An Introduction to Correspondence Analysis

Phillip M. Yelland

Cross tabulations (also known as *cross tabs*, or *contingency tables*) often arise in data analysis, whenever data can be placed into two distinct sets of categories. In market research, for example, we might categorize purchases of a range of products made at selected locations; or in medical testing, we might record adverse drug reactions according to symptoms and whether the patient received the standard or placebo treatment.

The statistical technique presented in this article, *correspondence analysis*, provides a means of graphically representing the structure of cross tabulations so as to shed light on the underlying mechanisms. The article provides a practical introduction to correspondence analysis in the form of a "five-finger exercise" in textual analysis—identifying the author of a text given samples of the works of likely candidates.

■ 1. Introduction

Correspondence analysis is a statistical technique that provides a graphical representation of *cross tabulations* (which are also known as *cross tabs*, or *contingency tables*). Cross tabulations arise whenever it is possible to place events into two or more different sets of categories, such as product and location for purchases in market research or symptom and treatment in medical testing. This article provides a brief introduction to correspondence analysis in the form of an exercise in *textual analysis*—identifying the author of a text based on examination of its characteristics. The exercise is carried out using *Mathematica* (Version 5.2).

Perhaps the most illustrious exponent of textual analysis is the self-styled "literary detective" Donald Foster, whose 2001 book [1] describes how he identified the authors of several anonymous works, including the best-selling *roman-à-clef Primary Colors* [2], which satirized the 1992 Clinton presidential campaign. Foster's methodology examines a broad spectrum of text characteristics, including word choice, punctuation, grammatical structure, and the like. The aim of the exercise in this article is to emulate Foster, though naturally the literary aspects of the approach taken are much more basic—the intent is not to describe a realistic method of textual analysis, but rather to use it to illustrate correspondence analysis.

Consider the following list of seventeenth- and eighteenth-century writers.

```
authors = {"Charles Darwin", "Rene Descartes",
    "Thomas Hobbes", "Mary Shelley", "Mark Twain"};
```

Imagine that we are given two fragments of text written by one or two of these writers, and charged with identifying the true author(s). To make things interesting, imagine also that the only information we are given about an unidentified fragment of text is the frequency with which certain letters appear in it. Accordingly, I have taken three distinct samples of about 1000 characters each from the writings of each these authors, and added up the number of times each of the following characters appears in each of the samples (restricting ourselves to less than the complete alphabet prevents the tables in the rest of the discussion from becoming unwieldy; the characters chosen happen to occur with middling frequency in all the texts as a whole).

```
chars = {"B", "C", "D", "F", "G", "H", "I", "L", "M",
    "N", "P", "R", "S", "U", "W", "Y"};
```

This is the cross tabulation.

```
sampleCrosstab =
  {{34, 37, 44, 27, 19, 39, 74, 44, 27, 61, 12, 65, 69,
     22, 14, 21}, {18, 33, 47, 24, 14, 38, 66, 41, 36,
     72, 15, 62, 63, 31, 12, 18},
   {32, 43, 36, 12, 21, 51, 75, 33, 23, 60, 24, 68, 85,
     18, 13, 14}, {13, 31, 55, 29, 15, 62, 74, 43, 28,
     73, 8, 59, 54, 32, 19, 20},
   {8, 28, 34, 24, 17, 68, 75, 34, 25, 70, 16, 56, 72,
     31, 14, 11}, {9, 34, 43, 25, 18, 68, 84, 25, 32, 76,
     14, 69, 64, 27, 11, 18},
   {15, 20, 28, 18, 19, 65, 82, 34, 29, 89, 11, 47, 74,
     18, 22, 17}, {18, 14, 40, 25, 21, 60, 70, 15, 37,
     80, 15, 65, 68, 21, 25, 9},
   {19, 18, 41, 26, 19, 58, 64, 18, 38, 78, 15, 65, 72,
     20, 20, 11}, {13, 29, 49, 31, 16, 61, 73, 36, 29,
     69, 13, 63, 58, 18, 20, 25},
   {17, 34, 43, 29, 14, 62, 64, 26, 26, 71, 26, 78, 64,
     21, 18, 12}, {13, 22, 43, 16, 11, 70, 68, 46, 35,
     57, 30, 71, 57, 19, 22, 20},
   {16, 18, 56, 13, 27, 67, 61, 43, 20, 63, 14, 43, 67,
     34, 41, 23}, {15, 21, 66, 21, 19, 50, 62, 50, 24,
     68, 14, 40, 58, 31, 36, 26},
   {19, 17, 70, 12, 28, 53, 72, 39, 22, 71, 11, 40, 67,
     25, 41, 17}};
TableForm[sampleCrosstab,
 TableHeadings →
  {Flatten[Table[# <> ": " <> ToString@i, {i, 3}] & /@ authors],
    chars}, TableSpacing → .5]
```

	B	C	D	F	G	H	I	L	M	N	P	R	S	U	W	Y
Charles Darwin: 1	34	37	44	27	19	39	74	44	27	61	12	65	69	22	14	21
Charles Darwin: 2	18	33	47	24	14	38	66	41	36	72	15	62	63	31	12	18
Charles Darwin: 3	32	43	36	12	21	51	75	33	23	60	24	68	85	18	13	14
Rene Descartes: 1	13	31	55	29	15	62	74	43	28	73	8	59	54	32	19	20
Rene Descartes: 2	8	28	34	24	17	68	75	34	25	70	16	56	72	31	14	11
Rene Descartes: 3	9	34	43	25	18	68	84	25	32	76	14	69	64	27	11	18
Thomas Hobbes: 1	15	20	28	18	19	65	82	34	29	89	11	47	74	18	22	17
Thomas Hobbes: 2	18	14	40	25	21	60	70	15	37	80	15	65	68	21	25	9
Thomas Hobbes: 3	19	18	41	26	19	58	64	18	38	78	15	65	72	20	20	11
Mary Shelley: 1	13	29	49	31	16	61	73	36	29	69	13	63	58	18	20	25
Mary Shelley: 2	17	34	43	29	14	62	64	26	26	71	26	78	64	21	18	12
Mary Shelley: 3	13	22	43	16	11	70	68	46	35	57	30	71	57	19	22	20
Mark Twain: 1	16	18	56	13	27	67	61	43	20	63	14	43	67	34	41	23
Mark Twain: 2	15	21	66	21	19	50	62	50	24	68	14	40	58	31	36	26
Mark Twain: 3	19	17	70	12	28	53	72	39	22	71	11	40	67	25	41	17

■ 2. χ^2 Calculations

Is it possible to say with reasonable certainty that the distribution of letters differs significantly from sample to sample (i.e., from row to row in the cross tab)? The usual means of answering such questions is Pearson's χ^2 test for independence; it tests whether a cross tab deviates significantly from one in which rows and columns are independent. In our case, independence would imply that the letters occur with the same frequency in all of the text samples.

Assume that the cross tab under examination is described formally by the $I \times J$ matrix $F = [f_{ij}]$. We derive the *correspondence matrix P* from F by dividing its entries by their grand total:

$$P = [p_{ij}] = \left[\frac{f_{ij}}{n}\right], \ \text{where } n = \sum_{i=1}^{I}\sum_{j=1}^{J} f_{ij}. \tag{1}$$

Next, define row and column totals:

$$\begin{aligned} p_{i+} &= \sum_{j=1}^{J} p_{ij} \\ p_{+j} &= \sum_{i=1}^{I} p_{ij}, \end{aligned} \tag{2}$$

The χ^2 statistic, X^2, is calculated:

$$X^2 = n \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left(p_{ij} - \mu_{ij}\right)^2}{\mu_{ij}}. \tag{3}$$

Here μ_{ij} is an estimate of an entry's value assuming independence:

$$\mu_{ij} = p_{i+}\, p_{+j}. \tag{4}$$

For our example, the calculations may be expressed in *Mathematica* as follows.

```
grandTotal = Total[sampleCrosstab, 2];
correspondenceMatrix = sampleCrosstab / grandTotal;

rowTotals = Total /@ correspondenceMatrix;
columnTotals = Total /@ Transpose@correspondenceMatrix;

independenceModel = Outer[Times, rowTotals, columnTotals];

chiSquaredStatistic =
   grandTotal ×
    Total[(correspondenceMatrix - independenceModel)^2 /
      independenceModel, 2];
N@chiSquaredStatistic
```

```
448.497
```

If rows and columns really are independent (i.e., "under the null hypothesis"), X^2 should follow a χ^2 distribution with $(I - 1) \times (J - 1)$ degrees of freedom. We can compare the actual value computed for the example cross tab with its distribution under the null hypothesis as follows.

```
nullDistribution = ChiSquareDistribution@
   Apply[Times, Dimensions@sampleCrosstab - 1];
1 - N@CDF[nullDistribution, chiSquaredStatistic]
Quantile[nullDistribution, .99]
```

```
0.
```

```
260.595
```

Thus there is (almost) no probability under the null hypothesis of observing a statistic as large as the one actually observed, and indeed only a 1% probability of seeing a value about half as large. According to the χ^2 test, therefore, there is a statistically significant difference in the distribution of letters across the samples.

Unfortunately, the χ^2 test by itself does not provide a solution to the problem of distinguishing the works of the different authors. Though it establishes that the distribution of letters differs significantly from one sample to another, it does not tell us whether the samples of one author differ from those of other authors more than they differ from each other, nor does it allow us to characterize the authors in terms of the distribution of letters in their works. Answers to these questions are provided by correspondence analysis.

■ 3. χ^2 **Distances**

For the purposes of correspondence analysis, the differences between the distributions of letters in the text samples—which you will recall are given in the rows of the cross tab— are measured by so-called χ^2 *distances*, which are weighted Euclidean distances between normalized rows (calculated by dividing row entries by their respective row total), with weights inversely proportional to the square roots of the column totals. In symbols, the χ^2 distance between row *i* and row *k* is given by the expression:

$$d_{ik} = \sqrt{\sum_{j=1}^{J} \frac{\left(p_{ij}/p_{i+} - p_{kj}/p_{k+}\right)^2}{p_{+j}}} \quad . \tag{5}$$

This computes the χ^2 distances between the text samples using the correspondence matrix and displays them in a reasonably compact table (after scaling up by 100 and rounding).

```
chisqd[row1_, row2_] :=
 Sqrt@Sum[(row1[[j]] / Total@row1 - row2[[j]] / Total@row2) ^ 2 /
    columnTotals[[j]],
   {j, Length@columnTotals}]

UpperDiagonalMatrix[f_, n_Integer ? NonNegative] :=
 Array[If[# ≤ #2, f[##], 0] &, {n, n}]

chiSquaredDistances = UpperDiagonalMatrix[
   Function[{
     i, k},
    chisqd[correspondenceMatrix[[i]],
     correspondenceMatrix[[k]]],
   Length@correspondenceMatrix];

abbreviatedAuthors =
  StringReplace[#, RegularExpression@"[[:lower:]]|\\s" →
     ""] & /@ authors;
abbreviatedSampleTitles =
  Flatten[Table[# <> ToString@i, {i, 3}] & /@
    abbreviatedAuthors];

distanceTable[uddistances_] := TableForm[
```

```
Map[
 NumberForm[#, {2, 0}, NumberPoint → ""] &,
 uddistances + Transpose@uddistances,
 {2}],
TableHeadings → {abbreviatedSampleTitles,
   abbreviatedSampleTitles}, TableSpacing → .6]
```

```
distanceTable[100. chiSquaredDistances]
```

	CD1	CD2	CD3	RD1	RD2	RD3	TH1	TH2	TH3	MS1	MS2	MS3	MT1	MT2	MT3
CD1	0	21	24	29	35	33	36	39	34	27	31	38	43	38	42
CD2	21	0	32	20	26	24	32	33	28	23	26	31	40	33	41
CD3	24	32	0	40	34	35	37	39	35	37	29	37	46	48	47
RD1	29	20	40	0	22	21	29	33	30	15	29	32	32	26	35
RD2	35	26	34	22	0	16	23	28	26	24	24	30	37	38	41
RD3	33	24	35	21	16	0	26	27	24	19	22	31	42	41	43
TH1	36	32	37	29	23	26	0	25	24	26	33	34	35	37	36
TH2	39	33	39	33	28	27	25	0	8	29	26	35	39	41	37
TH3	34	28	35	30	26	24	24	8	0	26	23	33	40	41	39
MS1	27	23	37	15	24	19	26	29	26	0	23	27	35	29	36
MS2	31	26	29	29	24	22	33	26	23	23	0	26	43	42	45
MS3	38	31	37	32	30	31	34	35	33	27	26	0	38	36	42
MT1	43	40	46	32	37	42	35	39	40	35	43	38	0	18	17
MT2	38	33	48	26	38	41	37	41	41	29	42	36	18	0	20
MT3	42	41	47	35	41	43	36	37	39	36	45	42	17	20	0

Certain characteristics of the samples can be detected in the table above. For example, it appears that the Mark Twain (MT) texts form a relatively isolated group, in that the distances from the MT samples to each other are considerably smaller than from the MT samples to those of other authors. By itself, however, the table does little to make apparent the overall pattern of the distances—something done in the next section. Before that, however, here is a little more on the nature of χ^2 distances.

As their name suggests, χ^2 distances are closely related to the χ^2 statistic of the previous section. To show how they are related, consider the "average" row—termed the *centroid* or *barycenter* in correspondence analysis—whose entries are simply the column totals:

$$z = \left[p_{+1}, \ldots, p_{+j}, \ldots, p_{+J} \right]. \tag{6}$$

From equation (5), since the row total for the centroid is 1 (by the definition of P), the χ^2 distance of row i to the centroid is:

$$d_{iz} = \sqrt{\sum_{j=1}^{J} \frac{\left(p_{ij} / p_{i+} - p_{+j} \right)^2}{p_{+j}}} \; . \tag{7}$$

Now with μ_{ij} as defined in (4):

$$d_{iz}^2 = \frac{1}{p_{i+}} \sum_{j=1}^{J} \frac{\left(p_{ij} - p_{i+} p_{+j} \right)^2}{p_{i+} p_{+j}} = \frac{1}{p_{i+}} \sum_{j=1}^{J} \frac{\left(p_{ij} - \mu_{ij} \right)^2}{\mu_{ij}}. \tag{8}$$

Drawing an analogy with the physical concept of angular inertia, correspondence analysis defines the *inertia* of a row as the product of the row total (which is referred to as the row's *mass*) and the square of its distance to the centroid, $p_{i+} d_{iz}^2$. Comparing the expression for d_{iz}^2 in (5) with definition of the χ^2 statistic in (3), it follows that the total inertia of all the rows in a contingency matrix is equal to the χ^2 statistic divided by n, a quantity known as Pearson's *mean-square contingency*, denoted ϕ^2:

$$\sum_{i=1}^{I} p_{i+} d_{iz}^2 = \frac{X^2}{n} = \phi^2. \tag{9}$$

The total inertia of a table is used to assess the quality of its graphical representation in correspondence analysis. For future reference, we can calculate ϕ^2 for our dataset.

```
phiSquared = Sum[
    rowTotals[[i]] chisqd[correspondenceMatrix[[i]], columnTotals]^
        2,
    {i, Length@correspondenceMatrix}];
N@phiSquared
```

```
0.0498662
```

■ 4. Calculating Row Scores

Correspondence analysis provides a means of representing a table of χ^2 distances in a graphical form, with rows represented by points, so that the distances between points approximate the χ^2 distances between the rows they represent.

To compute such a representation, we begin with a matrix of *standardized residuals*, which are the square roots of the terms comprising the χ^2 statistic in (3):

$$\Omega = \left[\frac{p_{ij} - \mu_{ij}}{\sqrt{\mu_{ij}}} \right]. \tag{10}$$

Next, we compute the *singular value decomposition* of Ω, which is to say that we find orthogonal matrices V and W, together with a diagonal matrix Λ, such that (with the transpose of matrix M denoted M^T, and writing I for the identity matrix):

$$\begin{aligned} \Omega &= V \Lambda W^T, \\ V V^T &= W W^T = I. \end{aligned} \tag{11}$$

The *scores* of the rows—whose interpretation we discuss later—are given by the expression:

$$R = \delta_r V \Lambda. \tag{12}$$

Here δ_r is the diagonal matrix comprising the reciprocals of the square roots of the row totals:

$$\delta_r = \begin{pmatrix} \dfrac{1}{\sqrt{p_{1+}}} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \dfrac{1}{\sqrt{p_{I+}}} \end{pmatrix}. \tag{13}$$

The scores of the rows in our sample cross tab are computed in the following (left multiplication by δ_r being more conveniently carried out in *Mathematica* by row-wise division).

```
standardizedResiduals =
  (correspondenceMatrix - independenceModel) /
   Sqrt@independenceModel;
{leftSingularMatrix, singularValuesMatrix,
   rightSingularMatrix} =
   SingularValueDecomposition[N@standardizedResiduals];
rowScores = leftSingularMatrix.singularValuesMatrix /
   Sqrt@rowTotals;
```

The row scores may be thought of as the coordinates of points in a high-dimensional space (14-dimensional, as it turns out in this case).

```
MatrixRank@rowScores
```

```
14
```

These points are arranged so that the Euclidean distance between two points is equal to the χ^2 distance between the two rows to which they correspond. To show how the χ^2 distances between the rows are reflected in their scores, the following reconstitutes the χ^2 distances in the previous section from Euclidean distances between the scores computed above. As you can see, the χ^2 distances are recovered perfectly.

```
scoreDistances = UpperDiagonalMatrix[
  Function[{
    i, k},
   Norm[rowScores[[i]] - rowScores[[k]]],
  Length@rowScores];
distanceTable[100 scoreDistances]
distanceTable[100 Chop[chiSquaredDistances - scoreDistances]]
```

	CD1	CD2	CD3	RD1	RD2	RD3	TH1	TH2	TH3	MS1	MS2	MS3	MT1	MT2	MT3
CD1	0	21	24	29	35	33	36	39	34	27	31	38	43	38	42
CD2	21	0	32	20	26	24	32	33	28	23	26	31	40	33	41
CD3	24	32	0	40	34	35	37	39	35	37	29	37	46	48	47
RD1	29	20	40	0	22	21	29	33	30	15	29	32	32	26	35
RD2	35	26	34	22	0	16	23	28	26	24	24	30	37	38	41
RD3	33	24	35	21	16	0	26	27	24	19	22	31	42	41	43
TH1	36	32	37	29	23	26	0	25	24	26	33	34	35	37	36
TH2	39	33	39	33	28	27	25	0	8	29	26	35	39	41	37
TH3	34	28	35	30	26	24	24	8	0	26	23	33	40	41	39
MS1	27	23	37	15	24	19	26	29	26	0	23	27	35	29	36
MS2	31	26	29	29	24	22	33	26	23	23	0	26	43	42	45
MS3	38	31	37	32	30	31	34	35	33	27	26	0	38	36	42
MT1	43	40	46	32	37	42	35	39	40	35	43	38	0	18	17
MT2	38	33	48	26	38	41	37	41	41	29	42	36	18	0	20
MT3	42	41	47	35	41	43	36	37	39	36	45	42	17	20	0

	CD1	CD2	CD3	RD1	RD2	RD3	TH1	TH2	TH3	MS1	MS2	MS3	MT1	MT2	MT3
CD1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CD2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CD3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RD1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RD2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RD3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TH1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TH2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TH3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MS1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MS2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MS3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MT1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MT2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MT3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

■ 5. Plotting Rows

Although the row scores faithfully reproduce the χ^2 distances between rows in the original table, as coordinates their dimensionality is far too high for them to be presented graphically. Thanks to the properties of the singular value decomposition, however, taking just the first two components of each row's score usually produces a reasonable approximation to the χ^2 distances, and yields coordinates that can be placed on a two-dimensional plot. (Below we have labeled the components "X" and "Y" to highlight their role as 2D coordinates.)
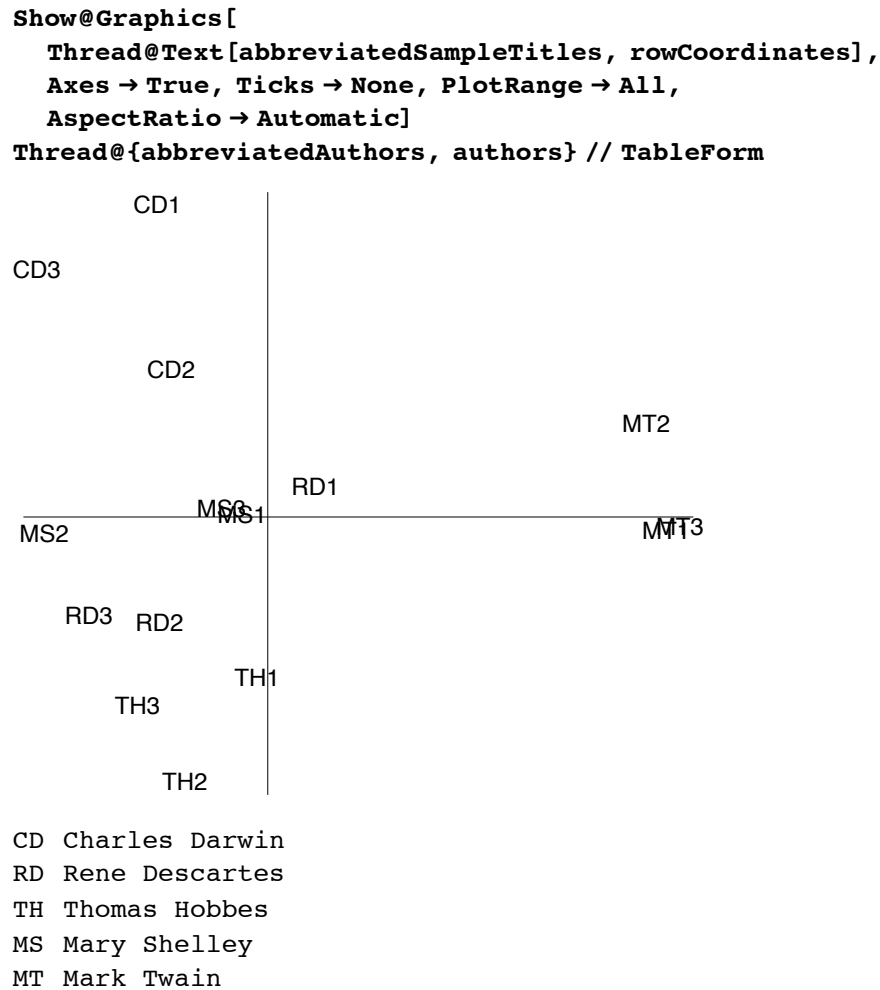
```
rowCoordinates = Take[rowScores, All, 2];
TableForm[rowCoordinates,
 TableHeadings → {abbreviatedSampleTitles, {"X", "Y"}}]
```

	X	Y
CD1	-0.0709773	0.20062
CD2	-0.0621094	0.0945122
CD3	-0.148509	0.158889
RD1	0.0306974	0.0190283
RD2	-0.0695518	-0.0683818
RD3	-0.115119	-0.0638048
TH1	-0.0068896	-0.103594
TH2	-0.0533825	-0.170423
TH3	-0.0838222	-0.121758
MS1	-0.016408	0.00138313
MS2	-0.143838	-0.0108755
MS3	-0.0298193	0.00545408
MT1	0.256214	-0.00919409
MT2	0.243356	0.0597298
MT3	0.265072	-0.00617881

The following displays each row's (abbreviated) label at the position given by its coordinates and returns a key to the abbreviations.

```
Show@Graphics[
  Thread@Text[abbreviatedSampleTitles, rowCoordinates],
  Axes → True, Ticks → None, PlotRange → All,
  AspectRatio → Automatic]
Thread@{abbreviatedAuthors, authors} // TableForm
```



```
CD Charles Darwin
RD Rene Descartes
TH Thomas Hobbes
MS Mary Shelley
MT Mark Twain
```

▲ **Figure 1.** Row plot for text samples.

The plot gives a much clearer picture of the way in which the letters are distributed across the text samples. For example, it is quite evident that—as we concluded from the original cross tab—the Mark Twain samples differ significantly as a group from those of the other writers. The text samples of Darwin and Hobbes also appear to be *sui generis*, though the Descartes and Shelly samples appear less distinct. The plot suggests that it may be possible, therefore, to distinguish between the works of at least some of the authors using correspondence analysis of their letter distributions.

□ Diagnostics

Since it uses only the first two components of the row scores, the plot above only approximates the true configuration of the rows in the cross tab. Before using it to make firm inferences, we might try to gauge the quality of the representation it provides. One indicator is derived from the inertia of the rows defined in (9). Recall that ϕ^2, the total inertia of the rows, is calculated from the row totals and the χ^2 distances of the rows to the centroid:

$$\phi^2 = \sum_{i=1}^{I} p_{i+} \, d_{iz}^2. \tag{14}$$

It may be shown that for any contingency matrix, the procedure of the previous section always places the centroid at the origin of the plot. Therefore, since Euclidean distances on the plot are supposed to approximate χ^2 distances, replacing each χ^2 distance d_{iz} in the right-hand side of (14) by the distance of the corresponding row point to origin should yield an approximation to ϕ^2. The following derives this quantity and computes its ratio to the true value of ϕ^2.

```
plotInertia = Total[rowTotals rowCoordinates^2, 2]
plotInertia / phiSquared
```

```
0.0280082
```

```
0.561666
```

Thus our two-dimensional plot captures about 56% of the total inertia of the table rows. While this seems hardly an impressive fraction, Murtagh ([3], p. 39) points out that ratios like this are not uncommon in correspondence analysis, and do not necessarily point to a bad representation. Nonetheless, we might want to exercise a modicum of caution before drawing categorical conclusions from our analysis.

As an aside, it turns out that the total inertia of the contingency matrix *P*—which was calculated "longhand" in (9)—is equal to the sum of the squares of the diagonal elements of the matrix Λ in (11). The latter comprise the *singular values* of the matrix Ω. Furthermore, the inertia retained in the two-dimensional plot is simply the sum of the squares of the first two singular values in Λ. Thus the following is an equivalent expression of the plot's inertia.

```
phiSquared1 = Tr[singularValuesMatrix^2]
plotInertia1 =
 Total@Take[Tr[singularValuesMatrix^2, List], 2]
plotInertia1 / phiSquared1
```

0.0498662

0.0280082

0.561666

■ 6. Plotting Columns

We have seen how correspondence analysis can be used to derive a visual representation of the relationships between the rows of a contingency matrix. We can also use correspondence analysis to illustrate the relationship between the rows and the columns of a correspondence matrix—between the texts and letters in our example. Since our primary concern is with the text samples, the rows of the cross tab, it might seem a digression to look at the cross tab columns (the characters appearing in the texts), but we will see in the next section that the geometry of the columns is central to the identification of the mystery texts.

As with the rows, we begin by deriving scores for the columns from the singular value decomposition in (11). With reference to (11), the matrix C, whose rows are the column scores, is calculated as follows:

$$C = D_c W, \tag{15}$$

where

$$\delta_c = \begin{pmatrix} \frac{1}{\sqrt{p_{+1}}} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \frac{1}{\sqrt{p_{+J}}} \end{pmatrix}. \tag{16}$$

Again, left multiplication by the diagonal matrix δ_c is more conveniently expressed as element-wise division in *Mathematica*.

```
columnScores = rightSingularMatrix / Sqrt@columnTotals;
```

As before, the two-dimensional column coordinates are simply the first two components of the scores.

```
columnCoordinates = Take[columnScores, All, 2];
TableForm[columnCoordinates,
 TableHeadings → {chars, {"X", "Y"}}]
```

	X	Y
B	-0.497367	2.59439
C	-1.37018	2.12847
D	1.38702	0.471271
F	-0.975604	-0.482018
G	1.04217	-0.294257
H	-0.0303792	-1.27463
I	-0.273091	-0.181811
L	1.10116	1.80015
M	-0.743416	-0.891027
N	-0.104025	-0.937001
P	-1.10752	0.365288
R	-1.25913	0.0127804
S	-0.278963	-0.060832
U	0.82417	0.108349
W	2.96486	-0.903185
Y	1.14266	1.30959

We can display both columns and rows on the same plot with a slight elaboration of the method used to plot the rows alone. The column coordinates are scaled so that the column points occupy roughly the same region of the plot as the row points.

```
columnPlotScale = .1;
gr = Graphics[{
    Thread@Text[abbreviatedSampleTitles, rowCoordinates],
    Red,
    Thread@Text[chars, columnPlotScale columnCoordinates]},
   Axes → True, Ticks → None, PlotRange → All,
   AspectRatio → Automatic]

Thread@{abbreviatedAuthors, authors} // TableForm
```

CD Charles Darwin
RD Rene Descartes
TH Thomas Hobbes
MS Mary Shelley
MT Mark Twain
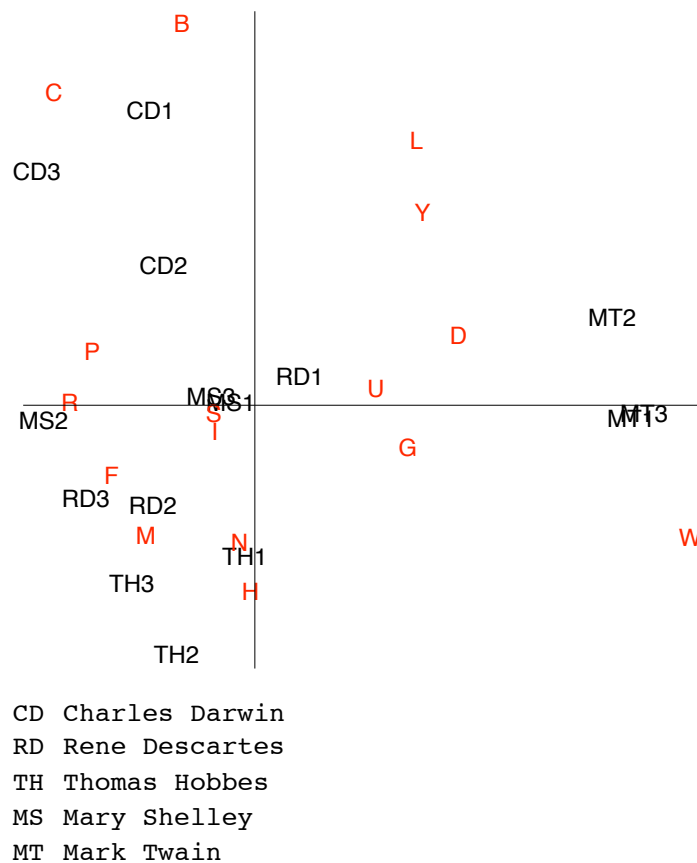
▲ **Figure 2.** Row and column plot for text samples.

Interpreting the relationships between rows and columns from a plot such as this is not as straightforward as it was for the previous plot with the rows only. For example, it is not true in general that the closer a column appears to a row, the greater the prevalence of the corresponding letter in the corresponding text sample.

To show how such relationships are actually represented, consider the text sample "MT2" (a row) and characters "P" and "Y" (columns).

```
{{isample}} = Position[abbreviatedSampleTitles, "MT2"];
{{ichar1}, {ichar2}} = Position[chars, "P" | "Y"];
```

Possibly the simplest way to determine the relationship between a text sample and a character is to draw lines from their corresponding points in the plot to the origin. If the angle between the two lines is *acute*, then the character occurs more often in the sample than it does on average in the texts as a whole. Conversely, if the angle is *obtuse*, the character occurs less often than overall. The following draws the appropriate lines for our chosen text sample and characters; it appears the character "Y" occurs more often than average in "MT2", while "P" occurs less often.

```
With[{
  pr = rowCoordinates[isample],
  pc1 = columnPlotScale columnCoordinates[ichar1],
  pc2 = columnPlotScale columnCoordinates[ichar2]},
 Show[
  gr,
  Graphics[{
    {AbsoluteDashing@{2, 4},
     Line /@ {{pr, {0, 0}}, {pc1, {0, 0}}, {pc2, {0, 0}}}},
    {Green, Circle[{0, 0}, .2 Norm@pr, ArcTan @@@ {pr, pc1}]},
    {Blue, Circle[{0, 0}, .3 Norm@pr, ArcTan @@@ {pr, pc2}]}}]]]
```
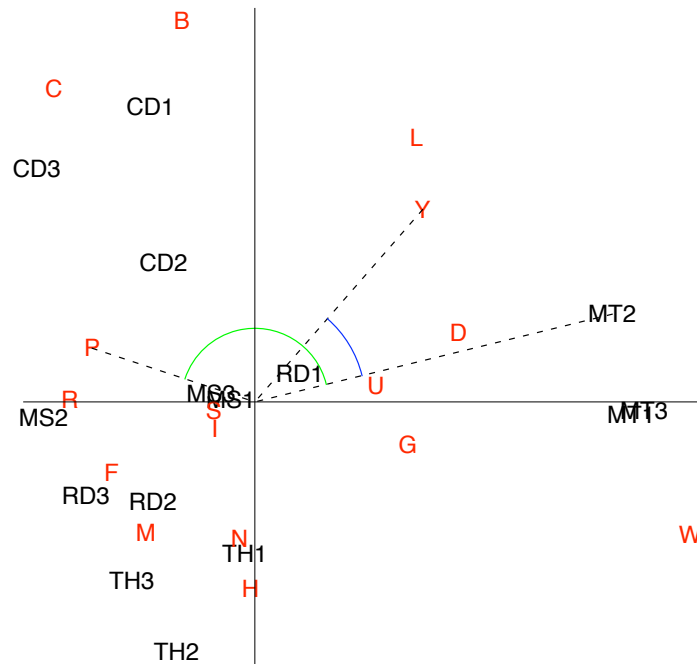

▲ **Figure 3.** Simple analysis of row/column plot.

Unfortunately, the method described above only tells us if a character appears more or less often than average in a text sample, not whether one character appears more often than another in a sample. In particular, an angle that is more acute does not signify a character that is more prevalent in a text.

A rather more complicated method that does illustrate the relative frequencies of characters in a text sample entails first drawing a line on the plot through the origin and the point corresponding to the text sample in question. Perpendiculars to this line are dropped from each character's position on the plot. The following draws such a construction for the selected text sample "MT2".

```
rowPoint = rowCoordinates[isample];
rowUnit = rowPoint / Norm@rowPoint;
columnProjections =
   Function[c, {rowUnit (rowUnit.c), c}] /@
     (columnPlotScale columnCoordinates);
projectionPoints =
   Sort@Append[columnProjections, {rowPoint, rowPoint}];
zp = {0, 0};
lineSegments =
   Partition[Sort[Append[projectionPoints[{1, -1}, 1]], zp]],
     2, 1];
segmentColors = If[rowPoint[1] < 0, {Green, Blue},
     {Blue, Green}];
Show[
 Graphics[{
   {AbsoluteDashing@{1, 5}, Line /@ projectionPoints},
   {Green, Line@{zp, rowPoint}},
   {AbsoluteDashing@{3, 3},
     MapThread[{#1, Line@#2} &,
       {segmentColors, lineSegments}]}},
   Axes → True, Ticks → None, PlotRange → All,
   AspectRatio → Automatic],
 gr]
```


▲ **Figure 4.** Comprehensive analysis of row/column plot.

The relative frequencies of the characters in the text sample can be read off by traversing the line through the text sample (colored blue and green on the plot above), looking at the positions at which the perpendiculars from the characters intersect it. A character with an intersection on the green line segment (i.e., on the same side of the origin as the text sample) occurs more often in the sample than the average in the texts overall, whereas one on the blue line segment (on the other side of the origin) occurs less frequently than the average. In addition, the further from the origin on the green line segment such an intersection occurs, the greater the frequency of the character in the sample. Conversely, the further out on the blue segment an intersection falls, the less frequent the character in the sample.

So from the plot above, it appears that the character "W" occurs most often in the sample text, and that characters "L", "D", "Y", "G", "U", "B", "S", "I", "N", "H", "C", "M", "P", "F", and "R" occur successively less often; characters "W" through "B" in the ranking appear more often than average, while "S" through "R" appear less often than average.

■ 7. Supplementary Points: Identifying the Mystery Texts

Finally, we return to the problem we faced at the outset: identifying the author or authors of the unidentified text fragments. We have seen how the application of simple correspondence analysis to the text samples allows us to view them graphically in terms of their letter distributions. In Section 5 we saw that it was generally possible to distinguish the authors of the text samples based upon the locations of the corresponding row points—with a few exceptions, samples of work by the same writer tended to occupy the same area of the plot. One might logically surmise that if we were to plot the mystery texts on the same correspondence plot as the samples, we would be able to determine their authorship by looking at the authors of the nearest samples. To begin, we need to calculate an additional cross tab containing the distribution of the selected characters in the mystery texts.

```
mysteryTextTab =
  {{24, 26, 80, 17, 32, 91, 86, 54, 32, 91, 19, 58, 93,
    50, 58, 30}, {19, 33, 35, 22, 40, 96, 116, 39, 40,
    129, 17, 72, 104, 30, 25, 24}};
mysteryTextXTitles = {"TextX1", "TextX2"};
TableForm[mysteryTextTab,
 TableHeadings → {mysteryTextXTitles, chars},
 TableSpacing → .5, TableAlignments → {Center}]
```

	B	C	D	F	G	H	I	L	M	N	P	R	S	U	W	Y
TextX1	24	26	80	17	32	91	86	54	32	91	19	58	93	50	58	30
TextX2	19	33	35	22	40	96	116	39	40	129	17	72	104	30	25	24

We could proceed by simply appending these frequencies as new rows to the original text samples cross tab given in Section 1 and recalculating the scores and coordinates for all the rows (that is, both the original samples and the mystery texts) in the resulting table. In principle, however, it is possible that the unidentified texts overlap one or more of the text samples, and if this were the case, appending the new rows to the cross tab would distort the analysis by "double-counting" some of the samples.

A more satisfactory approach derives from the fact that the row scores computed in Section 4 are actually weighted sums of the column scores calculated in Section 6. In matrix terms, recalling that P is the correspondence matrix and C is the matrix of column scores, it can be shown that:

$$R = \delta_r^2 \, P \, C, \tag{17}$$

where

$$\delta_r^2 = \delta_r \, \delta_r = \begin{pmatrix} \dfrac{1}{p_{1+}} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \dfrac{1}{p_{I+}} \end{pmatrix}. \tag{18}$$

If we replace the original correspondence matrix P in (17) with a new correspondence matrix formed from the cross tab of the unidentified texts, we derive a set of row scores for the unidentified texts according to the transformation determined by the text samples only (since they alone produced the row scores C), eliminating the risk of double-counting. Treated in this way, the unidentified texts comprise *supplementary points* in the terminology of correspondence analysis.

The following calculates row scores for the mystery texts as supplementary points (straightforward algebra vindicates the direct use of the new cross tab without the need to derive a new correspondence matrix).
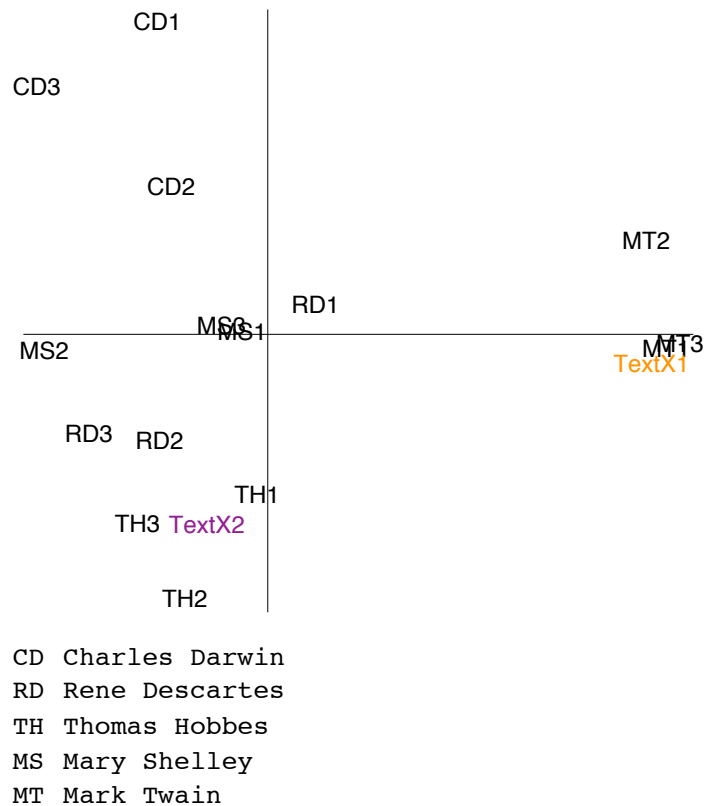
```
mysteryTextScores =
  mysteryTextTab.columnScores / (Total /@ mysteryTextTab)

{{0.246795, -0.0192125, 0.0430524, -0.0376722,
   0.0229245, -0.0444126, 0.0215733, -0.027965,
   0.0342692, 0.026139, -0.0318141, 0.0105992,
   0.0261775, 0.0248936, -0.758437, 0.651876},
 {-0.0391549, -0.122718, 0.073196, 0.0548724,
   0.147596, 0.0509551, -0.00784333, 0.011904,
   0.00672404, 0.044779, -0.0238026, -0.0271661,
   -0.0294296, -0.0392927, -0.72211, 0.693047}}
```

Lastly, as with the rows and columns, we take the first two elements of the scores above to produce the coordinates of the supplementary points. In the following, they are displayed on the same plot as the original rows.

```
Graphics[
 Join[
  MapThread[Text, {abbreviatedSampleTitles,
    rowCoordinates}],
  {Orange,
   Text["TextX1", mysteryTextScores[1, {1, 2}]]],
   Purple,
   Text["TextX2", mysteryTextScores[2, {1, 2}]]]}],
 Axes → True, Ticks → None, PlotRange → All,
 AspectRatio → Automatic]


Thread@{abbreviatedAuthors, authors} // TableForm
```

CD1

CD3

CD2

MT2

RD1

MS2 MS1 MS1   MT3
                     MMT3
                     TextX1

RD3 RD2

TH1

TH3 TextX2

TH2

```
CD  Charles Darwin
RD  Rene Descartes
TH  Thomas Hobbes
MS  Mary Shelley
MT  Mark Twain
```

▲ **Figure 5.** Plot of the mystery texts as supplementary points.

All points on this plot represent texts, or rows, and distances between points can be interpreted directly as degrees of similarity, just as with the row plot in Section 5. On this basis, judging by their closeness to the authors' other works, it appears that mystery texts 1

and 2 belong to Mark Twain and Thomas Hobbes respectively. While the manifest isolation of the Mark Twain texts on the plot leaves little doubt as to the provenance of the first unidentified text, the author of the second is a little less clearly defined—particularly given the middling diagnostic ratio calculated in Section 5. Nonetheless, I am sure you will agree that considering the rather scant literary information on which the analysis was based (amounting to no more than a table of letter frequencies), the results are encouraging.

■ 8. Conclusion

Correspondence analysis has a long and storied history that can be traced as far back as the 1930s. We have only scratched the surface of the subject in this brief introductory article. Of course, I have omitted proofs of the various assertions I have made in the course of the presentation. Furthermore, I have glossed over an important choice concerning the scaling of row and column scores and coordinates; I have used so-called *row principal scoring* (which preserves χ^2 distances between rows, but not columns), but there are other approaches that are equally valid.

A number of extensions exist to the so-called *simple* correspondence analysis presented here. Most important are *multiple* and *joint* correspondence analysis, which apply to contingency tables involving three or more variables or sets of categories (see [4] for details). For a comprehensive examination of correspondence analysis and related techniques, Greenacre's early book [5] remains among the best texts (in the English language, at least), though it is unfortunately currently out of print. Later books by Greenacre [6] and coeditor Blasius [7] explore applications of correspondence analysis and extensions to the basic methodology. Benzécri's treatise [8] is notable in that its author championed the use of correspondence analysis for many years, developing many of the geometric underpinnings that inform modern practice and establishing a seminal school of statistical analysis in France; unfortunately, translation from the original French and a prodigious price detract from the appeal of the text itself. Most recently, Murtagh [3] gives a thorough (if somewhat telegraphic) treatment of the subject, with an emphasis on the coding of data for analysis. Sections devoted to correspondence analysis also appear in the books by Agresti [9], Borg and Groenen [10], and Legendre and Legendre [11].

In his forward to [3], Benzécri writes of the immense opportunities afforded statisticians by "inexpensive means of computation that could not be dreamed of just thirty years ago" (indeed, correspondence analysis of realistically sized datasets is all but impossible without a computer). I hope that this article has demonstrated that *Mathematica* can play a valuable role in allowing all of us—statistician and non-statistician alike—to take advantage of these opportunities.

■ References

[1] D. W. Foster, *Author Unknown: Tales of a Literary Detective*, New York: Henry Holt & Company, 2001.

[2] [J. Klein], *Primary Colors: A Novel of Politics*, New York: Random House, 1996.

[3] F. Murtagh, *Correspondence Analysis and Data Coding with Java and R*, Boca Raton: Chapman & Hall/CRC, 2005.

[4] M. J. Greenacre, "Multiple and Joint Correspondence Analysis," *Correspondence Analysis in the Social Sciences* (M. J. Greenacre and J. Blasius, eds.), London: Academic Press, 1994 pp. 141–161.

[5] M. J. Greenacre, *Theory and Applications of Correspondence Analysis*, London: Academic Press, 1984.

[6] M. J. Greenacre, *Correspondence Analysis in Practice*, London: Academic Press, 1993.

[7] M. J. Greenacre and J. Blasius, eds., *Correspondence Analysis in the Social Sciences: Recent Developments and Applications*, London: Academic Press, 1994.

[8] J. P. Benzécri, *Correspondence Analysis Handbook*, New York: Marcel Dekker, 1992.

[9] A. Agresti, *Categorical Data Analysis*, 2nd ed., New York: Wiley, 2002.

[10] I. Borg and P. Groenen, *Modern Multidimensional Scaling: Theory and Applications*, New York: Springer, 1997.

[11] P. Legendre and L. Legendre, *Numerical Ecology*, 2nd English ed., New York: Elsevier Science, 1998.

About the Author

Phillip Yelland is a data analyst at Facebook, Inc., where his work centers on the use of statistical techniques in fraud detection and risk management. He has an M.A. and a Ph.D. in computer science from the University of Cambridge in England, and an M.B.A. from the University of California at Berkeley.

Phillip Yelland
Facebook, Inc.
1601 South California Avenue
Palo Alto, CA 94304
phillip.yelland@gmail.com