1}\right).$$

Moreover, if $b \sim \mathcal{N}\left(b; \mu_b, \lambda_b^{-1}\right)$, then the posterior distribution of b becomes

$$\mathcal{N}\left(b; \frac{\dfrac{\lambda}{w}\sum\limits_{k=1}^{K} y_k + \mu_b\lambda_b}{K\lambda + \lambda_b}, (K\lambda + \lambda_b)^{-1}\right).$$

Proof. See Appendix A.5. □

Applying Theorem 4.5 to the setting in Equation 4.22, we find the posterior of b_j as

$$\mathcal{N}\left(b_j; \frac{\lambda_j \sum\limits_{k:j_k=j}(a_j y_k - w_j x_{i_k}) + \dfrac{\mu_b}{s_B^2}}{N_j\lambda_j + \frac{1}{s_B^2}}, \left(N_j\lambda_j + \frac{1}{s_B^2}\right)^{-1}\right). \tag{4.29}$$

The mode of this distribution is the update equation of b_j, which is

$$b_j = \frac{a_j \displaystyle\sum_{k:j_k=j} y_k - w_j \sum_{k:j_k=j} x_{i_k} + \frac{\mu_B}{\lambda_j s_B^2}}{N_j + \frac{1}{\lambda_j s_B^2}}. \tag{4.30}$$

4.4. M-CBS: Consensus Bias Sensitive Model

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, b_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 $\mu_{\theta_j}(x_i) = a_j w_j(x_i + b_j)$. The prior on b_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 $\theta = \{\Lambda, A, W, B\}$. Thus, we have

$$\begin{aligned} p(Y, X, \theta) &= \prod_{k=1}^{K} p(y_k | x_{i_k}, \lambda_{j_k}, a_{j_k}, w_{j_k}, b_{j_k}) \prod_{j=1}^{R} p(\lambda_j) p(a_j) p(w_j) p(b_j) \prod_{i=1}^{N} p(x_i) \\ &\propto \prod_{k=1}^{K} \mathcal{N}\left(y_k; a_{j_k} w_{j_k}(x_{i_k} + b_{j_k}), \frac{w_{j_k}^2}{\lambda_{j_k}} \right) \prod_{j=1}^{R} \mathcal{G}\left(\lambda_j; \alpha_\lambda, \beta_\lambda\right) \\ & \prod_{j=1}^{R} \mathcal{G}\left(w_j; \beta_w + 1, \beta_w\right) \prod_{j=1}^{R} \mathcal{N}\left(b_j; \mu_B, s_B^2\right). \end{aligned} \tag{4.31}$$

For calculating the update equations of x_i, λ_j, a_j, w_j, and b_j, we make use of Theorems 4.1 to 4.3, 4.5 and 4.7.

4.4.1. Update Equation for the Consensus Value x in M-CBS

Theorem 4.6 (Posterior distribution of M-CBS x)**.** *Let the distribution of y_k be*

$$\mathcal{N}\left(y_k; a_{j_k} w_{j_k}(x_{i_k} + b_{j_k}), \frac{w_{j_k}^2}{\lambda_{j_k}}\right).$$

Then, the posterior distribution of x_i is

$$x_i | \{y_k, \theta_{j_k} : k \in \mathcal{K}_i\} \sim \mathcal{N}\left(x_i; \frac{\sum\limits_{k:i_k=i} \lambda_{j_k}(w_{j_k}^{-1} a_{j_k} y_k - b_{j_k})}{\sum\limits_{k:i_k=i} \lambda_{j_k}}, \left(\sum\limits_{k:i_k=i} \lambda_{j_k}\right)^{-1}\right)$$

where $\theta_j = \{a_j, w_j, b_j, \lambda_j\}$ is the set of parameters of annotator j and $\mathcal{K}_i = \{k \in \mathcal{K} : i_k = i\}$ is the set of annotations of sample i.

Proof. See Appendix A.6. □

Using Theorem 4.6 for the setting in Equation 4.31, we find the posterior distribution of x_i as

$$\mathcal{N}\left(x_i; \frac{\sum\limits_{k:i_k=i} \lambda_{j_k}\left({w_{j_k}}^{-1} a_{j_k} y_k - b_{j_k}\right)}{\sum\limits_{k:i_k=i} \lambda_{j_k}}, \left(\sum\limits_{k:i_k=i} \lambda_{j_k}\right)^{-1}\right). \tag{4.32}$$

Then, the mode of this distribution is the update equation of x_i, which is

$$x_i = \frac{\sum\limits_{k:i_k=i} \lambda_{j_k}\left(\frac{a_{j_k} y_k}{w_{j_k}} - b_{j_k}\right)}{\sum\limits_{k:i_k=i} \lambda_{j_k}}. \tag{4.33}$$

4.4.2. Update Equation for the Precision Parameter λ in M-CBS

By using Theorem 4.2 for the setting in Equation 4.31, we find the posterior of λ_j as

$$\mathcal{G}\left(\lambda_j; \frac{N_j}{2} + \alpha_\lambda, \beta_\lambda + \frac{1}{2}\sum_{k:j_k=j}\left(\frac{y_k}{w_j} - a_j(x_{i_k} + b_j)\right)^2\right). \tag{4.34}$$

The mode of this distribution is the update equation of λ_j, which is

$$\lambda_j = \frac{2(\alpha_\lambda - 1) + N_j}{2\beta_\lambda + \sum\limits_{k:j_k=j}\left(\frac{y_k}{w_j} - a_j(x_{i_k} + b_j)\right)^2}. \tag{4.35}$$

4.4.3. Update Equation for the Adverseness Parameter a in M-CBS

Using Theorem 4.3 for the setting in Equation 4.31, we find the update equation of a_j as

$$a_j = \operatorname{sgn}\left(\sum_{k:j_k=j} y_k(x_{i_k} + b_j)\right). \tag{4.36}$$

4.4.4. Update Equation for the Opinion Scale Parameter w in M-CBS

Theorem 4.7 (Mode of M-CBS w). *Let $y_k \in \mathbb{R}, \forall k \in \{1, \ldots, K\}$, $a \in \mathbb{R}$, $w > 0$, and $\lambda > 0$. If the distribution of y_k is $y_k \sim \mathcal{N}(y_k; awx_k, w^2\lambda^{-1})$ and $w \sim \mathcal{G}(w; \alpha_w, \beta_w)$, then the value w^* maximizing the posterior probability is a root of the equation*

$$w^{-3}\underbrace{\left(\lambda\sum_{k=1}^{K} y_k^2\right)}_{V_3} + w^{-2}\underbrace{\left(-\lambda a\sum_{k=1}^{K} y_k x_k\right)}_{V_2} + w^{-1}\underbrace{(\alpha_w - 1 - K)}_{V_1} + \underbrace{(-\beta_w)}_{V_0} = 0$$

Proof. See Appendix A.7. □

Applying Theorem 4.7 for the setting in Equation 4.31, we find that a root of the following cubic equation gives the desired value of w_j:

$$\begin{aligned} & V_3\left(\frac{1}{w_j}\right)^3 + V_2\left(\frac{1}{w_j}\right)^2 + V_1\left(\frac{1}{w_j}\right) + V_0 = 0 \text{ where} \\ & V_0 = -\beta_w \\ & V_1 = \beta_w - N_j \\ & V_2 = -\lambda_j a_j \sum_{k:j_k=j} y_k(x_{i_k} + b_j) \\ & V_3 = \lambda_j \sum_{k:j_k=j} y_k^2 \end{aligned} \tag{4.37}$$

Out of the solutions of Equation 4.37, the root maximizing the posterior is selected for the update of w_j.

4.4.5. Update Equation for the Bias Parameter b in M-CBS

Applying Theorem 4.5 to the setting in Equation 4.31, we find the posterior of b_j as

$$\mathcal{N}\left(b_j; \frac{\frac{\lambda_j a_j}{w_j}\sum_{k:j_k=j} y_k - \lambda_j \sum_{k:j_k=j} x_{i_k} + \frac{\mu_b}{s_B^2}}{N_j\lambda_j + \frac{1}{s_B^2}}, \left(N_j\lambda_j + \frac{1}{s_B^2}\right)^{-1}\right). \tag{4.38}$$

The mode of this distribution is the update equation of b_j, which is

$$b_j = \frac{\frac{a_j}{w_j}\sum_{k:j_k=j} y_k - \sum_{k:j_k=j} x_{i_k} + \frac{\mu_B}{\lambda_j s_B^2}}{N_j + \frac{1}{\lambda_j s_B^2}}. \tag{4.39}$$

Table 4.2. Table of update equations

Model	x_i	λ_j	a_j	w_j	b_j
M-AH	$\dfrac{\sum\limits_{k:i_k=i} \lambda_{j_k} a_{j_k} y_k}{\sum\limits_{k:i_k=i} \lambda_{j_k}}$	$\dfrac{2(\alpha_\lambda - 1) + N_j}{2\beta_\lambda + \sum\limits_{k:j_k=j} (y_k - a_j x_{i_k})^2}$	$\operatorname{sgn}\left(\sum\limits_{k:j_k=j} y_k x_{i_k}\right)$	N/A	N/A
M-SH	$\dfrac{\sum\limits_{k:i_k=i} \lambda_{j_k} w_{j_k} a_{j_k} y_k}{\sum\limits_{k:i_k=i} \lambda_{j_k} w_{j_k}^2}$	$\dfrac{2(\alpha_\lambda - 1) + N_j}{2\beta_\lambda + \sum\limits_{k:j_k=j} (y_k - a_j w_j x_{i_k})^2}$	$\operatorname{sgn}\left(\sum\limits_{k:j_k=j} y_k x_{i_k}\right)$	Equation 4.20	N/A
M-ABS	$\dfrac{\sum\limits_{k:i_k=i} \lambda_{j_k} w_{j_k} (a_{j_k} y_k - b_{j_k})}{\sum\limits_{k:i_k=i} \lambda_{j_k} w_{j_k}^2}$	$\dfrac{2(\alpha_\lambda - 1) + N_j}{2\beta_\lambda + \sum\limits_{k:j_k=j} (y_k - a_j(w_j x_{i_k} + b_j))^2}$	$\operatorname{sgn}\left(\sum\limits_{k:j_k=j} y_k (w_j x_{i_k} + b_j)\right)$	Equation 4.28	$\dfrac{a_j \sum\limits_{k:j_k=j} y_k - w_j \sum\limits_{k:j_k=j} x_{i_k} + \dfrac{\mu_B}{\lambda_j s_B^2}}{N_j + \frac{1}{\lambda_j s_B^2}}$
M-CBS	$\dfrac{\sum\limits_{k:i_k=i} \lambda_{j_k} \left(\dfrac{a_{j_k} y_k}{w_{j_k}} - b_{j_k}\right)}{\sum\limits_{k:i_k=i} \lambda_{j_k}}$	$\dfrac{2(\alpha_\lambda - 1) + N_j}{2\beta_\lambda + \sum\limits_{k:j_k=j} \left(\dfrac{y_k}{w_j} - a_j(x_{i_k} + b_j)\right)^2}$	$\operatorname{sgn}\left(\sum\limits_{k:j_k=j} y_k (x_{i_k} + b_j)\right)$	Equation 4.37	$\dfrac{\dfrac{a_j}{w_j} \sum\limits_{k:j_k=j} y_k - \sum\limits_{k:j_k=j} x_{i_k} + \dfrac{\mu_B}{\lambda_j s_B^2}}{N_j + \frac{1}{\lambda_j s_B^2}}$

4.5. Performance of Crowd Consensus Estimation Models

In this section, we first evaluate the performance of our models on annotation datasets with ground truth. We show how accurately the consensus values found by our models estimate the ground truth.

Then, 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.

4.5.1. Results on the Age Annotations Dataset

4.5.1.1. Accuracy of the Models in Estimating 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:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} |g(x_i) - z_i| \tag{4.40}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (g(x_i) - z_i)^2} \tag{4.41}$$

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.

Table 4.3. 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*

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 4.3 shows mean errors and standard deviations for the proposed and reference models. M-CBS outperforms all other models for all sets. Simpler 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 we can learn about the annotator's behavior. 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.

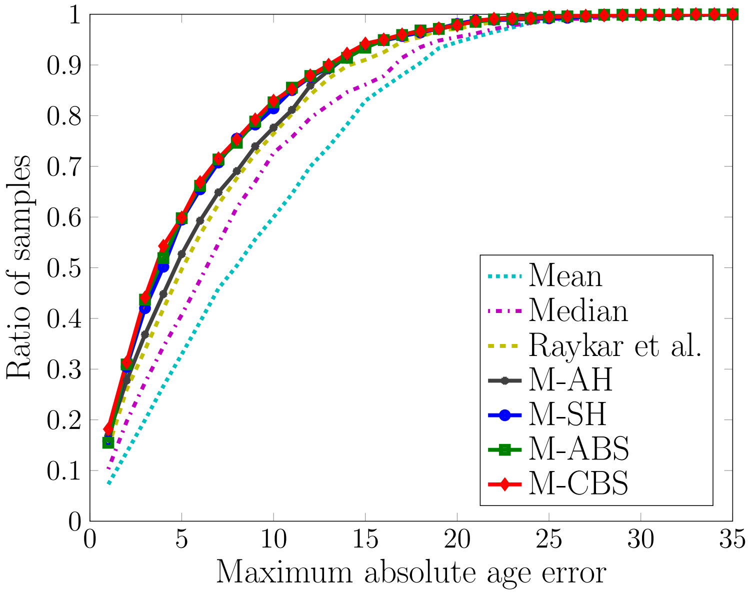


Figure 4.2. Cumulative match curves for the models

In Figure 4.2, 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 corresponding 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 4.3 shows the models' ground truth estimation performances of every 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

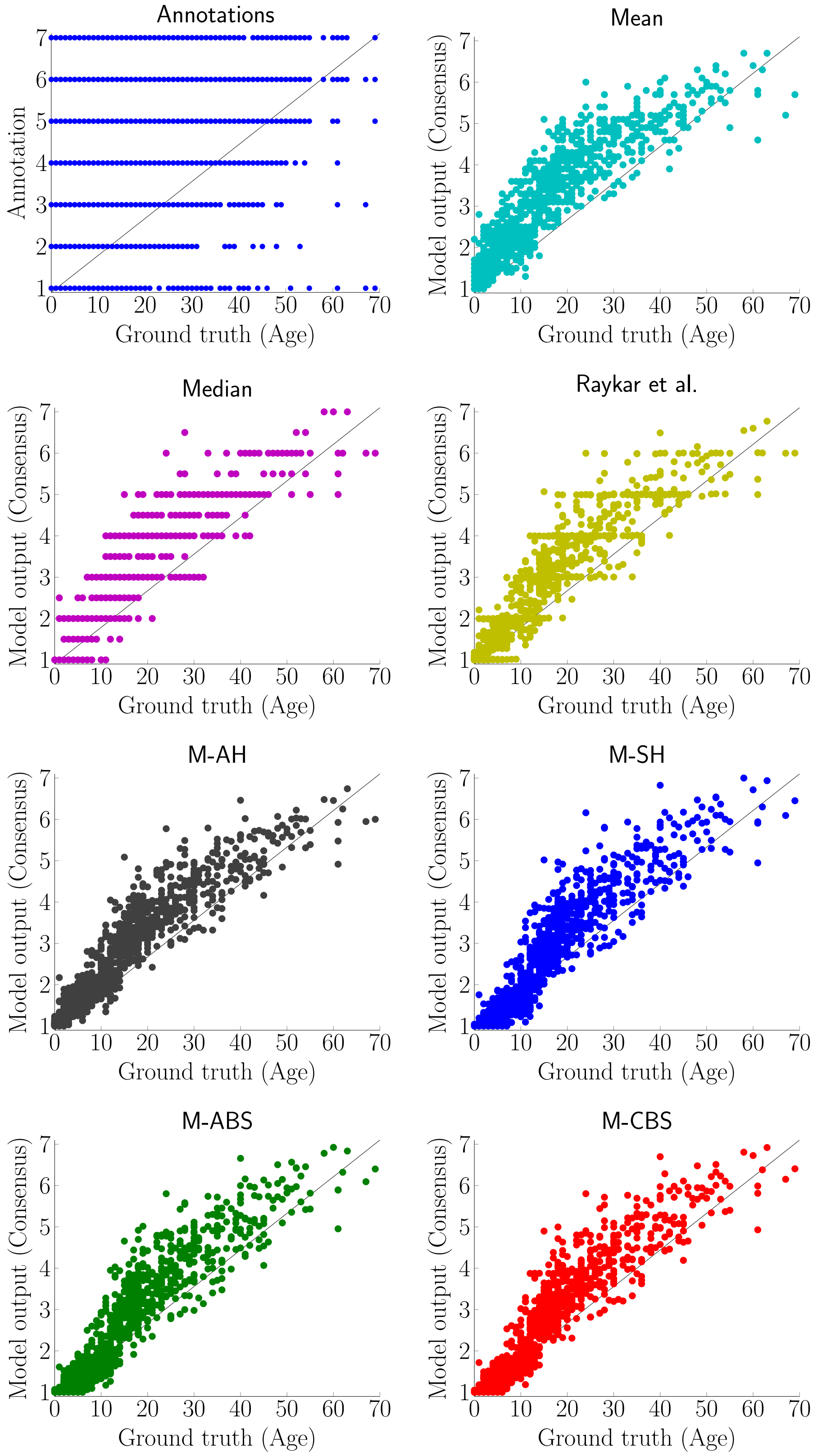


Figure 4.3. Ground truth estimation performance of models on joint set annotation data (The perfect fit would be on the diagonal)

power of the median model. The four models that we have proposed perform better as the model complexity increases.

4.5.1.2. 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 [12] work. Their method is suitable for 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 choose the median value during the binarization process. This is the reason for choosing 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 Age Annotations Dataset to be 'young' when they are less than 35, and 'old' otherwise.

In Table 4.4, 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

Table 4.4. The Matthews correlation coefficient, sensitivity, specificity, and accuracy measures for binarized results. For Welinder [12] 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 [12]	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

negative classes. Its value is between -1 and 1 where 1 is a result of perfect prediction. It is calculated as

$$\mathrm{MCC} = \frac{\mathrm{TP} * \mathrm{TN} - \mathrm{FP} * \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}} \tag{4.42}$$

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 [12] performs worse than the methods that accept continuous annotations, since it ignores lots of valuable information when binarizing the input annotations.

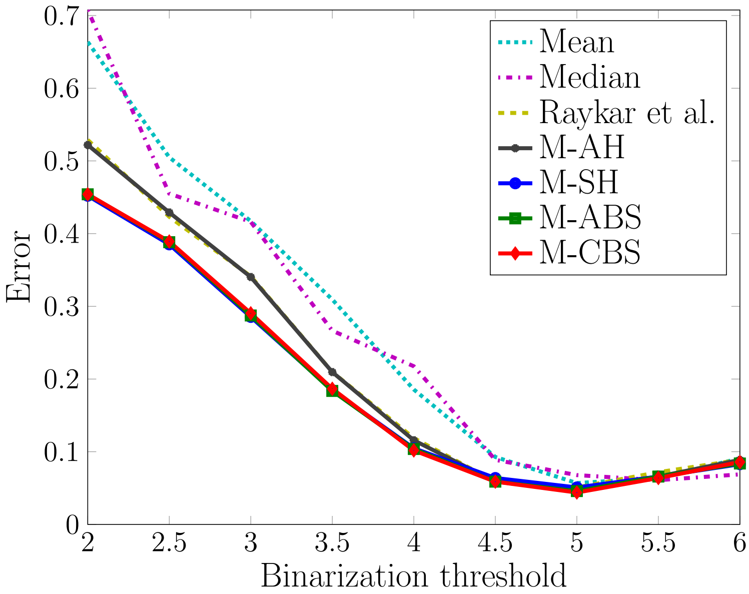


Figure 4.4. Change in error with respect to the change in consensus binarization threshold

In addition, we investigate the effect of different values of the consensus threshold for binarization in Figure 4.4. 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.

4.5.1.3. Discussion on Global Bias. With a careful look into Figures 4.3 and 4.4, 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_B) of the bias parameter(b)'s prior accordingly, we can decrease the global bias effect of the annotators. We empirically found that by setting $\mu_B = 0.7$, 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 4.5 shows errors when the global bias is compensated for. The errors reduce drastically when this effect is removed.

In Figure 4.5, we observe that the models estimate the ground truth better after we take the global bias into account. In Figure 4.5a and Figure 4.5b, the estimated

Table 4.5. Compensating for global bias: Errors of M-ABS and M-CBS with $\mu_B = 0.7$. The results are presented as mean and standard deviation for 100 repetitions.

(a) Mean absolute error

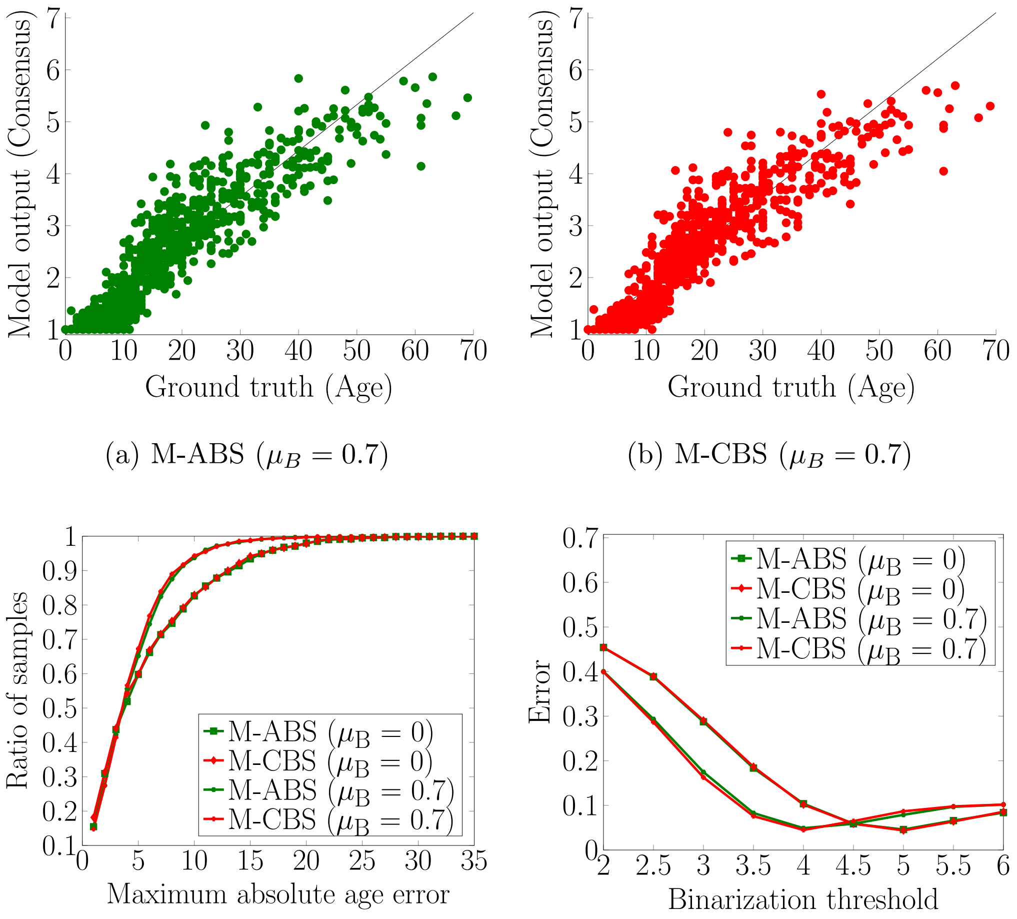
Model	**Set 1**	**Set 2**	**Joint**
M-ABS ($\mu_B = 0.7$)	4.52 ± 0.120	4.69 ± 0.102	4.24 ± 0.077
M-CBS ($\mu_B = 0.7$)	*4.44 ± 0.012*	*4.57 ± 0.007*	*4.14 ± 0.010*

(b) Root mean square error

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

consensus scores are closer to the ideal fit line. In the CMC plot in Figure 4.5c, 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 $\mu_B = 0.7$. Moreover, the binarization threshold shifts to four as one would expect (see Figure 4.5d).

An explanation of why this global bias exists for the Age 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 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.



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

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

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

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

Figure 4.5. Effect of removing global bias on the consensus scores

4.5.2. Results on ELEA Personality Impressions Data

In this section, we analyze the performance of the annotator models on a real dataset where there is no ground truth. We use the personality impressions as the 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.

4.5.2.1. Predicting Personality Impressions Using Nonverbal Cues. The nonverbal cues that we display in our everyday life, particularly during our interaction with others, contain significant information regarding our personality [86]. Psychologists have long investigated the links between the nonverbal cues that we display and our personality traits and have shown that several dimensions of personality are expressed through voice, face, body in the nonverbal channel [87]. In social computing literature, predicting personality using automatically extracted nonverbal cues has been addressed in several recent studies [17, 73, 78].

We use the data that is used in [73], where a large set of audio-visual nonverbal features are extracted and used in the prediction of personality. The set of features include attributes 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 [73]. For the current study, we use a concatenation of all the features used in [73] 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:

$$RAE = \frac{\sum_{i=1}^{N} |p_i - a_i|}{\sum_{i=1}^{N} |\bar{a}_i - a_i|} \tag{4.43}$$

where p is the value predicted by the regression model and a is the sample consensus value as estimated by the 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.

4.5.2.2. 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 4.6. Regression and classification results on extraversion prediction

Model Name	**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 4.6. 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 in agreement 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.

5. ACTIVE CROWD-LABELING METHODOLOGY

Passive crowd-labeling systems evaluate annotations after the completion of the acquisition phase. Thus, they are easily affected by erroneous annotations given by spammers and inattentive labelers. Each erroneous annotation means money wasted. It is important to be able to distinguish competent labelers from spammers and inattentive labelers early on in the labeling process for acquiring better annotations. Therefore, the most important questions would be: Which sample's label needs to be improved and which annotator should give the annotation? Active crowd-labeling is the process of collecting annotations with such concerns in mind. Smart selection of annotations also result in reduced annotation costs in addition to improved label qualities.

Carrying out a hands-on approach during the annotation acquisition process is in essence similar to active learning from the machine learning domain. In the classical sense, active learning draws its power from selecting the sample to be included in the learning process in a smart manner, thereby producing a well-trained algorithm with fewer samples. In classical active learning, the label of a sample is assumed to be provided by an annotator who always gives correct answers. In contrast, crowd-labeled instances may suffer from low quality annotations. The main motivation behind active crowd-labeling is to simultaneously select the most beneficial annotator-sample pair.

The process of active crowd-labeling is two-fold: One has to make good use of collected annotations, and also make a smart choice about which annotation to request next. The first part, which we call crowd consensus estimation, can be carried out by any of the models described in Chapter 4. The second stage has two components: how to select the sample to be annotated (Section 5.1) and how to select the annotator to annotate that sample (Section 5.2). Our primary concern is to improve every sample's consensus evenly. Therefore, we select the sample with the lowest consensus quality to be annotated. Once a sample is selected, we select the highest quality annotator for annotating it. This process is repeated with each new annotation in order to even out the sample consensus qualities across the whole dataset.

```
Input:
    Sets of all samples $\mathcal{I}$, all annotators $\mathcal{J}$, current annotations $\mathcal{K}$, currently active annotators $\mathcal{J}'$
function ACL($\mathcal{I}, \mathcal{J}, \mathcal{J}', \mathcal{K}$)
    EstimateLabels($\mathcal{I}, \mathcal{J}, \mathcal{K}$)
    repeat
        $k \leftarrow$ RequestAnnotation($\mathcal{I}, \mathcal{J}, \mathcal{J}', \mathcal{K}, \dots$)
        $\mathcal{K} \leftarrow \mathcal{K} \cup k$        ▷ Add the newly acquired annotation to the annotations set
        EstimateLabels($\mathcal{I}, \mathcal{J}, \mathcal{K}$)        ▷ Estimate consensus and relearn annotators
    until Budget limit or other stopping criteria are met
end function
```

Figure 5.1. ACL: Active Crowd-Labeling

Our approach consists of iteratively estimating crowd consensus and acquiring new annotations, as outlined in Figure 5.1. In this work, we denote the set of all samples to be annotated, the set of all annotators, and the set of current annotations as $\mathcal{I}$, $\mathcal{J}$, and $\mathcal{K}$, respectively. $\mathcal{J}'$ denotes the annotators that are currently in the system. Any of the models described in Chapter 4 can be used as the EstimateLabels($\cdot$) function used in Figure 5.1, which performs sample consensus estimation and annotator modeling. In Figures 6.1 and 7.1, we present two different approaches for the RequestAnnotation($\cdot$) function, the details of which are given in Chapters 6 and 7, respectively.

5.1. Which Sample Needs a New Label?

Since we want to improve our consensus estimations for the samples, we are in need of acquiring more annotations. Instead of randomly selecting samples for requesting annotations, a smarter strategy would reduce annotation costs while attaining high quality consensuses. The process of choosing which sample to annotate in a timely manner is of utmost importance since active crowd-labeling is a real-time process. Calculating the utility of all possible sample-annotator pairings for finding the optimal solution is often computationally very complex (at least $\mathcal{O}(nm)$) and poses scalability problems for large datasets and open annotator marketplaces. To this end, we opt for adopting a sub-optimal yet still beneficial approach to predict samples with low consensus quality by making use of readily available parameters inferred during the active crowd-labeling process.

During active crowd-labeling, our knowledge of a sample's consensus is gathered in its posterior distribution. Our motivation comes from the observation that a sample's quality may roughly be assessed by the variance of this posterior distribution. Using Bayesian rule on the full joint probabilities of the four models that we propose in Chapter 4, we find the posterior distributions of the consensus x_i as follows

M-AH:

$$x_i|\{y_k, a_{j_k}, \lambda_{j_k} : k \in \mathcal{K}_i\} \sim \mathcal{N}\left(x_i; \frac{\sum\limits_{k:i_k=i} \lambda_{j_k} a_{j_k} y_k}{\sum\limits_{k:i_k=i} \lambda_{j_k}}, \left(\sum\limits_{k:i_k=i} \lambda_{j_k}\right)^{-1} \right) \tag{5.1}$$

M-SH:

$$x_i|\{y_k, a_{j_k}, w_{j_k}, \lambda_{j_k} : k \in \mathcal{K}_i\} \sim \mathcal{N}\left(x_i; \frac{\sum\limits_{k:i_k=i} \lambda_{j_k} w_{j_k} a_{j_k} y_k}{\sum\limits_{k:i_k=i} {w_{j_k}}^2 \lambda_{j_k}}, \left(\sum\limits_{k:i_k=i} {w_{j_k}}^2 \lambda_{j_k}\right)^{-1} \right) \tag{5.2}$$

M-ABS:

$$x_i|\{y_k, \theta_{j_k} : k \in \mathcal{K}_i\} \sim \mathcal{N}\left(x_i; \frac{\sum\limits_{k:i_k=i} \lambda_{j_k} w_{j_k} \left(a_{j_k} y_k - b_{j_k}\right)}{\sum\limits_{k:i_k=i} {w_{j_k}}^2 \lambda_{j_k}}, \left(\sum\limits_{k:i_k=i} {w_{j_k}}^2 \lambda_{j_k}\right)^{-1} \right) \tag{5.3}$$

M-CBS:

$$x_i|\{y_k, \theta_{j_k} : k \in \mathcal{K}_i\} \sim \mathcal{N}\left(x_i; \frac{\sum\limits_{k:i_k=i} \lambda_{j_k} \left({w_{j_k}}^{-1} a_{j_k} y_k - b_{j_k}\right)}{\sum\limits_{k:i_k=i} \lambda_{j_k}}, \left(\sum\limits_{k:i_k=i} \lambda_{j_k}\right)^{-1} \right) \tag{5.4}$$

where $\theta_j = \{a_j, w_j, b_j, \lambda_j\}$ is the inferred set of annotator j's parameters and $\mathcal{K}_i = \{k \in \mathcal{K} : i_k = i\}$ is the set of sample i's annotations. For M-AH, the result is obtained by setting $b_{j_k} = 0$ and $w_{j_k} = 1$ for all j_k in either Theorem 4.1 or Theorem 4.6. For M-SH, the result is obtained by setting $b_{j_k} = 0$ for all j_k in Theorem 4.1. For M-ABS and M-CBS, the resulting distributions are obtained using Theorems 4.1 and 4.6, respectively.

Table 5.1. Sample score formulas for the proposed models. In these scores, λ_{j_k} are the precision parameters and w_{j_k} are the opinion scale parameters of every annotator j that has annotated sample i.

Model Name	**M-AH**	**M-SH**	**M-ABS**	**M-CBS**
$\boldsymbol{S_S(i)}$	$\sum\limits_{k:i_k=i} \lambda_{j_k}$	$\sum\limits_{k:i_k=i} {w_{j_k}}^2 \lambda_{j_k}$	$\sum\limits_{k:i_k=i} {w_{j_k}}^2 \lambda_{j_k}$	$\sum\limits_{k:i_k=i} \lambda_{j_k}$

The smaller the variance of the posterior distribution, the more confident we are on the inferred consensus and we want to request new annotations for the samples that we are less confident about. Thus, we use the reciprocal of the variance as a measure of consensus quality, namely the consensus quality score $S_S(i)$ of sample i. The consensus quality scores for all models are presented in Table 5.1, where λ_{j_k} are the precision parameters and w_{j_k} are the opinion scale parameters of every annotator j that has annotated sample i. These types of formulations are equivalent to counting the annotations of a sample weighted by its annotators' precision and opinion scale. Thus, a sample's consensus quality is only as good as the annotators' precision and opinion scale that have annotated it. Additionally, it also ensures that a sample's annotation count is also incorporated into its quality assessment. Note that adding a new annotation to an existing sample will definitely increase the sum and decrease the variance since w and λ values are positive. From a budget minimization point of view, it would be more beneficial to concentrate on those samples with the lowest scores. The approach that we present here is a fast (with complexity $\mathcal{O}(n)$) and reasonable way to reduce annotation costs and improve on the consensus values.

5.2. Who Annotates Better?

During the active crowd-labeling process, we need to identify competent annotators to utilize for new annotations. Thus, we need to rate annotators based on their competences. 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. So far, 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.

We would like to 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 4.1, we defined $\mu_\theta(\cdot)$ as the annotator function and in Table 5.2, 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:

$$\begin{aligned} p(x = \mu_\theta^{-1}(y), y|\theta) &= p(x = \mu_\theta^{-1}(y)|y, \theta)p(y|\theta) \\ &= p(y|x = \mu_\theta^{-1}(y), \theta)p(x = \mu_\theta^{-1}(y)) \\ &= \mathcal{N}\left(y; \mu_\theta(\mu_\theta^{-1}(y)), \sigma_\theta^2\right) \frac{1}{2c} \\ &= \frac{1}{2c\sigma_\theta\sqrt{2\pi}} \end{aligned} \tag{5.5}$$

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 4.1. Therefore, we have the following constraints:

$$-c \leq y \leq c \tag{5.6}$$

$$-c \leq x = \mu_\theta^{-1}(y) \leq c \tag{5.7}$$

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

$$\begin{aligned} -c \leq \mu_\theta^{-1}(y) \leq c \implies & \begin{cases} \mu_\theta(-c) \leq y \leq \mu_\theta(c) & \text{if } a_\theta = 1 \\ \mu_\theta(-c) \geq y \geq \mu_\theta(c) & \text{if } a_\theta = -1 \end{cases} \\ \implies & a_\theta\mu_\theta(-c) \leq a_\theta y \leq a_\theta\mu_\theta(c) \end{aligned} \tag{5.8}$$

By symmetry, we also have

$$-c \leq y \leq c \implies -c \leq a_\theta y \leq c \tag{5.9}$$

From Inequalities (5.8) and (5.9), we have

$$\underbrace{\min\{c, \max\{a_\theta \mu_\theta(-c), -c\}\}}_{d_\theta} \leq \underbrace{a_\theta y}_{r} \leq \underbrace{\max\{-c, \min\{a_\theta \mu_\theta(c), c\}\}}_{e_\theta} \tag{5.10}$$

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

$$\begin{aligned} l : [d_\theta, e_\theta] &\to \mathbb{R}^2 \\ r \mapsto (x(r), y(r)) &\implies r \mapsto (\mu_\theta^{-1}(a_\theta r), a_\theta r) \end{aligned} \tag{5.11}$$

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 $S_A(\theta)$ as the sum of the joint probabilities along the path l:

$$\begin{aligned} S_A(\theta) &= \int_l p(x, y|\theta) ds \\ &= \int_{d_{\mu_\theta}}^{e_{\mu_\theta}} p(\mu_\theta^{-1}(a_\theta r), a_\theta r|\theta) \|l'(r)\| dr \\ &= \frac{1}{2c\sigma_\theta \sqrt{2\pi}} \int_{d_{\mu_\theta}}^{e_{\mu_\theta}} \|l'(r)\| dr \end{aligned} \tag{5.12}$$

Figure 5.2 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.

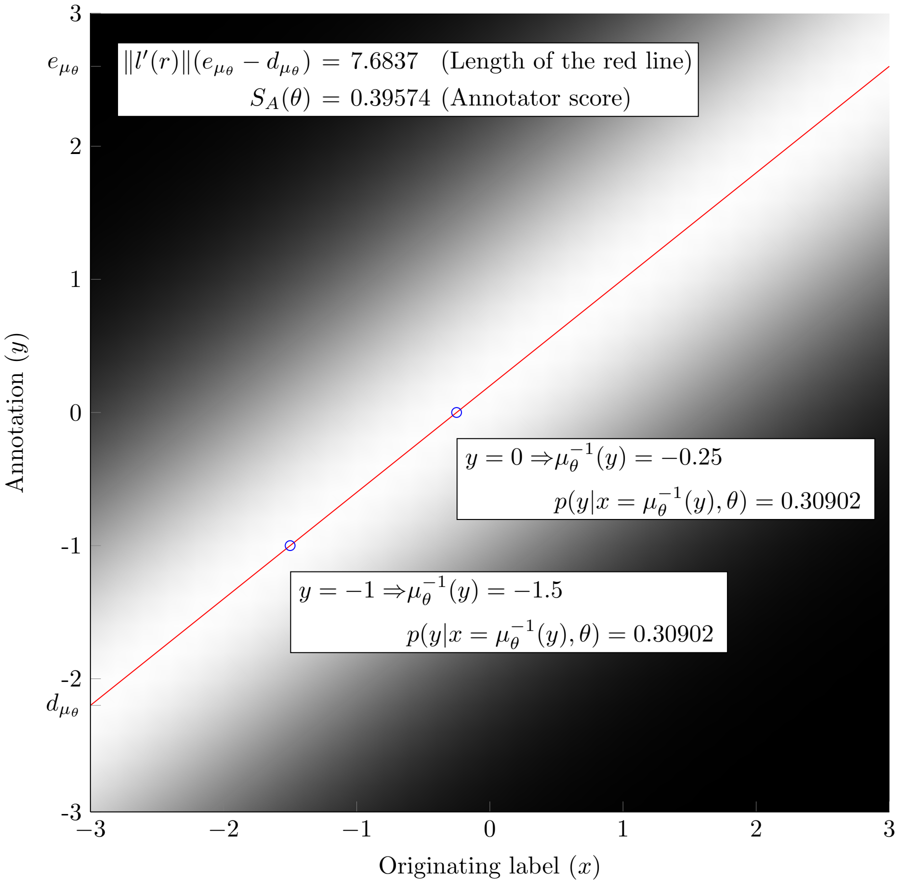
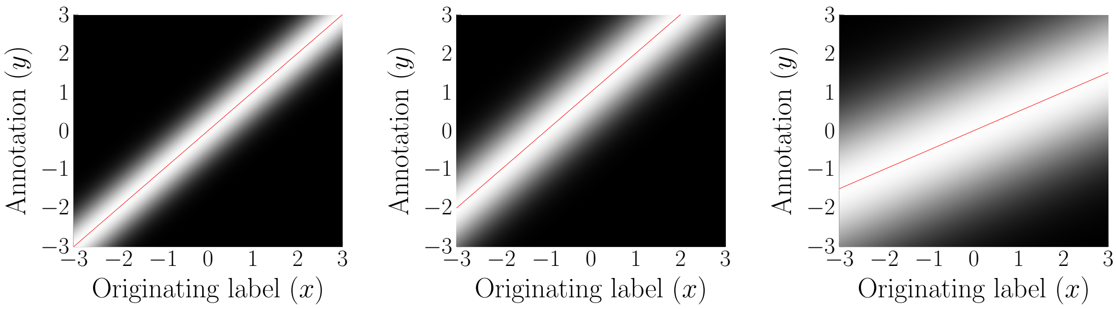


Figure 5.2. Score calculation for an annotator with parameters $a = 1$, $w = 0.8$, $b = 0.2$, $\lambda = 0.6$. The annotator is modeled using M-ABS. The intensity values depict the probability of the annotator rating a sample with respect to the ground truth. Brighter areas indicate a higher probability.

In Figure 5.3, we present three examples of annotators commonly encountered in crowd-labeling problems. Grayscale values represent posterior probability of annotation value ($p(y|x, \theta)$); the higher the intensity, the higher the probability. The red line is the peak of this distribution. For very competent annotators, w_j is close to 1 and b_j is close to 0. Additionally, they have high λ_j values resulting in a concentrated band of annotations around the peak. In contrast, inattentive annotators have lower λ_j values which result in more scattered annotations.



(a) Very competent annotator

(b) Positively biased annotator

(c) Inattentive annotator

Figure 5.3. Three examples of annotators: Very competent, positively biased, and inattentive.

Table 5.2 shows the derived annotator score formulas for the proposed models. A specific annotator j's score is denoted as $S_A(\theta_j)$, or in shorthand notation, $S_A(j)$. Note that d_{μ_j} and e_{μ_j} depend on the related model's $\mu_\theta(\cdot)$ function and their definition is given in Equation 5.10. It is also notable to mention that $S_A(j)$ does not depend on annotations or samples; it only depends on the parameters of the annotator.

Table 5.2. Annotator score formulas for the proposed models

Model Name	$\boldsymbol{\mu_j}$	$\boldsymbol{\sigma_j^2}$	$\boldsymbol{\lVert l'(r) \rVert}$	$\boldsymbol{S_A(j)}$
M-AH	$a_j x$	$\frac{1}{\lambda_j}$	$\sqrt{2}$	$\sqrt{\frac{\lambda_j}{\pi}}\left(e_{\mu_j} - d_{\mu_j}\right)$
M-SH	$a_j w_j x$	$\frac{1}{\lambda_j}$	$\sqrt{1 + \frac{1}{w_j^2}}$	$\frac{1}{w_j}\sqrt{\frac{\lambda_j\left(1 + {w_j}^2\right)}{2\pi}}\left(e_{\mu_j} - d_{\mu_j}\right)$
M-ABS	$a_j w_j x + b_j$	$\frac{1}{\lambda_j}$	$\sqrt{1 + \frac{1}{w_j^2}}$	$\frac{1}{w_j}\sqrt{\frac{\lambda_j\left(1 + {w_j}^2\right)}{2\pi}}\left(e_{\mu_j} - d_{\mu_j}\right)$
M-CBS	$a_j w_j\left(x + b_j\right)$	$\frac{{w_j}^2}{\lambda_j}$	$\sqrt{1 + \frac{1}{w_j^2}}$	$\frac{1}{{w_j}^2}\sqrt{\frac{\lambda_j\left(1 + {w_j}^2\right)}{2\pi}}\left(e_{\mu_j} - d_{\mu_j}\right)$

In Figure 5.4, we demonstrate the change in annotator scores using the formulas for M-CBS with respect to w and b when the variance is fixed. When selecting our priors, we preferred w to be around 1 and b to be around 0. By examining Figure 5.4, 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,

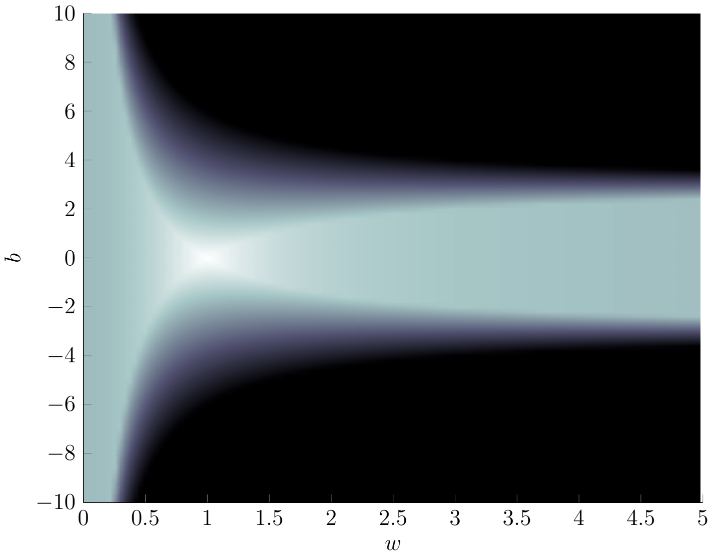


Figure 5.4. The change in annotator scores with respect to w and b parameters of M-CBS when the variance is fixed. Higher intensities correspond to higher annotator scores.

it does not matter which rate they give. In this case, the effect of b diminishes and the annotator scores do not vary for different b. 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.

5.2.1. How Beneficial Is Annotator Scoring?

We discussed the importance of identifying competent annotators and proposed a scoring metric. Now, we elaborate on the annotator scores calculated on real data for different models and how to make use of these scores.

First, we show the robustness of our scoring mechanism across different models. Figure 5.5 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

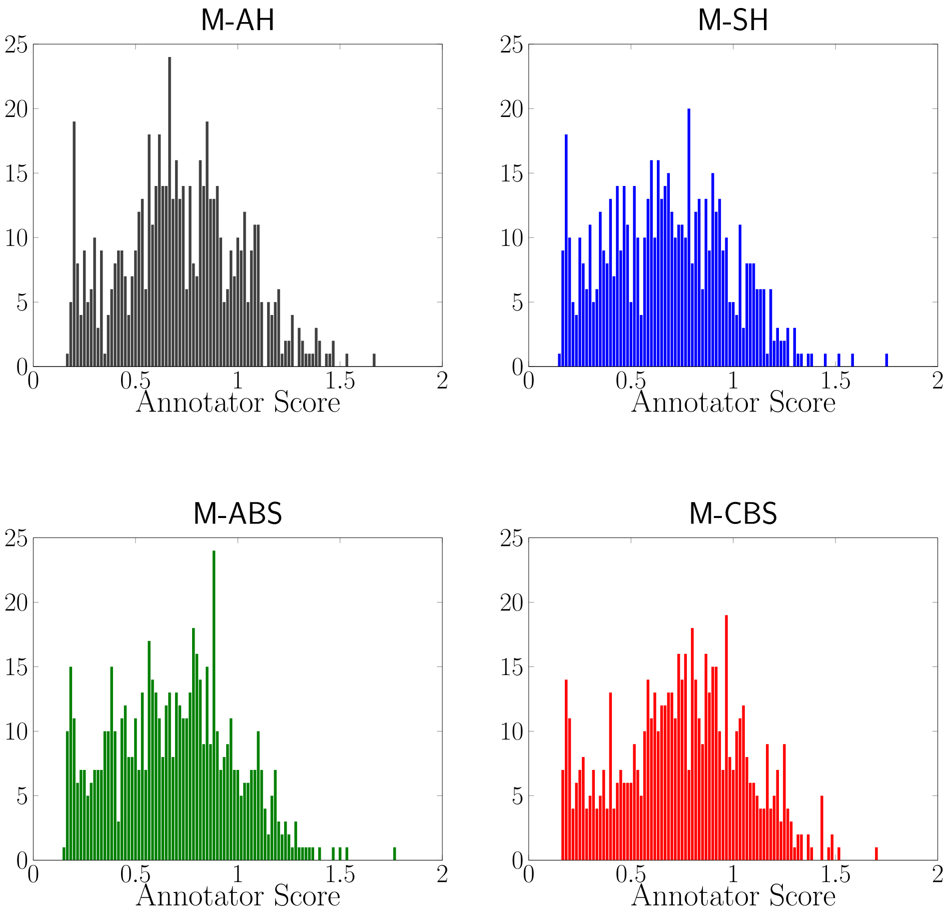


Figure 5.5. Annotator score histograms for the proposed models

behavior is that, a higher complexity model finds a higher quality consensus in which the annotators' individual opinions are represented better.

In Figure 5.6, 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 an 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.

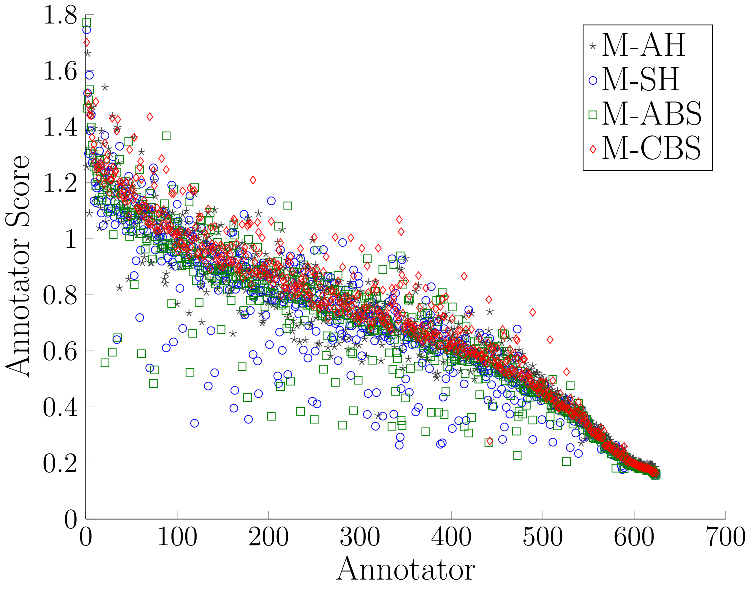
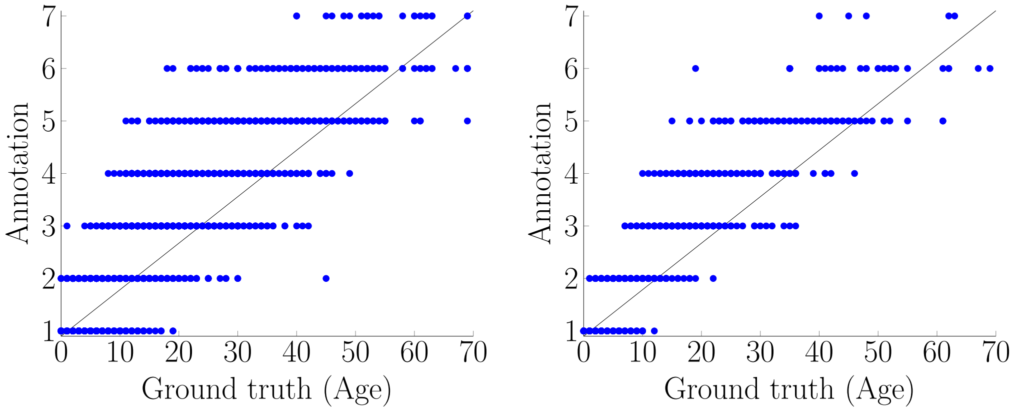


Figure 5.6. Annotator score comparison for the proposed models



(a) Top 50%

(b) Top 10%

Figure 5.7. The annotations of top scoring annotators.

Figure 5.7 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 4.3. 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 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 5.3. 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 5.3 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 ($\in \{0, \dots, 69\}$) 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.

6. O-CBS: IMPROVING THE EXISTING CONSENSUS USING ACTIVE CROWD-LABELING

When dealing with annotation problems, the task at hand often requires working with a limited pool of annotators, especially when the subject requires expert annotators. However, due to budget and/or time constraints, each annotator annotates only a subset of all samples. Although we can infer a preliminary consensus, later on we may want to reconsult the same annotators for the samples that they did not annotate beforehand in order to improve the consensuses.

In this section, we propose an annotation collection and consensus improvement method for the situation mentioned above, which we call O-CBS (Online M-CBS). Figure 6.1 gives the details of the annotation requesting mechanism for improving the existing consensus. We first need to identify which sample's consensus is not satisfactory and needs to be improved the most. The algorithm expects a sample consensus quality scoring function which measures trustworthiness of the consensus estimation and gives higher results when the estimation on the consensus is more trustworthy. Then, the sample with the least consensus quality score is selected to be improved. The related sample consensus quality score function introduced in Table 5.1 is a suitable choice.

The second part of the problem is the selection of the most suitable annotator for the selected sample. For this, we need an annotator competence scoring function that gives higher scores for more competent annotators. Finally, we ask the annotator with the highest competence score for a new annotation for the selected sample.

O-CBS is based on Figure 5.1 with M-CBS as the EstimateLabels($\cdot$) function and Figure 6.1 as the RequestAnnotation($\cdot$) function. In this setting, RequestAnnotation($\cdot$) employs $S_S(i) = \sum\limits_{k:i_k=i} \lambda_{j_k}$ (M-CBS row in Table 5.1) as the sample consensus quality scoring function. We investigate a family of annotator competence

```
Input:
    Sets of all samples $\mathcal{I}$, all annotators $\mathcal{J}$, current annotations $\mathcal{K}$, currently active annotators $\mathcal{J}'$
    $i_k$ and $j_k$ are the sample and annotator of annotation $k$, respectively
    $S_S(\cdot)$ and $S_A(\cdot)$ are the sample consensus quality function and annotator competence scoring function, respectively. (We assume that $S_S$ and $S_A$ are intrinsically aware of the annotator parameters $a$, $w$, $b$, and $\lambda$)
Output: New annotation $k$
function RequestAnnotation($\mathcal{I}, \mathcal{J}, \mathcal{J}', \mathcal{K}, S_S(\cdot), S_A(\cdot)$)
    for all $i \in \mathcal{I}$ do
        $\mathcal{K}_i \leftarrow \{k \in \mathcal{K} : i_k = i\}$    ▷ Annotations of sample $i$
        $\mathcal{J}_i \leftarrow \{j_k \in \mathcal{J} : k \in \mathcal{K}_i\}$    ▷ Annotators of sample $i$
    end for
    $i \leftarrow \underset{i' \in \mathcal{I} \text{ s.t. } \mathcal{J}' \setminus \mathcal{J}_{i'} \neq \emptyset}{\operatorname{argmin}} S_S(i')$    ▷ Select the sample with the worst consensus quality such that at least one of the currently active annotators has no annotations for that sample
    $j \leftarrow \underset{j' \in \mathcal{J}' \setminus \mathcal{J}_i}{\operatorname{argmax}} S_A(j')$    ▷ Select the most competent annotator from the set of active annotators who had not annotated sample $i$
    $k \leftarrow$ Request an annotation for sample $i$ from annotator $j$
    return $k$
end function
```

Figure 6.1. RequestAnnotation: Requesting annotation for improving the existing consensus

scoring functions, and we denote O-CBS with such different functions $(S_A, S_A^{\mathcal{K}}, S_A^1, \dots)$ as O-CBS$(\cdot)$. As a baseline method, we use O-CBS$\left(S_A^{\mathcal{R}}\right)$ which employs S_S for sample selection but selects annotators randomly. As another baseline method, we use O-CBS$(Random)$ which is a special case where the sample consensus quality scoring and the annotator competence scoring functions are both replaced with random selection.

6.1. Effectiveness of the Sample Scoring Function S_S

Since S_S is our choice of sample selection strategy in O-CBS, we start with presenting its performance by comparing it against random sample selection. In Figures 6.2 and 6.3, we observe the effectiveness of using the sample scoring function S_S across nine datasets. We report the MAE on the Age Annotations and the Head Pose Annotations datasets. On the Affective Text Analysis datasets, we report the accuracy. The graphs show that S_S is a favorable sample selection strategy across all datasets in terms of mean absolute error and accuracy. Especially in *pan*, *anger*, *joy*, and *sadness* datasets, there is a significant improvement over random sample selection. Although O-CBS$\left(S_A^{\mathcal{R}}\right)$ falls behind O-CBS$(Random)$ in the *fear* and *surprise* datasets

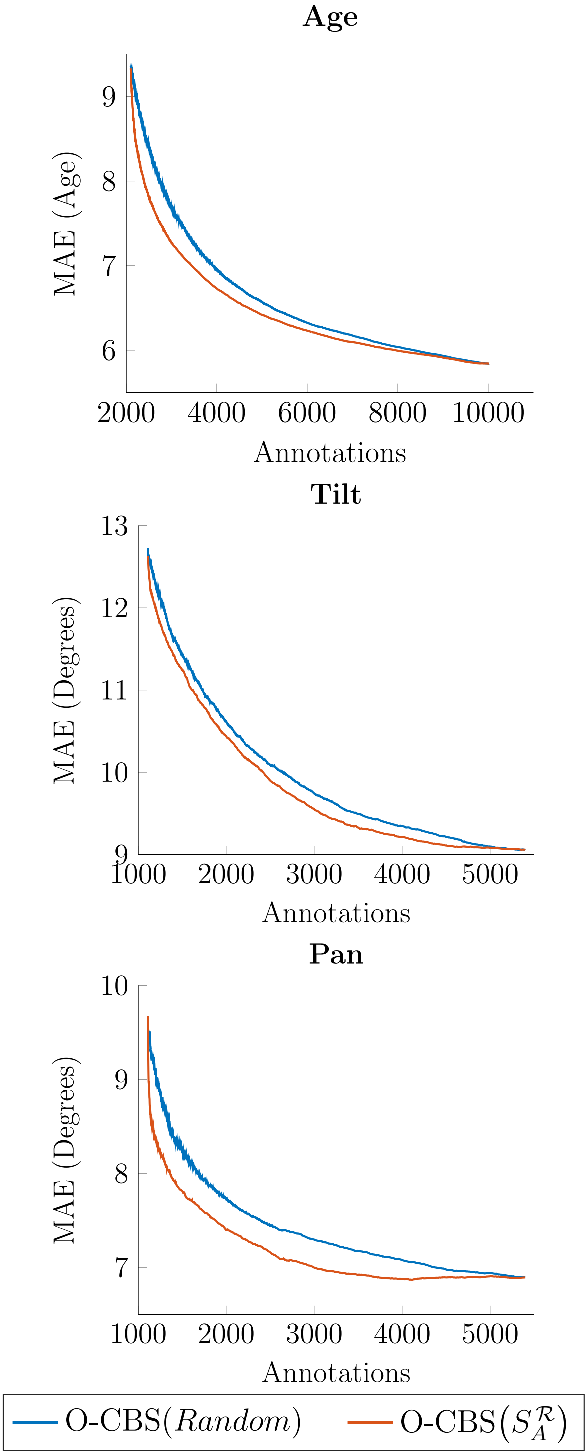


Figure 6.2. Effect of using S_S for sample selection on the Age Annotations and the Head Pose Annotations datasets, averaged over 100 runs with different starting subsets. O-CBS(*Random*) employs both random sample and random annotator selection, whereas O-CBS$\left(S_A^{\mathcal{R}}\right)$ employs random selection only for annotators and uses S_S for sample selection.

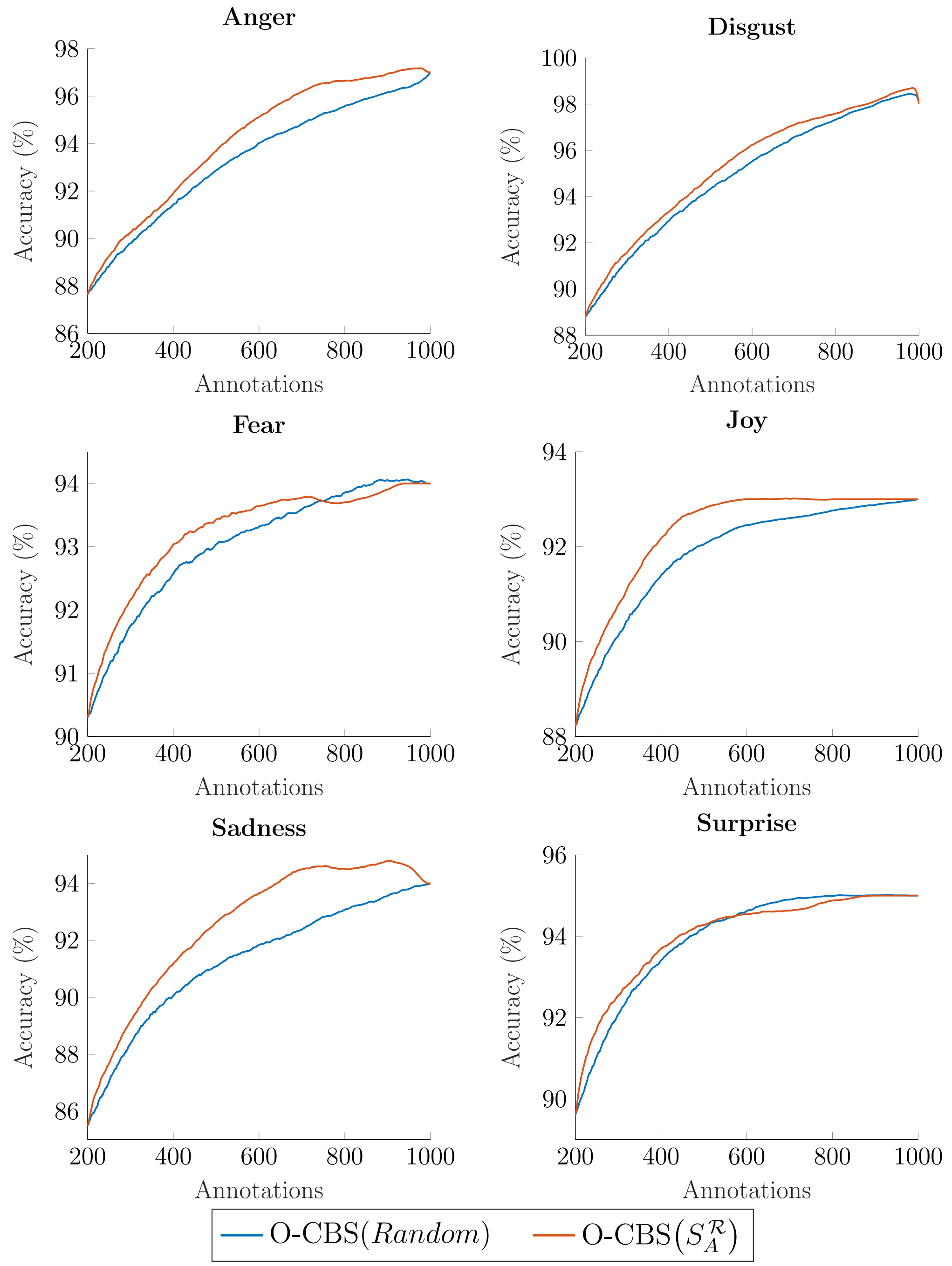


Figure 6.3. Effect of using S_S for sample selection on the Affective Text Analysis datasets, averaged over 100 runs with different starting subsets. O-CBS($Random$) employs both random sample and random annotator selection, whereas O-CBS$\left(S_A^{\mathcal{R}}\right)$ employs random selection only for annotators and uses S_S for sample selection.

as the number of annotations increases, the overall performance of S_S is beneficial. Even in the absence of an annotator selection strategy, S_S by itself provides significant improvement to active crowd-labeling performance.

6.2. Balancing the Scales: Suppressing Annotator Domination

The annotator competence scoring function for M-CBS described in Table 5.2 satisfies the aforementioned requirement of giving higher scores for more competent annotators. In this section, we discuss the shortcomings of the said annotator competence scoring function and propose several updates to alleviate these shortcomings.

Since our focus is on crowd annotation problems without any gold standard, we trust the consensus of the crowd to be true. However, it is possible that the majority of the crowd might be wrong or ill-intentioned. Moreover, ill-intentioned annotators are inclined to annotate more samples for gaining more money, resulting in an unbalanced system.

The stability of a crowd grows when more people are in it and the crowd-labeling approach is more susceptible to the actions of said people when the crowd is small. If the system is dominated by incompetent annotators, whenever a competent annotator joins the system, their opinion will be treated as an outlier and good annotators will have a low annotator competence score due to the mechanism introduced in Section 5.2. Since the active crowd-labeling method is inclined to acquire new annotations from the high scoring annotators, the method will continue requesting annotations mainly from incompetent annotators. Even if more truly competent annotators join the system, it may prove to be challenging to balance the scales in favor of them. Therefore, it is crucial to prevent annotator overloading early on and to let the method concentrate on competent annotators later on.

For overcoming these issues, we introduce a weighting factor to the annotator scoring mechanism proposed in Section 5.2. The idea is to suppress the annotator scores $S_A(j)$ proportionally to the annotator workloads so that the score of highly loaded annotators are suppressed. Additionally, we want to reduce this effect as the system gets more reliable in terms of annotations. We call this weighting factor *the dominance suppression factor*, which is

$$\left|\mathcal{K}^j\right|^{-\varphi\frac{\left|\mathcal{J}^1\right|}{|\mathcal{K}|}} \tag{6.1}$$

where $\varphi > 0$ is *the dominance suppression coefficient* which controls the effect of the weight, $|\mathcal{K}|$ is the current number of annotations, $|\mathcal{K}^j|$ is the number of annotations of annotator j, and $|\mathcal{J}^1|$ is the number of annotators that have at least one annotation.

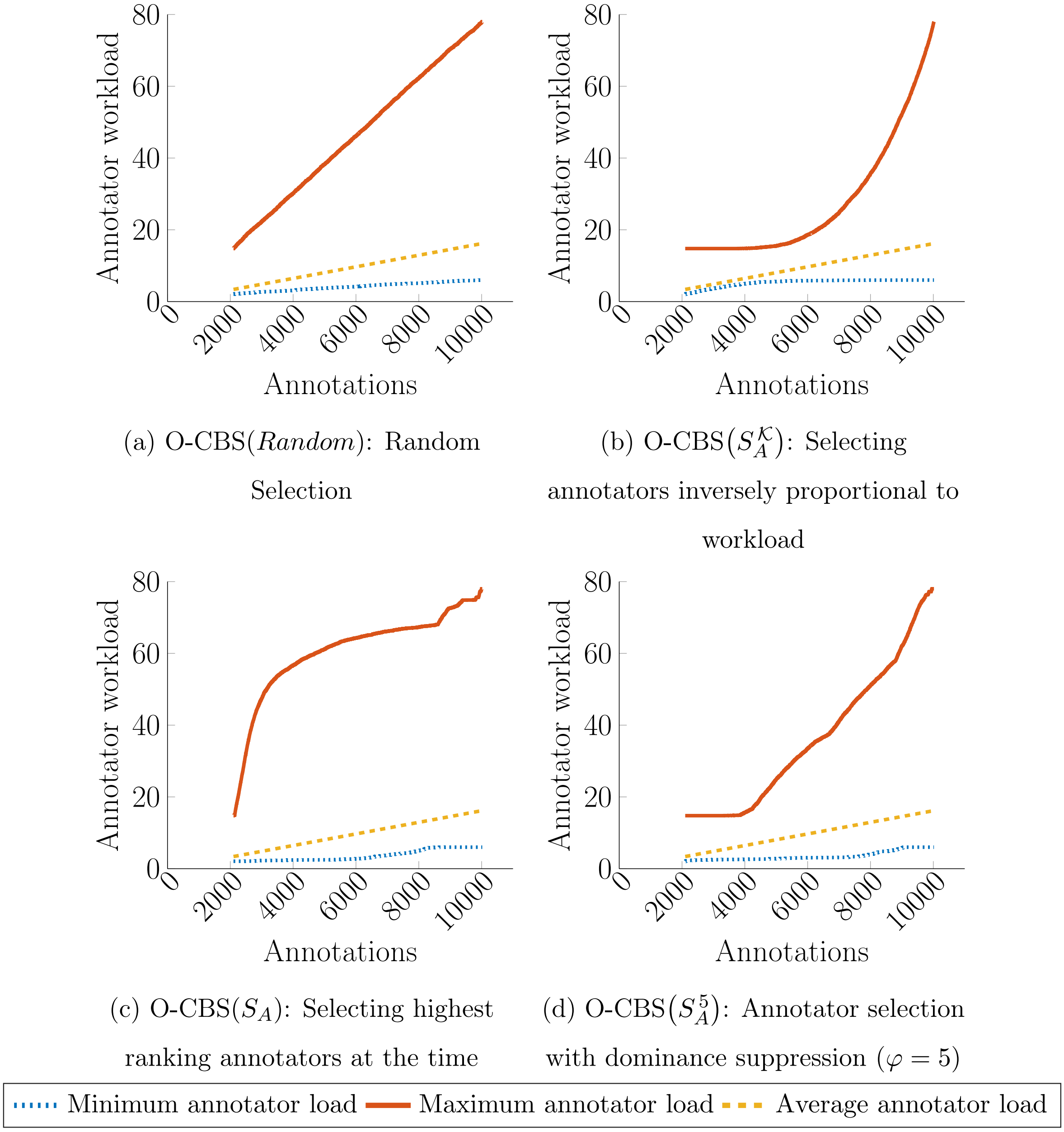
$\frac{|\mathcal{K}|}{|\mathcal{J}^1|}$ is the average number of annotations per annotator. With each new annotation, this factor increases; with each new annotator, it decreases momentarily. New annotator introduction to the system is rarer than adding new annotations to the annotation pool from current annotators. Thus, the suppression effect of the newly introduced dominance suppression factor almost always decreases as the active crowd-labeling process progresses.

Thus, we introduce a dominance suppression based annotator competence score as the product of the annotator competence score and the dominance suppression factor (Equation 6.1):

$$S_A^{\varphi}(j) = S_A(j)\left|\mathcal{K}^j\right|^{-\varphi\frac{\left|\mathcal{J}^1\right|}{|\mathcal{K}|}} \tag{6.2}$$

As a baseline method, we also introduce a simple annotator score based only on the annotator's workload:

$$S_A^{\mathcal{K}}(j) = |\mathcal{K}^j|^{-1} \tag{6.3}$$



(a) O-CBS(*Random*): Random Selection

(b) O-CBS($S_A^{\mathcal{K}}$): Selecting annotators inversely proportional to workload

(c) O-CBS(S_A): Selecting highest ranking annotators at the time

(d) O-CBS(S_A^5): Annotator selection with dominance suppression ($\varphi = 5$)

Figure 6.4. Change in the minimum, maximum, and average annotator workloads during the active crowd-labeling process. The results are provided for the Age Annotations dataset.

Figure 6.4 shows the load of minimum, maximum and averagely loaded annotators. The horizontal axis represents the total number of annotations currently in the system. The vertical axis represents the number of annotations (workload) of the annotator in question. Note that each point on the plots may represent a different annotator. Depending on the annotator selection criterion, the maximally and minimally loaded annotators will change during the annotation process. In Figure 6.4a, where new annotations are randomly selected, maximum annotator load increases linearly and diverges quickly from the average load. This means that only a handful of annotators are dominating the system. This is a tendency that we aim to avoid as mentioned before.

If $S_A^{\mathcal{K}}(j)$ is used as the annotator score, we see that the maximum annotator load tends to stay the same for a long time (Figure 6.4b). Although this behavior is desired since it prevents domination by a group of annotators, this scoring mechanism by its very nature does not incorporate the behavior of the annotator and fails to pinpoint competent annotators.

When the scoring function $S_A(j)$ (Section 5.2) is used, the active crowd-labeling system tends to overload the high scoring annotators and the maximum load increases rapidly (Figure 6.4c). However, this is risky due to the problems described earlier.

When dominance suppression is active, the scores of highly loaded annotators are weighted down for obtaining the desired behavior. In Figure 6.4d, we choose the dominance suppression coefficient $\varphi = 5$ and it is clear that we reach a more stable annotator load distribution. Early on in the active crowd-labeling process, the maximum annotator load holds steady while the system gets acquainted with the annotators in an objective manner. After a while the maximum workload starts to increase with the diminishing effect of the dominance suppression factor, thereby utilizing high quality annotators.

6.3. Effects of Annotator Dominance Suppression

In this section, we will discuss the results of improving the existing consensus by using active crowd-labeling under several different dominance suppression criteria. However, the data described in Chapter 2 was not collected considering active crowd-labeling. Thus, first we need to create starting subsets of the annotation data for evaluating O-CBS. We present our results on nine datasets, namely the Age Annotations, the Head Pose Annotations, and the six Affective Text Analysis datasets.

6.3.1. Selecting Starting Subset for Active Crowd-Labeling:

Assume that annotations are already collected for a fixed sample set and we want to improve the consensus values without adding new annotators to the system. This is a common case in many institutions where a dataset is collected and annotated in-house. In this setting, the problem of extending the annotation dataset boils down to asking an annotator to annotate a sample that they have not annotated before. In order to emulate this, we create annotation subsets for each dataset that satisfy the following conditions:

- The resulting subset should have ν annotations,
- Minimum sample count of the resulting subset should be ρ,
- Minimum annotator count of the resulting subset should be η,
- Every annotator in the resulting subset should have at least ζ annotations,
- Every sample in the resulting subset should have at least δ annotations,
- Annotations of an annotator should not be disconnected from the rest of the data. Every annotator must have an annotation for a sample that also has an annotation from another annotator.

Figure 6.5 gives the details of the starting subset creation process.

Specific to our problem, we employ the algorithm in Figure 6.5 such that the following conditions are satisfied:

- Every sample has an annotation ($\delta = 1$)
- Every annotator has at least $\zeta = 2$ annotations
- Every annotator has an annotation for a sample that also has an annotation from another annotator (this is needed for being able to compare annotators)

```
Input:
    Sets of samples $\mathcal{I}$, annotators $\mathcal{J}$, annotations $\mathcal{K}$
    Target annotation count $\nu$, minimum annotations per annotator $\zeta$, minimum annotations per sample $\delta$, minimum sample count $\rho$, minimum annotator count $\eta$
Output:
    Subset of annotations $\mathcal{K}$

function CreateSubset($\mathcal{I}, \mathcal{J}, \mathcal{K}, \nu, \zeta, \delta, \rho, \eta$)
    Shuffle($\mathcal{K}$)
    for all $k \in \mathcal{K}$ do
        for all $j \in \mathcal{J}$ do
            $\mathcal{K}^j \leftarrow \{k \in \mathcal{K} : j_k = j\}$    ▷ Annotations of the annotator $j$
        end for
        for all $i \in \mathcal{I}$ do
            $\mathcal{K}_i \leftarrow \{k \in \mathcal{K} : i_k = i\}$    ▷ Annotations of the sample $i$
        end for
        if $|\mathcal{K}^{j_k}| < \zeta$ then    ▷ If the annotator $j_k$ of the annotation $k$ has less than $\zeta$ annotations
            $\mathcal{D} \leftarrow \mathcal{K}^{j_k}$    ▷ Mark all annotations of $j_k$ to be removed
        else
            $\mathcal{D} \leftarrow \{k\}$    ▷ Mark only the annotation $k$ to be removed
        end if
        $\mathcal{T}_s \leftarrow \{i \in \mathcal{I} : |\mathcal{K}_i \setminus \mathcal{D}| > 0\}$    ▷ Samples with at least 1 annotation
        $\mathcal{T}_a \leftarrow \{j \in \mathcal{J} : |\mathcal{K}^j \setminus \mathcal{D}| > 0\}$    ▷ Annotators with at least 1 annotation
        if $\exists i \in \{i_k : k \in \mathcal{D}\}$ s.t. $|\mathcal{K}_i \setminus \mathcal{D}| < \delta$ then    ▷ If any sample has less than $\delta$ annotations
            continue    ▷ Reject
        else if $|\mathcal{T}_s| < \eta$ or $|\mathcal{T}_a| < \rho$ then    ▷ If number of samples or annotators are below limits
            continue    ▷ Reject
        else if $\exists j$ s.t. $|\mathcal{K}_i \setminus \mathcal{D}| = 1, \forall i \in \mathcal{I}_j$ then    ▷ If an annotator does not have a common sample annotated with another annotator
            continue    ▷ Reject
        else    ▷ Accept the removal of the annotation(s) in $\mathcal{D}$
            $\mathcal{K} \leftarrow \mathcal{K} \setminus \mathcal{D}$    ▷ Update $\mathcal{K}$ by removing $\mathcal{D}$
        end if
        if $|\mathcal{K}| < \nu$ then    ▷ Break if target annotation count is reached
            break
        end if
    end for
    return $\mathcal{K}$
end function
```

Figure 6.5. Create Starting Set By Elimination

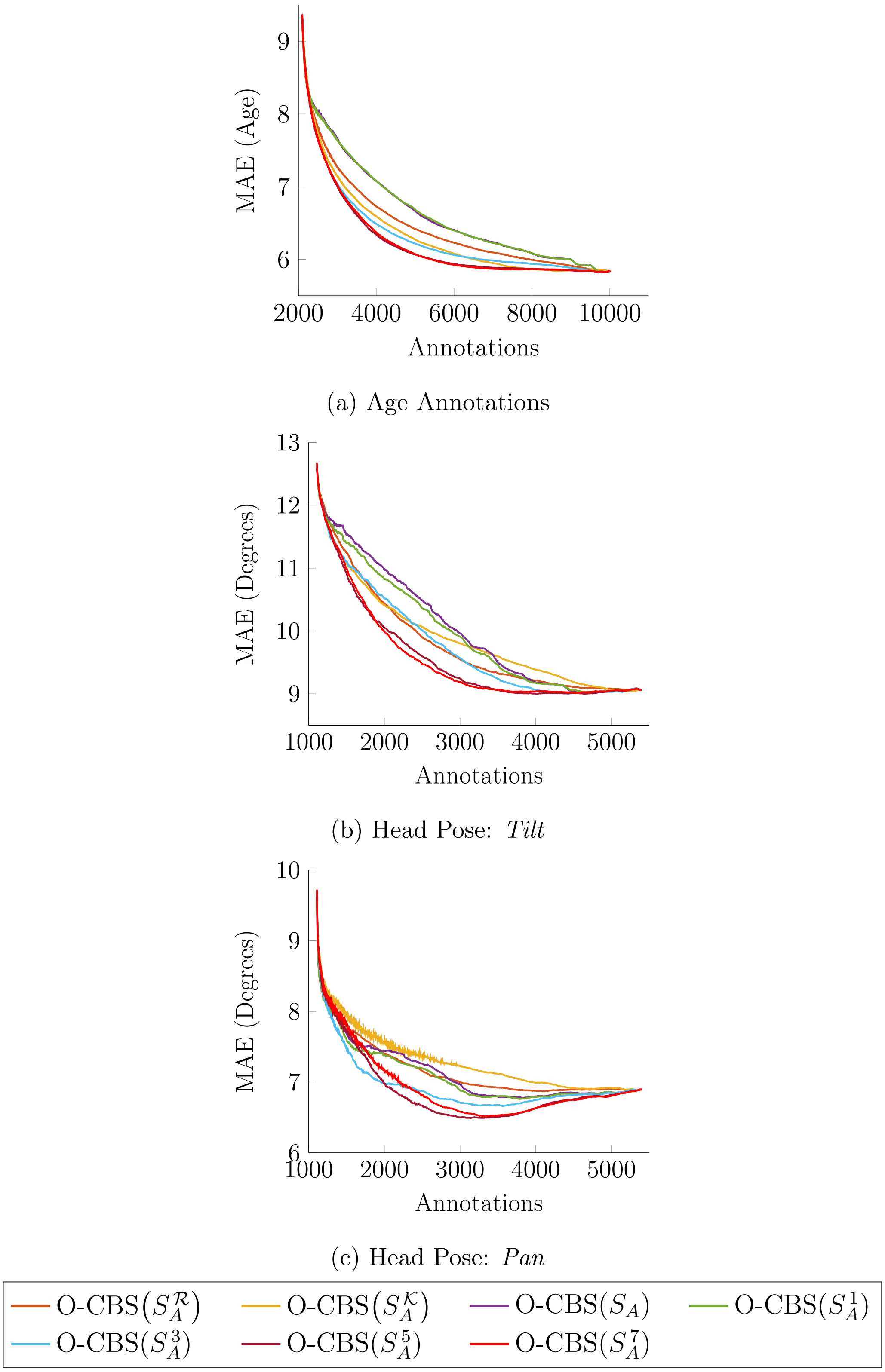
For each dataset, we prepare 100 different subsets satisfying these conditions. We fix the subset sizes, *i.e.* number of annotations, to 2100 for the Age Annotations dataset, 1110 for the Head Pose Annotations datasets, and 200 for the Affective Text Analysis datasets. In Table 6.1, we give pairwise inter-set similarity statistics of the created subsets. We observe that there is approximately 20% overlap between the resulting subsets on average. This similarity is low enough to ensure that the results of our active crowd-labeling scheme do not depend on initial conditions.

Table 6.1. Details of the created subsets

Dataset	Subset size	Inter-Set Similarity (%)		
		Min	Average	Max
Head Pose Annotations	1110	15.68	21.47 ± 1.17	26.13
Age Annotations	2100	18.43	21.13 ± 0.85	24.81
Affective Text Analysis	200	11	19.95 ± 2.63	29.5

6.3.2. Mean Absolute Age Error Improvement on the Age Annotations Dataset:

In Figure 6.6a, we present the results of our method's effect on mean absolute error in terms of age, by trying out different dominance suppression coefficients φ on the Age Annotations dataset. We have two baseline methods that we compare our approach with. The first is O-CBS$\big(S_A^{\mathcal{R}}\big)$ where the annotator is selected randomly. The second is where the sample with the worst consensus quality score is annotated by the annotator with the least annotation count (O-CBS$\big(S_A^{\mathcal{K}}\big)$). We do not plot O-CBS(*Random*) curves in Figure 6.6, since we already gave their comparison with O-CBS$\big(S_A^{\mathcal{R}}\big)$ in Figure 6.2. When φ is small, our method fails to suppress low-quality annotators as we describe in Section 6.2, resulting in even lower performance than the baseline methods. When $\varphi \geq 3$ our method outperforms the baseline approaches significantly. Instead of collecting 10000 annotations, roughly 6000 annotations are sufficient to drop below 6 years in terms of mean absolute error.

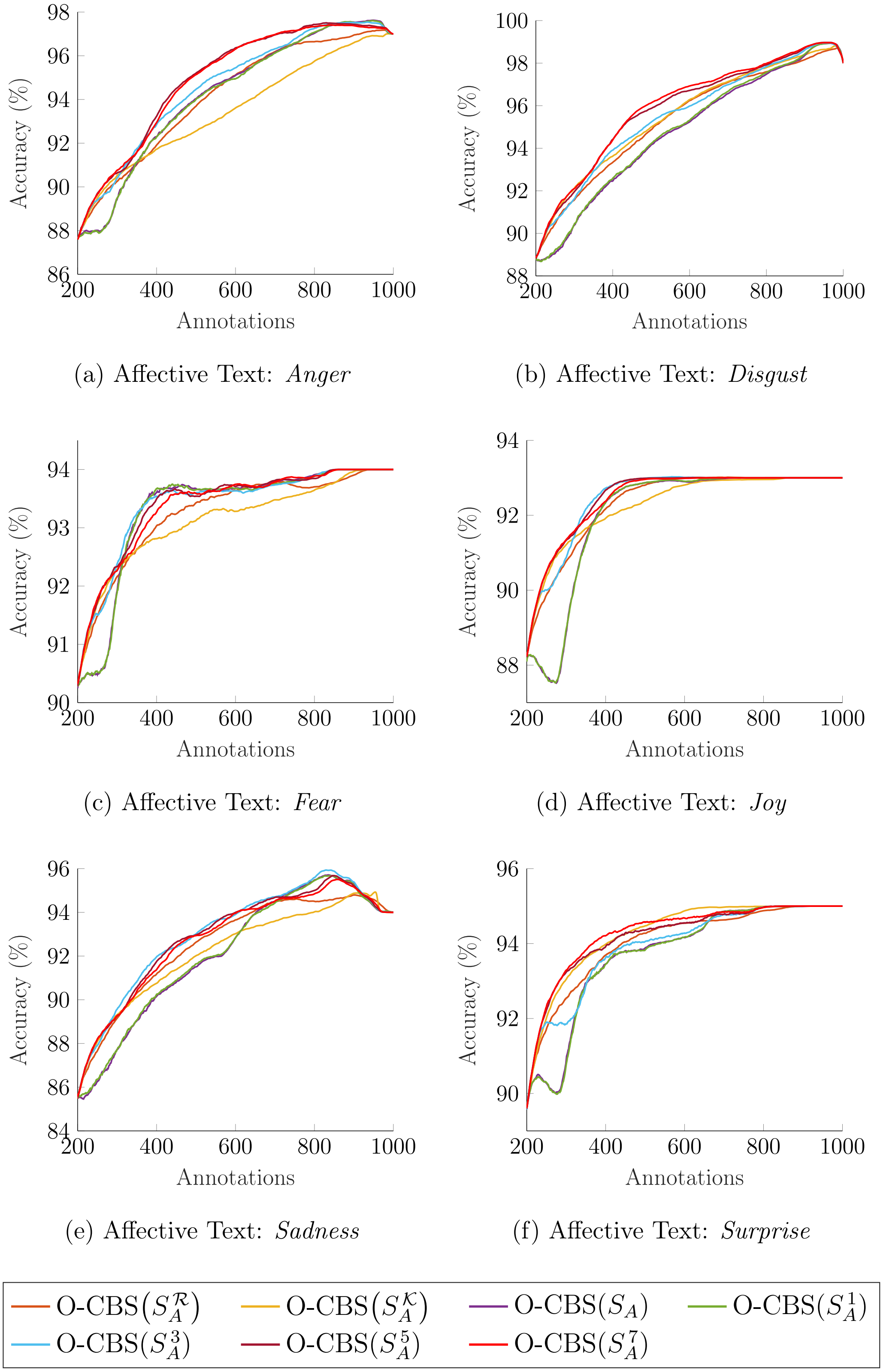


(a) Age Annotations

(b) Head Pose: *Tilt*

(c) Head Pose: *Pan*

Figure 6.6. Improving the existing consensus on the Age Annotations and the Head Pose Annotations datasets. The curves are averaged over 100 runs with different starting subsets.



(a) Affective Text: *Anger*

(b) Affective Text: *Disgust*

(c) Affective Text: *Fear*

(d) Affective Text: *Joy*

(e) Affective Text: *Sadness*

(f) Affective Text: *Surprise*

Figure 6.7. Improving the existing consensus on the Affective Text Analysis datasets. The curves are averaged over 100 runs with different starting subsets.

6.3.3. Mean Absolute Degree Error Improvement on the Head Pose Annotations Dataset:

We further test the performance of O-CBS on the Head Pose Annotations *tilt* and *pan* datasets. Figures 6.6b and 6.6c show the change in the mean absolute error in degrees, according to different dominance suppression coefficients. Similar to the performance on the Age Annotations dataset, O-CBS performs subpar when the dominance suppression coefficient φ is small, or the non-suppressed annotator scoring mechanism $S_A(j)$ is used. On the *tilt* dataset, the MAE achieved at the end of the annotation procedure can be achieved earlier on with much fewer annotations by using $\varphi \geq 5$. For the *pan* dataset, we also observe that the curves with $\varphi \geq 5$ have a trough shape around 3000 annotations. This trend is due to the fact that high-quality annotators are distinguished early on, resulting in low error. Additional annotations provided by lower quality annotators result in degrading the system performance. Note that we let the system to use all annotations for examining the total effect of the annotations on consensus quality. Every point on these graphs actually show the performance at the corresponding annotation limit. Therefore, it is also possible to interpret Figure 6.6 as what the performance of the system will be, should a budget limit be enforced.

6.3.4. Accuracy Improvement on the Affective Text Analysis Datasets:

We also test our method on the six Affective Text Analysis datasets, which present a more challenging problem since the datasets are much smaller than both the Age Annotations and Head Pose Annotations datasets. We do not plot O-CBS($Random$) curves in Figure 6.7, since we already gave their comparison with O-CBS$\big(S_A^{\mathcal{R}}\big)$ in Figure 6.3. Our first observation in Figures 6.7a to 6.7f is that each dataset belonging to an emotion results in different baseline method characteristics, presenting diverse conditions in which we test our method.

In consort with the results in Figures 6.6a to 6.6c, a higher dominance suppression coefficient $\varphi \geq 5$ helps to achieve high accuracy with fewer annotations. This effect is most prominent in *fear*, *joy*, and *surprise* datasets where roughly 400 out of

1000 annotations are sufficient for achieving near-maximum accuracy. Additionally, introducing the dominance suppression factor helps us to outperform the two baseline methods significantly, specifically in the *anger*, *fear*, and *sadness* datasets.

6.4. Speeding Up the Inference Process

In passive crowd consensus estimation, we randomly initialize the annotator parameters and iteratively infer the resulting annotator parameters using M-CBS. In an active crowd-labeling process, this inference process is repeated with each new annotation and the computational cost increases duly. However, we expect a small change in annotator parameters since there is only a small change in the annotations set. Thus, we can use our previous knowledge about the annotator parameters to reduce the complexity of the process.

M-CBS describes an annotator using a linear map and a noise parameter. When there are only a few annotations of an annotator, the model might infer a wrong conclusion about the behavior of the annotator in question. This is a very common case especially in the early phases of the active crowd-labeling scheme.

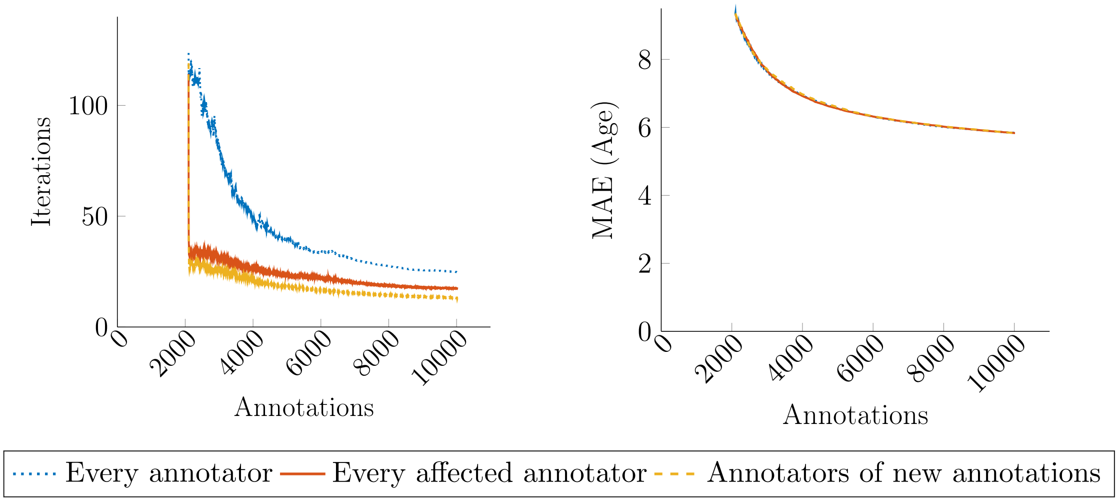


Figure 6.8. The effect of three different random initialization approaches on the number of iterations for O-CBS(*Random*) (random annotation addition) on the Age Annotations dataset. Reinitializing the annotator parameters of only those providing new annotations results in much fewer iterations with the same MAE.

In Figure 6.8, we present three random initialization approaches and their effect on iteration count and MAE. The first approach is to initialize every annotator's parameters each time a new annotation is acquired, thus avoiding sticking to a local extremum. This is actually a baseline approach which results in high iteration counts, especially early on in the active crowd-labeling process. Alternatively, we may initialize the parameters of every annotator that has provided an annotation for the newly annotated sample, since the new annotation will affect the sample's consensus. It is also possible to take a more conservative approach and reinitialize the parameters of only the new annotation's annotator. Both of these approaches still have the advantage of avoiding being stuck at local extrema. Results show that both of these approaches result in a significantly decreased number of iterations, with the latter approach being lower in iteration numbers. There is no change in the MAE, which confirms that these time-saving methods do not affect the quality of the consensus estimation process.

7. O-CBS+: STARTING ACTIVE CROWD-LABELING FROM SCRATCH

When the task giver has full control over the label collection process, it is more beneficial to identify the annotator quality as soon as possible. Timely evaluation of annotator quality results in saving both money and time by achieving high quality consensuses using fewer annotations. Thus, it is important to use the active crowd-labeling process from scratch.

O-CBS handles the case when we are already acquainted with the annotators, thus have an opinion about their annotation behaviors. However, for using active crowd-labeling at the start of the crowd-labeling process, we need to not only utilize current annotators, but also assess new annotators.

Even though the sample pool is fixed at the end, every sample seems to be new at the early stages of active crowd-labeling since we do not have annotations for them. O-CBS is not designed for the addition of new samples. When a new sample needs to be annotated, it is crucial to have an opinion about its consensus in a timely fashion.

In Figure 7.1, we take these concerns into account. We first check whether there is a new sample or not. If there are new samples that have not been annotated before, we randomly select a sample to be annotated. Otherwise, we select the sample with the worst consensus quality score, similar to O-CBS. Upon the selection of the sample, we need to decide if we want to have this sample annotated by a known annotator (exploit) or a new annotator (explore). If we decide to exploit an annotator, we request an annotation for the selected sample from the highest scoring available annotator. When exploring a new annotator, we want to have at least two annotations of the annotator since we want to have an opinion about their behavior and one of the annotations should be of an already annotated sample. Thus, we request two annotations from the new annotator accordingly.

```
Input:
    Sets of all samples $\mathcal{I}$, all annotators $\mathcal{J}$, current annotations $\mathcal{K}$, currently active annotators $\mathcal{J}'$
    $i_k$ and $j_k$ are the sample and the annotator of the annotation $k$, respectively
    $S_S(i)$ is the consensus quality score of sample $i$, $S_A(j)$ is the competence score of annotator $j$ (We
    assume that $S_S$ and $S_A$ are intrinsically aware of the annotator parameters $a$, $w$, $b$, and $\lambda$)
    $\mathcal{E}$ defines the probability of exploring a new annotator
Output: New annotation(s) $\{k\}$ or $\{k, k'\}$

function RequestAnnotationExp($\mathcal{I}, \mathcal{J}, \mathcal{J}', \mathcal{K}, S_S(\cdot), S_A(\cdot), \mathcal{E}$)
    for all $i \in \mathcal{I}$ do
        $\mathcal{K}_i \leftarrow \{k \in \mathcal{K} : i_k = i\}$    ▷ Annotations of sample $i$
        $\mathcal{J}_i \leftarrow \{j_k \in \mathcal{J} : k \in \mathcal{K}_i\}$    ▷ Annotators of sample $i$
    end for
    for all $j \in \mathcal{J}$ do
        $\mathcal{K}^j \leftarrow \{k \in \mathcal{K} : j_k = j\}$    ▷ Annotations of annotator $j$
    end for
    $\mathcal{U}_s \leftarrow \{i \in \mathcal{I} : |\mathcal{K}_i| = 0\}$    ▷ Samples without any annotation
    $\mathcal{U}_a \leftarrow \{j \in \mathcal{J} : |\mathcal{K}^j| = 0\}$    ▷ Annotators without any annotation
    if $|\mathcal{U}_s| > 0$ then    ▷ If there is a sample without any annotation
        $i \leftarrow$ Randomly select from $\mathcal{U}_s$
    else
        $i \leftarrow \operatorname*{argmin}_{i' \in \mathcal{I} \text{ s.t. } \mathcal{J}' \setminus \mathcal{J}_{i'} \neq \emptyset} S_S(i')$    ▷ Select the sample with the worst consensus quality such that at least one of the currently active annotators has no annotations for that sample
    end if
    $\mathcal{R} \leftarrow \mathcal{U}_a \cap \mathcal{J}'$    ▷ Set of explorable annotators
    $\mathcal{T} \leftarrow \mathcal{J}' \setminus (\mathcal{J}_i \cup \mathcal{U}_a)$    ▷ Set of exploitable annotators
    if $|\mathcal{R}| > 0$ and $|\mathcal{T}| > 0$ then    ▷ If there are both explorable and exploitable annotators
        explore $\leftarrow$ true with probability $\mathcal{E}$    ▷ Randomly decide whether to explore a new annotator or exploit an existing annotator
    else if $|\mathcal{R}| > 0$ then    ▷ If there are only explorable annotators
        explore $\leftarrow$ true
    else if $|\mathcal{T}| > 0$ then    ▷ If there are only exploitable annotators
        explore $\leftarrow$ false
    end if
    if explore then
        $j \leftarrow$ Randomly select from $\mathcal{R}$    ▷ Select an annotator from explorable annotators
        $i' \leftarrow$ Randomly select from $\mathcal{I} \setminus \mathcal{U}_s$    ▷ Select a sample from previously annotated samples
        $k' \leftarrow$ Request an annotation for a random sample $i'$ from annotator $j$
    else
        $j \leftarrow \operatorname*{argmax}_{j' \in \mathcal{J}' \setminus \mathcal{J}_i} S_A(j')$    ▷ Select the most competent annotator from the set of active annotators who had not annotated sample $i$
    end if
    $k \leftarrow$ Request an annotation for the sample $i$ from annotator $j$
    if explore then
        return $\{k, k'\}$
    else
        return $\{k\}$
    end if
end function
```

Figure 7.1. RequestAnnotationExp: Requesting annotation for smart label collection from scratch

O-CBS+ is based on Figure 5.1 with M-CBS as the EstimateLabels(·) function and RequestAnnotationExp(·) of Figure 7.1 as the RequestAnnotation(·) function. In this setting, RequestAnnotationExp(·) employs S_S as the sample consensus quality scoring function, same as O-CBS. Since in Section 6.3 we observe that O-CBS(S_A^5) performs to our satisfaction, we fix the dominance suppression coefficient as $\varphi = 5$ and use S_A^5 as the annotator competence scoring function for O-CBS+. We denote O-CBS+ with different exploration parameters ($\mathcal{E}$) as O-CBS+($\mathcal{E}$). As a baseline method, we use O-CBS+(*Random*) which is similar to O-CBS(*Random*). In O-CBS+(*Random*), the annotators are selected randomly regardless of whether they are already known or new. Note that if there are samples without any annotation, the random selection is performed among them. As soon as all samples have annotations, full random selection commences.

In the remainder of this section, we thoroughly study the performance of O-CBS+. First, we investigate the effect of the exploration parameter $\mathcal{E}$ for all datasets and discuss the risks and benefits of incorporating new annotators into the system.. Then, we compare the performance of O-CBS+ with two methods [37,53] from the literature. Note that the work of Welinder and Perona [37] provides the only directly comparable method to O-CBS+ as we have previously mentioned in Section 1.1. Raykar and Agrawal [53] provide comparative results with the binary method of Welinder and Perona [37] on the six Affective Text Analysis datasets using active crowd-labeling with binarized inputs. Although the method of Raykar and Agrawal [53] is not directly comparable to our work, for the sake of completeness we also provide comparative results by binarizing our continuous-valued consensuses. Finally, we investigate the effect of enforcing a sample score related stopping criterion and provide further comparative results with Welinder and Perona [37] and Raykar and Agrawal [53].

7.1. Effect of Annotator Exploration

In this section, we will discuss the results of starting active crowd-labeling from scratch under several different exploration parameters. We present our results on nine datasets, namely the Age Annotations, the Head Pose Annotations, and the six Affective Text Analysis datasets.

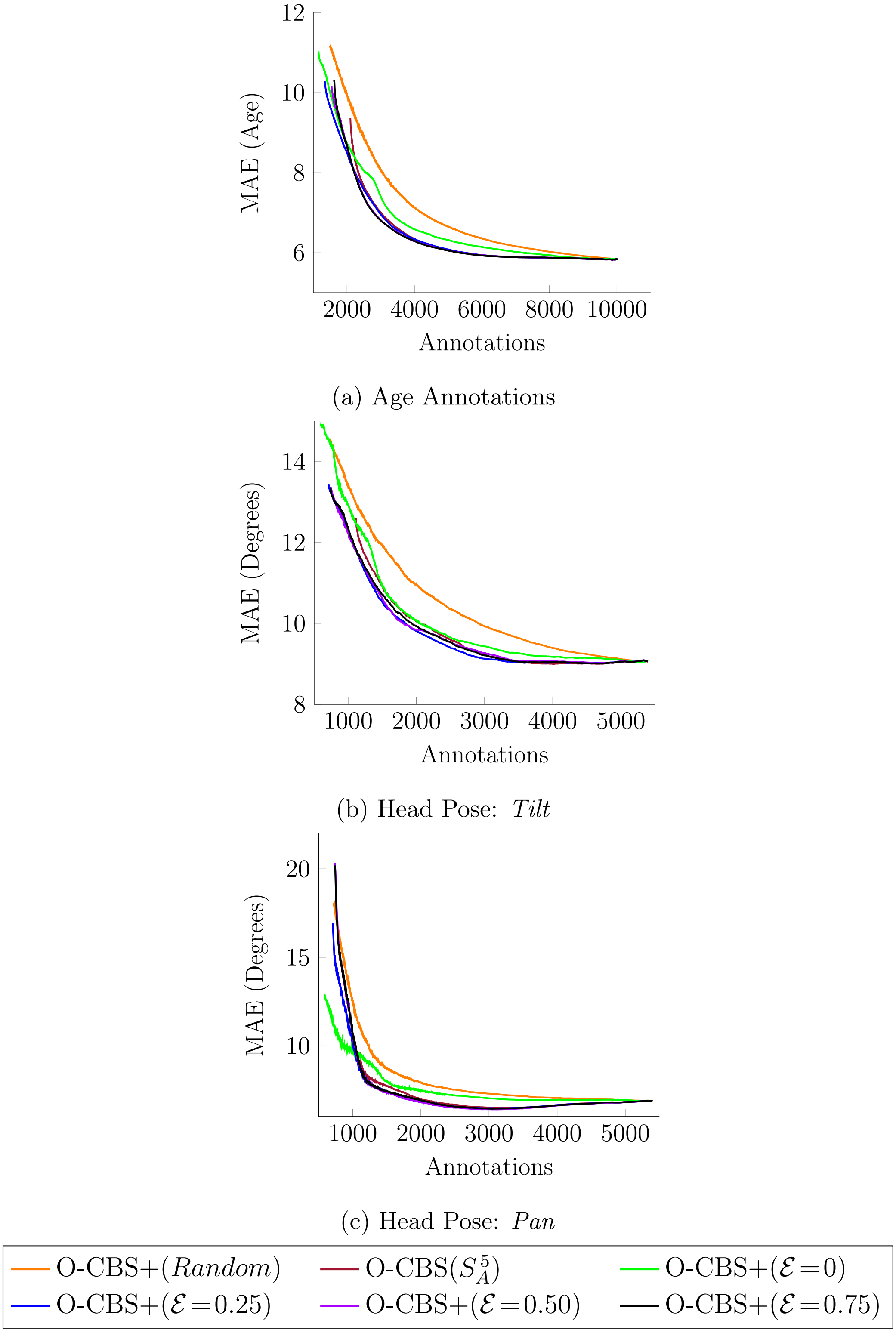
7.1.1. Mean Absolute Age Error Improvement on the Age Annotations Dataset

In Figure 7.2a, we present the effect of changing the exploration parameter $\mathcal{E}$ on the Age Annotations dataset. Figure 7.2a shows the reduction in the mean absolute error in terms of age, while the active crowd-labeling is started from scratch. For the analysis to be meaningful, we start reporting the error once each sample has a consensus estimation. Therefore, the curves do not start from zero annotations. Additionally, due to the fact that the active crowd-labeling process has a random nature, the moment where every sample has a consensus is different for each trial. Thus, the starting point of the curves also differ from one another in the figures.

In Figure 7.2a, we compare O-CBS+ with fixed dominance suppression coefficient of $\varphi = 5$ for different $\mathcal{E}$ values. We also compare with O-CBS(S_A^5) from Figure 6.6a and the random annotation selection mentioned in Section 6.3, as baseline comparisons. It is evident that active learning from scratch with exploration performs better than the random selection method. We also observe that starting from scratch ensures the same success with fewer annotations.

An important point worth mentioning is that using O-CBS+$(\mathcal{E}=0)$ is not the same as using O-CBS with an empty set of initial annotations. Although $\mathcal{E} = 0$ seems like no exploration takes place in the process, inevitably exploration is done when there is no annotator to exploit. This case may also happen for any $\mathcal{E} < 1$. Similarly for $\mathcal{E} > 0$, when the system runs out of annotators to explore, it goes on full-exploitation mode until a new annotator joins the system.

When we observe Figure 7.2a, we see that the results get better and the gain eventually diminishes with higher exploration coefficient $\mathcal{E}$. Note that the annotator set is limited in the dataset, and thus the systems with large $\mathcal{E}$ values learn all an-

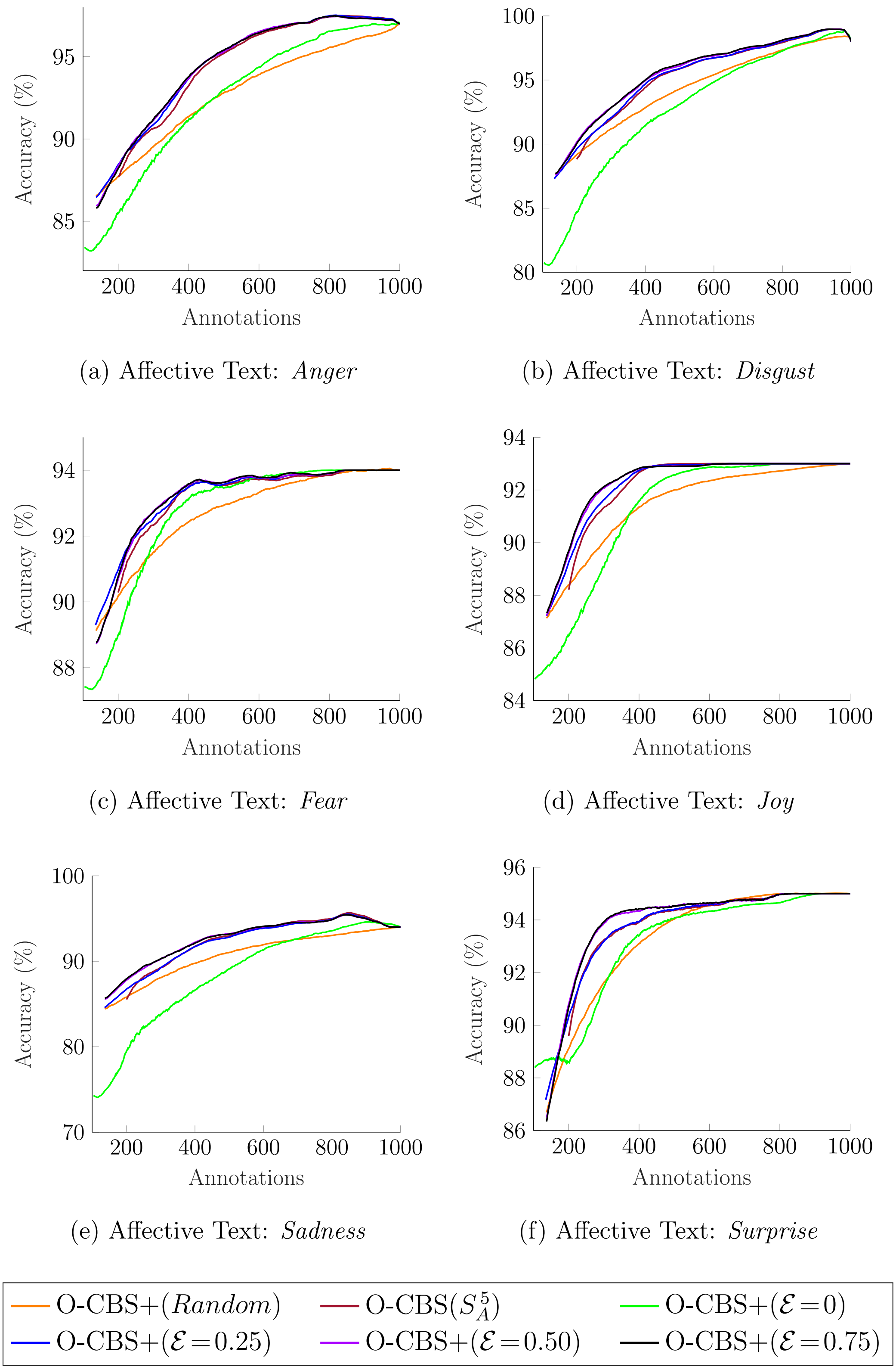


(a) Age Annotations

(b) Head Pose: *Tilt*

(c) Head Pose: *Pan*

Figure 7.2. Effect of changing the exploration parameter $\mathcal{E}$ on the Age Annotations and the Head Pose Annotations datasets. The results are presented for $\varphi = 5$ and are the averages of 100 repetitions.



(a) Affective Text: *Anger*

(b) Affective Text: *Disgust*

(c) Affective Text: *Fear*

(d) Affective Text: *Joy*

(e) Affective Text: *Sadness*

(f) Affective Text: *Surprise*

Figure 7.3. Effect of changing the exploration parameter $\mathcal{E}$ on the Affective Text Analysis datasets. The results are presented for $\varphi = 5$ and are the averages of 100 repetitions.

notators rapidly. When there are no new annotators to explore, the system begins to exploit high quality annotators early on. Therefore, better results are achieved faster. We have to keep in mind that the essence is in exploitation of high quality annotators, and this is achieved by exploration. Since the Age Annotations dataset is a fairly large dataset, the difference between choosing different exploration coefficients quickly becomes indistinguishable after all annotators are explored. However, exploration should be used moderately on open ended annotation problems (*i.e.* where the annotator pool is considered to be unlimited).

7.1.2. Mean Absolute Degree Error Improvement on the Head Pose Annotations Datasets

Figures 7.2b and 7.2c show the effect of the exploration parameter $\mathcal{E}$ on the Head Pose Annotations *tilt* and *pan* datasets. Similar to the Age Annotations dataset, we compare the O-CBS+ results with the two baseline methods O-CBS(S_A^5) and O-CBS+(*Random*). On both datasets, increasing the exploration coefficient $\mathcal{E}$ results in marginal decrease in terms of mean absolute degree error. The results in Figures 7.2a to 7.2c suggest that the effect of $\mathcal{E}$ is difficult to observe on large datasets and call for a closer inspection on smaller datasets. The advantage of annotator selection over random selection is more apparent in the *tilt* dataset.

7.1.3. Accuracy Improvement on the Affective Text Analysis Datasets

In Figures 7.3a to 7.3f, we present the effect of the exploration parameter $\mathcal{E}$ on the Affective Text Analysis datasets, which are significantly smaller datasets compared to the other three datasets. Overall, the results are in concord with those of the Age Annotations dataset (Figure 7.2a) and the Head Pose Annotations dataset (Figures 7.2b and 7.2c). In addition, the advantage of using a higher exploration parameter such as $\mathcal{E} = 0.75$ results in higher accuracies.

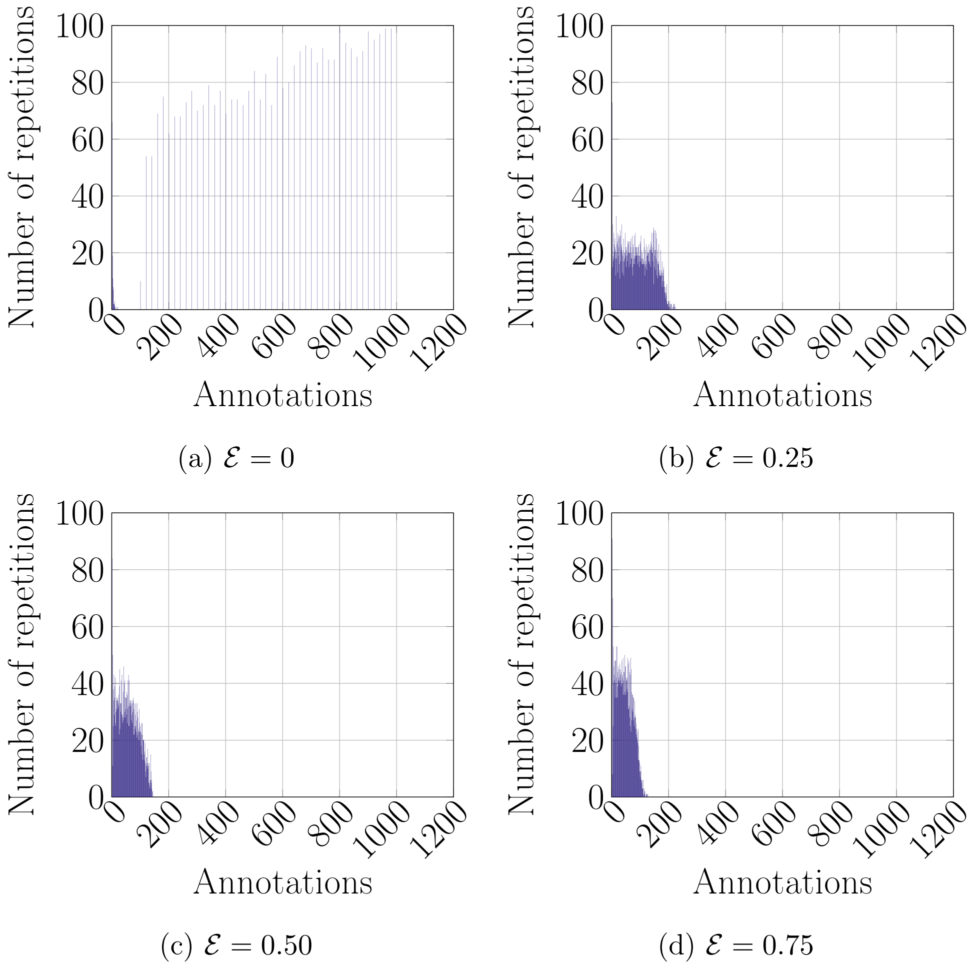
Since the annotation set is limited, all curves converge to the same point toward the end of the active crowd-labeling process. Therefore, well-performing methods which reach a higher accuracy with fewer annotations converge to the same point with the weaker methods at the end. An example for this can be observed in Figure 7.3f, where the exploration-based methods outperform O-CBS(S_A^5) but end up with the same accuracy at the end.

A striking difference from the Age Annotations dataset is the performance of the $\mathcal{E} = 0$ curve. In the six Affective Text Analysis datasets, it significantly falls behind its counterparts. The strict imposition of annotator exploitation results in the late integration of high-quality annotators to the system. Since the Affective Text Analysis datasets are much smaller than the Age Annotations dataset, timely exploration of high-quality annotators is much more critical for the success of the active learning process and the tardiness caused by selecting $\mathcal{E} = 0$ becomes evident in the graphs.

On specifically three datasets, namely *fear*, *joy*, and *surprise*, our method quickly reaches high accuracies with a small number of annotations. This is due to the fact that our method succeeds in selecting high-quality annotators faster. Another remark is about the peaks observed in the *anger*, *disgust*, and *sadness* datasets. These peaks indicate that the system has to exploit low-quality annotators when it runs out of annotations from the high-quality ones. The reason is that we are working with a limited annotation set and we force the system to use every annotation for observing the complete behavior. Therefore, the active learning performance degrades in these three datasets with an increasing number of annotations toward the end.

7.2. Is It Wise to Take Risks by Incorporating New Annotators?

Although it is apparent that a system without exploration would suffer when the starting annotation set is small, the intuitive expectation is that a conservative approach to exploration would be better. This is due to the fact that there is a risk associated with new annotators and we can always select the better annotators among the annotators we know. However, the results in Section 7.1 show otherwise.



(a) $\mathcal{E} = 0$

(b) $\mathcal{E} = 0.25$

(c) $\mathcal{E} = 0.50$

(d) $\mathcal{E} = 0.75$

Figure 7.4. New annotator exploration times on the Affective Text Analysis - *Anger* dataset for O-CBS+(S_A^5)

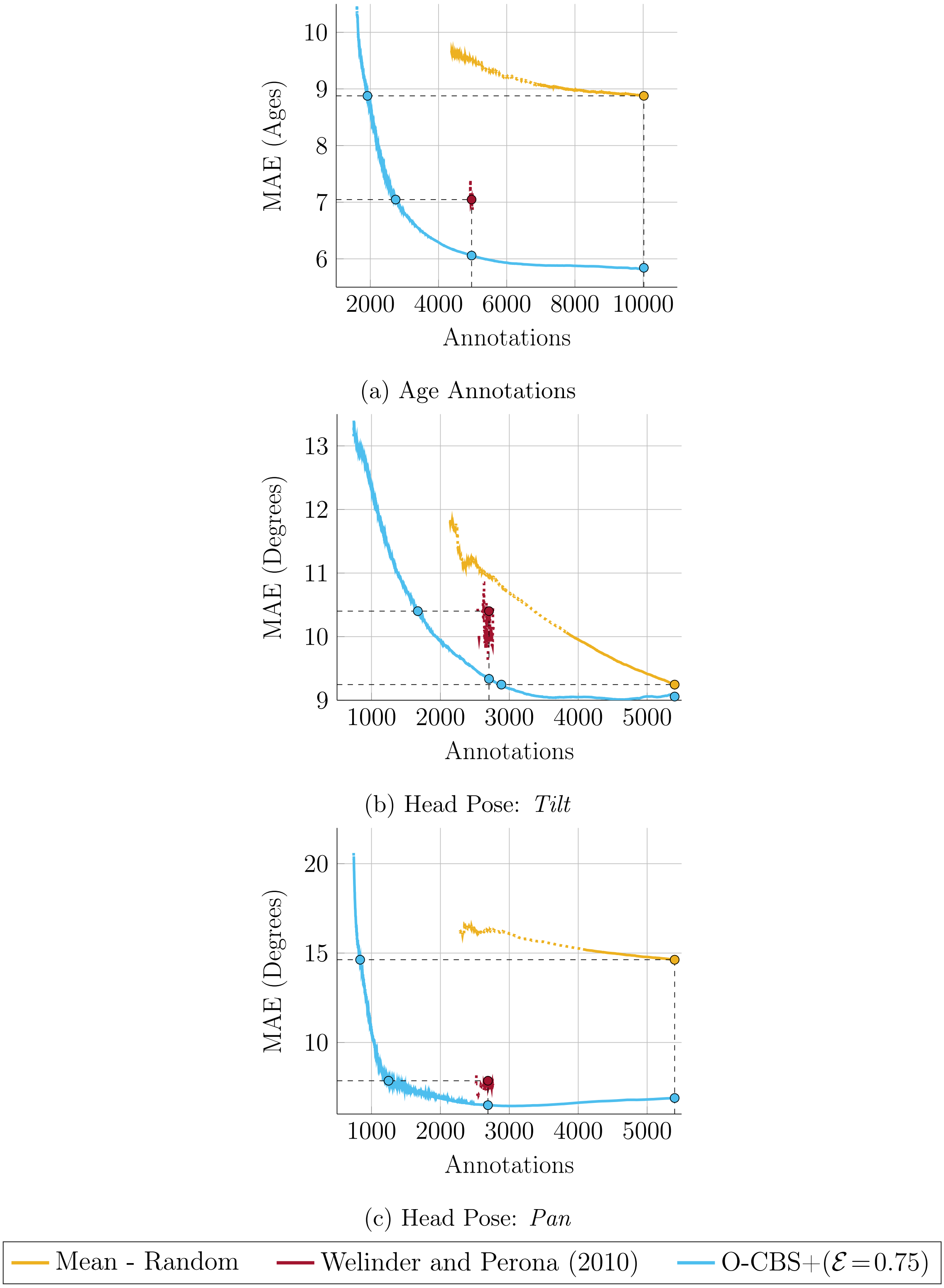
When we observe the exploration times shown in Figure 7.4, we see that the system exhausts new annotators quickly since our datasets contain finite number of annotators. When working with a limited annotator set, it is wise to assess all annotators quickly so that the active crowd-labeling approach starts to utilize better annotators early on. The results presented in Figure 7.4 and Section 7.1 validate this observation. A larger $\mathcal{E}$ results in the addition and assessment of new annotators to the system very quickly and therefore better results are achieved with fewer annotations by utilizing good annotators.

Note that these results are obtained from readily available datasets with a limited number of annotators. In a live and open-ended active crowd-labeling process, it would be wise to concentrate more on exploiting the existing good annotators and choose a smaller $\mathcal{E}$ value, instead of constantly exploring new annotators.

7.3. Comparative Performance of O-CBS+ Under Annotation Count Limitations

So far, we have deduced that S_A^5 is a good annotator competence scoring function choice and fixed it in O-CBS+. Figures 7.2 and 7.3 shows that fast exploration of annotators is preferable, especially for small datasets. Thus, we present the results using O-CBS+($\mathcal{E}=0.75$) for the comparative performance evaluation of O-CBS+ with the existing methods, namely Welinder and Perona [37] and Raykar and Agrawal [53]. The experiments with both opponent methods and our method O-CBS+($\mathcal{E}=0.75$) are repeated 100 times.

In Figure 7.5, we compare our method with the Mean-Random baseline method and the method of Welinder and Perona [37] on the Age Annotations and Head Pose Annotations datasets. Additionally, we make similar comparisons with the method of Raykar and Agrawal [53] on the Affective Text Analysis datasets in Figure 7.6. By the very nature of active crowd-labeling, annotations of the samples are acquired gradually. Thus, in the early steps of the process, not every sample has an estimated label. Moreover, the required number of annotations for obtaining consensus label of every sample varies depending on the sample selection strategy of the method in question. However, for the mean absolute error (MAE) and accuracy comparisons to make sense, every sample's consensus error must contribute to the mean. For this reason, we represent the initial part of the process where some sample labels do not have estimations by dotted lines in the plots. Additionally, both methods by Welinder and Perona [37] and Raykar and Agrawal [53] employ stopping criteria which results in the algorithms stopping at different annotation counts among 100 repetitions. Therefore, the ends of the curves are also shown in dotted lines when the MAE or the accuracy is calculated with fewer than 100 repetitions. The middle portions of the curves are shown in solid lines.

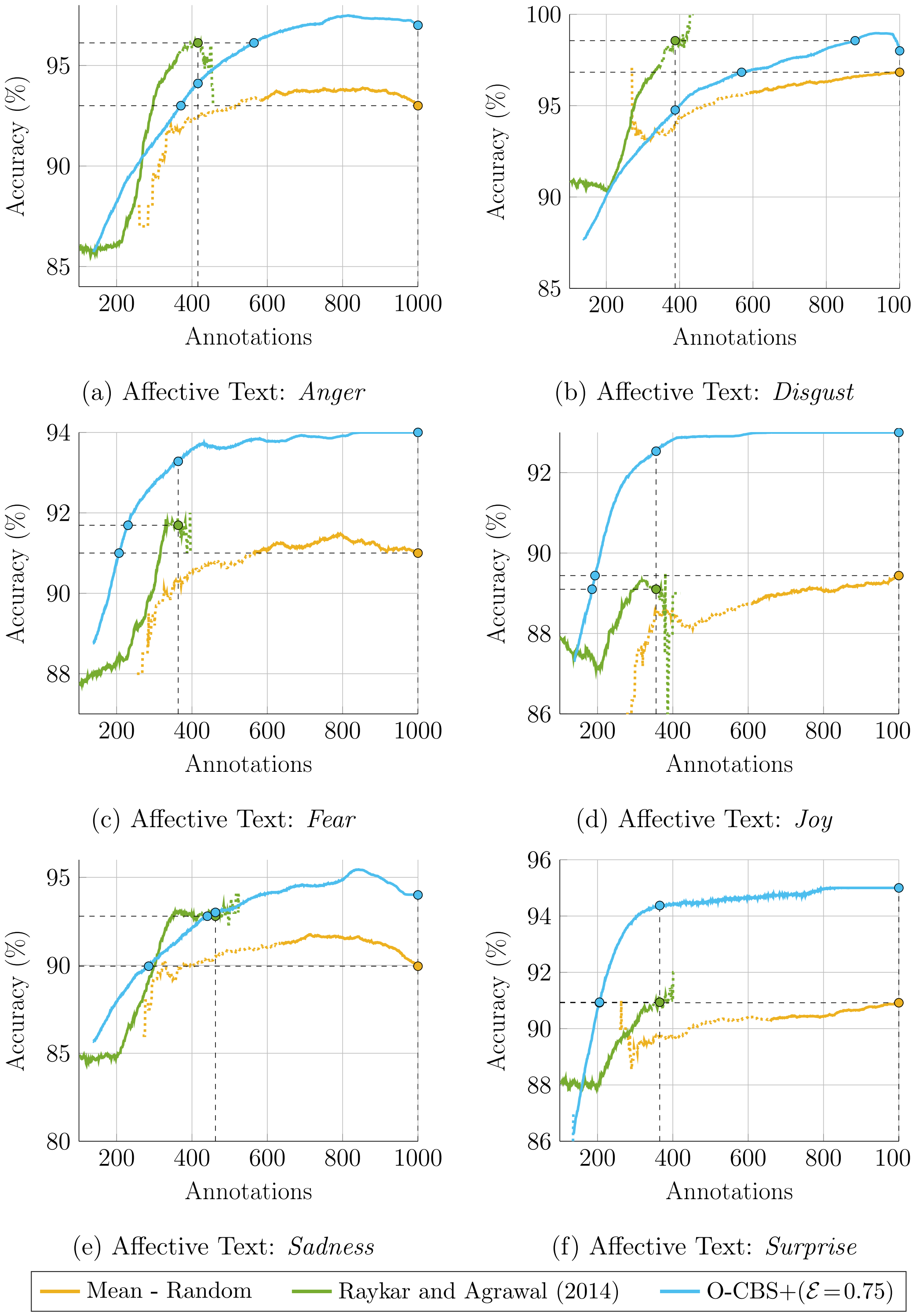


(a) Age Annotations

(b) Head Pose: *Tilt*

(c) Head Pose: *Pan*

Figure 7.5. Comparison of O-CBS+ with the method of Welinder and Perona [37] on the Age Annotations and the Head Pose Annotations datasets. The circles mark the required annotation counts for our method to reach the performances of the contender methods. The horizontal black dashed lines provide visual guide.



(a) Affective Text: *Anger*

(b) Affective Text: *Disgust*

(c) Affective Text: *Fear*

(d) Affective Text: *Joy*

(e) Affective Text: *Sadness*

(f) Affective Text: *Surprise*

Figure 7.6. Comparison of O-CBS+ with the method of Raykar and Agrawal [53] on the Affective Text Analysis datasets. The circles mark the required annotation counts for our method to reach the performances of the contender methods. The horizontal black dashed lines provide visual guide.

In the Mean-Random baseline method, the annotations are added randomly and the mean of the annotations of a sample are used as the resulting label. In Figure 7.5a, we observe that the mean absolute age error achieved by this baseline method on the Age Annotations dataset using all 10020 annotations can be matched by O-CBS+ with 1796 annotations (∼18% of all annotations). Figures 7.5b and 7.5c show that our method can match the performance of the Mean-Random baseline method on the Head Pose Annotations *tilt* and *pan* datasets with 2886 and 836 annotations (∼53% and ∼15% of all annotations), respectively.

Figures 7.6a to 7.6f present the performance of O-CBS+($\mathcal{E}=0.75$) against the method of Raykar and Agrawal [53] on the Affective Text Analysis datasets, accompanied with the Mean-Random method as the baseline. Similar to Figures 7.5a to 7.5c, O-CBS+($\mathcal{E}=0.75$) outperforms the Mean-Random method across all six datasets. Our method matches the end result of the Mean-Random method, using a minimum of 193 and a maximum of 569 annotations across the six datasets, and thereby resulting in a ∼70% cost reduction on average.

We support the findings of Figures 7.5 and 7.6 with a more detailed breakdown of the comparative results, presented in Table 7.1. We perform t-test for validating the statistical significance of the results presented in Figures 7.5 and 7.6. For comparison, we take the number of annotations at which an opponent algorithm stops, and use this as a stopping criterion for O-CBS+($\mathcal{E}=0.75$) to report the MAE or accuracy. Additionally, we also take the MAE or accuracy at which an opponent algorithm stops, and report the mean number of annotations needed to reach this target using O-CBS+($\mathcal{E}=0.75$). Significance test results against opponent methods are reported under the rightmost two columns, where the underlined values indicate that our method is significantly superior than the opponent method. Values written in regular font indicate a tie and italic values indicate that the opponent method is better. The results for the opponent methods are given in regular script as reference values.

In Table 7.1a, we observe the significance test results of O-CBS+($\mathcal{E}=0.75$) against the method of Welinder and Perona [37]. On the Age Annotations dataset, the al-

Table 7.1. The effect of enforcing annotation count or MAE/accuracy limit. The tables indicate the results of the t-test with significance level 0.01 across 100 repetitions, using underlined font when our method performs better, regular font when the test is inconclusive, and italic font when our method performs *worse.*

(a) Comparison with Welinder and Perona [37] on the Age Annotations and the Head Pose Annotations datasets

	Welinder and Perona [37]		O-CBS+($\mathcal{E}=0.75$)	Required annotations for O-CBS+($\mathcal{E}=0.75$) to reach target MAE
Dataset	Annotations	MAE	MAE at target annotations	
Age	4969.77	7.02 ages	6.06 ages	2775.98
Tilt	2705.03	10.10 degrees	9.33 degrees	1892.16
Pan	2689.77	7.58 degrees	6.49 degrees	1387.88

(b) Comparison with Raykar and Agrawal [53] on the Affective Text Analysis datasets

	Raykar and Agrawal [53]		O-CBS+($\mathcal{E}=0.75$)	Required annotations for O-CBS+($\mathcal{E}=0.75$) to reach target acc.
Dataset	Annotations	Accuracy (%)	Accuracy at target annotations (%)	
Anger	415.86	96.07	*94.11*	*535.81*
Disgust	387.78	98.92	*94.76*	*726.82*
Fear	363.49	91.50	93.28	247.32
Joy	355.51	89.17	92.53	196.22
Sadness	462.34	93.31	93.01	*522.80*
Surprise	365.22	91.60	94.38	231.41

gorithm of Welinder and Perona [37] stops at 4970 annotations on average and a little more than half on the annotations are unused because they come from annotators marked as spammers. At this point, the lowest mean absolute error is reached. For matching the same MAE, our method requires 2776 annotations on average, and achieves better overall performance as more annotators are employed. Similar results are also observed for the the Head Pose Annotations *tilt* and *pan* datasets, where O-CBS+ proves to be an effective algorithm both in terms of achieving significantly lower error with annotation count limitations and by using significantly fewer annotations for a targeted MAE.

Note that Welinder and Perona [37] do not employ sample prioritization. They acquire annotations for each sample one by one. For each sample, they acquire as many annotations as they can and move onto the next sample. Thus, the point where each sample has a consensus value occurs later in the annotation process. This is why in Figures 7.5a to 7.5c the red curves preceding the red dots are almost invisible since the annotation acquisition process stops after a very short while. Once their algorithm flags an annotator as a spammer, that annotator is not consulted anymore.

Compared to the method of Welinder and Perona [37], our method uses a more complex scheme. First, we employ sample prioritization by sample consensus quality scoring. Second, instead of grouping the annotators into two discrete groups as spammers and non-spammers, we rank them according to four parameters for each annotator. This way, better annotators are also ranked among themselves while low-quality annotators are ignored until the end of the annotation process. Low-quality annotators may also be completely excluded from the annotation process by a simple thresholding mechanism on the annotator competence score.

An additional observation about these methods' performances on the *tilt* dataset is that the algorithm of Welinder and Perona [37] falls short of achieving the Mean-Random baseline method's performance. This is due to the fact that many annotators are marked as spammers and the annotation process stops very early. Another reason is that the *tilt* dataset is actually quite a challenging dataset in the sense that the baseline method achieves a close performance to our method O-CBS+($\mathcal{E} = 0.75$), albeit using all annotations.

We present the performance of O-CBS+($\mathcal{E} = 0.75$) against the method of Raykar and Agrawal [53] on the Affective Text Analysis datasets in Figures 7.6a to 7.6f and Table 7.1b. In contrast to Welinder and Perona, Raykar and Agrawal [53] employ a more intricate annotation selection algorithm and the change in the accuracy over time (the green lines in Figures 7.6a to 7.6f) is observable since all samples have annotations. Our method succeeds to achieve higher accuracies at the targeted number of annotations in the *fear*, *joy*, and *surprise* datasets with a significant margin and is tied on the *sadness* dataset. Although our method seems to struggle in the *anger* and *disgust* datasets, observing Figures 7.6a and 7.6b shows that the overall performance of our method in the long run (*i.e.* without annotation count limit) is capable of achieving a higher or similar accuracy. These findings confirm that O-CBS+ is overall a better approach to the active crowd-labeling problem with significant gains on annotation expenses.

7.4. Comparative Performance of O-CBS+ While Enforcing a Sample Score Related Stopping Criterion

In Section 5.1, S_S is defined as the precision (reciprocal of the variance) of the posterior distribution of the sample consensus. In both O-CBS and O-CBS+, the aim is to reduce this variance value (*i.e.* increase S_S) for each sample. The algorithms are designed to choose the sample with the lowest S_S to be annotated in each annotation step. Thus, the overall direction is the enhancement of every sample's score (*i.e.* reducing the sample consensus posterior variance) during the course of active crowd-labeling.

So far, we were not concerned with the question of how high S_S should be for having a satisfactory sample consensus. Our aim was to increase consensus quality as much as possible within the annotation budget limit. In Figures 6.2, 6.3, 6.6, 6.7, 7.2, 7.3, 7.5 and 7.6, we show the performance of the proposed methods with only the budget limit as an enforceable stopping criterion. Every point on those graphs actually show the performance of the corresponding method for every possible annotation budget limit. However, this approach does not consider the adequacy of sample consensus values, and is at risk of prematurely ending the active crowd-labeling process or overspending by collecting excessive annotations.

To address this concern, we aim to stop the annotation process upon attaining satisfactory sample consensus values for all samples by setting a target on the sample

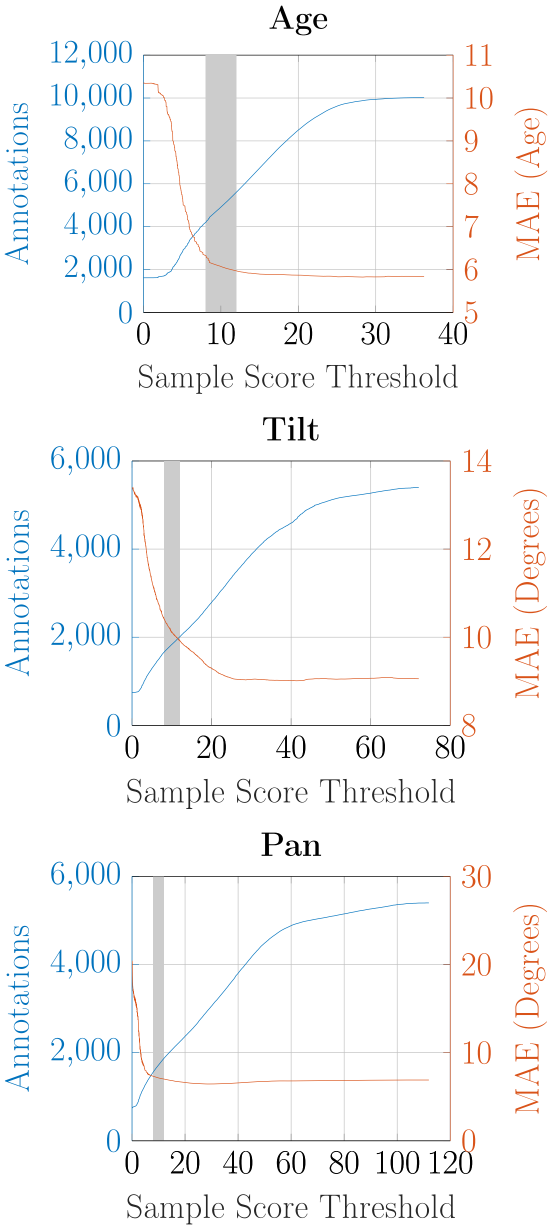


Figure 7.7. The effect of enforcing the sample scoring threshold τ on the Age Annotations and Head Pose Annotations Datasets. Blue curves show the final annotation count when τ is enforced and red curves show the MAE at the end of the annotation process for a given τ. The gray bands depict the region where $8 \leq \tau \leq 12$.

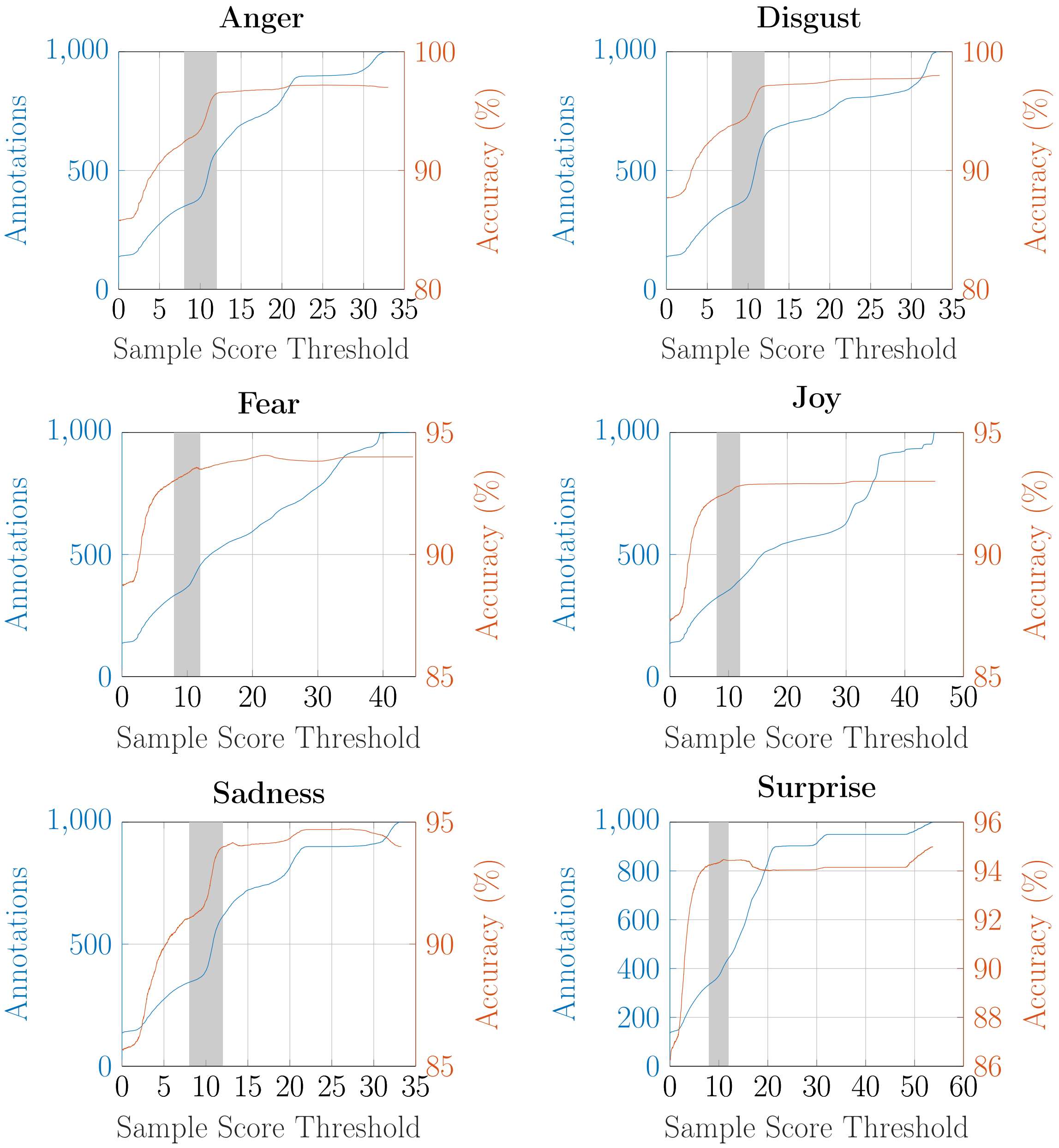


Figure 7.8. The effect of enforcing the sample scoring threshold τ on the Affective Text Analysis Datasets. Blue curves show the final annotation count when τ is enforced and red curves show the MAE at the end of the annotation process for a given τ. The gray bands depict the region where $8 \leq \tau \leq 12$.

consensus posterior variance, namely δ. This is equivalent to stopping the active crowd-labeling process when every sample has a satisfactory score S_S since S_S is the reciprocal of the posterior variance, *i.e.*

$$\min_i S_S(i) > \underbrace{\frac{1}{\delta}}_{\tau} \tag{7.1}$$

Therefore, τ signifies the target lower limit on S_S.

The cost associated with the active crowd-labeling systems consists of not only the annotation budget, but also the cost of reaching erroneous consensuses (which may also have monetary repercussions). System designers are often faced with making a trade-off between performance and budget to find a sensible operation range. In our case, collecting more annotations often result in reduced error while increasing expenses. Due to the nature of the sample score S_S and O-CBS+, choosing a high τ value would result in lower error and is preferable if the cost of making error is high. In contrast, system designers working with very limited budgets may resort to using a lower τ value. A reasonably low value for the posterior variance of a sample's consensus is 0.1. Enforcing a stopping criterion to reach this goal for each sample corresponds to choosing $\tau = 10$.

In Figures 7.7 and 7.8, we show the performance of O-CBS+($\mathcal{E}=0.75$) for varying τ values. Blue curves show the final annotation count (*i.e.* cost) when τ is enforced and red curves show the performance at the end of the annotation process for a given τ. The gray bands in the plots show the region around $\tau = 10$; specifically, the bands rest between $\tau = 8$ and $\tau = 12$. The plots show promising performance and annotation count values inside the gray bands. The results verify our previous deductions. Especially, for *anger*, *disgust*, and *sadness* datasets where our methods struggle, $\tau = 10$ presents a turning point for both error and budget. Additionally, in the remaining datasets the gray band areas signify very preferable operation ranges.

Table 7.2. The effect of enforcing various sample scoring thresholds. The tables indicate the results of the t-test with significance level 0.01 across 100 repetitions, using underlined font when our method performs better, regular font when the test is inconclusive, and italic font when our method performs *worse*.

(a) Comparison with Welinder and Perona on the Age Annotations and the Head Pose Annotations datasets

	Welinder and Perona [37]		O-CBS+($\mathcal{E}=0.75$)					
			$\tau = 8$		$\tau = 10$		$\tau = 12$	
Dataset	Ann.	MAE	Ann.	MAE	Ann.	MAE	Ann.	MAE
Age	4969.77	7.02	4189.93	6.33	4911.37	6.07	*5607.13*	5.97
Tilt	2705.03	10.10	1657.70	*10.42*	1836.39	10.11	2009.94	9.92
Pan	2689.77	7.58	1560.16	7.32	1721.22	7.13	1868.02	7.01

(b) Comparison with Raykar and Agrawal [53] on the Affective Text Analysis datasets

	Raykar and Agrawal [53]		O-CBS+($\mathcal{E}=0.75$)					
			$\tau = 8$		$\tau = 10$		$\tau = 12$	
Dataset	Ann.	Acc.(%)	Ann.	Acc.(%)	Ann.	Acc.(%)	Ann.	Acc.(%)
Anger	415.86	96.07	347.83	*93.38*	386.20	*94.58*	*564.59*	97.24
Disgust	387.78	98.92	346.12	*94.64*	392.72	*95.53*	*625.24*	*97.41*
Fear	363.49	91.50	331.49	93.45	365.74	93.77	*458.29*	93.74
Joy	355.51	89.17	323.10	92.59	352.96	92.79	*394.22*	92.98
Sadness	462.34	93.31	343.58	*91.96*	390.84	*92.72*	*603.89*	94.50
Surprise	365.22	91.60	334.87	94.60	371.00	94.67	*447.00*	94.64

In Table 7.2, we give the results of O-CBS+($\mathcal{E}=0.75$) for different τ values compared to the methods of Welinder and Perona [37] and Raykar and Agrawal [53]. The experiments with both the opponent methods and our method O-CBS+($\mathcal{E}=0.75$) are repeated 100 times. We perform t-test for validating the statistical significance of the results. We report the number of annotations and the error/accuracy when our algorithm stops for the τ values 8, 10, and 12. Significance test results against opponent methods are reported under the O-CBS+($\mathcal{E}=0.75$) heading, where underlined values indicate that our method is significantly superior than the opponent method. Values written in regular font indicate a tie and italic values indicate that the opponent method is better. The results for the opponent methods are given in regular script as reference values.

The results show that for $\tau = 8$, the number or annotations at which our algorithm stops are always significantly lower than its contenders, with acceptable error or accuracy values. When $\tau = 10$, our algorithm is tied with or better than its contenders in terms of annotation count and the accuracies improve, especially for the *tilt*, *anger*, *disgust*, and *sadness* datasets. For $\tau = 12$, our algorithm achieves significantly superior performance across all datasets except *disgust* in terms of error/accuracy at the expense of increasing cost.

8. A VARIATIONAL BAYESIAN APPROACH TO CROWD-LABELING WITH MULTIVARIATE ANNOTATIONS

Some annotation problems have multiple attributes annotated by the same annotator, as observed in the Head Pose and the Affective Text annotations datasets. In these cases, one might expect to observe that different attributes correlate with each other. Multivariate models which take these correlations into account may prove useful in consensus estimation. In this chapter, we propose a multivariate annotation model and give its variational Bayesian solution. We test our method on the Head Pose Annotations dataset.

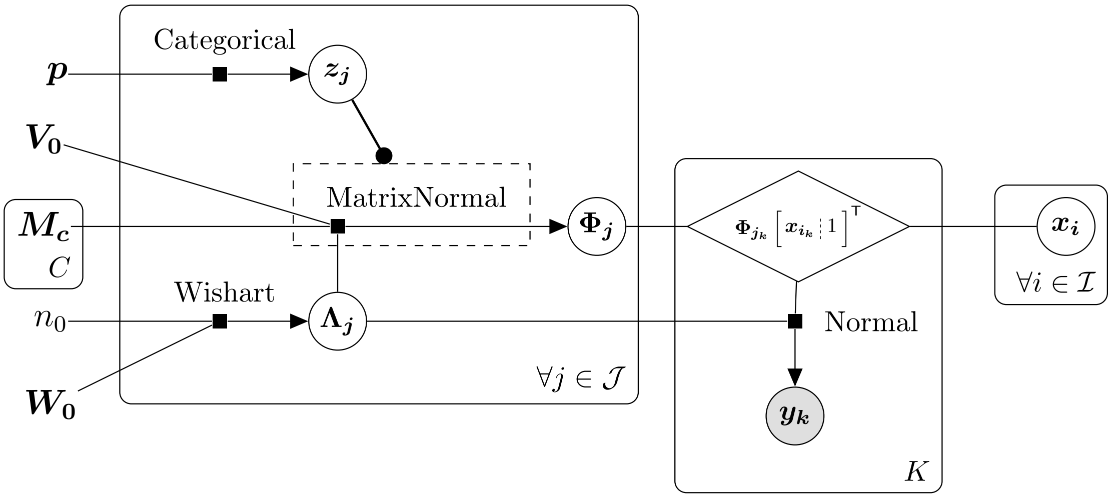


Figure 8.1. Directed factor graph of the proposed multivariate annotation model.

8.1. The Multivariate Annotation Model

In Figure 8.1, we illustrate the proposed multivariate annotation model. In the figure, $\boldsymbol{y_k}$ denotes an observed annotation, $\boldsymbol{x_i}$ denotes the sought consensus value of the sample i, and the parameters with subscript j are the latent parameters of the annotator j. We denote the observed set of annotations by $Y = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_K\}$ where K is the number of annotations and $\boldsymbol{y_k} \in \mathbb{R}^d$, $\forall k$. The set of latent consensus values is denoted by $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ where N is the number of samples and $\boldsymbol{x_i} \in \mathbb{R}^d$, $\forall i$. The sets of latent annotator parameters are denoted by $\Phi = \{\boldsymbol{\Phi_1}, \ldots, \boldsymbol{\Phi_R}\}$, $\Lambda = \{\boldsymbol{\Lambda_1}, \ldots, \boldsymbol{\Lambda_R}\}$, and $Z = \{\boldsymbol{z_1}, \ldots, \boldsymbol{z_R}\}$ where R is the number of annotators, $\boldsymbol{\Phi_j} \in \mathbb{R}^{d\times d+1}$, $\boldsymbol{\Lambda_j}$ is a $d{\times}d$ positive definite matrix, and $\boldsymbol{z_j}$ is a 1-of-C binary vector (*i.e.*, one element of the vector is one and the rest are zero) with elements $\boldsymbol{z_j}_c$. The prior variables of the model are $\boldsymbol{p}$, $\boldsymbol{V_0}$, $\boldsymbol{W_0}$, n_0, and $\{\boldsymbol{M_c} | c \in \{1, \ldots, C\}\}$ where $\boldsymbol{p} \in [0,1]^C$, such that $\sum_{c=1}^{C} \boldsymbol{p}_c = 1$, $\boldsymbol{V_0}$ is a $(d{+}1){\times}(d{+}1)$ positive definite matrix, $\boldsymbol{M_c} \in \mathbb{R}^{d\times(d+1)}, \forall c$, $n_0 > d{-}1$ is a real scalar, and $\boldsymbol{W_0}$ is a $d{\times}d$ positive definite matrix.

We model the conditional distribution of the observed annotations given the latent annotator parameters and the consensus as

$$p(Y|\Phi,\Lambda,X) = \prod_{k=1}^{K} \mathcal{N}_d\left(\boldsymbol{y_k}; \boldsymbol{\Phi_{j_k}} \boldsymbol{\chi_{i_k}}, \boldsymbol{\Lambda_{j_k}^{-1}}\right) \tag{8.1}$$

where $\boldsymbol{\chi_{i_k}} = \begin{bmatrix} \boldsymbol{x_{i_k}} \\ \hdashline 1 \end{bmatrix}$.

We choose a flat prior over X and introduce priors over the annotator parameters as

$$p(\Phi,\Lambda,Z) = p(\Phi|\Lambda,Z)p(\Lambda)p(Z) \tag{8.2}$$

where

$$\log p(\Phi|\Lambda,Z) = \prod_{j=1}^{R}\prod_{c=1}^{C} \mathcal{MN}_{d,d+1}\left(\boldsymbol{\Phi_j}; \boldsymbol{M_c}, \boldsymbol{\Lambda_j^{-1}}, \boldsymbol{V_0}\right)^{\boldsymbol{z_j}_c} \tag{8.3}$$

$$p(\Lambda) = \prod_{j=1}^{R} \mathcal{W}_d\left(\boldsymbol{\Lambda_j}; \boldsymbol{W_0}, n_0\right) \tag{8.4}$$

$$p(Z) = \prod_{j=1}^{R} \mathcal{C}\left(\boldsymbol{z_j}; \boldsymbol{p}\right) = \prod_{j=1}^{R}\prod_{c=1}^{C} \boldsymbol{p}_c^{\boldsymbol{z_j}_c} \tag{8.5}$$

Then, the joint distribution of all of the random variables is given by

$$p(Y,\Phi,\Lambda,Z,X)=p(Y|\Phi,\Lambda,X)p(\Phi|\Lambda,Z)p(\Lambda)p(Z)p(X) \tag{8.6}$$

8.2. Variational Distribution

Our model defines the joint distribution $p(Y,\Phi,\Lambda,Z,X)$. We aim to find the posterior distribution of the latent variables Φ, Λ, Z, and X given the observed variables Y. Using the variational Bayes approach, we can approximate the said posterior distribution, $p(\Phi,\Lambda,Z,X|Y)$. Let us decompose $\log p(Y)$ as

$$\log p(Y)=\mathcal{L}(q)+\mathrm{KL}(q\|p) \tag{8.7}$$

where

$$\mathcal{L}(q)=\sum_{Z}\iiint q(\Phi,\Lambda,Z,X)\log\frac{p(Y,\Phi,\Lambda,Z,X)}{q(\Phi,\Lambda,Z,X)}d\Phi d\Lambda dX \tag{8.8}$$

$$\mathrm{KL}(q\|p)=-\sum_{Z}\iiint q(\Phi,\Lambda,Z,X)\log\frac{p(\Phi,\Lambda,Z,X|Y)}{q(\Phi,\Lambda,Z,X)}d\Phi d\Lambda dX \tag{8.9}$$

We want $q(\Phi,\Lambda,Z,X)$ to be a good approximation of $p(\Phi,\Lambda,Z,X|Y)$. Thus, we want to minimize the KL divergence, $\mathrm{KL}(q\|p)$, or equivalently, maximize the lower bound value, $\mathcal{L}(q)$. For obtaining a tractable solution to our model, we consider a variational distribution factorizing the latent variables into three partitions such as

$$q(\Phi,\Lambda,Z,X)=q(\Phi,\Lambda)q(Z)q(X) \tag{8.10}$$

Moreover, these factors can be further factorized:

$$q(\Phi,\Lambda)=\prod_{j=1}^{R}q(\boldsymbol{\Phi_j},\boldsymbol{\Lambda_j})=\prod_{j=1}^{R}q(\boldsymbol{\Phi_j}|\boldsymbol{\Lambda_j})q(\boldsymbol{\Lambda_j}) \tag{8.11}$$

$$q(Z) = \prod_{j=1}^{R} q(\boldsymbol{z_j}) \tag{8.12}$$

$$q(X) = \prod_{i=1}^{N} q(\boldsymbol{x_i}) \tag{8.13}$$

The logarithm of an optimized factor ($\log q^*(\cdot)$) is obtained by calculating the expected value of $\log p(Y, \Phi, \Lambda, Z, X)$ with respect to the distributions of the other factors. Now, we derive the logarithm of the conditional annotation distribution which is

$$\begin{aligned}
\mathcal{L}_y = \log p(Y|\Phi, \Lambda, X) &= \sum_{k=1}^{K} \log \mathcal{N}_d \left(\boldsymbol{y_k}; \boldsymbol{\Phi_{j_k}} \boldsymbol{\chi_{i_k}}, \boldsymbol{\Lambda_{j_k}^{-1}}\right) \\
&= \sum_{k=1}^{K} \left(-\frac{1}{2} \log |2\pi \boldsymbol{\Lambda_{j_k}^{-1}}| - \frac{1}{2} \left(\boldsymbol{y_k} - \boldsymbol{\Phi_{j_k}} \boldsymbol{\chi_{i_k}}\right)^{\top} \boldsymbol{\Lambda_{j_k}} \left(\boldsymbol{y_k} - \boldsymbol{\Phi_{j_k}} \boldsymbol{\chi_{i_k}}\right) \right) \\
&= \sum_{k=1}^{K} \left(-\frac{1}{2} \log |2\pi \boldsymbol{\Lambda_{j_k}^{-1}}| - \frac{1}{2} \boldsymbol{y_k^{\top}} \boldsymbol{\Lambda_{j_k}} \boldsymbol{y_k} + \boldsymbol{\chi_{i_k}^{\top}} \boldsymbol{\Phi_{j_k}^{\top}} \boldsymbol{\Lambda_{j_k}} \boldsymbol{y_k} - \frac{1}{2} \boldsymbol{\chi_{i_k}^{\top}} \boldsymbol{\Phi_{j_k}^{\top}} \boldsymbol{\Lambda_{j_k}} \boldsymbol{\Phi_{j_k}} \boldsymbol{\chi_{i_k}} \right)
\end{aligned} \tag{8.14}$$

Similarly, the logarithms of the priors of Φ, Λ, and Z are

$$\begin{aligned}
\mathcal{L}_\Phi = \log p(\Phi|\Lambda, Z) &= \sum_{j=1}^{R} \sum_{c=1}^{C} \boldsymbol{z_j}_c \log \mathcal{MN}_{d,d+1} \left(\boldsymbol{\Phi_j}; \boldsymbol{M_c}, \boldsymbol{\Lambda_j^{-1}}, \boldsymbol{V_0}\right) \\
&= \sum_{j=1}^{R} \sum_{c=1}^{C} \boldsymbol{z_j}_c \left[\frac{d+1}{2} \log |\boldsymbol{\Lambda_j}| - \frac{d(d+1)}{2} \log(2\pi) - \frac{1}{2} \operatorname{Tr} \left(\boldsymbol{V_0^{-1}} \boldsymbol{\Phi_j^{\top}} \boldsymbol{\Lambda_j} \boldsymbol{\Phi_j}\right) \right. \\
&\qquad \left. - \frac{d}{2} \log |\boldsymbol{V_0}| + \operatorname{Tr} \left(\boldsymbol{V_0^{-1}} \boldsymbol{\Phi_j^{\top}} \boldsymbol{\Lambda_j} \boldsymbol{M_c}\right) - \frac{1}{2} \operatorname{Tr} \left(\boldsymbol{V_0^{-1}} \boldsymbol{M_c^{\top}} \boldsymbol{\Lambda_j} \boldsymbol{M_c}\right) \right] \\
&= \sum_{j=1}^{R} \left[\frac{d+1}{2} \log |\boldsymbol{\Lambda_j}| - \frac{d(d+1)}{2} \log(2\pi) - \frac{d}{2} \log |\boldsymbol{V_0}| - \frac{1}{2} \operatorname{Tr} \left(\boldsymbol{V_0^{-1}} \boldsymbol{\Phi_j^{\top}} \boldsymbol{\Lambda_j} \boldsymbol{\Phi_j}\right) \right. \\
&\qquad \left. + \operatorname{Tr} \left(\boldsymbol{V_0^{-1}} \boldsymbol{\Phi_j^{\top}} \boldsymbol{\Lambda_j} \sum_{c=1}^{C} \boldsymbol{z_j}_c \boldsymbol{M_c} \right) - \frac{1}{2} \operatorname{Tr} \left(\sum_{c=1}^{C} \boldsymbol{z_j}_c \left(\boldsymbol{M_c} \boldsymbol{V_0^{-1}} \boldsymbol{M_c^{\top}}\right) \boldsymbol{\Lambda_j} \right) \right]
\end{aligned} \tag{8.15}$$

$$\mathcal{L}_\Lambda = \log p(\Lambda) = \sum_{j=1}^{R} \log \mathcal{W}_d \left(\boldsymbol{\Lambda_j}; \boldsymbol{W_0}, n_0\right)$$

$$= \sum_{j=1}^{R} \left(\frac{n_0-d-1}{2} \log |\boldsymbol{\Lambda_j}| - \frac{1}{2} \operatorname{Tr}(\boldsymbol{W_0}^{-1} \boldsymbol{\Lambda_j}) - \frac{n_0 d}{2} \log 2 - \frac{n_0}{2} \log |\boldsymbol{W_0}| - \log \boldsymbol{\Gamma}_d \left(\frac{n_0}{2} \right) \right) \tag{8.16}$$

$$\mathcal{L}_z = \log p(Z) = \sum_{j=1}^{R} \sum_{c=1}^{C} \boldsymbol{z_j}_c \log \boldsymbol{p}_c \tag{8.17}$$

Since we choose the prior of X as flat, its logarithm $\mathcal{L}_x$ is constant. Then, the logarithm of the full joint is

$$\log p(Y, \Phi, \Lambda, Z, X) = \mathcal{L}_y + \mathcal{L}_\Phi + \mathcal{L}_\Lambda + \mathcal{L}_z + \mathcal{L}_x \tag{8.18}$$

8.2.1. The Factor $q(\Phi, \Lambda)$

Let us consider the factor $q(\boldsymbol{\Phi_j}, \boldsymbol{\Lambda_j})$. The logarithm of the optimal value can be expressed as:

$$\begin{aligned} \log q^*(\boldsymbol{\Phi_j}, \boldsymbol{\Lambda_j}) &= \mathbb{E}_{X,Z} \left[\log p(Y, \Phi, \Lambda, Z, X) \right] + const \\ &= \log q^*(\boldsymbol{\Phi_j} | \boldsymbol{\Lambda_j}) + \log q^*(\boldsymbol{\Lambda_j}) \end{aligned} \tag{8.19}$$

We only consider the terms with $\boldsymbol{\Phi_j}$ and take their expectation with respect to X and Z. Then, we have

$$\begin{aligned} \log q^*(\boldsymbol{\Phi_j} | \boldsymbol{\Lambda_j}) = \mathbb{E}_{X,Z} & \left[-\frac{1}{2} \operatorname{Tr} \left(\boldsymbol{V_0}^{-1} \boldsymbol{\Phi_j}^\intercal \boldsymbol{\Lambda_j} \boldsymbol{\Phi_j} \right) + \operatorname{Tr} \left(\boldsymbol{V_0}^{-1} \boldsymbol{\Phi_j}^\intercal \boldsymbol{\Lambda_j} \sum_{c=1}^{C} \boldsymbol{z_j}_c \boldsymbol{M_c} \right) \right] \\ & + \mathbb{E}_{X,Z} \left[\sum_{k: j_k = j} \left(\boldsymbol{\chi}_{\boldsymbol{i_k}}^\intercal \boldsymbol{\Phi_j}^\intercal \boldsymbol{\Lambda_j} \boldsymbol{y_k} - \frac{1}{2} \boldsymbol{\chi}_{\boldsymbol{i_k}}^\intercal \boldsymbol{\Phi_j}^\intercal \boldsymbol{\Lambda_j} \boldsymbol{\Phi_j} \boldsymbol{\chi}_{\boldsymbol{i_k}} \right) \right] + const \\ = & -\frac{1}{2} \operatorname{Tr} \left(\boldsymbol{V_0}^{-1} \boldsymbol{\Phi_j}^\intercal \boldsymbol{\Lambda_j} \boldsymbol{\Phi_j} \right) + \operatorname{Tr} \left(\sum_{c=1}^{C} \mathbb{E} \left[\boldsymbol{z_j}_c \right] \boldsymbol{M_c} \boldsymbol{V_0}^{-1} \boldsymbol{\Phi_j}^\intercal \boldsymbol{\Lambda_j} \right) \end{aligned}$$

$$\begin{aligned}
&+ \operatorname{Tr}\left(\sum_{k:j_k=j} \boldsymbol{y_k} \mathbb{E}\left[\boldsymbol{\chi_{i_k}}\right]^{\mathsf{T}} \boldsymbol{\Phi_j}^{\mathsf{T}} \boldsymbol{\Lambda_j} \right) \\
&- \frac{1}{2} \operatorname{Tr}\left(\sum_{k:j_k=j} \mathbb{E}_{\boldsymbol{x_{i_k}}} \left[\boldsymbol{\chi_{i_k}} \boldsymbol{\chi_{i_k}}^{\mathsf{T}}\right] \boldsymbol{\Phi_j}^{\mathsf{T}} \boldsymbol{\Lambda_{j_k}} \boldsymbol{\Phi_j} \right) + const \\
= const &- \frac{1}{2} \operatorname{Tr}\left(\left(\boldsymbol{V_0}^{-1} + \sum_{k:j_k=j} \mathbb{E}_{\boldsymbol{x_{i_k}}} \left[\boldsymbol{\chi_{i_k}} \boldsymbol{\chi_{i_k}}^{\mathsf{T}}\right] \right) \boldsymbol{\Phi_j}^{\mathsf{T}} \boldsymbol{\Lambda_j} \boldsymbol{\Phi_j} \right) \\
&+ \operatorname{Tr}\left(\left(\sum_{c=1}^{C} \mathbb{E}\left[\boldsymbol{z_{j}}_c\right] \boldsymbol{M_c} \boldsymbol{V_0}^{-1} + \sum_{k:j_k=j} \boldsymbol{y_k} \mathbb{E}\left[\boldsymbol{\chi_{i_k}}\right]^{\mathsf{T}} \right) \boldsymbol{\Phi_j}^{\mathsf{T}} \boldsymbol{\Lambda_j} \right)
\end{aligned} \tag{8.20}$$

We observe that the equation above is in the form of the matrix normal distribution, given by

$$q^*(\boldsymbol{\Phi_j} | \boldsymbol{\Lambda_j}) = \mathcal{MN}_{d,d+1}\left(\boldsymbol{\Phi_j}; \boldsymbol{M_j}, \boldsymbol{\Lambda_j}^{-1}, \boldsymbol{V_j}\right) \tag{8.21}$$

where

$$\boldsymbol{V_j}^{-1} = \boldsymbol{V_0}^{-1} + \sum_{k:j_k=j} \mathbb{E}_{\boldsymbol{x_{i_k}}} \left[\boldsymbol{\chi_{i_k}} \boldsymbol{\chi_{i_k}}^{\mathsf{T}}\right] \tag{8.22}$$

$$\boldsymbol{M_j} = \left(\sum_{c=1}^{C} \mathbb{E}\left[\boldsymbol{z_{j}}_c\right] \boldsymbol{M_c} \boldsymbol{V_0}^{-1} + \sum_{k:j_k=j} \boldsymbol{y_k} \mathbb{E}_{\boldsymbol{x_{i_k}}} \left[\boldsymbol{\chi_{i_k}}\right]^{\mathsf{T}} \right) \boldsymbol{V_j} \tag{8.23}$$

After removing the terms of $\log q^*(\boldsymbol{\Phi_j} | \boldsymbol{\Lambda_j})$ and considering the remaining terms with $\boldsymbol{\Lambda_j}$, we have

$$\begin{aligned}
\log q^*(\boldsymbol{\Lambda_j}) &= \mathbb{E}\left[\log p(Y, \Phi, \Lambda, Z, X)\right] - \log q^*(\boldsymbol{\Phi_j} | \boldsymbol{\Lambda_j}) + const \\
&= \frac{n_0 - d - 1}{2} \log |\boldsymbol{\Lambda_j}| + \sum_{k:j_k=j} \left(\frac{1}{2} \log |\boldsymbol{\Lambda_j}| - \frac{1}{2} \boldsymbol{y_k}^{\mathsf{T}} \boldsymbol{\Lambda_j} \boldsymbol{y_k} \right) \\
&\quad - \frac{1}{2} \operatorname{Tr}(\boldsymbol{W_0}^{-1} \boldsymbol{\Lambda_j}) - \frac{1}{2} \mathbb{E}_Z \left[\operatorname{Tr}\left(\sum_{c=1}^{C} \boldsymbol{z_{j}}_c \left(\boldsymbol{M_c} \boldsymbol{V_0}^{-1} \boldsymbol{M_c}^{\mathsf{T}} \right) \boldsymbol{\Lambda_j} \right) \right] \\
&\quad + \frac{1}{2} \operatorname{Tr}\left(\boldsymbol{M_j} \boldsymbol{V_j}^{-1} \boldsymbol{M_j}^{\mathsf{T}} \boldsymbol{\Lambda_j} \right) + const
\end{aligned} \tag{8.24}$$

This equation is in the form of the Wishart distribution, given by

$$q^*(\boldsymbol{\Lambda_j}) = \mathcal{W}_d\left(\boldsymbol{\Lambda_j}; \boldsymbol{W_j}, n_0 + N_j\right) \tag{8.25}$$

where

$$\boldsymbol{W_j}^{-1} = \boldsymbol{W_0}^{-1} + \sum_{k:j_k=j} \boldsymbol{y_k}\boldsymbol{y_k}^{\mathsf{T}} + \sum_{c=1}^{C} \mathbb{E}\left[\boldsymbol{z}_{\boldsymbol{j}c}\right]\left(\boldsymbol{M_c}\boldsymbol{V_0}^{-1}\boldsymbol{M_c}^{\mathsf{T}}\right) - \boldsymbol{M_j}\boldsymbol{V_j}^{-1}\boldsymbol{M_j}^{\mathsf{T}} \tag{8.26}$$

Alternatively, $\boldsymbol{W_j}^{-1}$ can also be represented as

$$\begin{aligned}\boldsymbol{W_j}^{-1} = \boldsymbol{W_0}^{-1} + &\sum_{c=1}^{C} \mathbb{E}\left[\boldsymbol{z}_{\boldsymbol{j}c}\right](\boldsymbol{M_c} - \boldsymbol{M}_{\boldsymbol{j_z}})\boldsymbol{V_0}^{-1}(\boldsymbol{M_c} - \boldsymbol{M}_{\boldsymbol{j_z}})^{\mathsf{T}} \\ &+ (\boldsymbol{M_j} - \boldsymbol{M}_{\boldsymbol{j_z}})\boldsymbol{V_0}^{-1}(\boldsymbol{M_j} - \boldsymbol{M}_{\boldsymbol{j_z}})^{\mathsf{T}} + \sum_{k:j_k=j} \mathbb{E}\left[(\boldsymbol{y_k} - \boldsymbol{M_j}\boldsymbol{\chi}_{\boldsymbol{i_k}})(\boldsymbol{y_k} - \boldsymbol{M_j}\boldsymbol{\chi}_{\boldsymbol{i_k}})^{\mathsf{T}}\right]\end{aligned} \tag{8.27}$$

By construction, this form shows that $\boldsymbol{W_j}$ is positive-definite. Derivation of this form is shown in Appendix B.

8.2.2. The Factor $q(Z)$

Now, for the factor $q(Z)$, we have

$$\begin{aligned}\log q^*(\boldsymbol{z_j}) &= \mathbb{E}_{X,\Phi,\Lambda}\left[\log p(Y, \Phi, \Lambda, Z, X)\right] + const \\ &= \sum_{c=1}^{C} \boldsymbol{z}_{\boldsymbol{j}c} \log \varrho_{jc} + const \\ \implies q^*(\boldsymbol{z_j}) &\propto \prod_{c=1}^{C} \varrho_{jc}^{\boldsymbol{z}_{\boldsymbol{j}c}}\end{aligned} \tag{8.28}$$

where

$$\log \varrho_{jc} = \Bigg[\log \boldsymbol{p}_c + \frac{d+1}{2}\mathbb{E}\left[\log|\boldsymbol{\Lambda_j}|\right] - \frac{d(d+1)}{2}\log(2\pi) - \frac{1}{2}\operatorname{Tr}\left(\boldsymbol{V_0}^{-1}\mathbb{E}\left[\boldsymbol{\Phi_j}^{\mathsf{T}}\boldsymbol{\Lambda_j}\boldsymbol{\Phi_j}\right]\right)$$

$$
-\frac{d}{2}\log|\boldsymbol{V_0}| + \operatorname{Tr}\left(\boldsymbol{V_0}^{-1}\mathbb{E}\left[\boldsymbol{\Phi}_{\boldsymbol{j}}^{\mathsf{T}}\boldsymbol{\Lambda_j}\right]\boldsymbol{M_c}\right) - \frac{1}{2}\operatorname{Tr}\left(\boldsymbol{V_0}^{-1}\boldsymbol{M}_{\boldsymbol{c}}^{\mathsf{T}}\mathbb{E}\left[\boldsymbol{\Lambda_j}\right]\boldsymbol{M_c}\right)\Bigg] . \tag{8.29}
$$

Then,

$$
q^*(\boldsymbol{z_j}) = \prod_{c=1}^{C} \rho_{jc}^{\boldsymbol{z}_{\boldsymbol{j}c}} \tag{8.30}
$$

where

$$
\rho_{jc} = \frac{\varrho_{jc}}{\sum\limits_{c'=1}^{C} \varrho_{jc'}} \tag{8.31}
$$

8.2.3. The Factor $q(X)$

Finally, let us consider the factor $q(X)$. First, we partition the matrix $\boldsymbol{\Phi_j}$ into two matrices $\boldsymbol{\Omega_j}$ ($d{\times}d$) and $\boldsymbol{b_j}$ ($d{\times}1$) as $\boldsymbol{\Phi_j} = \left[\boldsymbol{\Omega_j} \,\vdots\, \boldsymbol{b_j}\right]$. Now, we consider the terms with $\boldsymbol{x_i}$ and take their expectation with respect to the other factors. We have

$$
\begin{aligned}
\log q^*(\boldsymbol{x_i}) &= \mathbb{E}_{Z,\Phi,\Lambda}\left[\log p(Y,\Phi,\Lambda,Z,X)\right] + \mathit{const} \\
&= \mathbb{E}_{Z,\Phi,\Lambda}\left[\sum_{k:i_k=i}\left(\boldsymbol{\chi}_{\boldsymbol{i}}^{\mathsf{T}}\boldsymbol{\Phi}_{\boldsymbol{j_k}}^{\mathsf{T}}\boldsymbol{\Lambda_{j_k}}\boldsymbol{y_k} - \frac{1}{2}\boldsymbol{\chi}_{\boldsymbol{i}}^{\mathsf{T}}\boldsymbol{\Phi}_{\boldsymbol{j_k}}^{\mathsf{T}}\boldsymbol{\Lambda_{j_k}}\boldsymbol{\Phi_{j_k}}\boldsymbol{\chi_i}\right)\right] + \mathit{const} \\
&= \sum_{k:i_k=i}\mathbb{E}_{Z,\Phi,\Lambda}\left[\begin{bmatrix}\boldsymbol{x_i}\\ \hdashline 1\end{bmatrix}^{\mathsf{T}}\left[\boldsymbol{\Omega_{j_k}} \,\vdots\, \boldsymbol{b_{j_k}}\right]^{\mathsf{T}}\boldsymbol{\Lambda_{j_k}}\boldsymbol{y_k}\right] \\
&\quad - \frac{1}{2}\sum_{k:i_k=i}\mathbb{E}_{Z,\Phi,\Lambda}\left[\begin{bmatrix}\boldsymbol{x_i}\\ \hdashline 1\end{bmatrix}^{\mathsf{T}}\left[\boldsymbol{\Omega_{j_k}} \,\vdots\, \boldsymbol{b_{j_k}}\right]^{\mathsf{T}}\boldsymbol{\Lambda_{j_k}}\left[\boldsymbol{\Omega_{j_k}} \,\vdots\, \boldsymbol{b_{j_k}}\right]\begin{bmatrix}\boldsymbol{x_i}\\ \hdashline 1\end{bmatrix}\right] + \mathit{const} \\
&= \boldsymbol{x}_{\boldsymbol{i}}^{\mathsf{T}}\sum_{k:i_k=i}\mathbb{E}_{\Phi,\Lambda}\left[\boldsymbol{\Omega}_{\boldsymbol{j_k}}^{\mathsf{T}}\boldsymbol{\Lambda_{j_k}}\right]\boldsymbol{y_k} \\
&\quad - \frac{1}{2}\sum_{k:i_k=i}\mathbb{E}_{\Phi,\Lambda}\left[(\boldsymbol{x}_{\boldsymbol{i}}^{\mathsf{T}}\boldsymbol{\Omega}_{\boldsymbol{j_k}}^{\mathsf{T}} + \boldsymbol{b}_{\boldsymbol{j_k}}^{\mathsf{T}})\boldsymbol{\Lambda_{j_k}}(\boldsymbol{\Omega_{j_k}}\boldsymbol{x_i} + \boldsymbol{b_{j_k}})\right] + \mathit{const} \\
&= \boldsymbol{x}_{\boldsymbol{i}}^{\mathsf{T}}\sum_{k:i_k=i}\left(\mathbb{E}_{\Phi,\Lambda}\left[\boldsymbol{\Omega}_{\boldsymbol{j_k}}^{\mathsf{T}}\boldsymbol{\Lambda_{j_k}}\right]\boldsymbol{y_k} - \mathbb{E}_{\Phi,\Lambda}\left[\boldsymbol{\Omega}_{\boldsymbol{j_k}}^{\mathsf{T}}\boldsymbol{\Lambda_{j_k}}\boldsymbol{b_{j_k}}\right]\right)
\end{aligned}
$$

$$-\frac{1}{2}\boldsymbol{x_i}^{\mathsf{T}}\left(\sum_{k:i_k=i}\mathbb{E}_{\Phi,\Lambda}\left[\boldsymbol{\Omega}_{\boldsymbol{j_k}}^{\mathsf{T}}\boldsymbol{\Lambda}_{\boldsymbol{j_k}}\boldsymbol{\Omega}_{\boldsymbol{j_k}}\right]\right)\boldsymbol{x_i}{+}const \tag{8.32}$$

We observe that this equation is in the form of the multivariate normal distribution, given by

$$q^*(\boldsymbol{x_i}) = \mathcal{N}_d\left(\boldsymbol{x_i}; \boldsymbol{\mu_i}, \boldsymbol{\Sigma}_{\boldsymbol{i}}^{-1}\right) \tag{8.33}$$

where

$$\boldsymbol{\Sigma}_{\boldsymbol{i}}^{-1} = \sum_{k:i_k=i}\mathbb{E}_{\Phi,\Lambda}\left[\boldsymbol{\Omega}_{\boldsymbol{j_k}}^{\mathsf{T}}\boldsymbol{\Lambda}_{\boldsymbol{j_k}}\boldsymbol{\Omega}_{\boldsymbol{j_k}}\right] \tag{8.34}$$

$$\boldsymbol{\mu_i} = \boldsymbol{\Sigma_i}\left(\sum_{k:i_k=i}\mathbb{E}_{\Phi,\Lambda}\left[\boldsymbol{\Omega}_{\boldsymbol{j_k}}^{\mathsf{T}}\boldsymbol{\Lambda}_{\boldsymbol{j_k}}\right]\boldsymbol{y_k} - \sum_{k:i_k=i}\mathbb{E}_{\Phi,\Lambda}\left[\boldsymbol{\Omega}_{\boldsymbol{j_k}}^{\mathsf{T}}\boldsymbol{\Lambda}_{\boldsymbol{j_k}}\boldsymbol{b}_{\boldsymbol{j_k}}\right]\right) \tag{8.35}$$

8.2.4. Required Expectations for the Posterior Parameters

For calculating the posterior parameters found in Sections 8.2.1 to 8.2.3, we require the expectations of various random variables and their products. We start with $\mathbb{E}_{\boldsymbol{x_i}}\left[\boldsymbol{\chi_i}\right]$ which is required for Equation 8.23:

$$\mathbb{E}_{\boldsymbol{x_i}}\left[\boldsymbol{\chi_i}\right] = \left[\begin{array}{c}\mathbb{E}_{\boldsymbol{x_i}}\left[\boldsymbol{x_i}\right] \\ \hline 1\end{array}\right] = \left[\begin{array}{c}\boldsymbol{\mu_i} \\ \hline 1\end{array}\right] \tag{8.36}$$

since $\mathbb{E}_{\boldsymbol{x_i}}\left[\boldsymbol{x_i}\right] = \boldsymbol{\mu_i}$.

For Equation 8.22, we need $\mathbb{E}_{\boldsymbol{x_i}}\left[\boldsymbol{\chi_i}\boldsymbol{\chi}_{\boldsymbol{i}}^{\mathsf{T}}\right]$:

$$\mathbb{E}_{\boldsymbol{x_i}}\left[\boldsymbol{\chi_i}\boldsymbol{\chi}_{\boldsymbol{i}}^{\mathsf{T}}\right] = \left[\begin{array}{c|c}\mathbb{E}_{\boldsymbol{x_i}}\left[\boldsymbol{x_i}\boldsymbol{x}_{\boldsymbol{i}}^{\mathsf{T}}\right] & \mathbb{E}_{\boldsymbol{x_i}}\left[\boldsymbol{x_i}\right] \\ \hline \mathbb{E}_{\boldsymbol{x_i}}\left[\boldsymbol{x_i}\right]^{\mathsf{T}} & 1\end{array}\right] = \left[\begin{array}{c|c}\boldsymbol{\Sigma_i} + \boldsymbol{\mu_i}\boldsymbol{\mu}_{\boldsymbol{i}}^{\mathsf{T}} & \boldsymbol{\mu_i} \\ \hline \boldsymbol{\mu}_{\boldsymbol{i}}^{\mathsf{T}} & 1\end{array}\right] \tag{8.37}$$

since $\mathbb{E}_{\boldsymbol{x_i}}\left[\boldsymbol{x_i}\right] = \boldsymbol{\mu_i}$ and $\mathbb{E}_{\boldsymbol{x_i}}\left[\boldsymbol{x_i}\boldsymbol{x}_{\boldsymbol{i}}^{\mathsf{T}}\right] = \boldsymbol{\Sigma_i} + \boldsymbol{\mu_i}\boldsymbol{\mu}_{\boldsymbol{i}}^{\mathsf{T}}$.

For Equation 8.29, we need $\mathbb{E}_{\boldsymbol{\Phi}_j,\boldsymbol{\Lambda}_j}\left[\boldsymbol{\Phi}_j^\top\boldsymbol{\Lambda}_j\right]$:

$$\mathbb{E}_{\boldsymbol{\Phi}_j,\boldsymbol{\Lambda}_j}\left[\boldsymbol{\Phi}_j^\top\boldsymbol{\Lambda}_j\right] = \mathbb{E}_{\boldsymbol{\Lambda}_j}\left[\mathbb{E}_{\boldsymbol{\Phi}_j|\boldsymbol{\Lambda}_j}\left[\boldsymbol{\Phi}_j\right]^\top\boldsymbol{\Lambda}_j\right] = (n_0+N_j)\boldsymbol{M}_j^\top\boldsymbol{W}_j \tag{8.38}$$

For Equation 8.35, we need $\mathbb{E}_{\boldsymbol{\Phi}_j,\boldsymbol{\Lambda}_j}\left[\boldsymbol{\Omega}_j^\top\boldsymbol{\Lambda}_j\right]$:

$$\mathbb{E}_{\boldsymbol{\Phi}_j,\boldsymbol{\Lambda}_j}\left[\boldsymbol{\Phi}_j^\top\boldsymbol{\Lambda}_j\right] = \left[\begin{array}{c}\mathbb{E}_{\boldsymbol{\Phi}_j,\boldsymbol{\Lambda}_j}\left[\boldsymbol{\Omega}_j^\top\boldsymbol{\Lambda}_j\right]\\ \hdashline \mathbb{E}_{\boldsymbol{\Phi}_j,\boldsymbol{\Lambda}_j}\left[\boldsymbol{b}_j^\top\boldsymbol{\Lambda}_j\right]\end{array}\right] \tag{8.39}$$

$$\mathbb{E}_{\boldsymbol{\Phi}_j,\boldsymbol{\Lambda}_j}\left[\boldsymbol{\Omega}_j^\top\boldsymbol{\Lambda}_j\right] = (n_0+N_j)\left[\begin{array}{c:c}\boldsymbol{I_d} & \boldsymbol{O_{d1}}\end{array}\right]\boldsymbol{M}_j^\top\boldsymbol{W}_j \tag{8.40}$$

For Equation 8.29, we need $\mathbb{E}_{\boldsymbol{\Lambda}_j}\left[\boldsymbol{\Lambda}_j\right]$, $\mathbb{E}_{\boldsymbol{\Lambda}_j}\left[\log|\boldsymbol{\Lambda}_j|\right]$, and $\mathbb{E}_{\boldsymbol{\Phi}_j,\boldsymbol{\Lambda}_j}\left[\boldsymbol{\Phi}_j^\top\boldsymbol{\Lambda}_j\boldsymbol{\Phi}_j\right]$:

$$\mathbb{E}_{\boldsymbol{\Lambda}_j}\left[\boldsymbol{\Lambda}_j\right] = (n_0+N_j)\boldsymbol{W}_j \tag{8.41}$$

$$\mathbb{E}_{\boldsymbol{\Lambda}_j}\left[\log|\boldsymbol{\Lambda}_j|\right] = \boldsymbol{\psi}_d\left(\frac{n_0+N_j}{2}\right) + d\log 2 + \log|\boldsymbol{V}_j| \tag{8.42}$$

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\Phi}_j,\boldsymbol{\Lambda}_j}\left[\boldsymbol{\Phi}_j^\top\boldsymbol{\Lambda}_j\boldsymbol{\Phi}_j\right] &= \mathbb{E}_{\boldsymbol{\Lambda}_j}\left[\mathbb{E}_{\boldsymbol{\Phi}_j|\boldsymbol{\Lambda}_j}\left[\boldsymbol{\Phi}_j^\top\boldsymbol{\Lambda}_j\boldsymbol{\Phi}_j\right]\right]\\ &= \mathbb{E}_{\boldsymbol{\Lambda}_j}\left[d\boldsymbol{V}_j + \boldsymbol{M}_j^\top\boldsymbol{\Lambda}_j\boldsymbol{M}_j\right]\\ &= d\boldsymbol{V}_j + (n_0+N_j)\boldsymbol{M}_j^\top\boldsymbol{W}_j\boldsymbol{M}_j\end{aligned} \tag{8.43}$$

For Equations 8.34 and 8.35, we need $\mathbb{E}_{\boldsymbol{\Phi}_j,\boldsymbol{\Lambda}_j}\left[\boldsymbol{\Omega}_j^\top\boldsymbol{\Lambda}_j\boldsymbol{\Omega}_j\right]$ and $\mathbb{E}_{\boldsymbol{\Phi}_j,\boldsymbol{\Lambda}_j}\left[\boldsymbol{\Omega}_j^\top\boldsymbol{\Lambda}_j\boldsymbol{b}_j\right]$, respectively. Since $\boldsymbol{\Phi}_j = \left[\begin{array}{c:c}\boldsymbol{\Omega}_j & \boldsymbol{b}_j\end{array}\right]$, we have

$$\begin{aligned}\boldsymbol{\Phi}_j^\top\boldsymbol{\Lambda}_j\boldsymbol{\Phi}_j &= \left[\begin{array}{c}\boldsymbol{\Omega}_j^\top\\ \hdashline \boldsymbol{b}_j^\top\end{array}\right]\boldsymbol{\Lambda}_j\left[\begin{array}{c:c}\boldsymbol{\Omega}_j & \boldsymbol{b}_j\end{array}\right]\\ &= \left[\begin{array}{c:c}\boldsymbol{\Omega}_j^\top\boldsymbol{\Lambda}_j\boldsymbol{\Omega}_j & \boldsymbol{\Omega}_j^\top\boldsymbol{\Lambda}_j\boldsymbol{b}_j\\ \hdashline \boldsymbol{b}_j^\top\boldsymbol{\Lambda}_j\boldsymbol{\Omega}_j & \boldsymbol{b}_j^\top\boldsymbol{\Lambda}_j\boldsymbol{b}_j\end{array}\right]\end{aligned} \tag{8.44}$$

which gives us

$$\left[\begin{array}{c|c} \mathbb{E}_{\boldsymbol{\Phi}_j,\boldsymbol{\Lambda}_j}\left[\boldsymbol{\Omega}_{\boldsymbol{j}}^{\top}\boldsymbol{\Lambda}_{\boldsymbol{j}}\boldsymbol{\Omega}_{\boldsymbol{j}}\right] & \mathbb{E}_{\boldsymbol{\Phi}_j,\boldsymbol{\Lambda}_j}\left[\boldsymbol{\Omega}_{\boldsymbol{j}}^{\top}\boldsymbol{\Lambda}_{\boldsymbol{j}}\boldsymbol{b}_{\boldsymbol{j}}\right] \\ \hline \mathbb{E}_{\boldsymbol{\Phi}_j,\boldsymbol{\Lambda}_j}\left[\boldsymbol{b}_{\boldsymbol{j}}^{\top}\boldsymbol{\Lambda}_{\boldsymbol{j}}\boldsymbol{\Omega}_{\boldsymbol{j}}\right] & \mathbb{E}_{\boldsymbol{\Phi}_j,\boldsymbol{\Lambda}_j}\left[\boldsymbol{b}_{\boldsymbol{j}}^{\top}\boldsymbol{\Lambda}_{\boldsymbol{j}}\boldsymbol{b}_{\boldsymbol{j}}\right] \end{array}\right] = d\boldsymbol{V}_{\boldsymbol{j}} + (n_0+N_j)\boldsymbol{M}_{\boldsymbol{j}}^{\top}\boldsymbol{W}_{\boldsymbol{j}}\boldsymbol{M}_{\boldsymbol{j}} \tag{8.45}$$

Then, we have

$$\mathbb{E}_{\boldsymbol{\Phi}_j,\boldsymbol{\Lambda}_j}\left[\boldsymbol{\Omega}_{\boldsymbol{j}}^{\top}\boldsymbol{\Lambda}_{\boldsymbol{j}}\boldsymbol{\Omega}_{\boldsymbol{j}}\right] = \left[\begin{array}{c|c}\boldsymbol{I}_{\boldsymbol{d}} & \boldsymbol{O}_{\boldsymbol{d1}}\end{array}\right]\left(d\boldsymbol{V}_{\boldsymbol{j}} + (n_0+N_j)\boldsymbol{M}_{\boldsymbol{j}}^{\top}\boldsymbol{W}_{\boldsymbol{j}}\boldsymbol{M}_{\boldsymbol{j}}\right)\left[\begin{array}{c}\boldsymbol{I}_{\boldsymbol{d}} \\ \hline \boldsymbol{O}_{\boldsymbol{1d}}\end{array}\right] \tag{8.46}$$

$$\mathbb{E}_{\boldsymbol{\Phi}_j,\boldsymbol{\Lambda}_j}\left[\boldsymbol{\Omega}_{\boldsymbol{j}}^{\top}\boldsymbol{\Lambda}_{\boldsymbol{j}}\boldsymbol{b}_{\boldsymbol{j}}\right] = \left[\begin{array}{c|c}\boldsymbol{I}_{\boldsymbol{d}} & \boldsymbol{O}_{\boldsymbol{d1}}\end{array}\right]\left(d\boldsymbol{V}_{\boldsymbol{j}} + (n_0+N_j)\boldsymbol{M}_{\boldsymbol{j}}^{\top}\boldsymbol{W}_{\boldsymbol{j}}\boldsymbol{M}_{\boldsymbol{j}}\right)\left[\begin{array}{c}\boldsymbol{O}_{\boldsymbol{d1}} \\ \hline 1\end{array}\right] \tag{8.47}$$

For Equations 8.23 and 8.26, we need $\mathbb{E}_{\boldsymbol{z}_j}\left[\boldsymbol{z}_{\boldsymbol{j}}\right]$ which is

$$\mathbb{E}_{\boldsymbol{z}_j}\left[\boldsymbol{z}_{\boldsymbol{j}}\right] = \rho_{jc} \tag{8.48}$$

8.2.5. The Update Equations

By plugging the expectations into the posterior parameters, we end up with the update equations below:

$$\boldsymbol{V}_{\boldsymbol{j}}^{-1} = \boldsymbol{V}_{\boldsymbol{0}}^{-1} + \sum_{k:j_k=j}\left[\begin{array}{c|c}\boldsymbol{\Sigma}_{\boldsymbol{i_k}} + \boldsymbol{\mu}_{\boldsymbol{i_k}}\boldsymbol{\mu}_{\boldsymbol{i_k}}^{\top} & \boldsymbol{\mu}_{\boldsymbol{i_k}} \\ \hline \boldsymbol{\mu}_{\boldsymbol{i_k}}^{\top} & 1\end{array}\right] \tag{8.49}$$

$$\boldsymbol{M}_{\boldsymbol{j}} = \left(\sum_{c=1}^{C}\rho_{jc}\boldsymbol{M}_{\boldsymbol{c}}\boldsymbol{V}_{\boldsymbol{0}}^{-1} + \sum_{k:j_k=j}\boldsymbol{y}_{\boldsymbol{k}}\left[\begin{array}{c|c}\boldsymbol{\mu}_{\boldsymbol{i_k}}^{\top} & 1\end{array}\right]\right)\boldsymbol{V}_{\boldsymbol{j}} \tag{8.50}$$

$$\boldsymbol{W}_{\boldsymbol{j}}^{-1} = \boldsymbol{W}_{\boldsymbol{0}}^{-1} + \sum_{k:j_k=j}\boldsymbol{y}_{\boldsymbol{k}}\boldsymbol{y}_{\boldsymbol{k}}^{\top} + \sum_{c=1}^{C}\rho_{jc}\left(\boldsymbol{M}_{\boldsymbol{c}}\boldsymbol{V}_{\boldsymbol{0}}^{-1}\boldsymbol{M}_{\boldsymbol{c}}^{\top}\right) - \boldsymbol{M}_{\boldsymbol{j}}\boldsymbol{V}_{\boldsymbol{j}}^{-1}\boldsymbol{M}_{\boldsymbol{j}}^{\top} \tag{8.51}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{i}}^{-1} = \left[\begin{array}{c|c}\boldsymbol{I}_{\boldsymbol{d}} & \boldsymbol{O}_{\boldsymbol{d1}}\end{array}\right]\sum_{k:i_k=i}\left(d\boldsymbol{V}_{\boldsymbol{j_k}} + (n_0+N_{j_k})\boldsymbol{M}_{\boldsymbol{j_k}}^{\top}\boldsymbol{W}_{\boldsymbol{j_k}}\boldsymbol{M}_{\boldsymbol{j_k}}\right)\left[\begin{array}{c}\boldsymbol{I}_{\boldsymbol{d}} \\ \hline \boldsymbol{O}_{\boldsymbol{1d}}\end{array}\right] \tag{8.52}$$

$$\boldsymbol{\mu}_{\boldsymbol{i}} = \boldsymbol{\Sigma}_{\boldsymbol{i}}\left[\begin{array}{c|c}\boldsymbol{I}_{\boldsymbol{d}} & \boldsymbol{O}_{\boldsymbol{d1}}\end{array}\right]\sum_{k:i_k=i}\left((n_0+N_{j_k})\boldsymbol{M}_{\boldsymbol{j_k}}^{\top}\boldsymbol{W}_{\boldsymbol{j_k}}\left(\boldsymbol{y}_{\boldsymbol{k}} - \boldsymbol{M}_{\boldsymbol{j_k}}\left[\begin{array}{c}\boldsymbol{O}_{\boldsymbol{d1}} \\ \hline 1\end{array}\right]\right) - d\boldsymbol{V}_{\boldsymbol{j_k}}\left[\begin{array}{c}\boldsymbol{O}_{\boldsymbol{d1}} \\ \hline 1\end{array}\right]\right) \tag{8.53}$$

$$\begin{aligned}\varrho_{jc} = \exp\Bigg(& \log \boldsymbol{p}_c + \frac{d+1}{2}\boldsymbol{\psi}_d\left(\frac{n_0+N_j}{2}\right) + \frac{d+1}{2}\log|\boldsymbol{V_j}| - \frac{d(d+1)}{2}\log(\pi) \\ & -\frac{d}{2}\operatorname{Tr}\left(\boldsymbol{V_0}^{-1}\boldsymbol{V_j}\right) - \frac{(n_0+N_j)}{2}\operatorname{Tr}\left(\boldsymbol{V_0}^{-1}\boldsymbol{M_j}^{\mathsf{T}}\boldsymbol{W_j}\boldsymbol{M_j}\right) - \frac{d}{2}\log|\boldsymbol{V_0}| \\ & +(n_0+N_j)\operatorname{Tr}\left(\boldsymbol{V_0}^{-1}\boldsymbol{M_j}^{\mathsf{T}}\boldsymbol{W_j}\boldsymbol{M_c}\right) - \frac{(n_0+N_j)}{2}\operatorname{Tr}\left(\boldsymbol{V_0}^{-1}\boldsymbol{M_c}^{\mathsf{T}}\boldsymbol{W_j}\boldsymbol{M_c}\right)\Bigg)\end{aligned} \tag{8.54}$$

8.2.6. Lower Bound

The lower bound of the model can be calculated as follows:

$$\mathcal{L}(q) = \sum_Z \iiint q(\Phi,\Lambda,Z,X)\log\frac{p(Y,\Phi,\Lambda,Z,X)}{q(\Phi,\Lambda,Z,X)}d\Phi d\Lambda dX \tag{8.55}$$

$$\begin{aligned} &= \mathbb{E}_{\Phi,\Lambda,X,Z}\left[\log\frac{p(Y,\Phi,\Lambda,X,Z)}{q(\Phi,\Lambda,X,Z)}\right] \\ &= \mathbb{E}\left[\log p(Y,\Phi,\Lambda,X,Z)\right] - \mathbb{E}\left[\log q(\Phi,\Lambda,X,Z)\right] \\ &= \mathbb{E}\left[\log p(Y|\Phi,\Lambda,X)\right] + \mathbb{E}\left[\log p(\Phi|\Lambda,X,Z)\right] + \mathbb{E}\left[\log p(\Lambda)\right] + \mathbb{E}\left[\log p(Z)\right] \\ &\quad + \mathbb{E}\left[\log p(X)\right] - \mathbb{E}\left[\log q(\Phi|\Lambda,X,Z)\right] - \mathbb{E}\left[\log q(\Lambda)\right] - \mathbb{E}\left[\log q(Z)\right] - \mathbb{E}\left[\log q(X)\right] \end{aligned} \tag{8.56}$$

$$\begin{aligned} = \; & -\frac{1}{2}\sum_{j=1}^{R}\operatorname{Tr}\Bigg(\Bigg[\sum_{k:j_k=j}\boldsymbol{y_k}\boldsymbol{y_k}^{\mathsf{T}} + \sum_{c=1}^{C}\mathbb{E}\left[\boldsymbol{z_j}_c\right]\boldsymbol{M_c}\boldsymbol{V_0}^{-1}\boldsymbol{M_c}^{\mathsf{T}} + \boldsymbol{W_0}^{-1} \\ & \qquad\qquad - \boldsymbol{M_j}\boldsymbol{V_j}^{-1}\boldsymbol{M_j}^{\mathsf{T}} - \boldsymbol{W_j}^{-1}\Bigg]\mathbb{E}\left[\boldsymbol{\Lambda_j}\right]\Bigg) \\ & + \sum_{j=1}^{R}\operatorname{Tr}\left(\left[\sum_{k:j_k=j}\mathbb{E}\left[\boldsymbol{\chi_{i_k}}\right]\boldsymbol{y_k}^{\mathsf{T}} + \boldsymbol{V_0}^{-1}\sum_{c=1}^{C}\mathbb{E}\left[\boldsymbol{z_j}_c\right]\boldsymbol{M_c} - \boldsymbol{V_j}^{-1}\boldsymbol{M_j}^{\mathsf{T}}\right]\mathbb{E}\left[\boldsymbol{\Lambda_j}\boldsymbol{\Phi_j}\right]\right) \\ & - \frac{1}{2}\sum_{j=1}^{R}\operatorname{Tr}\left(\left[\sum_{k:j_k=j}\mathbb{E}\left[\boldsymbol{\chi_{i_k}}\boldsymbol{\chi_{i_k}}^{\mathsf{T}}\right] + \boldsymbol{V_0}^{-1} - \boldsymbol{V_j}^{-1}\right]\mathbb{E}\left[\boldsymbol{\Phi_j}^{\mathsf{T}}\boldsymbol{\Lambda_j}\boldsymbol{\Phi_j}\right]\right) \\ & + \sum_{j=1}^{R}\left(\frac{d}{2}\log|\boldsymbol{V_j}| + \frac{(n_0+N_j)d}{2}\log 2 + \frac{n_0+N_j}{2}\log|\boldsymbol{W_j}| + \log\boldsymbol{\Gamma}_d\left(\frac{n_0+N_j}{2}\right)\right) \\ & - \frac{(K+R(d+1))d}{2}\log(2\pi) - \frac{R}{2}\left(d\log|\boldsymbol{V_0}| + n_0\log|\boldsymbol{W_0}| + n_0 d\log 2\right) \\ & - R\log\boldsymbol{\Gamma}_d\left(\frac{n_0}{2}\right) + \frac{Nd}{2}(\log(2\pi)+1) + \frac{1}{2}\sum_{i=1}^{N}\log|\boldsymbol{\Sigma_i}| \end{aligned}$$

$$+\sum_{j=1}^{R}\sum_{c=1}^{C}\mathbb{E}\left[\boldsymbol{z_j}_c\right]\left(\log \boldsymbol{p}_c - \log \rho_{jc}\right) \tag{8.57}$$

The calculations of the expectations appearing in Equation 8.57 are shown in Section 8.2.4.

8.3. Preliminary Experiments

In this section, we present some preliminary experiments on the Head Pose Annotations dataset using the proposed multivariate model. Recall that our model needs a set of hyperparameters to work. We incorporate adverseness behavior into the proposed multivariate model through $\boldsymbol{M_c}$ matrices. To this end, we prepare all possible (*i.e.* $C = 2^d$ many) $\boldsymbol{M_c}$ matrices of the form:

$$\boldsymbol{M_c} = \begin{bmatrix} \diagdown & & 0 & 0 \\ & \boldsymbol{m}_c & & \vdots \\ 0 & & \diagdown & 0 \end{bmatrix} \tag{8.58}$$

where $\boldsymbol{m}_c \in \{-1, 1\}^d$. Each element of $\boldsymbol{m}_c$ encodes the presence or absence of an annotator's adverseness for different attributes. By construction, $\boldsymbol{z_j}_c = 1$ if and only if annotator j belongs to category c. The choice of $\boldsymbol{M_c}$ regarding any annotator j is governed by the random variable $\boldsymbol{z_j}$, which depends on the hyperparameter $\boldsymbol{p}$. We assume that we have no prior knowledge about the annotator behaviors occurring in the dataset and therefore, we choose a flat prior over $\boldsymbol{p}$. The remaining hyperparameters are chosen as $\boldsymbol{V_0} = 10^{-4}\boldsymbol{I_{d+1}}$, $\boldsymbol{W_0} = 10^{4}\boldsymbol{I_d}$, and $n_0 = 2$ for encouraging $|\boldsymbol{\Lambda_j}|$ to be large and assisting $\boldsymbol{\Phi_j}$ to somewhat resemble its mean $\boldsymbol{M_c}$.

8.3.1. Observations on the Model Error

In Figure 8.2, we present the change in the lower bound value $L(q)$ and the mean absolute errors for the *pan* and *tilt* attributes during the model fitting process. The lower bound monotonically increases (*i.e.* it is non-decreasing) with each iteration step as expected, which is equivalent to decreasing the KL-divergence between our model and the actual full joint distribution of the problem. We observe that the increases in the lower bound values are accompanied with slight decreases in the mean absolute errors in both attributes. This means that the model's success at describing the problem is reflected by lowered errors, even though the model is unaware of any ground truth values.

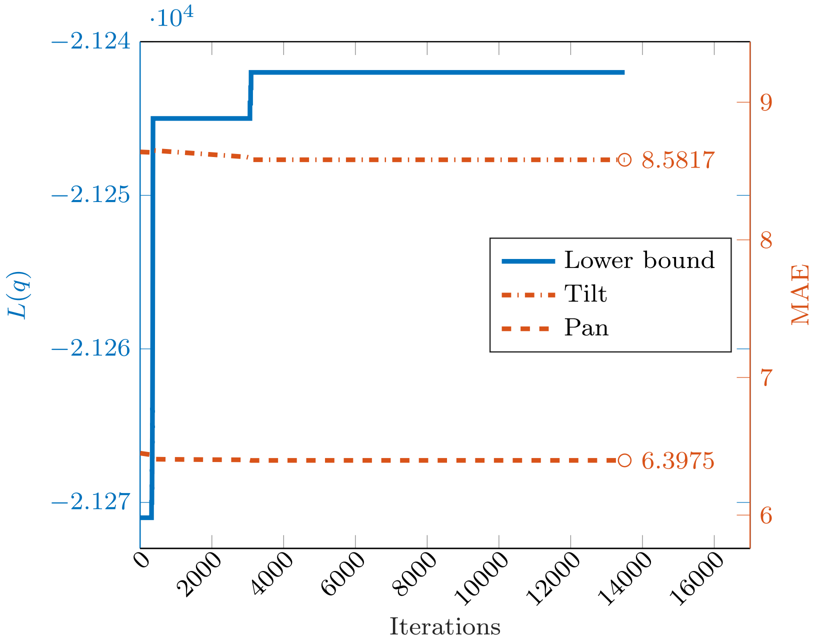


Figure 8.2. Change of the lower bound value ($L(q)$) and attribute error while fitting the model

In Figure 8.3, we present the cumulative match error curves for the proposed multivariate model and compare them with the mean model and M-CBS. Recall that both mean model and M-CBS are univariate methods. The results show that our multivariate method outperforms the mean model significantly in the *pan* attribute and it is marginally better in the *tilt* attribute. In the *pan* attribute, the multivariate model performs slightly better than M-CBS for errors less than 20 degrees. We also investigate the combined error by using vectoral distance (L_2-norm) and deduce similar conclusions. Considering that the multivariate model is in its preliminary phase, the results show potential for improvement.

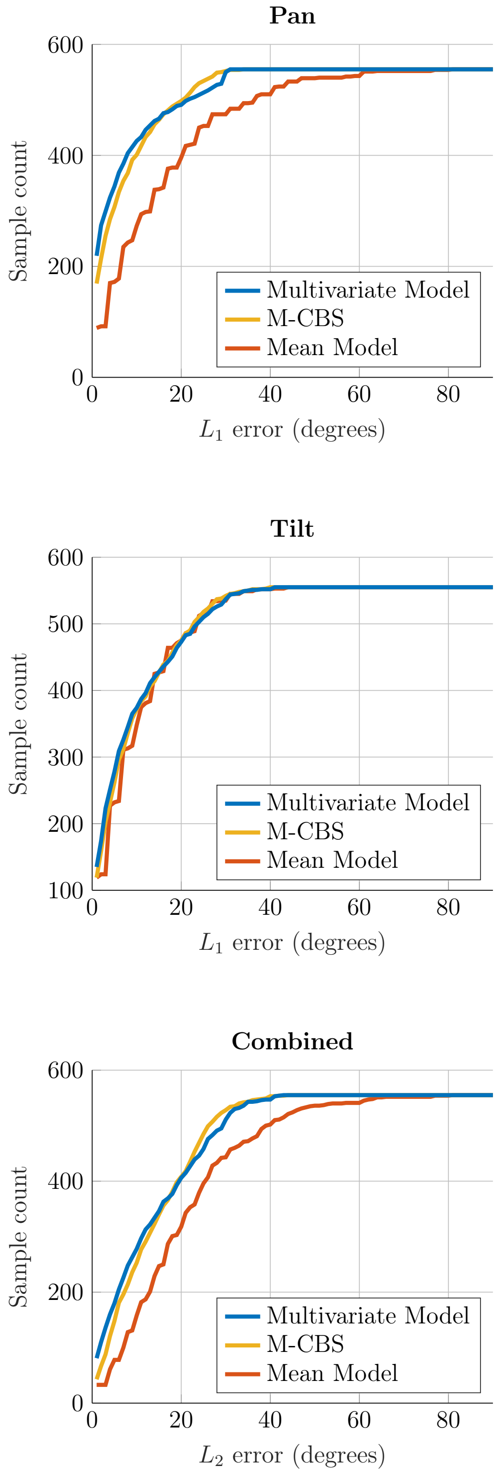


Figure 8.3. Cumulative match curves for the proposed multivariate model compared with two univariate models, namely mean model and M-CBS. Combined error is the Euclidean distance (L_2-norm) of a sample's 2-dimensional ground truth and its inferred consensus tuple.

8.3.2. Relation of the Sample Error and the Posterior Sample Variance

In Figure 8.4, we present the relation of the sample error and the posterior sample variance. We observe that low sample consensus error is often associated with a low determinant value of the posterior sample variance, indicated by the bright intensity area on the lower left corner of the figure. The slightly brighter patch located around 30 degrees of error with a low determinant is caused when multiple annotators mistakenly annotate the sample with the neighboring rate of the ground truth. Since each increment in our rating system corresponds to an increment of 30 degrees, annotator mistakes in neighboring rates translate into 30 degrees of error.

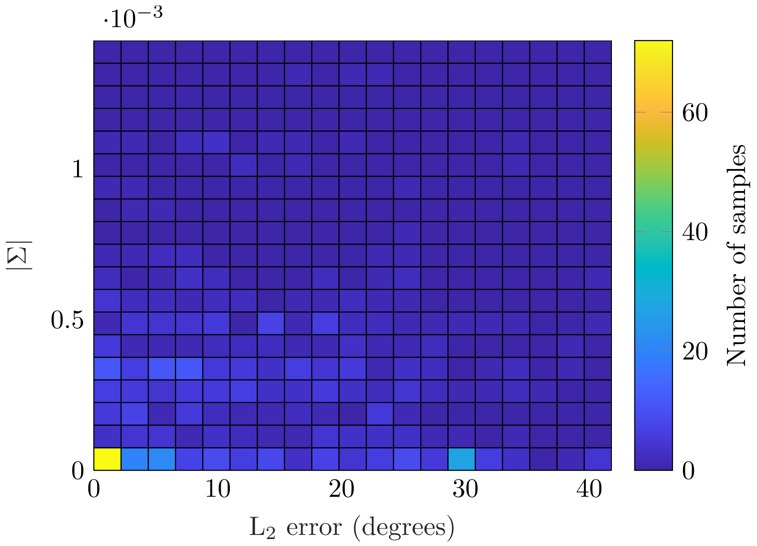


Figure 8.4. Heat map depicting the relation of the sample error and the posterior sample variance.

8.3.3. Observations on Annotators

In Figure 8.5, we give examples of competent, spammer, and adversary annotators and their annotations that we encounter in the Head Pose Annotations dataset. The variational distribution helps us to investigate the behaviors of the annotators. The most obvious observation would be the adverseness categories of the annotators, which we can find by $\mathbb{E}\left[\boldsymbol{z_j}\right]$. When we were collecting the Head Pose Annotations dataset on the CrowdFlower platform, we expected that some people would confuse left/right and give inverted scores for the *pan* attribute. We see that out of 189 annotators

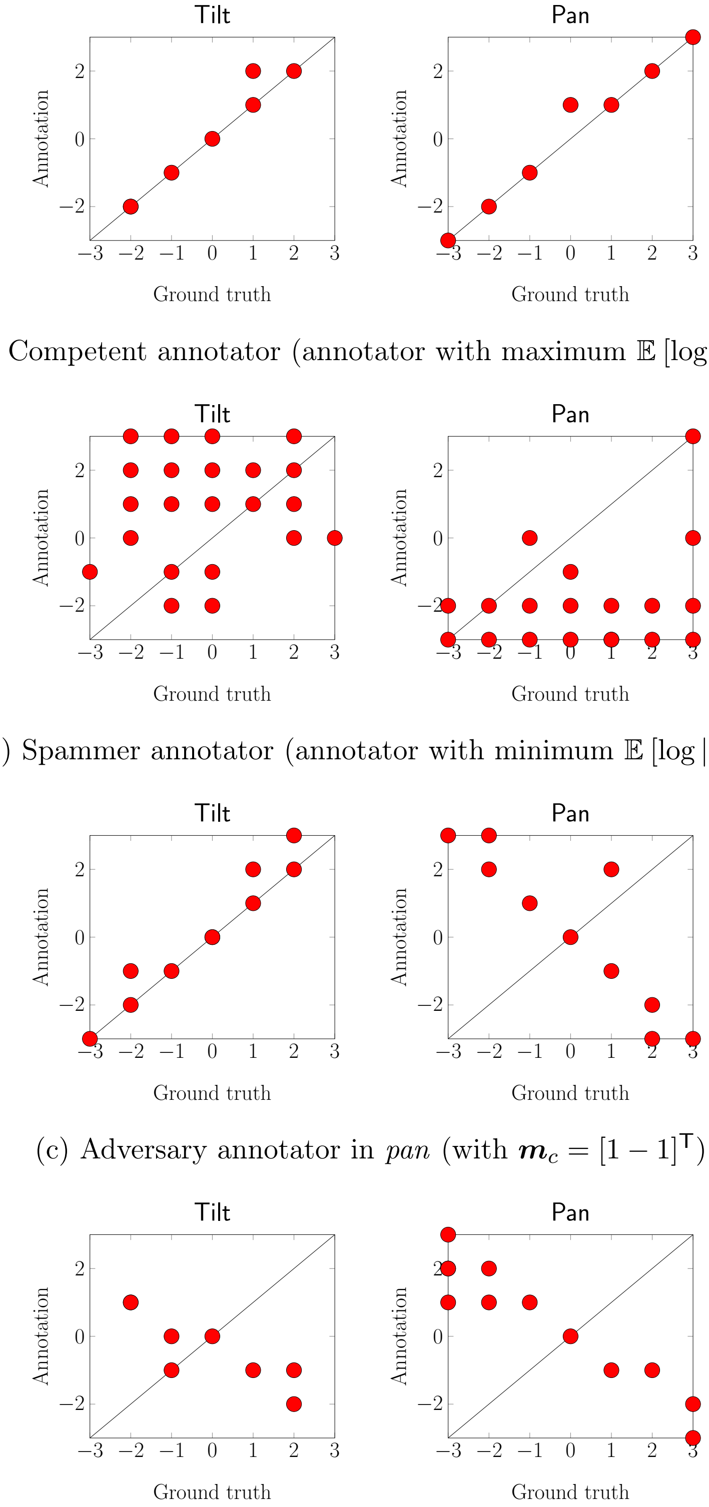
- Only one annotator is adverse in both *tilt* and *pan* attributes (see Figure 8.5d),
- 30 annotators are adverse in only *pan* attribute (see Figure 8.5c for an example),
- There are no annotators adverse in only *tilt* attribute.

Higher determinant value of an annotator's precision matrix means that the annotator is more precise with their annotations. Using these determinant values, we rank the annotators in the competent-spammer scale. Since we do not have the actual precision matrices, but their distributions, we use the expected values of their log-determinants ($\mathbb{E}\left[\log|\Lambda_j|\right]$). We observe that

$$\max_j \mathbb{E}\left[\log|\Lambda_j|\right] \approx 23.2652$$

$$\min_j \mathbb{E}\left[\log|\Lambda_j|\right] \approx -2.1495$$

In Figures 8.5a and 8.5b, we present the annotations of the annotators having the largest and smallest $\mathbb{E}\left[\log|\Lambda_j|\right]$ values, respectively. We see that using $\mathbb{E}\left[\log|\Lambda_j|\right]$ for ranking annotators in the competent-spammer scale is an appealing idea for investigating further.



(a) Competent annotator (annotator with maximum $\mathbb{E}\left[\log |\Lambda_j|\right]$)

(b) Spammer annotator (annotator with minimum $\mathbb{E}\left[\log |\Lambda_j|\right]$)

(c) Adversary annotator in *pan* (with $\boldsymbol{m}_c = [1 - 1]^\mathsf{T}$)

(d) Adversary annotator in both *tilt* and *pan* (with $\boldsymbol{m}_c = [-1 - 1]^\mathsf{T}$)

Figure 8.5. Examples of competent, spammer, and adversary annotators and their annotations. These annotators are easily revealed by investigating the posterior distribution parameters in detail.

9. CONCLUSIONS

The process of collecting annotations from crowds and using them for estimating consensus values is called *crowd-labeling*. In this context, every sample is annotated by a small subset of available annotators, where each person annotates a small subset of the dataset. The general aim of crowd-labeling is to make use of the resulting sparse set of annotations for inferring consensuses on sample labels, where the ground truth labels are unavailable and too costly to obtain. In this thesis, we tackle the problem of continuous-valued consensus estimation in crowd-labeling.

In the first part of this thesis, we introduce the Age Annotations dataset and the Head Pose Annotations dataset with *tilt* and *pan* attributes. Then, we propose four Bayesian models for obtaining consensus in continuous-valued crowd-labeling tasks by taking annotator behaviors into account. We also introduce a novel metric for measuring annotator quality. In addition, we adapt our methods to work with binary labeled data and reported their performance.

We observe various annotator behaviors and successfully compensate 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 show 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 observe that the consensus scores obtained with the proposed models lead to lower regression errors in comparison to the widely used methods.

We make several important observations in the course of this thesis. 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.

In the second part of this thesis, we introduce two active crowd-labeling algorithms for the crowdsourced labeling process, namely O-CBS and O-CBS+. We base our methods on selecting the most beneficial annotation by determining annotator and sample consensus qualities. In addition to a novel sample consensus quality score, we also introduce a family of competence scoring functions designed to prevent annotator domination. Both O-CBS and O-CBS+ are capable of utilizing a wide range of sample consensus quality and annotator competence scoring functions, inclusive of the two novel approaches that we introduce.

We investigate the effect of the dominance suppression factor and annotator exploration/exploitation trade-off over nine different real-world datasets. A thorough investigation of the dominance suppression factor in the annotator competence scoring function reveals that preventing annotator domination is of utmost importance in assessing the annotator quality correctly. The results also indicate that the timely exploration of new annotators is crucial for high quality consensus estimation. Additionally, we reduce the computational cost of the consensus estimation phase in the active crowd-labeling process, which constitutes a significant portion of the total CPU time.

We test O-CBS+ on the Age Annotations dataset, the Head Pose Annotations datasets, and the publicly available Affective Text Analysis datasets. Our method measures up to and surpasses the literature standards by using as few as one fifth of the annotations (*i.e.* ~80% cost reduction). We also investigate a sample score related stopping criterion so that the active crowd-labeling process is terminated automatically when the sample consensuses attain an acceptable quality.

In some annotation problems, annotators are asked to annotate multiple attributes for a single sample. This is the case for the Head Pose Annotations and the Affective Text Analysis datasets, which have two and six attributes, respectively.

In the earlier parts of this thesis, we handle the annotations of each attribute as separate and independent datasets. However, it could be beneficial to use those attributes together for understanding the behavior of the annotator better. For this purpose, we introduce a multivariate model and present a solution based on the variational Bayes approach. We conduct preliminary experiments, measure the model error, and investigate different annotator behaviors.

This thesis also sheds light on several open issues arising in the domain of crowd-labeling. Preliminary experiments on the proposed multivariate model show promising results and we believe that the multivariate model holds significant potential to be improved. In addition, an active-labeling approach that uses multivariate annotations should be investigated. Relaxing the homogeneous sample difficulty assumption by incorporating a heterogeneous sample difficulty parameter is another interesting future direction. Additionally, it may be worthwhile to investigate the effects of different sample consensus quality and annotator competence scoring functions on the active crowd-labeling system. Furthermore, addressing the issue of annotator competence fluctuation over time and distributing the tasks according to the recent performance of the annotators is also left to be explored in the future.

REFERENCES

1. Galton, F., “Vox populi”, *Nature*, 1907.

2. Amazon.com, Inc., *Amazon Mechanical Turk*, 2018, `https://www.mturk.com`, accessed at May 2018.

3. Figure Eight, Inc., *CrowdFlower*, 2018, `http://www.crowdflower.com`, accessed at May 2018.

4. Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, “ImageNet: A Large-Scale Hierarchical Image Database”, *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

5. Krasin, I., T. Duerig, N. Alldrin, V. Ferrari, S. Abu-El-Haija, A. Kuznetsova, H. Rom, J. Uijlings, S. Popov, S. Kamali, M. Malloci, J. Pont-Tuset, A. Veit, S. Belongie, V. Gomes, A. Gupta, C. Sun, G. Chechik, D. Cai, Z. Feng, D. Narayanan and K. Murphy, “OpenImages: A public dataset for large-scale multi-label and multi-class image classification.”, *Dataset available from https://storage.googleapis.com/openimages/web/index.html*, 2017.

6. Frenay, B. and M. Verleysen, “Classification in the Presence of Label Noise: A Survey”, *Neural Networks and Learning Systems, IEEE Transactions on*, Vol. 25, No. 5, pp. 845–869, May 2014.

7. Raykar, V. C., S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni and L. Moy, “Learning from crowds”, *Journal of Machine Learning Research*, Vol. 99, pp. 1297–1322, 2010.

8. Zhang, P. and Z. Obradovic, “Integration of Multiple Annotators by Aggregating Experts and Filtering Novices”, *2012 IEEE International Conference on Bioinformatics and Biomedicine*, pp. 1–6, Oct. 2012.

9. Kajino, H., Y. Tsuboi, I. Sato and H. Kashima, “Learning from Crowds and Experts”, *Proceedings of the 4th Human Computation Workshop (HCOMP)*, pp. 107–113, 2012.

10. Carpenter, B., “Multilevel Bayesian Models of Categorical Data Annotation”, *Unpublished manuscript*, 2008.

11. Rodrigues, F., F. Pereira and B. Ribeiro, “Learning from Multiple Annotators: Distinguishing Good from Random Labelers”, *Pattern Recognition Letters*, Vol. 34, No. 12, pp. 1428–1436, 2013.

12. Welinder, P., S. Branson, S. Belongie and P. Perona, “The Multidimensional Wisdom of Crowds”, *Advances in Neural Information Processing Systems*, Vol. 23, pp. 2424–2432, 2010.

13. Yan, Y., R. Rosales, G. Fung, M. Schmidt, G. Hermosillo, L. Bogoni, L. Moy and J. Dy, “Modeling Annotator Expertise: Learning When Everybody Knows a Bit of Something”, *International Conference on Artificial Intelligence and Statistics*, Vol. 9, pp. 932–939, 2010.

14. Raykar, V. and S. Yu, “Eliminating Spammers and Ranking Annotators for Crowdsourced Labeling Tasks”, *Journal of Machine Learning Research*, Vol. 13, pp. 491–518, 2012.

15. Bi, J. and X. Wang, “Min-Max Optimization for Multiple Instance Learning from Multiple Data Annotators”, *KDD’13 August*, 2013.

16. Chittaranjan, G., O. Aran and D. Gatica-Perez, “Exploiting Observers’ Judgements for Nonverbal Group Interaction Analysis”, *Face and Gesture 2011*, pp. 734–739, Mar. 2011.

17. Biel, J., O. Aran and D. Gatica-Perez, "You Are Known by How You Vlog: Personality Impressions and Nonverbal Behavior in YouTube", *Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

18. Dawid, A. P. and A. M. Skene, "Maximum likelihood estimation of observer error-rates using the EM algorithm", *Applied statistics*, pp. 20–28, 1979.

19. Raykar, V. C. and S. Yu, "Annotation Models for Crowdsourced Ordinal Data", *NIPS Workshop on Computational Social Science and the Wisdom of Crowds*, 2011.

20. Srivastava, G., J. A. Yoder, J. Park and A. C. Kak, "Using Objective Ground-Truth Labels Created by Multiple Annotators for Improved Video Classification: A Comparative Study", *Computer Vision and Image Understanding*, Vol. 117, No. 10, pp. 1384 – 1399, 2013.

21. Lakshminarayanan, B. and Y. Teh, "Inferring Ground Truth From Multi-Annotator Ordinal Data: A Probabilistic Approach", *arXiv preprint arXiv:1305.0015*, pp. 1–19, 2013.

22. Peng, J., Q. Liu, A. Ihler and B. Berger, "Crowdsourcing for Structured Labeling With Applications to Protein Folding", *ICML Workshop on Machine Learning Meets Crowdsourcing*, pp. 2008–2012, 2013.

23. Ok, J., S. Oh, J. Shin, Y. Jang and Y. Yi, "Iterative Bayesian Learning for Crowdsourced Regression", *arXiv:1702.08840 [cs.LG]*, pp. 1–22, 2017, http://arxiv.org/abs/1702.08840.

24. Settles, B., *Active Learning Literature Survey*, Tech. Rep. 1648, University of Wisconsin-Madison, Computer Sciences, 2010.

25. Fu, Y., X. Zhu and B. Li, "A Survey on Instance Selection for Active Learning", *Knowledge and Information Systems*, Vol. 35, No. 2, pp. 249–283, 2013.

26. Donmez, P., J. G. Carbonell and 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*, pp. 259–268, 2009.

27. Raykar, V. C. and S. Yu, "Ranking Annotators for Crowdsourced Labeling Tasks", J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira and K. Q. Weinberger (Editors), *Advances in Neural Information Processing Systems 24*, pp. 1809–1817, Curran Associates, Inc., 2011.

28. Fang, M., J. Yin and D. Tao, "Active Learning for Crowdsourcing Using Knowledge Transfer", *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pp. 1809–1815, 2014.

29. Li, H., B. Zhao and A. Fuxman, "The Wisdom of Minority: Discovering and Targeting the Right Group of Workers for Crowdsourcing", *Proceedings of the 23rd International Conference on World Wide Web*, pp. 165–175, 2014.

30. Jagabathula, S., L. Subramanian and A. Venkataraman, "Reputation-based Worker Filtering in Crowdsourcing", *Advances in Neural Information Processing Systems*, pp. 2492–2500, 2014.

31. Zhang, Q., Y. Wen, X. Tian, X. Gan and X. Wang, "Incentivize Crowd Labeling Under Budget Constraint", *2015 IEEE Conference on Computer Communications (INFOCOM)*, pp. 2812–2820, 2015.

32. Audhkhasi, K. and S. S. Narayanan, "A Globally-Variant Locally-Constant Model for Fusion of Labels from Multiple Diverse Experts Without Using Reference Labels", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 35, No. 4, pp. 769–783, 2013.

33. Liu, Q., J. Peng and A. T. Ihler, "Variational Inference for Crowdsourcing", F. Pereira, C. J. C. Burges, L. Bottou and K. Q. Weinberger (Editors), *Advances in*

Neural Information Processing Systems 25, pp. 692–700, Curran Associates, Inc., 2012.

34. Tian, Y. and 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*, pp. 226–234, 2012.

35. Wu, W., Y. Liu, M. Guo, C. Wang and X. Liu, "A Probabilistic Model of Active Learning with Multiple Noisy Oracles", *Neurocomputing*, Vol. 118, pp. 253 – 262, 2013.

36. Dutta, H. and W. Chan, "Using community structure detection to rank annotators when ground truth is subjective", *NIPS Workshop on Human Computation for Science and Computational Sustainability*, pp. 1–4, 2012.

37. Welinder, P. and P. Perona, "Online Crowdsourcing: Rating Annotators and Obtaining Cost-Effective Labels", *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pp. 25–32, Jun. 2010.

38. Zhang, P. and Z. Obradovic, "Learning from Inconsistent and Unreliable Annotators by a Gaussian Mixture Model and Bayesian Information Criterion", *Machine Learning and Knowledge Discovery in Databases*, pp. 553–568, 2011.

39. Whitehill, J., T. fan Wu, J. Bergsma, J. R. Movellan and P. L. Ruvolo, "Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise", Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams and A. Culotta (Editors), *Advances in Neural Information Processing Systems 22*, pp. 2035–2043, Curran Associates, Inc., 2009.

40. Ghosh, A., S. Kale and P. McAfee, "Who Moderates the Moderators? Crowdsourcing Abuse Detection in User-Generated Content Categories", *Proceedings of the 12th ACM Conference on Electronic Commerce*, pp. 167–176, 2011.

41. Wauthier, F. L. and M. I. Jordan, “Bayesian Bias Mitigation for Crowdsourcing”, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira and K. Q. Weinberger (Editors), *Advances in Neural Information Processing Systems 24*, pp. 1800–1808, Curran Associates, Inc., 2011.

42. Donmez, P. and J. G. Carbonell, “Paired-Sampling in Density-Sensitive Active Learning”, *The International Symposium on Artificial Intelligence and Mathematics*, 2008.

43. Sheng, V. S., F. Provost and P. G. Ipeirotis, “Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers”, *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 614–622, 2008.

44. Gao, J., X. Liu, B. C. Ooi, H. Wang and G. Chen, “An Online Cost Sensitive Decision-Making Method in Crowdsourcing Systems”, *Proceedings of the 2013 international conference on Management of data - SIGMOD ’13*, pp. 217–228, 2013.

45. Lin, C. H., Mausam and D. S. Weld, “Re-active Learning : Active Learning with Relabeling”, *Thirtieth AAAI Conference on Artificial Intelligence*, pp. 1845–1852, 2016.

46. Khetan, A. and S. Oh, “Achieving Budget-optimality with Adaptive Schemes in Crowdsourcing”, *Advances in Neural Information Processing Systems 30*, 2, pp. 4844–4852, 2016.

47. Donmez, P. and J. G. Carbonell, “Proactive Learning : Cost-Sensitive Active Learning with Multiple Imperfect Oracles”, *Proceedings of the 17th ACM conference on Information and knowledge management*, pp. 619–628, 2008.

48. Hsueh, P.-Y., P. Melville and V. Sindhwani, “Data Quality from Crowdsourcing: A Study of Annotation Selection Criteria”, *Proceedings of the NAACL HLT 2009*

Workshop on Active Learning for Natural Language Processing, June, pp. 27–35, 2009.

49. Tran-Thanh, L., M. Venanzi, A. Rogers and N. R. Jennings, “Efficient Budget Allocation With Accuracy Guarantees for Crowdsourcing Classification Tasks”, *Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems*, pp. 6–10, 2013.

50. Tran-Thanh, L., T. D. Huynh, A. Rosenfeld, S. Ramchurn and N. R. Jennings, “BudgetFix : Budget Limited Crowdsourcing for Interdependent Task Allocation With Quality Guarantees”, *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems*, pp. 477–484, 2014.

51. Nguyen, A. T., B. C. Wallace and M. Lease, “Combining Crowd and Expert Labels using Decision Theoretic Active Learning”, *Proceedings of the 3rd AAAI Conference on Human Computation (HCOMP)*, pp. 120–129, 2015.

52. Yan, Y., R. Rosales, G. Fung and J. G. Dy, “Active Learning from Crowds”, *Proceedings of the 28th International Conference on Machine Learning*, pp. 1161–1168, 2011.

53. Raykar, V. C. and P. Agrawal, “Sequential Crowdsourced Labeling as an Epsilon-Greedy Exploration in a Markov Decision Process”, *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics (AISTATS-14)*, Vol. 33, pp. 832–840, 2014.

54. Mozafari, B., P. Sarkar, M. Franklin, M. Jordan and S. Madden, “Scaling Up Crowd-Sourcing to Very Large Datasets: A Case for Active Learning”, *Proceedings of the VLDB Endowment*, Vol. 8, No. 2, pp. 125–136, 2014.

55. Zhuang, H. and J. Young, “Leveraging In-Batch Annotation Bias for Crowdsourced Active Learning”, *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM’15)*, pp. 243–252, 2015.

56. Zhu, C., H. Xu and S. Yan, "Online Crowdsourcing", *arXiv:1512.02393 [cs]*, 2015, `http://www.arxiv.org/pdf/1512.02393.pdf`.

57. Ho, C.-J., A. Slivkins and J. W. Vaughan, "Adaptive Contract Design for Crowdsourcing Markets: Bandit Algorithms for Repeated Principal-Agent Problems", *Journal of Artificial Intelligence Research*, Vol. 55, pp. 317–359, 2016.

58. Li, Q., F. Ma, J. Gao, L. Su and C. J. Quinn, "Crowdsourcing High Quality Labels With a Tight Budget", *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining (WSDM'16)*, pp. 237–246, 2016.

59. Karger, D. R., S. Oh and D. Shah, "Iterative Learning for Reliable Crowdsourcing Systems", J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira and K. Q. Weinberger (Editors), *Advances in Neural Information Processing Systems 24*, pp. 1953–1961, Curran Associates, Inc., 2011.

60. Karger, D. R., S. Oh and D. Shah, "Budget-Optimal Task Allocation for Reliable Crowdsourcing Systems", *Operations Research*, Vol. 62, No. 1, pp. 1–24, 2014.

61. Ho, C.-j., S. Jabbari and J. W. Vaughan, "Adaptive Task Assignment for Crowdsourced Classification", *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, Vol. 28, pp. 534–542, 2013.

62. Kamar, E., S. Hacker and E. Horvitz, "Combining Human and Machine Intelligence in Large-Scale Crowdsourcing", *AAMAS '12 Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems - Volume 1*, pp. 467–474, 2012.

63. Kamar, E., A. Kapoor and E. Horvitz, "Lifelong Learning for Acquiring the Wisdom of the Crowd", *International Joint Conference on Artificial Intelligence (IJCAI)*, Vol. 13, pp. 2313–2320, 2013.

64. Kamar, E., A. Kapoor and E. Horvitz, “Identifying and Accounting for Task-Dependent Bias in Crowdsourcing”, *Proceedings, The Third AAAI Conference on Human Computation and Crowdsourcing (HCOMP-15)*, pp. 92–101, 2015.

65. Venanzi, M., J. Guiver, P. Kohli and N. R. Jennings, “Time-Sensitive Bayesian Information Aggregation for Crowdsourcing Systems”, *Journal of Artificial Intelligence Research*, Vol. 56, pp. 517–545, 2016.

66. Snow, R., B. O’Connor, D. Jurafsky and A. Y. Ng, “Cheap and Fast—But is it Good?: Evaluating Non-Expert Annotations for Natural Language Tasks”, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 254–263, Association for Computational Linguistics, 2008.

67. Ambati, V., S. Vogel and J. Carbonell, “Active Learning and Crowd-Sourcing for Machine Translation”, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*, pp. 2169–2174, 2010.

68. Laws, F., C. Scheible and H. Schütze, “Active Learning With Amazon Mechanical Turk”, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1546–1556, 2011.

69. Marcus, A., D. Karger, S. Madden, R. Miller and S. Oh, “Counting with the Crowd”, *Proceedings of the VLDB Endowment*, Vol. 6, No. 2, pp. 109–120, 2012.

70. Guo, S., A. Parameswaran and H. Garcia-Molina, “So Who Won?: Dynamic Max Discovery with the Crowd”, *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pp. 385–396, ACM, 2012.

71. Kara, Y. E., G. Genc, O. Aran and L. Akarun, “Modeling Annotator Behaviors for Crowd Labeling”, *Neurocomputing*, Vol. 160, pp. 141–156, 2015.

72. Kara, Y. E., G. Genc, O. Aran and L. Akarun, "Actively Estimating Crowd Annotation Consensus", *Journal of Artificial Intelligence Research*, Vol. 61, pp. 363–405, 2018.

73. Aran, O. and D. Gatica-Perez, "One of a Kind: Inferring Personality Impressions in Meetings", *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, pp. 11–18, ACM, 2013.

74. Face and Gesture Recognition Working group, *The FGNet Aging Database*, 2018, `http://www-prima.inrialpes.fr/FGnet/html/benchmarks.html`, accessed at May 2018.

75. Gourier, N., D. Hall and J. L. Crowley, *Head Pose Image Database*, 2018, `http://www-prima.inrialpes.fr/perso/Gourier/Faces/HPDatabase.html`, accessed at May 2018.

76. Gourier, N., D. Hall and J. L. Crowley, "Estimating Face Orientation From Robust Detection of Salient Facial Structures", *FG Net Workshop on Visual Observation of Deictic Gestures*, pp. 1–9, FGnet (IST–2000–26434) Cambridge, UK, 2004.

77. Strapparava, C. and R. Mihalcea, "Semeval-2007 Task 14: Affective text", *Proceedings of the 4th International Workshop on Semantic Evaluations*, pp. 70–74, Association for Computational Linguistics, 2007.

78. Aran, O. and D. Gatica-Perez, "Cross-Domain Personality Prediction: From Video Blogs to Small Group Meetings", *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, pp. 127–130, ACM, 2013.

79. Sanchez-Cortes, D., O. Aran, M. Mast and D. Gatica-Perez, "A Nonverbal Behavior Approach to Identify Emergent Leaders in Small Groups", *IEEE Transactions on Multimedia*, Vol. 14, No. 3, pp. 816–832, 2012.

80. Sanchez-Cortes, D., O. Aran and D. Gatica-Perez, “An Audio Visual Corpus for Emergent Leader Analysis”, *ICMI-MLMI11: Workshop on Multimodal Corpora for Machine Learning: Taking Stock and Road Mapping the Future*, Nov 2011.

81. Gosling, S. D., P. J. Rentfrow and W. B. Swann, “A Very Brief Measure of the Big-Five Personality Domains”, *Journal of Research in Personality*, Vol. 37, pp. 504–528, 2003.

82. Zhu, D. and B. Carterette, “An Analysis of Assessor Behavior in Crowdsourced Preference Judgments”, *SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation*, pp. 17–20, 2010.

83. Kittur, A., E. H. Chi and B. Suh, “Crowdsourcing User Studies with Mechanical Turk”, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 453–456, ACM, 2008.

84. Abramowitz, M., I. A. Stegun *et al.*, “Handbook of Mathematical Functions”, *Applied Mathematics Series*, Vol. 55, No. 62, p. 39, 1966.

85. Gupta, A. K. and D. K. Nagar, *Matrix Variate Distributions*, Vol. 104, CRC Press, 1999.

86. Knapp, M. L. and J. A. Hall, *Nonverbal Communication in Human Interaction*, Wadsworth, Cengage Learning, 2008.

87. Gifford, R., *The SAGE Handbook of Nonverbal Communication*, chap. Personality and Nonverbal Behavior: A Complex Conundrum, pp. 159–181, SAGE Publications, Inc., 2006.

APPENDIX A: PROOFS OF THEOREMS

A.1. Proof of Theorem 4.1

Theorem 4.1 (Posterior distribution of x)**.** *Let the distribution of y_k be*

$$\mathcal{N}\left(y_k; a_{j_k}(w_{j_k}x_{i_k} + b_{j_k}), \frac{1}{\lambda_{j_k}}\right).$$

Then, the posterior distribution of x_i is

$$x_i|\{y_k, \theta_{j_k} : k \in \mathcal{K}_i\} \sim \mathcal{N}\left(x_i; \frac{\sum\limits_{k:i_k=i} \lambda_{j_k} w_{j_k}(a_{j_k}y_k - b_{j_k})}{\sum\limits_{k:i_k=i} w_{j_k}^2 \lambda_{j_k}}, \left(\sum_{k:i_k=i} w_{j_k}^2 \lambda_{j_k}\right)^{-1}\right)$$

where $\theta_j = \{a_j, w_j, b_j, \lambda_j\}$ is the set of parameters of annotator j and $\mathcal{K}_i = \{k \in \mathcal{K} : i_k = i\}$ is the set of annotations of sample i.

Proof. Let N, R, and K be number of samples, annotators, and annotations, respectively.

$$p(y_{1:K}|x_{1:N}, \theta_{1:R}) = \prod_{k=1}^{K} \mathcal{N}\left(y_k; a_{j_k}(w_{j_k}x_{i_k} + b_{j_k}), \frac{1}{\lambda_{j_k}}\right) \tag{A.1}$$

$$\log p(y_{1:K}|x_{1:N}, \theta_{1:R}) = \log \prod_{k=1}^{K} \mathcal{N}\left(y_k; a_{j_k}(w_{j_k}x_{i_k} + b_{j_k}), \frac{1}{\lambda_{j_k}}\right) \tag{A.2}$$

$$= \sum_{k=1}^{K} \log \mathcal{N}\left(y_k; a_{j_k}(w_{j_k}x_{i_k} + b_{j_k}), \frac{1}{\lambda_{j_k}}\right) \tag{A.3}$$

$$= \sum_{k=1}^{K} \left(-\frac{1}{2}\log 2\pi - \frac{1}{2}\log\frac{1}{\lambda_{j_k}} - \frac{1}{2}\left(\frac{(y_k - a_{j_k}(w_{j_k}x_{i_k} + b_{j_k}))^2}{\frac{1}{\lambda_{j_k}}}\right)\right) \tag{A.4}$$

$$= -\frac{K}{2}\log 2\pi - \frac{1}{2}\sum_{k=1}^{K}\log\frac{1}{\lambda_{j_k}} - \frac{1}{2}\sum_{k=1}^{K}\lambda_{j_k}\left(y_k - a_{j_k}(w_{j_k}x_{i_k} + b_{j_k})\right)^2 \tag{A.5}$$

Since $a_j \in \{-1, 1\}\ \forall j$, $a_j^2 = 1$:

$$= -\frac{K}{2}\log 2\pi - \frac{1}{2}\sum_{k=1}^{K}\log\frac{1}{\lambda_{j_k}} - \frac{1}{2}\sum_{k=1}^{K}\lambda_{j_k}(a_{j_k}y_k - b_{j_k} - w_{j_k}x_{i_k})^2 \tag{A.6}$$

From Bayes' rule we know that

$$p(x_i|y_{1:K}, x_{-i}, \theta_{1:R}) = \frac{p(y_{1:K}|x_{1:N}, \theta_{1:R})p(x_i)}{p(y_{1:K}|\theta_{1:R})} \tag{A.7}$$

Since the prior of x_i is flat

$$p(x_i|y_{1:K}, x_{-i}, \theta_{1:R}) \propto p(y_{1:K}|x_{1:N}, \theta_{1:R}) \tag{A.8}$$

By omitting independent variables, we get

$$p(x_i|\{y_k, \theta_{j_k} : k \in \mathcal{K}_i\}) \propto p(y_{1:K}|x_{1:N}, \theta_{1:R}) \tag{A.9}$$

Combining Equations A.6 and A.9 gives us

$$\begin{aligned}\log p(x_i|\{y_k, \theta_{j_k} : k \in \mathcal{K}_i\}) \propto -\frac{K}{2}\log 2\pi - \frac{1}{2}\sum_{k=1}^{K}\log\frac{1}{\lambda_{j_k}} \\ - \frac{1}{2}\sum_{k=1}^{K}\lambda_{j_k}(a_{j_k}y_k - b_{j_k} - w_{j_k}x_{i_k})^2\end{aligned} \tag{A.10}$$

By omitting the terms without x_i we get:

$$\propto -\frac{1}{2} \sum_{k:i_k=i} \lambda_{j_k} (a_{j_k} y_k - b_{j_k} - w_{j_k} x_i)^2 \tag{A.11}$$

$$\begin{aligned} \propto -\frac{1}{2} \sum_{k:i_k=i} \big(& \lambda_{j_k} (a_{j_k} y_k - b_{j_k})^2 + \lambda_{j_k} w_{j_k}^2 x_i^2 \\ & -2\lambda_{j_k} w_{j_k} x_i (a_{j_k} y_k - b_{j_k}) \big) \end{aligned} \tag{A.12}$$

Rearranging and omitting the terms without x_i:

$$\propto -\frac{1}{2} x_i^2 \underbrace{\sum_{k:i_k=i} w_{j_k}^2 \lambda_{j_k}}_{\sigma^{-2}} + x_i \underbrace{\sum_{k:i_k=i} \lambda_{j_k} w_{j_k} (a_{j_k} y_k - b_{j_k})}_{\mu\sigma^{-2}} \tag{A.13}$$

The equation is in the form of normal distribution. Therefore, we have

$$p(x_i | \{y_k, \theta_{j_k} : k \in \mathcal{K}_i\}) = \mathcal{N} \left(x_i ; \frac{\sum_{k:i_k=i} \lambda_{j_k} w_{j_k} (a_{j_k} y_k - b_{j_k})}{\sum_{k:i_k=i} w_{j_k}^2 \lambda_{j_k}}, \left(\sum_{k:i_k=i} w_{j_k}^2 \lambda_{j_k} \right)^{-1} \right) \tag{A.14}$$

□

A.2. Proof of Theorem 4.2

Theorem 4.2 (Posterior distribution of λ). *Let $x_k, y_k \in \mathbb{R}, \forall k \in \{1, \ldots, K\}$, $w > 0$, and $\lambda > 0$. If the distribution of y_k is $\mathcal{N}(y_k; wx_k, w^2\lambda^{-1})$, then the posterior distribution of λ is*

$$\mathcal{G}\left(\lambda; \frac{K}{2} + 1, \frac{1}{2} \sum_{k=1}^{K} \left(\frac{y_k}{w} - x_k \right)^2 \right).$$

Moreover, if the prior distribution of λ is $\mathcal{G}\left(\lambda; \alpha_\lambda, \beta_\lambda\right)$, then the posterior is

$$\mathcal{G}\left(\lambda; \frac{K}{2} + \alpha_\lambda, \beta_\lambda + \frac{1}{2}\sum_{k=1}^{K}\left(\frac{y_k}{w} - x_k\right)^2\right).$$

Proof.

$$\begin{aligned}
L &= \log \prod_{k=1}^{K} \mathcal{N}\left(y_k; wx_k, w^2\lambda^{-1}\right) && \text{(A.15)}\\
&= \sum_{k=1}^{K} \log \mathcal{N}\left(y_k; wx_k, w^2\lambda^{-1}\right) && \text{(A.16)}\\
&= \sum_{k=1}^{K}\left(-\frac{1}{2}\log\left(2\pi\frac{w^2}{\lambda}\right) - \frac{1}{2}\left(\frac{(y_k - wx_k)^2}{\frac{w^2}{\lambda}}\right)\right) && \text{(A.17)}\\
&= -\frac{K}{2}\log(2\pi w^2) + \frac{K}{2}\log\lambda - \lambda\frac{1}{2}\sum_{k=1}^{K}\left(\frac{y_k}{w} - x_k\right)^2 && \text{(A.18)}
\end{aligned}$$

The equation is in the form of Gamma distribution. Therefore, we have

$$\exp(L) \propto \mathcal{G}\left(\lambda; \frac{K}{2} + 1, \frac{1}{2}\sum_{k=1}^{K}\left(\frac{y_k}{w} - x_k\right)^2\right) \tag{A.19}$$

$$\begin{aligned}
P &= \log \mathcal{G}\left(\lambda; \alpha_\lambda, \beta_\lambda\right) && \text{(A.20)}\\
&= \alpha_\lambda \log \beta_\lambda + (\alpha_\lambda - 1)\log\lambda - \lambda\beta_\lambda - \log\Gamma\left(\alpha_\lambda\right) && \text{(A.21)}
\end{aligned}$$

Then, we have

$$\exp(L + P) \propto \mathcal{G}\left(\lambda; \frac{K}{2} + \alpha_\lambda, \beta_\lambda + \frac{1}{2}\sum_{k=1}^{K}\left(\frac{y_k}{w} - x_k\right)^2\right) \tag{A.22}$$

□

A.3. Proof of Theorem 4.3

Theorem 4.3 (Posterior distribution of a)**.** *Suppose that the values $x_k, y_k \in \mathbb{R}, \forall k \in \{1, \ldots, K\}$ and $\lambda > 0$ are given. Let $c \sim \mathcal{B}(c; p)$ and the distribution of y_k be $y_k \sim \mathcal{N}(y_k; ax_k, \lambda^{-1})$ where $a = 2c - 1$. Then the posterior distribution of c is*

$$\mathcal{B}\left(c; \left[1 + \exp\left(-2\lambda \sum_{k=1}^{K} y_k x_k\right)\right]^{-1}\right).$$

Moreover, the value a^* that maximizes this distribution is given by

$$a^* = \operatorname{sgn}\left(\sum_{k=1}^{K} y_k x_k\right).$$

Proof.

$$\begin{aligned}
L &= -\frac{1}{2}\sum_{k=1}^{K}\left(\log|2\pi\lambda^{-1}| + \lambda(y_k - ax_k)^2\right) && \text{(A.23)}\\
&= -\frac{1}{2}\sum_{k=1}^{K}\left(\log|2\pi\lambda^{-1}| + y_k\lambda y_k - ax_k\lambda y_k - ay_k\lambda x_k + a^2 x_k\lambda x_k\right) && \text{(A.24)}\\
&= -\frac{1}{2}\sum_{k=1}^{K}\left(\log|2\pi\lambda^{-1}| + y_k\lambda y_k - 2ay_k\lambda x_k + x_k\lambda x_k\right) && \text{(A.25)}\\
&= a\lambda\sum_{k=1}^{K} y_k x_k - \frac{1}{2}\sum_{k=1}^{K}\left(\log|2\pi\lambda^{-1}| + y_k\lambda y_k + x_k\lambda x_k\right) && \text{(A.26)}\\
&= (2c-1)\lambda\sum_{k=1}^{K} y_k x_k - \frac{1}{2}\sum_{k=1}^{K}\left(\log|2\pi\lambda^{-1}| + y_k\lambda y_k + x_k\lambda x_k\right) && \text{(A.27)}\\
&= c\underbrace{\left(2\lambda\sum_{k=1}^{K} y_k x_k\right)}_{\log p - \log(1-p)} - \lambda\sum_{k=1}^{K} y_k x_k - \frac{1}{2}\sum_{k=1}^{K}\left(\log|2\pi\lambda^{-1}| + y_k\lambda y_k + x_k\lambda x_k\right) && \text{(A.28)}
\end{aligned}$$

Then, we end up with

$$\exp(L) \propto \mathcal{B}\left(c; \left[1 + \exp\left(-2\lambda \sum_{k=1}^{K} y_k x_k\right)\right]^{-1}\right) \tag{A.29}$$

When $\left[1+\exp\left(-2\lambda\sum_{k=1}^{K} y_k x_k\right)\right]^{-1} < \frac{1}{2}$, the c value that maximizes the above distribution is $c^* = 0$, *i.e.* $a^* = -1$.

$$\begin{aligned}
a^* = -1 \iff\ & c^* = 0 && \text{(A.30)}\\
\iff\ & \left[1+\exp\left(-2\lambda\sum_{k=1}^{K} y_k x_k\right)\right]^{-1} < \frac{1}{2} && \text{(A.31)}\\
\iff\ & 1+\exp\left(-2\lambda\sum_{k=1}^{K} y_k x_k\right) > 2 && \text{(A.32)}\\
\iff\ & \exp\left(-2\lambda\sum_{k=1}^{K} y_k x_k\right) > 1 && \text{(A.33)}\\
\iff\ & -2\lambda\sum_{k=1}^{K} y_k x_k > 0 && \text{(A.34)}\\
\iff\ & \sum_{k=1}^{K} y_k x_k < 0 && \text{(A.35)}\\
\iff\ & \operatorname{sgn}\left(\sum_{k=1}^{K} y_k x_k\right) = -1 && \text{(A.36)}
\end{aligned}$$

The case for $a^* = 1$ is similar, resulting in $\operatorname{sgn}\left(\sum_{k=1}^{K} y_k x_k\right) = 1$. Therefore, we have $a^* = \operatorname{sgn}\left(\sum_{k=1}^{K} y_k x_k\right)$. □

A.4. Proof of Theorem 4.4

Theorem 4.4 (Posterior distribution of w). *Let $x_k, y_k \in \mathbb{R}, \forall k \in \{1,\ldots,K\}$, $w > 0$, and $\lambda > 0$. Let the distribution of y_k be $y_k \sim \mathcal{N}(y_k; wx_k, \lambda^{-1})$. Then, the posterior distribution of w is*

$$\mathcal{N}_{trunc}\left(w; \frac{\sum_{k=1}^{K} y_k x_k}{\sum_{k=1}^{K} x_k^2}, \left(\lambda\sum_{k=1}^{K} x_k^2\right)^{-1}, 0, \infty\right).$$

Moreover, if $w \sim \mathcal{G}\left(w; \alpha_w, \beta_w\right)$*, then the posterior distribution of* w *becomes*

$$\mathcal{GPTN}\left(x; \frac{\lambda \sum_{k=1}^{K} y_k x_k - \beta_w}{\lambda \sum_{k=1}^{K} x_k^2}, \left(\lambda \sum_{k=1}^{K} x_k^2\right)^{-1}, \alpha_w - 1\right).$$

Proof.

$$\begin{aligned}
L &= \log \prod_{k=1}^{K} \mathcal{N}\left(y_k; w x_k, \lambda^{-1}\right) \\
&= \sum_{k=1}^{K} \log \mathcal{N}\left(y_k; w x_k, \lambda^{-1}\right) \\
&= \sum_{k=1}^{K} \left(-\frac{1}{2} \log\left(2\pi\lambda^{-1}\right) - \frac{\lambda}{2}(y_k - w x_k)^2\right) \\
&= -\frac{K}{2} \log(2\pi) + \frac{K}{2} \log \lambda - \frac{\lambda}{2} \sum_{k=1}^{K} (y_k - w x_k)^2 \\
&= -\frac{K}{2} \log(2\pi) + \frac{K}{2} \log \lambda - \frac{\lambda}{2} \sum_{k=1}^{K} y_k^2 + w \underbrace{\lambda \sum_{k=1}^{K} y_k x_k}_{\sigma^{-2}\mu} - \frac{1}{2} w^2 \underbrace{\lambda \sum_{k=1}^{K} x_k^2}_{\sigma^{-2}}
\end{aligned}$$

Since $x > 0$, the above equation is in the form of the positively truncated normal distribution. Therefore, we have

$$\exp(L) \propto \mathcal{N}_{trunc}\left(w; \frac{\sum_{k=1}^{K} y_k x_k}{\sum_{k=1}^{K} x_k^2}, \left(\lambda \sum_{k=1}^{K} x_k^2\right)^{-1}, 0, \infty\right)$$

$$P = \log \mathcal{G}\left(w; \alpha_w, \beta_w\right)$$

$$= \alpha_w \log \beta_w + (\alpha_w - 1) \log w - w\beta_w - \log \Gamma(\alpha_w)$$

Then, we have

$$\exp(L+P) \propto w^{\alpha_w - 1} \exp\left(\underbrace{w\,\lambda \sum_{k=1}^{K} y_k x_k - \beta_w}_{\sigma^{-2}\mu} - \frac{1}{2} w^2 \underbrace{\lambda \sum_{k=1}^{K} x_k^2}_{\sigma^{-2}} \right)$$

$$\propto \mathcal{GPTN}\left(x; \frac{\lambda \sum_{k=1}^{K} y_k x_k - \beta_w}{\lambda \sum_{k=1}^{K} x_k^2}, \left(\lambda \sum_{k=1}^{K} x_k^2 \right)^{-1}, \alpha_w - 1 \right)$$

□

A.5. Proof of Theorem 4.5

Theorem 4.5 (Posterior distribution of b)**.** *Let $y_k \in \mathbb{R}, \forall k \in \{1, \ldots, K\}$, $b \in \mathbb{R}$, $w > 0$, and $\lambda > 0$. If the distribution of y_k is $y_k \sim \mathcal{N}(y_k; wb, w^2\lambda^{-1})$, then the posterior distribution of b is*

$$\mathcal{N}\left(b; \frac{1}{wK} \sum_{k=1}^{K} y_k, (K\lambda)^{-1} \right).$$

Moreover, if $b \sim \mathcal{N}\left(b; \mu_b, \lambda_b^{-1}\right)$, then the posterior distribution of b becomes

$$\mathcal{N}\left(b; \frac{\frac{\lambda}{w} \sum_{k=1}^{K} y_k + \mu_b \lambda_b}{K\lambda + \lambda_b}, (K\lambda + \lambda_b)^{-1} \right).$$

Proof.

$$\begin{aligned}
L &= \log \prod_{k=1}^{K} \mathcal{N}\left(y_k; wb, w^2\lambda^{-1}\right) && \text{(A.37)}\\
&= \sum_{k=1}^{K} \log \mathcal{N}\left(y_k; wb, w^2\lambda^{-1}\right) && \text{(A.38)}\\
&= \sum_{k=1}^{K} \left(-\frac{1}{2}\log\left(2\pi\frac{w^2}{\lambda}\right) - \frac{1}{2}\left(\frac{(y_k - wb)^2}{\frac{w^2}{\lambda}}\right)\right) && \text{(A.39)}\\
&= -\frac{K}{2}\log(2\pi w^2) + \frac{K}{2}\log\lambda - \lambda\frac{1}{2}\sum_{k=1}^{K}\left(\frac{y_k}{w} - b\right)^2 && \text{(A.40)}\\
&= -\frac{K}{2}\log(2\pi w^2) + \frac{K}{2}\log\lambda - \lambda\frac{1}{2}\sum_{k=1}^{K}\left(\frac{y_k}{w}\right)^2 + \lambda\sum_{k=1}^{K}\frac{y_k}{w}b - \lambda\frac{1}{2}\sum_{k=1}^{K} b^2 && \text{(A.41)}\\
&= -\frac{K}{2}\log(2\pi w^2) + \frac{K}{2}\log\lambda - \lambda\frac{1}{2}\sum_{k=1}^{K}\left(\frac{y_k}{w}\right)^2 + b\,\underbrace{\lambda\sum_{k=1}^{K}\frac{y_k}{w}}_{\sigma^{-2}\mu} - \frac{1}{2}b^2\underbrace{K\lambda}_{\sigma^{-2}} && \text{(A.42)}
\end{aligned}$$

The equation is in the form of normal distribution. Therefore, we have

$$\exp(L) \propto \mathcal{N}\left(b; \frac{1}{wK}\sum_{k=1}^{K} y_k, (K\lambda)^{-1}\right) \tag{A.43}$$

$$\begin{aligned}
P &= \log \mathcal{N}\left(b; \mu_b, \lambda_b^{-1}\right) && \text{(A.44)}\\
&= -\frac{1}{2}\log(2\pi\lambda_b^{-1}) - \frac{\lambda_b}{2}(b - \mu_b)^2 && \text{(A.45)}\\
&= -\frac{1}{2}\log(2\pi\lambda_b^{-1}) - \frac{1}{2}b^2\lambda_b - \frac{\lambda_b}{2}\mu_b^2 + b\lambda_b\mu_b && \text{(A.46)}
\end{aligned}$$

Then, we have

$$\exp(L+P) \propto \mathcal{N}\left(b; \frac{\frac{\lambda}{w}\sum_{k=1}^{K} y_k + \mu_b\lambda_b}{K\lambda+\lambda_b}, (K\lambda+\lambda_b)^{-1}\right) \tag{A.47}$$

□

A.6. Proof of Theorem 4.6

Theorem 4.6 (Posterior distribution of M-CBS x)**.** *Let the distribution of y_k be*

$$\mathcal{N}\left(y_k; a_{j_k}w_{j_k}(x_{i_k}+b_{j_k}), \frac{w_{j_k}^2}{\lambda_{j_k}}\right).$$

Then, the posterior distribution of x_i is

$$x_i|\{y_k, \theta_{j_k} : k \in \mathcal{K}_i\} \sim \mathcal{N}\left(x_i; \frac{\sum_{k:i_k=i} \lambda_{j_k}(w_{j_k}^{-1}a_{j_k}y_k - b_{j_k})}{\sum_{k:i_k=i} \lambda_{j_k}}, \left(\sum_{k:i_k=i} \lambda_{j_k}\right)^{-1}\right)$$

where $\theta_j = \{a_j, w_j, b_j, \lambda_j\}$ is the set of parameters of annotator j and $\mathcal{K}_i = \{k \in \mathcal{K} : i_k = i\}$ is the set of annotations of sample i.

Proof. Let N, R, and K be number of samples, annotators, and annotations, respectively.

$$p(y_{1:K}|x_{1:N}, \theta_{1:R}) = \prod_{k=1}^{K} \mathcal{N}\left(y_k; a_{j_k}w_{j_k}(x_{i_k}+b_{j_k}), \frac{w_{j_k}^2}{\lambda_{j_k}}\right) \tag{A.48}$$

$$\log p(y_{1:K}|x_{1:N}, \theta_{1:R}) = \log \prod_{k=1}^{K} \mathcal{N}\left(y_k; a_{j_k}w_{j_k}(x_{i_k}+b_{j_k}), \frac{w_{j_k}^2}{\lambda_{j_k}}\right) \tag{A.49}$$

$$= \sum_{k=1}^{K} \log \mathcal{N}\left(y_k; a_{j_k}w_{j_k}(x_{i_k}+b_{j_k}), \frac{w_{j_k}^2}{\lambda_{j_k}}\right) \tag{A.50}$$

$$= \sum_{k=1}^{K} \left(-\frac{1}{2}\log 2\pi - \frac{1}{2}\log\frac{w_{j_k}^2}{\lambda_{j_k}} - \frac{1}{2}\left(\frac{(y_k - a_{j_k} w_{j_k}(x_{i_k} + b_{j_k}))^2}{\frac{w_{j_k}^2}{\lambda_{j_k}}} \right) \right) \tag{A.51}$$

$$= -\frac{K}{2}\log 2\pi - \frac{1}{2}\sum_{k=1}^{K}\log\frac{w_{j_k}^2}{\lambda_{j_k}} - \frac{1}{2}\sum_{k=1}^{K}\left(\lambda_{j_k}\frac{(y_k - a_{j_k} w_{j_k}(x_{i_k} + b_{j_k}))^2}{w_{j_k}^2} \right) \tag{A.52}$$

$$= -\frac{K}{2}\log 2\pi - \frac{1}{2}\sum_{k=1}^{K}\log\frac{w_{j_k}^2}{\lambda_{j_k}} - \frac{1}{2}\sum_{k=1}^{K}\lambda_{j_k}(w_{j_k}^{-1} y_k - a_{j_k}(x_{i_k} + b_{j_k}))^2 \tag{A.53}$$

Since $a_j \in \{-1, 1\}\ \forall j$, $a_j^2 = 1$:

$$= -\frac{K}{2}\log 2\pi - \frac{1}{2}\sum_{k=1}^{K}\log\frac{w_{j_k}^2}{\lambda_{j_k}} - \frac{1}{2}\sum_{k=1}^{K}\lambda_{j_k}(w_{j_k}^{-1} a_{j_k} y_k - b_{j_k} - x_{i_k})^2 \tag{A.54}$$

From Bayes' rule we know that

$$p(x_i|y_{1:K}, x_{-i}, \theta_{1:R}) = \frac{p(y_{1:K}|x_{1:N}, \theta_{1:R})p(x_i)}{p(y_{1:K}|\theta_{1:R})} \tag{A.55}$$

Since the prior of x_i is flat

$$p(x_i|y_{1:K}, x_{-i}, \theta_{1:R}) \propto p(y_{1:K}|x_{1:N}, \theta_{1:R}) \tag{A.56}$$

By omitting independent variables, we get

$$p(x_i|\{y_k, \theta_{j_k} : k \in \mathcal{K}_i\}) \propto p(y_{1:K}|x_{1:N}, \theta_{1:R}) \tag{A.57}$$

Combining Equations A.54 and A.57 gives us

$$\log p(x_i|\{y_k,\theta_{j_k}:k\in\mathcal{K}_i\}) \propto -\frac{K}{2}\log 2\pi - \frac{1}{2}\sum_{k=1}^{K}\log\frac{w_{j_k}^2}{\lambda_{j_k}} \\ - \frac{1}{2}\sum_{k=1}^{K}\lambda_{j_k}(w_{j_k}^{-1}a_{j_k}y_k - b_{j_k} - x_{i_k})^2 \tag{A.58}$$

By omitting the terms without x_i we get:

$$\propto -\frac{1}{2}\sum_{k:i_k=i}\lambda_{j_k}(w_{j_k}^{-1}a_{j_k}y_k - b_{j_k} - x_i)^2 \tag{A.59}$$

$$\propto -\frac{1}{2}\sum_{k:i_k=i}\big(\lambda_{j_k}(w_{j_k}^{-1}a_{j_k}y_k - b_{j_k})^2 + \lambda_{j_k}x_i^2 \\ -2x_i\lambda_{j_k}(w_{j_k}^{-1}a_{j_k}y_k - b_{j_k})\big) \tag{A.60}$$

Rearranging and omitting the terms without x_i:

$$\propto -\frac{1}{2}x_i^2\underbrace{\sum_{k:i_k=i}\lambda_{j_k}}_{\sigma^{-2}} + x_i\underbrace{\sum_{k:i_k=i}\lambda_{j_k}(w_{j_k}^{-1}a_{j_k}y_k - b_{j_k})}_{\mu\sigma^{-2}} \tag{A.61}$$

The equation is in the form of normal distribution. Therefore, we have

$$p(x_i|\{y_k,\theta_{j_k}:k\in\mathcal{K}_i\}) = \mathcal{N}\left(x_i;\frac{\sum_{k:i_k=i}\lambda_{j_k}(w_{j_k}^{-1}a_{j_k}y_k - b_{j_k})}{\sum_{k:i_k=i}\lambda_{j_k}},\left(\sum_{k:i_k=i}\lambda_{j_k}\right)^{-1}\right) \tag{A.62}$$

□

A.7. Proof of Theorem 4.7

Theorem 4.7 (Mode of M-CBS w)**.** *Let* $y_k\in\mathbb{R},\forall k\in\{1,\ldots,K\}$*,* $a\in\mathbb{R}$*,* $w>0$*, and* $\lambda>0$*. If the distribution of* y_k *is* $y_k\sim\mathcal{N}(y_k;awx_k,w^2\lambda^{-1})$ *and* $w\sim\mathcal{G}(w;\alpha_w,\beta_w)$*,* *then the value w^* maximizing the posterior probability is a root of the equation*

$$w^{-3}\underbrace{\left(\lambda\sum_{k=1}^{K}y_k^2\right)}_{V_3}+w^{-2}\underbrace{\left(-\lambda a\sum_{k=1}^{K}y_kx_k\right)}_{V_2}+w^{-1}\underbrace{(\alpha_w-1-K)}_{V_1}+\underbrace{(-\beta_w)}_{V_0}=0$$

Proof.

$$\begin{aligned}
L &= \log\prod_{k=1}^{K}\mathcal{N}\left(y_k;awx_k,\frac{w^2}{\lambda}\right)\\
&= \sum_{k=1}^{K}\log\mathcal{N}\left(y_k;awx_k,\frac{w^2}{\lambda}\right)\\
&= \sum_{k=1}^{K}\left(-\frac{1}{2}\log(2\pi\lambda^{-1})-\frac{1}{2}\log w^2-\frac{1}{2}\left(\frac{(y_k-awx_k)^2}{\frac{w^2}{\lambda}}\right)\right)\\
&= -\frac{K}{2}\log(2\pi\lambda^{-1})-K\log w-\frac{\lambda}{2}\sum_{k=1}^{K}(w^{-1}y_k-ax_k)^2\\
&= -\frac{K}{2}\log(2\pi\lambda^{-1})-K\log w-w^{-2}\frac{\lambda}{2}\sum_{k=1}^{K}y_k^2\\
&\quad+w^{-1}\lambda a\sum_{k=1}^{K}y_kx_k-\frac{\lambda a^2}{2}\sum_{k=1}^{K}x_k^2
\end{aligned}$$

$$\begin{aligned}
P &= \log\mathcal{G}\left(w;\alpha_w,\beta_w\right)\\
&= \alpha_w\log\beta_w+(\alpha_w-1)\log w-w\beta_w-\log\Gamma\left(\alpha_w\right)
\end{aligned}$$

$$\begin{aligned}
\frac{dL+P}{dw} &= \frac{d}{dw}\left((\alpha_w-1-K)\log w-w^{-2}\frac{\lambda}{2}\sum_{k=1}^{K}y_k^2+w^{-1}\lambda a\sum_{k=1}^{K}y_kx_k-w\beta_w\right)\\
&= w^{-3}\underbrace{\left(\lambda\sum_{k=1}^{K}y_k^2\right)}_{V_3}+w^{-2}\underbrace{\left(-\lambda a\sum_{k=1}^{K}y_kx_k\right)}_{V_2}+w^{-1}\underbrace{(\alpha_w-1-K)}_{V_1}+\underbrace{(-\beta_w)}_{V_0}
\end{aligned}$$

□

APPENDIX B: POSITIVE DEFINITENESS OF $\boldsymbol{W_j}$

In this chapter, we show that $\boldsymbol{W_j}$ defined in Equation 8.26 is positive definite. Let us define

$$\boldsymbol{M_{j_z}} = \sum_{c=1}^{C} \mathbb{E}\left[\boldsymbol{z_{j_c}}\right] \boldsymbol{M_c} \tag{B.1}$$

$$\boldsymbol{X_j} = \sum_{k:j_k=j} \mathbb{E}\left[\boldsymbol{\chi_{i_k}} \boldsymbol{\chi_{i_k}^{\mathsf{T}}}\right] \tag{B.2}$$

$$\boldsymbol{Y_j} = \sum_{k:j_k=j} \boldsymbol{y_k} \mathbb{E}\left[\boldsymbol{\chi_{i_k}}\right]^{\mathsf{T}} \tag{B.3}$$

Then, by definitions of $\boldsymbol{V_j}$ (Equation 8.22) and $\boldsymbol{M_j}$ (Equation 8.23) we have

$$\boldsymbol{V_j^{-1}} = \boldsymbol{V_0^{-1}} + \boldsymbol{X_j} \tag{B.4}$$

$$\boldsymbol{M_j} \boldsymbol{V_j^{-1}} = \boldsymbol{M_{j_z}} \boldsymbol{V_0^{-1}} + \boldsymbol{Y_j} \tag{B.5}$$

Left multiplying Equation B.4 by $\boldsymbol{M_j}$ gives us

$$\boldsymbol{M_j} \boldsymbol{V_j^{-1}} = \boldsymbol{M_j} \boldsymbol{V_0^{-1}} + \boldsymbol{M_j} \boldsymbol{X_j} \tag{B.6}$$

Equations B.5 and B.6 give us

$$\boldsymbol{M_{j_z}} \boldsymbol{V_0^{-1}} + \boldsymbol{Y_j} = \boldsymbol{M_j} \boldsymbol{V_0^{-1}} + \boldsymbol{M_j} \boldsymbol{X_j} \tag{B.7}$$

Rearrange the terms

$$(\boldsymbol{M_j} - \boldsymbol{M_{j_z}}) \boldsymbol{V_0^{-1}} = \boldsymbol{Y_j} - \boldsymbol{M_j} \boldsymbol{X_j} \tag{B.8}$$

Take the transpose

$$\boldsymbol{V_0}^{-1}(\boldsymbol{M_j} - \boldsymbol{M_{j_z}})^\mathsf{T} = (\boldsymbol{Y_j} - \boldsymbol{M_j}\boldsymbol{X_j})^\mathsf{T} \tag{B.9}$$

Left multiply by $(\boldsymbol{M_j} - \boldsymbol{M_{j_z}})$

$$\begin{aligned}(\boldsymbol{M_j} - \boldsymbol{M_{j_z}})\boldsymbol{V_0}^{-1}(\boldsymbol{M_j} - \boldsymbol{M_{j_z}})^\mathsf{T} &= (\boldsymbol{M_j} - \boldsymbol{M_{j_z}})(\boldsymbol{Y_j} - \boldsymbol{M_j}\boldsymbol{X_j})^\mathsf{T} \qquad \text{(B.10)}\\ &= \boldsymbol{M_j}\boldsymbol{Y_j}^\mathsf{T} - \boldsymbol{M_j}\boldsymbol{X_j}\boldsymbol{M_j}^\mathsf{T} - \boldsymbol{M_{j_z}}\boldsymbol{Y_j}^\mathsf{T} + \boldsymbol{M_{j_z}}\boldsymbol{X_j}\boldsymbol{M_j}^\mathsf{T} \qquad \text{(B.11)}\end{aligned}$$

Let us define $\boldsymbol{A_j}$, $\boldsymbol{B_j}$, and $\boldsymbol{C_j}$:

$$\begin{aligned}\boldsymbol{A_j} &= (\boldsymbol{M_j} - \boldsymbol{M_{j_z}})\boldsymbol{V_0}^{-1}(\boldsymbol{M_j} - \boldsymbol{M_{j_z}})^\mathsf{T} && \text{(B.12)}\\ &= \boldsymbol{M_j}\boldsymbol{Y_j}^\mathsf{T} - \boldsymbol{M_j}\boldsymbol{X_j}\boldsymbol{M_j}^\mathsf{T} - \boldsymbol{M_{j_z}}\boldsymbol{Y_j}^\mathsf{T} + \boldsymbol{M_{j_z}}\boldsymbol{X_j}\boldsymbol{M_j}^\mathsf{T} && \text{(B.13)}\\ \boldsymbol{B_j} &= \sum_{k:j_k=j} \mathbb{E}\left[(\boldsymbol{y_k} - \boldsymbol{M_j}\boldsymbol{\chi_{i_k}})(\boldsymbol{y_k} - \boldsymbol{M_j}\boldsymbol{\chi_{i_k}})^\mathsf{T}\right] && \text{(B.14)}\\ &= \sum_{k:j_k=j} \boldsymbol{y_k}\boldsymbol{y_k}^\mathsf{T} - \boldsymbol{Y_j}\boldsymbol{M_j}^\mathsf{T} - \boldsymbol{M_j}\boldsymbol{Y_j}^\mathsf{T} + \boldsymbol{M_j}\boldsymbol{X_j}\boldsymbol{M_j}^\mathsf{T} && \text{(B.15)}\\ \boldsymbol{C_j} &= \sum_{c=1}^{C} \mathbb{E}\left[\boldsymbol{z_{j_c}}\right](\boldsymbol{M_c} - \boldsymbol{M_{j_z}})\boldsymbol{V_0}^{-1}(\boldsymbol{M_c} - \boldsymbol{M_{j_z}})^\mathsf{T} && \text{(B.16)}\\ &= \sum_{c=1}^{C} \boldsymbol{z_{j_c}}\boldsymbol{M_c}\boldsymbol{V_0}^{-1}\boldsymbol{M_c}^\mathsf{T} - \boldsymbol{M_{j_z}}\boldsymbol{V_0}^{-1}\boldsymbol{M_{j_z}}^\mathsf{T} && \text{(B.17)}\end{aligned}$$

By summing Equations B.13 and B.15 side by side, we get

$$\begin{aligned}\boldsymbol{A_j} + \boldsymbol{B_j} &= \sum_{k:j_k=j} \boldsymbol{y_k}\boldsymbol{y_k}^\mathsf{T} - \boldsymbol{Y_j}\boldsymbol{M_j}^\mathsf{T} - \boldsymbol{M_{j_z}}\boldsymbol{Y_j}^\mathsf{T} + \boldsymbol{M_{j_z}}\boldsymbol{X_j}\boldsymbol{M_j}^\mathsf{T} && \text{(B.18)}\\ &= \sum_{k:j_k=j} \boldsymbol{y_k}\boldsymbol{y_k}^\mathsf{T} - \boldsymbol{Y_j}\boldsymbol{M_j}^\mathsf{T} - \boldsymbol{M_{j_z}}\boldsymbol{Y_j}^\mathsf{T} + \boldsymbol{M_{j_z}}(\boldsymbol{V_j}^{-1} - \boldsymbol{V_0}^{-1})\boldsymbol{M_j}^\mathsf{T} && \text{(B.19)}\\ &= \sum_{k:j_k=j} \boldsymbol{y_k}\boldsymbol{y_k}^\mathsf{T} - \boldsymbol{Y_j}\boldsymbol{M_j}^\mathsf{T} - \boldsymbol{M_{j_z}}\boldsymbol{Y_j}^\mathsf{T} + \boldsymbol{M_{j_z}}\boldsymbol{V_j}^{-1}\boldsymbol{M_j}^\mathsf{T} - \boldsymbol{M_{j_z}}\boldsymbol{V_0}^{-1}\boldsymbol{M_j}^\mathsf{T} && \text{(B.20)}\end{aligned}$$

$$= \sum_{k:j_k=j} \boldsymbol{y_k y_k}^\mathsf{T} - \boldsymbol{M_j V_j}^{-1} \boldsymbol{M_j}^\mathsf{T} - \boldsymbol{M_{j_z} Y_j}^\mathsf{T} + \boldsymbol{M_{j_z} V_j}^{-1} \boldsymbol{M_j}^\mathsf{T} \tag{B.21}$$

$$= \sum_{k:j_k=j} \boldsymbol{y_k y_k}^\mathsf{T} - \boldsymbol{M_j V_j}^{-1} \boldsymbol{M_j}^\mathsf{T} + \boldsymbol{M_{j_z} V_0}^{-1} \boldsymbol{M_{j_z}}^\mathsf{T} \tag{B.22}$$

By adding $\boldsymbol{C_j}$ and $\boldsymbol{W_0}^{-1}$ to $\boldsymbol{A_j} + \boldsymbol{B_j}$, we end up with

$$\boldsymbol{W_0}^{-1} + \boldsymbol{A_j} + \boldsymbol{B_j} + \boldsymbol{C_j} = \underbrace{\boldsymbol{W_0}^{-1} + \sum_{k:j_k=j} \boldsymbol{y_k y_k}^\mathsf{T} + \sum_{c=1}^{C} \mathbb{E}\left[\boldsymbol{z_{j_c}}\right] \boldsymbol{M_c V_0}^{-1} \boldsymbol{M_c}^\mathsf{T} - \boldsymbol{M_j V_j}^{-1} \boldsymbol{M_j}^\mathsf{T}}_{\boldsymbol{W_j}^{-1} \text{(Equation 8.26)}} \tag{B.23}$$

Therefore,

$$\boldsymbol{W_j}^{-1} = \boldsymbol{W_0}^{-1} + \sum_{k:j_k=j} \boldsymbol{y_k y_k}^\mathsf{T} + \sum_{c=1}^{C} \mathbb{E}\left[\boldsymbol{z_{j_c}}\right] \boldsymbol{M_c V_0}^{-1} \boldsymbol{M_c}^\mathsf{T} - \boldsymbol{M_j V_j}^{-1} \boldsymbol{M_j}^\mathsf{T} \tag{B.24}$$

$$\begin{aligned} &= \boldsymbol{W_0}^{-1} + \sum_{c=1}^{C} \mathbb{E}\left[\boldsymbol{z_{j_c}}\right] (\boldsymbol{M_c} - \boldsymbol{M_{j_z}}) \boldsymbol{V_0}^{-1} (\boldsymbol{M_c} - \boldsymbol{M_{j_z}})^\mathsf{T} \\ &\quad + (\boldsymbol{M_j} - \boldsymbol{M_{j_z}}) \boldsymbol{V_0}^{-1} (\boldsymbol{M_j} - \boldsymbol{M_{j_z}})^\mathsf{T} + \sum_{k:j_k=j} \mathbb{E}\left[(\boldsymbol{y_k} - \boldsymbol{M_j \chi_{i_k}})(\boldsymbol{y_k} - \boldsymbol{M_j \chi_{i_k}})^\mathsf{T}\right] \end{aligned} \tag{B.25}$$

Since $\boldsymbol{W_0}^{-1}$ is positive definite and the remaining summands of the equation above are positive semidefinite, $\boldsymbol{W_j}$ is also positive definite. □