

---

# Subjective Ordinal Labels: Implications for Hypothesis Testing and Machine Learned Prediction Models

---

Gordon Rios\*

Hongyuan Zha†

## Abstract

In practice, due to the cost of labeling or tagging, subjective data is often used without adjustments for label uncertainty. This problem falls into the gap between textbook applications of machine learning techniques and research level methodology. A variety of approaches have been discussed in detail throughout the statistical research literature but more effort could be made to motivate this problem for the practitioner. This paper lays out basic mathematical arguments and remedies for the analyzing and modeling subjective ordinal response data in the presence of label uncertainty.

## 1 Introduction

Modeling and analysis of subjective ordinal data is pervasive in medical and industrial settings. Self reporting degree of symptom severity, evaluating search engine relevance, judging product or service quality, are all examples where subjective ordinal response data is collected. Often, only one response per case is collected and the data is aggregated for discriminating between alternative programs or used for learning prediction functions that can be used to make better decisions. The problem of error in the label assignment process is typically acknowledged but seems to fall into the gap between research and practice since the problem typically isn't treated or discussed in commonly used textbooks on applied statistics and modeling.

In practice, one finds commonly held opinions about error in subjective ordinal data that create barriers to pursuing more advanced methods of analysis and modeling. For hypothesis testing of the superiority of one program or another, one typically encounters the belief that label assignment error will be captured by measuring an increase in the observed or nominal variance. This variance, or general dispersion, is expected to appropriately reduce the significance of statistical tests applied to the data. In the next section, a set of direct arguments are used to illustrate why this practice is inappropriate. Some of these arguments are simplifications of the thorough discussion of bayesian identifiability presented for this problem in [8]. Further, there has been a lot of work in modeling the error process via maximum likelihood and applying EM algorithm [3], computing measures such as  $\kappa$

---

\*Yahoo! Inc.

†Department of Computer Science and Engineering, The Pennsylvania State University, University Park, PA 16802, zha@cse.psu.edu. The work of this author was supported in part by NSF grants DMS-0311800.

for observer disagreement [1], or attempting to correct for observer bias [9]. Some of this work has made it into practical settings but often with incomplete success.

For modeling purposes, the error in the data is known to exist but it's assumed that one can simply "power through the noise" and devise models that faithfully reflect the true conditional expectation of the label or associated score given the features or input variables. In a later section, it will be shown directly that prediction error is strictly increasing in the bayes error of the label assignments and thus will create a permanent impediment to modeling that no amount of data will remove. Remedies are discussed in the final section but good research has been done on strongly related problems such as "multiple-label" [5] (and cited works) or by specializing the approach described as "learning from an imprecise teacher" set forth in [2].

The widespread practice of associating scores with the ordinal labels and treating the results as continuous is presented without critique since this paper is concerned with other aspects of working with subjective labels. Details on this practice is reasonable for machine learning tasks can be found in [4]. In the following discussion, scores and labels, both observed and true, are drawn from the sets  $S$  and  $L$  respectively. Observed elements from these sets will be designated  $\tilde{s}$  and  $\tilde{l}$  when necessary to clarify exposition.

## 2 Label Assignment Error and System Uncertainty

A simple characterization of assignment error is the matrix  $Y$  of conditional probabilities where, for row  $i$  and column  $j$ ,  $Y_{i,j} = p(l_j|l_i)$  for observed label  $l_j$ , true label  $l_i$ , and with  $Y_{i,i} = 1 - \sum_{j \neq i} Y_{i,j}$ . The matrix  $Y$  can be estimated using a control group of perfectly labeled cases where for When crossed with the prior this matrix measures the bayes error of the labels. The matrix  $Y$  can be used to calculate the effects of increasing error on two important aspects of the distribution: nominal variance of the associated scores and total system entropy. A contrast can also be made between *nominal entropy* and system entropy but later it will be shown that labeling error guarantees positive divergence between them.

### 2.1 Nominal Variance

The formula for nominal variance of the score distribution given labeling error  $P(S|Y)$  requires the prior distribution of labels  $\pi$ . One can recover the value of the true prior  $\pi$  from  $Y$  and the observed prior  $\pi^*$  by solving the linear system  $Y^T \pi = \pi^*$ . Since  $Y^T$  is a stochastic matrix and assuming that  $\det(Y) \neq 0$  the existence of a unique solution is guaranteed.<sup>1</sup> Also, since the observed prior is a probability distribution (e.g. nonnegative with elements summing to one) the solution will also be a probability distribution. Using the quantities developed one can write the formula for nominal variance as

$V(S; Y, \pi) = E[S^2; Y, \pi] - E[S; Y, \pi]^2$ . To examine the derivative of  $V(S; Y, \pi)$  w.r.t. the individual cross error terms  $y_{i,j}$  for  $j \neq i$ :

$$\begin{aligned} \frac{\partial V(S; Y, \pi)}{\partial y_{i,j}} &= \frac{\partial E[S^2; Y, \pi]}{\partial y_{i,j}} - \frac{\partial E[S; Y, \pi]^2}{\partial y_{i,j}} \\ &= \frac{\partial E[S^2; Y, \pi]}{\partial y_{i,j}} - 2E[S; Y, \pi] \frac{\partial E[S; Y, \pi]}{\partial y_{i,j}} \end{aligned}$$

And, using the formulas for the two expected value components (rewriting  $E[S; Y, \pi]$ ):

$$E[S; Y, \pi] = \sum_i \pi_i \sum_j y_{i,j} s_j = \sum_i \pi_i \left\{ s_i + \sum_{j \neq i} y_{i,j} (s_j - s_i) \right\}$$

<sup>1</sup>This is covered in any linear algebra textbook (e.g. [7]).

and analogously for  $E[S^2; Y, \pi] = \sum_i \pi_i \{s_i^2 + \sum_{j \neq i} y_{i,j} (s_j^2 - s_i^2)\}$ . the expression for the derivative is given by

$$\frac{\partial V(S; Y, \pi)}{\partial y_{i,j \neq i}} = \pi_i ((s_j - s_i)(s_j + s_i) - 2E[S; Y, \pi]).$$

and with  $\pi_i \geq 0$  there is simple intuition behind the effects of error in  $y_{i,j \neq i}$ . When  $s_j > s_i$  and  $s_j + s_i < 2E[S; Y, \pi]$  an increase in error will actually *decrease* nominal variance. It's easy to see the various conditions in which an increase in error would increase, decrease, or result in no change in, the variance.

## 2.2 System Entropy

The effects of labeling error on system entropy are more direct. When labeling error is introduced into the system new states are added since the previous state for the true label is split up into multiple compound states  $z(i, j)$  where true label is  $l_i$  but observed label is  $l_j$ . Introducing this type of error must always increase system entropy. The initial system entropy over the labels is  $H(L) = -\sum_i \pi_i \log_2(\pi_i)$  so it follows that replacing one of the terms  $H(L)_i = \pi_i \log_2(\pi_i)$  in the sum with

$$H(L)_i^* = \alpha \pi_i \log_2(\alpha \pi_i) + (1 - \alpha) \pi_i \log_2((1 - \alpha) \pi_i)$$

for  $\alpha \in (0, 1)$ , will increase the sum since  $H(L)_i^* < -H(L)_i$ :

$$H(L)_i^* = \alpha \pi_i \log_2(\alpha \pi_i) + (1 - \alpha) \pi_i \log_2((1 - \alpha) \pi_i) \quad (2.1)$$

$$= H(L)_i + \pi_i \{\alpha \log_2(\alpha) + (1 - \alpha) \log_2(1 - \alpha)\} \quad (2.2)$$

and since the second part of the  $H(L)_i^*$  is strictly negative the claim is demonstrated. If  $\alpha$  is the amount of label error accruing to a specific erroneous state and  $(1 - \alpha)$  remaining in the true state then the effects on entropy ( $-H(L)_i^*$ ) are strictly increasing for  $\alpha \in (0, 0.5)$ .<sup>2</sup> Finally, this same argument shows that divergence between nominal entropy and system entropy is true by construction as labeling error is introduced into the system.

$$\frac{\partial H(L)_i^*}{\partial \alpha} = \pi_i (\log_2(\alpha) - \log_2(1 - \alpha))$$

Conceptually, it seems clear that while the effects of introducing labeling error into the nominal variance are ambiguous and dependent on associated scores, priors, etc., the effects on system entropy are direct and increasing over the range of interesting cases.

## 3 Contrasting Nominal Variance and System Entropy

It's useful to briefly consider three types of symmetric true label distributions: uniform, strong central tendency, and bimodal, to get an intuitive feel for the proposed divergence between nominal variance and system entropy. If changes are made evenly to both ends of the distribution then analysis can be done without factoring in the change in the mean. Essentially, I'm arguing that even with the most well behaved distributions the practice of using nominal variance for hypothesis testing is faulty. Analyzing the candidate distributions can be done with a simple exchange argument since label error results in density transfer between two ordinal labels ( $L$ ) and their associated scores ( $S$ ). Most of the effects will be at the extremes of the distribution so focus on pairs of labels on the same side of the mean. Choose two pairs of labels:  $(l_i, l_j)$  and  $(l_{K-j}, l_{K-i})$  with  $K$  the number of labels,  $s_i < s_j$ ,  $s_{K-j} < s_{K-i}$ , and  $s_i - s_j = s_{K-j} - s_{K-i}$  to ensure that changes in  $E[S]$  don't introduce bias. Focusing on the first pair, and assuming that the equivalent action is taken on the second pair, the effects on variance of the pairwise error between two labels ( $i, j$ ) is proportional<sup>3</sup> to a simple exchange equation:  $f(i, j) = \pi_i y_{i,j} - \pi_j y_{j,i}$ . When  $f(i, j) > 0.0$

<sup>2</sup>It's reasonable to consider pairwise error only upto 0.50 since error beyond that indicates more errors are made than correct assignments and the system should just reverse those labels.

<sup>3</sup>Since the action is mirrored in the second pair the equation captures half the density exchange.

variance will decrease because there is a net flow of event mass from the outer range of the distribution towards the center. With this simple equation it's easy to see the effects of label error on the three symmetric distributions:

1. Central Tendency (Unimodal): since  $\pi_i < \pi_j$  variance will increase unless  $y_{i,j}$  is proportionally larger than  $y_{j,i}$ .<sup>4</sup>
2. Uniform: since  $\pi_i = \pi_j$  then symmetric error  $y_{i,j} = y_{j,i}$  will have no effect on variance while  $y_{i,j} > y_{j,i}$  cases will *reduce* variance.
3. Bimodal: with  $\pi_i > \pi_j$  symmetric error  $y_{i,j} = y_{j,i}$  will actually *reduce* variance.

While system entropy continues to be strictly increasing in all cases by the same arguments used in section 2.2 above. Although I'll defer my discussion of bias until a later section I hope it's clear that asymmetric changes in labeling error, priors, or scores, will all create bias between the nominal and true distributions since in general one has no reason to expect that  $S^T \pi = S^T \pi^*$ .

## 4 Label Uncertainty and Hypothesis Testing

At this point it should be clear that doing analysis directly on the observed label and score distributions requires care. This includes nonparametric statistics performed on the nominal distribution. Though nonparametric statistics make no distributional assumptions all statistics require that random fluctuations in the data reflect the uncertainty in the system.

Consider hypothesis testing, typically choosing between treatments based on better expected scores or ordinal labels, by explicitly taking into account the matrix of cross error  $Y$ . One way to do that is to avoid direct comparisons between nominal scores (or functions thereof) and calculate the probability that, for each case in question, one treatment is better than the other.

### 4.1 Direct Comparisons of Single Labels

A basic example of testing over the nominal labels is to compare the results of two treatments,  $T_1$  and  $T_2$ , over  $n$  trials where  $T_{1,i}$  produces score  $s_{1,i}$  and  $T_{2,i}$  produces score  $s_{2,i}$  (with scores associated via the assigned labels  $l_{1,i}$  and  $l_{2,i}$ .) Since each score is part of a *related pair* the series to be tested is  $s_{2,i} - s_{1,i}$  for  $i = 1 \dots n$ .

A common practice is to treat the difference as continuous and test whether it is significantly different than zero using a *t-test*.<sup>5</sup> Since the test uses the nominal variance of the score distribution it will likely overestimate significance in the presence of label assignment error.<sup>6</sup>

A slightly more conservative practice is to create a series of score differences for each case and then test for significant difference from zero using a nonparametric statistical test. A popular choice for this application is the *Wilcoxon Signed Ranks Test for Related Pairs* which ranks the difference series and sums the ranks for positive and negative cases

---

<sup>4</sup>Informally, one often sees central tendency in subjective ordinal label assignments precisely because human assigners are hesitant to apply extreme labels which effectively results in  $y_{i,j} > y_{j,i}$ .

<sup>5</sup>Letting  $d = s_{2,i} - s_{1,i}$  the statistic is  $\frac{\bar{d}}{\sigma_d/\sqrt{n}}$  where  $n$  is the sample size and  $\sigma_d$  is the standard deviation of the differences.

<sup>6</sup>Since the variance of the  $d$  is a function of the nominal variance of the scores:  $\sigma_{S_2-S_1}^2 = \sigma_{S_2}^2 + \sigma_{S_1}^2 - 2\sigma_{S_2,S_1}$ .

(throwing away the zeros) to create  $T^+$  and  $T^-$  test values.<sup>7</sup> With  $d^+$  and  $d^-$  the positive and negative differences, compute the statistics (starting with  $T^+$ ) without ranks with

$$T^+ = \sum_{i \in 1, \dots, n^+} \left\{ \sum_{j \in 1, \dots, n^+} I(|d_i^+| > |d_j^+|) + \sum_{k \in 1, \dots, n^-} I(|d_i^+| > |d_k^-|) \right\}$$

with

$$\sum_{j \in 1, \dots, n^+} I(|d_i^+| > |d_j^+|) = \sum_{j \in 1, \dots, n^+} I(s_{2,i}^+ - s_{1,i}^+ > s_{2,j}^+ - s_{1,j}^+)$$

and

$$\sum_{k \in 1, \dots, n^-} I(|d_i^+| > |d_k^-|) = \sum_{k \in 1, \dots, n^-} I(s_{2,i}^+ - s_{1,i}^+ > s_{1,k}^- - s_{2,k}^-)$$

and add  $n^+$ , where  $n^+ = \#\{d^+\}$ ,  $I(x)$  the indicator function, and  $T^-$  calculated analogously. The order of the scores is reversed in the rhs of 4.1 since those differences  $d^-$  are negative. When the labels (and associated scores) have error the indicator functions become probabilities. Calculating simple probabilities of the true labels given those observed is taken up in the next section.

Another assumption of the Wilcoxon (and other nonparametric tests) is that of *i.i.d.* data (in the related pairs case one expects that the differences will be *i.i.d.*) However, even this simple assumption could be violated since the distribution of the difference depends on the specific labeling error for the associated scores. Even with symmetrical (and equal) error, the differences between an edge label and its neighbor will have different distribution than the differences between interior labels. Of course, without the assumption of either symmetric or equal error there's no reason to expect the differences to be *i.i.d.*<sup>8</sup> For example, the most conservative nonparametric procedure is the sign *sign test*<sup>9</sup> which in this setting requires only the *i.i.d.* assumption.

#### 4.1.1 Computing $P(T_2 > T_1)$ Directly

The correct approach, given label uncertainty, is to compute the probability that  $T_2$  is better than  $T_1$  given their respective nominal labels  $P(T_{2,i} > T_{1,i} | \tilde{l}_{1,i}, \tilde{l}_{2,i})$  for each case  $i$  and then take the expectation over those results to get the final probability. Letting  $l$  and  $\tilde{l}$  be labels drawn from  $L$  and the rows and columns of  $Y$  the equation for each case's probability is (after dropping the case subscript  $i$ ):

$$P(T_2 > T_1 | \tilde{l}_1, \tilde{l}_2) = \sum_{l_2, l_1 | l_2 > l_1} p(l_1, l_2 | \tilde{l}_1, \tilde{l}_2) = \sum_{l_2, l_1 | l_2 > l_1} p(l_1 | \tilde{l}_1) p(l_2 | \tilde{l}_2)$$

where  $l_1$  and  $l_2$  represent the true ordinal labels, assuming true labels between treatments are independent given the observed labels, calculating  $p(l_i | \tilde{l}_j) = \pi_i y_{i,j} / \sum_i \pi_i y_{i,j}$  using Bayes rule.

<sup>7</sup>Technically, the *Wilcoxon* test requires that the distribution of the scores be symmetric and continuous and since we're taking differences of a finite set of scores the application of the test is incorrect. However, in practice one often uses continuous functions of sets of values to compute each score. More serious is the frequent violation of symmetry in practice which is known to violate the test since it requires that the median be equal to the mean.

<sup>8</sup>Concretely, the density of a particular difference value (say 1.5) depends on which exact observed labels are used to generate the value. In general, even with regular interval scores, one can't expect  $P(s_2 - s_1) P(s_3 - s_2)$ .

<sup>9</sup>Binomial test on the counts of positive differences versus negative differences.

## 5 Prediction Models and Subjective Labeling Error

The discussion of learning prediction models over the nominal scores has several pieces. First, there are the direct effects from the bayes error captured in the  $Y$  matrix. I will show, perhaps trivially, that pointwise prediction error is strictly increasing in the amount of bayes error mass in the off diagonal entries of  $Y$ . I will discuss this issue in some detail since it imposes the most direct limitations on learning with subjective labels. Second, subjective labeling error increases the chance of suboptimal variable selection even under myopic evaluation criteria. For this issue, I will propose more of a thought experiment and follow up the analysis in later work. Third, if symmetry in error conditions are seriously violated then significant bias may be introduced into the learned function with particular unpredictability when the training cases are observed over high dimensional feature spaces. This final issue I simply raise as a possibility and defer detailed discussion entirely.

### 5.1 Bayes Error of Labeling and Prediction Error

The goal of learning a prediction function for the scores  $S$  (associated with the ordinal labels in  $L$ ) given data measured on a feature space  $X$  is to approximate  $E[S|X = x]$  with a function  $f(x)$  that is unbiased and efficient as possible. However, the best one can do in the predicting the observed scores  $\tilde{S}$  is to provide an efficient unbiased estimate  $E[\tilde{S}|X]$ .<sup>10</sup> After using the notational shortcut  $E_{S|X}[S|X = x] \equiv E[S|x]$  and  $V_{S|X}(S|X = x) \equiv V(S|x)$ , the expected pointwise prediction error  $PPE(x)$  of this estimate can be written as

$$\begin{aligned} E[PPE(x)] &= E \left[ (S^2 - 2SE[\tilde{S}|x] - E[\tilde{S}|x]^2) | x \right] \\ &= V(S|x) + E \left[ \left( E[S|x] - E[\tilde{S}|x] \right)^2 \right] \end{aligned}$$

now removing  $V(S|x)$  and focusing on the excess prediction error, over the variance of the true scores given  $x$ ,  $V_{S|X}(S|x)$ ,  $EPE(x)$  in the second part of 5.3 one gets

$$EPE(x) = E \left[ \left( \sum_{s_i \in S} p(s_i|x) \left\{ s_i - \sum_{\tilde{s}_j \in S} p(\tilde{s}_j|s_i) \tilde{s}_j \right\} \right)^2 \right]$$

Isolating the  $EPE(x)$  and identifying  $p(\tilde{s}_j|s_i)$  value as  $Y(i, j)$  for  $i$  and  $j$  indexed into  $S$ , one can demonstrate  $EPE(x)$  is a strictly increasing function of  $Y(i, j)$ ,  $j \neq i$ . Proceed by focusing on the subset of bayes error between any two scores  $s_i$  and  $s_j$ , let  $p_j = Y_{i,j}$ , use  $p(\tilde{s}_i|s_i) = Y_{i,i} = 1 - p_j$ , and let  $\alpha_i = p(s_i|x)$  for  $EPE_{i,j}(x)$  is

$$EPE_{i,j}(x) = E \left[ \left\{ \alpha_i(s_i + p_j(s_j - s_i)) + \alpha_j(s_j + p_i(s_i - s_j)) \right\}^2 \right]$$

which, after canceling cross terms  $\alpha_i \alpha_j p_i p_j ((s_i - s_j) + \alpha_j \alpha_i p_j p_i ((s_j - s_i))$  leaves only positive factors for the bayes error terms  $p_i$  and  $p_j$

$$EPE_{i,j}(x) = E \left[ \left\{ \alpha_j^2 p_j^2 (s_j - s_i)^2 + \alpha_i^2 p_i^2 (s_i - s_j)^2 \right\} \right]$$

making  $EPE_{i,j}$  strictly increasing in  $p_i$  and  $p_j$  for scores  $s_i \neq s_j$ .

<sup>10</sup>The regression function which minimizes squared error loss given  $x$ . For purposes of this paper I accept the formulation of the problem as a regression and acknowledge that problems may arise in implementation. This entire formulation of the problem follows the standard sorts of reasoning about optimal prediction functions explained in detail in [4].

## 6 Possible Remedies in Learning Prediction Functions

There are several immediate remedies that the practicing machine learning developer can try in the case of subjective labeling error. One of the clearest is to model the uncertainty in a likelihood setting and then estimate the model parameters. In the following, the assignment error information in  $Y$  is used to perform logistic regression in a multinomial classification directly on the labels.

### 6.1 Logistic Regression with Noisy Labels

For a case with feature vector  $X$ , denote the assigned score as  $S$  and the *true* label as  $L$ ; then consider the conditional probability  $P(S, L|X)$  and decompose it as

$$P(S, L|X) = P(S|L, X)P(L|X) = P(S|L)P(L|X),$$

where it's assumed that conditional on the true label, the assigned score is independent of the feature vector. Therefore, the conditional probability for the observed pair  $(X, S)$  is

$$P(S|X) = \sum_L P(S|L)P(L|X).$$

For logistic regression, the conditional probability for true label is given by [4],

$$P(L = j|X) = \frac{\exp(\beta_j^T X)}{1 + \sum_{i=1}^{K-1} \exp(\beta_i^T X)}$$

and

$$P(L = K|X) = \frac{1}{1 + \sum_{i=1}^{K-1} \exp(\beta_i^T X)}.$$

Now given a set of training pairs  $(x_i, s_i)$ ,  $i = 1, \dots, N$ , write the conditional likelihood as

$$\mathcal{L}(\{\beta_j\}_{j=1}^{K-1}, \{(x_i, s_i)\}_{i=1}^N) = \prod_{i=1}^N P(s_i|x_i) = \prod_{i=1}^N \left( \sum_{l_j=1}^K P(s_i|l_j)P(l_j|x_i) \right).$$

The parameters of the model can be estimated using variations of the EM algorithm [6].

In the above formulation, it's assumed that  $P(s_i|l_j)$  has been estimated from subjective scores against true labels.

### 6.2 Other Heuristics

Several simple heuristic remedies could be applied that might improve modeling performance. One method is to break up each case into a set of weighted sub-cases with response based on the true label  $l_i$  and observed label  $\tilde{l}_j$  with weight  $w_i = p(l_i|\tilde{l}_j)$  and  $\sum_i w_i = 1$ . For the sum of log-likelihood or sum of squares loss this reweighting is equivalent to substituting the expected objective (over the observed labels) for the true objective.

Adding an indicator variable to the features for each of the label assigners could improve model estimation. Since  $Y$  is the aggregate error in practice it's highly probable that individual biases vary on the part of the label assigners ([9]). With an indicator for each assigner the primary features might be less confounded by the effects of labeling error. Removing or clustering label classes with high cross error should result in a smaller number

of ordinal classes with lower total cross error. Of course, the conclusions and predictions must still be rich enough to enable the required decision making tasks.

At the design stage, one might build up an ordinal response that intrinsically has lower cross error. For example, eliciting independent votes from multiple observers (e.g. positive, *no opinion*, or negative) and counting positives minus negatives. Elicitation costs will likely be lower for this type of vote since there are fewer degrees of assignment and opt-out. Error is reduced since *no opinion* will catch some of the ambiguous, presumably high error, cases and independent votes geometrically reduces the probability of errors across multiple classes.

## 7 Conclusions and Further Work

Verifiably correct handling of subjective ordinal response data requires one to measure or account for assignment errors. This paper provides clear demonstrations of why one should characterize and adjust for assignment errors for any hypothesis testing or modeling tasks involving this type of data. Further work can build on this paper for an accessible introduction and motivation for the problem before going into detailed remedies for some of the issues discussed. Specifically, for discriminative modeling, the methods explored in [5] seem very promising for providing a comprehensive modeling solution for this problem. And finally, more work needs to be done to explore and demonstrate for practitioners the effects of assignment errors on variable selection and model estimation for popular off-the-shelf machine learning techniques such as SVM and boosted trees.

## References

- [1] V. Abraira and A. P. De Vargas. Generalization of the kappa coefficient for ordinal categorical data, multiple observers, and incomplete designs. *Questio* Vol. 23, No. 3, 1999, pp. 561-571
- [2] C. Ambroise, T. Denceux, G. Govaert, and P. Smets. Learning from an imprecise teacher: probabilistic and evidential approaches. *10th International Symposium on Applied Stochastic Models and Data Analysis*, Vol. 1, pp. 101-105, June 2001.
- [3] A.P. Dawid and A. M. Skene. Maximum Likelihood estimation of observer error rates using the *EM* algorithm. *Applied Statistics* 28:20-28
- [4] T. Hastie, R. Tibshirani, J. Friedman. *The Elements of Statistical Learning*. Springer, 2001.
- [5] R. Jin and Z. Ghahramani. Learning with Multiple Labels. The Sixteenth Annual Conference on Neural Information Processing Systems (NIPS 2002).
- [6] G.J. McLachlan and T. Krishnan. *The EM algorithm and Extensions*. Wiley, 1997.
- [7] G. Strang. *Linear Algebra and its Applications*. Harcourt, Brace, Jovanovich, 1988. Third Edition.
- [8] T. Swartz, Y. Haitovsky, A. Vexler, and T. Yang. Bayesian identifiability and misclassification in multinomial data *The Canadian Journal of Statistics* Vol. 32, No. 3, 2004, pp. 1-18
- [9] J. M. Wiebe, R. F. Bruce, and T. P. O'Hara. Development and Use of a Gold-Standard Data Set for Subjectivity Classifications. *Proc. 37th Annual Meeting of the Assoc. for Computational Linguistics (ACL-99)*, June 23-26, University of Maryland, pp. 246-253.