

EMPIRICAL BAYES APPROACH TO TRUTH DISCOVERY

Tsviel Ben Shabat, Reshef Meir, David Azriel
Technion - Israel Institute of Technology

Truth Discovery

In the early 20-th century, the 84 year old Sir Francis Galton [1907], had 787 guesses by different people of an ox's weight, he noticed that the median guess was surprisingly close to the truth. Sir Galton's discovery is what is known today as the *wisdom of the crowds*.



Imagine n workers (people/algorithms) who answer m questions (guessing the weights on m oxen), under the additive white Gaussian (AWG) noise model, the guess of the i -th worker to the j -th question is denoted by $X_{ij} \sim \mathcal{N}(\mu_j, \sigma_i^2)$. We call $1/\sigma_i^2$ the *competence* of the i -th worker. We would like to *discover* (estimate) the *truth* (the true oxen's weights) $\mu = (\mu_1, \dots, \mu_m)$. Can we do a better aggregation than the median guess?

Truth Discovery Algorithms

Usually, truth discovery algorithms aim at estimating workers' *competences* and apply a weighted mean i.e., the estimated truth of the j -th question would be $\sum_{i=1}^n w_i x_{ij}$

Meir et al. [2021] show that the average pairwise distance between a worker to the other workers' answers can estimate her competence. Denote the answer of the i -th person as \bar{x}_i and d_{ik} as the distance metric between the i -th and the k -th person then

$$w_i \propto \frac{1}{n-1} \sum_{k \neq i} d_{ik}$$

Li et al. [2014a] weigh workers' answers proportionally to the upper confidence interval limit of their estimated competence, denote the mean answer to the j -th question as \bar{x}_j then,

$$w_i \propto \frac{\mathcal{X}^2(\frac{\alpha}{2}, m)}{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}$$

where $\mathcal{X}^2(\frac{\alpha}{2}, m)$ is the appropriate $\frac{\alpha}{2}$ quartile of an m -degree chi-square distribution

To compare between algorithms or statistical estimators we use the square euclidean loss denoted by $\mathcal{L}(\mu, \hat{\mu}) = \|\mu - \hat{\mu}\|^2$, since estimators are random variables we compare the expected loss $E[\mathcal{L}(\mu, \hat{\mu})]$.

Statistical Estimators

First we introduce the *best linear unbiased estimator* for the case where multiple workers answer a single question and thus, can be applied to each question separately

$$A_B^{\sigma}(\vec{X}_j) := \left(\sum_{i=1}^n \frac{1}{\sigma_i^2} \right)^{-1} \sum_{i=1}^n \frac{X_{ij}}{\sigma_i^2}$$

Theorem - [Aitkin, 1934] The inverse variance weighting of the observations, is the best linear unbiased estimator (BLUE) for μ under the square loss function and the AWG model.

The *Empirical Bayes Estimator (EBE)* is for the case of a single worker answering multiple questions Denote $\vec{X} = (X_1, \dots, X_m)$, $\bar{X} = \frac{1}{m} \sum_{i=1}^m X_j$

$$\phi_{EB}(\vec{X}, \sigma) := \bar{X} \bar{1} + \left[1 - \frac{(m-3)\sigma^2}{\|\vec{X} - \bar{X} \bar{1}\|^2} \right] (\vec{X} - \bar{X} \bar{1})$$

Theorem - [Lehman and Casella, 1998] $E[L(\phi_{EB}(\vec{X}, \sigma), \mu)] < E[L(\vec{X}, \mu)] \forall \mu \in \mathbb{R}^m$ under the square loss and the AWG model

The above result is surprising because even with a single worker answering multiple questions, a simple manipulation of her answers outputs a better result.

Conclusion - Utilizing the square Euclidean loss metric, Truth Discovery algorithms serve as an approximation to the Best Linear Unbiased Estimator (BLUE) for addressing the challenge of multiple workers providing responses to numerous queries. In contrast, Lehman and Casella present a solution pertaining to a single worker responding to multiple inquiries. This study elucidates the possibility of integrating both approaches under specific mathematical constraints.

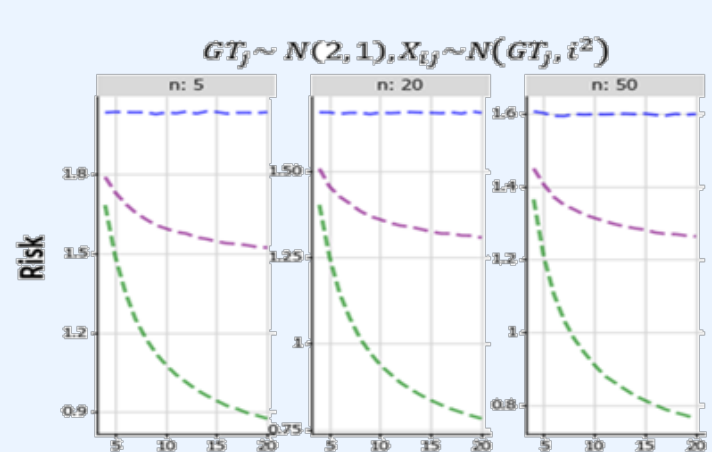
Known Competences

When we assume that workers' competences are known, we can use both of the aforementioned statistical estimators together. We justify it by proving that when applying *BLUE* we do not lose any information for estimating the truth thus, $A_B^{\sigma}(\mathbf{X})$ can be seen as reducing the problem to a single worker answer multiple questions.

Proposition - $A_B^{\sigma}(\mathbf{X})$ is a sufficient statistic for $\vec{\mu} = (\mu_1, \dots, \mu_m)$ under the AWG model.

Applying Lehman and Casella [1998] to the above proposition results in the following corollary

Corollary - Denote the loss function $\mathcal{L}(\mu, \hat{\mu}) = \|\mu - \hat{\mu}\|^2$, then $E[L(\phi_{EB}(A_B^{\sigma}(\mathbf{X}), \sigma), \mu)] < E[L(A_B^{\sigma}(\vec{X}_j), \mu)] \forall \mu \in \mathbb{R}^m$



Algorithm 1: EBLUE^σ for Known Competence
Input: Dataset $\mathbf{X} \in \mathbb{R}^{n \times m}$, variances $\vec{\sigma} \in \mathbb{R}_+^n$
 $\vec{X}^B \leftarrow A_B^{\sigma}(\mathbf{X})$;
 $\hat{\sigma}^2 \leftarrow \left(\sum_{i=1}^n \frac{1}{\sigma_i^2} \right)^{-1}$;
return $\phi_{EB}(\vec{X}^B, \hat{\sigma})$;

Applying the above corollary, we formulated algorithm 1 in the case of known competences. We first apply BLUE, thus reducing the problem to a single worker answering multiple questions. To demonstrate EBLUE^σ, we generated an estimation problem whose truths are normally distributed and workers' answers follow the AWG model, we then calculated the loss of the different estimators. We repeated these steps 100,000 time and estimated the expected loss (the risk). We compared BLUE estimator to EBLUE^σ (EBE) and it's root - the James–Stein estimator (Stein). We repeated the process for different numbers of workers (n) and truths (m) to produce the above graphs. As expected, since BLUE is applied to every question separately, it does not improve as the number of questions grow while both EBE and Stien significantly improve.

Estimated Competences

Truth discovery algorithms typically estimate the truth by estimating workers' competence and then aggregate answers, weighing them accordingly. In general we may not know the distribution of the aggregated answer, either since the initial observations depart from the AWG model, or because the algorithm A is complicated or unknown. We thus relax any assumption on the input in this section, except that the output of some algorithm A is *unbiased*. Thus $\vec{\mu} = E[\vec{X}^A]$. We also denote the true (unknown) variance by $\sigma^2 := \text{Var}[\vec{X}_j^A]$ (identical for all j). Similarly to the Known Competences case, we consider the aggregated answer vector, however rather than A_B^{σ} (which requires the actual workers' variances), we now assume an arbitrary truth discovery algorithm $A : \mathbb{R}^{n \times m} \rightarrow \mathbb{R}^m$ is used, together with some estimator of the variance $\psi : \mathbb{R}^{n \times m} \rightarrow \mathbb{R}_+$. Then, our general EBE algorithm simply applies ϕ_{EB} to modify the output of algorithm A, using the estimated variance $\hat{\sigma}^2 = \psi(\mathbf{X})$.

Theorem - $E[L(\phi_{EB}(\vec{X}^A), \sigma), \mu] < E[L(\vec{X}^A, \mu)] \forall \mu \in \mathbb{R}^m$ if and only if

$$2(m-3) \sum_{j=1}^m \text{Cov} \left(X_j^A, \frac{\psi(\mathbf{X})(X_j^A - \bar{X}^A)}{\|\vec{X}^A - \bar{X}^A \bar{1}\|^2} \right) - (m-3)^2 E_{\vec{\mu}} \left(\frac{(\psi(\mathbf{X}))^2}{\|\vec{X}^A - \bar{X}^A \bar{1}\|^2} \right) > 0.$$

Under further assumptions, we show that -

$E[L(\phi_{EB}(\vec{X}^A), \sigma), \mu] < E[L(\vec{X}^A, \mu)] \forall \mu \in \mathbb{R}^m$ if and only if $E_{\vec{\mu}, \sigma}(\psi(\mathbf{X})) < 2\sigma^2$

