Latent Confusion Analysis by Normalized Gamma Construction

Issei Sato SATO@R.DL.ITC.U-TOKYO.AC.JP
The University of Tokyo

Hisashi Kashima KASHIMA@I.KYOTO-U.AC.JP
Kyoto University

Hiroshi Nakagawa N3@DL.ITC.U-TOKYO.AC.JP
The University of Tokyo

Abstract

We developed a flexible framework for modeling the annotation and judgment processes of humans, which we called "normalized gamma construction of a confusion matrix." This framework enabled us to model three properties: (1) the abilities of humans, (2) a confusion matrix with labeling, and (3) the difficulty with which items are correctly annotated. We also provided the concept of "latent confusion analysis (LCA)," whose main purpose was to analyze the principal confusions behind human annotations and judgments. It is assumed in LCA that confusion matrices are shared between persons, which we called "latent confusions", in tribute to the "latent topics" of topic modeling. We aim at summarizing the workers' confusion matrices with the small number of latent principal confusion matrices because many personal confusion matrices is difficult to analyze. We used LCA to analyze latent confusions regarding the effects of radioactivity on fish and shellfish following the Fukushima Daiichi nuclear disaster in 2011.

1. Introduction

An important theme in collective intelligence is modeling the annotation and judgment processes of humans. We focus on modeling a confusion matrix with labeling. Extracting a confusion matrix is useful for not just obtaining better (closer to the ground truth) aggregation of labels but also obtaining diagnostic information on human annotation and judgments.

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

Dawid and Skene (1979) proposed a probabilistic generative model for subjective labeling. Their model can estimate individual confusion matrices even when the true label is not available. Each worker in this model has a confusion matrix in which if an item (e.g., an image) has true label u, worker j can assign another label l with probability $\pi_{u,l}^{(j)}$. Smyth et al. (1994) applied the Dawid and Skene (DS) model to the image labeling problem. Snow et al. (2008) applied the DS model to the analysis of opinions in natural language processing. Liu and Wang (2012) applied the DS model to judge the quality of (query, URL) pairs.

Whitehill et al. (2009) proposed the Generative model of Labels, Abilities, and Difficulties (GLAD), which simultaneously estimated the expertise of each worker and the difficulty of each task. It is beneficial to use GLAD, unlike the DS model, in that it models the difficulty with which items are correctly annotated. However, it suffers from a critical issue that when we apply GLAD to a task with multiple labels, the confusion matrix of a worker cannot be constructed (see Sec.3.2 for the details).

Contributions: This paper makes three contributions.

(1) We propose a normalized gamma construction (NGC) of a confusion matrix to model the annotation and judgment process of humans. This framework easily enables us to model a confusion matrix with labeling in a multi-label setting like the DS model and to take into account a task's difficulty like that with GLAD.

(2) We provide a novel concept in data science, *latent confusion analysis (LCA)*, which was developed with the NGC framework and latent Dirichlet enhanced modeling. The main aim of LCA is to extract latent (principal) confusions behind the annotation and judgment processes of humans. LCA summarizes the workers' confusion matrices with the small number of latent principal confusion matrices because many personal confusion matrices is difficult to analyze.

(3) The proposed learning algorithm was based on the

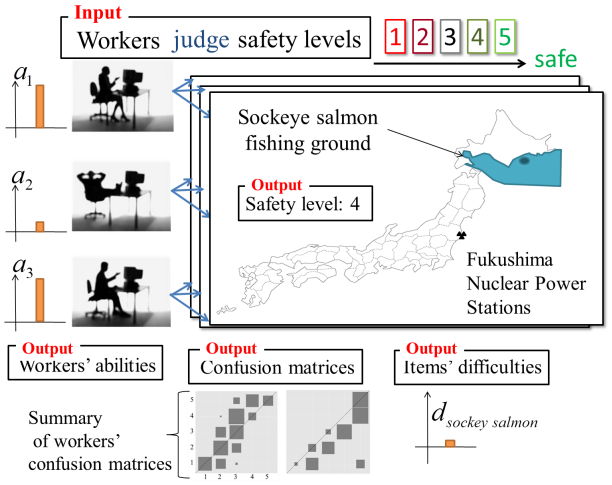

Figure 1. Motivational example and input/output of our model. We created rating data in which the safety level of fish and shellfish was annotated by using Japanese crowdsourcing. We asked people in Japanese crowds to assess the safety levels of items that consisted of fish names and their fishing grounds in Japan. We aimed at gaining insights into what kind of confusions Japanese people had about the effect of radioactivity on fish and shellfish. Unlike Dawid and Skene (1979), we modeled confusion matrices for the workforce, not for each worker, i.e., we shared latent confusion matrices between workers. When the workers' confusions could be described with a combination of a small number of principal confusion matrices, the number of principal confusion matrices was smaller than the number of workers. Therefore, we expected to only have to analyze data in the small principal confusion matrices. Moreover, we modeled the worker's ability, denoted by a, and the difficulty in correctly annotating items, denoted by d. This information helped us understand the properties of rating data. The details are described in the experimental section 6.

variational Bayes inference. Due to the normalization term of NGC, we had to devise a way of optimizing the variational lower bound, which enabled us to obtain closed form solutions in the M-step (see Sec.5). Moreover, we provide point estimations of prior parameters, i.e., we did not need to tune prior parameters for each dataset

Motivation behind LCA: Fukushima Daiichi nuclear disaster

We are often interested in obtaining diagnostic information on the types of confusions people experience. It is more useful in this situation to extract the shared confusions behind people than extract the individual confusion matrix for each person in the existing work. It is ideal in this situation to analyze the confusion matrix of each person by using a combination of latent confusions.

Figure 1 outlines our motivation for latent confusion analysis. On March 11 2011, the Tohoku earthquake and tsunami occurred, followed by a series of equipment failures, nuclear meltdowns, and the release of radioactive materials at the Fukushima I Nuclear Power Plant, which was called the Fukushima Daiichi disaster. This disaster is considered to be the largest nuclear disaster since the Chernobyl disaster of 1986 and was only the second disaster along with Chernobyl to measure Level 7 on the International Nuclear Event Scale.

Unsurprisingly, there was a great deal of concern in Japan about the risk to health and the food chain caused by radioactivity. A huge social issue emerged called "Fuhyo Higai" in Japanese that was related to trustworthiness. Farmers, fishermen, and related businesses face this risk because consumers stopped buying products that might be affected by radioactivity. There is now a growing need to analyze how people are confused about the effect radioactivity has on foods. We are therefore under pressure to analyze latent confusions from a questionary investigation into what effect radioactivity has on foods. The details on the dataset are described in the experimental section.

This paper is organized as follows. We describe existing models in Sec.3. We propose our novel framework in Sec.4. We provide the variational Bayes inference for LCA in Sec.5. We present comparative experimental results in Sec.6.

2. Preliminaries and notations

The **bold** notation of a variable indicates a set of the variables in terms of its subscripts, e.g, $\boldsymbol{z}_j = \{z_{j,i}\}_{i=1}^{n_j}$ and $\boldsymbol{z} = \{\boldsymbol{z}_j\}_{j=1}^{N}$. $\mathbb{E}[x]$ denotes the expectation of x by its distribution. In particular, $\mathbb{E}_q[x]$ denotes the expectation of x by its variational posterior. $\mathrm{KL}[\cdot||\cdot]$ denotes the Kullback-Leibler (KL) divergence. $\mathrm{Multi}(\cdot)$ denotes the multinomial distribution. $\mathrm{Dir}(\cdot)$ denotes the Dirichlet distribution. $\mathrm{Gamma}(\cdot)$ denotes the gamma distribution. The probability function of the gamma distribution is $\mathrm{Gamma}(x; a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}$. The expectation of x and $\log x$ are $\mathbb{E}[x] = a/b$ and $\mathbb{E}[\log x] = \Psi(a) - \log b$. $x \sim P$ expresses that a random variable x is distributed according to the probability distribution P. $\Psi(\cdot)$ is the digamma (psi) function. $\delta(c)$ is the delta function that takes a value of one if condition c is satisfied, and zero otherwise. $\sum_{l(\neq u)}$ means $\sum_{l=1}^{L} \delta(u \neq l)$.

Suppose that we have M items to annotate and L annotation labels. N denotes the number of workers. Each item has a true label from a set of labels $\{1, 2, \cdots, L\}$ where the true label is not available in fact. For example, in the case of customers rating books on a scale from one to five stars, we have M books and L is five.

Since the true labels cannot be observed, we formulate the

true labels as latent variables. $\tau_m = l$ denotes that the true label of item m is $l \in \{1, \cdots, L\}$. $x_{j,i}$ is the i-th item that worker j labels. $y_{j,i} \in \{1, \cdots, L\}$ denotes the label assigned by worker j to item $x_{j,i}$. $\boldsymbol{x}$ is a bag of $x_{j,i}$, $\boldsymbol{y}$ is a bag of $y_{j,i}$, and $\boldsymbol{\tau}$ is a bag of τ_m. n_j is the number of items that worker j annotates.

3. Existing models

Here, we describe the two models most related to our work.

3.1. Dawid and Skene (DS) model

Dawid and Skene (1979) considered the problem of measuring observer error and analyzed a patient record in which the patient was seen by different clinicians and different responses may be obtained to the same questions. They proposed a model that allowed an individual confusion matrix to be estimated even when the true response was not available. We call this model the DS model.

The key idea is to introduce the confusion matrix given by the probability, $\pi^{(j)}_{u,l}$, that worker j will assign label l when u is the true label. That is, worker j assigns label $y_{j,i}$ to item $x_{j,i}$ by

$$y_{j,i} = \begin{cases} u & \text{with probability} \quad \pi^{(j)}_{u,u} \\ l(\neq u) & \text{with probability} \quad \pi^{(j)}_{u,l} \end{cases}. \qquad (1)$$

The probabilities $\pi^{(j)}_{u,l}$ $(u \neq l)$ indicate the individual error rates for worker j, and $\pi^{(j)}_{u,u}$ is the probability that worker j will annotate the true label u. Note that the error rates are conditional probabilities where $\sum_{l=1}^{L} \pi^{(j)}_{u,l} = 1$ for each j and u. $\boldsymbol{\pi}$ is a set of $\pi^{j}_{u,l}$.

The likelihood of workers' annotation data $\boldsymbol{y}$ and true label $\boldsymbol{\tau}$, given $\boldsymbol{\pi} = \{\boldsymbol{\pi}^{(j)}\}_{j=1}^{N}$ is

$$p(\boldsymbol{y}, \boldsymbol{\tau} | \boldsymbol{x}, \boldsymbol{\pi}, \boldsymbol{\mu}) = \prod_{m=1}^{M} \mu_{\tau_m} \prod_{j=1}^{N} \prod_{i=1}^{n_j} \pi^{(j)}_{\tau_{x_{j,i}}, y_{j,i}}. \qquad (2)$$

where $\boldsymbol{\mu}$ is the true label prior, i.e, $\mu_l = p(\tau_m = l)$, and we denote $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_L)$. Dawid and Skene (1979) used the Expectation-Maximization (EM) algorithm to estimate $p(\tau_m | \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{p}, \boldsymbol{\mu})$, $\pi^{(j)}_{u,l}$, and μ_l.

3.2. Generative Model of Labels, Abilities, and Difficulties (GLAD)

Whitehill et al. (2009) formulated a probabilistic model of the binary-labeling process, i.e., $L = 2$, by modeling the true labels, workers' abilities, and the difficulty with which items were correctly annotated, called the Generative model of Labels, Abilities, and Difficulties (GLAD).

The ability (expertise) of each worker j is modeled by the parameter $a_j \in (-\infty, \infty)$. $a_j = \infty(-\infty)$ means the worker always labels items correctly (incorrectly). $a_j = 0$ means that the worker has no information about the true label. The difficulty of annotating item m to be annotated correctly is modeled by d_m, which is positive and $d_m = \infty$ means the item is very ambiguous and hence even the highest skilled worker has only a 50% chance of labeling it correctly. $1/d_m = \infty$ means the item is so easy to annotate that most of workers always label it correctly.

Label $y_{j,i} = l$ that worker j assigns to i-th item $x_{j,i} = m$ given true label τ_m is generated according to

$$p(y_{j,i} = l | a_j, d_m, \tau_m) = \sigma(a_j/d_m)^{\delta(\tau_m = l)} \left(1 - \sigma(a_j/d_m)\right)^{\delta(\tau_m \neq l)}, \qquad (3)$$

where $\sigma(\cdot)$ is the sigmoid function. Equation (3) means that if true label $\tau_m = l$, the probability of $y_{j,i} = l$ is $\sigma(a_j/d_m)$; otherwise, $1 - \sigma(a_j/d_m)$. It is ideal for us to have closed-form solutions in the M-step. However, we have to numerically solve an optimization problem, e.g., by gradient ascent, to estimate worker j's ability a_j and task m's difficulty d_m for each M-step in the EM algorithm, which requires tuning a step-size parameter.

Whitehill et al. formulated a multiple-label variant of GLAD (mGLAD) in their paper's supplementary material. It is assumed with mGLAD that the probability of worker j assigning label l given true label τ_m is

$$p(y_{j,i} = l | a_j, d_m, \tau_m) = \sigma(a_j/d_m)^{\delta(\tau_m = l)} \left(\frac{1 - \sigma(a_j/d_m)}{L - 1}\right)^{\delta(\tau_m \neq l)}. \qquad (4)$$

Equation (4) means that if true label $\tau_m = u$, the probability of $y_{j,i} = u$ is $\sigma(a_j/d_m)$ and that of other labels is $(1 - \sigma(a_j/d_m))/(L-1)$, respectively. The workers' ability parameters a_j $(j = 1, \cdots, N)$ are shared in all items. The problem is that the workers' labeling confusion cannot be modeled because it is modeled as a uniform, i.e., $1/(L-1)$.

3.3. Other Related Work

Various studies have investigated crowdsourcing. The following studies differ from our work in that they are not aimed at analyzing the confusion matrices, in particular, latent confusions and the most cases are binary labeling.

Classification: Raykar et al. (2010) studied a binary classifier via estimating the annotator accuracy and the actual true label. Yan et al. modeled annotators' expertise as a function of the item's feature vector (Yan et al., 2010b;a; 2011). Welinder et al. (2010) modeled a binary annotation process by considering the low-dimensional feature vector of each image in an image labeling task. Wauthier et al.

(2011) proposed a Bayesian model to account for annotator bias. Liu et al. (2012) connected the aggregation of binary labels in crowdsourcing with belief propagation and a mean field algorithm. These studies seemed to be along the lines of the DS and GLAD models but their setting was binary labeling and did not deal with the confusion matrix. Zhou et al. (2012) proposed a minimax entropy principle to improve the quality of noisy labels.

Clustering and Ranking: Gomes et al. (2011) presented "crowdclustering," which clusters items by using a worker's label (same or different) on a pair of items. Yi et al. (2012) combined a metric learning and the manual annotations obtained via crowdsourcing for clustering. Raykar and Yu (2011) proposed a score to rank the annotators.

Confusion Matrix Modeling: Venanzi et al. (2014) proposed a community-based Bayesian aggregation model, which assume that each worker belongs to a certain community and the worker ' s confusion matrix is similar to the community ' s confusion matrix. (Venanzi et al., 2014) is the most similar work to our work[1]. The difference between their model and ours is as follows. (1) Each worker belongs to one community. (2) They model one confusion matrix for each worker. (3) They do not model the abilities of workers and the difficulty of items simultaneously.

4. Proposed framework

We describe the proposed framework in this section. First, we present the normalized gamma construction (NGC) of a confusion matrix. Then, we propose latent confusion analysis (LCA).

4.1. Normalized Gamma Construction of Confusion Matrix

The confusion matrix of each worker is constructed with probabilistic vectors in the DS model. It is common to assume that a probability vector is distributed according to the Dirichlet distribution. That is, the process to generate the confusion matrix in Eq.(1) for worker j in a Bayesian manner is

$$\boldsymbol{\pi}_u^{(j)} \sim \mathrm{Dir}(\gamma_1, \cdots, \gamma_L),\ (u = 1, \cdots, L), \tag{5}$$

where γ_l $(l = 1, \cdots, L)$ is a parameter of the Dirichlet distribution. However, in this formulation, we cannot model the difficulty with which items are correctly annotated. Therefore, we need a novel process of generating probabilistic vectors to introduce the difficulty with which items are correctly annotated, as with GLAD.

[1](Venanzi et al., 2014) was published after this paper was submitted.

Here, we consider the following relationship between the Dirichlet distribution and the gamma distribution (p.594 of (Devroye, 1986)).

If g_ℓ $(\ell = 1, \cdots, L)$ is independently distributed according to $\mathrm{Gamma}(\gamma_i, 1)$ respectively, i.e.,

$$g_\ell \sim \mathrm{Gamma}(\gamma_\ell, 1), \tag{6}$$

then the vector $(g_1/s, \cdots, g_\ell/s)$, where $s = \sum_{\ell=1}^{L} g_i$, follows the Dirichlet distribution with parameters $\gamma_1, \cdots, \gamma_L$, i.e.,

$$s = \sum_{\ell=1}^{L} g_i \sim \mathrm{Gamma}(\sum_{\ell=1}^{L} \gamma_\ell, 1) \tag{7}$$

$$(\pi_1, \cdots, \pi_L) = (\frac{g_1}{s}, \cdots, \frac{g_L}{s}) \sim \mathrm{Dir}(\gamma_1, \cdots, \gamma_L). \tag{8}$$

This reformulation inspired us to use the construction of each worker's confusion matrix $\boldsymbol{\pi}^{(j)}$ by using random variables distributed according to the gamma distribution, which presents a flexible framework for modeling the annotation process of workers in the next session and an efficient inference algorithm based on the VB inference.

The DS model's generation process in Eq.(1), and the generation process of $\boldsymbol{\pi}_u^{(j)}$ with the Dirichlet distribution in Eq.(5), can be reformulated as follows. Let $c_{j,u,l}$ be a confusion variable for worker j to assign label l to an item that has true label u and

$$c_{j,u,l} \sim \mathrm{Gamma}(\gamma_c, 1)\ (u, l = 1, \cdots, L). \tag{9}$$

Given true label u, we have

$$y_{j,i} = \begin{cases} u & \text{with probability} \quad \pi_{u,u}^{(j)} = \dfrac{c_{j,u,u}}{\sum_{v=1}^{L} c_{j,u,v}} \\ l(\neq u) & \text{with probability} \quad \pi_{u,l}^{(j)} = \dfrac{c_{j,u,l}}{\sum_{v=1}^{L} c_{j,u,v}} \end{cases} \tag{10}$$

This idea enables us to easily introduce the ability of humans, a confusion matrix and the difficulty with which items are correctly annotated into modeling human annotation and judgment processes, and to model the concept of latent confusion analysis (LCA), which is described in the next section.

4.2. Latent Confusion Analysis

It is assumed with the DS model that a confusion matrix is formulated for each worker. In this section, we consider that confusion matrices are shared between workers, which we call "latent confusions" in tribute to the "latent topics" of latent Dirichlet allocation (Blei et al., 2003). Figure 2 outlines the graphical model of the proposed model.

Worker j has a latent variable for the i-th item to be annotated, denoted by $z_{j,i}$, and $z_{j,i} = k$ indicates that worker j is affected by the k-th latent confusion matrix when annotating the i-th item. Let K be the number of latent confusions, which are given by, for $k = 1, \cdots, K$,

$$c_{k,u,l} \sim \text{Gamma}(\gamma_c, 1)\ (u, l = 1, \cdots, L). \tag{11}$$

Moreover, we introduce the ability of each worker j, denoted by a_j, and the difficulty with which item m can be correctly annotated, denoted by d_m as

$$a_j \sim \text{Gamma}(\gamma_a, 1)\ (j = 1, \cdots, N), \tag{12}$$
$$d_m \sim \text{Gamma}(\gamma_d, 1)\ (m = 1, \cdots, M). \tag{13}$$

Therefore, worker j assigns label $y_{j,i}$ to item $x_{j,i} = m$ when $z_{j,i} = k$ and the true label is $u \in \{1, \cdots, L\}$ as

$$y_{j,i} = \begin{cases} u \text{ with probability} & \pi^{(j,m,k)}_{u,u} \propto a_j c_{k,u,u} \\ l(\neq u) \text{ with probability} & \pi^{(j,m,k)}_{u,l} \propto d_m c_{k,u,l} \end{cases} \tag{14}$$

A large a_j and $c_{k,u,u}$ mean that worker j tends to correctly label items when the true label is u. A large $c_{k,u,l}$ means that a worker tends to label l when the true label is u. A large d_m means that item m is so difficult to correctly annotate that even the most expert worker will usually label it incorrectly according to its confusion matrix $\boldsymbol{c}_k$.

The remaining problem is how to model latent variables $\boldsymbol{z}$. We use latent Dirichlet modeling for latent variables. That is, for each worker j, $\boldsymbol{\theta}_j \sim \text{Dir}(\boldsymbol{\alpha})$, where $\boldsymbol{\alpha}$ is the K-dimensional parameter vector of the Dirichlet distribution and for the i-th item to be annotated, $z_{j,i} \sim \text{Multi}(\boldsymbol{\theta}_j)$.

Moreover, we model the probability distribution over item $x_{j,i}$ by $\phi_{k,m}$ which indicates the probability that the item that a worker annotates is m when the worker is affected by the k-th confusion matrix, i.e. $\sum_{m=1}^{M} \phi_{k,m} = 1$, as follows. When $z_{j,i} = k$, $x_{j,i} \sim \text{Multi}(\boldsymbol{\phi}_k)$. It is easy to understand this generation process when $x_{j,i}$ is regarded as a word in latent Dirichlet allocation.

The reason we modeled the process for the generation of items is that we wanted to analyze the relationship between items and latent confusion. It is useful to gain insights into what types of items are affected by the k-th latent confusion in the annotation process. When $\phi_{k,m}$ takes a large value, we find that the annotation of item m is greatly affected by the k-th latent confusion. We assume that $\boldsymbol{\phi}_k \sim \text{Dir}(\beta)$, $(k = 1, \cdots, K)$, where $\text{Dir}(\beta)$ is a symmetric Dirichlet prior with scalar parameter β.

5. Variational Bayes Inference for LCA

We provide the variational Bayes (VB) inference in the proposed model. Due to the normalization term of NGC, we have to devise a way of optimizing the variational lower bound.

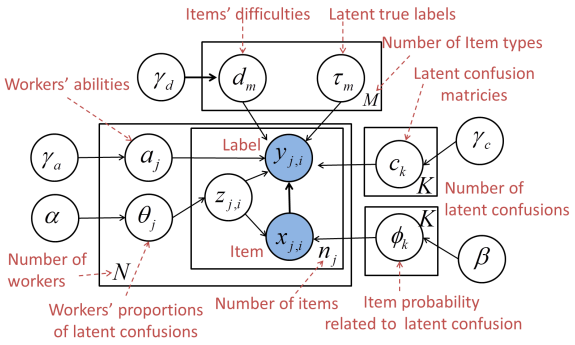

Figure 2. Graphical model of LCA

We present the key idea to derive the VB inference for the proposed model. This is promising to enable this NGC framework to be applied to many applications.

For simplicity, we consider the simple form $\frac{g_\ell}{\sum_\ell g_\ell}$ and g_ℓ is distributed by the gamma distribution. We typically need the expectation calculation in the VB inference, i.e., $\mathbb{E}[\log \frac{g_\ell}{\sum_\ell g_\ell}] = \mathbb{E}[\log g_\ell] - \mathbb{E}[\log \sum_\ell g_\ell]$. The problem is that we cannot calculate the expectation of a log function of the normalization term $\mathbb{E}[\log \sum_\ell g_\ell]$.

Here, we will return to the definition of the gamma distribution. The gamma distribution $p(\xi; a, b) = \frac{b^a}{\Gamma(a)} \xi^{a-1} e^{-b\xi}$ indicates that $1 = \int p(\xi; a, b) d\xi = \int \frac{b^a}{\Gamma(a)} \xi^{a-1} e^{-b\xi} d\xi, b^{-a} = \int \frac{1}{\Gamma(a)} x^{a-1} e^{-b\xi} d\xi$. When we set $b = \sum_\ell g_\ell$ and $a = 1$, we have $\frac{1}{\sum_\ell g_\ell} = \int e^{-(\sum_\ell g_\ell)\xi} d\xi$. Therefore, $\log \frac{1}{\sum_\ell g_\ell} = \log \int e^{-(\sum_\ell g_\ell)\xi} d\xi$ By introducing probability distribution $q(\xi)$ and Jensen's inequality, we have

$$\log \frac{1}{\sum_\ell g_\ell} = \log \int q(\xi) \frac{e^{-(\sum_\ell g_\ell)\xi}}{q(\xi)} d\xi \geq \int q(\xi) \log \frac{e^{-(\sum_\ell g_\ell)\xi}}{q(\xi)} d\xi. \tag{15}$$

The expectation of this lower bound has an analytic solution, and thus, this lower-bound and the estimation of $q(\xi)$ enable us to obtain closed form solutions for the VB inference of NGC and LCA.

We estimate variational posteriors $q(\boldsymbol{\tau})$, $q(\boldsymbol{a})$, $q(\boldsymbol{c})$, $q(\boldsymbol{d})$, $q(\boldsymbol{z})$, $q(\boldsymbol{\theta})$, $q(\boldsymbol{\phi})$ and $q(\boldsymbol{\xi})$. and use the point estimation for γ_a, γ_c, γ_d, $\boldsymbol{\mu}$, $\boldsymbol{\alpha}$ and β because we do not want to tune the hyper-parameters for each task. The estimate of $\boldsymbol{\mu}$ is the same as that in the Dawid and Skene model (Dawid & Skene, 1979).

The details are described in the supplementary material.

Table 1. Basic information on dataset:
N denotes the number of workers. M denotes the number of item types. L denotes the number of label types. $\mathbb{E}[n_{\text{worker}}] = \sum_{j=1}^{N} n_j/N$ denotes the average number of items that a worker annotated. N_{item} denotes the number of workers who annotated an item.

Dataset	N	M	L	$\mathbb{E}[n_{\text{worker}}]$	N_{item}
Safety-level	62	97	5	78.2	50
Preposition	47	100	5	63.8	30
Bluebird	39	108	2	108	39

5.1. Computational Complexity

Let T be the total number of labeled items. The computational costs per iteration in DS model and LCA are $\mathcal{O}(NL^2 + TL + ML)$ and $\mathcal{O}(TKL^2 + ML + MK)$, respectively. Seemingly, LCA is not that scalable because it can reveal a much more informative latent structure than existing models. However, the scalability of LCA is the same as that of LDA, and our learning algorithm is deterministic. Therefore, we can easily apply recent advances in scaling-up LDA into LCA such as those in the literature (Hoffman et al., 2010; Zhai et al., 2012).

6. Experiments

We empirically analyzed the proposed model in this section.

Since our problem setting was unsupervised, i.e., the true labels and confusion matrices were not available, it was difficult to evaluate the models. Therefore, we use datasets in which the correct answers (labels or scores) were known. Here, we call a "gold label" a correct label that is actually known in the datasets. We only use a gold label to evaluate an estimated label that has a maximum probability of $q(\tau_m)$ for each model, i.e., $\tau_m^* = \text{argmax}_{\tau_m}\, q(\tau_m)$.

MV indicates majority voting. DS indicates the Dawid and Skene model, and GLAD/mGLAD (multi-label variant of GLAD described in Sec. 3). LCA is our model described in Sec. 4.2.

6.1. Datasets and Evaluation Metrics

We applied the models to three datasets: (1) safety-level data, (2) preposition data, and (3) bluebird data. We created datasets (1) and (2) by using crowdsourcing and published the datasets[2]. Table 1 summarizes the basic information on the datasets.

[2]http://www.r.dl.itc.u-tokyo.ac.jp/ sato/icml2014/

Safety-Level Data: Analyzing safety-level data is the main purpose of this study. We analyzed public confusion regarding the effects of radioactivity on food products following the Fukushima Daiichi nuclear disaster. We used Japanese crowdsourcing to prepare this dataset. We asked crowd workers to judge the safety level of an item by using two pieces of information on each item: "the name of the fish or shellfish" and "where its fishing grounds are in Japan," as outlined in Fig.1. The number of labels for annotation was $L = 5$, in which "safety level 1" meant "It's dangerous. I will not eat this food" and "safety level 5" meant "It's safe. I will eat this food." We used items described in a safety manual on the effects of radioactivity on food products published in 2012[3]. The number of items was $M = 97$. These items in the safety manual had a safety level from 1 (dangerous) to 100 (safe), which was calculated by using radioactivity measurement and expert knowledge. We used this information to evaluate models as a gold safety level. Each item was annotated by 50 workers. The evaluation metric was the correlation coefficient between the gold safety level and the estimated safety level with the maximum probability of $q(\tau_m)$.

We used other datasets to compare our model with the other models in several settings.

Preposition Data: The use of prepositions in English is often a headache for non-native English speakers. We analyzed public confusion in the use of prepositions. We collected 100 sentences as fill-in-the-blank questions from the Special English of Voice of America (VOA) [4], where $M = 100$. We asked crowd workers to select a preposition by choosing from the labels "on," "at," "in," "for," and "to," which were the prepositions that cause confusion for Japanese people, i.e., $L = 5$. The number of workers was $N = 47$ and each item was annotated by 30 workers. The evaluation metric was the accuracy measured by using a gold label and an estimated answer that had a maximum probability of $q(\tau_m)$, i.e., accuracy=the number of correct answers / M.

Bluebird Data: We used a dataset called "bluebird," published by Welinder et al. (2010). This dataset included $M = 108$ items and $N = 39$ workers on a fully connected bipartite assignment graph, where the workers were asked whether the presented images contained the Indigo Bunting or Blue Grosbeak, i.e., $L = 2$. Each item was annotated by 39 workers. The evaluation metric was accuracy.

[3]"Complete manual on the effects of radioactivity on food products" (in Japanese) ISBN-10: 4796696857

[4]http://learningenglish.voanews.com/

Table 2. Empirical results.

Larger values indicate better performance. If there is an equality of votes in majority voting (MV), we select a label at random. We tried five random seeds in MV. Note that we could only determine one label in other models even if there was an equality of votes because we used the maximum $q(\tau_m)$ for labeling items. K^+ denotes the effective number of latent confusion which is estimated by the implicit sparsity of the VB inference. Note that we set $K = N$ in these experiments.

Dataset	Evaluation Metric	MV(five random seeds)	LCA (K^+/K)	DS	GLAD/mGLAD
Safety-level	Correlation coefficient	$0.525, 0.525, 0.508, 0.510, 0.528$	$\mathbf{0.571\ (20/62)}$	0.505	0.472
Preposition	Accuracy	$0.709, 0.719, 0.700, 0.710, 0.710$	$\mathbf{0.770\ (7/47)}$	0.739	0.750
Bluebird	Accuracy	0.759 (No equality of votes)	$\mathbf{0.898\ (6/39)}$	$\mathbf{0.898}$	0.722

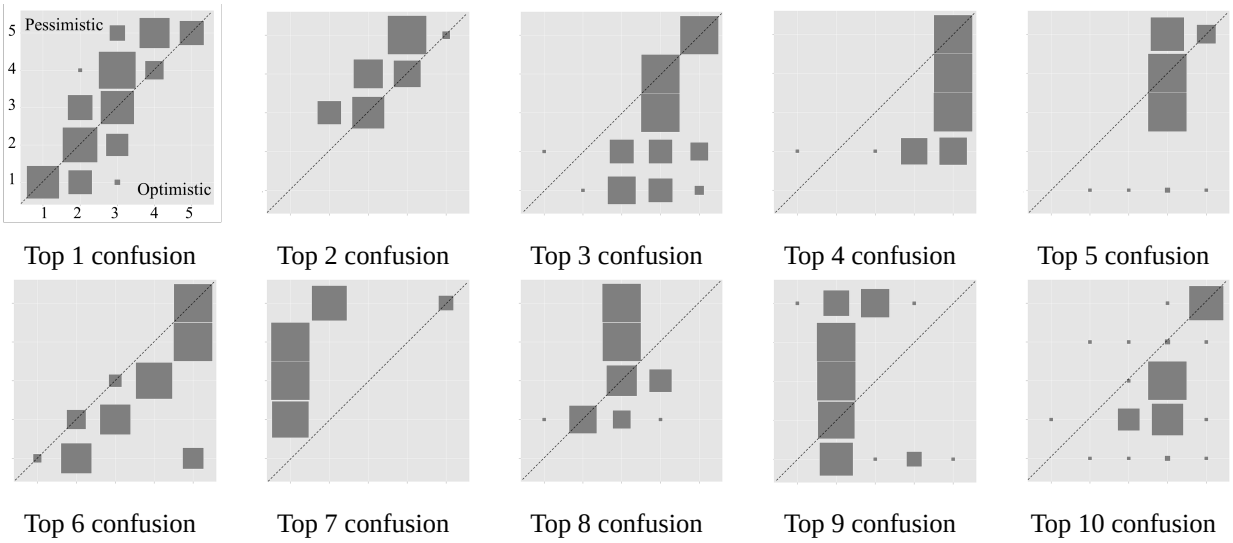

Figure 3. Top ten frequent latent confusions behind safety-level data. The size of the gray squares indicates the size of the values of the confusion probabilities. For simple visualization, a row with no gray squares means that the confusion probability is uniform (see Sec.6.5).

6.2. Initialization

The results obtained from the DS and GLAD models only depended on the initialization of $q(\tau_m)$. We initialized $q(\tau_m)$ with an empirical distribution by using worker voting as Dawid and Skene (1979) did in their study, who observed that this initialization was more effective than random initializations. We also found in pilot experiments that voting initialization was more effective than random initializations in the DS and GLAD models. When we used voting initialization for $q(\tau_m)$ and not random initialization, we only had to do the experiment once. The results from LCA depended on the initialization of $q(z_{j,i})$ (or $q(c_{k,u,v})$) as well as $q(\tau_j)$ and the number of latent confusions K. Therefore, it was ideal for LCA that we did not use randomization for the initialization. We devised the following strategy to initialize LCA. Note that we actually considered initializing $q(c_{k,u,v})$ instead of $q(z_{j,i})$.

(1) We set $K = N$ and $q(z_{j,i} = j) = 1$ (0 otherwise), which means that each worker had a personal confusion matrix as well as the DS model. We initialized $q(\tau_m)$ with an empirical distribution by voting like that in the DS model.

(2) We ran the VB inference with $q(z_{j,i} = j) = 1$ being fixed, which meant that we did not use the latent Dirichlet enhanced modeling. The results in this step only depended on the initialization of $q(\tau_m)$ as with the DS model.

(3) We reset $q(z_{j,i} = j) = 1/K$ and initialized $q(\boldsymbol{\theta}_j)$ and $q(\boldsymbol{\phi}_k)$ with their prior distributions.

After these three steps, we ran the VB inference for LCA. This initialization scheme only depended on $q(\tau_m)$ as with the DS and GLAD models.

When we set $K < N$, we select the workers' personal confusions in descending order of their expected ability of $\mathbb{E}_q[a_j]$, which is pre-estimated in initialization step 2.

We initialized $\gamma_a = \gamma_c = \gamma_d = 1$.

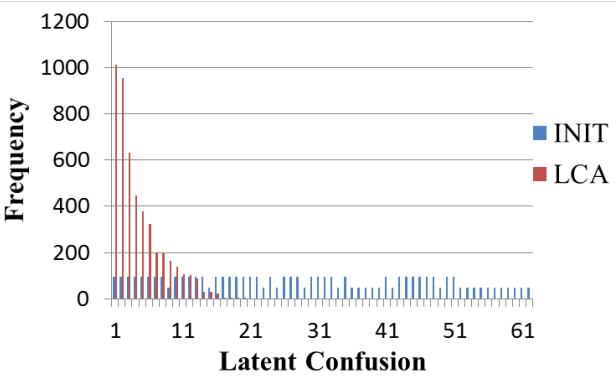

Figure 4. Expected frequency of latent confusions in safety-level data.

6.3. Empirical results for evaluation metric

Table 2 summarizes the experimental results. Our model, LCA, outperformed or was competitive with the other models in terms of each evaluation metric for each dataset. The results revealed that using a family of confusion matrices helped to recover the ground truth. We think one of the reasons is that our approach is intuitively a kind of multi-task learning. We could complement the latent judgment tendencies of workers who assessed the small number of items by sharing confusion matrices among users. Moreover, when a dataset has a genre or category, e.g., that in rating movies, a worker has a biased knowledge, which can be modeled by a family of confusion matrices. In this case, the confusion matrix of a worker cannot be represented as one confusion matrix but a combination of various latent confusion matrices. For example, there can be bias in a worker's knowledge due to where they lives in savefy-level data.

6.4. Effective number of latent confusions

It was well known that the VB inference induced implicit sparcity, which is called a zero-forcing effect (Minka, 2005). This effect enables us to estimate the effective number of latent confusions. This property is seen in the update equation of $q(z_{j,i})$.

Let K^+ be the effective number of latent confusions, i.e., the number of principal confusions. We calculated K^+ by using the number of latent confusions whose expected frequency $\mathbb{E}[n_k] = \sum_{j,i} q(z_{j,i} = k)$ was greater than 0.5 in Table 2. The performance of LCA and the DS model was the same with the bluebird data; however, it was found that LCA used a smaller number of confusion matrices than the DS model did.

Figure 4 plots the frequency of latent confusions before and after the VB inference of LCA in the safety-level data. "INIT" indicates the frequency in the initialization step, i.e., each frequency indicates the number of items that each worker annotated (see Sec.6.2). "LCA" means that the frequency was given after the VB inference, where the frequency was the expected frequency, i.e., $\mathbb{E}[n_k]$. The latent confusions are sorted in descending order in terms of the frequency of "LCA."

6.5. Visualizing Top 10 latent confusions

We analyzed the top 10 latent confusions by frequency extracted by using LCA in Fig.3, which revealed "pessimistic" and "optimistic" confusions in the safety-level data.

We normalized the k-th confusion matrix, $\boldsymbol{c}_k$, to make each row a confusion probability, i.e., $\tilde{c}_{k,u,\ell} = \mathbb{E}[c_{k,u,\ell}] / \sum_{\ell=1}^{L} \mathbb{E}[c_{k,u,\ell}]$. The element of a confusion matrix in the u-th row and ℓ-th column, $\tilde{c}_{k,u,\ell}$, expresses the probability that if an item has true label, u, label ℓ will be annotated. The size of the gray squares indicates the size of the values of their elements. We deducted $\min_\ell \tilde{c}_{k,u,\ell}$ from each row to enable simple visualization. Therefore, a row with no gray squares means that the confusion probability is uniform.

If there are gray squares below the dashed line, confusion indicates the safety level has been overestimated, which means "optimistic." The top 1, 2, 7, 8, and 9 confusions seem to be pessimistic, and the others seem to be optimistic. The top 1 confusion was less unbiased than the other confusions. The top 4 confusion was much more optimistic because safety levels 2, 3, and 4 could be confused as safety level 5. The top 7 confusion was, in comparison, much more pessimistic. The frequency of each confusion is plotted in Fig.4.

7. Conclusion

We proposed modeling the annotation and judgment processes of humans by using the normalized gamma construction (NGC) of a confusion matrix. The NGC framework flexibly enabled various properties of data to be modeled and it provided an efficient learning algorithm based on the variational Bayes inference and the fixed point iteration algorithm to estimate prior parameters. Therefore, it had a wide range of applications besides those that we described in this paper. We also provided the concept of "latent confusion analysis (LCA)," which was used to analyze the principal confusions behind human annotations and judgments. We modeled LCA by using NGC and latent Dirichlet modeling. LCA seemed to be increasingly more important because there is an information overload in real life and people are too taken in by it.

References

Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3: 993–1022, 2003.

Dawid, A P and Skene, A M. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied Statistics*, 28(1):20–28, 1979.

Devroye, Luc. *Non-Uniform Random Variate Generation*. Springer-Verlag, 1986.

Gomes, Ryan G., Welinder, Peter, Krause, Andreas, and Perona, Pietro. Crowdclustering. In *Advances in Neural Information Processing Systems 24*, pp. 558–566. 2011.

Hoffman, Matthew D., Blei, David M., and Bach, Francis R. Online learning for latent dirichlet allocation. In *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010*, pp. 856–864, 2010.

Liu, Chao and Wang, Yi-Min. Truelabel + confusions: A spectrum of probabilistic models in analyzing multiple ratings. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

Liu, Qiang, Peng, Jian, and Ihler, Alex. Variational inference for crowdsourcing. In *Advances in Neural Information Processing Systems 25*, pp. 701–709. 2012.

Minka, Thomas. Divergence measures and message passing. Technical report, Microsoft Research, 2005.

Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L., and Moy, L. Learning from crowds. *Journal of Machine Learning Research*, 11:1297–1322, 2010.

Raykar, Vikas C. and Yu, Shipeng. Ranking annotators for crowdsourced labeling tasks. In *Advances in Neural Information Processing Systems 24*, pp. 1809–1817. 2011.

Smyth, P., Fayyad, U., Burl, M., Perona, P., and Baldi, P. Inferring ground truth from subjective labelling of venus images. In *Advances in Neural Information Processing Systems 7*, pp. 1085–1092, 1994.

Snow, Rion, O'Connor, Brendan, Jurafsky, Daniel, and Ng, Andrew Y. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, pp. 254–263, 2008.

Venanzi, Matteo, Guiver, John, Kazai, Gabriella, Kohli, Pushmeet, and Shokouhi, Milad. Community-based bayesian aggregation models for crowdsourcing. In *Proceedings of the 23rd International World Wide Web Conference*, pp. 155–164, 2014.

Wauthier, Fabian L. and Jordan, Michael I. Bayesian bias mitigation for crowdsourcing. In Shawe-Taylor, J., Zemel, R.S., Bartlett, P., Pereira, F.C.N., and Weinberger, K.Q. (eds.), *Advances in Neural Information Processing Systems 24*, pp. 1800–1808. 2011.

Welinder, Peter, Branson, Steve, Belongie, Serge, and Perona, Pietro. The multidimensional wisdom of crowds. In *Advances in Neural Information Processing Systems 23*, pp. 2424–2432, 2010.

Whitehill, Jacob, Ruvolo, Paul, fan Wu, Ting, Bergsma, Jacob, and Movellan, Javier. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems 22*, pp. 2035–2043, 2009.

Yan, Yan, Rosales, Romer, Fung, Glenn, and Dy, Jennifer. Modeling multiple annotator expertise in the semi-supervised learning scenario. In *Proc. of the Proceedings of the 26th Conference Annual Conference on Uncertainty in Artificial Intelligence*, pp. 674–682, 2010a.

Yan, Yan, Rosales, Romer, Fung, Glenn, Schmidt, Mark W., Valadez, Gerardo Hermosillo, Bogoni, Luca, Moy, Linda, and Dy, Jennifer G. Modeling annotator expertise: Learning when everybody knows a bit of something. *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, pp. 932–939, 2010b.

Yan, Yan, Rosales, Romer, Fung, Glenn, and Dy, Jennifer. Active learning from crowds. In *Proc. of the 28th International Conference on Machine Learning*, pp. 1161–1168, 2011.

Yi, Jinfeng, Jin, Rong, Jain, Anil, Jain, Shaili, and Yang, Tianbao. Semi-crowdsourced clustering: Generalizing crowd labeling by robust distance metric learning. In *Advances in Neural Information Processing Systems 25*, pp. 1781–1789. 2012.

Zhai, Ke, Boyd-Graber, Jordan L., 0001, Nima Asadi, and Alkhouja, Mohamad L. Mr. lda: a flexible large scale topic modeling package using variational inference in mapreduce. In *ACM International Conference on World Wide Web*, pp. 879–888, 2012.

Zhou, Dengyong, Platt, John, Basu, Sumit, and Mao, Yi. Learning from the wisdom of crowds by minimax entropy. In *Advances in Neural Information Processing Systems 25*, pp. 2204–2212. 2012.