COMPUTATIONAL MODELS FOR MULTIDIMENSIONAL ANNOTATIONS OF AFFECT

by

Anil Ramakrishna

A Dissertation Presented to the
FACULTY OF THE GRADUATE SCHOOL
UNIVERSITY OF SOUTHERN CALIFORNIA
In Partial Fulfillment of the
Requirements for the Degree
DOCTOR OF PHILOSOPHY
(Computer Science)

December 2019

To all the people who inspire me to keep learning;
past, present, and future.

Acknowledgments

I have been fortunate to have met some of the brightest minds of my generation. For this, I am indebted to Dr. Shrikanth Narayanan who gave me the opportunity, mentorship and freedom to learn and grow as a researcher as well as an individual.

I would like to thank Dr. Aiichiro Nakano, Dr. Morteza Dehghani, Dr. Panayiotis Georgiou and Dr. Jonathan Gratch for serving on my dissertation and qualifier committees. All the suggestions I received from you were critical to the formation of this dissertation.

I would also like to thank my colleagues from the Signal Analysis and Interpretation Lab (SAIL) at USC for being an important part of my journey. The countless discussions and debates I have had with so many of you were instrumental for my work.

Finally, I would like to thank my parents, siblings and friends for standing by me through the good and bad times. This would not have been possible without your support.

Contents

List of Figures

List of Tables

Abstract

Affect is an integral aspect of human psychology which regulates all our interactions with external stimuli. It is highly subjective, with different stimuli leading to different affective responses in people due to varying personal and cultural artifacts. Computational modeling of affect is an important problem in Artificial Intelligence, which often involves supervised training of models using a large number of labeled data points. However, training labels are difficult to obtain due to the inherent subjectivity of affective dimensions. The most common approach to obtain the training labels is to collect subjective opinions or ratings from expert or naive annotators, followed by a suitable aggregation of the ratings.

In this dissertation, we will present our contributions towards building computational models for aggregating the subjective ratings of affect, specifically in the multidimensional setting. We propose latent variable models to capture annotator behaviors using additive Gaussian noise and matrix factorization models, which show improved performance in estimating the dimensions of interest. We then apply our matrix factorization model to the task of sentence level estimation of psycholinguistic normatives. Finally, we set up future work in estimating agreement on multidimensional annotations.

Chapter 1

Introduction

Affect is an abstract entity which is said to manifest prior to the realm of personal awareness or consciousness [2]. According to [3], it is fundamental in nature and subsumes several other related concepts such as sentiment, feelings and emotion, along with higher order mental constructs such as humor and mood. An important characteristic of affect and other related concepts is their inherent subjectivity. For example, a given image may evoke different emotions in people depending on their backgrounds. Similarly, different people may react differently to humorous situations. This subjectivity often leads to challenges in building models for recognizing affect.

Modeling affect is an important problem in Artificial Intelligence (AI). Incorporating affect can enrich the quality of interactions with AI agents, and it is relevant in all of the modalities commonly encountered in AI such as speech, vision and language. Modeling of affect spans the interdisciplinary field of Affective Computing (AC), which includes tasks such as emotion recognition, sentiment analysis and opinion mining, along with recognizing higher order constructs such as humor and mood. Typical approaches used in these tasks involve training supervised machine learning models, which assumes the availability of a dataset with training labels. However, these labels are not easy to obtain thanks to the subjectivity of affective dimensions. Further, we may not always have a clearly identifiable ground truth unlike typical machine learning tasks. For example, when developing an AI system to identify physical attributes of a person (such as race, gender, etc.) from images, often we can easily and reliably identify these attributes without ambiguity. However, affective dimensions such as emotion are social constructs, with culture playing an important role in their recognition [4], because of which there may not be an objective ground truth.

Commonly used strategies to collect training labels for the affective dimensions include: (i) use an approximate proxy from the corpus to identify the labels; for ex, if we are building models to predict humor, laughter cues may be used as a proxy, or (ii) combine noisy labels using an annotation fusion model.

As a case study for using a proxy to identify the training labels, we explored the computational modeling of humor from conversations in psychotherapy sessions[1]. We used shared occurrence of laughter between the client and the therapist as a proxy for occurrence of humorous utterances. To capture context, we used a hierarchical 2 layer LSTM network and showed improved performance in recognizing humor compared to a standard baseline. However, similar attempts to use canned laughter from television sitcoms as a

[1] Anil Ramakrishna, Timothy Greer, David Atkins, Shrikanth Narayanan, Computational modeling of conversational humor in psychotherapy, in: Proceedings of Interspeech, Hyderabad, India, 2018.

proxy to humor failed due to low agreement on which utterances were humorous, between human annotations and those followed by the canned laughter. This highlights the limitations of using a proxy as we may not always have access to a reliable approximation to the label of interest. Further, identifying a suitable proxy assumes domain knowledge which may not always be easily accessible.

An alternate approach to obtain training labels is to collect noisy judgments from human annotators who may be trained experts (such as medical professionals working on diagnostics data) or untrained workers from crowdsourcing platforms such as Amazon Mechanical Turk[2] (MTurk) and Crowdflower[3]. Given noisy labels from such annotators, the typical approach is to aggregate them to obtain the label of interest. Common aggregation strategies include majority voting or simple averaging, but they assume uniform reliability among the annotators which may not be true with crowdsourcing platforms. To address this, several authors have developed annotation fusion models to capture the behavior of individual annotators to improve the quality of estimated labels. However, most existing works model the annotation dimensions individually even in settings where we collect annotator ratings on multiple dimensions. For example, while collecting annotations on affective dimensions, it is common to collect ratings on dimensions such as valence, arousal and dominance and it maybe beneficial to model the annotations jointly while aggregating them. In this dissertation, we will explore this hypothesis and propose models to perform annotation fusion, which make use of correlations between the different annotation dimensions.

[2] www.mturk.com

[3] www.crowdflower.com

1.1 Contributions

The specific contributions made in this dissertation are as follows.

- We propose two multidimensional annotation fusion models with latent ground truth vectors to capture relationships between the dimensions. In both models, the annotator parameters and the ground truth vectors are estimated jointly using the Expectation Maximization [5] algorithm.
  - The first model assumes additive Gaussian noise for the annotators' distortion function.
  - The second model assumes a matrix factorization structure for the distortion function.
- We develop a novel strategy to estimate psycholinguistic normatives at sentence level by making use of the matrix factorization based annotation fusion model.

1.2 Organization

The overall organization of this dissertation is as follows:

- In Chapter 2, we introduce the problem of annotation fusion and discuss prior works in this domain along with their weaknesses. We also motivate the need for multidimensional annotation fusion.
- In Chapter 3, we present our first model for multidimensional annotation fusion which uses additive Gaussian noise.

- In Chapter 4, we present the matrix factorization based model to capture multidimensional annotations. Derivations for the model are presented in Appendix A.

- In Chapter 5, we apply the annotation fusion model described in Chapter 4 to the task of estimating sentence level psycholinguistic norms.

- We conclude in Chapter 6 and highlight future directions in the estimation of agreement for multidimensional annotations.

Chapter 2

Multidimensional Annotation Fusion: Preliminaries

2.1 Introduction

Crowdsourcing is a popular tool used in collecting human judgments on affective constructs such as emotion and engagement. Typical examples included annotations of images or video clips with categorical emotion labels or with continuous dimensions such as *valence* or *arousal*. Online platforms such as Amazon Mechanical Turk (Mturk) have recently risen in popularity owing to their inexpensive label costs and also their ability to scale efficiently.

Crowdsourcing is also a popular approach in collecting labels for use in the training of supervised machine learning algorithms. Such labels are typically obtained from domain experts, which can be slow and expensive. For example, in the medical domain, it is often expensive to collect diagnosis information given laboratory tests since this requires judgments of trained professionals. On the other hand, unlabeled patient data may be easily available. Crowdsourcing has been particularly successful in such settings with easy availability of unlabeled data instances since we can collect a large number of annotations from untrained and inexpensive workers over the Internet, which when combined together may be comparable or even better than expert annotations [6].

A typical crowdsourcing setting involves collecting annotations from a large number of workers and hence there is a need to robustly combine them to estimate the ground truth. The most common approach for this is to take simple averages for continuous labels or perform majority voting for categorical labels. However, this assumes uniform competency across all the workers which is not always guaranteed or justified. Several alternative approaches have been proposed to address this challenge, each with a specific structure to the function modeling the annotators' behavior. In practice, it is common to collect annotations on multiple questions for each data instance being labeled in order to reduce costs or annotators' mental load or even to improve annotation accuracy. For example, while collecting emotion annotations for a given data instance (such as a single image or video segment), collecting labels on dimensions such as valence or arousal together (concurrently or one after another) may be preferred over collecting valence annotations for all instances followed by arousal annotations.

Such a joint annotation task may entail *task specific* or *annotator specific* dependencies between the annotated dimensions. In the emotion annotation example, task specific dependencies may occur due to inherent correlations between the valence and arousal dimensions depending on the experimental setup. Annotator specific dependencies may occur due to a given annotator's (possibly incorrect or incomplete) understanding of the annotation dimensions. Hence it is of relevance to jointly model the different annotation dimensions. However, most state of the art models in annotation fusion combine the annotations by modeling the different dimensions independently. The focus of this dissertation is to highlight the benefits of modeling them jointly. Joint modeling of the annotation dimensions may result in more accurate estimates of the ground truth as well as giving a better picture of the annotators' behavior. In this chapter, we will present prior work in the domain of annotation fusion and motivate the need for multidimensional annotation fusion models.

2.2 Keywords

We list a few important key words and their definitions below.

- Annotators These are workers from crowdsourcing platforms such as Mturk who provide their judgments on the subjective construct under discussion.

- Annotations These are the noisy judgments we obtain from the annotators. We use the terms ratings and annotations interchangeably.

- Ground truth The objective value of labels for cases in which they can be clearly and unambiguously identified (for example, height of people).

- Reference labels In cases when there are no unambiguous ground truth values (for example, emotion or humor), we aggregate expert opinions to obtain a reference label against which predictions of our fusion models are compared.

- Data instance The individual data point for which the subjective ratings are being collected

Since the annotation fusion models are applicable while aggregating annotations from tasks both with or without a well defined ground truth, in the following sections and chapters, we overload the term ground truth and use it when we refer to the hidden variable of interest to be estimated by the annotation fusion models, even in problems without a well defined ground truth. In such cases, the predicted estimates from the models are compared with reference labels obtained from experts instead.

2.3 Related work

Several authors, most notably [6], assert the benefits of aggregating opinions from many people which is often believed to be better than those from a small number of experts, under certain conditions. Often referred to as the *wisdom of crowds*, this approach has been remarkably popular in recent times, specially in fields such as psychology where a ground truth may not be easily accessible or may not exist. This popularity can be largely attributed to online crowdsourcing platforms such as Mturk that connect researchers with low cost workers from around the globe. Along with cost, scalability of annotations is another major appeal with such tools leading to their use in machine learning in large scale labeling of data instances such as images [7], audio/video clips [8] and text snippets [9].

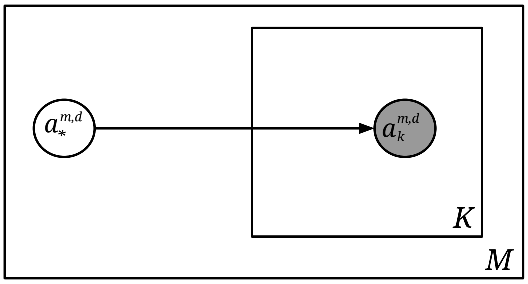


Figure 2.1: Plate notation for a basic annotation model. $a_*^{m,d}$ is the latent ground truth for the given data point (for the d^{th} question) and $a_k^{m,d}$ is the rating provided by the k^{th} annotator.

Figure 2.1 shows a common setting in the crowdsourcing paradigm. For each data point m, annotator k provides a noisy label $a_k^{m,d}$ which depends on the ground truth $a_*^{m,d}$ where d is the dimension being annotated. Since we collect several annotations for each data point, we need to aggregate them to estimate the unknown ground truth. The most common technique used in aggregating these opinions is to take the average value in case of numeric labels or perform majority voting in the case of categorical labels as shown in Equation 2.1.

$$a_*^{m,d} = \underset{j}{\text{argmax}} \sum_k \mathbb{1}\{a_k^{m,d} == j\} \tag{2.1}$$

where, $\mathbb{1}\{\}$ is the indicator function

While simple and easy to implement, this approach assumes consistent reliability among the different annotators which seems unreasonable, especially in online platforms such as Mturk. To avoid this, several approaches have been suggested that account for annotator reliability in estimating the ground truth. We explain a few in detail below.

Early efforts to capture reliability in annotation modeling [10], [11] assumed specific structure to the functions modeled by each annotator. Given a set of annotations $a_k^{m,d}$ along with the corresponding function parameters, the ground truth is estimated using the MAP estimator

$$a_*^{m,d} = \operatorname*{argmax}_j \sum_k \log p(a_k^{m,d} | a_*^{m,d} = j) + \log p(a_*^{m,d} = j) \tag{2.2}$$

where $p(a_*^{m,d})$ is the prior probability of ground truth.

In [10], the categorical ground truth label $a_*^{m,d} = i$ is modified probabilistically by annotator k using a stochastic matrix Π_k as shown in Equation 2.3 in which each row is a multinomial conditional distribution given the ground truth.

$$P(a_k^{m,d} = j | a_*^{m,d} = i) = \pi_{ij}^k \tag{2.3}$$

Given annotations from K different annotators, their parameters Π_k and prior distribution of labels $p_j = P(a_*^{m,d} = j)$, the ground truth is estimated using MAP estimation as before.

$$a_*^{m,d} = \operatorname*{argmax}_j \sum_k \log \pi_{j(a_k^{m,d})} + \log p_j \tag{2.4}$$

The above expression makes a conditional independence assumption for annotations given the ground truth label. Since we do not typically have the annotator parameters Π^k, these are estimated using the EM algorithm.

Figure 2.2 shows an extension of the model in Figure 2.1 in which we learn a predictor (classifier/regression model) for the ground truth jointly

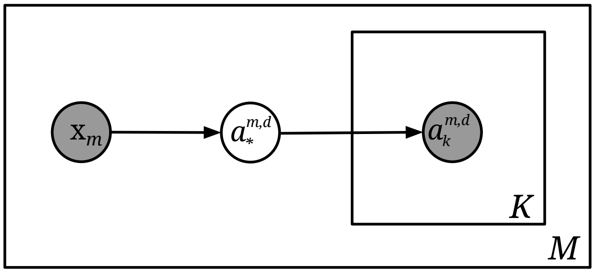


Figure 2.2: Annotation model proposed by [1] with a jointly learned predictor. $\mathbf{x}_m$ is the set of features for the m^{th} data point; $a_*^{m,d}$ is the d^{th} dimension of the latent ground truth which is modeled as a function of $\mathbf{x}_m$; $a_k^{m,d}$ is the rating provided by the k^{th} annotator.

with annotator parameters. Such a predictor may be used to obtain ground truth for new data points. This strategy of jointly modeling the annotator functions as well as the ground truth predictor has been shown to have better performance when compared to classifiers trained independently of the estimated ground truth [1]. The ground truth estimate in this model is given by

$$a_*^{m,d} = \underset{a_*^{m,d}}{\operatorname{argmax}} \sum_k \log p(a_k^{m,d} | a_*^{m,d}) + \log p(a_*^{m,d} | \mathbf{x}_m) \tag{2.5}$$

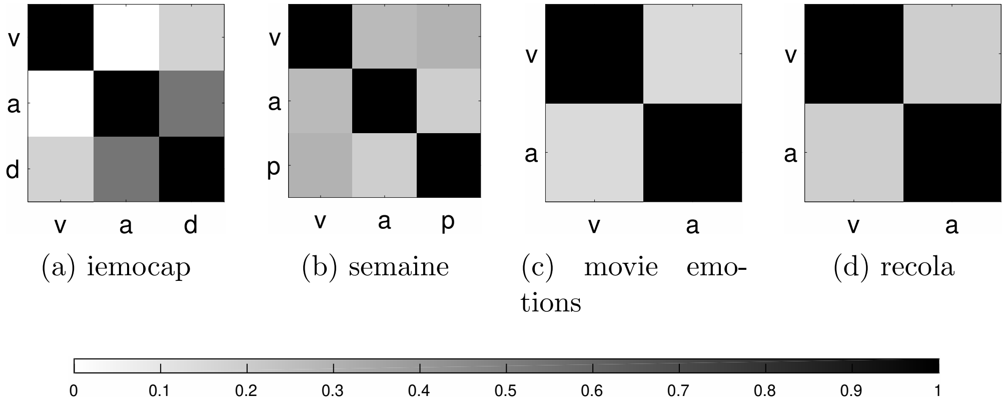
Recently several additional extensions have been proposed to Figure 2.2; For example in [12], the authors assume varying regions of annotator expertise in the data feature space and account for this using different probabilities for label confusion for each region. The authors show that this leads to a better estimation of annotator reliability and ground truth.

The models described so far were designed for annotation tasks in which the task is to rate some global property of the data point. For example, in image based emotion annotation, the task may be to provide annotations on dimensions such as valence and arousal conveyed by each image. However, human interactions may often involve continuous variations of these dimensions over time [13] which are captured using time series annotations from audio/video clips. In this context, the previous models are applicable only if annotations from each frame are treated independently. However, this entails several unrealistic assumptions such as independence between frames, zero lag in the annotators and synchronized response in the annotators to the underlying stimulus.

Several works have been proposed to capture the underlying reaction lag in the annotators. [14] proposed a generalization of Probabilistic Canonical Correlation Analysis (PCCA) [15] named Dynamic PCCA which captures temporal dependencies of the shared ground truth space in a generative setting. They further extend this model by incorporating a latent time warping process to implicitly handle the reaction lags in annotators. This work is extended in [16] where they also jointly learn a function to capture dependence of the latent ground truth signal with the data points' features in both generative and discriminative settings similar to the setting of [1]. [17] address the reaction lag by explicitly finding the time shift that maximizes the mutual information between expressive behaviors and the annotations. [18] generalize the work of [17] by using a linear time invariant (LTI) filter which can also handle any bias or scaling the annotators may introduce.

More recent works in annotation fusion include [19] in which the authors propose a variant of the model in Figure 2.1 with various annotator functions to capture four specific types of annotator behavior. [20] describe a mechanism named approval voting that allows annotators to provide multiple answers instead of one for instances where they are not confident. [21] use repeated sampling for opinions from annotators over the same data instances to increase reliability in annotations.

Most of the models described above focus on combining annotations on each dimension separately. The model proposed in [16] can indeed be generalized to combine the different annotation dimensions together but that is not the focus of their work and as such they do not evaluate on this task. However, in many practical applications, annotation tasks are multi-dimensional. For example, while collecting emotion ratings it is routine to collect annotations on valence, arousal, dominance and other related dimensions. In these cases, it may be beneficial to model the different dimensions together since they may be closely related. Further, there may be dependencies between the internal definitions the annotators hold for the annotation dimensions. For example, while annotating emotional dimensions, a given annotator may associate certain valence values with only a certain range of arousal. It is therefore of relevance to model such annotator specific relationships between the different dimensions as part of the annotator distortion function and predictor modeling paradigm. In this dissertation, we address this gap by proposing latent variable models for multidimensional annotation fusion. We motivate this problem further in the next section.



(a) iemocap (b) semaine (c) movie emotions (d) recola

Figure 2.3: Correlation heatmaps for annotations from a representative sample of emotion annotated datasets; v - valence, a - arousal, d - dominance, p - power

2.4 Motivation

To examine the relationships between annotation dimensions, we created a plot of absolute values of correlation scores between annotation dimensions from four commonly studied emotion corpora in Figure 2.3: iemocap [22], semaine [23], recola [24] and the movie emotion corpus from [25]. Each of these corpora include annotations over emotion dimensions such as valence, arousal, dominance and power. For the iemocap corpus we used global annotations while the others include time series annotations of the affective dimensions from videos. In each case, the correlations were computed between concatenated annotation values from annotators who provide ratings on all the dimensions.

As is evident, in almost all cases, the annotation dimensions exhibit nonzero correlations, highlighting the need for fusion models that take into account such correlations. The models we propose in the next chapters are aimed at capturing this form of dependency. We attribute the inconsistent correlations between the dimensions across corpora to varying underlying affective narratives as well as differences in perceptions and biases introduced by individual annotators themselves.

Chapter 3

Additive Gaussian noise model

3.1 Introduction

In this chapter, we will present our first model to capture multidimensional annotations which assumes that each annotator's distortion function uses an additive Gaussian noise vector to capture the relationship between annotation dimensions. Similar to the models described in Section 2.3, we assume that the ground truth vector for each data point is hidden while the feature vector corresponding to each file and the annotation vector are available. With this formulation, we use the EM algorithm to estimate the model parameters.

The chapter is organized as follows: in Section 3.2, we describe the proposed model and provide equations for parameter estimation using EM algorithm. We describe the data used to evaluate the model in Section 3.3, experiments in sections 3.4, and results 3.5 before concluding in Section 3.6.

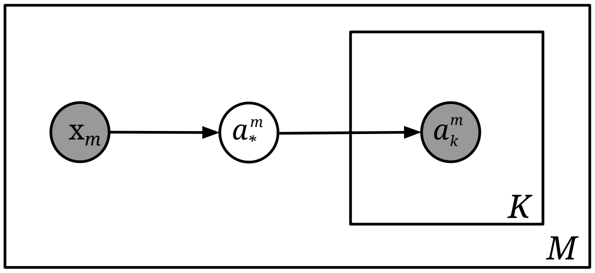


Figure 3.1: Graphical model representation for the proposed model. $\mathbf{x}_m$ is the set of features for the m^{th} instance, $\mathbf{a}_*^m$ is the latent ground truth and $\mathbf{a}_k^m$ is the rating provided by the k^{th} annotator for that instance. $\mathbf{x}_m$ and $\mathbf{a}_k^m$ are observed variables, $\mathbf{a}_*^m$ is latent.

3.2 Model

Consider a set of M data points with features $\{\mathbf{x}_1, .., \mathbf{x}_m\}$; $\mathbf{x}_m$ being the feature vector corresponding to the m^{th} point. Each data point is associated with a D dimensional ground truth vector for which ratings from several annotators are pooled. In this work, we assume that each datapoint is annotated by a subset of K annotators. This is a more general setting than assuming that the ratings are available from every annotator (as assumed in [1]), and is often the case with data collection over online platforms such as Mturk. We represent the set of ratings for the m^{th} data point by a set $\mathbf{A}_m$. For example, if annotators 1, 2 and 5 provided their ratings (out of K annotators), $\mathbf{A}_m$ would be the set $\{\mathbf{a}_1^m, \mathbf{a}_2^m, \mathbf{a}_5^m\}$, where $\mathbf{a}_k^m$ is the multidimensional rating from the k^{th} annotator. The vector $\mathbf{a}_k^m$ is a D-dimensional vector, represented as $\{a_k^{m,1}, .., a_k^{m,d}, .., a_k^{m,D}\}$, where $a_k^{m,d}$ is the rating by the k^{th} annotator for the d^{th} dimension corresponding to the data point m. Armed with this notation, we train annotation fusion model shown as a graphical model in Figure 3.1. This model is inspired from the works of Raykar et al. [1] and Gupta et al. [18]. The model assumes that there exists a latent ground truth $\mathbf{a}_*^m$ (also of dimensionality D), which is conditioned on the data features. The relationship between the features and $\mathbf{a}_*^m$ is captured by the function $f(\mathbf{x}_m|\boldsymbol{\theta})$, with parameter $\boldsymbol{\theta}$. We assume f to be an affine projection of the feature vectors as shown in Equation 3.1, with $\boldsymbol{\theta}$ being the projection matrix.

$$\mathbf{a}_*^m = f(\mathbf{x}_m|\boldsymbol{\theta}) = \boldsymbol{\theta}^T \begin{bmatrix} \mathbf{x}_m \\ 1 \end{bmatrix} \tag{3.1}$$

The model further assumes that each annotator's ratings are noisy modifications of the ground truth $\mathbf{a}_*^m$. We assume these modifications to be the addition of an D-dimensional Gaussian noise with distribution $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, as shown in Equation 3.2. $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ represent the mean and co-variance matrix of this distribution, respectively.

$$\mathbf{a}_k^m = \mathbf{a}_*^m + \eta_k, \text{ where } \eta_k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{3.2}$$

Model training

We estimate the model parameters by maximizing the data log-likelihood. Since the model contains a latent variable (the ground truth $\mathbf{a}_*^m$), we adopt the Expectation Maximization algorithm [5] widely used for similar settings. During model training, our objective is to estimate the model parameters $\boldsymbol{\Phi} = \{\boldsymbol{\theta}, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, .., \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K\}$ that maximize the log-likelihood $\mathcal{L}$ of the observed annotator ratings, given the features. Assuming independent data points, $\mathcal{L}$ is given by

$$\mathcal{L} = \log \prod_{n=1}^{N} p(\mathbf{A}_m | \mathbf{x}_m, \mathbf{\Phi}) = \sum_{n=1}^{N} \log p(\mathbf{A}_m | \mathbf{x}_m, \mathbf{\Phi}) \tag{3.3}$$

The EM algorithm iteratively performs an E-step followed by an M-step. A detailed derivation of these steps for the EM algorithm can be referred from various resources as [5], [10] and [12]. We specifically refer the reader to the EM algorithm derivation in [18] for a fusion model similar to the one presented in this chapter. The authors in [18] perform a hard version of EM algorithm where in the E-step an estimate of ground truth $\mathbf{a}_*^m$ is computed. This is followed by parameter update in the M-step based on the estimated $\mathbf{a}_*^m$. Popular methods such as Viterbi training [26] and K-means clustering [27] are variants of the hard EM algorithm for training Hidden Markov Models and clustering, respectively. Borrowing formulations from the aforementioned research studies, we summarize the E and M steps for obtaining the parameters for the graphical model shown in Figure 3.1.

EM algorithm

The EM algorithm involves iteratively executing the Expectation and Maximization steps listed below.

Initialize the model parameters $\mathbf{\Phi}$

E-step We estimate the ground truth $\mathbf{a}_*^m$ $\forall m = 1..m$ by solving the optimization equation shown below. $||.||_2$ represents the l^2-norm.

$$\mathbf{a}_*^m = \underset{\mathbf{a}_*^m}{\operatorname{argmin}} \sum_{\substack{k=\text{Set of}\\ \text{annotators in } \mathbf{A}_m}} \left|\left|\boldsymbol{\Sigma}_k^{-\frac{1}{2}}(\mathbf{a}_k^m - \mathbf{a}_*^m - \boldsymbol{\mu}_k)\right|\right|_2^2 + \left|\left|\mathbf{a}_*^m - \boldsymbol{\theta}^T \begin{bmatrix} \mathbf{x}_m \\ 1 \end{bmatrix} \right|\right|_2^2 \tag{3.4}$$

M-step Given $\mathbf{a}_*^m$, we estimate the model parameters $\boldsymbol{\Phi}$ using the following equations. M_k is the number of datapoints annotated by annotator k.

$$\boldsymbol{\mu}_k = \frac{1}{M_k} \sum_{\substack{m'=\text{Set of datapoints}\\ \text{rated by annotator } k}} \left(\mathbf{a}_k^{m'} - \mathbf{a}_*^{m'}\right) \tag{3.5}$$

$$\boldsymbol{\Sigma}_k = \frac{1}{M_k - 1} \sum_{\substack{m'=\text{Set of datapoints}\\ \text{rated by annotator } k}} \left((\mathbf{a}_k^{m'} - \mathbf{a}_*^{m'} - \boldsymbol{\mu}_k) * (\mathbf{a}_k^{m'} - \mathbf{a}_*^{m'} - \boldsymbol{\mu}_k)^T\right) \tag{3.6}$$

$$\boldsymbol{\theta} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \sum_m \left(\left|\left|\mathbf{a}_*^m - \boldsymbol{\theta}^T \begin{bmatrix} \mathbf{x}_m \\ 1 \end{bmatrix} \right|\right|_2^2\right) \tag{3.7}$$

Termination We run the algorithm until convergence of data log-likelihood $\mathcal{L}$.

Model testing

To evaluate our model, we use the task of predicting back the annotator rating given parameter estimates. We show later how this task can also be used to address the issue of annotation cost and reducing cognitive load on the annotator by partial prediction of the ratings. Note that though our model estimates the latent values for above dimensions, it is hard to evaluate the quality of these estimates as they are unobserved and often subjective in the dataset of interest (as is true for several datasets in the Behavioral Signal Processing domain [28]). In order to predict the rating for the m^{th} file from the k^{th} annotator, we first predict $\mathbf{a}_*^m$ using Equation 3.1 and then add the mean $\boldsymbol{\mu}_k$ of the noise distribution $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, corresponding to the k^{th} annotator. Note that adding $\boldsymbol{\mu}_k$ to $\mathbf{a}_*^m$ provides the maximum likelihood estimate of $\mathbf{a}_k^m$ thanks to Equation 3.2 and the Gaussian noise assumption [29].

We use Mean Squared Error (MSE) computed per dimension, averaged over all the annotators as our evaluation metric. For the dimension d (out of D dimensions), we compute the MSE $\mathcal{E}_d$ as shown in Equation 3.8. I_{mk} is an indicator variable marking if the k^{th} rater annotated the data point m (Equation 3.9). $a_k^{m,d}$ is the true rating obtained from the rater k on data point m and $\hat{a}_k^{m,d}$ is the model prediction.

$$\mathcal{E}_d = \frac{\sum_{m=1}^{M} \sum_{k=1}^{K} I_{mk} (a_k^{m,d} - \hat{a}_k^{m,d})^2}{\sum_{m=1}^{M} \sum_{k=1}^{K} I_{mk}} \tag{3.8}$$

where

$$I_{mk} = \begin{cases} 1 & \text{if annotator } k \text{ annotates data point } m \\ 0 & \text{otherwise} \end{cases} \tag{3.9}$$

We choose this metric as it allows for evaluation on each dimension independently. Such a metric is particularly relevant in the Behavioral Signal Processing domain where an evaluation on each dimension of rating is desired. In the next section, we describe the dataset used in this study.

3.3 Data

We evaluate our model using the SafariBob dataset [30]. The dataset contains multimodal recordings of children watching and imitating video stimuli, each corresponding to a different emotional expression. We extract audio clips from each of these recordings which are annotated over M-Turk. For the purpose of our experiments, we use a set of 244 audio clips (each approximately 25-30 seconds) which were rated over M-Turk by a set of 124 naive annotators. The annotators provide a four dimensional rating ($D = 4$), providing their judgments on expressiveness, naturalness, goodness of pronunciation and engagement of the speaker in each audio clip. The numeric values of these attributes lies in the range of 1 to 5. Each utterance in the data set is annotated by a subset of 15 (out of 124) annotators. This setting is subsumed by the model proposed in Section 3.2. For further details on the dataset, we refer the reader to [30].

Feature set

We use various statistical functionals computed over a set of acoustic-prosodic properties of the utterance resulting in a set of 474 features ($\mathbf{x}_m$) per file. These features are inspired by prior works in speech emotion recognition [31, 32]. The list of the signals and their statistical functionals used as features is shown in Table 3.1. In the next section, we describe our experimental setup including the baseline model and test different variants of the model described in Section 3.2.

Acoustic-prosodic signals	Audio intensity, mel-frequency band, mel-frequency cepstral coefficients and pitch
Statistical functionals	Mean, median, standard deviation, range, skewness and kurtosis

Table 3.1: Acoustic prosodic signals and their statistical functionals used as features $\mathbf{x}_m$ in this study.

3.4 Experiments

Based on the approach described in Section 3.2, we train models with different assumptions. Since our goal in these experiments is to predict the annotator ratings, we initially train a baseline system individually modeling every annotator. This is followed by various modifications of the proposed model to predict annotator ratings. We discuss these models in detail below.

3.4.1 Baseline: Individual annotator modeling

For the baseline, we train individual models for each annotator, instead of the joint model described in Section 3.2. We use an affine projection scheme, for which the relationship between the k^{th} annotator's ratings and features is shown in Equation 3.10. $\boldsymbol{\theta}_k$ is the projection matrix for the k^{th} annotator. The parameter $\boldsymbol{\theta}_k$ is obtained using minimum mean squared error criterion on the training set, using data points that the annotator rated.

$$\mathbf{a}_k^m = f(\mathbf{x}_m | \boldsymbol{\theta}_k) = \boldsymbol{\theta}_k^T \begin{bmatrix} \mathbf{x}_m \\ 1 \end{bmatrix} \tag{3.10}$$

3.4.2 Joint annotator - Independent rating (Joint-Ind) modeling

In this scheme, we train the joint annotator model assuming independence between each dimension in the multidimensional rating. This is achieved by training a separate model for each annotator dimension entry $a_k^{m,d}$. The training procedure is same as presented in Section 3.2, with the special case of ratings being scalar. Consequently, we end up with $D = 4$ different models, one for each dimension. This model acts as a strong baseline and can help shed light on the benefits of modeling the dimensions jointly.

3.4.3 Joint annotator - Joint rating (Joint-Joint) modeling

We next model both the annotators and the ratings jointly as described in Section 3.2. For each annotator we end up with multidimensional parameters $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ spanning all four dimensions, which are in turn used to predict the annotator's rating for each data instance. We expect this model to capture any joint relationship between the different dimensions in the ratings, which was not modeled by the previous Joint-Ind model.

3.4.4 Joint annotator - Conditional rating (Joint-Cond) modeling

The Joint-Cond model is an extension of the model described in Section 3.4.3. In this scheme, we assume partial availability of annotator ratings on

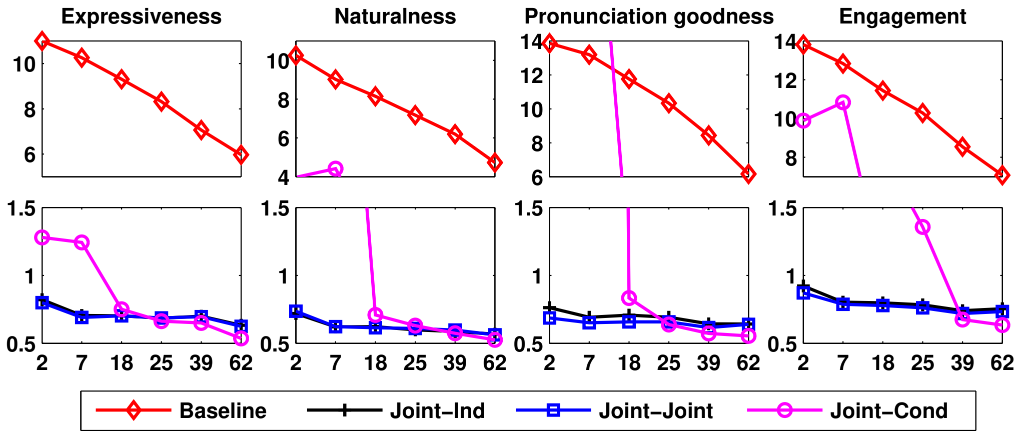


Figure 3.2: MSE $\mathcal{E}_d$ for the four (baseline, Joint-Ind, Joint-Joint and Joint-Cond) modeling schemes as annotators with less than a threshold count of ratings are dropped. Y-axis represents $\mathcal{E}_d$ and X-axis represents the minimum number of annotations (cutoff threshold).

a few dimensions. We then use the known distribution parameters for that annotator and the available partial rating to predict the missing dimension. For the sake of brevity we focus on the case when only one of the rating dimensions is missing, noting however that other cases with more than one missing dimension are entirely straightforward. The primary goal of this model is to reduce cognitive load on the annotator by asking him/her to annotate a subset of the rating dimensions.

We represent the available subset of rating dimensions in the vector $\mathbf{a}_k^m$, barring rating $a_k^{m,d}$ of dimension d as $\mathbf{a}_k^{m,/d}$. Further, we represent the means and co-variance matrix entries corresponding to the dimensions barring dimension d as $\boldsymbol{\mu}_k^{m,/d}$ and $\boldsymbol{\Sigma}_k^{m,/d}$. In our specific case, $\boldsymbol{\mu}_k^{m,/d}$ and $\boldsymbol{\Sigma}_k^{m,/d}$ would be of dimensionalities 3×1 and 3×3, respectively. Also, the entries within $\boldsymbol{\Sigma}_k^m$ storing the co-variances between the dimension d and other dimensions is represented as $\boldsymbol{\Gamma}_k^{m,d}$. $\boldsymbol{\Gamma}_k^{m,d}$ is a vector of dimensionality 1×3. Now, given that the Joint annotator - Joint rating model prediction for the rating at dimension d was given by $\hat{a}_k^{m,d}$, we update it to $\hat{a}_k^{m,d+}$ with the availability of $\boldsymbol{a}_k^{m,/d}$ as shown in Equation 3.11. This equation follows from the computation of conditional Gaussian distribution from a joint Gaussian distribution, given partial availability of some of the variables [29].

$$\hat{a}_k^{m,d+} = \hat{a}_k^{m,d} + \boldsymbol{\Gamma}_k^{m,d}(\boldsymbol{\Sigma}_k^{m,/d})^{-1}(\boldsymbol{a}_k^{m,/d} - \boldsymbol{\mu}_k^{m,/d}) \tag{3.11}$$

We report the MSE $\mathcal{E}_d, \forall d \in 1, .., 4$ separately.

3.5 Results

We report results from two different experiment settings for the models described above. In the first setting, we use ratings from all annotators over the entire data. However, as some of the annotators only annotated a handful of data points (as few as 2 data points), in the second setting we discard annotators with fewer than a threshold number of ratings. This allows for a more robust estimation of parameters $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ per annotator. We use a 10 fold cross validation scheme over each annotator for all the models.

3.5.1 Setting 1: Training on data from all annotators

We first compare the different models by including all the annotators in our corpus irrespective of the amount of data they annotated. The metric $\mathcal{E}_d$ for every dimension d is shown in table 3.2.

From the table, we observe that the Joint-Ind and Joint-Joint models outperform the chosen baseline predictor in all the cases. The Joint-Joint

Dimension d	1 (Ex)	2 (Na)	3 (Go)	4 (En)
Baseline	11.00	10.25	13.86	13.81
Joint-Ind	0.82	0.72	0.76	0.92
Joint-Joint	0.80	0.74	0.69	0.87
Joint-Cond	1.28	3.97	26.08	9.89

Table 3.2: MSE $\mathcal{E}_d$ for annotator label prediction on the four rating dimensions; Ex: Expressiveness, Na: Naturalness, Go: Goodness of Pronunciation, En: Engagement

model shows the best performance in 3 out of 4 cases. It makes use of the joint information in the data to make accurate predictions on the annotator ratings rendering confidence in the model's ability to reliably estimate the hidden ground truth in multidimensional annotation settings, making this the model of choice in most cases including when the number of ratings per annotator are low. The Joint-Cond model does better than baseline for expressiveness, naturalness and engagement but fares much worse on pronunciation goodness. We attribute this to poor parameter estimation particularly on annotators with a small number of ratings. In particular the co-variance matrix $\mathbf{\Sigma}_k$ is poorly estimated for most annotators, which plays an important role in determining the Joint-Cond estimate. We expect the model to do well when a sufficient amount of rating is available from every annotator, which is discussed in the next section.

3.5.2 Setting 2: Training on annotators with more than a threshold count of ratings

In this setting, we iteratively remove annotators if they rated fewer than a threshold number of data samples. The metric $\mathcal{E}_d$ is then computed only on the retained annotators. The progression of $\mathcal{E}_d$ as we increase the threshold is shown in Figure 3.2.

From Figure 3.2, we observe similar performance trends as the previous section when the cutoff threshold is low. However, as the minimum number of annotations is increased, the baseline and Joint-Cond models show marked improvements in performance, while the Joint-Ind and Joint-Joint models' performance remains more or less consistent. The improvement is significantly better for the Joint-Cond model and it outperforms the Joint-Ind and Joint-Joint beyond a certain threshold for all the rating dimensions. Hence we can use the Joint-Cond model to reduce the dimensionality of queries made to a given annotator, after a sufficient number of ratings are collected for him/her, in turn reducing the annotator's cognitive load and overall annotation cost.

3.6 Conclusion

Ratings from multiple annotators are often pooled in several applications to obtain the ground truth. Several previous works [1] have proposed methods for modeling these ratings from multiple annotators. However, such models were not investigated in the case of multidimensional annotations. In this work, we presented a model for multidimensional annotation fusion and proposed variants which were applied to the task of predicting back annotator labels. We tested the fusion model on the SafariBob dataset with four dimensional ratings and observed that the proposed model outperformed two baselines by making label predictions with low MSE. A further extension was proposed which was shown to be useful in reducing the dimension of ratings presented to annotators after we obtain sufficiently confident parameters.

The model described in this chapter uses additive Gaussian noise to capture the relationship between annotation dimensions. However, such a model fails to capture more nuanced structural relationships between the dimensions. For example, if a given annotator's perception of a dimension scales with one or more of the actual ground truth values, such relationships are not easily captured by the model presented in this chapter. To address this, in the next chapter we propose a matrix factorization based model for multidimensional annotation fusion.

Chapter 4

Matrix factorization model

4.1 Introduction

In the previous chapter, we addressed the need for joint modeling of annotation dimensions by proposing a model that uses additive joint multidimensional Gaussian noise. We evaluated the model on Mturk annotations collected for audio clips of children diagnosed with autism. However, this model fails to capture nuanced relationships between ground truth values and the annotation dimensions. In this chapter, we address this shortcoming by proposing a matrix factorization based multidimensional annotation fusion model, which decomposes annotation vectors into a data point specific ground truth vector and an annotator specific linear transformation matrix. The model we propose is an extension of the Factor Analysis model and is applicable to both the global annotation setting (such as while collecting emotion annotations on a picture, judgment about the overall tone of a conversation, etc.) as well as time series annotations (for example, annotations of audio/video clips). Similar to the model proposed in the previous chapter, this model treats the hidden ground truth as latent variables and estimates them jointly along with the annotator parameters using the Expectation Maximization algorithm [5]. We evaluate the model in both settings on synthetic and real emotion corpora. We also create an artificial annotation task with controlled ground truth which is used in the model evaluation for both settings.

The rest of the chapter is organized as follows. In Section 4.2 we describe the proposed model and provide equations for parameter estimation using EM algorithm. We evaluate the model in Section 4.3 and provide conclusions in Section 4.4.

4.2 Model

4.2.1 Setup

The proposed model is shown in Figure 4.1. Each data point m has feature vector $\mathbf{x}_m$ and an associated multidimensional ground truth $\mathbf{a}_*^m$, which is defined as follows,

$$\mathbf{a}_*^m = f(\mathbf{x}_m; \Theta) + \boldsymbol{\epsilon}_m \tag{4.1}$$

We assume that from a pool of K annotators, a subset operates on each data point and provides their annotation $\mathbf{a}_k^m$.

$$\mathbf{a}_k^m = g(\mathbf{a}_*^m; F_k) + \boldsymbol{\eta}_k \tag{4.2}$$

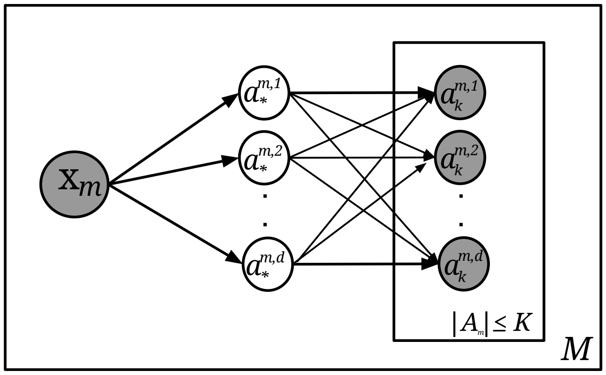


Figure 4.1: Proposed model. $\mathbf{x}^m$ is the set of features for the m^{th} data point, $a_*^{m,d}$ is the latent ground truth for the d^{th} dimension and $a_k^{m,d}$ is the rating provided by the k^{th} annotator. Vectors $\mathbf{x}^m$ and $\mathbf{a}_k^m$ (shaded) are observed variables, while $\mathbf{a}_*^m$ is latent. $\mathbf{A}_m$ is the set of annotator ratings for the m^{th} instance.

where index k corresponds to the k^{th} annotator; F_k is an annotator specific matrix that defines his/her linear weights for each output dimension; $\boldsymbol{\epsilon_m}$ and $\boldsymbol{\eta_k}$ are noise terms defined individually in the next sections along with the functions f and g. In the global annotation setting, both $\mathbf{a}_*^m$ and $\mathbf{a}_k^m \in \mathbb{R}^{\mathrm{D}}$ where D is the number of items being annotated; for the time series setting $\mathbf{a}_*^m$ and $\mathbf{a}_k^m \in \mathbb{R}^{\mathrm{T\times D}}$, where T is the total duration of the data point (audio/video signal). In all subsequent definitions, we use uppercase letters M, K, T, D to denote various counts and lowercase letters m, k, t, d to denote the corresponding index variables.

We make the following assumptions in our model.

A1 Annotations are independent for different data points.

A2 The annotations for a given data point are independent of each other given the ground truth.

A3 The model ground truths for different annotation dimensions are assumed to be conditionally independent of each other given the features $\boldsymbol{x}_m$.

4.2.2 Global annotation model

In this setting, the ground truth and annotations are d dimensional vectors for each data point. We define the ground truth $\mathbf{a}_*^m$ and annotations $\mathbf{a}_k^m$ as follows.

$$\mathbf{a}_*^m = \Theta^T \mathbf{x}_m + \boldsymbol{\epsilon}_m \tag{4.3}$$

$$\mathbf{a}_k^m = F_k \mathbf{a}_*^m + \boldsymbol{\eta}_k \tag{4.4}$$

where, $\mathbf{x}_m \in \mathrm{I\!R}^{\mathrm{P}}$; $\Theta \in \mathrm{I\!R}^{\mathrm{P \times D}}$; $\boldsymbol{\epsilon}_m \sim N(\mathbf{0}, \sigma^2 I)$; $\sigma^2 \in \mathrm{I\!R}$. The annotator noise $\boldsymbol{\eta}_k$ is defined as $\boldsymbol{\eta}_k \sim N(\mathbf{0}, \tau_k^2 I)$; $\tau_k^2 \in \mathrm{I\!R}$. $F_k \in \mathrm{I\!R}^{\mathrm{D \times D}}$ is the annotator specific weight matrix. Each annotation dimension value $a_k^{m,d}$ for annotator k is defined as a weighted average of the ground truth vector $\mathbf{a}_*^m$ with weights given by the vector $F_k(d,:)$.

Parameter Estimation

The model parameters $\Phi = \{F_k, \Theta, \sigma^2, \tau_k^2\}$ are estimated using Maximum Likelihood Estimation (MLE) in which they are chosen to be the values that maximize the likelihood function $\mathcal{L}$.

$$\begin{aligned}\log \mathcal{L} &= \sum_{m=1}^{M} \log p(\mathbf{a}_1^m \ldots \mathbf{a}_K^m; \Phi) \\ &= \sum_{m=1}^{M} \log \int_{\mathbf{a}_*^m} p(\mathbf{a}_1^m \ldots \mathbf{a}_K^m | \mathbf{a}_*^m; F_k, \tau_k^2) p(\mathbf{a}_*^m; \Theta, \sigma^2)\, d\mathbf{a}_*^m \qquad (4.5)\end{aligned}$$

Optimizing Equation 4.5 directly is intractable because of the presence of the integral within the log term, hence we use the EM algorithm. Note that the model we propose assumes that only some random subset of all available annotators provide annotations on a given data point, as shown in Figure 4.1. However, for ease of exposition, we overload the variable K and use it here to indicate the number of annotators that attempt to judge the given data point m.

EM algorithm

The Expectation Maximization (EM) algorithm to estimate the model parameters is shown below. It is an iterative algorithm in which the E and M-steps are executed repeatedly until an exit condition is encountered. Complete derivations for the model can be found in Appendix A.1.

Initialization We initialize by assigning the expected values and covariance matrices for the m ground truth vectors $\mathbf{a}_*^m$ to their sample estimates (i.e. sample mean and sample covariance) from the corresponding annotations. We then estimate the parameters as described in the maximization step using these estimates.

E-step In this step we take expectation of the log likelihood function with respect to $p(\mathbf{a}_*^m|\mathbf{a}_1^m \dots \mathbf{a}_K^m)$ and the resulting objective is maximized with respect to the model parameters in the M-step. Equations to compute the expected value and covariance matrices for the latent variable $\mathbf{a}_*^m$ in the E-step are listed below.

$$\mathbb{E}_{\mathbf{a}_*^m|\mathbf{a}_1^m \dots \mathbf{a}_K^m}[\mathbf{a}_*^m] = \Theta^T \mathbf{x}_m + \Sigma_{\mathbf{a}_*^m, \mathbf{a}_1^m \dots \mathbf{a}_K^m} \Sigma^{-1}_{\mathbf{a}_1^m \dots \mathbf{a}_K^m, \mathbf{a}_1^m \dots \mathbf{a}_K^m} (\mathbf{a}^m - \boldsymbol{\mu}^m)$$

$$\Sigma_{\mathbf{a}_*^m|\mathbf{a}_1^m \dots \mathbf{a}_K^m}[\mathbf{a}_*^m] = \Sigma_{\mathbf{a}_*^m, \mathbf{a}_*^m} - \Sigma_{\mathbf{a}_*^m, \mathbf{a}_1^m \dots \mathbf{a}_K^m} \Sigma^{-1}_{\mathbf{a}_1^m \dots \mathbf{a}_K^m, \mathbf{a}_1^m \dots \mathbf{a}_K^m} \Sigma_{\mathbf{a}_1^m \dots \mathbf{a}_K^m, \mathbf{a}_*^m}$$

The Σ terms are covariance matrices between the subscripted random variables. $\mathbf{a}^m$ and $\boldsymbol{\mu}^m$ are DK dimensional vectors obtained by concatenating the K annotation vectors $\mathbf{a}_1^m, \dots \mathbf{a}_K^m$ and their corresponding expected values.

M-step In this step, we compute current estimates for the parameters as follows. The expectations shown below are over the conditional distribution $\mathbf{a}_*^m|\mathbf{a}_1^m \dots \mathbf{a}_K^m$.

$$\Theta = (\mathrm{X}^T \mathrm{X})^{-1} (\mathrm{X}^T \, \mathbb{E}[\mathrm{a}_*^m])$$

$$F_k = \left(\sum_{m=1}^{M_k} \mathbf{a}_K^m \, \mathbb{E}[(\mathbf{a}_*^m)^T] \right) \left(\sum_{m=1}^{M_k} \mathbb{E}[\mathbf{a}_*^m \, (\mathbf{a}_*^m)^T] \right)^{-1}$$

$$\sigma^2 = \frac{1}{md} \sum_{m=1}^{M} \left(\mathbb{E}[(\mathbf{a}_*^m)^T \mathbf{a}_*^m] - 2tr \left(\Theta'^T \mathbf{x}_m \, \mathbb{E}[(\mathbf{a}_*^m)^T] \right) + tr(\mathbf{x}_m^T \Theta' \Theta'^T \mathbf{x}_m) \right)$$

$$\tau_k^2 = \frac{1}{m_k d} \sum_{m=1}^{M_k} \left((\mathbf{a}_K^m)^T \mathbf{a}_K^m - 2tr \left(F_k'^T \mathbf{a}_K^m \, \mathbb{E}[(\mathbf{a}_*^m)^T] \right) + tr \left(F_k'^T F_k' \, \mathbb{E}[\mathbf{a}_*^m (\mathbf{a}_*^m)^T] \right) \right)$$

Note the similarity of the update equation for Θ with the familiar normal equations. We are using the soft estimate of $\mathbf{a}_*^m$ to find the expression for Θ in each iteration. Here, X is the feature matrix for all data points; it includes individual feature vectors x_m in its rows. Θ' and F_k' are parameters from the previous iteration.

Termination We run the algorithm until convergence, the criterion for which was chosen to be when the change in model log-likelihood reduces to less than 0.001% from the previous iteration.

4.2.3 Time series annotation model

In this setting, the ground truth and the annotations are matrices with T rows (time) and D columns (annotation dimensions). The ground truth matrix $\mathbf{a}_*^m$ is defined as follows.

$$\mathrm{vec}(\mathbf{a}_*^m) = \mathrm{vec}(\mathrm{X}_m\Theta) + \boldsymbol{\epsilon}_m \tag{4.6}$$

where $\mathbf{a}_*^m \in \mathrm{I\!R}^{\mathrm{T}\times\mathrm{D}}$, $\mathrm{X}_m \in \mathrm{I\!R}^{\mathrm{T}\times\mathrm{P}}$ and $\Theta \in \mathrm{I\!R}^{\mathrm{P}\times\mathrm{D}}$; T represents the time dimension and is the length of the time series. X_m is the feature matrix where each row corresponds to features extracted from the data point for one particular time stamp. vec(.) is the vectorization operation which flattens the input matrix in column first order to a vector. $\boldsymbol{\epsilon}_m \sim \mathcal{N}(\mathbf{0}, \sigma^2 I) \in \mathrm{I\!R}^{\mathrm{TD}}$ is the additive noise vector with $\sigma \in \mathrm{I\!R}$.

In [18], the authors propose a linear model where the annotation function $g(\mathbf{a}_*^m; F_k)$ is a causal linear time invariant (LTI) filter of fixed width. The advantage of using an LTI filter is that it can capture scaling and time-delay biases introduced by the annotators.

Since the filter width W is chosen such that $W \ll T$ where T is the number of time stamps for which we have the annotations, the annotation function for dimension d' can be viewed as the left multiplication of a filter matrix $B_k^{d'} \in \mathrm{I\!R}^{\mathrm{T}\times\mathrm{T}}$ as shown in Equation 4.7.

$$B_k^{d'} = \begin{bmatrix} b_1^{d'} & 0 & 0 & 0 & 0 & \dots & 0 \\ b_2^{d'} & b_1^{d'} & 0 & 0 & 0 & \dots & 0 \\ b_3^{d'} & b_2^{d'} & b_1^{d'} & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & & \\ 0 & b_W^{d'} & \dots & b_1^{d'} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & & \\ 0 & 0 & 0 & 0 & b_W^{d'} & \dots & b_1^{d'} \end{bmatrix} \tag{4.7}$$

We extend this model in our work to combine information from all of the annotation dimensions. Specifically, the ground truth is left multiplied by D horizontally concatenated filter matrices each $\in \mathrm{I\!R}^{\mathrm{T}\times\mathrm{T}}$ corresponding to a different dimension as shown below.

$$\mathbf{a}_k^{m,d} = F_k^d \text{vec}(\mathbf{a}_*^m) + \boldsymbol{\eta}_k \tag{4.8}$$

where,

$$F_k^d = [B_k^{d,1}, B_k^{d,2}, \dots, B_k^{d,D}] \tag{4.9}$$

$F_k^d \in \mathrm{I\!R}^{\mathrm{T}\times\mathrm{TD}}$ with WD unique parameters. $\boldsymbol{\eta}_k \sim \mathcal{N}(\mathbf{0}, \tau_k^2 I) \in \mathrm{I\!R}^{\mathrm{T}}$ with $\tau_k^2 \in \mathrm{I\!R}$.

Parameter Estimation

Estimating the model parameters similar to the global model requires computing the expectations over a vector of size TD. Since T is the number of time stamps in the task and can be arbitrarily long, this may not be feasible in all tasks. For example, in the movie emotions corpus [25], annotations are computed at a rate of 25 frames per second with each file of duration ∼30 minutes or of ∼45k annotation frames. To avoid this we use a variant of EM named *Hard EM* in which instead of taking expectations over the entire conditional distribution of $\mathbf{a}_*^m$ we find its mode. This variant has been shown to be comparable in performance to the classic EM (*Soft EM*) despite being significantly faster and simple [33]. This approach is similar to the parameter estimation strategy devised by [18] in their time series annotation model.

The likelihood function is similar to the global model in Equation 4.5 as shown below.

$$\log \mathcal{L} = \sum_{m=1}^{M} \log \int_{\mathbf{a}_*^m} p(\mathbf{a}_1^m \dots \mathbf{a}_K^m | \mathbf{a}_*^m; F_k, \tau_k^2) p(\mathbf{a}_*^m; \Theta, \sigma^2)\, d\mathbf{a}_*^m$$

However the integral here is with respect to the flattened vector vec($\mathbf{a}_*^m$).

EM algorithm

The EM algorithm for the time series annotation model is listed below. Complete derivations for the model can be found in Appendix A.2

Initialization Unlike the global annotation model, we initialize $\mathbf{a}_*^m$ randomly since we observed better performance when compared to initializing it with the annotation means. Given this $\mathbf{a}_*^m$, the model parameters are estimated as described in the maximization step below.

E-step In this step we assign $\mathbf{a}_*^m$ to the mode of the conditional distribution $q(\mathbf{a}_*^m) = p(\mathbf{a}_*^m | \mathbf{a}_1^m, \dots, \mathbf{a}_K^m)$. Since this distribution is normal finding the mode is equivalent to minimizing the following expression.

$$\mathbf{a}_*^m = \underset{\mathbf{a}_*^m}{\operatorname{argmin}} \sum_k \sum_d ||\mathbf{a}_k^{m,d} - F_k^d \text{vec}(\mathbf{a}_*^m)||_2^2 + ||\text{vec}(\mathbf{a}_*^m) - \text{vec}(\mathrm{X}_m \Theta)||_2^2$$

M-step Given the estimate for $\mathbf{a}_*^m$ from the E-step, we substitute it in the likelihood function and maximize with respect to the parameters in the M-step. The estimates for the different parameters are shown below.

$$\begin{aligned}
\Theta &= \left(\sum_{m=1}^{M} \mathrm{X}_m^T \mathrm{X}_m \right)^{-1} \left(\sum_{m=1}^{M} \mathrm{X}_m^T \mathbf{a}_*^m \right) \\
f_k^d &= \left(\sum_{m=1}^{M_k} A^T A \right)^{-1} \left(\sum_{m=1}^{M_k} A^T \mathbf{a}_k^{m,d} \right) \\
\sigma^2 &= \frac{1}{MTD} \sum_{m=1}^{M} ||\text{vec}(\mathbf{a}_K^m) - \text{vec}(\mathrm{X}_m \Theta)||_2^2 \\
\tau_k^2 &= \frac{1}{M_k TD} \sum_{m=1}^{M_k} \sum_d ||\mathbf{a}_k^{m,d} - F_k^d \text{vec}(\mathbf{a}_*^m)||_2^2
\end{aligned}$$

A is a matrix obtained by reshaping vec($\mathbf{a}_*^m$).

Termination We run the algorithm until convergence, the criterion for which was chosen to be when the change in model log-likelihood reduces to less than 0.5% from the previous iteration.

4.3 Experiments and Results

We evaluate the models described above on three different types of data: synthetic data, an artificial task with human annotations, and finally with real data. We describe these individually below. We compare our joint models with their *independent* counterparts in which each annotation dimension is modeled separately. Update equations for the independent model can be obtained by running the models described above for each dimension separately with $D = 1$. Note that the independent model is similar in the global setting to the regression model proposed in [1] (with ground truth scaled by the singleton f_k^d). In the time series setting it is identical to the model proposed by [18].

The models are evaluated by comparing the estimated $\mathbf{a}_*^m$ with the actual ground truth. We report model performance using two metrics: the Concordance correlation coefficient (ρ_c) [34] and the Pearson's correlation coefficient (ρ). ρ_c measures any departures from the *concordance line* (line passing through the origin at 45° angle). Hence it is sensitive to rotations or rescaling in the predicted ground truth. Given two samples x and y, the sample concordance coefficient $\hat{\rho_c}$ is defined as shown below.

$$\hat{\rho_c} = \frac{2s_{xy}}{s_x^2 + s_y^2 + (\bar{x} - \bar{y})^2} \tag{4.10}$$

We also report results in Pearson's correlation to highlight the accuracy of the models in the presence of rotations.

As noted before, the models proposed in this paper are closely related to the Factor Analysis model, which is vulnerable to issues of unidentifiability [35], due to the matrix factorization. Different types of unidentifiability have been studied in literature, such as factor rotation, scaling and label switching. In our experiments, we handle label switching through manual judgment (by reassigning the estimated ground truth between dimensions if necessary) as is common in psychology [36], but defer the task of choosing an appropriate prior on the rotation matrix F_k to address other unidentifiabilities for future work.

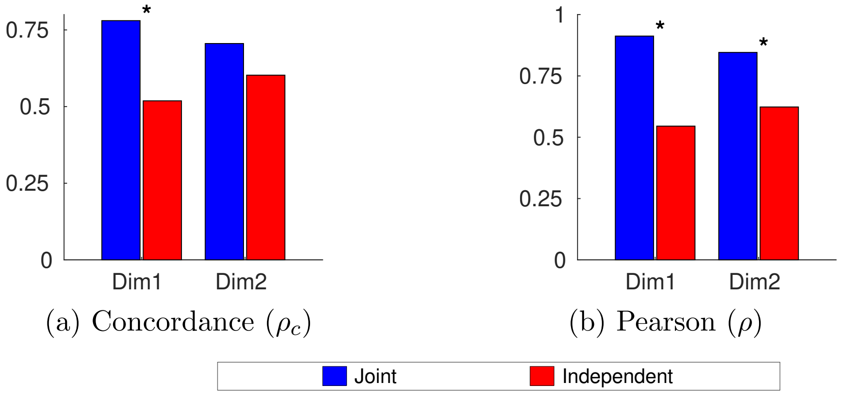
We report aggregate test set results using C-fold cross validation. To address overfitting, within each fold, we evaluate the parameters obtained after each iteration of the EM algorithm by estimating the ground truth on a disjoint validation set, and pick those with the highest performance in concordance correlation ρ_c as the parameter estimates of the model. We then estimate the performance of this parameter set in predicting the ground truth from a separate held out test set for that fold. Finally, we also report statistically significant differences between the joint and independent models at 5% false-positive rate ($\alpha = 0.05$) in all our experiments.

4.3.1 Global annotation model

The global annotation model uses the EM algorithm described in Section 4.2.2 to estimate the ground truth for discrete annotations. We evaluate the model in three different settings described below. Statistical significance tests were run by computing bootstrap confidence intervals [37] on the differences in model performances across the C-folds.

Synthetic data

We created synthetic data according to the model described in Section 4.2.2 with random features $\mathrm{X} \in \mathbb{R}^{500}$ for 100 data points each with 2 dimensions of annotations (i.e. D=2). 10 artificial annotators, each with unique random F_k matrices were used to produce annotations for all the files. Elements of the feature matrices were sampled from the standard normal distribution, while



(a) Concordance (ρ_c)

(b) Pearson (ρ)

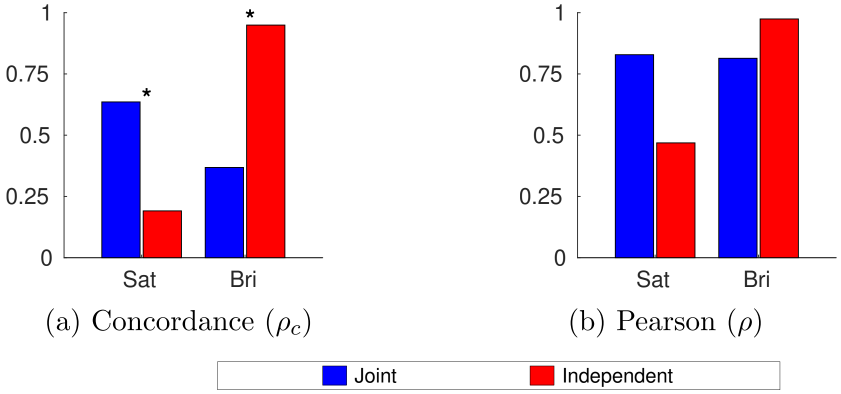
Figure 4.2: Performance of global annotation model on synthetic dataset; **-statistically significant*

the elements of F_k matrices were sampled from $\mathcal{U}(0,1)$. Elements of ground truth $\mathbf{a}_*^m$ were sampled from $\mathcal{U}(-1,1)$ and θ was estimated from $\mathbf{a}_*^m$ and X. Since its off diagonal elements are non-zero, our choice of F_k represents tasks in which the annotation dimensions are related to each other.

Figure 4.2 shows the performance of joint and independent models in predicting the ground truth $\mathbf{a}_*^m$. For both dimensions, the proposed joint model predicts the $\mathbf{a}_*^m$ with considerably higher accuracy as shown by the higher correlations, highlighting the advantages of modeling the annotation dimensions jointly when they are expected to be related to each other.

Artificial data

Since crowdsourcing experiments typically involve collecting subjective annotations, they seldom have well defined ground truth. As a result, most annotation models are evaluated on expert annotations collected by specially trained users. For example, while collecting annotations on medical data the ground truth estimated by fusing annotations from naive users may be eval-



(a) Concordance (ρ_c)

(b) Pearson (ρ)

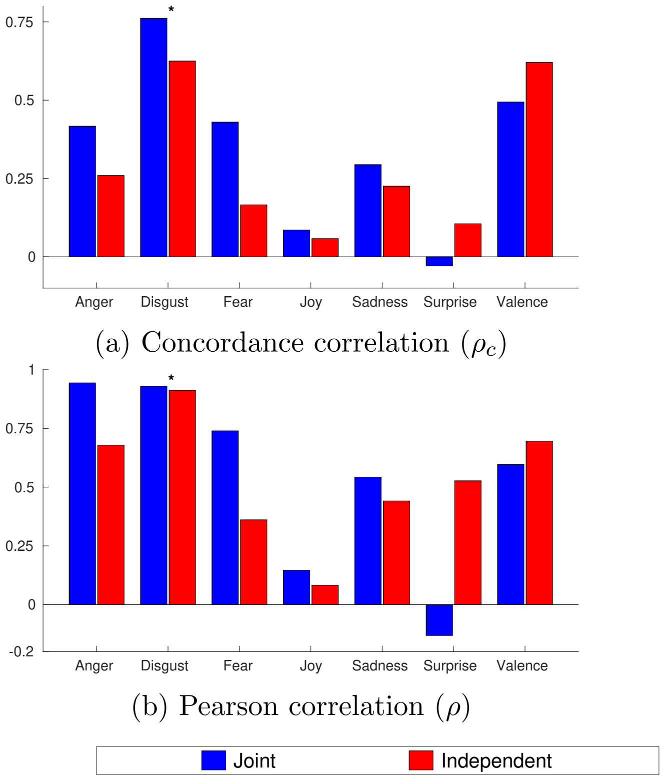
Figure 4.3: Performance of global annotation model on artificial dataset; *Sat-Saturation, Bri-Brightness; *-statistically significant*

uated against reference labels provided by experts such as doctors. However, this poses a circular problem since the expert annotations themselves may be subjective and combining them to may not be straightforward. To address this, we created an artificial task with controlled ground truth on which we collect annotations from multiple annotators and evaluate the fused annotation values with the known ground truth values, similar to [38]. In our task, the annotators were asked to provide their best estimates on saturation and brightness values for monochromatic images. The relationship between perceived saturation and brightness is well known as the Helmholtz—Kohlrausch effect, according to which, increasing the saturation of an image leads to an increase in the perceived brightness, even if the actual brightness was constant [39].

In our experiments, we collected annotations on images from two regimes: one with fixed saturation and varying brightness, and vice versa. This approach was chosen since it would allow us to evaluate the impact of change in either brightness or saturation while the other was held constant. The color of the images were chosen randomly (and independent of the image's saturation and brightness) between green and blue. Annotations were collected on Mturk and the annotators were asked to familiarize themselves with saturation and brightness using an online interactive tool before providing their ratings. In both experiments, a reference image with fixed brightness and saturation was inserted after every ten annotation images to prevent any bias in the annotators. The reference images were hidden from the annotators and appeared as regular annotation images. For parameter estimation, RGB values were chosen as the features for each image.

We used the joint model to estimate the ground truth for the two regimes separately since we expect the relationship between saturation and brightness to be dissimilar in the two cases. From each experiment, predicted values of the underlying dimension being varied was compared with the actual $\mathbf{a}_*^m$ values. For example, in the experiment with varying saturation and fixed brightness, the joint model was run on full annotations, but only estimated values of saturation were compared with the ground truth. For the independent model, we use annotation values of the underlying dimension being varied from each regime, and compare the estimated values with ground truth.

Figure 4.3 shows the performance of the joint and independent models for this experiment. The joint model leads to better estimates of saturation when compared to the independent model by making use of the annotations on brightness. This agrees with the Helmholtz—Kohlrausch phenomenon described above, since the annotators can perceive the changing saturation as a change in brightness, leading to correlated annotations for the two dimen-



(a) Concordance correlation (ρ_c)

(b) Pearson correlation (ρ)

Figure 4.4: Performance of global annotation model on the text emotions dataset; **-statistically significant*

sions. On the other hand, the independent model leads to better estimates of brightness, which seems to have no effect on perceived saturation annotations. This experiment highlights the benefits of jointly modeling annotations in cases where the annotation dimensions may be correlated or dependent on each other.

Real data

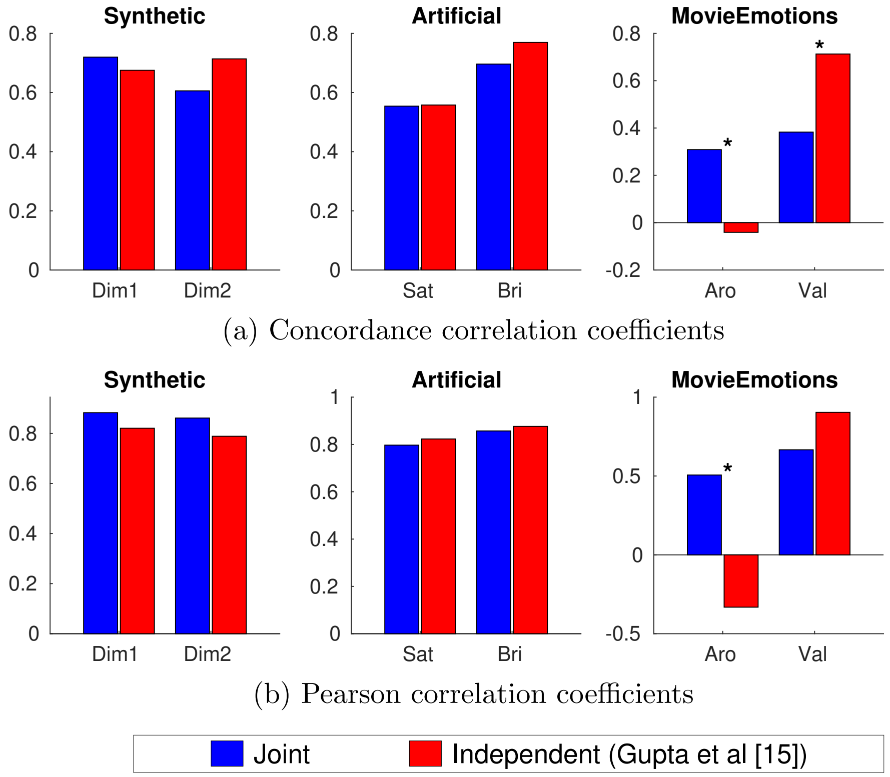
Our final experiment for the global model was on the task of annotating news headlines in which the annotators provide numeric ratings for various emotions. This dataset was first described in the 2007 SemEval task on affective text [40]. Numeric ratings from the original task were labeled *in house* and we treat these as expert annotations since the annotators were trained with examples. We use Mturk annotations from [9] as the actual input to our model using which the ground truth estimates are computed. Sentence level annotations are provided on seven dimensions (D=7): *anger*, *disgust*, *fear*, *joy*, *sadness*, *surprise* and *valence* (positive/negative polarity). In our experiments, we use sentence level embeddings computed using the pre-trained sentence embedding model sent2vec[1] [41] as features.

Figure 4.4 shows the performance of the joint and independent models on this task. The joint model shows better performance in predicting the ground truth for *anger*, *disgust*, *fear*, *joy* and *sadness*, but performs worse than the independent model in predicting *surprise* and *valence*.

4.3.2 Time series annotation model

In this setting, the annotations are collected on data with a temporal dimension, such as time series data, video or audio signals. Similar to the global model, we evaluate this model in 3 settings: synthetic, artificial and on real data. The evaluation metrics ρ_c and ρ are computed over estimated and actual ground truth vectors $\mathbf{a}_*^m$ by concatenating the data points into a single vector. The time series models have the window size W as an additional hyperparameter, which is selected using a validation set. In each fold of the dataset, we train model parameters for different window sizes from the set $\{5, 10, 20, 50\}$, and pick W and related parameters with the highest concor-

[1] https://github.com/epfml/sent2vec



(a) Concordance correlation coefficients

(b) Pearson correlation coefficients

Figure 4.5: Concordance and Pearson correlation coefficients between ground truth and model predictions for the time series annotation model; *-*statistically significant*

dance correlation ρ_c on the validation set. These are then evaluated on a disjoint test set, and we repeat the process for each fold. In each experiment, the parameters were initialized randomly, and the process was repeated 20 times at different random initializations, selecting the best starting point using the validation set. To identify significant differences, we compute the test set performance of the two models for each fold, and run the paired t-test between the two C sized samples. We avoid computing bootstrap confidence intervals due to smaller test set sizes.

Synthetic data

The synthetic dataset was created using the model described in Section 4.2.3. Elements of the feature matrix were sampled from the standard normal distribution while elements of F_k and ground truth were sampled from $\mathcal{U}(0, 1)$. In this setting each data point includes T feature vectors, one for each time stamp. The time dependent feature matrices were created using a random walk model without drift but with *lag* to mimic a real world task. In other words, while creating the P dimensional time series, the features vectors were held fixed for a time period arbitrarily chosen to be between 2 to 4 time stamps. This was done because in most tasks the underlying dimension (such as emotion) is expected to remain fixed at least for a few seconds. In addition, the transition between changes in the feature vectors were linear and not abrupt. In our experiments, we chose $P = 500$, $T = 350$, $D = 2$, $M = 18$ and the number of annotators $K = 6$.

Figure 4.5 shows the aggregate results across C-folds ($C = 5$) for the joint and independent models in the 3 settings. In the synthetic dataset, the joint model achieves higher values for Pearson's correlation ρ for both the dimensions and higher value for ρ_c for dimension 1. For dimension 2 however, the independent model achieves better ρ_c.

Artificial data

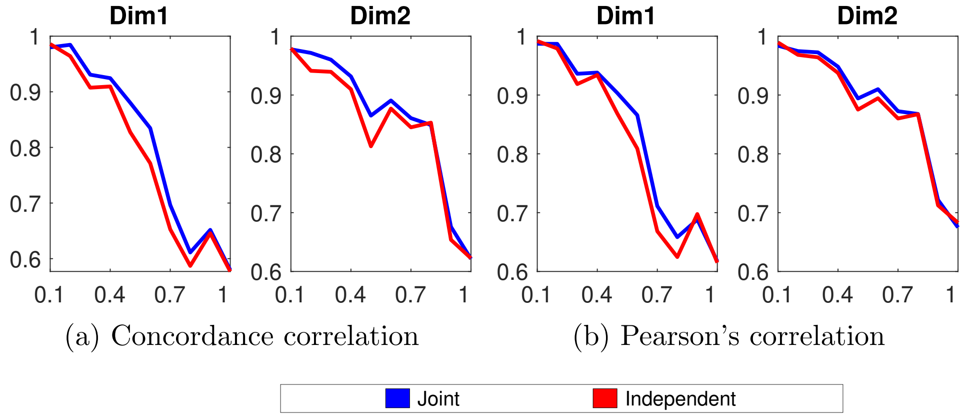
We collected annotations on videos with the artificial task of identifying saturation and brightness, described in the previous section. The videos consisted of monochromatic images with the underlying saturation and brightness varied independent of each other. The dimensions were created using a random walk model with lag. The annotations were collected in house using an annotation system developed using the Robot Operating System [42]. 10 graduate students gave their ratings on the two dimensions. Each dimension was annotated independently using a mouse controlled slider. For parameter estimation, the feature vectors for each time stamp were RGB values.

As seen in Figure 4.5, both models achieve similar performance in predicting the ground truth for saturation and brightness in terms of ρ, as well as in predicting saturation in terms of ρ_c. The independent model achieves slightly better performance in predicting brightness in terms of concordance correlation (though not statistically significant); however, their performance in terms of ρ suggests that the joint model output differs only in terms of a linear scaling. The joint model appears to be at par with the independent model for the most part, suggesting that the transformation matrix F_k connecting the two dimensions for each annotator, is unable to accurately capture the dependencies between the dimensions, likely due to the fact that, unlike the global annotation model, the underlying brightness and saturation were varied simultaneously and independent of each other (leading to non-linear dependencies between them), and that we limit F_k to only capture linear relationships.

Real data

We finally evaluate our model on a real world task with time series annotations. We chose the task of predicting the emotion dimensions of valence and arousal from movie clips, first explained in [25]. The associated corpus includes time series annotations of the emotion dimensions on contiguous 30 minute video segments from 12 Academy Award winning movies. This task was chosen because the data set includes both expert annotations as well as annotations from naive users. We treat the expert annotations as reference and evaluate the estimated ground truth dimensions against them; however, we note that the expert labels were provided by just one annotator, which may itself be noisy.

For each movie clip, 6 annotators provide annotations on their *perceived* valence and arousal using the Feeltrace [43] annotation tool. The features used in our parameter estimation include combined audio and video features extracted separately. The audio features were estimated using emotion recognition baseline features from Opensmile [44] at 25 fps (same frame rate as the video clips) and aggregated at a window size of 5 seconds using the following statistical functionals: mean, max, min, std, range, kurtosis, skewness and inter-quartile range. The video features were extracted using OpenCV [45] and included frame level luminance, intensity, Hue-Saturation-Value (HSV) color histograms and optical flow [46], which were also aggregated to 5 sec-



(a) Concordance correlation

(b) Pearson's correlation

Figure 4.6: Effect of varying dependency between annotation dimensions for the synthetic model

onds using simple averaging. The combined features were of size $P = 1225$ for each frame.

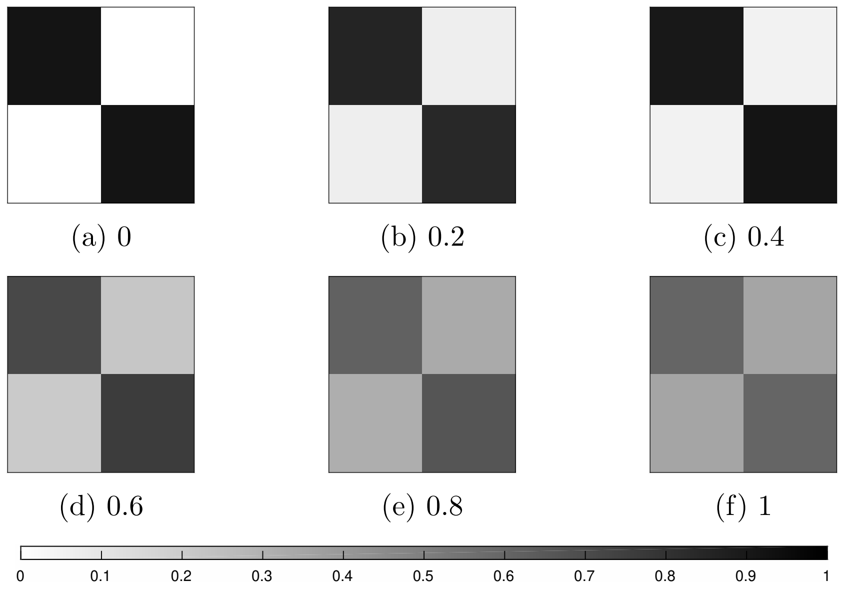
Figure 4.5 shows the performance of the two models for the dataset. The joint model seems to considerably outperform the independent model while estimating arousal while the independent models seem to produce better estimates of valence from the annotations. The independent model seems to perform poorly in arousal prediction, but shows strong performance with valence suggesting higher agreement between the annotators and the expert's opinions on valence. The joint model, however shows a balanced performance, where the information from valance seems to help in predicting arousal.

4.3.3 Effect of dependency among dimensions

To evaluate the impact of the magnitude of dependency between the annotation dimensions on the performance of the models, we created a set of synthetic annotations for the global model similar to Section 4.3.1. We created

10 synthetic datasets, each with constant F_k matrices across all annotators. The principal diagonal elements were fixed to 1 while the off diagonal elements were increased between 0.1 to 1 with a step size of 0.1. Similar to the previous setting, we created 100 annotators, each operating on 10 files. Note that despite the annotators having identical F_k matrices, their annotations on a given file were different because of the noise term $\boldsymbol{\eta}_k$ in Equation 4.2.

Figure 4.6 shows the 5-fold cross validated performance of the joint and independent models on this task. As seen in the figure, the joint model consistently outperforms the independent model in both metrics. Both the models start with similar performance when the off diagonal elements are close to zero since this implies no dependency between the annotation dimensions, and the performance of both models continues to degrade as the off diagonal elements increase. However, the joint model is able to make better predictions of the ground truth by making using of the dependency between the dimensions, highlighting the benefits of modeling the annotation dimensions jointly. We also created a plot of averages of all predicted F_k matrices for different step sizes (off diagonal elements of synthetic annotators) in Figure 4.7. In each case, the predicted F_k matrices close resemble the actual matrices for the annotators highlighting the accuracy of the joint model. However, as we get closer to step size 1, the estimated F_k matrices appear to be washed out with all terms of the estimated F_k close to 0.5 instead of 1 (Figure 4.7f). We attribute this to unidentifiability due to scaling that may have been introduced by the model during parameter estimation. Addressing this is an important part of our proposed future work.


(a) 0

(b) 0.2

(c) 0.4

(d) 0.6

(e) 0.8

(f) 1

Figure 4.7: Average F_k plots estimated from the joint model at different step sizes for off diagonal elements of the annotator's F_k matrices

4.4 Conclusion

We presented a model to combine multidimensional annotations from crowdsourcing platforms such as Mturk. The model assumes the ground truth to be latent and distorted by the annotators. The latent ground truth and the model parameters are estimated using the EM algorithm. EM updates are derived for both global and time series annotation settings. We evaluate the model on synthetic and real data. We also propose an artificial task with controlled ground truth and evaluate the model.

Weaknesses of the model include vulnerability to unidentifiability issues like most variants of factor analysis [35]. Typical strategies to address this issue involve adapting a suitable prior constraint on the factor matrix. For example, in PCA, the factors are ordered such that they are orthogonal to each other and arranged in decreasing order of variance. In our experiments, the most severe form of unidenfiability observed was due to label switching, which we addressed using manual judgments. We defer the task of choosing an appropriate prior constraint on F_k for future work.

Future work also includes generalizing the model with Bayesian extensions, in which case the parameters can be estimated using variational inference. Providing theoretical bounds to the model performance, specially with respect to the sample complexity may be possible since we have assumed normal distributions throughout the model.

Chapter 5

Estimation of psycholinguistic norms for sentences

5.1 Introduction

Psycholinguistic norms are numeric ratings assigned to linguistic cues such as words or sentences to measure various psychological constructs. Examples include dimensions such as valence, arousal, and dominance which are used to analyze the affective state of the author. Other examples include norms of higher order mental constructs such as concreteness and imagability which have been associated with improvements in learning [47]. The ease of computing the norms has enabled their application in a variety of tasks in natural language processing such as information retrieval [48], sentiment analysis [49], text based personality prediction [50] and opinion mining. The norms are typically annotated at the word level by psychologists who provide numeric scores to a curated list of seed words, which are then extrapolated to a larger vocabulary using either semantic relationships such as synonymy and hyponymy or using word occurrence based contextual similarity.

Most NLP applications of psycholinguistic norms use sentence or document level scores, but manual annotation of the norms at sentence level is difficult and not straightforward to generalize. In these cases, estimation of sentence level norms is done by aggregating the word level scores using simple averaging [51, 52], or by using distribution statistics of the word level scores [53]. However, such aggregation strategies may not be accurate at estimating the sentence level scores of the norms. In this work, we propose a new approach to estimate sentence level norms using the joint multidimensional model presented in Chapter 4 along with partial sentence level annotations.

Annotation of the normatives at the sentence level is a challenging task when compared to word level annotations since it involves evaluating the underlying semantics of the sentence in the abstract space of the corresponding dimension, with some dimensions in particular such as dominance being more difficult than others. Dominance is a measure of how dominant or submissive the object behind the word is. Being one of the three basic dimensions that are frequently used to describe emotional states (along with pleasure and arousal), dominance is commonly used in affective computing but annotating this dimension at the sentence level is considerably difficult. For example, it is relatively easy to estimate the dominance score for the words *happy* or *angry* but assigning a score for dominance for the sentence *I'm happy to know that the earthquakes are behind us* would be considerably difficult as it involves words with extreme values of dominance (happy and earthquake). On the other hand, some norms are easier to annotate at the sentence level (for example, valence). We use this fact along with the joint parameters learned from the fusion model to predict norms at sentence level given partial annotations.

The joint multidimensional annotation fusion model presented in Section 4.2.2 assumes a matrix factorization scheme to capture annotator behaviors. The annotations are assumed to be obtained by left multiplying the ground truth vector $\mathbf{a}_*^m$ with an annotator specific linear transformation matrix denoted as F_k, which captures the individual contributions of ground truth values for each dimension in the annotation output. In our work, we make use of this parameter F_k to estimate sentence level normative scores. We start by training the joint global annotation model using the EM algorithm listed in Section 4.2.2 at the word level to estimate the annotator parameters, and use the word level estimates for F_k on sentence level ratings from the same set of annotators. To make model predictions on a given dimension, we make use of partial annotations on the remaining dimensions along with F_k. Our proposed approach shows improved performance in predicting the sentence level norms when compared to various word level normative aggregation strategies.

The rest of the chapter is organized as follows. In Section 5.2, we briefly introduce the parameters of the joint multidimensional annotation fusion model and explain our data collection strategy in Section 5.3. We present our experimentats and results in Sections 5.4 and Section 5.5 before concluding in Section 5.6.

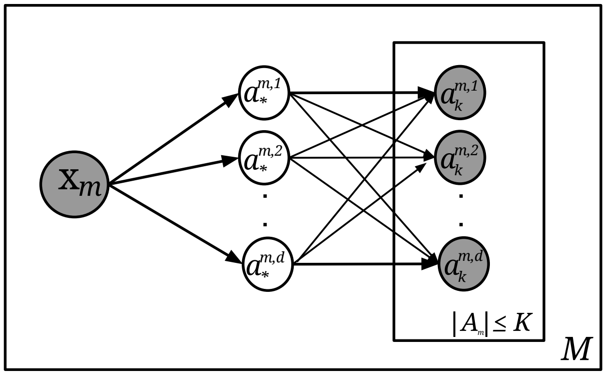


Figure 5.1: Joint multidimensional annotation fusion model from Section 4.2.3. F_k is estimated from word level annotations of psycholinguistic norms, which is used in predicting norms at the sentence level

5.2 Model

$$\begin{aligned} \mathbf{a}_*^m &= \Theta^T \mathbf{x}_m + \boldsymbol{\epsilon}_m \\ \mathbf{a}_k^m &= F_k \mathbf{a}_*^m + \boldsymbol{\eta}_k \end{aligned} \tag{5.1}$$

A plate notation diagram describing joint multidimensional annotation model introduced in Section 4.2.2 is shown in Figure 5.1. In this model, each annotator is assumed to operate on the ground truth vector $\mathbf{a}_*^m$ by left multiplying a matrix F_k as shown in Equation 5.1. This matrix captures the relationship between all dimensions of the ground truth vector with those in the annotation vector $\mathbf{a}_k^m$. The key idea used in our approach is that for a given annotator, the relationships between the annotation dimensions is similar for both word and sentence level annotations. In other words, the matrix F_k is assumed to be identical for both word and sentence annotations of the norms. Given multidimensional ratings from a set of annotators at word and sentence levels, we use the joint fusion model to estimate annotator specific parameters F_k at the word level, which are then used at the sentence level to estimate the psycholinguistic norms.

5.3 Data

We collected word level annotations on the affective norms of Valence, Arousal and Dominance using Mturk for words sampled from [54]. This corpus was chosen because it provides expert ratings on Valence, Arousal and Dominance for nearly 14,000 English words. Annotators were asked to provide numeric ratings between 1 to 5 (inclusive) for each dimension, on assignments consisting of a set of 20 randomly sampled words. In total, we collected annotations on 200 words. Instructions for the annotation assignments included definitions along with examples for each of the dimensions being annotated. After filtering incomplete and noisy submissions, we retained only those annotators who provided ratings for at least 100 words in the subsequent sentence level annotation task, to ensure sufficient training data.

Sentence level annotations were collected on sentences from the Emobank corpus [55], which includes expert ratings on valence, arousal and dominance for 10000 English sentences. 21 annotators from the word level annotation task were invited to provide labels for 100 sentences randomly sampled from this corpus. The assignments were presented in a similar fashion as word level annotations, with each assignment including 10 sentences and the workers providing numeric ratings for valence, arousal and dominance for each sentence. We use the annotator specific parameters F_k estimated at the word level to predict psycholinguistic norms for the sentences given partial annotations, using the approach described in the next section.

5.4 Experiments

Given annotator parameters F_k estimated at the word level, we use partial annotator ratings at the sentence level to predict the norms. For example, while predicting sentence level scores of valence, we use the sentence level annotator ratings on arousal and dominance along with the word level parameter matrix F_k^{word}, and repeat the process for each dimension. The use of partial annotations enables us to predict sentence level norms on challenging psycholinguistic dimensions using ratings on dimensions which maybe easier to annotate.

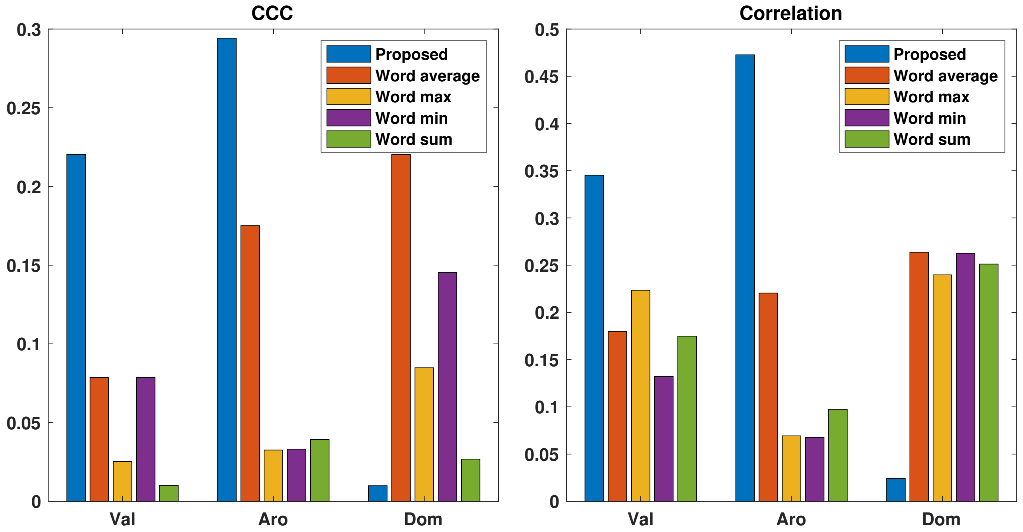
In our experiments, we make use of the IID Gaussian noise assumption in Equation 5.1, which reduces the task of predicting the sentence level norm to a linear regression problem shown in Equation 5.2. Rows of the matrix F_k^{word} are treated as features of the regression model with vector $\mathbf{a}_*^m$ as the regression parameter. Given partial annotations $\mathbf{a}_k^{m,\setminus d}$ and matrix F_k^{word}, the regression parameter vector $\mathbf{a}_*^m$ can be estimated using normal equations or gradient descent.

$$\begin{bmatrix} \cdot \\ a_1^{m,\backslash d} \\ \vdots \\ a_K^{m,\backslash d} \\ \cdot \end{bmatrix} = \begin{bmatrix} \cdot \\ F_1^{\text{word}}[\backslash d, :] \\ \vdots \\ F_K^{\text{word}}[\backslash d, :] \\ \cdot \end{bmatrix} \begin{bmatrix} \\ \\ \mathbf{a}_*^m \\ \\ \\ \end{bmatrix} \tag{5.2}$$

where, d is the dimension to predict

We compare the predicted norms with expert ratings from the Emobank corpus which acts as our reference to evaluate model performance. For baselines, we compute different aggregations of word level normative scores after filtering out non-content words as is common in literature [51]. The aggregation functions we evaluated are: unweighted average, maximum, minimum and sum of the word level norms. We use Concordance Correlation Coefficient (CCC; Equation 4.10) and Pearson's Correlation as our evaluation metrics. We report our results in the next section.

Finally, we repeat the above experiments with three other dimensions: pleasantness, imagability and genderladenness. Since we do not have reference labels for these to evaluate the predictions from different models, we simply report training errors from several regression models when trained on the model predictions. In this case, low training errors imply higher learnability which can act as a proxy to evaluate the quality of predictions. For regression models, we used support vector regression with l_1 and l_2 losses, as well as ridge regression. Model hyperparameters were tuned using 5 fold cross validation. In each dimension, the lowest mean squared error (MSE)



(a) CCC

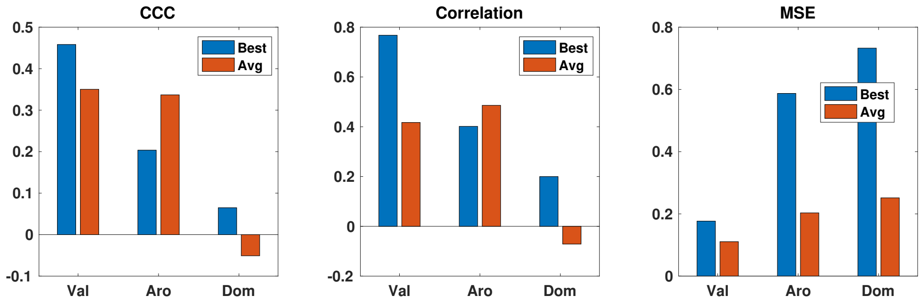
(b) Correlation

Figure 5.2: Performance of proposed and baseline models in predicting sentence level norms

across all regression models explored is reported.

5.5 Results

Figure 5.2 shows the performance of the proposed model and the the different baselines. As seen from the figure, the proposed model outperforms the baselines in predicting valence and arousal in both evaluation metrics, suggesting the efficacy of the approach. Using partial ratings at sentence level along with matrix F_k which captures relationships between the dimensions, the proposed approach seems to outperform the baseline word aggregation schemes in these two dimensions. On the other hand, model performance on dominance appears considerably low in both metrics. To further investigate the reason for this, we created plots for each dimension comparing the best

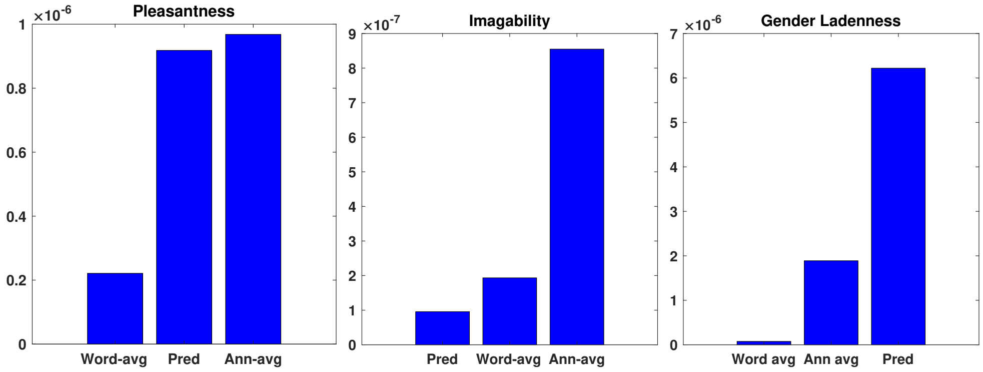


(a) CCC (b) Correlation (c) MSE

Figure 5.3: Performance of best annotator in our dataset and annotator average

possible annotator in our dataset and the average rating across all annotators as shown in Figure 5.3. Evidently, for dominance, we notice very low values for the two correlation metrics and high MSE, suggesting a high disagreement between our annotators and those from the Emobank corpus for this dimension. This may have been due to a possibly differing definition and/or interpretation of dominance between the two sets of annotators. Addressing this is likely to improve the quality of performance of the proposed model in predicting dominance.

Figure 5.4 shows the training set MSE for our experiment on pleasantness, imagability and genderladenness. As seen in the figure, the proposed approach achieves the best performance in at least one dimension (imagability), warranting further explorations for these dimensions, perhaps by collecting expert ratings.



(a) Pleasantness (b) Imagability (c) Genderladenness

Figure 5.4: Training error on labels predicted; Pred: Predictions from proposed model, Word-avg: Average of word level norm scores; Ann avg: annotator avg

5.6 Conclusion

We presented a novel approach to estimate sentence level psycholinguistic norms and showed improvements over standard baselines. We evaluate our approach on annotations of valence, arousal and dominance. Future work includes evaluating the model on other dimensions such as pleasantness. The primary challenge lies in obtaining expert ratings on these dimensions at the sentence level. Recently, alternate schemes to evaluate the model in the absence of a reliable ground truth or reference have been proposed, such as the evaluation strategy used in the AVEC 2018 challenge [56]. The challenge organizers proposed a scheme where annotation fusion models are evaluated by training and testing baseline regression models on the predicted labels from the fusion models on disjoint sets. High performance on the test set suggests *consistent learnability* of the predicted models and can act as a proxy for label quality. We aim to expand our annotation experiments on other psycholinguistic norms and use this strategy to evaluate our approach in future work.

Chapter 6

Conclusions and Future Work

In this dissertation, we presented our work on multidimensional annotation fusion for subjective ratings on affective dimensions. We presented two latent variable models which used additive Gaussian noise and a matrix factorization model respectively to capture the annotators' distortion functions. We then applied the matrix factorization model to the task of predicting psycholinguistic norms at the sentence level and showed improved results compared to baseline models which aggregate word level scores. We also recognized appropriate future works for some of the tasks described above. We now present our proposed future work to the task of computing agreement on multidimensional annotations.

6.1 Multidimensional annotation agreement

Computing agreement between annotators is an important step in most data collection projects as it provides a measure of reliability of the annotators.

Agreement is usually measured using a single numeric score with a high value suggesting high quality labels. However, most existing strategies to compute agreement are limited to univariate settings. In the case of multivariate annotations, the common practice is to compute agreement for each dimension separately and report an array of agreement scores which can be cumbersome or to report a suitable aggregate such as the average or median agreement score which discards useful information. To address this, we propose to develop a new metric which provides a single numeric score to capture the agreement between annotators in the multidimensional setting. Specifically, we propose to extend Cohen's κ [57], one of the most frequently used metric to compute agreement between annotators.

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{6.1}$$

where, P_o is the observed agreement

P_e is the expected agreement due to chance

Equation 6.1 shows the formula to compute kappa, which is given by the ratio of observed agreement beyond chance over the best possible observable agreement (equal to 1) beyond chance. To extend this, we use a reformulation of κ given by [58] shown in Equation 6.2. This variant of κ is obtained by subtracting the ratio of observed disagreement δ over the expected disagreement due to chance μ_δ from 1.

$$\kappa = 1 - \frac{\delta}{\mu_\delta} \tag{6.2}$$

$$\text{where, } \delta = \frac{1}{m}\sum_{m} \Delta(\mathbf{a}_1^m, \mathbf{a}_2^m)$$

$$\mu_\delta = \frac{1}{m^2}\sum_{m1}\sum_{m2} \Delta(\mathbf{a}_1^{m1}, \mathbf{a}_2^{m2})$$

$$\Delta \text{ is a distance measure}$$

As seen in Equation 6.2, both δ and μ_δ use a distance measure Δ to compute disagreement. In the formulation of [58], Euclidean (l_2) distance is used to measure the disagreement but it is not clear if it is the optimal choice. We propose to expand on this work by evaluating other distance measures such as l_1 (Equation 6.3) or l_∞ (Equation 6.4) distances. Each of these have specific advantages over the l_2 distance. For example, use of l_∞ avoids over penalizing differences in annotator scales in cases where the annotators operate on different internal ranges. Similarly, use of l_1 distance may lead to κ distributions with lower entropy, and this could lead to distinctly defined regions which is often desirable in agreement metrics. Our proposed work includes exploring various distance measures to compute multidimensional agreement from the formula listed in Equation 6.2 and draw comparisions between them.

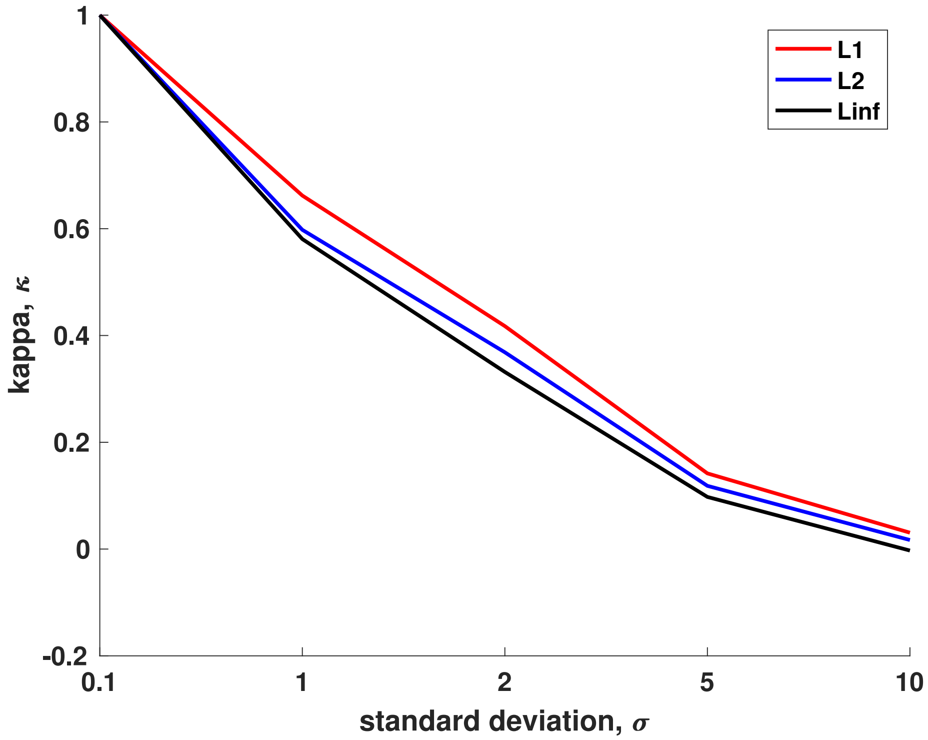


Figure 6.1: Comparision of κ for various distance measures

$$L_1 = \sum_d |a^m_{1,d} - a^m_{2,d}| \tag{6.3}$$

$$L_\infty = \max_d |a^m_{1,d} - a^m_{2,d}| \tag{6.4}$$

Evaluating the quality of agreement obtained from the different distance measures is challenging since they often behave comparably. To illustrate this, we created two synthetic annotators who differ only by additive Gaussian noise and created plots of estimated κ as we increase the standard deviation of the additive noise, for the different distance measures described above. As seen in Figure 6.1, in this annotation experiment, all three distance measures lead to similar decrease in agreements and it is unclear if one is superior to another. To drawing better comparisons, we would need more carefully designed annotation experiments which highlight the key differences between the distance measures, and we propose to explore this further in our future work.

Bibliography

[1] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy, "Learning from crowds," *The Journal of Machine Learning Research*, vol. 11, pp. 1297–1322, 2010.

[2] M. D. Munezero, C. S. Montero, E. Sutinen, and J. Pajunen, "Are they different? affect, feeling, emotion, sentiment, and opinion detection in text," *IEEE transactions on affective computing*, vol. 5, no. 2, pp. 101–111, 2014.

[3] K. S. Fleckenstein, "Defining affect in relation to cognition: A response to susan mcleod," *Journal of Advanced Composition*, vol. 11, no. 2, pp. 447–453, 1991. [Online]. Available: http://www.jstor.org/stable/20865808

[4] U. Schimmack, S. Oishi, and E. Diener, "Cultural influences on the relation between pleasant emotions and unpleasant emotions: Asian dialectic philosophies or individualism-collectivism?" *Cognition & Emotion*, vol. 16, no. 6, pp. 705–719, 2002.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

[6] J. Surowiecki, *The wisdom of crowds.* Anchor, 2005.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on.* IEEE, 2009, pp. 248–255.

[8] C. Vondrick, D. Patterson, and D. Ramanan, "Efficiently scaling up crowdsourced video annotation," *International Journal of Computer Vision*, vol. 101, no. 1, pp. 184–204, 2013.

[9] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng, "Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks," in *Proceedings of the conference on empirical methods in natural language processing.* Association for Computational Linguistics, 2008, pp. 254–263.

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

[11] P. Smyth, U. M. Fayyad, M. C. Burl, P. Perona, and P. Baldi, "Inferring ground truth from subjective labelling of venus images," in *Advances in neural information processing systems*, 1995, pp. 1085–1092.

[12] K. Audhkhasi and S. Narayanan, “A globally-variant locally-constant model for fusion of labels from multiple diverse experts without using reference labels,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 4, pp. 769–783, 2013.

[13] A. Metallinou and S. S. Narayanan, “Annotation and processing of continuous emotional attributes: Challenges and opportunities,” in *2nd International Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE 2013)*, apr 2013.

[14] M. A. Nicolaou, V. Pavlovic, and M. Pantic, “Dynamic probabilistic cca for analysis of affective behaviour,” in *European Conference on Computer Vision*. Springer, 2012, pp. 98–111.

[15] F. R. Bach and M. I. Jordan, “A probabilistic interpretation of canonical correlation analysis,” University of California, Berkeley, Tech. Rep., 2005.

[16] M. A. Nicolaou, V. Pavlovic, and M. Pantic, “Dynamic probabilistic cca for analysis of affective behavior and fusion of continuous annotations,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 7, pp. 1299–1311, 2014.

[17] S. Mariooryad and C. Busso, “Correcting time-continuous emotional labels by modeling the reaction lag of evaluators,” *IEEE Transactions on Affective Computing*, vol. 6, no. 2, pp. 97–108, 2015.

[18] R. Gupta, K. Audhkhasi, Z. Jacokes, A. Rozga, and S. Narayanan, "Modeling multiple time series annotations based on ground truth inference and distortion," *IEEE Transactions on Affective Computing*, 2016.

[19] Y. E. Kara, G. Genc, O. Aran, and L. Akarun, "Modeling annotator behaviors for crowd labeling," *Neurocomputing*, vol. 160, pp. 141–156, 2015.

[20] N. B. Shah, D. Zhou, and Y. Peres, "Approval voting and incentives in crowdsourcing," *arXiv preprint arXiv:1502.05696*, 2015.

[21] V. S. Sheng, F. Provost, and P. G. Ipeirotis, "Get another label? improving data quality and data mining using multiple, noisy labelers," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008, pp. 614–622.

[22] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, p. 335, 2008.

[23] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 5–17, 2012.

[24] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *Automatic Face and Gesture Recognition (FG), 2013 10th*

IEEE International Conference and Workshops on. IEEE, 2013, pp. 1–8.

[25] N. Malandrakis, A. Potamianos, G. Evangelopoulos, and A. Zlatintsi, "A supervised approach to movie emotion tracking," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on.* IEEE, 2011, pp. 2376–2379.

[26] M. Franzini, K.-F. Lee, and A. Waibel, "Connectionist viterbi training: A new hybrid method for continuous speech recognition," in *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on.* IEEE, 1990, pp. 425–428.

[27] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[28] S. Narayanan and P. G. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1203–1233, 2013.

[29] C. M. Bishop, *Pattern Recognition and Machine Learning.* Springer, 2006.

[30] R. B. Grossman, L. R. Edelson, and H. Tager-Flusberg, "Emotional facial and vocal expressions during story retelling by children and adolescents with high-functioning autism," *Journal of Speech, Language, and Hearing Research*, vol. 56, no. 3, pp. 1035–1044, 2013.

[31] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge." in *INTERSPEECH*, vol. 2009. Citeseer, 2009, pp. 312–315.

[32] R. Gupta, C.-C. Lee, and S. S. Narayanan, "Classification of emotional content of sighs in dyadic human interactions," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2012.

[33] V. I. Spitkovsky, H. Alshawi, D. Jurafsky, and C. D. Manning, "Viterbi training improves unsupervised dependency parsing," in *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, 2010, pp. 9–17.

[34] I. Lawrence and K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, pp. 255–268, 1989.

[35] L. R. Fabrigar, D. T. Wegener, R. C. MacCallum, and E. J. Strahan, "Evaluating the use of exploratory factor analysis in psychological research." *Psychological methods*, vol. 4, no. 3, p. 272, 1999.

[36] H. F. Kaiser, "The varimax criterion for analytic rotation in factor analysis," *Psychometrika*, vol. 23, no. 3, pp. 187–200, 1958.

[37] B. Efron and R. J. Tibshirani, *An introduction to the bootstrap*. CRC press, 1994.

[38] B. M. Booth, K. Mundnich, and S. S. Narayanan, "A novel method for human bias correction of continuous-time annotations," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 3091–3095.

[39] D. Corney, J.-D. Haynes, G. Rees, and R. B. Lotto, "The brightness of colour," *PloS one*, vol. 4, no. 3, p. e5091, 2009.

[40] C. Strapparava and R. Mihalcea, "Semeval-2007 task 14: Affective text," in *Proceedings of the 4th International Workshop on Semantic Evaluations*, ser. SemEval '07. Stroudsburg, PA, USA: Association for Computational Linguistics, 2007, pp. 70–74. [Online]. Available: http://dl.acm.org/citation.cfm?id=1621474.1621487

[41] M. Pagliardini, P. Gupta, and M. Jaggi, "Unsupervised learning of sentence embeddings using compositional n-gram features," *CoRR*, vol. abs/1703.02507, 2017. [Online]. Available: http://arxiv.org/abs/1703.02507

[42] M. Quigley, B. Gerkey, K. Conley, J. Faust, T. Foote, J. Leibs, E. Berger, R. Wheeler, and A. Ng, "Ros: an open-source robot operating system," in *Proc. of the IEEE Intl. Conf. on Robotics and Automation (ICRA) Workshop on Open Source Robotics*, Kobe, Japan, May 2009.

[43] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder, "FEELtrace: An instrument for recording perceived emotion in real time," in *ISCA tutorial and research workshop (ITRW) on speech and emotion*, 2000.

[44] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, ser. MM '10.

New York, NY, USA: ACM, 2010, pp. 1459–1462. [Online]. Available: http://doi.acm.org/10.1145/1873951.1874246

[45] G. Bradski, “The OpenCV Library,” *Dr. Dobb’s Journal of Software Tools*, 2000.

[46] L. S. Chen, T. S. Huang, T. Miyasato, and R. Nakatsu, “Multimodal human emotion/expression recognition,” in *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on.* IEEE, 1998, pp. 366–371.

[47] A. Paivio, J. C. Yuille, and S. A. Madigan, “Concreteness, imagery, and meaningfulness values for 925 nouns.” *Journal of experimental psychology*, vol. 76, no. 1p2, p. 1, 1968.

[48] S. Tanaka, A. Jatowt, M. P. Kato, and K. Tanaka, “Estimating content concreteness for finding comprehensible documents,” in *Proceedings of the sixth ACM international conference on Web search and data mining.* ACM, 2013, pp. 475–484.

[49] F. Å. Nielsen, “A new anew: Evaluation of a word list for sentiment analysis in microblogs,” *arXiv preprint arXiv:1103.2903*, 2011.

[50] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore, “Using linguistic cues for the automatic recognition of personality in conversation and text,” *Journal of artificial intelligence research*, vol. 30, pp. 457–500, 2007.

[51] A. Ramakrishna, V. R. Martínez, N. Malandrakis, K. Singla, and S. Narayanan, “Linguistic analysis of differences in portrayal of movie

characters," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 1669–1678.

[52] N. Malandrakis and S. S. Narayanan, "Therapy language analysis using automatically generated psycholinguistic norms." in *INTERSPEECH*, 2015, pp. 1952–1956.

[53] J. Gibson, N. Malandrakis, F. Romero, D. C. Atkins, and S. S. Narayanan, "Predicting therapist empathy in motivational interviews using language features inspired by psycholinguistic norms," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[54] A. B. Warriner, V. Kuperman, and M. Brysbaert, "Norms of valence, arousal, and dominance for 13,915 english lemmas," *Behavior research methods*, vol. 45, no. 4, pp. 1191–1207, 2013.

[55] S. Buechel and U. Hahn, "Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 2017, pp. 578–585.

[56] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, H. Kaya, M. Schmitt, S. Amiriparian, N. Cummins, D. Lalanne, A. Michaud *et al.*, "Avec 2018 workshop and challenge: Bipolar disorder and cross-cultural af-

fect recognition," in *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop.* ACM, 2018, pp. 3–13.

[57] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[58] K. J. Berry and P. W. Mielke Jr, "A generalization of cohen's kappa agreement measure to interval measurement and multiple raters," *Educational and Psychological Measurement*, vol. 48, no. 4, pp. 921–933, 1988.

Appendix A

Derivations for the matrix factorization model

A.1 EM update equations for global annotation model

A.1.1 Components of the joint distribution $p(\mathbf{a}_1^m \ldots \mathbf{a}_K^m, \mathbf{a}_*^m)$

To help with the model formulation, we first derive parameters of the joint distribution $p(\mathbf{a}_1^m \ldots \mathbf{a}_K^m, \mathbf{a}_*^m)$. Since the product of two normal distributions is also normal [29], this joint distribution is also normal and is given by,

$$\begin{bmatrix} \mathbf{a}_*^m \\ \mathbf{a}_1^m \\ \vdots \\ \mathbf{a}_K^m \end{bmatrix} \sim N \left(\begin{bmatrix} \Theta^T \mathbf{x}_m \\ F_1 \Theta^T \mathbf{x}_m \\ \vdots \\ F_K \Theta^T \mathbf{x}_m \end{bmatrix}, \begin{bmatrix} \Sigma_{**} & \Sigma_{*1} & \dots & \Sigma_{*K} \\ \Sigma_{1*} & \Sigma_{11} & \dots & \Sigma_{1K} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{K*} & \Sigma_{K1} & \dots & \Sigma_{KK} \end{bmatrix} \right) \tag{A.1}$$

The different components of the covariance matrix from Equation A.1 are derived below.

$$\begin{aligned}
\Sigma_{**} &= Cov(\mathbf{a}_*^m) \\
&= \sigma_*^2 I \\
\Sigma_{k*} &= \mathbb{E}[\mathbf{a}_k^m (\mathbf{a}_*^m)^T] - \mathbb{E}[\mathbf{a}_k^m]\, \mathbb{E}[(\mathbf{a}_*^m)^T] \\
&= \mathbb{E}[(F_k \mathbf{a}_*^m + \boldsymbol{\eta}_k)(\mathbf{a}_*^m)^T] - \mathbb{E}[F_k \mathbf{a}_*^m + \boldsymbol{\eta}_k]\, \mathbb{E}[(\mathbf{a}_*^m)^T] \\
&= F_k(\sigma_*^2 I) \\
\Sigma_{kk} &= Cov(F_k \mathbf{a}_*^m + \boldsymbol{\eta}_k) \\
&= Cov(F_k \mathbf{a}_*^m) + \tau_k^2 I \\
&= F_k \Sigma_{**} F_k^T + \tau_k^2 I \\
&= \sigma_*^2 F_k F_k^T + \tau_k^2 I \\
\Sigma_{k_i k_j} &= \mathbb{E}_{\mathbf{a}_*^m}[Cov(\mathbf{a}_{k_1}^m, \mathbf{a}_{k_2}^m | \mathbf{a}_*^m)] + Cov(\mathbb{E}[\mathbf{a}_{k_1}^m | \mathbf{a}_*^m], \mathbb{E}[\mathbf{a}_{k_2}^m | \mathbf{a}_*^m]) \\
&= Cov(\mathbb{E}[\mathbf{a}_{k_1}^m | \mathbf{a}_*^m], \mathbb{E}[\mathbf{a}_{k_2}^m | \mathbf{a}_*^m]) \\
&= Cov(F_{k_1} \mathbf{a}_*^m, F_{k_2} \mathbf{a}_*^m) \\
&= F_{k_1} \Sigma_{**} (F_{k_2})^T \\
&= \sigma_*^2 F_{k_1} F_{k_2}^T
\end{aligned}$$

In the derivation of $\Sigma_{k_i k_j}$, the first equation is a direct application of the law of total covariance and the second equation is because of the conditional independence assumption of annotation values $\mathbf{a}_{k_i}^m$ given the ground truth $\mathbf{a}_*^m$

Finally, owing to the jointly normal distributions, $p(\mathbf{a}_*^m|\mathbf{a}_1^m \dots \mathbf{a}_K^m)$ is also normal:

$$p(\mathbf{a}_*^m|\mathbf{a}_1^m \dots \mathbf{a}_K^m) \sim N(\mu_{\mathbf{a}_*^m|\mathbf{a}_1^m \dots \mathbf{a}_K^m}|\Sigma_{\mathbf{a}_*^m|\mathbf{a}_1^m \dots \mathbf{a}_K^m})$$

Also, by definitions of conditional normal distributions, given a normal vector of the form

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right)$$

the conditional distribution $p(x_1|x_2) \sim N(\mu_{x_1|x_2}, \Sigma_{x_1|x_2})$ has the following form.

$$\mu_{x_1|x_2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \tag{A.2}$$

$$\Sigma_{x_1|x_2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \tag{A.3}$$

A.1.2 Model formulation

We begin by introducing a new distribution $q(\mathbf{a}_*^m)$ in Equation 4.5. We drop the parameters Φ from the likelihood function expansion for convenience.

$$\log \mathcal{L} = \sum_{m=1}^{M} \log \int_{\mathbf{a}_*^m} q(\mathbf{a}_*^m) \frac{p(\mathbf{a}_1^m \dots \mathbf{a}_K^m|\mathbf{a}_*^m)p(\mathbf{a}_*^m)}{q(\mathbf{a}_*^m)}\, d\mathbf{a}_*^m \tag{A.4}$$

Using Jensen's inequality over log of expectation, we can write the above as follows,

$$\log \mathcal{L} \geq \sum_{m=1}^{M} \int_{\mathbf{a}_*^m} q(\mathbf{a}_*^m) \log \frac{p(\mathbf{a}_1^m \ldots \mathbf{a}_K^m | \mathbf{a}_*^m) p(\mathbf{a}_*^m)}{q(\mathbf{a}_*^m)} \, d\mathbf{a}_*^m \tag{A.5}$$

The bound above becomes tight when the expectation is taken over a constant value, i.e.

$$\frac{p(\mathbf{a}_1^m \ldots \mathbf{a}_K^m | \mathbf{a}_*^m) p(\mathbf{a}_*^m)}{q(\mathbf{a}_*^m)} = c$$

Solving for the constant c, we have

$$q(\mathbf{a}_*^m) = \frac{p(\mathbf{a}_1^m \ldots \mathbf{a}_K^m, \mathbf{a}_*^m)}{p(\mathbf{a}_1^m \ldots \mathbf{a}_K^m)} = p(\mathbf{a}_*^m | \mathbf{a}_1^m \ldots \mathbf{a}_K^m)$$

E-Step

The E-step involves simply assuming $q(\mathbf{a}_*^m)$ to follow the conditional distribution $p(\mathbf{a}_*^m | \mathbf{a}_1^m \ldots \mathbf{a}_K^m)$.

To help with future computations, we also compute the following expectations, where the first two are a result of equations A.2 and A.3; third equation is by definition of covariance and the last one is a standard result.

$$\mathbb{E}_{\mathbf{a}_*^m | \mathbf{a}_1^m \ldots \mathbf{a}_K^m}[\mathbf{a}_*^m] = \Theta^T \mathbf{x}_m + \Sigma_{\mathbf{a}_*^m, \mathbf{a}_1^m \ldots \mathbf{a}_K^m} (\Sigma_{\mathbf{a}_1^m \ldots \mathbf{a}_K^m, \mathbf{a}_1^m \ldots \mathbf{a}_K^m})^{-1} (\mathbf{a}^m - \boldsymbol{\mu}^m)$$

$$\Sigma_{\mathbf{a}_*^m | \mathbf{a}_1^m \ldots \mathbf{a}_K^m}[\mathbf{a}_*^m] = \Sigma_{\mathbf{a}_*^m, \mathbf{a}_*^m} - \Sigma_{\mathbf{a}_*^m, \mathbf{a}_1^m \ldots \mathbf{a}_K^m} (\Sigma_{\mathbf{a}_1^m \ldots \mathbf{a}_K^m, \mathbf{a}_1^m \ldots \mathbf{a}_K^m})^{-1} \Sigma_{\mathbf{a}_1^m \ldots \mathbf{a}_K^m, \mathbf{a}_*^m}$$

$$\mathbb{E}_{\mathbf{a}_*^m | \mathbf{a}_1^m \ldots \mathbf{a}_K^m}[\mathbf{a}_*^m (\mathbf{a}_*^m)^T] = \Sigma_{\mathbf{a}_*^m | \mathbf{a}_1^m \ldots \mathbf{a}_K^m}[\mathbf{a}_*^m] + \mathbb{E}_{\mathbf{a}_*^m | \mathbf{a}_1^m \ldots \mathbf{a}_K^m}[\mathbf{a}_*^m] \, \mathbb{E}_{\mathbf{a}_*^m | \mathbf{a}_1^m \ldots \mathbf{a}_K^m}[(\mathbf{a}_*^m)^T]$$

$$\mathbb{E}_{\mathbf{a}_*^m | \mathbf{a}_1^m \ldots \mathbf{a}_K^m}[(\mathbf{a}_*^m)^T \mathbf{a}_*^m] = trace(\Sigma_{\mathbf{a}_*^m | \mathbf{a}_1^m \ldots \mathbf{a}_K^m}[\mathbf{a}_*^m]) + \mathbb{E}_{\mathbf{a}_*^m | \mathbf{a}_1^m \ldots \mathbf{a}_K^m}[(\mathbf{a}_*^m)^T] \, \mathbb{E}_{\mathbf{a}_*^m | \mathbf{a}_1^m \ldots \mathbf{a}_K^m}[\mathbf{a}_*^m]$$

$\mathbf{a}^m$ and $\boldsymbol{\mu}^m$ are DK dimensional vectors obtained by concatenating the K annotation vectors $\mathbf{a}_1^m, \ldots \mathbf{a}_K^m$ and their corresponding expected values $F_1 \Theta^T \mathbf{x}_m \ldots F_K \Theta^T \mathbf{x}_m$.

M-step

In the M-step, we find the parameters of the model by maximizing Equation A.5. We first write this equation as an expectation and an equality. The expectation below is with respect to $q(\mathbf{a}_*^m) = p(\mathbf{a}_*^m | \mathbf{a}_1^m \ldots \mathbf{a}_K^m)$; we drop the subscript for ease of exposition

$$\begin{aligned}
\log \mathcal{L} &= \sum_{m=1}^{M} \mathbb{E}_{\mathbf{a}_*^m | \mathbf{a}_1^m \ldots \mathbf{a}_K^m} \big[\log \frac{p(\mathbf{a}_1^m \ldots \mathbf{a}_K^m | \mathbf{a}_*^m) p(\mathbf{a}_*^m)}{q(\mathbf{a}_*^m)} \big] \\
\log \mathcal{L} &= \sum_{m=1}^{M} \mathbb{E} \log p(\mathbf{a}_1^m \ldots \mathbf{a}_K^m | \mathbf{a}_*^m) + \mathbb{E} \log p(\mathbf{a}_*^m) + \mathcal{H} \\
\log \mathcal{L} &= \sum_{m=1}^{M} \big(\sum_{k=1}^{K} \mathbb{E} \log p(\mathbf{a}_k^m | \mathbf{a}_*^m) + \mathbb{E} \log p(\mathbf{a}_*^m) + \mathcal{H} \big) \qquad \text{(A.6)}
\end{aligned}$$

where $p(\mathbf{a}_*^m)$ and $p(\mathbf{a}_k^m | \mathbf{a}_*^m)$ are given by equations 4.3 and 4.4 respectively. The last equation above uses that fact that we assume independence among annotators given the ground truth. Also expectation commutes with the linear sum over the K terms.

Here, $\mathcal{H}$ is the entropy of $p(\mathbf{a}_*^m | \mathbf{a}_1^m \ldots \mathbf{a}_K^m)$. We maximize Equation A.6 with respect to each of the parameters to obtain the M-step updates.

Estimating $\boldsymbol{F_k}$ Differentiating Equation A.6 with respect to F_k and equating the derivative to 0

$$\begin{aligned}
\Delta_{F_k} Q &= 0 \\
\Delta_{F_k} \sum_{m=1}^{M_k} \mathbb{E}[(\mathbf{a}_k^m - F_k \mathbf{a}_*^m)^T (\tau_k^2 I)^{-1} (\mathbf{a}_k^m - F_k \mathbf{a}_*^m)] &= 0 \\
\Delta_{F_k} \frac{1}{\tau_k^2} \sum_{m=1}^{M_k} \mathbb{E}[(\mathbf{a}_k^m - F_k \mathbf{a}_*^m)^T (\mathbf{a}_k^m - F_k \mathbf{a}_*^m)] &= 0 \\
\sum_{m=1}^{M_k} -2\mathbf{a}_k^m \, \mathbb{E}[(\mathbf{a}_*^m)^T] + 2F_k \, \mathbb{E}[\mathbf{a}_*^m (\mathbf{a}_*^m)^T] &= 0 \\
\therefore F_k = \left(\sum_{m=1}^{M_k} \mathbf{a}_k^m \, \mathbb{E}[(\mathbf{a}_*^m)^T] \right) &\left(\sum_{m=1}^{M_k} \mathbb{E}[\mathbf{a}_*^m (\mathbf{a}_*^m)^T] \right)^{-1}
\end{aligned}$$

where, M_k is the number of points annotated by user k.

We used the following facts in the above derivation: $trace(x) = x$ for scalar x; $trace(AB) = trace(BA)$; $\Delta_A trace(A^T x) = x$ and $\Delta_A trace(A^T AB) = AB + AB^T$ for matrix A. We also make use of the fact that expectation and trace of a matrix are commutative since trace is a linear sum.

Estimating $\mathbf{\Theta}$ Similarly, to find Θ, we differentiate Equation A.6 with respect to Θ and equate it to 0.

$$\begin{aligned}
\Delta_\Theta Q &= 0 \\
\Delta_\Theta \sum_{m=1}^{M} \mathbb{E}[(\mathbf{a}_*^m - \Theta^T \mathbf{x}_m)^T (\sigma^2 I)^{-1} (\mathbf{a}_*^m - \Theta^T \mathbf{x}_m)] &= 0 \\
\Delta_\Theta \frac{1}{\sigma^2} \sum_{m=1}^{M} \mathbb{E}[(\mathbf{a}_*^m - \Theta^T \mathbf{x}_m)^T (\mathbf{a}_*^m - \Theta^T \mathbf{x}_m)] &= 0 \\
\sum_{m=1}^{M} -2\mathbf{x}_m \, \mathbb{E}[(\mathbf{a}_*^m)^T] + 2\mathbf{x}_m \mathbf{x}_m^T \Theta &= 0
\end{aligned}$$

$$\Theta = \Big(\sum_{m=1}^{M} \mathbf{x}_m \mathbf{x}_m^T\Big)^{-1} \Big(\sum_{m=1}^{M} \mathbf{x}_m \, \mathbb{E}[(\mathbf{a}_*^m)^T]\Big)$$

$$\therefore \Theta = (\mathrm{X}^T\mathrm{X})^{-1}(\mathrm{X}^T \, \mathbb{E}[\mathbf{a}_*^m])$$

which looks like the familiar normal equation except we use the expected value of $\mathbf{a}_*$. Here, X is the matrix of features of the M data points; it includes individual feature vectors x_m in its rows.

Estimating $\boldsymbol{\sigma}$ Differentiating Equation A.6 with respect to σ and equating to 0, we have

$$\Delta_\sigma Q = 0$$

$$\Delta_\sigma \sum_{m=1}^{M} \Big(- D \log \sigma - \frac{1}{2\sigma^2}\big(\mathbb{E}[(\mathbf{a}_*^m)^T \mathbf{a}_*^m] - 2tr(\Theta^T \mathbf{x}_m \, \mathbb{E}[(\mathbf{a}_*^m)^T]) + tr(\mathbf{x}_m^T \Theta \Theta^T \mathbf{x}_m)\big)\Big) = 0$$

$$\sum_{m=1}^{M} -\frac{D}{\sigma} + \frac{1}{\sigma^3}\Big(\mathbb{E}[(\mathbf{a}_*^m)^T \mathbf{a}_*^m] - 2tr(\Theta^T \mathbf{x}_m \, \mathbb{E}[(\mathbf{a}_*^m)^T]) + tr(\mathbf{x}_m^T \Theta \Theta^T \mathbf{x}_m)\Big) = 0$$

$$\frac{MD}{\sigma} = \frac{1}{\sigma^3} \sum_{m=1}^{M} \Big(\mathbb{E}[(\mathbf{a}_*^m)^T \mathbf{a}_*^m] - 2tr\big(\Theta^T \mathbf{x}_m \, \mathbb{E}[(\mathbf{a}_*^m)^T]\big) + tr(\mathbf{x}_m^T \Theta \Theta^T \mathbf{x}_m)\Big)$$

$$\therefore \sigma^2 = \frac{1}{MD} \sum_{m=1}^{M} \Big(\mathbb{E}[(\mathbf{a}_*^m)^T \mathbf{a}_*^m] - 2tr\big(\Theta^T \mathbf{x}_m \, \mathbb{E}[(\mathbf{a}_*^m)^T]\big) + tr(\mathbf{x}_m^T \Theta \Theta^T \mathbf{x}_m)\Big)$$

Estimating $\boldsymbol{\tau_k}$ Differentiating Equation A.6 with respect to τ_k and equating to 0, we have

$$\Delta_{\tau_k} Q = 0$$

$$\Delta_{\tau_k} \sum_{m=1}^{M_k} \bigg(- D \log \tau_k - \frac{1}{2\tau_k^2} \big((\mathbf{a}_k^m)^T \mathbf{a}_k^m - 2tr(F_k^T \mathbf{a}_k^m \, \mathbb{E}[(\mathbf{a}_*^m)^T]) + \\ tr(F_k^T F_k \, \mathbb{E}[\mathbf{a}_*^m (\mathbf{a}_*^m)^T]) \big) \bigg) = 0$$

$$\sum_{m=1}^{M_k} \bigg(- \frac{D}{\tau_k} + \frac{1}{\tau_k^3} \big((\mathbf{a}_k^m)^T \mathbf{a}_k^m - 2tr(F_k^T \mathbf{a}_k^m \, \mathbb{E}[(\mathbf{a}_*^m)^T]) + tr(F_k^T F_k \, \mathbb{E}[\mathbf{a}_*^m (\mathbf{a}_*^m)^T]) \big) \bigg) = 0$$

$$\therefore \tau_k^2 = \frac{1}{DM_k} \sum_{m=1}^{M_k} \bigg((\mathbf{a}_k^m)^T \mathbf{a}_k^m - 2tr(F_k^T \mathbf{a}_k^m \, \mathbb{E}[(\mathbf{a}_*^m)^T]) + tr(F_k^T F_k \, \mathbb{E}[\mathbf{a}_*^m (\mathbf{a}_*^m)^T]) \bigg)$$

A.2 EM update equations for time series annotation model

A.2.1 Model formulation

Similar to the process described in Appendix A.1, the log likelihood function for the time series model is shown below (similar to Equation A.5).

$$\log \mathcal{L} \geq \sum_{m=1}^{M} \int_{\mathbf{a}_*^m} q(\mathbf{a}_*^m) \log \frac{p(\mathbf{a}_1^m \ldots \mathbf{a}_K^m | \mathbf{a}_*^m) p(\mathbf{a}_*^m)}{q(\mathbf{a}_*^m)} \, d\mathbf{a}_*^m \tag{A.7}$$

The bound becomes tight when $q(\mathbf{a}_*^m) = p(\mathbf{a}_*^m | \mathbf{a}_1^m \ldots \mathbf{a}_K^m)$.

E-step

Computing the expectation function over the entire distribution of $q(\mathbf{a}_*^m)$ is computationally expensive since $\mathbf{a}_*^m$ is a matrix. To avoid this, we instead use *Hard-EM* in which we assume a dirac-delta distribution for $\mathbf{a}_*^m$ which is centered at the mode of $q(\mathbf{a}_*^m)$. This is a common practice in latent models and is the approach followed by [18] in estimating the annotator filter parameters. We assign this value to $\mathbf{a}_*^m$ in the E-step:

$$\begin{aligned}
\mathbf{a}_*^m &= \operatorname*{argmax}_{\mathbf{a}_*^m} q(\mathbf{a}_*^m) \\
&= \operatorname*{argmax}_{\mathbf{a}_*^m} p(\mathbf{a}_*^m | \mathbf{a}_1^m, \ldots \mathbf{a}_K^m) \\
&= \operatorname*{argmax}_{\mathbf{a}_*^m} \frac{p(\mathbf{a}_*^m, \mathbf{a}_1^m, \ldots \mathbf{a}_K^m)}{p(\mathbf{a}_1^m, \ldots \mathbf{a}_K^m)} \\
&= \operatorname*{argmax}_{\mathbf{a}_*^m} p(\mathbf{a}_1^m, \ldots \mathbf{a}_K^m | \mathbf{a}_*^m) p(\mathbf{a}_*^m) \\
&= \operatorname*{argmax}_{\mathbf{a}_*^m} \log p(\mathbf{a}_1^m, \ldots \mathbf{a}_K^m | \mathbf{a}_*^m) p(\mathbf{a}_*^m) \\
\therefore \mathbf{a}_*^m &= \operatorname*{argmax}_{\mathbf{a}_*^m} \big(\log p(\mathbf{a}_1^m, \ldots \mathbf{a}_K^m | \mathbf{a}_*^m) + \log p(\mathbf{a}_*^m | \mathbf{x}^m) \big)
\end{aligned}$$

Since we assume that each annotator is independent of the others given the ground truth, we have

$$\begin{aligned}
\mathbf{a}_*^m &= \operatorname*{argmax}_{\mathbf{a}_*^m} \log \prod_k p(\mathbf{a}_k^m | \mathbf{a}_*^m) + \log p(\mathbf{a}_*^m) \\
\mathbf{a}_*^m &= \operatorname*{argmax}_{\mathbf{a}_*^m} \sum_k \log p(\mathbf{a}_k^m | \mathbf{a}_*^m) + \log p(\mathbf{a}_*^m)
\end{aligned}$$

Further, since each annotation dimension $\mathbf{a}_k^{m,d}$ is assumed to independent given $\mathbf{a}_*^m$, we have

$$\mathbf{a}_*^m = \operatorname*{argmax}_{\mathbf{a}_*^m} \sum_k \sum_d \log p(\mathbf{a}_k^{m,d} | \mathbf{a}_*^m) + \log p(\mathbf{a}_*^m)$$

Finally, since both $\mathbf{a}_k^{m,d}$ and $\mathbf{a}_*^m$ are defined using iid Gaussian noise, the above maximization problem is equivalent to the following minimization.

$$\mathbf{a}_*^m = \underset{\mathbf{a}_*^m}{\operatorname{argmin}} \sum_k \sum_d ||\mathbf{a}_k^{m,d} - F_k^d \text{vec}(\mathbf{a}_*^m)||_2^2 + ||\text{vec}(\mathbf{a}_*^m) - \text{vec}(\mathrm{X}_m \Theta)||_2^2$$

For convenience, we reshape $\mathbf{a}_*^m$ into a vector and optimize with respect to the flattened vector. If we choose $\text{vec}(\mathbf{a}_*^m) = v$ and $\text{vec}(\mathrm{X}_m\Theta) = y$, the objective becomes,

$$Q(v) = \sum_k \sum_d ||\mathbf{a}_k^{m,d} - F_k^d v||_2^2 + ||v - y||_2^2$$

Differentiating Q with respect to v and equating the gradient to 0, we get

$$\begin{aligned}
\Delta_v Q &= 0 \\
\Delta_v \sum_k \sum_d (\mathbf{a}_k^{m,d} - F_k^d v)^T (\mathbf{a}_k^{m,d} - F_k^d v) + \ (v-y)^T(v-y) &= 0 \\
\Delta_v \sum_k \sum_d (\mathbf{a}_k^{m,d})^T \mathbf{a}_k^{m,d} + v^T (F_k^d)^T F_k^d v - 2(\mathbf{a}_k^{m,d})^T F_k^d v + (v^T v - 2y^T v + y^T y) &= 0 \\
\sum_k \sum_d 2(F_k^d)^T F_k^d v - 2(F_k^d)^T \mathbf{a}_k^{m,d} + (2v - 2y) &= 0 \\
\therefore v = \left(\sum_k \sum_d (F_k^d)^T F_k^d + I \right)^{-1} \left(\sum_k \sum_d (F_k^d)^T \mathbf{a}_k^{m,d} + y \right)
\end{aligned}$$

We can extract $\mathbf{a}_*^m$ by reshaping v back into a matrix.

M-step

Given the point estimate for $\mathbf{a}_*^m$, the log-likelihood Equation A.7 can now be written as a function of the model parameters.

$$\log \mathcal{L} = \sum_{m=1}^{M} \sum_{k=1}^{K} \log p(\mathbf{a}_k^m | \mathbf{a}_*^m; F_k^d, \tau_k) + \log p(\mathbf{a}_*^m; \Theta, \sigma)$$

In the M-step, we optimize the above equation with respect to the parameters $\Phi = \{F_k, \tau_k, \Theta, \sigma\}$.

$$Q(F_k, \tau_k, \Theta, \sigma) = \sum_{m=1}^{M} \sum_{k=1}^{K} \log p(\mathbf{a}_k^m | \mathbf{a}_*^m; F_k^d, \tau_k) + \log p(\mathbf{a}_*^m; \Theta, \sigma) \qquad \text{(A.8)}$$

Estimating $\boldsymbol{F_k^d}$: Since each F_k^d is a filter matrix constructed from a vector $f_k^d \in \mathrm{I\!R}^{\mathrm{WD}}$, we differentiate Equation A.8 with respect to f_k^d.

$$\begin{aligned} \Delta_{f_k^d} Q &= 0 \\ \Delta_{f_k^d} \sum_{m=1}^{M_k} \log p(\mathbf{a}_k^m | \mathbf{a}_*^m; F_k^d, \tau_k) &= 0 \\ \Delta_{f_k^d} \sum_{m=1}^{M_k} ||\mathbf{a}_k^{m,d} - F_k^d \mathrm{vec}(\mathbf{a}_*^m)||_2^2 &= 0 \end{aligned}$$

In the last step we make use of the fact that $\mathbf{a}_k^m$ depends on $\mathbf{a}_*^m$ through Gaussian noise. We also discard all other dimensions $d' \neq d$ since these do not depend on f_k^d. To estimate f_k^d, we can rearrange $F_k^d \mathrm{vec}(\mathbf{a}_*^m)$ such that f_k^d is now the parameter vector of a linear regression problem with the independent variables represented by matrix A which is obtained by creating a filtering matrix out of vec($\mathbf{a}_*^m$). Hence, the optimization problem becomes

$$\Delta_{f_k^d} \sum_{m=1}^{M_k} ||\mathbf{a}_k^{m,d} - Af_k^d||_2^2 = 0$$
$$\therefore f_k^d = \Bigg(\sum_{m=1}^{M_k} A^T A \Bigg)^{-1} \Bigg(\sum_{m=1}^{M_k} A^T \mathbf{a}_k^{m,d} \Bigg)$$

Estimating $\boldsymbol{\tau_k}$ Differentiating Equation (A.8) with respect to τ_k and equating the gradient to 0, we have,

$$\Delta_{\tau_k} Q = 0$$
$$\Delta_{\tau_k} \sum_{m=1}^{M_k} \log p(\mathbf{a}_k^m | \mathbf{a}_*^m; F_k^d, \tau_k) = 0$$
$$\Delta_{\tau_k} \sum_{m=1}^{M_k} \sum_d \log \frac{1}{|2\pi\tau_k^2 I|^{\frac{1}{2}}} e^{-\frac{1}{2\tau_k^2}||\mathbf{a}_k^{m,d} - F_k^d \text{vec}(\mathbf{a}_*^m)||_2^2} = 0$$
$$\Delta_{\tau_k} \sum_{m=1}^{M_k} \sum_d -T \log \tau_k - \frac{1}{2\tau_k^2} ||\mathbf{a}_k^{m,d} - F_k^d \text{vec}(\mathbf{a}_*^m)||_2^2 = 0$$
$$\frac{-M_k DT}{\tau_k} + \frac{1}{\tau_k^3} \sum_{m=1}^{M_k} \sum_d ||\mathbf{a}_k^{m,d} - F_k^d \text{vec}(\mathbf{a}_*^m)||_2^2 = 0$$
$$\therefore \tau_k^2 = \frac{1}{M_k DT} \sum_{m=1}^{M_k} \sum_d ||\mathbf{a}_k^{m,d} - F_k^d \text{vec}(\mathbf{a}_*^m)||_2^2$$

Estimating $\boldsymbol{\Theta}$ Differentiating Equation A.8 with respect to Θ and equating the gradient to 0, we have.

$$\Delta_\Theta Q = 0$$

$$\Delta_\Theta \sum_{m=1}^{M} ||\text{vec}(\mathbf{a}_*^m) - \text{vec}(\mathrm{X}_m\Theta)||_2^2 = 0$$

By definition, each column of Θ is independent of each other. Hence we can estimate each θ^d separately (taking derivatives with respect to above equation would cancel all terms except those in θ^d).

$$\Delta_{\theta^d} \sum_{m=1}^{M} (\mathbf{a}_*^{m,d} - \mathrm{X}_m\theta^d)^T(\mathbf{a}_*^{m,d} - \mathrm{X}_m\theta^d) = 0$$

$$\Delta_\Theta \sum_{m=1}^{M} (\mathbf{a}_*^{m,d})^T(\mathbf{a}_*^{m,d}) - 2(\mathbf{a}_*^{m,d})^T\mathrm{X}_m\theta^d + (\theta^d)^T\mathrm{X}_m^T\mathrm{X}_m\theta^d = 0$$

$$\theta^d = \left(\sum_{m=1}^{M} \mathrm{X}_m^T\mathrm{X}_m\right)^{-1}\left(\sum_{m=1}^{M} \mathrm{X}_m^T\mathbf{a}_*^{m,d}\right)$$

We can combine the estimation of all the columns of Θ as follows.

$$\therefore \Theta = \left(\sum_{m=1}^{M} \mathrm{X}_m^T\mathrm{X}_m\right)^{-1}\left(\sum_{m=1}^{M} \mathrm{X}_m^T\mathbf{a}_*^{m}\right)$$

Estimating $\boldsymbol{\sigma}$ Differentiating Equation A.8 with respect to σ and equating the gradient to 0, we have.

$$\Delta_\sigma Q = 0$$

$$\Delta_\sigma \sum_{m=1}^{M} \log p(\mathbf{a}_*^m; \Theta, \sigma) = 0$$

From Equation 4.6, $\mathbf{a}_*^m$ was defined by adding zero mean Gaussian noise to $\text{vec}(\mathbf{a}_*^m)$. Assuming $v = \text{vec}(\mathbf{a}_k^m)$ and $y = \text{vec}(\mathrm{X}_m\Theta)$, we have

$$\Delta_\sigma \sum_{m=1}^{M} \log \frac{1}{|2\pi\sigma^2 I|^{\frac{1}{2}}} e^{-\frac{1}{2}(v-y)^T(\sigma^2 I)^{-1}(v-y)} = 0$$

$$\Delta_\sigma \sum_{m=1}^{M} -TD \log \sigma - \frac{1}{2\sigma^2}||v-y||_2^2 = 0$$

$$\sum_{m=1}^{M} \frac{-TD}{\sigma} + \frac{1}{\sigma^3}||v-y||_2^2 = 0$$

$$\therefore \sigma^2 = \frac{1}{MTD} \sum_{m=1}^{M} ||\text{vec}(\mathbf{a}_k^m) - \text{vec}(\mathrm{X}_m\Theta)||_2^2$$