Mahalanobis distance-based record linkage revisited

Krish Muralidhar, Price College of Business, University of Oklahoma

Josep Domingo-Ferrer, UNESCO Chair in Data Privacy, Department of Computer Science and Mathematics, Universitat Rovira i Virgili

Preliminaries

- We use $\boldsymbol{Y}$ (a matrix of dimension $n{\times}m$, where n is the number of records and m the number of attributes) to represent the original confidential data, with
  - y_{ij} (i = 1, 2,…, n; j = 1, 2,…,m) denoting the confidential value of the j-th attribute for the i-th person,
  - $\boldsymbol{y_i}$ (of dimension 1×m) denoting a single row from $\boldsymbol{Y}$, and
  - $\boldsymbol{\Sigma_{YY}}$ is the covariance matrix of $\boldsymbol{Y}$
- We use $\widetilde{\boldsymbol{Y}}$ to represent the masked data but use index k to represent the fact that while $\widetilde{\boldsymbol{Y}}$ is generated from $\boldsymbol{Y}$, the exact linkage of the records in $(\boldsymbol{Y}, \widetilde{\boldsymbol{Y}})$ is unknown to the user

Distance based record linkage: Euclidean

- Given an original record $\boldsymbol{y_i}$, compute

$$d_{ik}^2 = \sum_{j=1}^{m} \left(y_{ij}^* - \tilde{y}_{kj}^*\right)^2$$

  - where $\left(y_{ij}^*, \tilde{y}_{kj}^*\right)$ represent attributes standardized to have zero mean and unit variance
- Repeat the process for every masked record $k = 1, 2, \dots, n$.
- Link $\boldsymbol{y_i}$ to that record $\tilde{\boldsymbol{y}}_{\boldsymbol{k}}$ with $Min(d_{ik}^2)$.

Distance based record linkage: Mahalanobis (Torra et al 2006)

- $d_{ik}^2 = (\boldsymbol{y}_i - \widetilde{\boldsymbol{y}}_k)\boldsymbol{S}^{-1}(\boldsymbol{y}_i - \widetilde{\boldsymbol{y}}_k)^T$
- $\boldsymbol{S}$ is the estimate of $\boldsymbol{\Sigma}_{\boldsymbol{ee}}$
- Computation of $\boldsymbol{S}$ requires knowledge of $\boldsymbol{\Sigma}_{\boldsymbol{Y}\widetilde{\boldsymbol{Y}}}$, which in turn requires the true linkage between $\boldsymbol{Y}$ and $\widetilde{\boldsymbol{Y}}$
- Alternative estimate $\boldsymbol{S} := \boldsymbol{\Sigma}_{\boldsymbol{YY}} + \boldsymbol{\Sigma}_{\widetilde{\boldsymbol{Y}}\widetilde{\boldsymbol{Y}}}$

Torra, V., Abowd, J.M., Domingo-Ferrer, J. (2006) Using Mahalanobis distance-based record linkage for disclosure risk assessment. In: *Privacy in Statistical Databases - PSD 2006*, LNCS 4302, pp. 233–242. Springer.

Information preserving statistical obfuscation (Burridge 2003)

- IPSO is a masking method for generating synthetic data where the mean vector and covariance matrix of the masked data are *identical* to the original data.
  - a set of m_1 quasi-identifier attributes ($\boldsymbol{Y}$) which are masked into ($\tilde{\boldsymbol{Y}}$) and then released
  - a set of m_2 confidential attributes ($\boldsymbol{X}$) which are released unmasked
- $\tilde{\boldsymbol{y}}_i = \boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{e}_i$
  - $\boldsymbol{\beta}$ represent the regression coefficients to predict $\boldsymbol{Y}$ from $\boldsymbol{X}$
  - $\boldsymbol{e}_i \sim \boldsymbol{Normal}(\boldsymbol{0}, \boldsymbol{\Sigma}_{ee})$ $\boldsymbol{\mu}_e \equiv \boldsymbol{0}, \boldsymbol{\Sigma}_{ee} \equiv \boldsymbol{\Sigma}_{YY} - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}$, and orthogonal to $\boldsymbol{x}_i\boldsymbol{\beta}$.
- This ensures that $\boldsymbol{\mu}_{\tilde{Y}} \equiv \boldsymbol{\mu}_Y$, $\boldsymbol{\Sigma}_{\tilde{Y}\tilde{Y}} \equiv \boldsymbol{\Sigma}_{YY}$, and $\boldsymbol{\Sigma}_{X\tilde{Y}} \equiv \boldsymbol{\Sigma}_{XY}$.

IPSO Variations

- IPSO-A: Preserves the characteristics only asymptotically
- IPSO-B: Preserves the regression coefficients of $(\tilde{\boldsymbol{Y}}\ on\ \boldsymbol{X})$ to be the same as $(\boldsymbol{Y}\ on\ \boldsymbol{X})$
- IPSO-C: $\boldsymbol{\mu}_{\tilde{\boldsymbol{Y}}} \equiv \boldsymbol{\mu}_{\boldsymbol{Y}}$, $\boldsymbol{\Sigma}_{\tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}} \equiv \boldsymbol{\Sigma}_{\boldsymbol{YY}}$, and $\boldsymbol{\Sigma}_{\boldsymbol{X}\tilde{\boldsymbol{Y}}} \equiv \boldsymbol{\Sigma}_{\boldsymbol{XY}}$

Burridge, J. (2003) Information preserving statistical obfuscation. *Statistics and Computing*, 13:321–327.

Application of Mahalanobis distance to IPSO (Torra et al 2006) – EIA data set

Masking Method	**Mahalanobis Distance (Torra et al)**	**Euclidean Distance**
IPSO-A	3206	66
IPSO-B	3194	65
IPSO-C	773	65

- Mahalanobis distance easily outperforms Euclidean distance by a wide margin

But …

- Why does IPSO-C provide so much better protection than IPSO-A or IPSO-B?

Masking Method	Mahalanobis Distance (Torra et al)	Euclidean Distance
IPSO-A	3206	66
IPSO-B	3194	65
IPSO-C	773	65

The problem with Torra et al (2006) approach

- Torra et al (2006) used the same specification $d_{ik}^2 = (\boldsymbol{y_i} - \tilde{\boldsymbol{y}}_{\boldsymbol{k}})\boldsymbol{S}^{-1}(\boldsymbol{y_i} - \tilde{\boldsymbol{y}}_{\boldsymbol{k}})^T$ and $\boldsymbol{S} := \boldsymbol{\Sigma}_{\boldsymbol{YY}} + \boldsymbol{\Sigma}_{\tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}}$ in all three cases (IPSO-A, IPSO-B, IPSO-C)
- For IPSO-A and IPSO-B, $(\boldsymbol{y}_i - \tilde{\boldsymbol{y}}_i) \cong (\boldsymbol{y}_i - \boldsymbol{x}_i\boldsymbol{\beta})$ and $\boldsymbol{S}$ is negligible. So the Torra et al procedure works well.
- This is not the case for IPSO-C. Hence, the approach suggested by Torra et al (2006) for IPSO-C performs poorly (relative to IPSO-A and IPSO-B)

Different specifications are required for the IPSO method

- For IPSO-C, the appropriate computation should be
  - $\left(\boldsymbol{y}_i - E(\tilde{\boldsymbol{y}}_i)\right) = (\boldsymbol{y}_i - \boldsymbol{x}_k\boldsymbol{\beta})$
  - $\boldsymbol{S} = \boldsymbol{\Sigma}_{\boldsymbol{ee}} \equiv \boldsymbol{\Sigma}_{\tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}} - \boldsymbol{\Sigma}_{\tilde{\boldsymbol{Y}}\boldsymbol{X}}\boldsymbol{\Sigma}_{\boldsymbol{XX}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{X}\tilde{\boldsymbol{Y}}}$
  - All estimates for IPSO-C can be computed using only the released data. Other than the target record, no access to the original data is necessary

Record Linkage Results

Masking Method	Modified Mahalanobis Distance	Mahalanobis Distance (Torra et al)	Euclidean Distance
IPSO-A	3219	3206	66
IPSO-B	3196	3194	65
IPSO-C	**3206**	**773**	65

- The three variations of IPSO are inherently similar. We **should** expect to see similar record linkage results for all three IPSO variations.

Some experimental results

- Theoretically, Mahalanobis distance based measure should always outperform Euclidean distance
- In practice, this may not be the case
  - Euclidean distance requires no estimation
  - Mahalanobis distance requires estimation of $\boldsymbol{\Sigma}_{ee}$
    - Inaccuracy in estimation could result in poor performance

Simulation experiment

- Compare performance of Euclidean and Mahalanobis distance record linkage for
  - Different types of data (Low, medium, high, and mixed correlation among variables)
  - Two masking methods (simple versus correlated noise)
  - Three different perturbation levels (low, medium, high)
  - For a given data set, apply both simple and correlated noise. Evaluate performance of Euclidean and Mahalanobis distance based record linkage

What we expected

- For independent noise, by default, Euclidean distance will be the best record linkage procedure since $\boldsymbol{\Sigma_{ee}}$ is a diagonal matrix

- For highly correlated data masked using correlated noise, Mahalanobis distance will perform better since the (non-diagonal) structure of $\boldsymbol{\Sigma_{ee}}$ will have a significant impact on record linkage performance

Results showing best record linkage performance

Correlation Structure of Original Data	Perturbation Level	Independent Noise	Correlated Noise
Low	Low	Euclidean	Both
Low	Medium	Euclidean	Euclidean
Low	High	Both	Euclidean
Medium	Low	Euclidean	Mahalanobis
Medium	Medium	Euclidean	Both
Medium	High	Euclidean	Both
High	Low	Both	Mahalanobis
High	Medium	Both	Mahalanobis
High	High	Both	Mahalanobis
Mixed	Low	Euclidean	Mahalanobis
Mixed	Medium	Euclidean	Mahalanobis
Mixed	High	Euclidean	Mahalanobis

General conclusion

- Record linkage techniques must take advantage of the knowledge of the masking method
- The success of the record linkage techniques will depend on their ability to **accurately reverse engineer the masking method**

Further research

- Does every masking method have a corresponding record linkage procedure that results in best possible linkages?
  - Independent noise – Euclidean distance (this study)
  - Correlated noise – Mahalanobis distance (this study)
  - IPSO style procedures – Mahalanobis distance (this study)
  - Data swapping – Rank based record linkage (Nin, et al 2008)
  - Multiplicative noise – Distance measure based on Geometric mean?
  - Other masking methods?

Questions?

ти благодарам