On Parametric Models for Pairwise Comparisons

with Applications to Estimation of Random Match Probabilities

Donald Gantz, Department of Applied IT, George Mason University
John Miller, Department of Statistics, George Mason University
Chris Saunders, Department of Applied IT, George Mason University

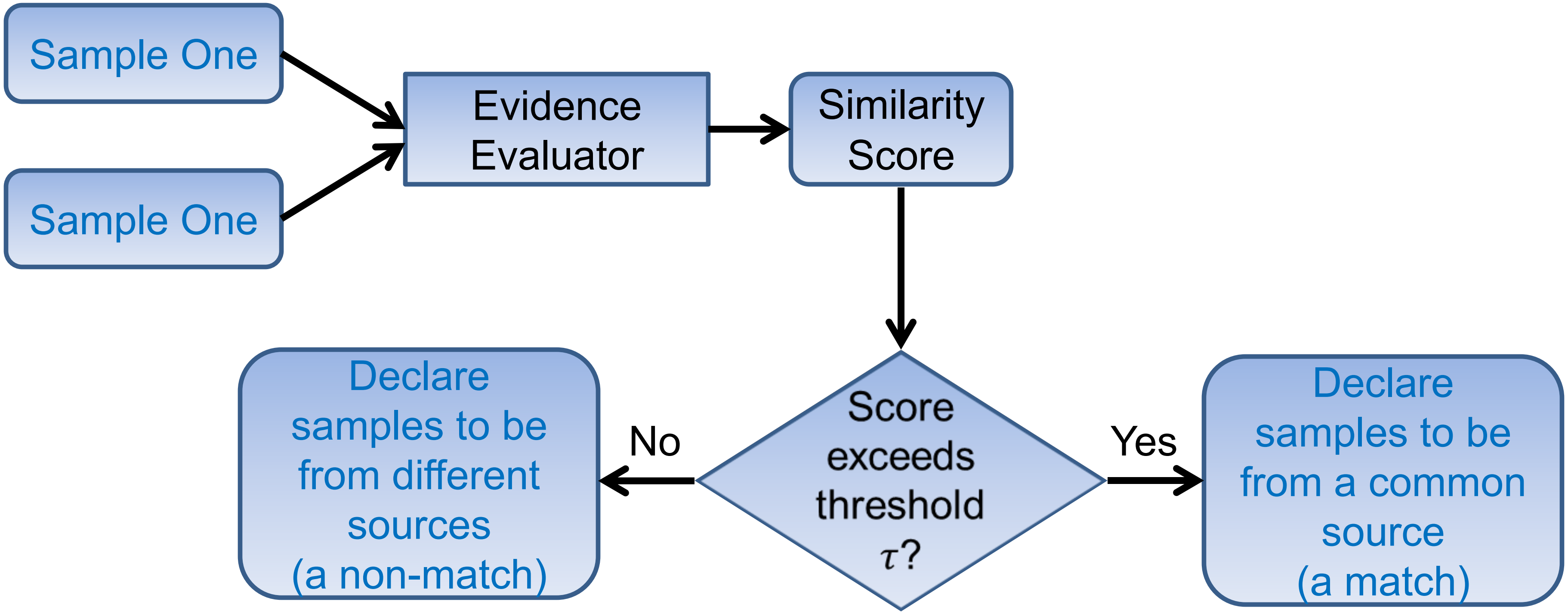
August 11, 2011

Trace Evidence Symposium
Kansas City, MO

Acknowledgement and Disclaimer

The research detailed in this presentation was supported in part by Award No. 2009-DN-BX-K234 awarded by the National Institute of Justice, Office of Justice Programs, US Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect those of the Department of Justice.

A Common Situation in Forensics


Sample One
Sample One
Evidence Evaluator
Similarity Score
Score exceeds threshold τ?
No
Yes
Declare samples to be from different sources (a non-match)
Declare samples to be from a common source (a match)

The similarity score is a numerical value measuring association of the samples, with higher values suggesting that the samples come from a common source.

A Sampling Model

Suppose we assume that we have some population of objects which can lead to samples like those on the preceding slide.

A very important consideration of the evidence evaluation method is to know the probability that two randomly sampled objects from different sources would lead (erroneously) to a conclusion of a match. This quantity is called the random match probability.

(The goal of this presentation is to obtain a valid upper confidence limit for the random match probability based on a sample.)

One way to assess the random match probability is to obtain a random sample of n objects known to have each come from a different source and calculate the similarity score for each pair of objects. This is called taking all pairwise comparisons.

Let us call the score resulting from comparing sample i to sample j

$$s_{ij}.$$

There are $N = n(n-1)/2$ possible comparisons*.

* We are assuming that $s_{ij} = s_{ji}$. That is the similarity of object i to object j is the same as the similarity of object j to object i.

A Plausible Model for the Correlation Structure of the s_{ij}

The correlation between s_{ij} and s_{kl} should be:

- One if there are two subscripts in common
- A positive quantity if there is one subscript in common
- Zero if there are no subscripts in common

An Important Quantity

The mean of the scores from all possible pairs of different source objects in the population is a very important quantity in calculations of the random match probability based on scores. We will call this parameter θ.

A plausible estimate for this quantity is the mean of the N scores from the pairwise comparisons in the sample.

It will be important to have proper estimates of the variability of this mean in forming upper confidence limits for the random match probability.

Issues

Some researchers simply ignore the correlation structure and proceed as if there is a sample of N independent scores.

Other researchers believe that the correlation structure forces one to use only uncorrelated pairs (such as s_{12}, s_{34}, s_{56}, etc.).

We will show that it is possible to account for the correlation structure to create a confidence limit for the random match probability.

A Mathematical Model for a Score

We assume that

$$s_{ij} = \theta + a_i + a_j + e_{ij},$$

where θ is an unknown parameter, the a_i are i.i.d. $N(0,{\sigma_a}^2)$ $i = 1, \ldots, n,$ and the e_{ij} are i.i.d. $N(0,{\sigma_e}^2)$, $i = 1, \ldots n-1; j = i+1, \ldots n.$

Our goal is to use this model to form a valid upper confidence limit for the random match probability.

The Model in Matrix Terms

We can rewrite this model in matrix terms by writing the scores (s_{ij}) in lexicographic order as a vector $\mathbf{y}$. The errors (e_{ij}) are listed in the same order and the a's are listed in order of their subscripts. There is a design matrix $\mathbf{P}$ (for pairwise) which has N rows and n columns. $\mathbf{P}$ is mostly composed of zeroes but has a one in the ith and jth columns for the row corresponding to s_{ij}.

Thus our model becomes

$$\mathbf{y} = \theta\mathbf{1}_N + \mathbf{Pa} + \mathbf{e},$$

where $\mathbf{y}$ and $\mathbf{e}$ are as described above, $\mathbf{a}$ is the vector of the a_i, and $\mathbf{1}_N$ is an N by 1 vector of ones.

The Expected Value and Covariance Matrix of the Vector of Scores

The N by 1 expected value of the vector of scores is just $\theta\mathbf{1}_N$.

The N by N covariance matrix of the vector of scores is

$$\boldsymbol{\Sigma} = {\sigma_e}^2\mathbf{I}_N + {\sigma_a}^2\mathbf{PP}'.$$

We can convert the covariance matrix to a correlation matrix if we wish. The resulting correlation matrix contains mostly zeroes but has non-zero correlations of $r = {\sigma_a}^2/({\sigma_e}^2 + 2{\sigma_a}^2)$ in some positions.

Our Model Expressed Using Vectors and Matrices for $n = 8$

$$y = \begin{bmatrix} \mathtt{S12} \\ \mathtt{s13} \\ \mathtt{s14} \\ \mathtt{s15} \\ \mathtt{s16} \\ \mathtt{s17} \\ \mathtt{s18} \\ \mathtt{s23} \\ \mathtt{s24} \\ \mathtt{s25} \\ \mathtt{s26} \\ \mathtt{s27} \\ \mathtt{s28} \\ \mathtt{s34} \\ \mathtt{s35} \\ \mathtt{s36} \\ \mathtt{s37} \\ \mathtt{s38} \\ \mathtt{s45} \\ \mathtt{s46} \\ \mathtt{s47} \\ \mathtt{s48} \\ \mathtt{s56} \\ \mathtt{s57} \\ \mathtt{s58} \\ \mathtt{s67} \\ \mathtt{s68} \\ \mathtt{s78} \end{bmatrix} \qquad P = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

Correlation Matrix for n = 8

$$r = {\sigma_a}^2/({\sigma_e}^2 + 2{\sigma_a}^2)$$

```
1 r r r r r r r r r r r r 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
r 1 r r r r r r 0 0 0 0 0 r r r r r 0 0 0 0 0 0 0 0 0 0
r r 1 r r r r 0 r 0 0 0 0 r 0 0 0 0 r r r r 0 0 0 0 0 0
r r r 1 r r r 0 0 r 0 0 0 0 r 0 0 0 r 0 0 0 r r r 0 0 0
r r r r 1 r r 0 0 0 r 0 0 0 0 r 0 0 0 r 0 0 r 0 0 r r 0
r r r r r 1 r 0 0 0 0 r 0 0 0 0 r 0 0 0 r 0 0 r 0 r 0 r
r r r r r r 1 0 0 0 0 0 r 0 0 0 0 r 0 0 0 r 0 0 r 0 r r
r r 0 0 0 0 0 1 r r r r r r r r r r 0 0 0 0 0 0 0 0 0 0
r 0 r 0 0 0 0 r 1 r r r r r 0 0 0 0 r r r r 0 0 0 0 0 0
r 0 0 r 0 0 0 r r 1 r r r 0 r 0 0 0 r 0 0 0 r r r 0 0 0
r 0 0 0 r 0 0 r r r 1 r r 0 0 r 0 0 0 r 0 0 r 0 0 r r 0
r 0 0 0 0 r 0 r r r r 1 r 0 0 0 r 0 0 0 r 0 0 r 0 r 0 r
r 0 0 0 0 0 r r r r r r 1 0 0 0 0 r 0 0 0 r 0 0 r 0 r r
0 r r 0 0 0 0 r r 0 0 0 0 1 r r r r r r r r 0 0 0 0 0 0
0 r 0 r 0 0 0 r 0 r 0 0 0 r 1 r r r r 0 0 0 r r r 0 0 0
0 r 0 0 r 0 0 r 0 0 r 0 0 r r 1 r r 0 r 0 0 r 0 0 r r 0
0 r 0 0 0 r 0 r 0 0 0 r 0 r r r 1 r 0 0 r 0 0 r 0 r 0 r
0 0 r 0 0 0 r r 0 0 0 0 r r r r r 1 0 0 0 r 0 0 r 0 r r
0 0 r r 0 0 0 0 r r 0 0 0 r r 0 0 0 1 r r r r r r 0 0 0
0 0 r 0 r 0 0 0 r 0 r 0 0 r 0 r 0 0 r 1 r r r 0 0 r r 0
0 0 r 0 0 r 0 0 r 0 0 r 0 r 0 0 r 0 r r 1 r 0 r 0 r 0 r
0 0 r 0 0 0 r 0 r 0 0 0 r r 0 0 0 r r r r 1 0 0 r 0 r r
0 0 0 r r 0 0 0 0 r r 0 0 0 r r 0 0 r r 0 0 1 r r r r 0
0 0 0 r 0 r 0 0 0 r 0 r 0 0 r 0 r 0 r 0 r 0 r 1 r r 0 r
0 0 0 r 0 0 r 0 0 r 0 0 r 0 r 0 0 r r 0 0 r r r 1 0 r r
0 0 0 0 r r 0 0 0 0 r r 0 0 0 r r 0 0 r r 0 r r 0 1 r r
0 0 0 0 r 0 r 0 0 0 r 0 r 0 0 r 0 r 0 r 0 r r 0 r r 1 r
0 0 0 0 0 r r 0 0 0 0 r r 0 0 0 r r 0 0 r r 0 r r r r 1
```

The Eigenstructure of Σ

There are N eigenvectors of Σ:

- 1 eigenvector ($\mathbf{v}_1 = \mathbf{1}_N/\sqrt{N}$)
  with eigenvalue $\lambda_1 = {\sigma_e}^2 + 2(n-1){\sigma_a}^2$
- $(n-1)$ eigenvectors ($\mathbf{v}_2$ to $\mathbf{v}_n$)
  with eigenvalue $\lambda_2 = {\sigma_e}^2 + (n-2){\sigma_a}^2$
- $(N-n)$ eigenvectors ($\mathbf{v}_{n+1}$ to $\mathbf{v}_N$)
  with eigenvalue $\lambda_3 = {\sigma_e}^2$

Because eigenvectors are orthogonal, we have $\mathbf{v}_k'\mathbf{1}_N = 0$ for all $k > 1$.

The Likelihood for the Vector of Scores

The log-likelihood can be written as

$$-2lnL = ln(2\pi) + ln|\boldsymbol{\Sigma}| + (\mathbf{y} - \theta\mathbf{1}_N)'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \theta\mathbf{1}_N)$$

$$= ln(2\pi) + ln\lambda_1 + (n-1)ln\lambda_2 + (N-n)ln\lambda_3$$
$$+ \frac{N(\bar{y} - \theta)^2}{\lambda_1} + \frac{\mathbf{y}'(\sum_{k=2}^{n}\mathbf{v}_k\mathbf{v}_k{}')\mathbf{y}}{\lambda_2} + \frac{\mathbf{y}'(\sum_{l=n+1}^{N}\mathbf{v}_l\mathbf{v}_l{}')\mathbf{y}}{\lambda_3}$$

$$= ln(2\pi) + ln\lambda_1 + (n-1)ln\lambda_2 + (N-n)ln\lambda_3$$

$$+ \frac{N(\bar{y} - \theta)^2}{\lambda_1} + \frac{SS_a}{\lambda_2} + \frac{SS_e}{\lambda_3}$$

Unbiased Estimates of the Parameters of the Model

We can find unbiased estimators for all parameters in our model: First let $MS_a = SS_a/(n-1)$ and $MS_e = SS_e/(N-n)$.

$$\hat{\theta} = \bar{y}$$

$$\hat{\sigma}_a{}^2 = \frac{MS_a - MS_e}{n-2}$$

$$\hat{\sigma}_e{}^2 = MS_e$$

These estimates are closely related to REML estimates. We have derived closed form versions of all three of these estimates.

(For future reference, let $SS_t = SS_a + SS_e$ and $MS_t = SS_t/(N-1)$.)

Some More Important Quantities

The variance of a randomly selected score (the similarity score for two randomly selected objects) is given by

$$\sigma_s{}^2 = \sigma_e{}^2 + 2\sigma_a{}^2.$$

The variance of the mean of all scores in the sample ($\bar{y}$) is given by

$$\sigma_{\bar{y}}{}^2 = \frac{\sigma_e{}^2}{N} + \frac{4\sigma_a{}^2}{n}.$$

We can obtain unbiased estimates for each of these quantities by plugging in the unbiased estimates of the variance components. These will be designated by "hats".

Some Other Related Quantities

The expected value of MS_t is given by

$$\sigma_e{}^2 + 2\frac{n-1}{n+1}\sigma_a{}^2$$

which is almost the same as $\sigma_s{}^2$.

The expected value of MS_t/N is given by

$$\frac{\sigma_e{}^2}{N} + \left(\frac{1}{n+1}\right)\frac{4\sigma_a{}^2}{n}$$

which is *NOT AT ALL* the same as $\sigma_{\bar{y}}{}^2$!

The Random Match Probability Based on Our Model

For a given cutoff τ, the random match probability is the probability that a randomly selected s_{ij} will exceed τ. That is

$$RMP = P\{s_{ij} > \tau\}.$$

For our model

$$P\{s_{ij} > \tau\} = 1 - \Phi\left(\frac{\tau - \theta}{\sigma_s}\right) \equiv \pi$$

where Φ is the standard normal CDF.

Some Equivalent Mathematical Statements

The following statements are equivalent for any value of B (either random or not random).

$$\pi < B \Leftrightarrow 1 - \Phi\left(\frac{\tau - \theta}{\sigma_s}\right) < B \Leftrightarrow \Phi\left(\frac{\tau - \theta}{\sigma_s}\right) > 1 - B$$

$$\Leftrightarrow \frac{\tau - \theta}{\sigma_s} > \Phi^{-1}(1 - B) \equiv L$$

Thus we require a random quantity L such that

$$P\left\{\frac{\tau - \theta}{\sigma_s} > L\right\} = 1 - \alpha$$

It is straightforward to convert such an interval to an upper confidence bound for π.

A Hypothetical

Suppose for a moment that we knew σ_e and σ_a but that we estimated θ by $\bar{y}$. Then

$$P\left\{\frac{\tau - \bar{y}}{\sigma_{\bar{y}}} - \frac{\tau - \theta}{\sigma_{\bar{y}}} < z_\alpha\right\} = 1 - \alpha$$

BUT

$$\frac{\tau - \bar{y}}{\sigma_{\bar{y}}} - \frac{\tau - \theta}{\sigma_{\bar{y}}} < z_\alpha \Leftrightarrow \frac{\tau - \theta}{\sigma_s} > \frac{\tau - \bar{y}}{\sigma_s} - z_\alpha \frac{\sigma_{\bar{y}}}{\sigma_s}$$

We could use this to get our upper confidence bound for π. The resulting interval would have exactly correct coverage probability.

What would happen if we ignored the correlation structure?

If we ignored the correlation structure in this model, we could use the same formula from the previous slide but substituting the expected value of MS_t/N for $\sigma_{\bar{y}}$. This upper confidence bound has some very bad properties.

Actual Coverage Probability for Hypothetical Method

(Nominal Coverage Probability 0.95, True RMP = 0.000001)

n	ρ =.1	ρ =.5	ρ =1	ρ =2	ρ =10
10	.855	.758	.730	.712	.695
50	.708	.627	.611	.602	.593
100	.653	.592	.580	.573	.567
500	.571	.541	.536	.533	.530
1000	.551	.529	.525	.523	.521

$\rho = {\sigma_a}^2/{\sigma_e}^2$

A Confidence Interval Based on Fieller’s Theorem

Results for New Method Based on Fieller’s Theorem

(Nominal Coverage Probability 0.95; True RMP=0.001; 1,000,000 Simulations per Cell)

n	ρ =.1	ρ =.5	ρ =1	ρ =2	ρ =10
50	.9465 .0005 .0022	.9433 .0002 .0042	.9442 .0002 .0055	.9457 .0001 .0068	.9481 .0001 .0085
100	.9474 .0006 .0017	.9467 .0004 .0027	.9469 .0003 .0034	.9479 .0002 .0040	.9495 .0002 .0048
500	.9497 .0008 .0012	.9490 .0006 .0016	.9489 .0006 .0017	.9495 .0005 .0019	.9498 .0005 .0021
1000	.9501 .0009 .0012	.9497 .0007 .0014	.9495 .0007 .0015	.9497 .0006 .0016	.9497 .0006 .0017

Coverage Probability
Average Lower Bound
Average Upper Bound

Conclusions

- We have found a method which yields an approximately correct confidence interval for the random match probability based on pairwise comparisons.

- This method can be used in any situation where the scores can be (monotonically) transformed to approximate normality.

- As long as the model is valid, an estimate of the random match probability can be made even if no matches are observed in the sample.

- We are continuing research to obtain an upper bound (rather than a two-sided interval) for the random match probability.