Bayesian non parametric LR assessment for the rare type match

Giulia Cereda

University of Lausanne, Institute of Forensic Science.
Leiden University, Mathematical Institute.

The problem

A match between suspect’s DNA profile and crime stain’s DNA profile.

Data: Do the match and a database of reference.

Two hypotheses of interest:
1. Hypothesis of interest: The suspect left the stain.
2. Hypothesis of interest: Someone else left the stain.

The situation in which the suspect’s DNA type is not in the database.

Assumptions of the model

Assumption 1

There are infinitely many DNA types in Nature.

Assumption 2

Each DNA type is just a category: no structure is assumed.

Reduction of data

Assumption 2 allows to reduce a database of size n and the two new observations to the partition of the set \([n+2]\) obtained by the classes of the equivalence relation “to have the same DNA type”.

Example:

Database:

\[
\{(1,3,6), (2,4,5), (7,8)\}
\]

Suspect:

Crime stain

The reading of the reduced data only depends on \(p = (p_1, p_2, ..., p_2) \geq p_2 \geq ...\) the ranked vector of DNA frequencies in Nature.

Two parameter Poisson Dirichlet prior

Two parameter GEM distribution

Given \(0 < \alpha < 1\) and \(\theta > -\alpha\), the vector \(W = (W_1, W_2, ..., )\) is said to be distributed according to the GEM(\(n, \theta\)), if \(W = W_n \prod_{i=1}^{\infty} (1 - U_i)\), where \(V_i, U_i, \ldots\) are independent random variables with \(V_i \sim B(1 - \alpha, \theta + \alpha)\).

Two parameter Poisson Dirichlet distribution

Given \(0 < \alpha < 1\), \(\theta > -\alpha\), and a vector \(W = (W_1, W_2, ..., ) \sim \text{GEM}(\alpha, \theta)\), the random vector \(P = (P_1, P_2, ..., )\) obtained by ordering \(W\) such that \(P_1 \geq P_2\), has the two parameter Poisson Dirichlet distribution \(P(\alpha, \theta)\).

The model

\((X_1, X_2, ..., X_{n+2}) \sim (\pi, \theta)\)

The equivalence classes of the indices \(\{1, 2, ..., n + 2\}\) generate the same partition of \([n + 2]\) as that obtained with the database.

The distribution of \(D\)

The following result simplifies the model:

\[P(\alpha, \theta) \sim \text{PD}(\alpha, \theta)\]

\[X_1, X_2, ..., X_{n+1} \sim P(\alpha, \theta)\]

The trend of the reduced data only depends on \(p = (p_1, p_2, ..., p_2) \geq p_2 \geq ...\) the ranked vector of DNA frequencies in Nature.

The likelihood ratio

Figure 1

Rank observed frequencies from a database of Y-STR haplotypes from N=18925 individuals.

Figure 2

Loglikelihood function reparametrized with \(\hat{\phi} = \frac{1}{\theta + \alpha}\) and \(\hat{\theta}\) using the same database, compared to Gaussian distribution centered at the MLE.

Conclusions

- Study of the asymptotic behavior of the estimators \(\hat{\phi}_{MLE}\) and \(\hat{\theta}_{MLE}\) (available theory contradicts the Gaussian behavior).
- Comparison with a nonparametric frequentist estimator for the LR (Cereda (2015b)).
- Extension to the general case (not only rare type match).

Future research questions

- Study of the asymptotic behavior of the estimators \(\hat{\phi}_{MLE}\) and \(\hat{\theta}_{MLE}\) (available theory contradicts the Gaussian behavior).
- Comparison with a nonparametric frequentist estimator for the LR (Cereda (2015b)).
- Extension to the general case (not only rare type match).

References


FNS grant n.: 10531A-156146/1.