fabl vs fastLink: An Attempt at Mathematical Rigor

Brian Kundinger

April 9, 2024

Here are a couple of realizations about fabl and record linkage in general. First, a surprising result with an incredibly straightforward proof.

1 EXPECTED WEIGHT FOR NON-MATCHING RECORD PAIRS

Let $p \in [P]$ denote the patterns that comparison vectors can take. Let $w_p = \frac{m_p}{u_p}$ where $m_p = \prod_{p=1}^{P} m_p^{I(\gamma_{ij}=h_p)}$ and u_p is defined similarly. Let $E_u[w]$ be the expected value of the weight for nonmatching comparison vector, so $E_u[w] = E\left[\prod_{p=1}^{P} w_p^{I(\gamma_{ij}=h_p)} | \Delta_{ij} = 0\right]$. We have

$$E_u[w] = \sum_{p=1}^{P} \frac{m_p}{u_p} u_p = \sum_{p=1}^{P} m_p = 1.$$

Surprising! This is true regardless of which Fellegi Sunter style model we choose, for any choice of m and u, by definition. This seems to just be a property of likelihood ratios, but its interesting that it comes up on our setting.

Another interesting fact:

$$\begin{aligned} V_u[w] &= E_u[w^2] - E_u[w]^2 \\ &= \sum_{p=1}^{P} \frac{m_p^2}{u_p^2} u_p - 1 \\ &= \sum_{p=1}^{P} \frac{m_p^2}{u_p} - 1 \\ &= \sum_{p=1}^{P} \frac{m_p}{u_p} m_p - 1 \\ &= E_m[w] - 1 \end{aligned}$$

This is a pretty result! It also suggests that for any reasonable Fellegi Sunter linkage task, we must have $E_m[w] > 1$.

Moreover, we gain some insight around something you mentioned in our call. Using the actual definition of variance, we have

$$V_u[w] = E_u[w - E_u[w]]^2 = \sum_{p=1}^{P} (w-1)^2 u_p$$

This means that the spread of the weights around 1 is the actual definition of the variance for the nonmatching record pairs, and that increasing the variability of the nonmatching pairs is equivalent to increasing $E_m[w]$.

Interesting stuff!

2 FIXED AND TRUE M AND U PARAMETERS

We compare the estimates of the fitted match probability under base Fellegi Sunter (FS) and fabl (fabl). In order to compare, we assume that m and u are the same across the two models, and use the correspondance $\lambda = \frac{\pi}{n_1}$. We will also assume there for every record in X_2, there is at most one matching record in X_1. In both models, we are looking for $p(\Delta_{ij} = 1 | \Gamma, m, u)$.

Because of the independent record pair assumption in standard Fellegi Sunter, the

estimate under fastLink can be expressed as.

$$\begin{aligned} p_{\text{fastLink}}(\Delta_{ij} = 1 | \gamma_{ij} = h_p, m, u) &= \frac{\lambda m_p}{\lambda m_p + (1-\lambda) u_p} \\ &= \frac{\frac{\pi}{n_1} w_p}{\frac{\pi}{n_1} w_p + (1 - \frac{\pi}{n_1})} \\ &= \frac{w_p}{w_p + (\frac{n_1}{\pi} - 1)} \end{aligned}$$

Under fabl, the match probability for the (i, j) record pair is dependent on the weights for all record pairs involving record $j \in X_2$. Letting $\Gamma_{(-j)} = \{\gamma_{qj} | \forall q \neq i\}$, we can more explicitly notate this dependency:

$$\begin{aligned} p_{\text{fabl}}(\Delta_{ij} = 1 | \gamma_{ij} = h_p, \Gamma_{(-j)} = \gamma_{(-j)}) &= \frac{\frac{\pi}{n_1} w_p}{\frac{\pi}{n_1} \sum_{q=1}^{n_1} \prod_{p=1}^{P} w_p^{I(\gamma_{qj} = h_p)} + (1 - \pi)} \\ &= \frac{w_p}{w_p + \sum_{q \neq j} \prod_{p=1}^{P} w_p^{I(\gamma_{qj} = h_p)} + (\frac{n_1}{\pi} - n_1)} \end{aligned}$$

With this explicit form, we gain a better since of how fabl operates. Let γ_{ij} be a nonmatching record pair. When $Z_j > 0$, we know that $\sum_{q \neq j} \prod_{p=1}^{P} w_p^{I(\gamma_{qj} = h_p)}$ will contain a matching record pair, and will therefore likely contain a large w_p, and therefore $p_{\text{fabl}}(\Delta_{ij} = 1 | \gamma_{ij} = h_p, \Gamma_{(-j)} = \gamma_{(-j)})$ will be shrunk towards zero.

Unfortunately, we cannot simply take an expectation of this quantity over the possible orientations of $\Gamma_{(-j)}$, because it is in the denominator.

3 EXPECTED RATIO OF MATCH PROBABILITIES

I'm not sure if this is hacky or not, but I did find a way to construct an interesting quantity: the expected ratio of match probabilities between the two models. I believe this quantity would characterize the relationship between the update for the m parameters (and by extension, the u parameters as well) for the two models, leading to different estimates of those parameters (and then of course)

Let γ_{ij} be a nonmatching record pair. For any pattern h_p, we have

$$
\begin{aligned}
&E_{\Gamma_{(-j)}}\left[\frac{p_{\text{fastLink}}(\Delta_{ij}=1|\gamma_{ij}=h_p)}{p_{\text{fabl}}(\Delta_{ij}=1|\gamma_{ij}=h_p,\Gamma_{(-j)}=\gamma_{(-j)})}\right] \\
&\quad= E_{\Gamma_{(-j)}}\left[\frac{\frac{w_p}{w_p+(\frac{n_1}{\pi}-1)}}{\frac{w_p}{w_p+\sum_{q\neq j}\prod_{p=1}^{P} w_p^{I(\gamma_{qj}=h_p)}+(\frac{n_1}{\pi}-n_1)}}\right] \\
&\quad= E_{\Gamma_{(-j)}}\left[\frac{w_p+\sum_{q\neq j}\prod_{p=1}^{P} w_p^{I(\gamma_{qj}=h_p)}+(\frac{n_1}{\pi}-n_1)}{w_p+(\frac{n_1}{\pi}-1)}\right] \\
&\quad= E_{\Gamma_{(-j)}}\left[E\left[\frac{w_p+\sum_{q\neq j}\prod_{p=1}^{P} w_p^{I(\gamma_{qj}=h_p)}+(\frac{n_1}{\pi}-n_1)}{w_p+(\frac{n_1}{\pi}-1)}\middle| I(Z_j>0)\right]\right] \\
&\quad= \pi\left(\frac{w_p+E_m[w]+n_1-2+(\frac{n_1}{\pi}-n_1)}{w_p+(\frac{n_1}{\pi}-1)}\right)+(1-\pi)\left(\frac{w_p+n_1-1+(\frac{n_1}{\pi}-n_1)}{w_p+(\frac{n_1}{\pi}-1)}\right) \\
&\quad= \pi\left(\frac{w_p+E_m[w]-2+(\frac{n_1}{\pi})}{w_p+(\frac{n_1}{\pi}-1)}\right)+(1-\pi)(1) \\
&\quad> 1
\end{aligned}
$$

because $E_m[w] > 1$.

If we assume γ_{ij} is a matching record pair, conditioning on $I(Z_j > 0)$ doesn't give any useful information, and the expected ratio is 1. So on average, it seems like fastLink attributes more match probability to nonmatching pairs than fabl does. Immediately, this suggests that the record level rate of matching parameter $\pi_{\text{fastLink}} = \lambda n_1$ would be greater on average than π_{fabl}.

Moreover, I used the quotient rule to check that the quantity $\frac{w_p+E_m[w]-2+(\frac{n_1}{\pi})}{w_p+(\frac{n_1}{\pi}-1)}$ is increasing in w_p. This means that smaller weights proportionally exhibit more intense shrinkage. Interesting!

Caveat: I fear there may be something wrong with this math. It just weirds me out that I was able to calculate that quantity, but not the reverse,

$$E_{\Gamma_{(-j)}}\left[\frac{p_{\text{fabl}}(\Delta_{ij}=1|\gamma_{ij}=h_p)}{p_{\text{fastlink}}(\Delta_{ij}=1|\gamma_{ij}=h_p,\Gamma_{(-j)}=\gamma_{(-j)})}\right],$$

or the difference

$$E_{\Gamma_{(-j)}}\left[p_{\text{fastlink}}(\Delta_{ij}=1|\gamma_{ij}=h_p,\Gamma_{(-j)}=\gamma_{(-j)})-p_{\text{fabl}}(\Delta_{ij}=1|\gamma_{ij}=h_p)\right]$$

because the randomness was found in the demoninator. That's why I said it felt hacky.

4 LESS CONCRETE THOUGHTS

By removing match probability associated with low weight patterns, I anticipate that $m_{\hat{p}_{\text{fastLink}}} < m_{\hat{p}_{\text{fabl}}}$ for high weight patterns, and $m_{\hat{p}_{\text{fastLink}}} > m_{\hat{p}_{\text{fabl}}}$ for low weight patterns. I feel like this would have the effect of increasing $E_m[w]$.

Additionally, because of the normalization that occurs in $p_{\text{fabl}}(\Delta_{ij} = 1|\gamma_{ij} = h_p, \Gamma_{(-j)} = \gamma_{(-j)})$, we see that fabl is able to not match certain high weight patterns if they are in the company of even higher weight patterns. This is not possible in FS, where each record is viewed independently. Heuristically, this seems to diversify the set of records that are labeled nonmatches. And as we saw before, adding variability to the w_p associated with nonmatches is equivalent to increasing $E_m[w]$.

Lastly, if we are indeed increasing $E_m[w] = \sum_{p=1}^{P} \frac{m_p}{u_p} m_p$, it seems to me that we would also be increasing the KL divergence between the m and u distributions, which here is defined as

$$KL(m||u) = \sum_{p=1}^{P} \log\left(\frac{m_p}{u_p}\right) m_p$$

This would be pretty strong mathematical way of saying "fabl is better at distinguishing between the matching and nonmatching record pairs".