Discussion on FastLink

Using Probabilistic Model to Assist Merging of Large-Scale Administrative Records

Johan Lim

Statistics, Seoul National University

Jan. 2018

Prologue

- Merging of two data sets

- FastLink is scalable in speed and size. For the scalability
  - simple model
  - ex-post adjustments for auxiliary variables (eg. name and migration)
  - parallel processing

- Discussion:
  - mostly on the first 12 pages of the paper (Section 1 - Section 2.3)
  - no try both R packages "FastLink" and "RecordLinkage"
  - **optimality** of FastLink in the context of **multiple testing error control**
  - not consider missingness in the data for simplicity.

FastLink-1

- Observed similarity measure $\boldsymbol{\gamma}(i,j)$ is $1 \times K$ vector of similarity measure. $\boldsymbol{\Gamma} = \{\boldsymbol{\gamma}(i,j), i \in \mathcal{A}, j \in \mathcal{B}\}$.
- The unobserved M_{ij} (what we want to predict) is the 0/1 variable indicating the matching of $i \in \mathcal{A}$ and $j \in \mathcal{B}$. "1" implies the matching.
- Define

$$\xi_{ij} = \Pr\big(\mathrm{M}_{ij} = 1 \big| \boldsymbol{\gamma}(i,j)\big). \tag{1}$$

- **FastLink**: Decision rule is for some $\mathbf{c}$

$$\widehat{\mathrm{M}}_{ij} = \begin{cases} 1 & \xi_{ij} \geq \mathbf{c} \\ 0 & \xi_{ij} < \mathbf{c} \end{cases}$$

FastLink-2

- The threshold $\mathbf{c}$ is chosen as (in middle of page 8) to make the FDR be controlled at the aimed level α.

  Fortunately, the probabilistic model can estimate the false discovery rate (FDR) and the false discovery rate (FNR). We can estimate the FDR using our model parameter as

$$\Pr\big(\mathrm{M}_{ij}=1\big|\xi_{ij}\geq c\big)=\frac{\sum_{i=1}^{N_{\mathcal{A}}}\sum_{j=1}^{N_{\mathcal{B}}}\mathbf{1}\big\{\xi_{ij}\geq\mathbf{c}\big\}\big(1-\xi_{ij}\big)}{\sum_{i=1}^{N_{\mathcal{A}}}\sum_{j=1}^{N_{\mathcal{B}}}\mathbf{1}\big\{\xi_{ij}\geq\mathbf{c}\big\}}.$$

  and the FNR as

$$\Pr\big(\mathrm{M}_{ij}=1\big|\xi_{ij}<\mathbf{c}\big)=\frac{\sum_{i=1}^{N_{\mathcal{A}}}\sum_{j=1}^{N_{\mathcal{B}}}\xi_{ij}\mathbf{1}\big\{\xi_{ij}<\mathbf{c}\big\}}{\lambda N_{\mathcal{A}}N_{\mathcal{B}}}.$$

- No theory on optimality of FastLink is given in the paper.

- Here, we want to say it is (or its slight modification) is **asymptotically optimal** in **some sense**.

Multiple Testing-1

- Reformulate the problem into a **multiple testing problem**:

  The hypothesis for the pair (i, j), $i = 1, \ldots, N_{\mathcal{A}}$, $j = 1, \ldots, N_{\mathcal{B}}$ is:

  $$\underline{\mathcal{H}_{0,(i,j)} : \mathrm{M}_{ij} = 0 \text{ (no-match)}} \quad \text{vs} \quad \underline{\mathcal{H}_{1,(i,j)} : \mathrm{M}_{ij} = 1 \text{ (match)}}$$

  For notational simplicity, let $h = 1, 2, \ldots, N_{\mathcal{AB}} (= N_{\mathcal{A}} N_{\mathcal{B}})$,

  $$\underline{\mathcal{H}_{0,h} : \mathrm{M}_h = 0 \text{ (no-match)}} \quad \text{vs} \quad \underline{\mathcal{H}_{1,h} : \mathrm{M}_h = 1 \text{ (match)}}$$

  Remark that under **the model assumption (independence assumption)** of the paper, the procedure is permutable/exchangable in $\mathcal{A}$ and $\mathcal{B}$.

Multiple Testing-2

- The outcomes of multiple testing is summarized as: **given** $\mathbf{c}$,

Hypothesis	Claimed "no-match"	Claimed "match"	sum
True "no-match"	N_{00}(TN)	N_{10}(FD)	N_0
True "match"	N_{01}(FN)	N_{11}(TD)	N_1
Total	$\mathrm{S} = N_{\mathcal{AB}} - \sum_h \widehat{\mathrm{M}}_h$	$\mathrm{R} = \sum_h \widehat{\mathrm{M}}_h$	$N_{\mathcal{AB}}$

- Multiple testing errors:
  - $\mathrm{FDR} = \mathrm{E}\left((\mathrm{FD}/\mathrm{R}) \cdot \mathrm{I}(\mathrm{R} > 0)\right)$
  - $\mathrm{FNR} = \mathrm{E}\left((\mathrm{FN}/\mathrm{S}) \cdot \mathrm{I}(\mathrm{S} > 0)\right)$.
  - $\mathrm{mFDR} = \mathrm{E}(\mathrm{FD})/\mathrm{E}(\mathrm{R})$, marginal FDR
  - $\mathrm{mFNR} = \mathrm{E}(\mathrm{FN})/\mathrm{E}(\mathrm{S})$, marginal FNR

Optimal FastLink-1

- Under general dependence on γ_hs (equal to $\gamma(i,j)$s, $\mathbf{\Gamma}$)

- Consider

$$\xi_h = \Pr\big(\mathrm{M}_h = 1 \big| \mathbf{\Gamma}\big). \tag{2}$$

  and $\xi_{(h)}$ is the $h-$th largest order statistic of $\xi_1, \xi_2, \ldots, \xi_{N_{\mathcal{AB}}}$.

- **Optimal FastLink (o-FastLink)**:

  Rejects hypotheses $\mathcal{H}_{0,(h)}$ for $h = 1, 2, \ldots, h^*$, where

$$h^* = \max\left\{ h \;\middle|\; \frac{1}{h}\sum_{j=1}^{h} \xi_{(j)} \geq 1 - \alpha \right\}.$$

  The thresholding of o-FastLink is $\mathbf{c}_{\mathrm{opt}} = \xi_{(h^*)}$.

Optimal FastLink-2

Theorem

The above o-FastLink minimizes the mFNR over all decision rules depending on $\mathbf{\Gamma}$ at the mFDR being less than α.

- The proof is from the equivalence between the optimal weighted classification problem and optimal multiple testing problem under "monotone ratio condition (MRC)" proven by Sun and Cai (2007, 2009)[1] The above is more close to the corollary of their main theorems.
- The MRC is satisfied under the model (independnce) assumptions of the paper. Yes.
- Need consistent estimators of the model parameters, λ and $\boldsymbol{\pi}_{km}$. Yes.

[1](1) Sun, W. and Cai, T. T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association*, **102**, 901-912. (2) Sun, W. and Cai, T. (2009). Large-scale multiple testing under dependence. *Journal of the Royal Statistical Society: Series B*, **71**, 393-424.

Optimal FastLink-3

- We further claim that "FastLink" is asymptotically equivalent to "o-FastLink". Believe it should be.

- The theorem is true even **for the general dependence**.

$$\xi_h = \Pr\left(\mathrm{M}_h = 1 \middle| \mathbf{\Gamma}\right). \tag{3}$$

General Dependence-1

- FastLink makes the independence assumptions, for $m = 0, 1$, (in page 6)

$$\gamma_k(i,j)\big|\mathrm{M}_{ij} = m \quad \sim \text{ indep}, \quad \mathrm{Discrete}\big(\boldsymbol{\pi}_{km}\big)$$
$$\mathrm{M}_{ij} \quad \sim \text{i.i.d} \quad \mathrm{Bernoulli}(\lambda).$$

- Independence observations: $\boldsymbol{\gamma}(i,j) = (\gamma_1(i,j), \gamma_2(i,j), \ldots, \gamma_K(i,j))$s are independent for $i \in \mathcal{A}$ and $j \in \mathcal{B}$. Thus,

$$\xi_h = \Pr\big(\mathrm{M}_h = 1\big|\boldsymbol{\Gamma}\big) = \Pr\big(\mathrm{M}_h = 1\big|\boldsymbol{\gamma}_{ij}\big).$$

- Special dependence: one-to-one match (bottom of page 6):

  The iid assumption of the latent variable M_{ij} is necessarily violated if each record in the data set $\mathcal{A}$ should be matched with no more than one record in the data set $\mathcal{B}$.

  Solve additional linear sum assignment problem to enforce one-to-one match.

General Dependence-2

- Need a probabilistic model for dependent $\boldsymbol{\gamma}(i,j)$s that
  - provides a good approximation to the real
  - gives an easy computation of $\xi_{ij} = \Pr\left(\mathrm{M}_{ij} = 1 \middle| \boldsymbol{\Gamma}\right)$.
- A popular approach for the dependent $\boldsymbol{\gamma}(i,j)$s is the model with dependent latent variables M_{ij}s. (eg. hidden Markov model for the observations on time, d-lattice system, sparse graph).
- Any good model for the bipartite graph (completely dense graph)? How about M. Sadinle (2017)[2]?

[2]Sadinle, M. (2017) Bayesian estimation of bipartite matchings for record linkage.To appear in *Journal of the American Statistical Association*.

Epilogue

- No doubt that FastLink is a good frequentist alternative to many Bayesian ways.
- In Korea, the record linkage may not be an important issue due to "unique resident registration number". But, the Korean data can be a good test bed for new linking methdology.
- In statistical disclosure control, the record linkage provides a measure of disclosure risk.
- Thank you.