

---

## Multivariate matching Forensic Science

Dr. David Lucy

d.lucy@lancaster.ac.uk

Lancaster University

OOS seminar – p.1/33

---

## Multivariate matching

- Multi-variate equivalent to the univariate matching.
- More generalised approach using likelihood ratios.

A large part of making any likelihood ratio approach a practical proposition for forensic scientists is dealing with multivariate observations.

OOS seminar – p.2/33

---

## Multivariate matching

Current areas of interest are:

- Trace evidence - glass - elemental and isotopic compositions - Edinburgh, Cracow.
- Speaker identification - formant and cepstra coefficients - Canberra (Rose), Madrid.
- Facial identification - landmarks and procrustes coefficients - Nottingham (Dryden), Sheffield (Fieller, Moorcroft).

Of the three facial identification in the UK context will probably be the most readily taken up in practice.

OOS seminar – p.3/33

---

## Speaker identification

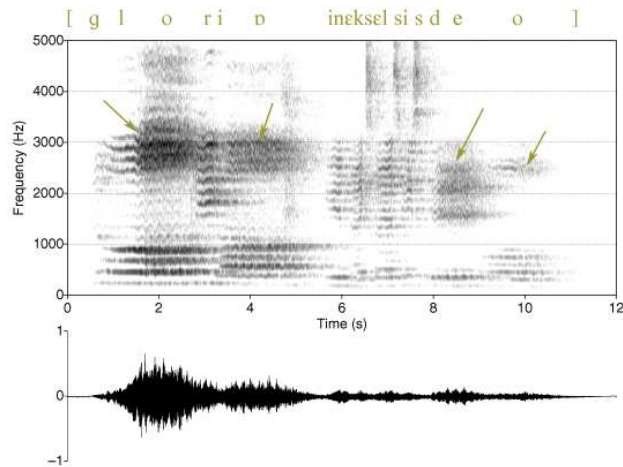
- Speech has four dimensions:
  1. time - continuous
  2. amplitude - continuous
  3. frequency - continuous
  4. sound type - categorical

The idea being to use variation between between individuals within these dimensions as a source of evidence.

Variation is caused by natural variation in the shape of the vocal tracts

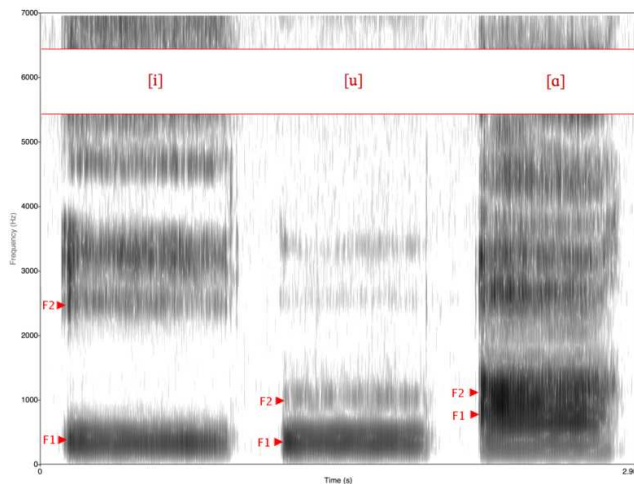
OOS seminar – p.4/33

## Speaker identification



OOS seminar – p.5/33

## Speaker identification



OOS seminar – p.6/33

## Speaker identification

Forensic speaker identification:

- Usually concerned with recorded data - telephone etc.
- Quality of recordings limited to F1 and F2, sometimes get F3.

The old Japanese land line system could get up to F5, but usually natural limit on how many variables available.

OOS seminar – p.7/33

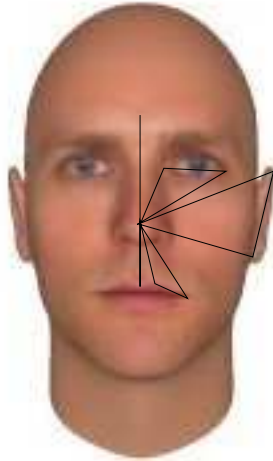
## Facial identification

Facial identification

- CCTV and recording of offences offer opportunity to directly convict offenders.
- CCTV provides a sequence of images for a model.
- Landmark measurements vary between individuals.
- Relatively few landmarks can be reliably located upon the human face.

OOS seminar – p.8/33

## Facial identification



OOS seminar – p.9/33

## Likelihood ratio calculation

Least squares estimate of the within item component of covariance:

- $\hat{U} = S_w / (N - m)$
- $S_w = \sum_{i=1}^m \sum_{j=1}^n (x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)^T$

Least squares estimate of the between item component of covariance:

- $\hat{C} = [S^* / (m - 1)] - (\hat{U} / n)$
- $S^* = \sum_{i=1}^m (\bar{x}_i - \bar{x}_{..})(\bar{x}_i - \bar{x}_{..})^T$

OOS seminar – p.11/33

## Likelihood ratio calculation

Essentially the same as univariate, except:

- $m$  items, so  $i = 1, \dots, m$
- $n$  replicated observations per item, so  $j = 1, \dots, n$
- $p$  variables

With means:

- $\bar{x}_{..} = 1/mn \sum_{i=1}^m \sum_{j=1}^n x_{ij}$ .
- $\bar{x}_{i.} = 1/n \sum_{j=1}^n x_{ij}$ .

OOS seminar – p.10/33

## Likelihood ratio calculation

Also for all  $p$  variables featured in the database:

- There are  $n_1$  replicated observations from the control item
  - with mean  $\bar{y}_1$
  - may be the series of measurements from an item from a crimescene, or intercepted voice formants.
- $n_2$  replicated observations from the recovered item
  - with mean  $\bar{y}_2$ .
  - might be observations from an item associated with a suspect, or, measurements taken from the formants of a tape of that suspect.

OOS seminar – p.12/33

## Likelihood ratio calculation

Let:

$$\alpha_1 = (2\pi)^{-p} |D_1|^{-1/2} |D_2|^{-1/2} |C|^{-1/2}$$

$$\alpha_2 = (mh^p)^{-1} |D_1^{-1} + D_2^{-1} + (h^2C)^{-1}|^{-1/2}$$

$$\alpha_3 = \exp \left\{ -\frac{1}{2} (\bar{y}_1 - \bar{y}_2)^T (D_1 + D_2)^{-1} (\bar{y}_1 - \bar{y}_2) \right\}$$

where:

$$D_1 = n_1^{-1} \hat{U}$$

$$D_2 = n_2^{-1} \hat{U}$$

OOS seminar – p.13/33

## Likelihood ratio calculation

and:

$$h = [4/(2p + 1)]^{1/(p+4)} m^{-1/(p+4)}$$

$$\beta_1 = (2\pi)^{-p} |C|^{-1} (mh^p)^{-2}$$

$$\gamma_1 = |D_1|^{-1/2} |D_1^{-1} + (h^2C)^{-1}|^{-1/2}$$

$$\gamma_2 = |D_2|^{-1/2} |D_2^{-1} + (h^2C)^{-1}|^{-1/2}$$

$$y^* = (D_1^{-1} + D_2^{-1})^{-1} (D_1^{-1} \bar{y}_1 + D_2^{-1} \bar{y}_2)$$

OOS seminar – p.14/33

## Likelihood ratio calculation

Then a suitable likelihood ratio may be calculated:

$$LR = \frac{\alpha_1 \alpha_2 \alpha_3 \sum_{i=1}^m \exp \left\{ -\frac{1}{2} (y^* - \bar{x}_{i.})^T ((D_1^{-1} + D_2^{-1})^{-1} + (h^2C)^{-1}) (y^* - \bar{x}_{i.}) \right\}}{\beta_1 \prod_{l=1}^2 [\gamma_l \sum_{i=1}^m \exp \{ -\frac{1}{2} (\bar{y}_l - \bar{x}_{i.})^T (D_l + h^2C)^{-1} (\bar{y}_l - \bar{x}_{i.}) \}]}$$

For the propositions:

1.  $H_p$  - the sets of observations  $y_2$  are drawn from the same source as those from which  $y_1$  are drawn.
2.  $H_d$  - the sets of observations  $y_2$  are drawn from some source in the population other than those from which  $y_1$  are drawn.

OOS seminar – p.15/33

## Glass data

Window	Si	K	Ca	Fe
1	53370	503	21903	1594
1	53615	427	21942	1568
1	53617	462	21580	1518
⋮	⋮	⋮	⋮	⋮
2	52346	271	36941	3240
2	52082	279	37109	3179
2	52416	301	36979	3214
⋮	⋮	⋮	⋮	⋮
62	52315	454	23234	1596
62	52168	425	23408	1676
62	52448	435	23520	1531

OOS seminar – p.16/33

## Glass data

Data are a 4 part subcomposition

- **Scale invariance** -  $y_1 = \{53370, 503, 21903, 1594\} \equiv y_2 = \{22238, 210, 9126, 664\}$  ( $\lambda = 2.4$ ) - ensure by taking ratios - either of one element or  $\sum$ .
- **Permutation invariance** - less fundamental than scale invariance
  - Any analysis must be invariant to the ordering of the variables in the composition.
  - Classical way of expunging singularity is to drop a variable - which variable?

Speech data are probably not compositional, but the face ratio data definitely are.

OOS seminar – p.17/33

## Glass data

Window	Log (Si/ $\Sigma$ )	Log (K/ $\Sigma$ )	Log (Ca/ $\Sigma$ )	Log (Fe/ $\Sigma$ )
1	0.3714	5.0358	1.262	3.8824
1	0.3691	5.2019	1.2625	3.9011
1	0.3642	5.1183	1.2743	3.9287
⋮	⋮	⋮	⋮	⋮
2	0.5725	5.8361	0.9211	3.3549
2	0.576	5.8054	0.915	3.3723
2	0.5724	5.7323	0.9213	3.3641
⋮	⋮	⋮	⋮	⋮
61	0.3593	6.7835	1.2808	3.7776
61	0.3595	6.2166	1.2798	3.8217
61	0.3637	6.002	1.2753	3.7701

OOS seminar – p.19/33

## Glass data

One final condition:

- Subcompositional coherence - the distances between two sub-compositions must be less than, or equal to, the distance for the full compositions.
- Euclidean distances do not guarantee this.
- Take logs which does.

The short answer to the problem of compositional data is to work with log ratio data, rather than the relative intensities of which the variables are composed.

OOS seminar – p.18/33

## Demonstration programs

These programs written in R

- R is an open source implementation of the S language specification.
- Open source good because it is available and consequently most innovations in statistical programming are tried in R first.
- R is available for Windows, Mackintosh and most Unix machines.

The disadvantage is that it can be seen as difficult to use - but really it isn't!

OOS seminar – p.20/33

## Demonstration programs

---

Graphical programs good, but limited.

- Need a more general framework.
- More flexible - can select different sets of variables.

Command line R is solution for the moment

OOS seminar – p.21/33

## Demonstration programs

---

```
require(Hmisc)
require(nlme)

# load up other functions
source("two-level-MVRA-functions.r")

# read in data
dat <- read.table("../data/Buscaglia-glass.txt", header=TRUE)

# these are constants - do not change
item.column <- 1
variable.labels <- c("Log Si", "Log K", "Log Ca", "Log Fe")
```

OOS seminar – p.23/33

## Demonstration programs

---

Open the file `command-line-demo.r` in a text editor

- Possibly TinnR for Windows.
- Notepad
- WordPad

From within R run the file `command-line-demo.r` in the same way that `graphical-demo.r` was run.

OOS seminar – p.22/33

## Demonstration programs

---

```
# can be changed if desired
control.indicies <- c(1,2)
recovered.indicies <- c(3,4,5)

# keep these between 2 and 5 for JoAnne's data
data.columns <- c(2,3,4)

# calculate the means and covariances for the population
UC <- two.level.U.C(dat, data.columns, item.column, REML=TRUE)

# select the control and recovered items change
# at will between 1 and 62
control.item <- 2
recovered.item <- 1
```

OOS seminar – p.24/33

## Demonstration programs

---

```
# extract the control and recovered items from the data and
calculate
# the appropriate means and number of replicates
control.inds <- which(dat[,item.column] ==
control.item)[control.indicies]
recovered.inds <- which(dat[,item.column] ==
recovered.item)[recovered.indicies]

control.bar <- two.level.y.bar(dat[control.inds,], data.columns)
recovered.bar <- two.level.y.bar(dat[recovered.inds,],
data.columns)
```

OOS seminar – p.25/33

## Demonstration programs

---

```
# calculate the smoothing parameter
h.opt <- calculate.h.opt(UC$n.items, length(data.columns))

# finally calculate the likelihood ratio for the comparison
LR <- two.level.density.LR(control.bar$y.bar, recovered.bar$y.bar,
control.bar$n.cs, recovered.bar$n.cs, UC$U, UC$C, UC$x.bar.i,
h.opt)

print(round(LR,4))
```

OOS seminar – p.26/33

## Demonstration programs

---

Try:

- Comparing likelihood ratios for differing combinations of variables for the same item.
- Notice how the likelihood varies with increasing numbers of variables if the control and recovered items are the same.

Using these functions this way is similar to how, at the moment, one would perform these calculations in case work.

OOS seminar – p.27/33

## Graphical models

---

In some instances:

- Data are highly multivariate (> 10 variables).
- Too few observations - for 10 variables really need many hundreds in background.

Solution - make more observations, or, break the problem down.

OOS seminar – p.28/33

## Graphical models

---

A useful tool is:

- The package MIM (multivariate interactive modelling).
- <http://www.hypergraph.dk/>
- Edwards, D. (1995) *Introduction to Graphical Modelling*, Springer.
- Work using the means of the observations for each item.
- Find independent sets, called *cliques*.

Can treat likelihood ratio of full set of variables as the product of the likelihood ratios of all the cliques divided by their separators.

OOS seminar – p.29/33

## Graphical models

---

Cliques - {Ca, K, Si} and {Fe, Si}.

- Running union is {Ca, K, Si}
- new clique is {Fe, Si}
- Separator is union of these, is Si.

The factorisation  $\Gamma$  is:

$$\Gamma = \frac{f(Ca, K, Si) f(Fe, Si)}{f(Si)}$$

OOS seminar – p.31/33

## Graphical models

---

Cliques - {Ca, K, Si} and {Fe, Si}.

- The set chain is an ordering of the cliques to guarantee full factorisation of the model.
- As we have only two cliques they are automatically in their set chain - for more complicated models the rules are simple.

OOS seminar – p.30/33

## Graphical models

---

Cliques - {Fe, Si} and {Ca, K, Si}.

- Running union is {Fe, Si}
- new clique is {Ca, K, Si}
- Separator is union of these, is Si.

The factorisation  $\Gamma$  is again:

$$\Gamma = \frac{f(Ca, K, Si) f(Fe, Si)}{f(Si)}$$

OOS seminar – p.32/33

# Graphical models

---

If we can find the likelihood ratios for the components:

- $\alpha_1 = f(\text{Ca}, \text{K}, \text{Si})$
- $\alpha_2 = f(\text{Fe}, \text{Si})$
- $\alpha_3 = f(\text{Si})$

$$\text{LR} = \frac{\alpha_1 \alpha_2}{\alpha_3}$$

This is illustrated in `graphical-model-demo.r`