Maximum Likelihood Estimation of Observer Error-rates using the EM algorithm

Dawid et al. (1979)

A walk-through explanation

Please write to ahmed.y.allam@gmail.com for reporting any error/mistakes or suggestions for improving this write-up.

In this document, we present a walk-through explanation for the probabilistic modeling and parameters estimation procedure based on the work of Dawid et al. [1] following the same notation used in the original paper.

1 Problem formulation

Suppose we have I number of patients who visit K number of clinicians at different times (i.e. different weeks or different medical visits). Denote k to be an index for the clinicians where $k \in \mathbb{N}_K$ and $\mathbb{N}_K = \{1, 2, \ldots, K\}$. Similarly, denote i to be an index for the patients where $i \in \mathbb{N}_I$ and $\mathbb{N}_I = \{1, 2, \ldots, I\}$. Every clinician k can ask a single question to any patient i. Denote l to be the patient's response where l could take a value from a finite set of possible responses (i.e. $l \in \mathbb{N}_J$ and $\mathbb{N}_J = \{1, 2, \ldots, J\}$).
A clinician k can see and ask patient i multiple times (i.e. during different medical visits) and not all patients need to see all the clinicians. Therefore, for every patient i seen by clinician k, we have a response l that is recorded by the clinician k while j represents the true response of patient i.

Given this setting, the objective is to estimate the error-rates of clinician k. The error-rates are represented by the parameters π^k_{jl} describing the probability a clinician k will record response l given j is the true response.

2 Assumptions

We assume that the responses of a patient to the different clinicians are independent given the true response. That is the response given by patient i to clinician A is independent from the response given to clinician B, for every $A, B \in \mathbb{N}_K$ where $A \neq B$.
Moreover, if a patient visited clinician k multiple times (i.e. multiple medical visits), the responses given in each visit are independent from each other $\forall k \in \mathbb{N}_K$. In addition, there is no clinician by patient interaction.

3 Scenarios

3.1 Scenario 1: True responses are known

We denote $\{T_{ij} : j = 1, \cdots, J\}$to be a set of indicator variables for patient i. That is for every patient i we have a set $\{T_{i1}, T_{i2}, \cdots, T_{iJ}\}$indexed by j (i.e. to access its j^{th} element) having an entry equal to 1 when j is the true response and 0 elsewhere. For example if J were the true

response for patient i then the variable $T_{iJ} = 1$ and all other variables are 0 such that $\{T_{ij}\} = \{0, 0, \cdots, 1\}$.

3.1.1 Single patient and single clinician

If we consider a single patient i and one clinician k where $q \in \mathbb{N}_J$ is the true response (i.e. $T_{iq} = 1$ and $T_{ij} = 0 \, \forall j \neq q$), the counts of responses of each type ($\{1, \cdots, J\}$) recorded by clinician k would follow a multinomial distribution:

$$\frac{(\sum_{l=1}^{J} n_{il}^k)!}{\prod_{l=1}^{J} n_{il}^k!} \prod_{l=1}^{J} (\pi_{ql}^k)^{n_{il}^k}$$

In the original paper, the normalizing constant $\frac{(\sum_{l=1}^{J} n_{il}^k)!}{\prod_{l=1}^{J} n_{il}^k!}$ was omitted and the distribution was reported as proportional to: $\prod_{l=1}^{J} (\pi_{ql}^k)^{n_{il}^k}$

The term n_{il}^k represents the counts of responses of type/value l recorded by clinician k for patient i.

Hence, the expression $(\sum_{l=1}^{J} n_{il}^k)!$ represents the permutation of all responses recorded by clinician k for patient i. The multinomial distribution is a probability distribution on counts data (i.e. counts of all response types recorded by clinician k for patient i). However, if we have a sequence of responses received from patient i and recorded by clinician k where q is the true response, then the probability of the sequence given the above parametrization becomes:

$$p(responses|T_{iq} = 1) = \prod_{l=1}^{J} (\pi_{ql}^k)^{n_{il}^k} \tag{1}$$

3.1.2 Single patient and multiple clinicians

Given the independence assumptions (i.e. the recordings of responses by clinician A are independent of the ones recorded by clinician B given the true response $\forall A, B \in \mathbb{N}_K$ where $A \neq B$), the probability of the sequence of responses recorded for one patient i by multiple clinicians K given that q is the true response:

$$p(responses|T_{iq} = 1) = \prod_{k=1}^{K} \prod_{l=1}^{J} (\pi_{ql}^k)^{n_{il}^k} \tag{2}$$

It could be noted that we are always conditioning the probability by $T_{iq} = 1$ where q represents the true response. Therefore, the unconditional probability of a sequence of responses recorded for a

patient i becomes

$$p(responses) = \sum_{j=1}^{J} p(responses|T_{ij} = 1)p(T_{ij} = 1)$$
$$= p(responses|T_{i1} = 1)p(T_{i1} = 1) + \cdots + p(responses|T_{iJ} = 1)p(T_{iJ} = 1)$$

Given that for each patient the indicator variable T_{ij} is equal to 1 when j is the true response and 0 elsewhere, then the unconditional probability will always be of the form:
$p(responses) = p(responses|T_{iq} = 1)p(T_{iq} = 1)$ where $q \in \mathbb{N}_J$ is the true response (i.e. $T_{iq} = 1$ and $T_{ij} = 0 \, \forall j \neq q$).

The probability of the true response being equal to j for patient i is equivalent to p_j which is equal to counting the number of patients having a true response j divided by the total number of patients I. Hence, $p(T_{ij} = 1) = p_j$
As a result, we rewrite the unconditional probability:

$$p(responses; parameters) = \prod_{j=1}^{J} \left\{ p_j \prod_{k=1}^{K} \prod_{l=1}^{J} (\pi_{jl}^{k})^{n_{il}^{k}} \right\}^{T_{ij}} \tag{3}$$

The above probability function consists of J terms of which $J - 1$ terms are equal to 1 ($T_{ij} = 0, j \neq q$) and one term of the form $p(responses|T_{iq} = 1)p(T_{iq} = 1)$

3.1.3 Multiple patients and multiple clinicians

As the responses from all patients are considered to be independent, if we have a sequence of responses recorded by Kclinicians for I patients, the unconditional likelihood of the responses:

$$l(\pi^k, p|responses) = \prod_{i=1}^{I} \prod_{j=1}^{J} \left\{ p_j \prod_{k=1}^{K} \prod_{l=1}^{J} (\pi_{jl}^{k})^{n_{il}^{k}} \right\}^{T_{ij}} \tag{4}$$

The likelihood in (4) is a function of the parameters π^k (error-rate matrix of size $J \times J$ for every clinician k) and p (a vector of size $J \times 1$ representing the probability of the true response being equal to j)

3.1.4 Parameters estimation:

In this current setting, the set of unknown parameters to be estimated are p and π^k. Changing the values of the parameters would result into different values of the likelihood function measuring how well the chosen values of the parameters fits the data/responses we have.

To estimate these parameters, we use maximum likelihood induction principle. In other words, we need to find the value of the parameters which maximizes the likelihood function in equation (4). These estimates are called the maximum likelihood estimates (MLE).

Since the natural log is strictly increasing monotonic function, maximizing the likelihood function is equivalent to maximizing the log-likelihood function. Therefore, we maximize the following log-likelihood function:

$$L(\pi^k, p|responses) = \sum_{i=1}^{I}\sum_{j=1}^{J} T_{ij}\left[\log p_j + \sum_{k=1}^{J}\sum_{l=1}^{J} n_{il}^k \log \pi_{jl}^k\right] \tag{5}$$

Moreover, there are two constraints on the parameters of the likelihood function:

$\sum_{l=1}^{J} \pi_{jl}^k = 1$, $\forall j \in \mathbb{N}_J$ and $\forall k \in \mathbb{N}_K$

$\sum_{j=1}^{J} p_j = 1$

As a result, we have a constrained optimization problem that can be solved using Lagrange multipliers. By taking the derivative of the log-likelihood function with respect to the parameters and set it equal to zero, we aim at finding the estimates of the parameters that maximize the log-likelihood function.

3.1.4.1 MLE estimates for p_j:

$$\frac{\partial L(\pi^k, p, \lambda|responses)}{\partial p_j} = \frac{\partial\left[\sum_{i=1}^{I}\sum_{j=1}^{J} T_{ij}\log p_j + \lambda(\sum_{j=1}^{J} p_j - 1)\right]}{\partial p_j} \tag{6}$$

The terms related to π^k in the log-likelihood function were omitted from equation (6) since they will be equal to 0 when taking the derivative with respect to p_j. Notice also the added constraint to the log-likelihood function with the Lagrangian multiplier λ.

$$\begin{aligned}\frac{\partial L(\pi^k, p, \lambda|responses)}{\partial p_j} &= \frac{\partial}{\partial p_j}\sum_{i=1}^{I} T_{ij}\log p_j + \frac{\partial}{\partial p_j}\lambda(\sum_{j=1}^{J} p_j - 1) \\ &= \frac{\sum_{i=1}^{I} T_{ij}}{p_j} + \lambda \end{aligned} \tag{7}$$

Another way for deriving the result in (7), is to expand the log-likelihood function in equation (6) in terms of its J parameters. We denote the expression we want to expand by:

$$f(p, \lambda) = \sum_{i=1}^{I}\sum_{j=1}^{J} T_{ij}\log p_j + \lambda(\sum_{j=1}^{J} p_j - 1)$$

$$f(p,\lambda) = \sum_i T_{i1} \log p_1 + \sum_i T_{i2} \log p_2 + \cdots + \sum_i T_{iJ} \log p_J + \lambda p_1 + \lambda p_2 + \cdots \lambda p_J - \lambda$$

$$f(p,\lambda) = \begin{bmatrix} \sum_i T_{i1} & \sum_i T_{i2} & \dots & \sum_i T_{iJ} \end{bmatrix} \begin{bmatrix} \log p_1 \\ \log p_2 \\ \vdots \\ \log p_J \end{bmatrix} + \lambda \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_J \end{bmatrix} \text{ - } \lambda$$

$$\frac{\partial f(p,\lambda)}{\partial p_j} = \begin{bmatrix} \frac{\partial f(p,\lambda)}{\partial p_1} \\ \frac{\partial f(p,\lambda)}{\partial p_2} \\ \vdots \\ \frac{\partial f(p,\lambda)}{\partial p_J} \end{bmatrix} = \begin{bmatrix} \frac{\sum_i T_{i1}}{p_1} + \lambda \\ \frac{\sum_i T_{i2}}{p_2} + \lambda \\ \vdots \\ \frac{\sum_i T_{iJ}}{p_J} + \lambda \end{bmatrix}$$

Given we have J response options then p represents a vector of J parameters indexed by j. More precisely, it is $J-1$ parameters where the last parameter $p_J = 1 - \sum_{j=1}^{J-1} p_j$ (this is a direct result from the constraint $\sum_{i=1}^{J} p_j = 1$).

To find the parameter p_j maximizing the log-likelihood function, we equate the derivative of the log-likelihood function with respect to p_j to 0.

$$\frac{\partial L(p,\lambda|responses)}{\partial p_j} = 0 \Rightarrow \frac{\sum_{i=1}^{I} T_{ij}}{p_j} + \lambda = 0$$

$$p_j = -\frac{\sum_{i=1}^{I} T_{ij}}{\lambda}$$

given the constraint $\sum_{j=1}^{J} p_j = 1$ then $\sum_{j=1}^{J} \left(\frac{-\sum_{i=1}^{I} T_{ij}}{\lambda} \right) = 1$

and hence $\lambda = -\sum_{i=1}^{I} \sum_{j=1}^{J} T_{ij} = -I$

This is because T_{ij} is an indicator variable for the true label/response for each patient (i.e. 1 for the true response and 0 elsewhere). The sum for all true labels for all patients will be equal to the number of patients I in the sample.

Therefore $\hat{p}_j = \frac{-\sum_{i=1}^{I} T_{ij}}{-I} = \frac{\sum_{i=1}^{I} T_{ij}}{I}$ which is equivalent to counting the number of patients

having true response j and dividing it by the total number of patients (i.e. prevalence of true response j in the sample of I patients we are working with).

3.1.4.2 MLE estimates of π_{jl}^k :

$$\frac{\partial L(\pi^k, p, \lambda | responses)}{\partial \pi_{jl}^k} = \frac{\partial \left[\sum_{i=1}^{I} \sum_{j=1}^{J} T_{ij} \left(\sum_{k=1}^{K} \sum_{l=1}^{J} n_{il}^k \log \pi_{jl}^k \right) + \sum_{k=1}^{K} \sum_{j=1}^{J} \lambda_j^k (\sum_{l=1}^{J} \pi_{jl}^k - 1) \right]}{\partial \pi_{jl}^k} \quad (8)$$

The terms related to p in the log-likelihood function were omitted from equation (8) since they will be equal to 0 when taking the derivative with respect to π_{jl}^k. Notice also the added constraint to the log-likelihood function with the Lagrangian multiplier λ_j^k.

$$\begin{aligned} \frac{\partial L(\pi^k, p, \lambda | responses)}{\partial \pi_{jl}^k} &= \frac{\partial}{\pi_{jl}^k} \sum_{i=1}^{I} T_{ij} n_{il}^k \log \pi_{jl}^k + \frac{\partial}{\pi_{jl}^k} \lambda_j^k (\sum_{l=1}^{J} \pi_{jl}^k - 1) \\ &= \frac{\sum_{i=1}^{I} T_{ij} n_{il}^k}{\pi_{jl}^k} + \lambda_j^k \end{aligned} \quad (9)$$

Again, to derive the expression in equation (9), it would be helpful to expand the log-likelihood function in equation (8) in terms of its parameters. We denote the expression we want to expand by:

$$\begin{aligned} g(\pi^k, \lambda) = &\sum_i T_{i1} \left(\sum_k n_{i1}^k \log \pi_{11}^k + \sum_k n_{i2}^k \log \pi_{12}^k + \cdots + \sum_k n_{iJ}^k \log \pi_{1J}^k \right) + \\ &\sum_i T_{i2} \left(\sum_k n_{i1}^k \log \pi_{21}^k + \sum_k n_{i2}^k \log \pi_{22}^k + \cdots + \sum_k n_{iJ}^k \log \pi_{2J}^k \right) + \\ &\vdots \\ &\vdots \\ &\sum_i T_{iJ} \left(\sum_k n_{i1}^k \log \pi_{J1}^k + \sum_k n_{i2}^k \log \pi_{J2}^k + \cdots + \sum_k n_{iJ}^k \log \pi_{JJ}^k \right) + \\ &\sum_{k=1}^{K} \lambda_1^k \left(\pi_{11}^k + \pi_{12}^k + \cdots + \pi_{1J}^k - 1 \right) + \\ &\sum_{k=1}^{K} \lambda_2^k \left(\pi_{21}^k + \pi_{22}^k + \cdots + \pi_{2J}^k - 1 \right) + \end{aligned}$$

$$\vdots$$
$$\vdots$$
$$\sum_{k=1}^{K} \lambda_J^k \left(\pi_{J1}^k + \pi_{J2}^k + \cdots + \pi_{JJ}^k - 1\right)$$

$$= \sum_i T_{i1} \left(\begin{bmatrix} n_{i1}^1 & n_{i1}^2 & \cdots & n_{i1}^K \end{bmatrix} \begin{bmatrix} \log \pi_{11}^1 \\ \log \pi_{11}^2 \\ \vdots \\ \log \pi_{11}^K \end{bmatrix} \right) + \begin{bmatrix} \lambda_1^1 & \lambda_1^2 & \cdots & \lambda_1^K \end{bmatrix} \begin{bmatrix} \pi_{11}^1 \\ \pi_{11}^2 \\ \vdots \\ \pi_{11}^K \end{bmatrix} +$$

$$\sum_i T_{i1} \left(\begin{bmatrix} n_{i2}^1 & n_{i2}^2 & \cdots & n_{i2}^K \end{bmatrix} \begin{bmatrix} \log \pi_{12}^1 \\ \log \pi_{12}^2 \\ \vdots \\ \log \pi_{12}^K \end{bmatrix} \right) + \begin{bmatrix} \lambda_1^1 & \lambda_1^2 & \cdots & \lambda_1^K \end{bmatrix} \begin{bmatrix} \pi_{12}^1 \\ \pi_{12}^2 \\ \vdots \\ \pi_{12}^K \end{bmatrix} +$$

$$\vdots$$

$$\sum_i T_{i1} \left(\begin{bmatrix} n_{iJ}^1 & n_{iJ}^2 & \cdots & n_{iJ}^K \end{bmatrix} \begin{bmatrix} \log \pi_{1J}^1 \\ \log \pi_{1J}^2 \\ \vdots \\ \log \pi_{1J}^K \end{bmatrix} \right) + \begin{bmatrix} \lambda_1^1 & \lambda_1^2 & \cdots & \lambda_1^K \end{bmatrix} \begin{bmatrix} \pi_{1J}^1 \\ \pi_{1J}^2 \\ \vdots \\ \pi_{1J}^K \end{bmatrix} - \begin{bmatrix} \lambda_1^1 \\ \lambda_1^2 \\ \vdots \\ \lambda_1^K \end{bmatrix} +$$

$$\sum_i T_{i2} \left(\begin{bmatrix} n_{i1}^1 & n_{i1}^2 & \cdots & n_{i1}^K \end{bmatrix} \begin{bmatrix} \log \pi_{21}^1 \\ \log \pi_{21}^2 \\ \vdots \\ \log \pi_{21}^K \end{bmatrix} \right) + \begin{bmatrix} \lambda_2^1 & \lambda_2^2 & \cdots & \lambda_2^K \end{bmatrix} \begin{bmatrix} \pi_{21}^1 \\ \pi_{21}^2 \\ \vdots \\ \pi_{21}^K \end{bmatrix} +$$

$$\sum_i T_{i2} \left(\begin{bmatrix} n_{i2}^1 & n_{i2}^2 & \cdots & n_{i2}^K \end{bmatrix} \begin{bmatrix} \log \pi_{22}^1 \\ \log \pi_{22}^2 \\ \vdots \\ \log \pi_{22}^K \end{bmatrix} \right) + \begin{bmatrix} \lambda_2^1 & \lambda_2^2 & \cdots & \lambda_2^K \end{bmatrix} \begin{bmatrix} \pi_{22}^1 \\ \pi_{22}^2 \\ \vdots \\ \pi_{22}^K \end{bmatrix} +$$

$$\vdots$$

$$\sum_i T_{i2} \left(\begin{bmatrix} n_{iJ}^1 & n_{iJ}^2 & \cdots & n_{iJ}^K \end{bmatrix} \begin{bmatrix} \log \pi_{2J}^1 \\ \log \pi_{2J}^2 \\ \vdots \\ \log \pi_{2J}^K \end{bmatrix} \right) + \begin{bmatrix} \lambda_2^1 & \lambda_2^2 & \cdots & \lambda_2^K \end{bmatrix} \begin{bmatrix} \pi_{2J}^1 \\ \pi_{2J}^2 \\ \vdots \\ \pi_{2J}^K \end{bmatrix} - \begin{bmatrix} \lambda_2^1 \\ \lambda_2^2 \\ \vdots \\ \lambda_2^K \end{bmatrix}$$

$$\vdots$$
$$\vdots$$

$$\sum_i T_{iJ} \left(\begin{bmatrix} n_{i1}^1 & n_{i1}^2 & \cdots & n_{i1}^K \end{bmatrix} \begin{bmatrix} \log \pi_{J1}^1 \\ \log \pi_{J1}^2 \\ \vdots \\ \log \pi_{J1}^K \end{bmatrix} \right) + \begin{bmatrix} \lambda_J^1 & \lambda_J^2 & \cdots & \lambda_J^K \end{bmatrix} \begin{bmatrix} \pi_{J1}^1 \\ \pi_{J1}^2 \\ \vdots \\ \pi_{J1}^K \end{bmatrix} +$$

$$\sum_i T_{iJ} \left(\begin{bmatrix} n_{i2}^1 & n_{i2}^2 & \cdots & n_{i2}^K \end{bmatrix} \begin{bmatrix} \log \pi_{J2}^1 \\ \log \pi_{J2}^2 \\ \vdots \\ \log \pi_{J2}^K \end{bmatrix} \right) + \begin{bmatrix} \lambda_J^1 & \lambda_J^2 & \cdots & \lambda_J^K \end{bmatrix} \begin{bmatrix} \pi_{J2}^1 \\ \pi_{J2}^2 \\ \vdots \\ \pi_{J2}^K \end{bmatrix} +$$

$$\vdots$$

$$\sum_i T_{iJ} \left(\begin{bmatrix} n_{iJ}^1 & n_{iJ}^2 & \cdots & n_{iJ}^K \end{bmatrix} \begin{bmatrix} \log \pi_{JJ}^1 \\ \log \pi_{JJ}^2 \\ \vdots \\ \log \pi_{JJ}^K \end{bmatrix} \right) + \begin{bmatrix} \lambda_J^1 & \lambda_J^2 & \cdots & \lambda_J^K \end{bmatrix} \begin{bmatrix} \pi_{JJ}^1 \\ \pi_{JJ}^2 \\ \vdots \\ \pi_{JJ}^K \end{bmatrix} - \begin{bmatrix} \lambda_J^1 \\ \lambda_J^2 \\ \vdots \\ \lambda_J^K \end{bmatrix}$$

For every clinician we have an error-rate parameters π^k that represents a matrix of size $J \times J$ where for every true response $j \in \mathbb{N}_J$ we have $J-1$ parameters to estimate (this is a direct result from the constraint $\sum_{l=1}^{J} \pi_{jl}^k = 1$ where the last parameter is estimated by $\pi_{jJ}^k = 1 - \sum_{l=1}^{J-1} \pi_{jl}^k$)
Therefore, the total number of parameters of π_{jl}^k for all clinicians is $J(J-1)K$ where $K$is the total number of clinicians and J is the total number of allowed response types/categories.

For example, to find the parameter estimate of π_{11}^1that is the probability of clinician 1 recording a response equal to 1 given the true response is equal to 1, we differentiate $g(\pi^k, \lambda)$ with respect to π_{11}^1 and equate to 0.

Hence, $\dfrac{\partial g(\pi^k, \lambda)}{\partial \pi_{11}^1} = \dfrac{\sum_{i=1}^{I} T_{i1} n_{i1}^1}{\pi_{11}^1} + \lambda_1^1 = 0$

$$\pi_{11}^1 = \frac{-\sum_{i=1}^{I} T_{i1} n_{i1}^1}{\lambda_1^1}$$

Similarly, we estimate the parameters π_{11}^k for every clinician k:

$$\pi_{11}^2 = \frac{-\sum_{i=1}^{I} T_{i1} n_{i1}^2}{\lambda_1^2}$$

$$\pi_{11}^3 = \frac{-\sum_{i=1}^{I} T_{i1} n_{i1}^3}{\lambda_1^3}$$

$$\vdots$$

$$\pi_{11}^K = \frac{-\sum_{i=1}^{I} T_{i1} n_{i1}^K}{\lambda_1^K}$$

This also applies for estimating the parameters π_{1l}^k that is probability of a clinician k recording a response $l \in \mathbb{N}_J$ given the true response is 1:

$$\pi_{1l}^1 = \frac{-\sum_{i=1}^{I} T_{i1} n_{il}^1}{\lambda_1^1}$$

$$\pi_{1l}^2 = \frac{-\sum_{i=1}^{I} T_{i1} n_{il}^2}{\lambda_1^2}$$

$$\vdots$$

$$\pi_{1l}^K = \frac{-\sum_{i=1}^{I} T_{i1} n_{il}^K}{\lambda_1^K}$$

More generally, to estimate the parameters π_{jl}^k that is the probability of a clinician k recording a response $l \in \mathbb{N}_J$ given the true response is $j \in \mathbb{N}_J$:

$$\frac{\partial g(\pi^k, \lambda)}{\partial \pi_{jl}^k} = 0 \Rightarrow \frac{\sum_{i=1}^{I} T_{ij} n_{il}^k}{\pi_{jl}^k} + \lambda_j^k = 0$$

$$\pi_{jl}^k = \frac{-\sum_{i=1}^{I} T_{ij} n_{il}^k}{\lambda_j^k}$$

To solve for λ_j^k we substitute for π_{jl}^k in the constraint $\sum_{l=1}^{J} \pi_{jl}^k = 1$ to get:

$$\sum_{l=1}^{J} \left(\frac{-\sum_{i=1}^{I} T_{ij} n_{il}^k}{\lambda_j^k} \right) = 1 \Rightarrow \lambda_j^k = -\sum_{l=1}^{J} \sum_{i=1}^{I} T_{ij} n_{il}^k$$

Therefore $\hat{\pi}_{jl}^k = \frac{-\sum_{i=1}^{I} T_{ij} n_{il}^k}{-\sum_{l=1}^{J} \sum_{i=1}^{I} T_{ij} n_{il}^k} = \frac{\sum_{i=1}^{I} T_{ij} n_{il}^k}{\sum_{l=1}^{J} \sum_{i=1}^{I} T_{ij} n_{il}^k}$

As a result, the MLE estimates of the error-rates parameters π_{jl}^k is basically counting the number of

times response $l \in \mathbb{N}_J$ is recorded by clinician k divided by the total number of responses recorded by the same clinician k given the true response $j \in \mathbb{N}_J$ is known. In other words, we divide our sample of patients into subsamples based on their true response j. Then we count for each subsample separately the number of times a clinician k records a response $l \in \mathbb{N}_J$ and we divide it by the total number of responses recorded by the same clinician for the same subsample (normalization step). As we already know the true response for each subsample (i.e. all patients have the same true response j), the estimation of π_{jl}^k is equivalent to finding a probability distribution on all responses recorded by clinician k for every known true response j.

3.2 Scenario 2: True responses are unknown

We started the problem by including in our definition our knowledge of patient's indicator variable T_{ij} indicating the true response for every patient $i \in \mathbb{N}_I$. However, in this scenario we consider this variable to be unknown. In other words, in this current formulation, the set of indicator variables $\{T_{ij}\}$is not anymore a set of binary variables where only the true response has value 1 and the rest is 0. Instead, the set of indicator variables will define a probability distribution on its J

elements/variables that sum to 1 (i.e. $\sum_{j=1}^{J} p(T_{ij} = 1) = 1$).

One way to view the modification of the set of indicator variables between the two scenarios is that in the first, the patient's true response is determined by being exclusively one of the possible J responses (i.e. “hard assignment”) whereas in the second scenario, the true response of a patient does not exclusively belong to only one response rather it is defined by a probability distribution on the J possible responses it might be assigned to (i.e. “soft assignment”).

Therefore, to estimate the probability distribution on the J possible responses (i.e. $p(T_{ij} = 1),\ \forall j \in \mathbb{N}_J$) for every patient i, we use Bayes theorem. Given the responses recorded by the Kclinicians for patient i, the probability of the true response for patient i being j is:

$$p(T_{ij} = 1|responses) = \frac{p(responses|T_{ij} = 1)p(T_{ij} = 1)}{p(responses)} \tag{10}$$

$$\propto p(responses|T_{ij} = 1)p(T_{ij} = 1)$$

$$\propto \prod_{k=1}^{K}\prod_{l=1}^{J}(\pi_{jl}^k)^{n_{il}^k} p_j$$

Two things to notice, the first is the proportional sign and the second is the value of the numerator

in equation (10). For the latter, the expression evaluated in the numerator is a result we already established earlier in equation (2) (the case of the probability of the responses recorded by multiple clinicians for a single patient given the true response value is known) multiplied by the marginal probability of the true response p_j (i.e. the prevalence of true response j in the sample of I patients).

As the set of indicator variables $\{T_{ij}\}$ for patient i represents a distribution on the Jpossible values a true response could be equal to, we could compute $p(T_{ij} = 1|responses)$for every $j \in \mathbb{N}_J$ without the normalizing constant $p(responses)$that is absorbed in the proportional sign. After computing $p(T_{ij} = 1|responses)$ for all j values, we divide each by the sum of all computed probabilities (normalization step). The denominator would be the $p(responses)$for patient i representing the normalization constant and is equivalent to $p(responses) = \sum_{q=1}^{J} \prod_{k=1}^{K} \prod_{l=1}^{J} (\pi_{ql}^{k})^{n_{il}^{k}} p_q$

Therefore, given a sequence of responses recorded by Kclinicians for single patient i, the likelihood function becomes

$$\sum_{q=1}^{J} p_q \prod_{k=1}^{K} \prod_{l=1}^{J} (\pi_{ql}^{k})^{n_{il}^{k}}$$

which is equivalent to the $p(responses)$ on patient i. By the independence assumption of the responses given by the different patients, the full likelihood function for all I patients is equivalent to

$$l(\pi_{jl}^{k}, p_j|responses) = \prod_{i=1}^{I} \left(\sum_{q=1}^{J} p_q \prod_{k=1}^{K} \prod_{l=1}^{J} (\pi_{ql}^{k})^{n_{il}^{k}} \right) \tag{11}$$

It is worth noting the difference in likelihood function in the case of the true responses are known (equation 4) and in the case when they are not (equation 11). The first is a multinomial distribution of the responses recorded from patients by the clinicians. While the second is a mixture of multinomial distributions weighted by the marginal probability of the true response p_j where $j \in \mathbb{N}_J$ (i.e. prevalence of true responses in the data/responses recorded by the clinicians from I patients).

3.2.1 Parameters estimation

In this current setting, the set of unknown parameters are p_j, π_{jl}^{k}, where $\{T_{ij}\}$ represents a set of "missing/hidden" variables.

Maximizing the log-likelihood function as before to get estimates of p_j, π_{jl}^k is a difficult task. However, the problem could be solved easily when assuming $\{T_{ij}\}$ variables are known. That is we go back to the first scenario where we could obtain estimates of p_j, π_{jl}^k by maximizing the log-likelihood function. Knowing these parameter estimates will allow us updating the probability of each indicator variable (i.e. $p(T_{ij} = 1)\ \forall j \in \mathbb{N}_J$) using Bayes theorem as in equation (10).

This scenario of dependency between the missing data variables and the parameters of the likelihood function is well-suited for the application of the expectation-maximization algorithm (EM) described in Dempster et al. [2].

3.2.2 Sketch of EM procedure

1- Initialize randomly the values of the indicator variables $T_{ij}\ \forall j \in \mathbb{N}_J$ and $\forall i \in \mathbb{N}_I$

$$T_{ij}^t = p(T_{ij} = 1)^t = random(0, 1) \text{ such that } \sum_{j=1}^{J} p(T_{ij} = 1)^t = 1$$

2- Use MLE estimation of the parameters π_{jl}^k and p_j using the generated values of T_{ij}^t in step (1) at iteration t

$$(\hat{\pi}_{jl}^k)^t = \frac{\sum_{i=1}^{I} p(T_{ij} = 1)^t n_{il}^k}{\sum_{l=1}^{J}\sum_{i=1}^{I} p(T_{ij} = 1)^t n_{il}^k}$$

$$(\hat{p}_j)^t = \frac{\sum_{i=1}^{I} p(T_{ij} = 1)^t}{\sum_{i=1}^{I}\sum_{j=1}^{J} p(T_{ij} = 1)^t}$$

3- Use the estimated parameters at iteration t in step (2) to update the values of the indicator variables $T_{ij}\ \forall j \in \mathbb{N}_J$ and $\forall i \in \mathbb{N}_I$

$$p(T_{ij} = 1)^{t+1} = \frac{\prod_{k=1}^{K}\prod_{l=1}^{J}((\pi_{jl}^k)^t)^{n_{il}^k}(p_j)^t}{\sum_{q=1}^{J}\prod_{k=1}^{K}\prod_{l=1}^{J}((\pi_{ql}^k)^t)^{n_{il}^k}(p_q)^t}$$

4- Loop between step (2) and (3) until the results converge. The results converge means either the absolute difference in the estimate of the parameters $\hat{\pi}_{jl}^k$ and $\hat{p}_j$ between iteration t and iteration $t+1$ is less than a threshold ϵ. Or the absolute difference in the likelihood of the responses/data given the parameters defined in equation (11) between iteration t and iteration $t+1$ is less than a

threshold ϵ. ϵ could be a small number such as 10^{-4} or 10^{-6}

The EM algorithm would run for multiple initializations for step (1) and continues until the convergence of the estimates of the parameters. To identify the best parameter estimates from all runs, we choose the ones that maximizes the full likelihood of the responses expressed in equation (11). Moreover, one possible initialization of the indicator variables in step (1) is to use the data representing the recorded responses by the K clinicians for every patient i such that

$$\hat{T}_{ij} = \frac{\sum_{k=1}^{K} n_{ij}^{k}}{\sum_{k=1}^{K}\sum_{l=1}^{J} n_{il}^{k}}$$

This is equivalent to counting separately the number of times each response j was recorded by the Kclinicians on patient i and divide each by the total number of responses recorded by the K clinicians for the same patient to normalize.

Another note about the parameter estimates $\hat{p}_j$ $\hat{\pi}_{jl}^{k}$ when using the EM procedure is that they are weighted version of the MLE estimates in the first scenario where the weights represent the probability of the true response j for every patient i.

4 Concluding remarks

The application of this modeling procedure extends beyond the presented application setting (i.e. clinicians and patients). This approach could be applied in many other settings where there are multiple raters who are coding/rating set of objects based on a finite set of options. By using this probabilistic modeling approach, multiple ratings for the same object could be taken into account to produce a “consensus rating” that performs much better than the majority voting strategy.
In such setting, the raters will play the role of the clinicians and the rated objects will play the role of the patients. Indeed, this proposed procedure by Dawid et al. [1] was applied for detecting small volcanoes on the planet Venus using the scientists' multiple labelings of the Magellan image database [3]. Moreover, the work of Dawid et al. [1] could be considered as one of the original probabilistic consensus models that inspired many subsequent modeling approaches aiming at finding consensus rating in a situation when multiple ratings are provided as in [4-6].

5 References

1. Dawid AP, Skene AM. Maximum likelihood estimation of observer error-rates using the EM algorithm. J R Stat Soc Ser C Appl Stat. 1979;28(1):20–8.

2. Dempster A, Laird N, Rubin D. Maximum Likelihood from Incomplete Data via the EM Algorithm. J R Stat Soc Ser B. 1977;39(1):1–38. Available from: http://www.jstor.org/stable/10.2307/2984875

3. Smyth P, Fayyad U, Burl M, Perona P, Baldi P. Inferring Ground Truth from Subjective Labelling of Venus Images [Internet]. The MIT Press; 1995. Available from: http://authors.library.caltech.edu/55562/1/949-inferring-ground-truth-from-subjective-labelling-of-venus-images.pdf

4. Raykar VC, Yu S, Zhao LH, Jerebko A, Florin C, Valadez GH, et al. Supervised learning from multiple experts. Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09. New York, New York, USA: ACM Press; 2009. p. 1–8. Available from: http://dl.acm.org/citation.cfm?id=1553374.1553488

5. Raykar VC, Yu S, Zhao LH, Valadez GH, Florin C, Bogoni L, et al. Learning From Crowds. J Mach Learn Res; 2010 Mar 1;11:1297–322. Available from: http://dl.acm.org/citation.cfm?id=1756006.1859894

6. Hovy D, Berg-Kirkpatrick T, Vaswani A, Hovy E. Learning Whom to Trust with MACE. NAACL-HLT '13. 2013. p. 1120–30. Available from: http://www.aclweb.org/anthology/N13-1132