A Correlated Worker Model for Grouped, Imbalanced and Multitask Data

An T. Nguyen
Department of Computer Science
University of Texas at Austin
atn@cs.utexas.edu

Byron C. Wallace
School of Information
University of Texas at Austin
byron.wallace@utexas.edu

Matthew Lease
School of Information
University of Texas at Austin
ml@utexas.edu

Abstract

We consider the important crowdsourcing problem of estimating worker *confusion matrices*, or sensitivities and specificities for binary classification tasks. In addition to providing diagnostic insights into worker performance, such estimates enable robust online task routing for classification tasks exhibiting imbalance and asymmetric costs. However, labeled data is often expensive and hence estimates must be made without much of it. This poses a challenge to existing methods. In this paper, we propose a novel model that captures the correlations between entries in confusion matrices. We applied this model in two practical scenarios: (1) an imbalanced classification task in which workers are known to belong to groups and (2) a *multitask* scenario in which labels for the same workers are available in more than one labeling task. We derive an efficient variational inference approach that scales to large datasets. Experiments on two real world citizen science datasets (biomedical citation screening and galaxy morphological classification) demonstrate consistent improvement over competitive baselines. We have made our source code available.

1 INTRODUCTION

Crowdsourcing is a popular approach to collecting annotations at comparatively low cost. However, crowdsourcing annotation work necessitates taking care to evaluate worker annotation quality, as this will likely be lower than that of a domain expert. The standard means of addressing this is to collect multiple labels for each item and then use an aggregation method to derive a consensus label. The simplest such method is majority voting, which selects the majority label for each item. More complex methods exist; see (Sheshadri and Lease, 2013) for a review.

Many crowd consensus methods posit some model of worker qualities, e.g., a worker's overall accuracy. However, the problem of modeling workers has typically been considered a secondary issue, with the primary concern being label aggregation. Here we argue that the problems of aggregating labels and of modeling workers are distinct (though related). A good method for aggregating labels will not necessarily provide reliable estimates of worker qualities. For example, majority voting assumes that workers are equally good; this assumption is almost certainly usually wrong, but nonetheless can yield high quality aggregated labels when workers do not make correlated errors. Consider, e.g., a scenario in which each item has been labeled by three workers, two of whom are always correct and one of whom is always wrong. Here, majority voting would provide perfect label aggregation, but plainly the workers are not equally good.

The simplest way to model worker skill is with the univariate metric of overall accuracy. This may be appropriate when classes are balanced and/or when false negatives and false positives are equally expensive. However, in most real-world tasks, we would prefer a more granular model of accuracies. The standard way is to model the worker confusion matrices, whose entries are the probabilities of a worker providing each possible label j, conditioned on the true label i: $A_{ij} = Pr(\text{Response} = j|\text{TrueLabel} = i)$. This class conditional approach posits two parameters for each worker, affording the flexibility to accurately capture worker performance. The caveat is that more data is needed to estimate more parameters.

For example, when the majority of items belong to the negative class, very few positive items will be available to reliably estimate the probabilities of responses given that the true label is positive, i.e. A_{01} and A_{11}. Explicitly modeling correlations between sensitivity and specificity is one potential means of improving estimates in this case: Workers who perform well on negative items are likely to also correctly classify positive ones. Another scenario in which this general approach may help is when data from multiple labeling tasks is available, for example if many workers

who performed a text classification task come back to work on a new image classification task. In such scenarios, we might expect a worker who does well on one task to also be likely to do so on the other.

Specific contributions of this work are as follows.

- Taking inspiration from previous work in modeling medical diagnostic test (Dahabreh et al., 2012; Reitsma et al., 2005), we propose modeling the correlation between worker sensitivity and specificity. A natural property of our model is the assumption that workers belong to one or more groups, each with its own mean sensitivity and specificity. Intuitively, these means allow us 'back-off' to group-level quality estimates when data from a specific worker is sparse.
- We extend our idea to model the correlations of the workers across multiple labeling tasks. This allows us to *transfer knowledge of worker quality from a task with more data to one with less*.
- We introduce an efficient variational method to scale parameter estimation to handle large data.

We are unaware of any previous work on using correlations to improve estimates of worker confusion matrices nor a generative model of worker performance in multiple tasks.

2 RELATED WORK

Dawid and Skene (1979) presented the classic crowd consensus model in which each item corresponds to a hidden 'true label' variable and each worker is modeled by a confusion matrix of class-conditional label probabilities. Raykar et al. (2010); Kim and Ghahramani (2012); Liu and Wang (2012) all presented Bayesian extensions, placing priors on the worker confusion matrices. Unique amongst this prior work, Liu and Wang (2012) emphasized getting good estimates of the confusion matrices to gain diagnostic insights into worker performance, while the other two focus on recovering the true labels and building a good classifier. Recently, Lakkaraju et al. (2015) further extended the model by clustering workers and items based on their features. Although very effective compared to Liu and Wang (2012)'s Hybrid Confusion approach, demographic features such as worker age, education or job are not always available, e.g., on Amazon Mechanical Turk.

The idea of detecting latent groups and communities of similar workers has been studied previously (Simpson et al., 2013), and incorporated into a generative model to improve the aggregated labels (Venanzi et al., 2014). By contrast, here we consider exploiting *known* worker groups, when such information is available, to improve our confusion matrix estimates.

While modeling individual confusion matrices is the most common approach to annotator modeling, recent work has also explored other strategies. Kajino et al. (2012)'s multitask formulation views each worker as a learning task. Bi et al. (2014) models each worker as a classifier, whose parameters deviate from the true parameters. While these methods have been shown to outperform the 'Two Coin' model (Raykar et al., 2010) for the task of label aggregation, they unfortunately do not provide direct estimates of worker sensitivity and specificity.

In terms of inference and learning algorithms, Dawid and Skene (1979) and Raykar et al. (2010) both used the Expectation-Maximization (EM) algorithm (Dempster et al., 1977) while Liu and Wang (2012) used Gibbs sampling. Variational inference has also been applied and shown to perform well for crowdsourcing models (Liu et al., 2012).

3 METHODS

We now present the details of our probabilistic graphical model, including details of representation, inference and learning (Koller and Friedman, 2009). We first define a joint probability model over all observed and hidden variables of interest, conditioned on the parameters. In Section 3.2, we present our inference method, which involves an efficient variational algorithm. This estimates the distribution over the hidden variables, assuming the parameters are known. In Section 3.3, we present an EM approach to learn the parameters from data. Finally, we extend our approach to the multitask setting in section 3.4.

3.1 MODEL

We assume that each worker has a latent sensitivity and specificity, and that these two quantities are correlated. This assumption has similarly been made in medicine for estimating diagnostic test performance (Dahabreh et al., 2012). Following this work, we explicitly model the correlation between sensitivity and the false positive rate (FPR) (=1− specificity).

Let Z_i be the (potentially unobserved) true label for instance i and L_{ij} be the label provided for i by worker j. We then model worker j using two hidden variables, U_j and V_j, these capture transformations of worker sensitivity and FPR, and are assumed to be drawn from a bivariate normal with a covariance matrix to be estimated.

More precisely, the generative process is as follows:

$$U_j, V_j \sim \mathcal{N}(\mu, C) \tag{1}$$
$$Z_i \sim \mathrm{Ber}(\theta) \tag{2}$$
$$L_{ij}|Z_i = 1 \sim \mathrm{Ber}(\mathcal{S}(U_j)) \tag{3}$$
$$L_{ij}|Z_i = 0 \sim \mathrm{Ber}(\mathcal{S}(V_j)) \tag{4}$$

$\mathrm{Ber}(p)$ is the Bernoulli distribution with parameter p en-

coding the probability of the variable taking the value 1. $\mathcal{N}(\mu, C)$ is the bivariate Normal distribution with mean vector μ and covariance C: the correlation between U and V is thus modeled by the off-diagonal entries in C (C is symmetric; the two off-diagonal entries are equal). $\mathcal{S}$ is the Sigmoid function: $\mathcal{S}(x) = 1/(1 + \exp(-x))$, which maps real numbers to the interval $[0, 1]$. U_j and V_j are thus the logit-transformed sensitivities and FPRs of workers (the logit function is the inverse of the sigmoid fuction).

Note that μ and C are group-level parameters, capturing expected sensitivity and FPRs across all workers. Thus ours may be viewed as a 'fixed effects' model (Hedges, 1994) of worker quality, as we assume individual worker parameters are drawn from a shared parent distribution. This is in contrast to much of the previous work on this task, which has often modeled individuals independently (although there have been exceptions to this, e.g., Liu and Wang (2012)). In one scenario we assume that workers belong to distinct groups: experienced workers and novices. We also assume that we know *a priori* to which groups workers belong. In this case we fit separate models for each group, deriving corresponding distinct estimates for mean sensitivities and FPRs (and covariances).

Taking a Bayesian view, one may place priors on the shared variables μ, C and θ. However this increases model complexity and introduces the need to specify appropriate priors. Intuitively, these parameters are informed by all or most of the items in the dataset, and we should therefore have considerably less uncertainty around our estimates of them, compared to the hidden variables Z, U and V (which are informed by one or a small number of items).

Putting the components together, the unnormalized joint posterior of our model has the form:

$$P(U_{1..m}, V_{1..m}, Z_{1..n}, L) = \prod_{j=1}^{m} \mathcal{N}(U_j, V_j|\mu, C)$$
$$\prod_{i=1}^{n} \mathrm{Ber}(Z_i|\theta) \prod_{Z_i=1} \mathrm{Ber}(L_{ij}|\mathcal{S}(U_i)) \prod_{Z_i=0} \mathrm{Ber}(L_{ij}|\mathcal{S}(V_i)) \quad (5)$$

Where we are denoting the number of workers by m and the number of items by n.

3.2 INFERENCE

We aim to recover the posterior distribution of all of the hidden variables given worker labels: $p(U_{1..m}, V_{1..m}, Z_{1..n}|L)$. Unfortunately, evaluating this analytically is intractable. One possibility is instead to perform approximate inference via sampling methods such as Markov chain Monte Carlo (MCMC). However, practical implementations of MCMC, such as BUGS (Spiegelhalter et al., 1995) and PyMC (Patil et al., 2010), do not scale to large datasets, rendering this approach infeasible for our application.

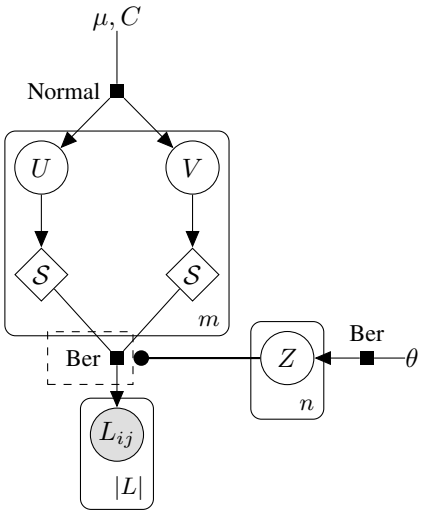


Figure 1: The Factor Graph of our model. Circles represent random variables (shaded variables are observed), squares depict factors, diamonds are deterministic mappings, edge endpoints (μ, C, θ) are parameters, plates denote repetitions, dotted plates are gates (in this case, the value of Z is used to select which one of the two $S(U)$ and $S(V)$ is used as the parameter for the Bernoulli distribution.

We therefore propose a novel variational inference algorithm for the model specified above. Variational approaches (Wainwright and Jordan, 2008) aim to approximate the true posterior p via a simpler distribution over the same variables: $q(u_{1..m}, v_{1..m}, Z_{1..n})$. The idea is to make q 'close' to p by minimizing the Kullback-Leibler (KL) divergence between the two, i.e., $\mathbb{KL}(q||p)$. By minimizing this divergence, we are maximizing a lower bound on the data log likelihood.

A typical strategy in variational inference is to make the *mean field* assumption, i.e., assume that q neatly factorizes:

$$q(U_{1..m}, V_{1..m}, Z_{1..n}) = \prod_{j=1}^{m} q(U_j)q(V_j) \prod_{i=1}^{n} q(Z_i) \quad (6)$$

where each distribution q on the right hand side is over one hidden variable (disambiguated by the argument), and has the form:

$$q(U_j) = \mathcal{N}(\tilde{\mu}_{uj}, \tilde{\sigma}^2_{uj}) \quad (7)$$
$$q(V_j) = \mathcal{N}(\tilde{\mu}_{vj}, \tilde{\sigma}^2_{vj}) \quad (8)$$
$$q(Z_i) = \mathrm{Ber}(\tilde{\theta}_i) \quad (9)$$

Here $\{\tilde{\mu}_{uj}, \tilde{\sigma}^2_{uj}, \tilde{\mu}_{vj}, \tilde{\sigma}^2_{vj}|j = 1...m\}$ and $\{\tilde{\theta}_i|i = 1...n\}$ are the variational parameters that should be selected to

minimize $\mathbb{KL}(q||p)$. This optimization problem can be solved via coordinate descent by updating each factor distribution while keeping all others fixed. The general mean field update for a vector X of hidden variables has the form:

$$q^*(X_i) \propto \exp\{\mathbb{E}_{-q_i} \log P(X)\} \tag{10}$$

$P(X)$ is the unnormalized posterior (here, Equation 5) and E_{-q_i} is the expectation with respect to all variables except X_i. By this equation, we can update $q(X_i)$ by making changes to its variational parameters. The equation updates our estimate over a variable based on the current belief over its neighbors (terms involving other variables are absorbed into the constant). Adapting this general form to our model, we can derive the following update equations:

$$q^*(Z_i = 1) \propto \exp\Big\{ \log \mathrm{Ber}(1|\theta) + \sum \mathbb{E}_{U_j \sim q(U_j)} \log \mathrm{Ber}(L_{ij}|\mathcal{S}(U_j)) \Big\} \tag{11}$$

$$q^*(Z_i = 0) \propto \exp\Big\{ \log \mathrm{Ber}(0|\theta) + \sum \mathbb{E}_{V_j \sim q(V_j)} \log \mathrm{Ber}(L_{ij}|\mathcal{S}(V_j)) \Big\} \tag{12}$$

$$q^*(U_j) \propto \exp\Big\{ \mathbb{E}_{V_j \sim q(V_j)} \log \mathcal{N}(U_j, V_j|\mu, C) + \sum q(Z_i = 1) \log \mathrm{Ber}(L_{ij}|\mathcal{S}(U_j)) \Big\} \tag{13}$$

$$q^*(V_j) \propto \exp\Big\{ \mathbb{E}_{U_j \sim q(U_j)} \log \mathcal{N}(U_j, V_j|\mu, C) + \sum q(Z_i = 0) \log \mathrm{Ber}(L_{ij}|\mathcal{S}(V_j)) \Big\} \tag{14}$$

We have elided indices in summations above for brevity. The sums in Equations 11 and 12 are over all of the workers that have provided labels for item i. The sums in Equation 13 and 14 are over all of the items that worker j has labeled. Recall that in Inference, the parameters μ, C and θ are assumed to be known.

Intuitively, Equations 11 and 12 consider item i and update our approximation of the posterior over Z_i by taking into account the prior θ and evidence from all of the worker labels provided for the item. Equation 13 concerns the (logit-transformed) sensitivity estimate for worker j, taking into account the bivariate Normal and the current approximation over the logit-transformed FPR V_j. The approximation is further updated using all of the items worker j has labeled, with respect to the current approximation over the true label Z_i of each. Equation 14 can be interpreted similarly, although here we consider the logit-transformed FPR estimate for worker j.

Although the update equations are available, evaluating them is difficult due to the model being non-conjugate. We thus applied Laplace Variational Inference (Wang and Blei, 2013), to directly approximate these equations. We first let $T_j = \begin{pmatrix} U_j \\ V_j \end{pmatrix}$ to treat U_j and V_j as a single variable and let $f(T_j)$ be the exponent in their update equation:

$$f(T_j) = \log \mathcal{N}(T_j|\mu, C) + \sum q(Z_i = 1) \log \mathrm{Ber}(L_{ij}|\mathcal{S}(U_j)) + \sum q(Z_i = 0) \log \mathrm{Ber}(L_{ij}|\mathcal{S}(V_j)) \tag{15}$$

By using a Laplace approximation on f, we can derive the approximate update for U_j and V_j:

$$q^*(T_j) \propto \exp(f(T_j)) \approx \mathcal{N}\left(\widehat{T}_j, \nabla^2 f(\widehat{T}_j)^{-1}\right) \tag{16}$$

where $\widehat{T}_j$ is the maximum of $f(T_j)$, can be found by numerical optimization, and the Hessian matrix $\nabla^2 f(T_j)$ can be derived analytically by using the result:

$$\nabla^2 \log \mathcal{N}(T|\mu, C) = C^{-1}(T - \mu)(T - \mu)^T C^{-1} - C^{-1} \tag{17}$$

For the variable Z_i, the expectations in their update equations (11 and 12) can also be approximated similarly, for example, let $g(u) = \log \mathcal{S}(u)$, we have

$$\mathbb{E}_{u \sim \mathcal{N}(\mu, \sigma)} g(u) \approx g(\mu) + \frac{1}{2} g''(\mu)\sigma \tag{18}$$

Again, the second derivative g'' can be derived analytically:

$$g''(u) = -e^u/(1 + e^u)^2 \tag{19}$$

The inference procedure initializes the variational distribution q at some value and then applies the update equations iteratively until convergence. In our implementation, we iterate until the average changes in the variational parameters is less than 0.01. To initialize the means of $q(U_j)$, $q(V_j)$ and $q(Z)$, we use majority voting. To initialize the variance of $q(U_j)$ and $q(V_j)$, we make use the Beta distribution. For example, suppose majority voting predicts that worker j has provided a True Positives and b False Negatives so that his sensitivity can be estimated as $a/(a + b)$. We initialize the variance $\tilde{\sigma}_{uj}$ of the logit-transformed sensitivity U_j as the logit-tranformed variance of $\mathrm{Beta}(a, b)$, which intuitively gives a smaller variance for workers who have provided more labels.

3.3 LEARNING

We consider μ, C and θ as parameters to be learned from data. The learning algorithm is a simple application of EM. In the E step, we perform variational inference to estimate the posterior over the hidden variable, given all of the the workers' labels. In the M step, we maximize the parameters μ, C and θ under that posterior distribution. For the

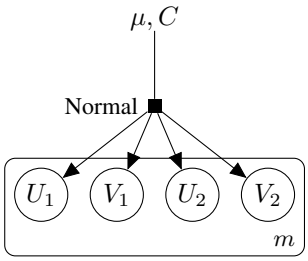


Figure 2: The Factor Graph highlights the extension to two tasks from the single task model in **Figure 1**.

parameters μ and C, the expected sufficient statistics (with respect to the variational distribution) can be evaluated:

$$\mathbb{E}(U_j) = \tilde{\mu}_j \tag{20}$$

$$\mathbb{E}(U_j^2) = \tilde{\mu}_{uj}^2 + \tilde{\sigma}_{uj}^2 \tag{21}$$

$$\mathbb{E}(U_J V_J) = \tilde{\mu}_{uj}\tilde{\mu}_{vj} \tag{22}$$

and then plugged into an estimator of multivariate Normal mean and covariance. For θ, it is simply set to the expected proportion of positive items, which is the average of $\{\tilde{\theta}_i | i = 1...n\}$.

3.4 MULTITASK MODEL

In addition to exploiting correlations between worker sensitivities and FPRs, our model can easily accommodate other sorts of worker performance correlations. For example, in this section we show that the same model can be adopted to capitalize on correlated performances across related labeling tasks. Specifically, we assume that workers have different sensitivities and FPRs in different tasks but that these values are correlated across tasks. Let U_1, V_1, U_2, V_2 be the logit-transformed sensitivities and FPRs in the first and second task. These are assumed to be generated from a four-dimensional Normal distribution:

$$\begin{pmatrix} U_1 \\ V_1 \\ U_2 \\ V_2 \end{pmatrix} \sim \mathcal{N}\left(\mu, C = \begin{pmatrix} A & X \\ X^T & B \end{pmatrix}\right) \tag{23}$$

where A, B and X are 2×2 matrices. A and B are the intra-task covariance matrices. X models the correlations across tasks, for instance X_{11} is the correlation between U_1 and U_2. $X_{11} > 0$ means that a worker with high sensitivity in the first task is likely to have high sensitivity in the second task. Our idea is that each task has its own mean sensitivity and specificity to represent its difficulty. On the other hand, the covariance matrix C represents how these sensitivites and specificities in two tasks are correlated. **Figure 2** is an illustration. Our inference and learning algorithms can be easily extended to this model.

4 EXPERIMENTS

We conducted experiments on two large 'citizen science' datasets to compare our proposed method to baselines. In citizen science, workers volunteer to contribute to science, without financial compensation.

We report the Root Mean Square Error (RMSE) of the predicted worker sensitivities and specificities:

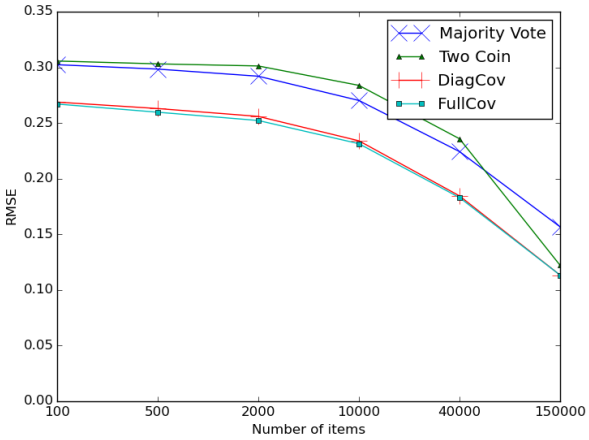
$$\text{RMSE} = \sqrt{\frac{1}{m}\sum_{j=1}^{m}(\text{Predicted}_j - \text{True}_j)^2} \tag{24}$$

where Predicted_j is the sensitivity or specificity of worker j that is predicted by a method and True_j is the corresponding true value. In practice, a worker true sensitivity and specificity are latent but can be accurately estimated given that the worker have labeled a large number of items and the true labels for those items are available: $\text{Sensitivity} = \text{TP}/(\text{TP}+\text{FN})$ and $\text{Specificity} = \text{TN}/(\text{TN}+\text{FP})$ where TP, TN, FP and FN are the number of true (false) positives (negatives) that the worker have labeled. We calculate the sensitivity and specificity estimates using all of the available data and treat those as the true values (or gold standard) for evaluation. Also, to ensure the quality of such gold standard, we only include in our evaluation the workers who have provided labels for at least 5 positives and 5 negatives (on the entire dataset). We note that the methods being evaluated are given a small portion of the dataset and still need to produce good estimates based on very few crowd labels and without access to the true labels. A similar approach to evaluate confusion matrix estimates has been used by Lakkaraju et al. (2015).

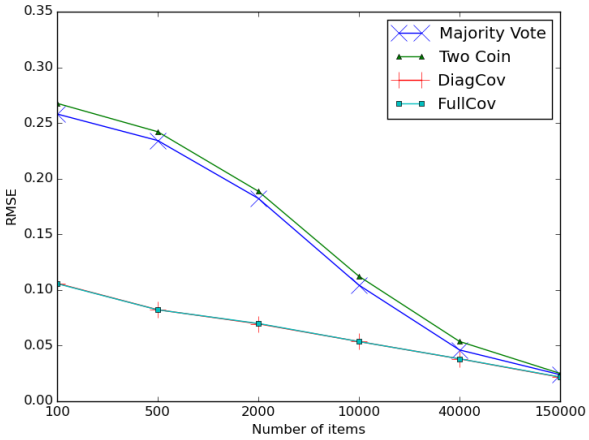
4.1 BIOMEDICAL CITATION SCREENING

We consider data from the EMBASE screening project (http://screening.metaxis.com/EMBASE/), which aims to identify reports of randomized controlled trials (RCTs) from EMBASE,[1] a biomedical literature database. The aim is to be comprehensive, thus placing an emphasis on sensitivity. An important property of this dataset is its imbalance: fewer than 5% of the items are positive. Our model aims to improve the estimates of sensitivities using their correlations with specificities (recall that sensitivity is the probability of being correct given a positive, and we expect this estimate to be poor given very few positives). Also, the screening project has relied on a mix of novice volunteer workers and domain experts with associated costs and levels of expertise. Our model can easily exploit such information on two groups of workers by using two different set of parameters (μ and C), one for each group. This is a large dataset with 151,224 items and 576 workers. Of

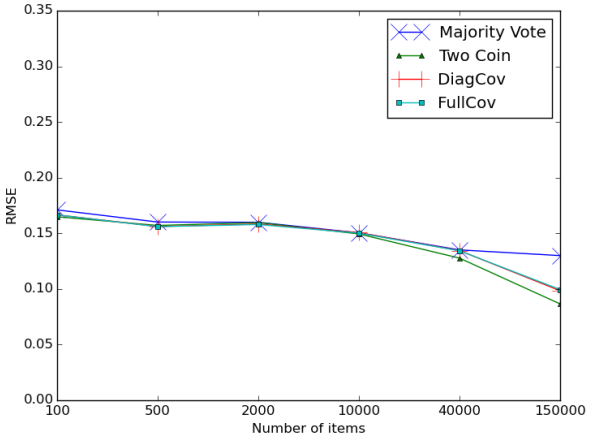
[1] RCTs are experiments in which participants are randomized to groups in which individuals are exposed to different interventions; one group is designated as a control.



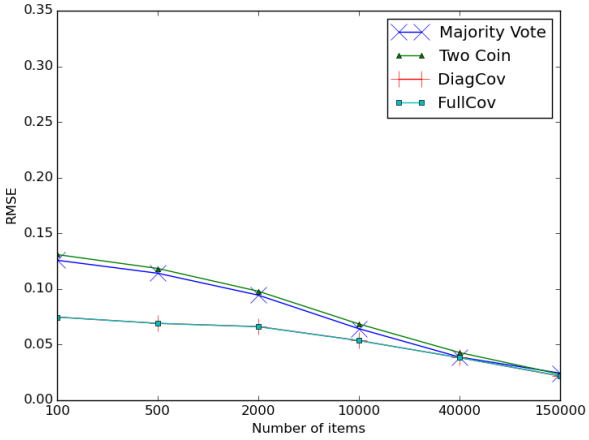
(a) Sensitivity with Uniform Prior (1,1)



(b) Specificity with Uniform Prior (1,1)



(c) Sensitivity with Informative Prior (4,1)



(d) Specificity with Informative Prior (4,1)

Figure 3: The RMSE in Sensitivity and Specificity with Uniform and Informative Prior of four methods (averaged over 5 runs). DiagCov and FullCov are two variants of our model; their curves *overlap* in (b), (c) and (d).

those workers, 156 have labeled at least 5 positives and 5 negatives and are used for evaluation. The number of labels per workers is generally very skewed: 525 workers (91%) provided less than the average of 603 labels.

We implemented two variants of our method: one with a full covariance matrix (FullCov) and one with a covariance matrix constrained to be diagonal (DiagCov) i.e., entries in the confusion matrices are assumed to be uncorrelated. This allows us to observe how much improvement is due to modeling correlation and how much is from modeling worker groups. We compare these two variants to two baselines: Majority Vote, in which sensitivities and specificities are estimated based on the majority labels, and the *Two Coin* model (without features) due to Raykar et al. (2010, sec. 2.7.4)

We also implemented the *Hybrid Confusion* model by Liu and Wang (2012), which is the same as Two Coin but with inference by Gibbs sampling. Because the results were very similar to Two Coin, we omit these for clarity.

All of the methods are given the same prior or initialized in the same way. The Two Coin model has Beta priors on worker sensitivities and specificities, which can be interpreted as smoothing constants. The same constants are given to Majority Vote to smooth its estimates. Our model has no prior on the parameters μ and C but those and the variational parameters are initialized using the outputs from Majority Vote (with smoothing constants). To explore the effect of these worker priors (or initialization), we did experiments with a uniform prior ($\mathrm{Beta}(1,1)$ as done by Raykar et al. (2010)) and an informative prior $\mathrm{Beta}(4,1)$ as in Liu and Wang (2012)). The prior on the class propor-

	Items	Workers	LPI	LPW
Task 1	17862	1242	16.8 ± 3.9	241 ± 288
Task 2	6476	198	5.0 ± 2.1	163 ± 130
Task 3	21951	681	6.7 ± 3.7	215 ± 243
Task 4	21915	679	6.7 ± 3.7	215 ± 243

Table 1: Statistics of four tasks we consider in the Galaxy Zoo 2 dataset, after pre-processing. 'LPI' stands for 'Labels per item' and 'LPW' for 'Labels per worker'. In these columns, we report the means and standard deviations of the number of labels per item (worker).

tion θ is always uniform ($\mathrm{Beta}(1,1)$ as done by both).

Figure 3 presents our results with RMSE on the Y-axis and the number of items on the X-axis. We average results over 5 runs, randomly sampling a number of items from the dataset for each. The two plots above are for uniform prior. We see that Two Coin's performance is surprisingly weak while two variants of our model achieve the best performance. On the plot for Sensitivity, we also see some small improvement of FullCov over DiagCov (but significant in our paired t-test). In the plot for Specificity, those two variants have the same performance (the curve for FullCov has overwritten the one for DiagCov). This is what we expect since the specificity estimates are for the majority (negative) class and the correlation has little effect given a large number of labels available. Surprisingly, much of the improvement of our method can be attributed to the 'group part' of our model (not the 'correlation part'). As discussed above, the 'group part' provides a 'back off' to the group level estimates when there is not enough data on a worker.

The two plots below show our results for Informative Prior, where all of the methods perform better, as expected. For Sensitivity, Two Coin performs well, slightly better than Majority Vote while comparable to ours for the most part and slightly better than ours for $40,000$ or more items. However, for specificity, it still performs worse than Majority Vote and ours. The difference is probably due to $\mathrm{Beta}(4,1)$ being a better prior for sensitivity than specificity[2]. This suggests that Two Coin and similarly Hybrid Confusion can perform well but their performance are dependent on good priors. In contrast, our method is robust across different settings of priors. This can probably be explained by the fact that our group level estimates (which play role in 'backing off' sparse workers) are learned from data while Two Coin's priors are set to constants.

4.2 GALAXY MORPHOLOGICAL CLASSIFICATION

The Galaxy Zoo 2 dataset (Willett et al., 2013) consists of labels provided by volunteer workers on morphological classification of galaxies. A worker is given an image of a galaxy and is asked multiple questions. We consider each question to be a labeling task. Specifically in our experiments, we consider the (simplified) first four questions/tasks:

1. Is the galaxy smooth or disk-like? If the answer is disk, proceed to task 2, otherwise stop. [3]
2. Is the disk viewed edge-on? If the answer is no, proceed to task 3, otherwise stop.
3. Is there a bar in the center? Proceed to task 4 regardless of the answer.
4. Is there a spiral arm pattern?

Since the tasks have varying difficulties and require varying skills, worker performance in each task can be very different from others, but we can naturally expect some degree of transferability between tasks. We aim to evaluate our multitask model on improving the estimates of worker sensitivities and specificities on a *target task*, given labels on a *source task* (for the same set of workers). We did experiments in two scenarios:

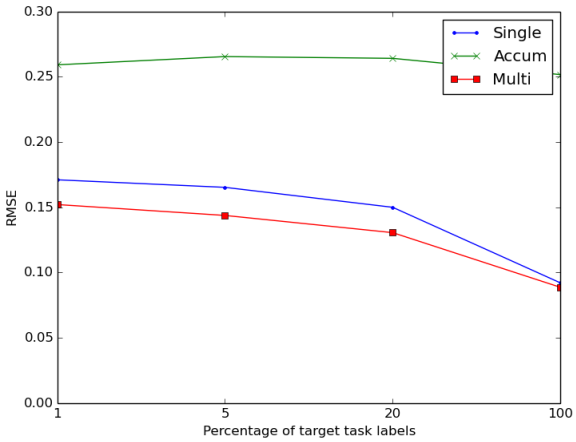
1. Conditional Task 1 $\rightarrow$ Task 2: The methods are given all of the labels in Task 1, a portion of labels in Task 2 and must estimate worker sensitivities and specificities in Task 2.
2. Independent Task 3 $\rightarrow$ Task 4: The methods are given all of the labels in Task 3, a portion of labels in Task 4 and must estimate worker sensitivities and specificities in Task 4. Here the two tasks are independent, while in the first scenario, Task 2 is asked only when the worker answers 'disk' in Task 1.

We compare our multitask model (*Multi*) to two baselines: *Single*, where only labels in the target task are considered, and *Accum*, where labels from the source task are merged with those from the target task. The Accum baseline is only applicable when the two tasks are questions with the same number of choices and a matching of these choices is available (otherwise the labels could not be merged). Our approach has no such restriction. For both baselines, we used the FullCov variant of our method.

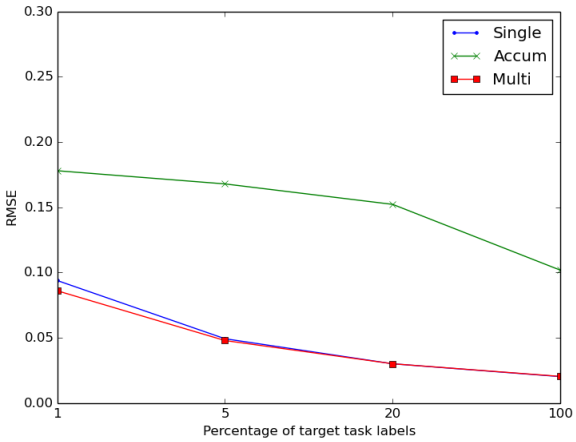
The dataset is extremely large, with nearly 60 million labels for 11 tasks from over 83 thousand workers. We do the following pre-processing to reduce the size of the dataset. (1) We take the first million labels (for all of 11 tasks). (2) For each of the first four tasks, we filter out workers with less than 100 labels and items with less than 3 labels. The statistics of the four tasks after pre-processing are in **Table 1**. We note that the gold standard worker sensitivities and specificities are estimated from the *entire* dataset (without pre-processing).

[2] $\mathrm{Beta}(4,1)$ has a mean of 0.8. The true means for sensitivities and specificities are 0.78 and 0.94

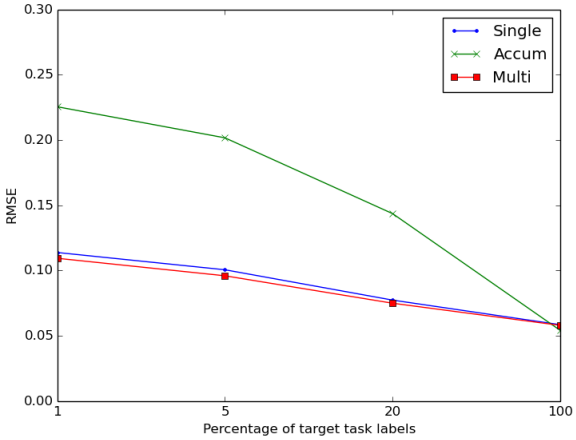
[3] 'Stop' means go to a question we don't consider. This question has a third answer ('star of artifact') which is very rare and we don't consider for simplicity.



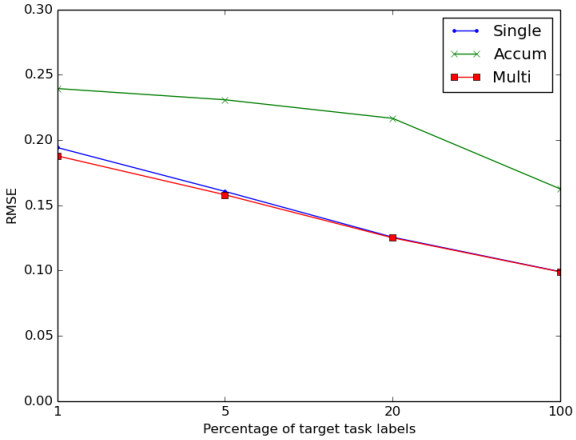
(a) Sensitivity: Conditional Task 1 → Task 2



(b) Specificity: Conditional Task 1 → Task 2



(c) Sensitivity: Independent Task 3 → Task 4



(d) Specificity: Independent Task 3 → Task 4

Figure 4: The RMSE against percentage of labels in the target task available of our method compared to two baselines (averaged over 5 runs).

In **Figure 4**, we report our results. Overall, Accum is surprisingly weak, giving estimates with much higher RMSE than Single. This shows that worker performance in two different tasks are sufficiently different that a naive transfer strategy is unlikely to work.

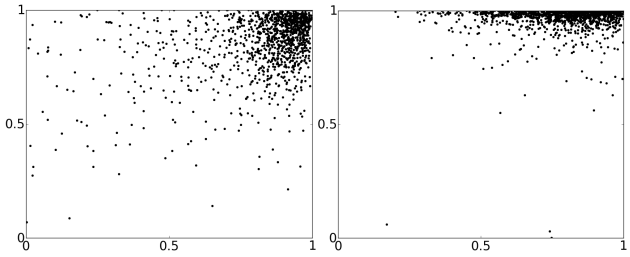
Compared to the Single baseline, our method has shown improvement for the case when a small percentage of labels in the target task is available. The improvement diminishes when more target task labels are available, as expected. On a close look at their differences, one might notice that the improvement is sometimes quite modest. Looking into the 'true' worker sensitivities and specificities (**Figure 5**), we found an overall positive correlation between tasks as expected. However, we also observe a surprisingly large number of workers who do better in the source task but worse in the second task (and vice versa). This may be because the tasks are somewhat subjective that the variations in workers performance are mostly due to their different perception and interpretation. In short, we believe the true multitask correlation plays a role in how much improvement we observe.

To further investigate this, we repeat the same experiments on simulated labels. We note that the purpose of our simulation is to complement, not to replace our results on the real data. The simulated labels are generated by our model from the following parameters:

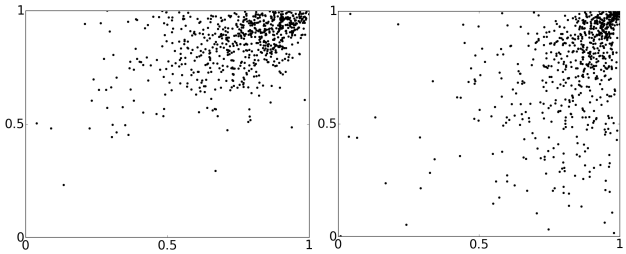
$$\mu = \begin{pmatrix} 1.49 \\ -1.45 \\ 2.18 \\ -2.59 \end{pmatrix} C = \begin{pmatrix} 1.80 & 0 & x & 0 \\ 0 & 1.30 & 0 & x \\ x & 0 & 1.06 & 0 \\ 0 & x & 0 & 1.89 \end{pmatrix}$$

The mean μ and the variances in C are set to the empirical estimates in Task 1 and Task 2 while x is the inter-task correlation and is varied in $\{0.5, 0.75, 1.0, 1.25\}$, some posi-

tive values in a reasonable range which keep the covariance matrix positive definite. We assume 5000 items in each task and 200 workers. Each item is labeled by 5 randomly selected workers. **Figure 6** shows that as the correlation x increases, we see a greater improvement of Multi over Single. That is, the more correlated the worker performances in different tasks are, the greater the improvement realized by our model.

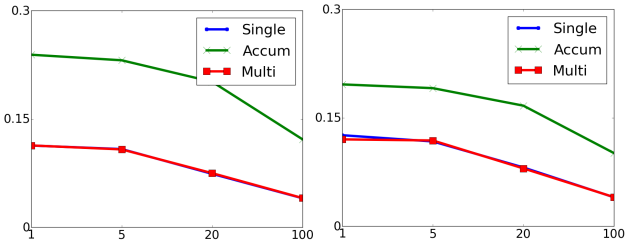


(a) Sensitivity Task 1 vs. 2 (b) Specificity Task 1 vs. 2

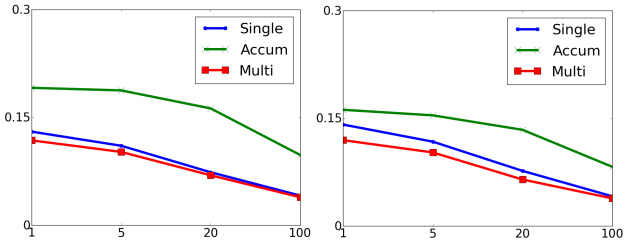


(c) Sensitivity Task 3 vs. 4 (d) Specificity Task 3 vs. 4

Figure 5: Worker sensitivities (specificities) in two tasks. Each point is a worker. In (a) and (b), the X-axis is task 1 and the Y-axis is task 2. In (c) and (d), the X-axis is task 3 and the Y-axis is task 4.



(a) $x = 0.50$ (b) $x = 0.75$



(c) $x = 1.00$ (d) $x = 1.25$

Figure 6: The RMSE in Sensitivity (averaged over 5 runs) on simulated labels for 4 values of the inter-task correlation x. The X-axis is the percentage of target task labels and the Y-axis is RMSE. The curves for Single and Multi overlap in (a). The results for Specificity are similar.

5 CONCLUSION

We have presented our approach to improve the estimates of worker confusion matrices and reported the results of our experiments on real and simulated data. Our main idea is to exploit the correlations in the matrix entries (sensitivities and specificities) and the knowledge of groups in the workers population. The idea also applies to the case when labels from multiple tasks are available. In all of the cases we consider, our method shows good performance compared to baselines. We have made our source code available[4]. We expect the datasets to be available on request from their owners.

While we have reported on binary classification tasks with no instance-level features, where a confusion matrix reduces to sensitivity and specificity, our approach can be easily generalized. For future work, we will extend our work to categorical tasks, with features when available. Such features can be modeled by additional variables associated with the instances and a variational algorithm can be derived similar to (Felt et al., 2015). We are also interested in a better model for the multitask setting, which can capture important factors such as item and task difficulties as well as worker skill and expertise. Also, we would like to take advantage of the the full posterior distribution over the worker confusion matrices in an application such as an online decision system (Nguyen et al., 2015; Werling et al., 2015), rather than using only point estimates. Finally, while we have favored variational inference over MCMC, recent probabilistic languages such as Stan (Carpenter et al., 2015) is an attractive alternative and interesting to compare to our approach.

[4]https://github.com/thanhan/code-uai16

Acknowledgments

We thank the EMBASE screening project for providing the RCT dataset, Kyle Willett for providing the Galaxy Zoo 2 dataset and the anonymous reviewers for valuable comments. We are also grateful for the volunteers who contribute to these datasets and make this work possible. This study was supported in part by National Science Foundation grant No. 1253413 and IMLS grant RE-04-13-0042-13. Any opinions, findings, and conclusions or recommendations expressed by the authors are entirely their own and do not represent those of the sponsoring agencies.

References

Bi, W., Wang, L., Kwok, J. T., and Tu, Z. (2014). Learning to predict from crowdsourced data. In *Uncertainty in Artificial Intelligence*.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., and Riddell, A. (2015). Stan: a probabilistic programming language. *Journal of Statistical Software*.

Dahabreh, I. J., Trikalinos, T. A., Lau, J., and Schmid, C. (2012). *An Empirical Assessment of Bivariate Methods for Meta-Analysis of Test Accuracy*. Agency for Healthcare Research and Quality (US).

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

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

Felt, P., Ringger, E., Seppi, K., Black, K., and Haertel, R. (2015). Early gains matter: A case for preferring generative over discriminative crowdsourcing models. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Hedges, L. V. (1994). Fixed effects models. *The handbook of research synthesis*, pages 285–299.

Kajino, H., Tsuboi, Y., and Kashima, H. (2012). A convex formulation for learning from crowds. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*.

Kim, H.-C. and Ghahramani, Z. (2012). Bayesian classifier combination. In *International conference on artificial intelligence and statistics*, pages 619–627.

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Lakkaraju, H., Leskovec, J., Kleinberg, J., and Mullainathan, S. (2015). A bayesian framework for modeling human evaluations. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pages 181–189.

Liu, C. and Wang, Y.-m. (2012). Truelabel+ confusions: A spectrum of probabilistic models in analyzing multiple ratings. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 225–232.

Liu, Q., Peng, J., and Ihler, A. T. (2012). Variational inference for crowdsourcing. In *Advances in Neural Information Processing Systems*, pages 692–700.

Nguyen, A. T., Wallace, B. C., and Lease, M. (2015). Combining crowd and expert labels using decision theoretic active learning. In *Proceedings of the 3rd AAAI Conference on Human Computation (HCOMP)*.

Patil, A., Huard, D., and Fonnesbeck, C. J. (2010). Pymc: Bayesian stochastic modelling in python. *Journal of statistical software*, 35(4):1.

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

Reitsma, J. B., Glas, A. S., Rutjes, A. W., Scholten, R. J., Bossuyt, P. M., and Zwinderman, A. H. (2005). Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of clinical epidemiology*, 58(10):982–990.

Sheshadri, A. and Lease, M. (2013). Square: A benchmark for research on computing crowd consensus. In *First AAAI Conference on Human Computation and Crowdsourcing*.

Simpson, E., Roberts, S., Psorakis, I., and Smith, A. (2013). Dynamic bayesian combination of multiple imperfect classifiers. In *Decision Making and Imperfection*, pages 1–35. Springer.

Spiegelhalter, D. J., Thomas, A., Best, N. G., and Gilks, W. R. (1995). Bugs: Bayesian inference using gibbs sampling, version 0.50. *MRC Biostatistics Unit, Cambridge*.

Venanzi, M., Guiver, J., Kazai, G., Kohli, P., and Shokouhi, M. (2014). Community-based bayesian aggregation models for crowdsourcing. In *Proceedings of the 23rd international conference on World wide web*, pages 155–164. ACM.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305.

Wang, C. and Blei, D. M. (2013). Variational inference in nonconjugate models. *The Journal of Machine Learning Research*, 14(1):1005–1031.

Werling, K., Chaganty, A. T., Liang, P. S., and Manning, C. D. (2015). On-the-job learning with bayesian decision theory. In *Advances in Neural Information Processing Systems*, pages 3447–3455.

Willett, K. W., Lintott, C. J., Bamford, S. P., Masters, K. L., Simmons, B. D., Casteels, K. R., Edmondson, E. M., Fortson, L. F., Kaviraj, S., Keel, W. C., et al. (2013). Galaxy zoo 2: detailed morphological classifications for 304 122 galaxies from the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society*, page stt1458.