Notes on correspondence Analysis

Chris Davey

24/12/2019

Overview

Correspondence analysis can be a useful data exploration tool when investigating the relationship between two types of factors. Multiple correspondence analysis provides a similar capability for multiple sets of factors, however this note will only review correspondence analysis. These sets of scripts are for my own learning about correspondence analysis. For more robust and accurate tools, the R package FactorMineR should be leveraged for this type of analysis (http://factominer.free.fr).

Data Set.

The data set used for this exploration is from the package NADA, called HgFish, this provides the mercury concentration in fish accross the united states based on land use (see HgFish in the NADA package). The property Hg is converted into a factor variable given a set of 15 levels, this is a simplistic method of discretising the continuous measure for mercury, and is based on the histogram.

```
require(NADA)
source("lib/ca_tools.R")

data(HgFish)

summary(HgFish)
```

```
##  LandUse                Species       Weight           Length
##  AF  :19   LargemouthBass   :44   Min.   :  11.0   Min.   : 40.0
##  Ag  :42   SmallmouthBass   :31   1st Qu.:  94.0   1st Qu.:188.0
##  Bkg :23   Sunfish          : 9   Median : 221.0   Median :253.0
##  Mine:15   SpottedBass      : 7   Mean   : 320.4   Mean   :258.1
##  Urb :34   MountainWhitefish: 6   3rd Qu.: 453.0   3rd Qu.:324.0
##            Pickerel         : 6   Max.   :1694.0   Max.   :550.0
##            (Other)          :30
##        Hg           HgCen            WatMeHg          WatTotHg
##  Min.   :0.0300   Mode :logical   Min.   :0.0060   Min.   :  0.270
##  1st Qu.:0.1170   FALSE:118       1st Qu.:0.0300   1st Qu.:  1.450
##  Median :0.2360   TRUE :15        Median :0.0610   Median :  2.375
##  Mean   :0.3584                   Mean   :0.1315   Mean   :  9.470
##  3rd Qu.:0.3730                   3rd Qu.:0.1295   3rd Qu.:  4.613
##  Max.   :4.2200                   Max.   :1.4810   Max.   :655.635
##                                   NA's   :6        NA's   :7
##     SedMeHg          SedTotHg          WatDOC          SedLOI
##  Min.   : 0.043   Min.   :   1.85   Min.   : 0.600   Min.   :0.002643
##  1st Qu.: 0.233   1st Qu.:  19.51   1st Qu.: 2.400   1st Qu.:0.021800
##  Median : 0.525   Median :  41.40   Median : 3.500   Median :0.034500
##  Mean   : 1.630   Mean   : 147.37   Mean   : 4.512   Mean   :0.111131
##  3rd Qu.: 2.665   3rd Qu.: 128.50   3rd Qu.: 5.022   3rd Qu.:0.078700
##  Max.   :10.851   Max.   :2477.50   Max.   :24.000   Max.   :0.920000
##                                     NA's   :5
##      SedAVS          PctWetland
```

```
##  Min.   :   0.00   Min.   : 0.000
##  1st Qu.:   3.00   1st Qu.: 0.000
##  Median :  48.38   Median : 1.000
##  Mean   : 245.66   Mean   : 9.032
##  3rd Qu.: 202.45   3rd Qu.: 4.000
##  Max.   :2633.00   Max.   :95.000
##  NA's   :13
```

```
data <- HgFish

## To be able to associate mercury levels we will generate 15 cuts.
hg_hist <- hist(HgFish$Hg, breaks=15, plot=FALSE)

hg_labels <- make_break_labels("Hg", hg_hist$breaks)

data$HgBand <- cut(HgFish$Hg, breaks=hg_hist$breaks, labels=hg_labels, include.lowest=TRUE)
```

The resulting properties of interest will be the LandUse variable and the HgBand variable from the data set.

```
knitr::kable(head(data[,c("LandUse", "HgBand")]))
```

LandUse	HgBand
Urb	Hgx0x0.2
Urb	Hgx0x0.2
Urb	Hgx0.2x0.4
Urb	Hgx0.2x0.4
Urb	Hgx0.2x0.4
Urb	Hgx0x0.2

Method

The method for obtaining the correspondence matrix and the associated chisq distances, as well as principle and standard coordinates is derived from chapter 17 of "Modern Multivariate Statistical Techniques" by Alan Izenman Izenman (2008).

The approach to performing the correspondence analysis converts each record into two matrices of one-hot encoded records for each factor X for a factor representing rows, and Y for a factor representing columns. For example a sample of the X matrix for Land Use.

Note that the notation here differs slightly from that used in Izenman (2008) where the matrices X and Y have columns corresponding to the variable indicator whereas in the text, the matrices X and Y had rows corresponding to the variable indicator. Hence the matrices X and Y in this example are the transpose of the matrices in the text. The notation has been adjusted in further notes below.

```
X <- model.matrix(~ LandUse -1, data)
knitr::kable(as.data.frame(X[5:15,]))
```

	LandUseAF	LandUseAg	LandUseBkg	LandUseMine	LandUseUrb
5	0	0	0	0	1
6	0	0	0	0	1
7	0	0	0	0	1
8	0	0	0	0	1
9	0	0	1	0	0
10	0	0	1	0	0

	LandUseAF	LandUseAg	LandUseBkg	LandUseMine	LandUseUrb
11	0	0	1	0	0
12	0	0	1	0	0
13	0	0	1	0	0
14	0	0	0	1	0
15	0	0	0	0	1
and	an example o	f the Y matr	ix for HgBand		

```
Y <- model.matrix(~ HgBand -1, data)
knitr::kable(as.data.frame(Y[5:15,]))
```

	HgBandHgx0x0.2	HgBandHgx0.2x0.4	HgBandHgx0.4x0.6	HgBandHgx0.6x0.8	HgBandHgx0.8x1	HgBandHgx
5	0	1	0	0	0	
6	1	0	0	0	0	
7	1	0	0	0	0	
8	1	0	0	0	0	
9	0	1	0	0	0	
10	1	0	0	0	0	
11	0	1	0	0	0	
12	0	1	0	0	0	
13	1	0	0	0	0	
14	1	0	0	0	0	
15	1	0	0	0	0	

Having obtained two separate matrices of indicator variables for rows X and columns Y we can then calculate the joint frequencies resulting in the observed cell frequencies table. Since the rows of X represent each observation we transpose X to obtain an $r \times s$ matrix for the joint frequency matrix N.

$$N = X'Y$$

The matrix N contains the frequency where the corresponding occurance of an indicator in X occurs with an indicator in Y. In this data set it is the count of the frequency where a Land Use category in X occurs in correspondence with a mercury level band from Y.

Also define the diagonal matrices

$$D_r = n^{-1}X'X$$
$$D_c = n^{-1}Y'Y$$

To represent the marginal frequencies of rows D_r and columns D_c respectively. Where n is the sum of the frequency table (the total number of samples in all categories) $n = \sum_{i=1}^{r}\sum_{j=1}^{s} N_{ij}$.

These matrices are combined to form the Burt matrix which is an analogue of the sample covariance matrix for the pair of discrete variables Izenman (2008).

$$\begin{pmatrix} nD_r & N \\ N' & nD_c \end{pmatrix}$$

The joint density estimate $P(A \wedge B)$ (where A represents row variables and B respresents column variables) can be derived from the frequency matrix N simply by dividing through by n.

$$P = n^{-1}N$$

The marginal density estimates can be given by summing either rows or columns of P or alternately, are given by the diagonals of D_r and D_c respectively.

The conditional density estimates for row profiles given column profiles $P(A|B)$ the likelihood of row attribute occuring given the presence of the column attribute, P_c are estimated as:

$$P_c = D_c^{-1} P$$

Similarly the conditional density estimates for column profiles given row profiles $P(B|A)$ the likelihood of a column attribute occuring given a row attribute occuring P_r are given as

$$P_r = D_r^{-1} P$$

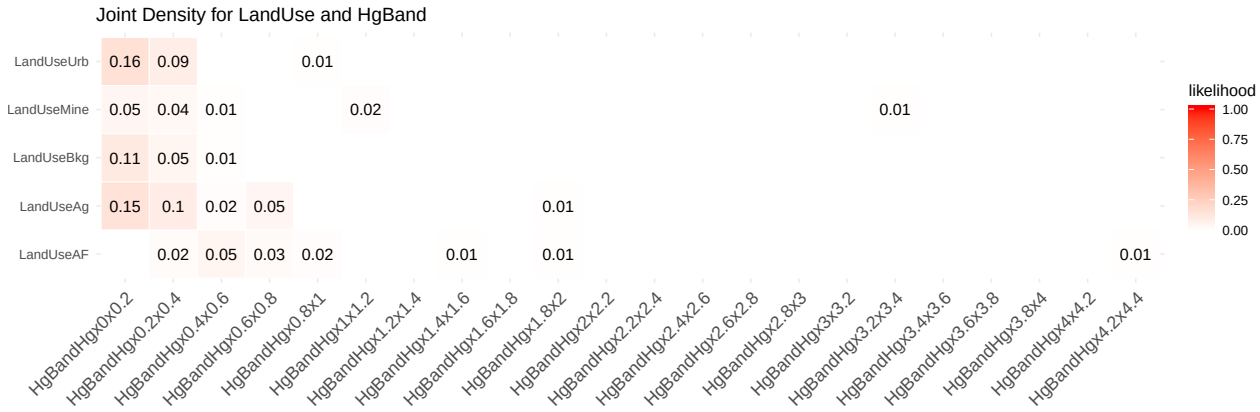
The matrices P, P_r and P_c are uniformly minimum variane unbiased estimators for their respective densities Izenman (2008). Pr is known as the row profiles and P_c the column profiles.

In the example data set the set of matrices are calculated using the convenience method from the ca_tools.R script "compute_correspondence_tables".

```
mat <- compute_correspondence_tables(data, "LandUse", "HgBand")
```

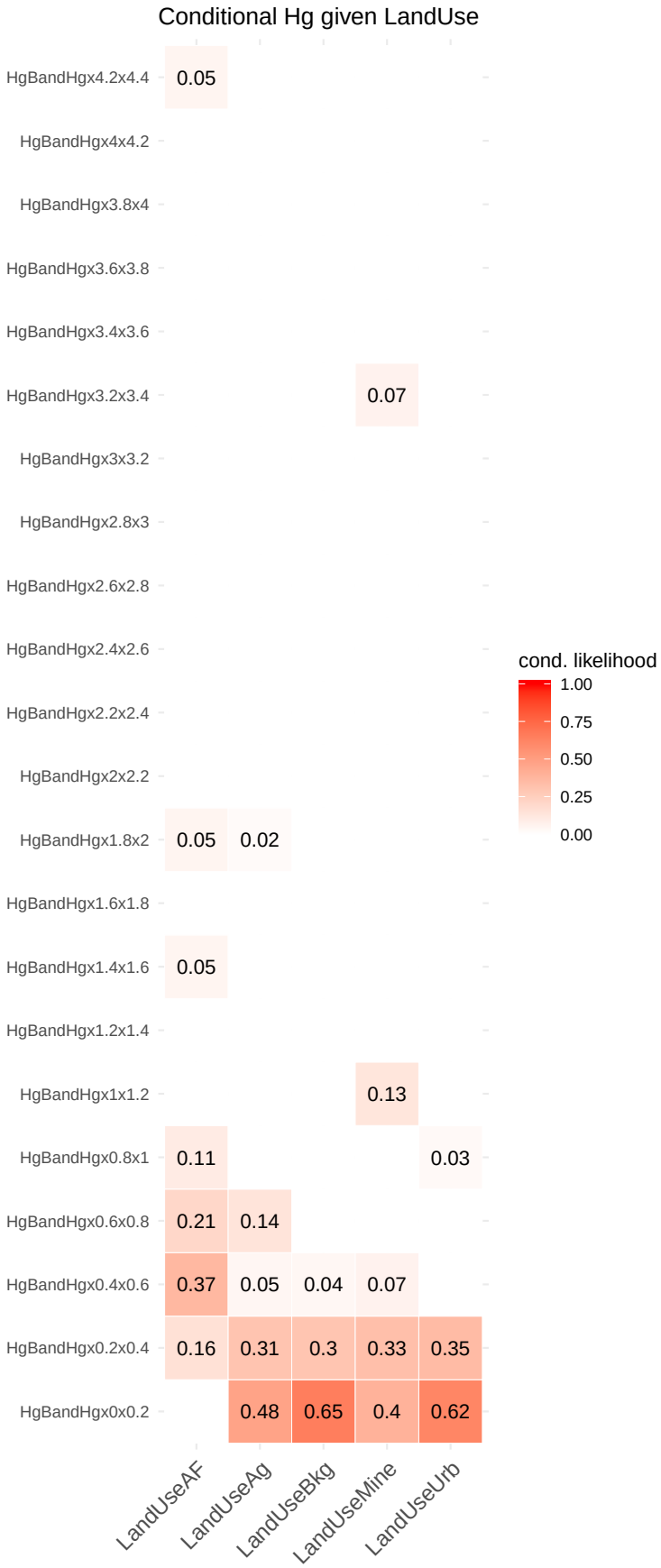
The joint density is shown below.

```
display_table(mat$P_joint, "Joint Density for LandUse and HgBand", "likelihood")
```



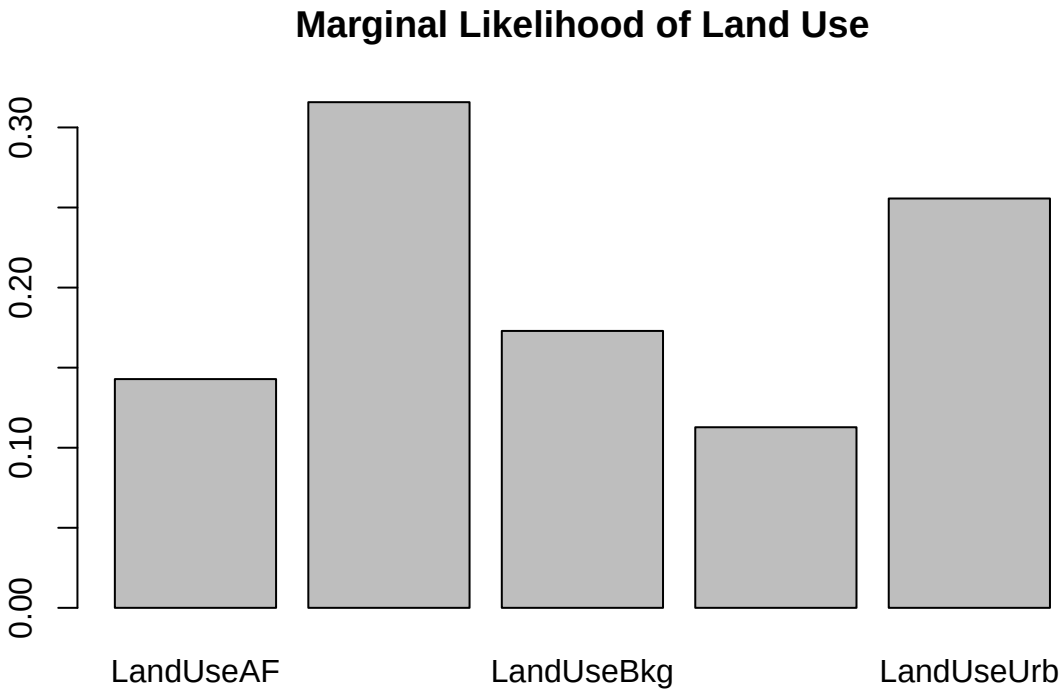
The conditional density estimate for HgBand given LandUse is shown below.

```
display_table(t(mat$P_r), "Conditional Hg given LandUse", "cond. likelihood")
```


Conditional Hg given LandUse
HgBandHgx4.2x4.4
HgBandHgx4x4.2
HgBandHgx3.8x4
HgBandHgx3.6x3.8
HgBandHgx3.4x3.6
HgBandHgx3.2x3.4
HgBandHgx3x3.2
HgBandHgx2.8x3
HgBandHgx2.6x2.8
HgBandHgx2.4x2.6
HgBandHgx2.2x2.4
HgBandHgx2x2.2
HgBandHgx1.8x2
HgBandHgx1.6x1.8
HgBandHgx1.4x1.6
HgBandHgx1.2x1.4
HgBandHgx1x1.2
HgBandHgx0.8x1
HgBandHgx0.6x0.8
HgBandHgx0.4x0.6
HgBandHgx0.2x0.4
HgBandHgx0x0.2
0.05
0.07
0.05 0.02
0.05
0.13
0.11 0.03
0.21 0.14
0.37 0.05 0.04 0.07
0.16 0.31 0.3 0.33 0.35
0.48 0.65 0.4 0.62
LandUseAF
LandUseAg
LandUseBkg
LandUseMine
LandUseUrb
cond. likelihood
1.00
0.75
0.50
0.25
0.00

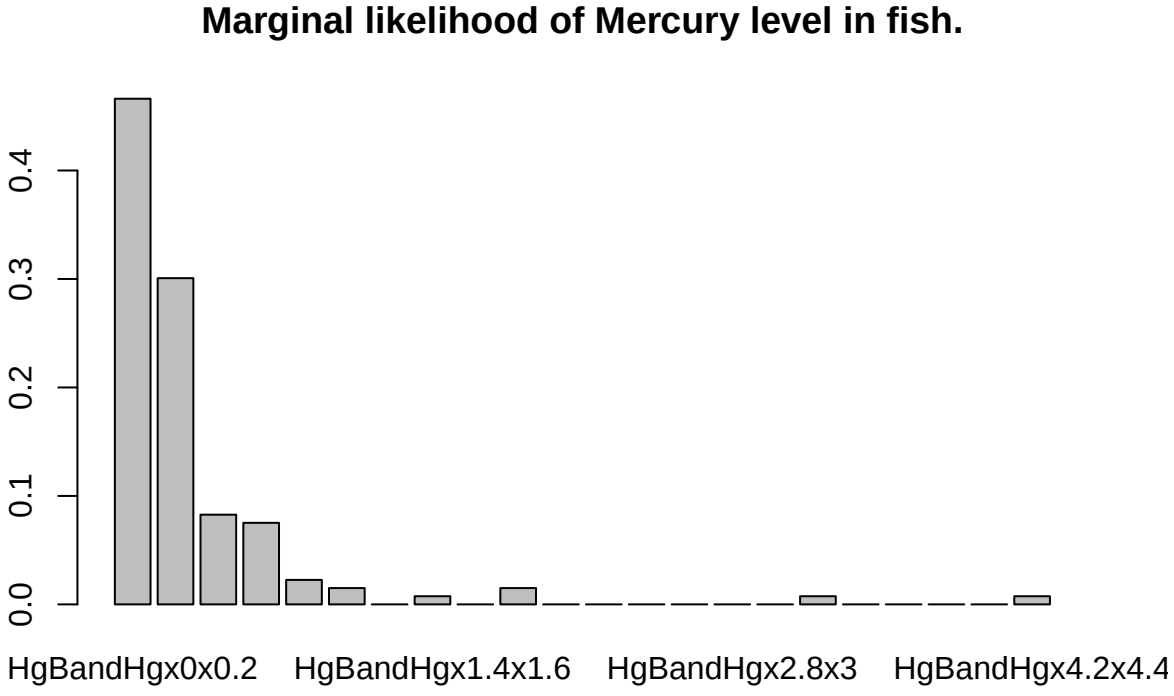
The row and column margin densities are shown as follows.

```
# row marginal probabilities
barplot(mat$P_row_margins, main="Marginal Likelihood of Land Use")
```



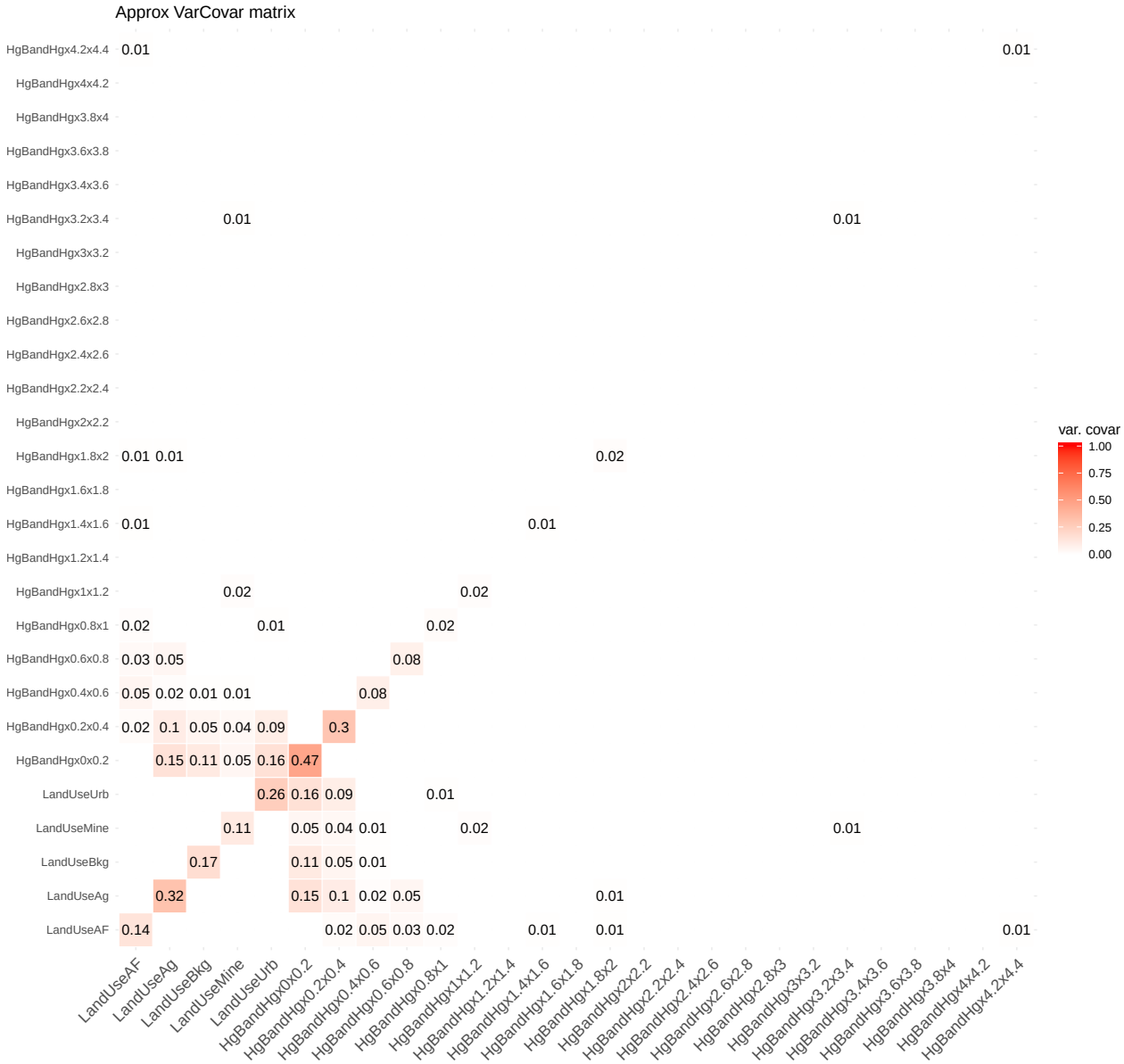
The marginal likelihood r is given by the diagonal of the matrix D_r, $r = diag(D_r)$ and likewise for the column marginal likelihood $c = diag(D_c)$.

```
# column marginal probabilities
barplot(mat$P_col_margins, main="Marginal likelihood of Mercury level in fish.")
```



The burt matrix divided by n gives approximation of the variance covariance matrix, shown below.

```
# Burt matrix analogue of variance-covariance matrix for discrete data.
display_table(mat$var_covar_mat, "Approx VarCovar matrix", "var. covar")
```



Testing Independence.

A chisq test with $(r-1)(s-1)$ degrees of freedom, where r represents the number of rows in the frequency matrix N and s the number of columns, can be used to test the hypothesis for independence between rows and columns.

There are two distance matrices required for the calculation of the test statistic. The first being the distance between ith and jth column profiles.

$$d^2(a_i, a_j) = (a_i - a_j)' D_c^{-1} (a_i - a_j)$$

and the distance from the row centroid c

$$d^2(a_i, c) = (a_i - c)' D_c^{-1} (a_i - c)$$

The first distance matrix provides larger distances for those rows that contribute more to the overall distance measure.

The second matrix is used in calculation of the overall χ^2 statistic for row profiles.

$$\chi^2 = n \sum_{i=1}^{r} p_{i+} d^2(a_i, c)$$

where p_{i+} is the ith row of P_r. The chisq statistic can be used to test whether there is any evidence to suggest that the rows are not independent.

```
row_chisq_stat <- mat$row_chisq_stat
df <- (nrow(mat$P_joint) - 1) * (ncol(mat$P_joint)-1)
pchisq(row_chisq_stat, df=df, lower.tail=FALSE)
```

```
## [1] 3.287882e-89
```

In this dataset the statistic suggests there is evidence to suggest the rows are not independent.

The test pvalue is also already calculated in the result from the method.

```
mat$row_chisq_pvalue
```

```
## [1] 3.287882e-89
```

We can look to the distances between rows to determine which rows influence the statistic the most.

```
display_table(mat$row_chisq_dist, "Row Distances", "Chisq")
```

Seemingly row 1 is far from the other rows, and row 4 appears to be further from other rows as well.

The same metrics can be calculated for the column distances for the ith row b_i and jth row b_j in the column profile matrix P_c.

$$d^2(b_i, b_j) = (b_i - b_j)' D_r^{-1} (b_i - b_j)$$

The distance from the centered column profiles r is also given as

$$d^2(b_i, r) = (b_i - r)' D_r^{-1} (b_i - r)$$

The chisq statistic sums over each row j in the column profile P_c

$$\chi^2 = n \sum_{j=1} p_{+j} d^2(b_j, r)$$

Where p_{+j} is the jth row of P_c in this instance.

The statistic produced can test the null hypothesis that the columns are independent at $(r-1)(s-1)$ degrees of freedom.

```
mat$col_chisq_stat
```

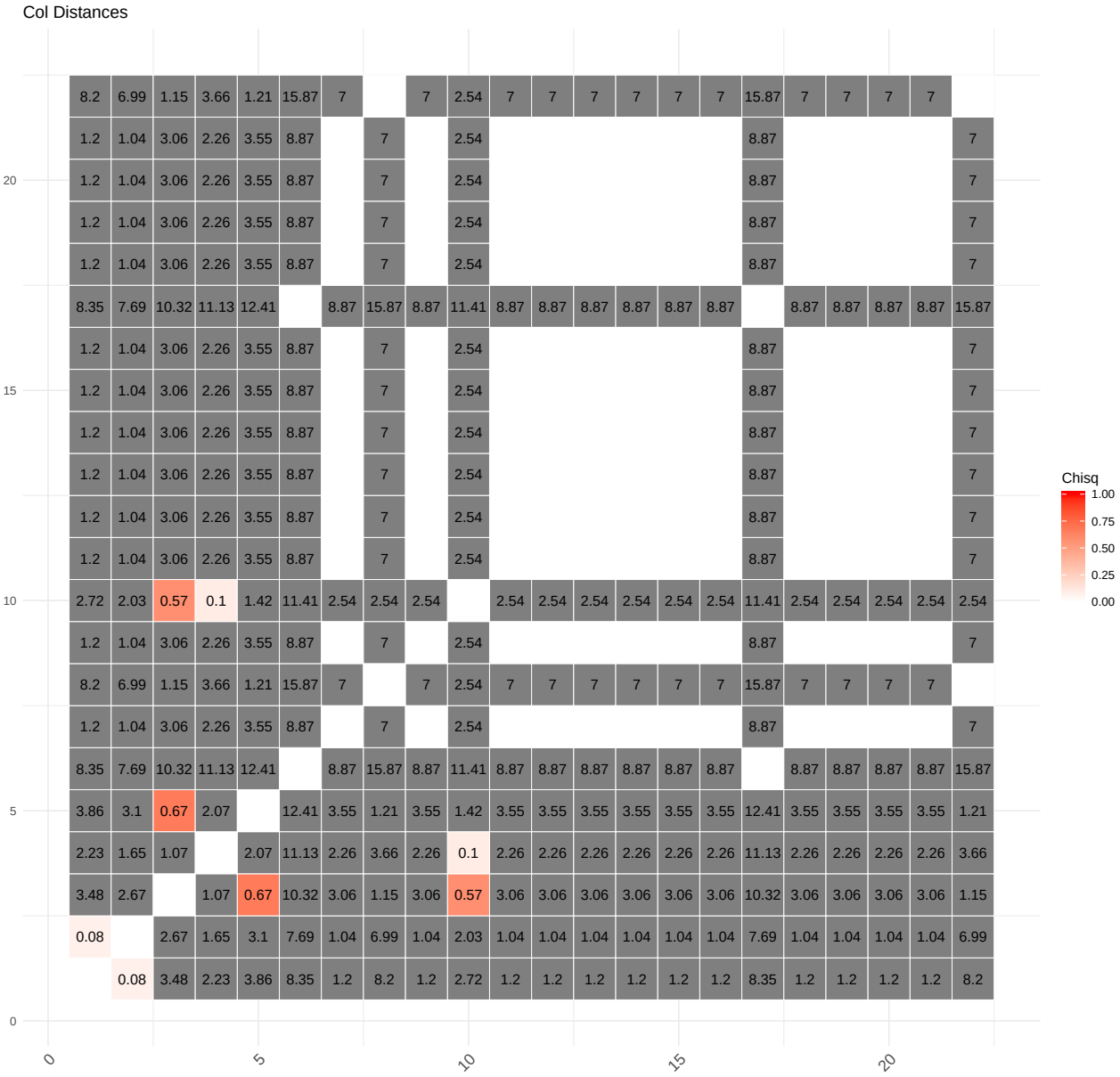
```
## [1] 4705.881
```

```
mat$col_chisq_pvalue
```

```
## [1] 0
```

In this case there is some evidence to suggest that the columns are not independent of each other.

```
display_table(mat$col_chisq_dist, "Col Distances", "Chisq")
```



In this example we can see from the column distances that there are many columns further away from each other.

Inertia

In computing the values for inertia we require the relative frequency matrix $\hat{P}$

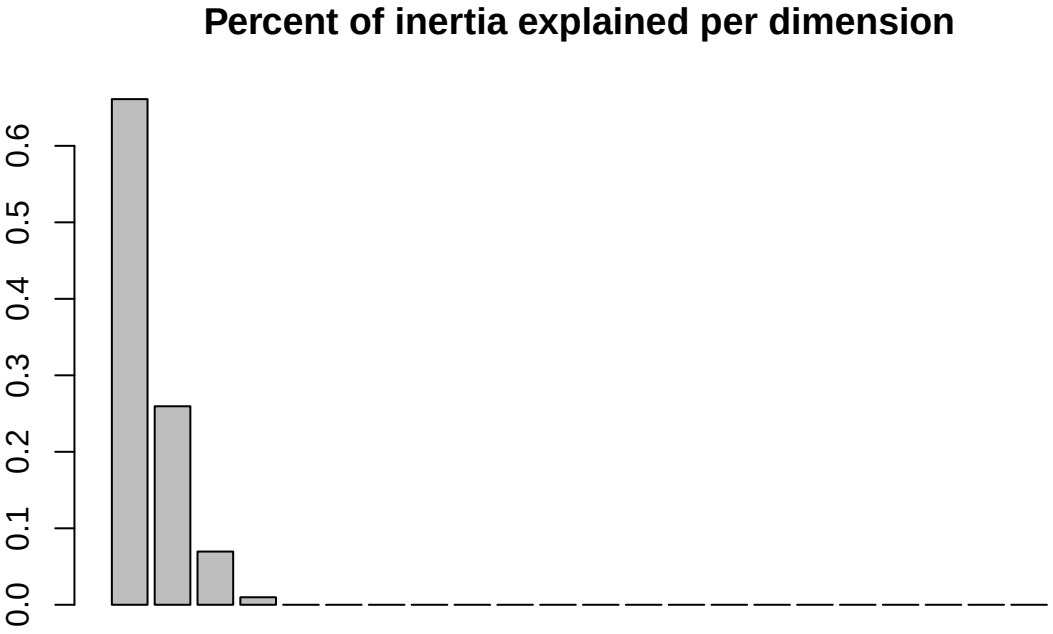
$$\hat{P} = P - rc'$$

The matrix $\hat{N} = n\hat{P}$ is referred to as the matrix of residuals because the $ijth$ entry shows the difference between observed cell frequency and expected cell frequency Izenman (2008). The total inertia is calculated from the eigendecomposition of the $(s \times s)$ matrix R_0

$$R_0 = D_c^{-1/2}\hat{P}'D_r^{-1}\hat{P}D_c^{-1/2}$$

The trace of the matrix R_0 provides the measure of total inertia in the contingency table, which is also given by the sum of the eigenvalues of R_0. The eigenvalues of R_0 indicate the principle inertias contributed by each

dimension of the decomposition, giving an analogue to the total amount of variance explained by principle component decomposition Izenman (2008).



The first three dimensions above indicate the largest contribution to the inertia of the contingency table, this also indicates that the first three dimensions contribute the most to the deviations from independence.

Coordinates for row and column profiles.

The matrix R_0 is decomposed into $R_0 = M'M$ the matrix M is calculated from

$$M = D_r^{-1/2} \hat{P} D_c^{-1/2}$$

and the singular value decomposition of this matrix is used to calculate the principle components $M = UD_\lambda V'$.

The matrices A and B further decompose the matrix $\hat{P}$ such that $\hat{P} = AD_\lambda B'$ these are given from the left and right eigenvalues as follows

$$A = D_r^{1/2} U$$
$$B = D_c^{1/2} V$$

Where A and B are the left and right hand eigenvectors resulting from the singular value decomposition of $\hat{P}$. The principle coordinates representing the squared χ^2 distances between the centered row profiles and B is given by

$$G'_p = D_r^{-1} A D_\lambda$$

and the principle coordinates representing the χ^2 distances between the centered column profiles and A is given by

$$H_p = D_c^{-1} B D_\lambda$$

The standard coordinates for row profiles are given by

$$G_s = U' D_r^{-1/2}$$

and likewise the standard coordinates for column profiles are given by

$$H_s = V' D_c^{-1/2}$$

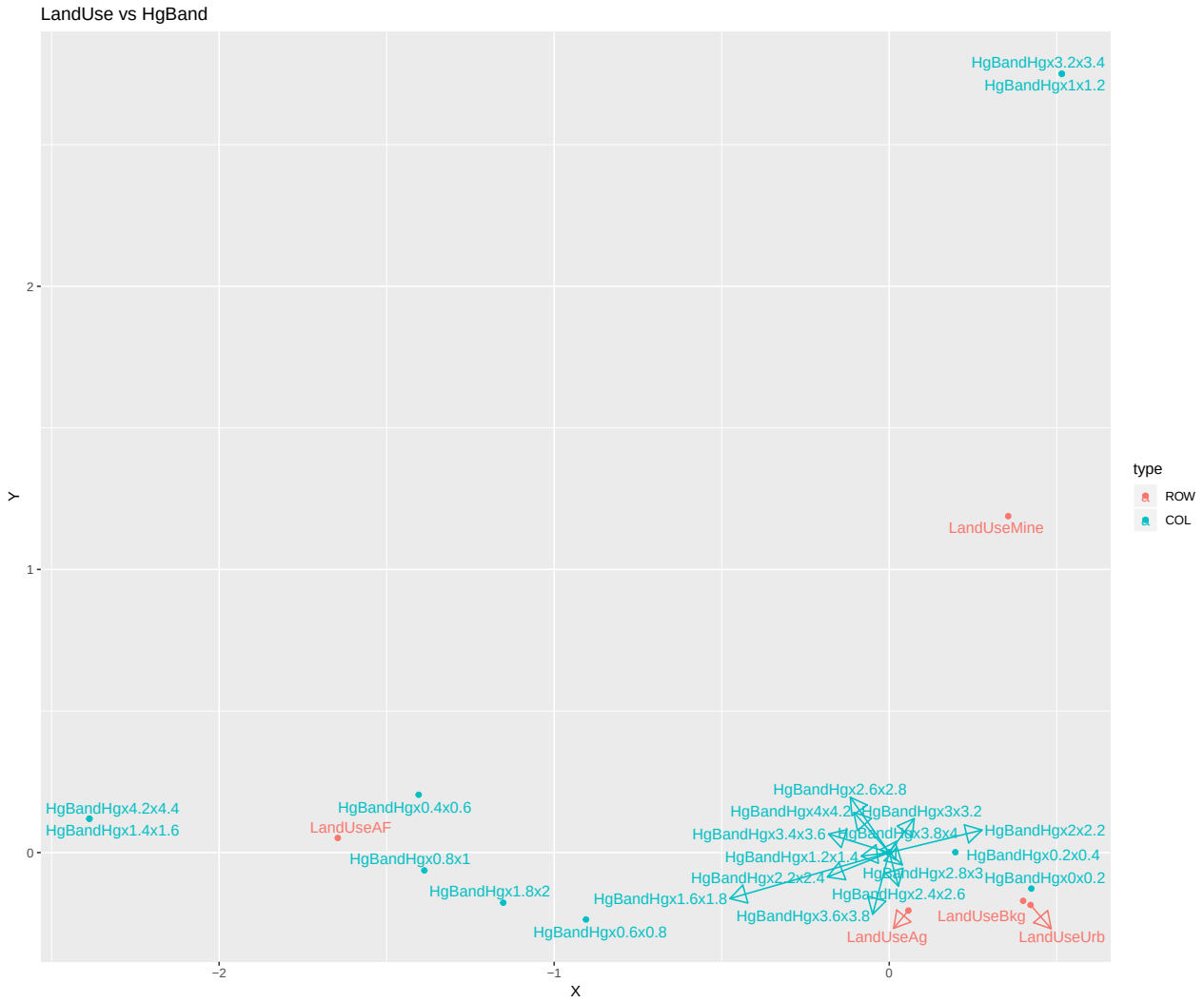
These coordinates are used in visualisation as either **Symmetric Map** where both the coordinates are taken from the principle coordinates, or as an **Asymmetric Map** where either column or row coordinates are taken from the principle coordinates and the remaining (row or column) are taken from the standard coordinates Izenman (2008).

The allows for the ordination of the relationship between the row categories and the column categories, these are known as **correspondence maps**. Points that appear close to each other indicate that the categories are closely related. The interpretation is given by Izenman Izenman (2008) as

- if row points are close, then those rows have similar conditional distributions accross columns
- if column points are close, then those columns have similar conditional distributions accross rows
- if a row point is close to a column point, then that configuration suggests a particular deviation from independence.

The visualisation of the principle coordinates for the data in our example is shown below.

```
p <- draw_profiles(mat, "LandUse vs HgBand")
print(p)
```



The mercury content bands 4.2-4.4 micrograms per gram $\mu g/g$, 1.4-1.6 Hg $\mu g/g$, 0.8-1 Hg $\mu g/g$, 0.4-0.6$\mu g/g$

and 1.8-2$\mu g/g$ appear to be related to mixed use Agriculture and Forest (AF). 0.6-0.8 Hg $\mu g/g$ appears to be between mixed use agriculture and forest (AF) and agriculture use (AG). Noteably 1-1.2 Hg $\mu g/g$ and 3.2-3.4 Hg $\mu g/g$ are closer to mining land use that the others. Low levels of mercury 0-0.2 Hg $\mu g/g$ appear to be closely related to land use Background (Bkg) and Urban (Urb). Note there is a broad cluster of data points close to the land use for Agriculture, these appear to be in the mid-ranges for mercury usage between 1.2 up to 4.2 Hg $\mu g/g$.

Summary

The correspondence analysis provides unbiased estimators for joint and conditional densities, and is able to test for independence of rows and columns. In addition the principle coordinates produced provide an ordination that can be useful in visualising correspondences between row and column categories and along with the measures of inertia can aid in the interpretation of the data.

References

Izenman, Alan J. 2008. *Modern Multivariate Statistical Techniques: Regression, Classification and Manifold Learning*. Springer Series in Statistics. Springer New York Inc.