Correlation and regression in contingency tables. A measure of association or correlation in nominal data (contingency tables), using determinants

Colignatus, Thomas

Thomas Cool Consultancy Econometrics

15 March 2007

Correlation and regression in contingency tables

A measure of association or correlation in nominal data (contingency tables), using determinants

Thomas Colignatus, June 20 2007

http://www.dataweb.nl/~cool

I thank anonymous referee(s) for their comments, a good friend at Leiden University who does a bit in statistics and who made an important remark, and a sympathetic mind in California who wondered about the relation to logistic regression.

■ Summary

Nominal data in contingency tables currently lack a correlation coefficient, such as has already been defined for real data. A measure can be designed using the determinant, with the useful interpretation that the determinant gives the ratio between volumes. A contingency table by itself gives all connections between the variables. Required operations are only normalization and aggregation by means of that determinant, so that, in fact, a contingency table is its own correlation matrix. The idea for the normalization is that the conditional probabilities given the row and column sums can also be seen as regression coefficients that hence depend upon correlations. With M a $m \times n$ contingency table with $m \geq n$, and A = Normalized[M], then $A'A$ is a square $n \times n$ matrix and the suggested measure is r = Sqrt[Det[$A'A$]]. The sign can be recovered from a generalization of the determinant to non-square matrices. With M an $n_1 \times n_2 \times ... \times n_k$ contingency matrix, then pairwise correlations can be collected in a $k \times k$ matrix **R**. A matrix of such pairwise correlations is called an association matrix. If that matrix is also positive semi-definite (PSD) then it is a proper correlation matrix. The overall correlation then is $R = f$[**R**] where f can be chosen to impose PSD-ness. An option is to use R = Sqrt[1 - Det[**R**]]. However, for both nominal and cardinal data the advisable choice is to take the maximal multiple correlation within **R**. The resulting measure of "nominal correlation" measures the distance between a main diagonal and the off-diagonal elements, and thus is a measure of strong correlation. Cramer's V measure for pairwise correlation can be generalized in this manner too. It measures the distance between all diagonals (including cross-diagonals and subdiagonals) and statistical independence, and thus is a measure of weaker correlation. Finally, when also variances are defined for the variables then aggregate regression coefficients for the variables can be determined from the variance-covariance matrix.

Table of contents

Introduction

■ Nominal versus cardinal data

Real-valued (cardinal) data are a researcher's paradise since you can do almost anything with them. For example, when we have data on temperatures {-2.4, 4.3, 10.5, 21.5, 29.9} in degrees Celsius and associated sales of icecream {10.5, 11.3, 8.7, 50.4, 70.8} in kilo's then it is easy to express the degree of association in a single number. A correlation value generally lies between -1 and +1 and in this case we find that temperature and icecream sales are highly and positively correlated, so that we are motivated to determine the true model that captures their relationship.

```
Needs["Statistics`MultiDescriptiveStatistics`"]
Correlation[{-2.4, 4.3, 10.5, 21.5, 29.9}, {10.5, 11.3, 8.7, 50.4, 70.8}]
```

0.928301

The situation is entirely different for nominal data. This kind of data only gives categories, such as Male versus Female, or Democrat versus Republican, that are just labels and aren't numbers, making it difficult to apply the Pearson formula for correlation and to find some "average value" and to square them. The following numerical example from Kleinbaum et al. (2003:277) will be used more often in the discussion below. Let there be two shops selling both blue and green hats, and a customer visiting both shops. The customer tries all hats and scores them as "fit" or "no fit". The resulting data are nominal since they only count the cases and there isn't an "average shop" or "average colour" or "average fit". There are nevertheless the same kind of questions of association, e.g. whether we can use colour to select a fitting hat or whether it is the shop that matters.

- We can retrieve a contingency table from a databank.

```
CT[Set, Default, "Hat shops"]
```

		Green	Blue
Shop1	Fit	5	8
	No fit	1	2
Shop2	Fit	2	1
	No fit	8	5

To determine associations, the nominal data are collected in tables, called contingency tables or crosstables. When we want to analyze these data and read the statistical reference guides for help, we discover that nominal data and in particular these contingency tables lack a standard coefficient of correlation. It is also interesting to observe that a common notion in statistics is that "correlation doesn't mean causation" and it is a bit paradoxical that this notion has no numerical expression for nominal data since there is no standard measure of correlation. The contingency table is meant to analyze association, and the data lie there on the disection table right in front of us, but we cannot express simple correlation, to great frustration. Instead of giving us a simple standard, the reference guides point to various alternatives that quickly become complex, increasing our frustration from being expelled from the paradise of those real-valued data.

Contingency tables are much used in psychology, epidemiology or experimental economics, and the frustration that we feel with respect to above table is just a small example of what these researchers must experience every day. Their statistical analyses quickly proceed with more complex approaches like the χ^2test on statistical independence, which tests are not only more complex but also require levels of statistical significance that tend to say little about the strength of association. When we collect sufficiently large numbers of data then a low association may still become statistically significant. The most dubious research outcome is when we start out with small numbers and a suggestion of high association that motivates us to collect more numbers, but with an end result that we find a low association that however differs statistically significantly from independence: what to make of that ? This is not to say that the concept of statistical significance isn't useful, merely that it is not sufficient. In the case just referred to, the true test question likely is whether the result also differs significantly from a specific higher association that would be relevant for some theory. But to discuss higher degrees of association other than mere independence we are served with a measure of association.

■ The new suggestion

This paper presents a new proposal for correlation in contingency tables. We will develop this measure below, starting with a 2×2 table, generalize this to a $n \times n$ table, then $m \times n$, and finally $n_1 \times n_2 \times ... \times n_k$. In this paper the term "nominal correlation" will be used to indicate the intended application. It should be understood that this term is intended for this paper only. The term thus should be read as "what happens if we apply this measure for the purpose of expressing association ?" and not as "this is an established statistical practice firmly grounded in statistical theory". The first draft of the discussion used the neutral label "volume ratio measure" but it appeared that this destroyed much of the added value of asking that key question, and it distracted the mind with irrelevant questions on what shops, hats and fits have to do with "volume". As the measure seems promising we can only conclude and advise that future papers investigate that promise, both in theory and with tests in practice, and readers are warned not to merely and uncritically apply the measure. It must be mentioned as well that the author has limited time and resources. He has used the literature in the list of references but there obviously exists more literature. Thus, since the suggested measure is new and has not been tested in years of statistical practice, it is hoped that theorists of linear algebra and statisticians working with nominal data look into the potential value of this suggested measure.

■ Analysis of the hat shops example

For the hat shops contingency table above we find the following number for the overall association. For the reader this currently is just a meaningless number between 0 and 1, and hence the objective of the discussion below is to explain what it means. A key question is whether ranges of values can be compared to ranges of values of the Pearson correlation coefficient for real-valued data.

```
NominalCorrelation[mat = CT[Data]] // N
```

```
0.665851
```

Though we don't know what the number means, let us see what happens when we interprete it like a Pearson correlation coefficient. Since 0.67 is closer to 1 than to 0, we conclude that there is quite some association in these data. If we would select a statistical significance level then we could decide on statistical independence. But given

this amount of correlation we certainly become more interested in what the relation between these variables is.

The example of the hat shops has been taken since it is an example of the Simpson paradox. In each separate shop the green hats fit relatively better, but for both shops combined the blue hats fit relatively better. For Shop 1 the fit / no-fit odds for green hats are 5/1 and for blue hats 8/2, thus the odds ratio (5/1) / (8/2) = 5/4. For Shop 2 we find the fit / no-fit odds ratio (1/4) / (1/5) = 5/4 too. For the two shops separately the odds ratio is above 1 but for the total (7/9 versus 9/7) it is below 1. The dispersion over the two shops is a confounder.

```
OddsRatio /@ Append[mat, Plus @@ mat]
```

$$\left\{\frac{5}{4}, \frac{5}{4}, \frac{49}{81}\right\}$$

As the tables only concern the problem of fitting hats then it makes sense to eliminate the confounder, add the two tables, and find a small association, but of a reverse sign.

```
NominalCorrelation[Plus @@ mat] // N
```

$$-0.125$$

Note that the analysis depends upon the case at hand. When these data tables don't concern hat shops but represent another kind of problem, then it might not be sensible to merely add the subtables. In that case the difference between -0.125 and 0.67 helps us to consider that there indeed is some relation, and, if the problem is serious enough, we might grow convinced that it could be worth while to find the true model. For example, we might do a meta-analysis on the findings of the separate shops, aggregating the problem in such a way that the overall direction reflects the individual ones. Or, for example in voting theory, there can also be paradoxes but on close inspection with an acceptable explanation, such that it would neither make sense to simply add the subtables. Indeed, in the current example, if Shop1 has better quality control then one would buy a blue hat from Shop1 and be happy that the shops are not aggregated. To be sure, though, one has to check whether there is cause for a systematic difference.

■ What nominal correlation stands for

Thus, this example shows that the notion of nominal correlation coefficient might help in the analysis. In this introductory discussion, we have been interpreting the number just like a Pearson coefficient of correlation. Can we really do so ? What is behind this number ? This question brings us to some leading ideas. But before looking into those leading ideas, we better present the key result of this paper, so that it is clear what the focus is on.

■ On the structure of this paper

The author has been struggling a bit both to develop the idea and how to present it while developing it. The paper is part of Colignatus (2007e), a work in progress on writing a book "Elementary Statistics and Causality". Colignatus (2007f) discusses causality in the 2 × 2 × 2 contingency table. The present paper arose from the observation that contingency tables didn't have an easy expression of the notion that "correlation is no causation" since there was no obvious measure for correlation. The first version of this paper presented nominal correlation, starting with the 2 × 2 table and building up to the $n_1 \times ... \times n_k$ case. Next versions of the paper started to insert general notions and overviews before that definition. The development created an ever increasing need for such insertions since we also had to resolve issues such as (i) overall correlation, also for real data, (ii) (approximating) positive semi-definiteness, (iii) relation to existing measures and the Cramer's V in particular, (iv) the variance, (v) the sign and determinant of a non-square matrix, (vi) relation to logistic regression. Inserting those topics, and then again giving overviews, actually reduced clarity. The current version of the paper returns to the original format, starts again with the 2 × 2 table and builds up to the $n_1 \times ... \times n_k$ case. Only then, once you have seen the construction and understand what the discussion is about, then it will be useful to extend on those mentioned issues. Thus, the discussion starts out with the geometric structure and later provides the other details.

It may help to observe that the author is an econometrician who is used to working with both micro data and macro aggregates. An economy has millions of products but there still is an aggregate price index. Thus it comes naturally to this author to concentrate on the k variables and be less impressed by the $n_1 \times ... \times n_k$ categories. Consider for example a 5 × 7 × 3 × 4 contingency table, thus with 420 cells. We might want to explain all individual cells and subsequently use statistical tests to determine the most

parsimonious model. Nominal correlation and regression however focus on the four variables only. This makes a sharper distinction between the collection of the data and the processing of the data. Proper measurement requires that we collect the data in all detail of the 420 cells but decision making might be guided by summary statistics that only concern the four variables. In the detail of the 420 cells we may find statistically significant relations but the overall correlation may still be too low to generate sufficient interest.

The basic ideas

Relation between correlation and regression

Correlation and regression for real-valued data have some properties that we retain for nominal data. A measure of correlation $\rho_{x,y} = \rho_{y,x}$ expresses how much the variables x and y are associated on a scale of -1 to 1. The regression coefficient of x in a pairwise regression for y is $\beta_{y,x} = \rho_{x,y}\ \sigma_y\ /\ \sigma_x$. By consequence, looking both ways, $\rho_{x,y}{}^2 = \beta_{y,x}\ \beta_{x,y}$. In multiple regression, the regression coefficients are derived as cofactors of the matrix of pairwise correlations. The interpretation of the regression coefficient is that it gives the contribution of one unit change of x to both $\hat{y}$ and y. The squared multiple correlation coefficient R^2 is the squared correlation between y and $\hat{y}$.

A contingency matrix as its own correlation matrix

Consider a $m \times n$ contingency table, thus with two nominal variables, the explanatory variable x with n categories put in the columns and y with m categories put in the rows. When counting occurrences of the variables there is a dependence on time since both variables are scored within the same time frame. The result of the counts is the contingency table C_t with t the time index. For example for two dichotomous variables $C_t = \{\{n_{11\,t},\ n_{12\,t}\},\ \{\{n_{21\,t},\ n_{22\,t}\}\}$. The data can be aggregated over a period so that t would rather stand not for an individual observation but for a period total. A summation over a category can be denoted with a "+" instead of the index, so that n_{1+t} is the summation of the first row for period t. The counts can be expressed as probabilities $p_{\mathrm{ijt}} = n_{\mathrm{ijt}}/\ n_{++t}$ and for the marginal probabilities we can write $p_{\mathrm{jt}} = p_{+\mathrm{jt}}$ for column variable x and $q_{\mathrm{it}} = p_{i+t}$ for row variable y. Probabilities look like real-valued data though for nominal data the only order is the order of presentation.

When the difference over time has meaning then we can do a regression over time, for example as $n_{ijt} = \alpha_{ij} + \beta_{ij}\, n_{+jt} + \epsilon_{ijt}$ which would give an estimate of the *i,j* cell consisting of a base value and a marginal contribution dependent upon the column sum. A vector format of regression uses x and y as vectors containing the border sums (marginals), seeing the variables as vectors with the data in the order of presentation, so that $y_t = \alpha + \beta\, x_t + \epsilon_t$, where α and β are matrices collecting those cell coefficients.

It may also be that the distinction between time periods is less informative or even disinformative so that the true result is C_+ only. In that case a regression over time ceases to be relevant and even becomes impossible. It is still possible to calculate the conditional probabilities $c_{i,j} = \mathrm{P}[C_{i,j} \mid x_j] = n_{i,j} / n_{+j}$ that can be seen as regression coefficients. Then by definition $\alpha = 0$. The example of the regression over time helps us to recognize that those conditional probabilities are regression coefficients indeed. Given that we target the cells, their values would be independent of the other cells. In that case $\rho_{\mathrm{xj,yi}}{}^2 = \beta_{\mathrm{yi,xj}}\, \beta_{\mathrm{xj,\,yi}} = n_{i,j}{}^2 / (n_{+j}\ n_{i+})$. The matrix that contains the square roots of those values can be denoted as C_ρ, and will be called in this paper the "normalized matrix" of the $m \times n$ contingency table.

```
NormalizedMatrix[{{a, b, c}, {d, e, f}}]
```

$$\begin{pmatrix} \frac{a}{\sqrt{a+b+c}\,\sqrt{a+d}} & \frac{b}{\sqrt{a+b+c}\,\sqrt{b+e}} & \frac{c}{\sqrt{a+b+c}\,\sqrt{c+f}} \\ \frac{d}{\sqrt{a+d}\,\sqrt{d+e+f}} & \frac{e}{\sqrt{b+e}\,\sqrt{d+e+f}} & \frac{f}{\sqrt{c+f}\,\sqrt{d+e+f}} \end{pmatrix}$$

The matrix C_ρ contains correlation coefficients but differs from a standard correlation matrix **R**. C_ρ concerns categories, has size $m \times n$, and if it would be square then it would not necessarily be symmetric. Since we are discussing nonnegative matrices, the correlations will also be nonnegative. C_ρ seems useless, but we can use C_ρ to say something about a pairwise correlation between x and y. Namely, we can say that the *variables* x and y are highly correlated when the categories of x translate into those of y. There would be full correlation when C_ρ would happen to be filled with zeros except for something that looks like a diagonal containing only ones. Of course, the $m \times n$ case is generally non-square, so we need some generalization of that (this would basically be an aggregation matrix).

Writing the regression in vector notation also comes with a shift in focus, away from the single cells and towards targetting the variables. We started out with the variables but the measurement created cells that started to attract all attention towards themselves. Refocussing, we wonder whether there might not be a "pairwise correlation" between x and y. This refocus may come along with an imposition of overall conditions. For

example, while the impact of a *category* is measured by *adding* one unit, the impact of a *variable* might be measured by *moving* a unit. Such conditions could be stated in terms of the regressions on the cells but derive their meaning from the interpretation in terms of the variables.

Moving a unit

The reasoning on moving a unit is as follows. It will be easiest when we take the context of a treatment-control study. Let S be the total number of participants in the study, T the number treated, C the controls, E the number with a (positive) effect. Then $E = E_T + E_C$ with the respective subgroups, such that $p_T = E_T / T$ and $p_C = E_C / C$ are the respective effect probabilities, so that $E = p_T\, T + p_C\ C$.

TreatmentControlMatrix[Table, 0, Set]

	Effective	Ineffective	Total
Treatment	ET	IT	ET + IT
Controls	EC	IC	EC + IC
Sum	EC + ET	IC + IT	EC + ET + IC + IT

In terms of our earlier discussion, moving one hat from shop 1 to shop 2, we increase the control group, and thus we would calculate the influence on the ineffectiveness. For this kind of study it makes more sense to consider the opposite, which gives the same regression coefficient. The regression coefficient for the effect of an increase of treatment T on the total effect E can be found as:

$$\mathrm{d}E / \mathrm{d}T = (p_T\,(T + \mathrm{d}T) + p_C\,(C - \mathrm{d}T) - E) / \mathrm{d}T$$

$$\mathrm{d}_T E = p_T - p_C \quad \text{taking} \quad \mathrm{d}T = 1$$

Note that $T^* = 1 / (p_T - p_C)$ is also called the “number needed to treat”, i.e. the number needed to have one single success above the control outcome.

This interpretation of a regression coefficient is only heuristic since it is formulated in terms of categories and not variables. There are only two categories here and when there are more categories then taking a unit from “one position” and placing it into “another position” would be averages or vector concepts.

For a numerical example, let us take the relation between shops and fitness. Let us sum out the colour to get the 2 × 2 data. The structure that we get is that of a treatment

control study, where shop 1 is the treatment, shop 2 the control, and the fit is the effect measure.

```
CT[Sum, "Colour"]
```

	Fit	No fit
Shop1	13	3
Shop2	3	13

We find a nominal correlation that neglects the effect of submatrices.

```
NominalCorrelation[%] // N
```

0.625

Let us move one hat from shop 1 to shop 2. The requirement is that this hat is typical of shop 1 when it is taken from shop 1, but suddenly becomes typical of shop 2 when it is put into shop 2. In other words, the rates of fitness per shop are kept constant. Their difference, called the risk difference, turns up as above correlation measure. Since the variances are 1 there is no difference between the correlation and regression coefficient.

```
Move1FromRow1To2[matsf = %%] // N
```

$$\left\{\text{Mat(In)} \to \begin{pmatrix} 13. & 3. & 16. \\ 3. & 13. & 16. \\ 16. & 16. & 32. \end{pmatrix}, \text{Mat(Out)} \to \begin{pmatrix} 12.1875 & 2.8125 & 15. \\ 3.1875 & 13.8125 & 17. \\ 15.375 & 16.625 & 32. \end{pmatrix},\right.$$
$$\left.\text{Row[1.]} \to \{-0.8125, -0.1875\}, \text{Row[2.]} \to \{0.1875, 0.8125\}, \text{Dif} \to \{-0.625, 0.625\}\right\}$$

Again, this is an example of a 2 × 2 case, also from a symmetric matrix. It is not correct to conclude that nominal correlation in general would be equal to risk difference. For a more general discussion, see Colignatus (2007g), that uses the risk difference regression as a bridge between nominal regression and logistic regression. (It has been used as a bridge here too.)

Linking up to the volume ratio

We are now ready to reconsider the 2 × 2 case, link up with the volume ratio idea, and build up nominal correlation to the $n_1 \times ... \times n_k$ case.

The 2 × 2 case

■ The layout

Contingency tables generally are presented with table-headings and / or border-sums. Calculations are normally done with the inner matrix. We will use the following formats.

- The core or inner matrix.

  mat2 = {{a, b}, {c, d}}

  $$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

- With border sums added.

  mat3 = {{a, b}, {c, d}} // AddBorderSums

  $$\begin{pmatrix} a & b & a+b \\ c & d & c+d \\ a+c & b+d & a+b+c+d \end{pmatrix}$$

- With table headings. The following case has men and women dieting or not. Observed frequencies are *a* to *d*. We wonder whether the behaviour of the groups differs.

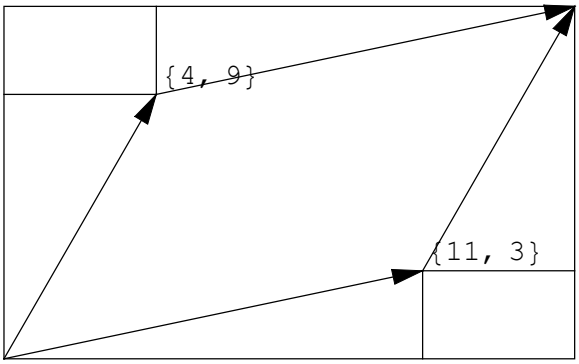
  DiseaseTestMatrix[Table, Set, mat3, "men", "dieting"]

	Disease	Not	Tested
Positive	a	b	$a+b$
Negative	c	d	$c+d$
Sum	$a+c$	$b+d$	$a+b+c+d$

■ An insight: diagram of the 2 × 2 case

The 2 × 2 matrix {{*a, b*}, {*c, d*}} contains two row vectors {*a, b*} and {*c, d*} that together span a parallelepiped. When we draw a diagram of this, we find that the parallelepiped is contained in a rectangle with sides (*a* + *c*) and (*b* + *d*) which are the column sums of the matrix. The following gives a numerical example. Recall that we discuss nonnegative matrices.

```
ShowDet[{{11, 3}, {4, 9}}];
```



The total area of the rectangle is given by (*a* + *c*)(*b* + *d*) while the area of the parallelepiped can be found by subtraction of the small rectangles and triangles, thus (*a* + *c*)(*b* + *d*) - 2 *bc* - 2 * ($\frac{1}{2}$*ab*) - 2 * ($\frac{1}{2}$*cd*) = *ad* - *bc*. This latter value is the determinant of the matrix.

```
(a + c) (b + d) - 2 b c - 2 (1/2 a b) - 2 (1/2 c d) // Simplify
```

$$a\,d - b\,c$$

```
Det[{{a, b}, {c, d}}]
```

$$a\,d - b\,c$$

When we take the ratio of the areas *cr* = (*ad* - *bc*) / ((*a* + *c*)(*b* + *d*)) then we find a number between -1 and 1.

Note that the determinant *ad* - *bc* also holds for the dual (transposed) matrix, giving a ratio *rr*.

Since there are two ways of looking at the matrix a more robust measure is the geometric average $\sqrt{cr * rr}$. The numerator remains *ad* - *bc* but the denominator becomes $\sqrt{}((a + c)(b + d)(a + b)(c + d))$. This gives us a "standardized volume ratio".

```
FullSimplify[CorrelationPr2By2[mat2], Assumptions → {a ≥ 0, b ≥ 0, c ≥ 0, d ≥ 0}]
```

$$\frac{a\,d - b\,c}{\sqrt{(a+b)\,(a+c)\,(b+d)\,(c+d)}}$$

We can easily check that a diagonal matrix with $b = c = 0$ gives outcome +1 and with $a = d = 0$ gives outcome -1. Nominal data have no natural order, but one cannot avoid an order of presentation and the sign of the correlation in this case reflects that.

■ Statistical independence means zero correlation

As might already have been obvious from the properties of determinants, algebraic dependence means that this measure shows zero association. The following routine constructs a matrix by multiplying the marginals. We can multiply with the total number of observations *N*.

```
mat4 = PrTable[t, p] N
```

$$\begin{pmatrix} N\,p\,t & N\,(1-p)\,t \\ N\,p\,(1-t) & N\,(1-p)\,(1-t) \end{pmatrix}$$

```
FullSimplify[CorrelationPr2By2[mat4], Assumptions → {t ≥ 0, p ≥ 0}]
```

0

Next step: square matrices

The 2 × 2 case can easily be extended to the $n \times n$ case. The determinant is only defined for square matrices. We directly get a measure if we apply the paradigm that we put the determinant in the numerator and the square root of the products of the border sums in the denominator.

In normalizing a matrix with the products of sums of columns and rows, it may be noted that the latter relate to determinants of diagonal matrices. Let A be any square matrix. Let Detr(A) = Det(diag(A 1)) and Detc(A) = Det(diag(A' 1)). Detr and Detc are just the

products of the diagonals but there may be an advantage to write them in this manner. The determinant of a normalized matrix can be resolved in subterms using the properties of the determinant with respect to multiplications of rows and columns. The suggested measure for square matrice M then gives $\text{Det}(M) \;/\; \sqrt{\text{Detr}(M)\ \text{Detc}(M)}$.

The case $n = 3$ is already a big deviation from $n = 2$. The following example is from Mood & Graybill (1963:325). Given the categories, and possibly an implied order related to the way of presentation, or perhaps even some true ordinality, one might interprete the negative association as "the more capable the less poorly clothed". The correlation measure shows little *overall* association though (note the emphasis on "overall").

```
mat5 = {{"", "Dull", "Intelligent", "Very Capable"},
        {"Very well clothed", 81, 322, 233},
        {"Well clothed", 141, 457, 153},
        {"Poorly clothed", 127, 163, 48}};
```

```
SquareMatrixNormedDet[Take[mat5, -3, -3]] // N
```

```
-0.0285548
```

The low overall association does not preclude that there might be more association when we aggregate over subgroups. However, in that case we need a theory that can handle $m \times n$ matrices.

```
mat6 = mat5 . Transpose[{{1, 0, 0, 0}, {0, 1, 1, 0}, {0, 0, 0, 1}}]
```

Null	Dull + Intelligent	Very Capable
Very well clothed	403	233
Well clothed	598	153
Poorly clothed	290	48

```
SquareMatrixNormedDet[Take[mat6, -3, -2]] // N
```

SquareMatrixNormedDet::dim : Must be a n by n matrix More…

Next step: *m* × *n* matrices

■ Theory

Take an arbitrary $m \times n$ contingency table, say $A \in \mathbb{R}^{m,n}$, not necessarily integers. With $f[x] = A\,x$, we have f: $\mathbb{R}^n \to \mathbb{R}^m$. A property in linear algebra is that if S is a subset of $\mathbb{R}^n$ with Volume[S] then Volume[f[S]] = Sqrt[Det[$A'.A$]] Volume[S]. Hence Sqrt[Det[$A'.A$]] gives a volume ratio. For a problem like the hat shops, one can interprete f as taking combinations of the nominal categories, where a combination is not quite an addition (though it is in linear algebra). Thus there is an interpretation that makes the volume ratio an interpretable summary of association. Note that there is also a dimensionality adjustment, since a 2D surface has zero 3D volume; however, each space has a unit metric, e.g. as 1 m^2 or 1 m^3.

To make the measure robust we include the following: (1) A is normalized by dividing column-wise and row-wise by the square roots of the sums of the columns and rows respectively, (2) of $A'.A$ and $A.A'$ we take the one with lower size, since the larger one will give a zero determinant. It makes sense to consider only the smaller matrix since all variation in the higher dimension is only relevant with respect to the smaller dimension. (3) While the above causes the loss of the sign of the association, we recover this by letting s[A] be the sign of the association in the matrix (see below).

Hence: let M be the $m \times n$ contingency table and $m \geq n$. The suggested pairwise VolumeRatio or NominalCorrelation measure is:

$r = s[M]$ Sqrt[Det[$A'A$]] with A = Normalized[M] and s[M] the sign

PM 1. For normalization, let dr = diag[M.1] and dc = diag[1'.M] with diag[...] the diagonal matrix. Then $A = dr^{-\frac{1}{2}}\, M\, dc^{-\frac{1}{2}}$. We can check that if M is square, then the $m \times n$ measure reduces to the measure already defined for square measures, with the only loss of the sign (due to the square root of the squares). PM 2. Note that this is a special kind of normalization. Repeated application results into different values. PM 3. Given that we are still considering nonnegative matrices, the only condition for the division would seem to be that every row or column contains at least one non-zero element. A row or column with only zero's would cause the determinant to be zero too so that

division can be left out of consideration. It might be too simple to merely drop such a row or column. PM 4. For a calculation procedure, it seems most efficient to first check the diagonal. Only for zero elements on the diagonal it is necessary to check whether also the row or column are all zero. PM 5. In OLS the coefficient of determination relates to the correlation between explained variable y and explanation $\hat{y}$. In this present case for nominal data we consider an overall correlation. See **Appendix C**.

■ Implementation

The examples above

We now can tackle above case. We first check that we reproduce the proper value of the square example that we could calculate above, and then produce the value for the case that we could not calculate before.

```
NominalCorrelation[Take[mat5, −3, −3]] // N
```

```
−0.0285548
```

```
NominalCorrelation[Take[mat6, −3, −2]] // N
```

```
−0.208759
```

Consistency with the 2 × 2 case

It may be noted that the measure is consistent with the 2 × 2 case. We lose the sign because of the quadratic term so that it has to be added explicitly.

```
NominalCorrelation[{{a, b}, {c, d}}] // Simplify
```

$$\sqrt{\frac{(b\,c-a\,d)^2}{(a+b)\,(a+c)\,(b+d)\,(c+d)}}\ \text{sgn}(a\,d-b\,c)$$

The sign of the correlation

There appears to be no standard solution to determine the direction of association in a non-square matrix. In writing this paper, the author has been experimenting with three approaches: (i) neglect the sign, use only the absolute value, (ii) aggregate the matrix in a 2 × 2 matrix and use its sign, (iii) use a generalized determinant for non-square matrices. The latter applies Sqrt[Det[*A'.A*]] for the level and NsqDet[*M*] for the sign. **Appendix L** explains the last approach and shows that it likely is best. The method of splitting up and aggregating does not always generate the right result for square matrices so it is not sufficiently general. For a $m \times n$ contingency table, the routine PairwiseSign determines the direction of the association. In case of doubt (or formal matrices) it is always possible to set the sign to 1.

```
NominalCorrelation[{{a, b}, {c, d}}, ForceSign → 1] // Simplify
```

$$\sqrt{\frac{(b\,c - a\,d)^2}{(a+b)\,(a+c)\,(b+d)\,(c+d)}}$$

Finally $n_1 \times n_2 \times \ldots \times n_k$

■ In general

Let us take x_i as nominal data, $i = 1, \ldots, k$, and let n_i be the number of nominal categories in the i^{th} variable, then a contingency matrix M is of size $n_1 \times n_2 \times \ldots \times n_k$. When we can define an association measure $r_{i,j}$ between two variables then we can collect all of those in a $k \times k$ correlation (association) matrix **R.** Subsequently, we can define an overal correlation as $R = f[\mathbf{R}]$. For this *f,* a default approach is to take the highest value of any multiple correlation in the data.

For nominal data, the crux is to find a good $r_{i,j}$. The relation between two variables x_i and x_j can be considered in two ways. (1) One way is to sum out all other variables, giving $B = M_{i,j}$, the border matrix of x_i and x_j. It may well be that some third variable in some category has most influence on the correlation score, but when we consider only two variables then the influence of the other variables and the manner of influence might be considered to no longer apply. (2) The other approach is to hold that all variation in the submatrices that generate the border matrix is important too. Consider all matrices

$B_p = M_{i,j}(p)$ used in the summing procedure to create $B = M_{i,j} = \sum_{p=1}^{n[i,j]} M_{i,j}(p)$ and note that $p = 1, ..., n[i, j] = n_1 \times n_2 \times ... \times n_k / (n_i \times n_j)$. With our measure r on an arbitrary $m \times n$ contingency table, we determine the r_p for each B_p, and determine a weighted average $\bar{r}_p$, using the total number of observations in B_p as the weight.

The overall method thus contains two steps that should not be confused:

1. nominal correlation for two variables, given a total border sum, using either that border sum or the weighted average of the inner matrices
2. overall correlation from a matrix of correlations for k variables (for which we take the maximal value of any multiple correlation rather than the determinant measure).

The first is nominal correlation strictly by itself, the second is overall correlation for any kind of correlation matrix. Each step has its own reason. But with the distinctions and categories we get various combinations that we can look into, see also **Appendix I**.

As said, we should be strict about association and correlation when this is relevant. The routine VolumeRatioMatrix produces the matrix of pairwise correlations based upon the volume ratio. The routine NominalCorrelationMatrix uses that as input and checks upon positive semi-definiteness. When there is a problem then it first determines the λ between the border sum result and the inner matrices result. When a problem remains then it uses a PSD approximator, by default setting negative eigenvalues to zero.

■ Using bordersums

Reconsider the shop case, its dimensions (1) Shop, (2) Colour and (3) Fitness, and its border matrices. For example, the border matrix for Shop versus Colour (the first two dimensions) is created by summing over Fitness (the third dimension).

CT[Show]

		Green	Blue
Shop1	Fit	5	8
	No fit	1	2
Shop2	Fit	2	1
	No fit	8	5

BorderMatrices[mat]

$$\left\{\{1, 2\} \to \begin{pmatrix} 6 & 10 \\ 10 & 6 \end{pmatrix}, \{1, 3\} \to \begin{pmatrix} 13 & 3 \\ 3 & 13 \end{pmatrix}, \{2, 3\} \to \begin{pmatrix} 7 & 9 \\ 9 & 7 \end{pmatrix}\right\}$$

When we use only these border matrices to determine the association between the variables then we neglect the variation that is in the submatrices. Application of the NominalCorrelation measure to these 2 × 2 bordermatrices gives the elements of the total correlation matrix.

```
case = BorderMatrices;

nc[case, Mat] = NominalCorrelationMatrix[mat, BordersOrAll → BorderMatrices]
```

$$\begin{pmatrix} 1 & -\frac{1}{4} & \frac{5}{8} \\ -\frac{1}{4} & 1 & -\frac{1}{8} \\ \frac{5}{8} & -\frac{1}{8} & 1 \end{pmatrix}$$

There are no messages and thus this volume ratio matrix is directly PSD of itself. Since we used the bordermatrices as the method, the value of λ should be 0. We can recover it from the Results.

```
Factor /. Results[NominalCorrelationMatrix, BorderMatrices, True]
```

```
0
```

Having produced the correlation matrix, we now can consider the measures for overall correlation. By default, OverallCorrelation[mat] only considers the maximal multiple correlation in this correlation matrix.

```
mrsq = MultipleRSquared;

nc[mrsq, case] = {Correlation → (OverallCorrelation[nc[case, Mat]] // N)}
```

```
{Correlation → 0.648564}
```

An alternative is to take the arithmetic average of all multiple correlations. One may experiment with such functions but it will be noted quickly that a single high correlation within a block of data might be sufficient to say that there is high correlation within that data.

- The default is Function → Max.

```
OverallCorrelation[nc[case, Mat], Function → Average] // N
```

```
0.540495
```

OverallCorrelation[nc[case, Mat], Mode → Det] would use the determinant measure. We can also calculate this directly.

```
nc[Det, case] = {Correlation → ((1 – Det[nc[case, Mat]]) // Sqrt // N)}
```

{Correlation → 0.655506}

■ Using inner submatrices

When we want to account for the variation in the inner submatrices, then we can determine all submatrices that are used in the sum for a pairwise border matrix, determine each separate VolumeRatio, and then add these outcomes. It will be sensible in this addition to use the weights given by the numbers of observations in each submatrix. Note that there is an element of arbitrariness in using the weighted sum. We might also weigh e.g. the *squared* measures of association, or use a geometric average, or take the maximal value, and so on. For the time being, the simple weighted average seems wise. It will also be instructive to see how the procedure works. In the following, the "VR" stands for the VolumeRatio measure, i.e. the NominalCorrelation, and the "Add" stands for taking the weighted sum.

```
NominalCorrelationMatrix[mat, BordersOrAll → Show]
```

$$\begin{pmatrix} 1 & \text{Add}\left[\left\{\text{VR}\left[\begin{pmatrix}5 & 8\\ 2 & 1\end{pmatrix}\right], \text{VR}\left[\begin{pmatrix}1 & 2\\ 8 & 5\end{pmatrix}\right]\right\}\right] & \text{Add}\left[\left\{\text{VR}\left[\begin{pmatrix}5 & 1\\ 2 & 8\end{pmatrix}\right], \text{V}\right. \\ \text{Add}\left[\left\{\text{VR}\left[\begin{pmatrix}5 & 8\\ 2 & 1\end{pmatrix}\right], \text{VR}\left[\begin{pmatrix}1 & 2\\ 8 & 5\end{pmatrix}\right]\right\}\right] & 1 & \text{Add}\left[\left\{\text{VR}\left[\begin{pmatrix}5 & 1\\ 8 & 2\end{pmatrix}\right], \text{V}\right. \\ \text{Add}\left[\left\{\text{VR}\left[\begin{pmatrix}5 & 1\\ 2 & 8\end{pmatrix}\right], \text{VR}\left[\begin{pmatrix}8 & 2\\ 1 & 5\end{pmatrix}\right]\right\}\right] & \text{Add}\left[\left\{\text{VR}\left[\begin{pmatrix}5 & 1\\ 8 & 2\end{pmatrix}\right], \text{VR}\left[\begin{pmatrix}2 & 8\\ 1 & 5\end{pmatrix}\right]\right\}\right] & 1 \end{pmatrix}$$

Actually doing the calculation gives this correlation matrix.

```
case = All;
```

```
nc[case, Mat] = NominalCorrelationMatrix[mat, BordersOrAll → All]  // N
```

$$\begin{pmatrix} 1. & -0.221917 & 0.61807 \\ -0.221917 & 1. & 0.0413449 \\ 0.61807 & 0.0413449 & 1. \end{pmatrix}$$

In this numerical example, using all variation in the submatrices, the total measure (maximal multiple correlation) appears to be a bit different from the one using only the border matrices (in particular due to some of the signs of the correlations).

```
nc[mrsq, case] = {Correlation → (OverallCorrelation[nc[case, Mat]]  // N)}
```

{Correlation → 0.665851}

Using the determinant measure of overall correlation.

```
nc[Det, case] = {Correlation → ((1 – Det[nc[case, Mat]]) // Sqrt // N)}
```

```
{Correlation → 0.666565}
```

■ Review

We calculated the nominal correlations using both the border matrices or all submatrices, and using only the maximal multiple RSquared or the determinant measure. The following table collects these results.

```
heading = TableHeadings → {{mrsq, Det}, {BorderMatrices, All}};

InsideTable[Set, nc, heading]

InsideTable[Show, Correlation]
```

	BorderMatrices	All
MultipleRSquared	0.648564	0.665851
Det	0.655506	0.666565

There are no general conclusions yet since all this is a result of this particular numerical example.

The difference between the correlation matrices is:

```
nc[All, Mat] – nc[BorderMatrices, Mat]
```

$$\begin{pmatrix} 0. & 0.0280832 & -0.00692995 \\ 0.0280832 & 0. & 0.166345 \\ -0.00692995 & 0.166345 & 0. \end{pmatrix}$$

■ The default

As already shown above, the default definition uses the inner submatrices and the maximal multiple RSquared. In that case we can directly call NominalCorrelation, that is smart enough to recognize that the input is a multidimensional contingency table so that it produces the underlying matrix itself.

```
NominalCorrelation[mat] // N
```

```
0.665851
```

The key result of this paper

In a nutshell

The key result of this paper is that correlation and regression coefficients are made available, based upon a determinant or volume ratio measure of association. Regression coefficients follow from both the correlation matrix and the variances. Though the nominal data have no order, there is an order of presentation. Regression coefficients then derive their interpretation from moving a unit in a positive or negative direction of that order of presentation. For the hat shop tables, "getting a higher shop number", "becoming blue" and "losing fit" are positive steps in the right direction.

The numerical example

For the discussion on nominal variance and its scale, see **Appendix K**. Nominal variance is defined by default on the border totals, with a maximal value of 1 when all categories for a variable have equal frequency. For the hat shops example we find that the border totals are all {16, 16} so that the variances are all equal to 1. Points to note are: (1) the nominal correlation matrix does not depend upon the standard deviations but instead the variance-covariance matrix is created with them (in reverse order as for real data), (2) for the regression coefficients the standard deviations drop from the expressions when they are all equal.

NominalStatistics[mat]

$$\Big\{\text{ContingencyTableQ} \to \text{True},\ \text{OverallCorrelation} \to 0.665851,$$
$$\text{Length} \to \{2, 2, 2\},\ \text{EffectiveNumberOfCategories} \to \{2., 2., 2.\},$$
$$\text{Variance} \to \{1., 1., 1.\},\ \text{Spread} \to \{1., 1., 1.\},\ \text{BorderTotals} \to \begin{pmatrix} 16 & 16 \\ 16 & 16 \\ 16 & 16 \end{pmatrix},$$
$$\text{BorderMatrices} \to \left\{\{1, 2\} \to \begin{pmatrix} 6 & 10 \\ 10 & 6 \end{pmatrix}, \{1, 3\} \to \begin{pmatrix} 13 & 3 \\ 3 & 13 \end{pmatrix}, \{2, 3\} \to \begin{pmatrix} 7 & 9 \\ 9 & 7 \end{pmatrix}\right\},$$
$$\text{NominalCorrelationMatrix} \to \begin{pmatrix} 1. & -0.221917 & 0.61807 \\ -0.221917 & 1. & 0.0413449 \\ 0.61807 & 0.0413449 & 1. \end{pmatrix},$$
$$\text{CovarMat} \to \begin{pmatrix} 1. & -0.221917 & 0.61807 \\ -0.221917 & 1. & 0.0413449 \\ 0.61807 & 0.0413449 & 1. \end{pmatrix},$$
$$\text{CovarRegress} \to \begin{pmatrix} 0. & -0.247895 & 0.628319 \\ -0.400445 & 0. & 0.288848 \\ 0.659735 & 0.187751 & 0. \end{pmatrix}\Big\}$$

The main result of this output are both the nominal correlations and the matrix *C* with regression coefficients.

NominalStatistics[Results, CovarRegress, CT[Variables]]

	Shop	Colour	Fitness
Shop	0.	−0.247895	0.628319
Colour	−0.400445	0.	0.288848
Fitness	0.659735	0.187751	0.

CovarRegress takes the matrix of pairwise correlations, uses the cofactors and standard deviations to determine the multiple regression coefficients, and puts out matrix *C*. For the matrix *C* and variables $x = \{x_1, x_2, x_3\}$, the relation is $x == C\,x + \epsilon$. The variable that is explained has a zero coefficient. The matrix *C* of regression coefficients ("CovarRegress") is not symmetric since it matters in regression what the explained variable is.

For these data, some of the conclusions are as follows. Take Fitness as the explained variable. The data are in the order Shop, Colour, Fitness, so we take the third row of *C*. Fitness and Shop have a correlation coefficient of 61.8% and a regression coefficient of 66%, so that if one hat in the study were replaced from Shop1 to Shop2 then Fitness would move 0.66 (the {3, 1} cell) from fit to no-fit. Fitness and colour have a correlation

coefficient of 4%. If one unit is moved from Green to Blue then the number of not-fitting hats would rise by 0.2 (the {3, 2} cell).

The influence of more variables

We can also see the effect of using larger matrices and more dimensions. Recalling the correlations on using the border matrices only:

```
NominalCorrelationMatrix[mat, BordersOrAll → BorderMatrices] // N
```

$$\begin{pmatrix} 1. & -0.25 & 0.625 \\ -0.25 & 1. & -0.125 \\ 0.625 & -0.125 & 1. \end{pmatrix}$$

The correlation coefficient for the influence of shops on fitness was 0.625, but when we consider the 2 × 2 × 2 model then it rises to 0.66. Replacing a "typical hat" now requires us to assume something on colour and this has some influence. More dramatically, the relation between colour and fitness in the 2 × 2 case is -0.125 but changes to 0.04 in the 2 × 2 × 2 case. In general, using larger sizes and dimensions will move us further from the agreeable notions of the averages (of the risk difference). It is an option indeed to define regression coefficients using only condensed 2 × 2 tables, and neglect the submatrices. All in all, having a general expression might be more practical in the end.

Restatement of what the new results are

Up to now there was no straightforward way to express these notions for nominal data. What currently appears to exist in the literature is fragmented. In the below, we will discuss some frequently used methods and their fragmentation. There are methods for 2 × 2 tables. There is Cramer's V that is limited to the $m \times n$ world. Given the link between correlation and regression one might surmise that this very link also exists in the literature, so that e.g. Cramer's V on association also has a pendant in some notion of regression, or that e.g. logistic regression has a pendant in correlation, but this does not seem to be the case. Given the number of pages already created, Colignatus (2007g) compares nominal regression to specifically logistic regression.

Now that we have seen what "nominal correlation and regression" mean in terms of geometry and numerical expression, let us delve deeper into the proper meaning, and also link up to existing methods while doing so. This brings us to the discussion of the leading ideas as seen from that angle.

Leading ideas

■ The volume ratio idea

The general idea is that the determinant of a matrix measures the change in volume when mapping a body from one space to another. We can also find a specific determinant that lies between -1 and 1. The value 1 means that the contingency table has a diagonal and that all off-diagonal terms are zero. Information on one variable is equivalent to information on the other variable. The value 0 means that the data are evenly spread across all cells. Information on one variable tells us nothing about the other one. Absolute values between 0 and 1 arise when the data move from statistical independence to full structure. Thus we have a normalized volume ratio that can be interpreted as correlation. Note that the correlation coefficient for real data has the geometric interpretation of the cosine of the angle between the vectors. For nominal data we thus don't find a cosine but the important point remains that there is a meaningful interpretation. Perhaps the true position should be the other way around that the standard is a volume ratio and that real data don't quite follow that standard but use a cosine instead. Perhaps there is a way that the volume ratio can be translated into a cosine interpretation. For now it suffices that we consider volumes only.

■ Existing options

With respect to the statistical reference guides - and see in particular the accessible sources on the internet Becker (1999), Garson (2007) and Losh (2004) - it must be observed that we find that there are actually various possible measures of association for contingency tables. This is another way of saying that there is no standard. The multitude of measures does not present paradise but rather a tropical forest or maze. We find Phi, (Pearson's) Contingency Coefficient, Tschuprow's T, Cramer's V, Lambda, (Theil's) Uncertainty Coefficient and tetrachoric correlation. **Appendix A** reviews these measures and rejects them for various reasons. A key point is that some measures are not symmetric. Another key point is that the most promising measures Phi, (Pearson's) Contingency Coefficient and Cramer's V all depend upon the χ^2statistic that is not elementary and requires quite a bit of statistical theory. Cramer's V is the most promising measure but is limited to the $m \times n$ case, and thus cannot handle the hat shops example above. However, it appears that the proposed measure for nominal correlation is equal to Cramer's V for the 2×2 case. We might start with that case and develop the story from there. However, it is better to develop the new measure by its own logic, and once we have done so, it appears that the volume ratio measure and Cramer's V have different philosophies and different outcomes. **Appendix D** discusses conventional approaches to the 2×2 case. **Appendix B** discusses the use of Cramer's V in relation to the χ^2test and shows its complexity. The determinant measure and Cramer's V both range (in absolute value) from 0 to 1 but their scores have a different meaning and gradient so that the numbers cannot be translated directly. Readers who are versed in statistics and want to start on familiar grounds might consider starting there. The key distinction is that the volume ratio measures the closeness to *a* diagonal while the χ^2 allows for *any* diagonal, also cross-diagonals and subdiagonals. We thus can say that nominal correlation is stronger than the χ^2 measure or Cramer's V. Note that it is always useful to have more measures. When we have a contingency table say of size $5 \times 3 \times 7$ then this block of 105 counts can usefully be summarized, not really in only one, but likely some more index numbers. One number for each separate aspect.

■ Strength of association

If one holds that Cramer's V is correlation, then one must accept this also for the determinant / volume ratio measure, since it is only stronger correlation. As said, the volume ratio measure tests on *a* diagonal (and not more) and Cramer's V tests on *any* diagonal. The determinant measure might be called strong correlation and Cramer's V weak correlation, given that the first has stricter conditions ("and") than the latter ("or"). Having *a* diagonal within the determinant measure is a sufficient condition for Cramer's V but not conversely. We may also note that statistical independence destroys the weak condition and henceforth the strong condition.

```
SquareTruthTable[StrongCondition ⇒ WeakCondition] // Transpose
```

(StrongCondition ⇒ WeakCondition)

	StrongCondition	¬ StrongCondition
WeakCondition	True	True
¬ WeakCondition	False	True

It might be a good research strategy to first test whether there is any association and if so then proceed with checking whether this concerns a main diagonal or not. If no association is found then one can save the trouble of checking for a main diagonal. One might also work conversely, check whether there is a main diagonal, and if not, proceed with a weaker condition. In practice, one would run both routines, just to be sure, anyhow.

■ The notion of overall correlation

It is useful to take real-valued data as our point of reference. Consider a block of data $X = \{x_1, ..., x_k\}$ and $x_i \in \mathbb{R}^n$ real data vectors, $i = 1, ..., k$. Let the correlation matrix **R** contain the pairwise correlations between x_i and x_j. Let $\mathcal{R}$ = Det[**R**] the determinant of **R** and let $\mathcal{R}_{i,j}$ be that specific co-factor. The coefficient of determination or the squared multiple correlation coefficient for the OLS regression of the first variable on the others is $(R^2)_{1.2,..,k} = 1 - \mathcal{R} / \mathcal{R}_{1,1}$, see Johnston (1972:134). Overall correlation is defined as $R_O = \sqrt{1 - \mathcal{R}}$, or in squared form as $R^2 = R^2{}_O = 1 - \mathcal{R}$. We require $0 \le R^2{}_O \le 1$ and thus require that $0 \le \mathcal{R} \le 1$. From this we can also derive that **R** must be a positive semi-definite (PSD) matrix. **Appendix C** contains some supporting notions for real data, including the proof that $0 \le \mathcal{R} \le 1$ for such a PSD correlation matrix. For example, when all variables are uncorrelated then the off-diagonal elements in **R** are zero and Det[**R**] = 1, and then we see an overall correlation of $R^2 = R^2{}_O = 0$. Similarly, for two variables and r the pairwise correlation then **R** = {{1, r}, {r, 1}} and then we find $R^2 = 1 - \text{Det}[\mathbf{R}] = 1 - (1 - r^2) = r^2$ as it should be.

NB. For ease of notation, this paper uses $R^2 = R^2{}_O$ for squared overall correlation and $(R^2)_{1.2,..,k}$ for the squared multiple correlation, or the coefficient of determination, "R-Squared". In general though, the notation for "R-Squared" must be maintained, and thus R_O is the better notation for overall correlation outside of this paper.

■ Two uses of the determinant

There is a key distinction here. In matrix **R** a linear dependence causes a zero determinant, as for example would happen with two equal rows (and henceforth two columns, given symmetry). When two variables have the same correlation pattern then they probably are correlated themselves too. In that case we want overall correlation $R = \sqrt{1 - \mathcal{R}}$ to express that **R** contains a strong correlation between (some of) the data indeed. Thus algebraic dependence in **R** gives overall correlation of 1 indeed. This is a bit different for contingency tables. In the square contingency table M a linear dependence causes a zero determinant as well, yet in this case the algebraic dependence means statistical independence. In that case we want to express that, with its categories, the data are not correlated. Hence there are two distinct realms where we use the determinant, not to be confused. See also **Appendix M.**

■ Association and correlation

Apparently, real data are such a paradise for analysis that the question about an overall correlation is not urgent and has slipped from much discussion. The issue is more urgent for nominal data since we quickly get higher order contingency tables, and those tables frequently come as a block of data without obvious order. One of the key questions is also that we might have "correlation matrices" that are *not* PSD to start with and can only be approximated by an associated PSD matrix that is "close" to it. This issue requires some sharp definitions. Suppose that we define a measure of "correlation" between two nominal data in a contingency matrix, and make sure that pairwise correlations are $-1 \le r \le 1$. Suppose that we have k such variables and collect the pairwise "correlations" in a matrix **R**. Are we allowed to call this matrix a "correlation matrix" ? Students from the school of real data analysis (one part of the brain of the author) will say "no" since there is no proof that this matrix will be PSD. Students from the school of nominal data analysis (another part of the brain of the author) wil say "why not ?", meaning that we generalize the concept and allow a matrix to contain pairwise correlations, subsequently impose $0 \le \mathcal{R} \le 1$ and thus approximate PSD-ness if the need may arise. The issue can be resolved by using the sharp definition that a collection of pairwise correlation is called an "association matrix" and is only accepted as a "correlation matrix" if it is also PSD. A consequence is that the "pairwise correlation" has two values, the original one and the one in the PSD approximation. This may still be close to blasphemy for the school of real data analysis yet there is nothing to it since those values are well-defined now. They only must be properly used and practice makes perfect.

The situation for both real and nominal data can be summarized in the general point of the choice of some f such that overall correlation within the data is $R = f[\mathbf{R}]$ with **R** now defined on pairwise correlations only.

Once the sharp distinction between association and correlation in above technical sense is clear and has been accepted, it appears in practice that this sharpness only is relevant for actual calculations and interpretations of the data. For general discussion the terms still can be and must be used interchangeably. To say for example that hat fitness and hat colour are associated but not correlated strains the use of the English language. Texts need to be lively and it helps when there are more terms to express one's thoughts. It would be pedantic to use the symbol **A** for the association matrix and **R** for the correlation matrix, except of course when we sit down for the formulas in order to specify the approximation algorithm for the computer. Also given the general scarcity of

symbols we still can use **R** for general use while it is kept in mind how we handle issues technically.

This generalization procedure on putting pairwise correlations in **R** has also been implemented for Cramer's V, generalizing its use from $m \times n$ to the higher dimensions. It needs to be seen how this relates to a more straightforward adjustment of the χ^2 for multidimensional tables.

■ Warranting positive semi-definiteness

For the volume ratio measure for contingency data there is some internal dependency such that a volume ratio in one dimension restricts ratios in other dimensions. **Appendix H** contains a simulation where cell values can differ from 1 to a million, and we find that 90% of the absolute volume ratio scores are between 0 and 1. If we neglect the inner variation of the table and base the pairwise correlations only on the border sum matrices then 100% of the cases are between 0 and 1. This is only indicative since there is no mathematical proof yet that those outcomes will always be like that. The simulation doesn't use zero entries yet to keep issues simple. These results nevertheless give some indication that the new proposed measure might be of some practical use. The latter will be enhanced when we find an acceptable method to approximate PSD-ness - which is complex way of saying that we make sure that $0 \leq R^2 \leq 1$ and by implication that $0 \leq$ Det[**R**] ≤ 1. Notably, when $\mathbf{R}_B$ is the association matrix containing the pairwise correlations using the border matrices only and when $\mathbf{R}_A$ is the association matrix containing the pairwise correlations that uses all inner variation, then we can maximize λ such that $0 \leq \lambda \leq 1$ and $0 \leq$ Det[$\mathbf{R} = (1 - \lambda)\ \mathbf{R}_B + \lambda\ \mathbf{R}_A$] ≤ 1. A statistical report would not only mention **R** but also λ. This gives just one example of the possibilities for adjustment. When it would appear that it is not guaranteed that $0 \leq$ Det[$\mathbf{R}_B$] ≤ 1 and that such a λ can be found then also technical approximations are possible. See **Appendix J** for such technical methods of PSD approximation. Higham (1989) presents the minimum of a Frobenius distance (Euclidean norm), we can also work directly on the eigenvalues, and there are also straightforward numerical methods that try to preserve some "grid".

■ Overall correlation versus multiple correlation

We are used to think in terms of multiple correlations. Since statistical practice on real data doesn't make much use of overall correlation $R^2 = = R^2_O = 1 - \mathcal{R}$ we have less experience with this notion. A score might be less informative when we are not used to interpreting such values. Given that we are most accustomed to the (squared) multiple correlation coefficient, an alternative measure of overall correlation is to consider the whole series $(R^2)_i$ and take the maximal value. When it holds for one i that $\mathcal{R}_{i,i} = 1$ then this again collapses to $1 - \mathcal{R}$. It seems to make most sense for current practice to say that the overall correlation within a block of data is given by the maximal value that we can find in it. Of course, we might consider a normal or geometric average too. Considering all the options it seems that using $R^2 = 1 - \mathcal{R}$ or such averages lose out against simply taking the maximal value of all multiple correlations. For both types of data, it is thus suggested that the default *f* just selects the highest value of any multiple correlation in the data.

PM. The routine OverallCorrelation allows one the choice and multiple correlation is only the default. The notion of OverallCorrelation thus is not to be confused with NominalCorrelation. NominalCorrelation only calls OverallCorrelation when the input is either an association or correlation matrix or a higher dimensional block of data that can be reduced to such.

■ Regression coefficients

Once we have a correlation matrix for contingency tables, it is straightforward to define a measure of variance as well, and then determine regression coefficients as we already can do for real data (Johnston (1972:133)). **Appendix K** discusses the choice of the variance.The main problem is that it requires a bit of care to properly interprete these regression coefficients. The interpretation should link up to the definition of the variance. See also Colignatus (2007g).

■ What the correlation coefficient is useful for

A correlation coefficient only gives some rough indication of association. The way of association, e.g. linear or non-linear, is a different issue and depends upon the model and its tests. However, the correlation coefficient has some psychological value for research on real data and undoubtedly will have the same effect for nominal data as well. A strong association for example might cause more curiosity as to what the true model is.

■ The (other) appendices

See **Appendix E** for an example in higher dimensions. See **Appendix F** for a note on the Frobenius theorems on nonnegative (contingency) matrices. See **Appendix G** for a note on causality, and how this paper originated. See **Appendix H** for a formal definition that the nominal correlation matrix is positive semidefinite. See **Appendix I** for an overview of the variants for the measure: {border matrices versus all submatrices}, {overall correlation versus multiple correlation}, {PSD adjustment or not}. See **Appendix J** for methods of PSD approximation. **Appendix K** discusses regression, though focusses on the selection of the variance and leaves a comparison with logistic regression for Colignatus (2007g). See **Appendix L** for the sign of a non-square matrix. **Appendix M** contains notes on regression. See **Appendix N** for an application to two more practical examples. See **Appendix O** for an explanation of a routine used to handle contingency tables. See **Appendix P** for a list of other routines used here.

Conclusion

The crux in this development lies in the $m \times n$ case. It covers the lower orders, and it forms the core for the $n_1 \times n_2 \times ... \times n_k$ generalization. Since the $m \times n$ case has a sound interpretation, the overall interpretation is sound. There is an element of arbitrariness in the methods of aggregation in the upward generalization. Namely, (i) there is the use of a weighted average for submatrices of pairwise correlations, (ii) the use of the maximal value of multiple correlations in the total correlation matrix, and (iii) the method to warrant positive semi-definiteness. Though arbitrary as this seems, each step has merit, and as a standard it is well defined.

The suggested measure has a useful interpretation as the volume ratio, with values between -1 and 1. The meaning of the index can be translated conceptually as measuring the distance from *a* diagonal (anywhere but unique). The main comparison to existing measures is to the χ^2 measure in general and Cramer's V in the $m \times n$ case, see **Appendix B**. Compared to the χ^2 scores and tests currently in use, the suggested measure has the added value of indicating the overall strength of association. That a deviation from independence is statistically significant or not, at some level of significance, need not be the most meaningful message when researching an issue. Compared to Cramer's V, it must be observed that the two indices measure different aspects of a contingency table and thus derive their usefulness from these two different angles. Cramer's V measures the distance to *any* diagonal (more at the same time, cross-diagonals and subdiagonals). This is weaker than the suggested measure for overall correlation.

Considering all aspects, it makes sense to call the volume ratio & determinant measure "nominal correlation". As said, this is a suggestion only, and only applies for the scope of this paper. The measure has been implemented as such in The Economics Pack and can be used to generate more experience.

Appendix A: Measures of association mentioned by common resources

There seems to be no standard and satisfactory measure of association for nominal data. Apart from the official books mentioned in the list of references, like Mood & Graybill (1963) and Kleinbaum et al. (2003), also the resources on the internet mentioned there, notably Becker (1999), Garson (2007) and Losh (2004), have been used to find measures for the association in nominal data. See Cool, Th. (1999, 2001), "The Economics Pack, Applications for *Mathematica*", and the website update, for an implementation of the LLR and Pearson tests, in the CrossTable package, and for some implementations in the Life Sciences packages (all created in 1993-2004). The following measures have been found.

(1) Fisher's exact test does a test and does not provide a measure of association. Similarly for the likelihood ratio test and the Pearson approximation. The measure then is yes/no with respect to passing the test.

(2) The Odds Ratio depends upon direction (column-wise versus row-wise) and it is unknown whether this is generalizable. For the 2 by 2 table, taking the default column direction:

OddsRatio[]

$$\frac{a\,d}{b\,c}$$

(3) "The tetrachoric correlation coefficient is essentially the Pearson product-moment correlation coefficient between the row and column variables, their values for each observation being taken as 0 or 1 depending on the category it falls into"

(4) Phi (Cramer's V in non-square tables): "Also in 2-by-2 tables, phi is identical to the correlation coefficient. In larger tables, where phi may be greater than 1.0, there is no simple intuitive interpretation, which is a reason why phi is often used only for 2-by-2 tables." (Garson (2007))

PM 4.1. The statement "Also in 2-by-2 tables, phi is identical to the correlation coefficient" is confusing since that correlation coefficient is not clearly defined for

nominal data. As with the "tetrachoric" measure, one takes {1, 0} assignments to the nominal values, but are these also the values for Phi ?

PM 4.2. There are different ways to determine a χ^2 value. The Pearson test statistic is $(o - t)^2 / t$, with o the observed and t the theoretical frequency. For a 2 by 2 table the theoretical frequency might come from the hypothesis of independence. But other hypotheses are possible too.

PM 4.3 Cramer's V is the most useful of all these possible measures. Yet its interpretation is the χ^2 from the hypothesis of statistical independence, and one wonders whether this captures the intuition of correlation as given by the real variables.

Cramer's V, defined for a $m \times n$ matrix, is the square root of the Pearson χ^2 value divided by the sample size p times q, where q is the smaller of (m - 1) and (n - 1). Thus $V = \sqrt{\chi^2 / (p\, q)}$. For a 2 × 2 case we find q = 1, and then V reduces to Phi (that always takes q = 1).

PM 4.4. Since the χ^2 measure can be extended in higher dimensions (see the CrossTable package), one feels that this might be possible with this V too. (The CrossTable package might be used to check upon value 1 for all diagonals.)

See **Appendix B** for a longer discussion.

(5) The Contingency Coefficient, Pearson's C = SQRT[$\chi^2/(\chi^2$ + n)]. "There is no easily intuited interpretation of C or C*, though C* may be viewed as the association between two variables as a percentage of their maximum possible variation. Pearson viewed C as a nominal approximation of Pearsonian correlation r." (Garson (2007))

(6) The Uncertainty Coefficient, UC or Theil's U - an asymmetric measure.

(7) Hoeffding's Dependence Coefficients - not looked into.

(8) Eta is an asymmetric correlation coefficient, and its dependent variable would be interval scaled.

Appendix B: Relation to the χ^2 measure

■ In general

A standard procedure in the analysis of contingency tables is to perform the Pearson test on independence, which is an approximation of the log-likelihood ratio test. Independence arises when the cells in the matrix are mere products of the marginals, given in the border sums. With o the observed data and t the theoretical frequencies, derived from those products given by the hypothesis of independence, the Pearson test statistic is $\Sigma\ (o - t)^2 / t$. When we introduce additional assumptions on the distribution then we can perform a test. A possible assumption is a multinomial distribution and for larger numbers this gets closer to the multivariate normal, such that the Pearson test statistic would have a χ^2distribution. Mood & Graybill (1963:314) clarify the procedure: "In casting about a test which may be used when the sample is not large, we may inquire how it is that a test criterion comes to have a unique distribution for large samples when the distribution actually depends on unknown parameters which may have any values in certain ranges. The answer is that the parameters are not really unknown; they can be estimated, and their estimates approach their true values as the sample size increases."

The most useful existing measure for association in crosstables is Cramer's V, defined for a $m \times n$ matrix, as the square root of the χ^2 value divided by the sample size p times q, where q is the smaller of (m - 1) and (n - 1). Thus $V = \sqrt{\chi^2 / (p\,q)}$. For a 2 × 2 case we find q = 1, and then V reduces to Phi (that always takes q = 1).

Thinking about association within a table along this route is not simple and plunges us into the deep waters of statistical hypothesis testing. We can recognize that the 2 × 2 case has the same form for Cramer's V as the suggested determinant measure for Nominal Correlation, yet, the reasoning behind it is quite different. To understand the relation between these two measures we must refresh our understanding of the χ^2 test on statistical independence. Their behaviour with respect to diagonals appears to be key. Some illuminating cases are cross diagonals and triangular matrices. We can also consider aggregating a higher order matrix and check the consistency or change of the values.

In the discussion below we will consider the 2 × 2 case, the 3 × 3 case and some higher dimensions where we focus on the role of diagonals. In this discussion symmetric matrices are used, just for simplicity, without affecting the generality of the conclusion.

■ The 2 × 2 case

The routine Test does the Pearson χ^2 test of $n_1 \times ... \times n_k$ contingency tables. The routine was written in 1993 - 1995 and has been in The Economics Pack since. We find fruitful employment for it again in this discussion, also using its feature that it can explain the various intermediate steps. In output, the rule **Do → Accept | Reject** expresses whether the hypothesis of independence is accepted or rejected at the stated significance level. The routine normally uses N[.] to prevent long expressions yet in this case we want to Rationalize[.] again. We presently consider the 2 × 2 case.

test =

```
Test[Chi2, Pearson, {{a, b}, {c, d}}] /. x_?NumberQ :> Rationalize[x] // Simplify
```

Observed

a	b
c	d

Theoretical

$\frac{(a+b)(a+c)}{a+b+c+d}$	$\frac{(a+b)(b+d)}{a+b+c+d}$
$\frac{(a+c)(c+d)}{a+b+c+d}$	$\frac{(b+d)(c+d)}{a+b+c+d}$

Pearson (o − t)^2 / t

$\frac{(a+b+c+d)\left(a-\frac{1.(a+b)(a+c)}{a+b+c+d}\right)^2}{(a+b)(a+c)}$	$\frac{(a+b+c+d)\left(b-\frac{1.(a+b)(b+d)}{a+b+c+d}\right)^2}{(a+b)(b+d)}$
$\frac{(a+b+c+d)\left(c-\frac{1.(a+c)(c+d)}{a+b+c+d}\right)^2}{(a+c)(c+d)}$	$\frac{(a+b+c+d)\left(d-\frac{1.(b+d)(c+d)}{a+b+c+d}\right)^2}{(b+d)(c+d)}$

$$\left\{\text{ArrayDepth} \to 2,\ \text{BorderSums} \to \begin{pmatrix} a+b & c+d \\ a+c & b+d \end{pmatrix},\right.$$

$$\text{Chi2PValue} \to 1 - Q\left(\frac{1}{2},\ 0,\ \frac{(a+b+c+d)(bc-ad)^2}{2(a+b)(a+c)(b+d)(c+d)}\right),\ \text{DegreesOfFreedom} \to 1,$$

$$\text{Dimensions} \to \{2, 2\},\ \text{Do} \to \text{If}\left[20\,Q\left(\frac{1}{2},\ 0,\ \frac{(a+b+c+d)(bc-ad)^2}{2(a+b)(a+c)(b+d)(c+d)}\right) \geq 19,\ \text{Reject},\ \text{Accept}\right],$$

$$\text{MarginalPr} \to \left\{\text{MarginalPr}(1) \to \left\{\frac{a+b}{a+b+c+d},\ \frac{c+d}{a+b+c+d}\right\},\right.$$

$$\left.\text{MarginalPr}(2) \to \left\{\frac{a+c}{a+b+c+d},\ \frac{b+d}{a+b+c+d}\right\}\right\},\ \text{NumberOfObservations} \to a+b+c+d,$$

$$\text{Partition} \to \{1, 2\},\ \text{ProbabilityMatrix} \to \begin{pmatrix} \frac{(a+b)(a+c)}{(a+b+c+d)^2} & \frac{(a+b)(b+d)}{(a+b+c+d)^2} \\ \frac{(a+c)(c+d)}{(a+b+c+d)^2} & \frac{(b+d)(c+d)}{(a+b+c+d)^2} \end{pmatrix},\ \text{SignificanceLevel} \to \frac{1}{20},$$

$$\left.\text{Test} \to \text{Pearson},\ \text{TestStatistic} \to \frac{(a+b+c+d)(bc-ad)^2}{(a+b)(a+c)(b+d)(c+d)}\right\}$$

The object of interest is the test statistic. We can select it from above output and give it a name that expresses its distribution under the null hypothesis of independence.

chi2 = (TestStatistic /. test)

$$\frac{(a+b+c+d)(bc-ad)^2}{(a+b)(a+c)(b+d)(c+d)}$$

Cramer's V then is (for the 2 × 2 case):

```
Sqrt[chi2 / (NumberOfObservations /. test)]
```

$$\sqrt{\frac{(b\,c-a\,d)^2}{(a+b)\,(a+c)\,(b+d)\,(c+d)}}$$

Note too, that, in general, it seems that one would be more interested in the measure of association and its confidence interval rather than the test on independence. Given that the χ^2 value and its distribution have been given and Cramer's V differs only by a constant, it should not be too difficult to determine these.

■ A 3 × 3 case

With respect to above 3 × 3 case, Cramer's V suggests some modest association.

```
CramersV[Take[mat5, -3, -3]]
```

0.197584

We find that the hypothesis of independence would be rejected at the standard 5% significance level. Note that this test result actually derives from the χ^2 test procedure in the background that is used to construct the Cramer's V measure.

```
Do /. Results[CramersV]
```

Reject

Note however this crucial observation, Cool (1995, 2001:368) discussing some other examples: "The χ^2 test is sensitive to the number of observations per degree of freedom. (... When ...) there are more observations per degree of freedom, (...) the test is stronger. (In the same manner, by increasing the sample size, acceptance tends to turn into rejection, and by reducing the size, rejection can turn into acceptance.)" Thus a statistically significant deviation from independence in a χ^2 test does not imply that there would be much of an association.

The Cramer V outcome of 0.197584 differs importantly from the outcome of Nominal Correlation of 0.0285548. At this point there seems little to be said about this. Different measures, different results. But one cannot evade the impression that Cramer's V is sensitive to the quirks of χ^2 testing. The point that Cramer's V is equal to the volume ratio measure for the 2 × 2 case saves it though, and we need to embark upon the longer discussion below.

■ The general principle

The key point

The determinant measure NominalCorrelation looks at the degree of dominance of *a* diagonal of the matrix (and not more), for any arbitrary permutation. Cramer's V looks at *any* diagonal, also subdiagonals or cross-diagonals. In the following we shall be using square matrices for simplicity. To emphasize neutrality and technicality we use the routine SquareMatrixNormedDet rather than VolumeRatio or NominalCorrelation (though they are technically similar). The determinant measure arrives at 1 or -1 and Cramer's V arrives at 1 when the categories in the dimensions can be sorted such that there appears a neat diagonal. In this respect the measures are similar. The key point is that they differ when the diagonal is less important.

```
perms = Permutations[DiagonalMatrix[Table[Random[Integer, {1, 10}], {i, 4}]]];

SquareMatrixNormedDet /@ perms
```

```
{1, −1, −1, 1, 1, −1, −1, 1, 1, −1, −1, 1, 1, −1, −1, 1, 1, −1, −1, 1, 1, −1, −1, 1}
```

```
CramersV /@ perms
```

```
{1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1.}
```

The measures differ however with respect to subdiagonals and cross diagonals.

Cross diagonals

Illuminating are the cross diagonal matrices, where a score in one dimension only allows some limited (and diverging, in the order of presentation) possibilities in the other dimension. The determinant measure concludes that these dimensions are not correlated (the cross-diagonal destroys the diagonal) yet Cramer's V picks up a pattern.

```
crossdiag[n_] := Table[If[i == j || i == n − j + 1, 1, 0], {i, n}, {j, n}];
```

```
(crossdiagonal =
    {{crossdiag[2], crossdiag[3], crossdiag[4], crossdiag[5], crossdiag[6]}
    }) // TableForm
```

1 1 1 1	1 0 1 0 1 0 1 0 1	1 0 0 1 0 1 1 0 0 1 1 0 1 0 0 1	1 0 0 0 1 0 1 0 1 0 0 0 1 0 0 0 1 0 1 0 1 0 0 0 1	1 0 0 0 0 1 0 1 0 0 1 0 0 0 1 1 0 0 0 0 1 1 0 0 0 1 0 0 1 0 1 0 0 0 0 1

```
Map[({SquareMatrixNormedDet[#], CramersV[#]} // N) &, crossdiagonal, {2}]
```

({0., 0.} {0., 0.707107} {0., 0.57735} {0., 0.707107} {0., 0.632456})

Triangular matrices

Another illuminating case is triangularity. The case of a triangular contingency table is intriguing. When we can rearrange a table so that it shows a triangle then we know that with an "increasing" score in one dimension we have more certainty in the sense of less (and also "increasing") categories in the other dimension. This is a specific kind of regularity and association but not necessarily "correlation" when we take the latter in the (strong) sense that a main diagonal in some symmetric manner wins out over the off-diagonal elements.

```
(triangular = {{Table[If[i ≥ j, 1, 0], {i, 4}, {j, 4}],
        Table[Which[ i == j, 1, i ≥ j, Random[Integer, {0, 10}], True, 0], {i, 4}, {j, 4
        Table[If[i ≥ j, Random[Integer, {0, 10}], 0], {i, 4}, {j, 4}]}
    }) // TableForm
```

1 0 0 0 1 1 0 0 1 1 1 0 1 1 1 1	1 0 0 0 5 1 0 0 4 4 1 0 7 0 10 1	6 0 0 0 2 8 0 0 7 2 9 0 8 6 10 10

```
res1 = Map[({SquareMatrixNormedDet[#], CramersV[#]} // N) &, triangular, {2}]
```

({0.0416667, 0.375771} {0.00104897, 0.419612} {0.0852573, 0.463423})

Aggregation

When we aggregate nominal categories from a $n \times n$ case to a 2×2 level then the most important effect is that the diagonal increases in value. What first was off-diagonal now becomes diagonal. Precisely for this reason the value for the determinant measure rises strongly. Cramer's V maintains a more stable value, since it already took this effect into account but taking along the subdiagonals. PM. For simplicity the following aggregation is symmetric but this does not have to be so. PM. Having the 2×2 case, the measures of course use the same formula. We do the operation to show the differential effect due to aggregation.

```
agg = {{1, 1, 1, 0}, {0, 0, 0, 1}};
TableForm[(triangular /. (x_?SquareMatrixQ :> agg.x.Transpose[agg]))]
```

```
6  0      16  0      34  0
3  1      17  1      24  10
```

```
res2 = Map[({SquareMatrixNormedDet[#], CramersV[#]} // N) &, %, {2}]
```

```
({0.408248, 0.408248}  {0.164122, 0.164122}  {0.415227, 0.415227})
```

```
res2 – res1
```

```
({0.366582, 0.0324775}  {0.163073, −0.25549}  {0.32997, −0.048196})
```

Zero diagonal

What is important in the example above is that the diagonals are not zero. With a zero diagonal, the determinant measure is indeterminate while Cramer's V still produces a value. Cramer's V namely looks at subdiagonals, which means that some aggregation can produce diagonal elements. We can check this by setting the diagonals of above matrices to zero and performing the relevant aggregation.

```
TableForm[(triangular /. (x_?SquareMatrixQ :> x – 10 IdentityMatrix[4])) /.
    (y_?NumberQ) :> If[y < 0, 0, y]]
```

```
0  0  0  0      0  0  0  0      0  0  0  0
1  0  0  0      5  0  0  0      2  0  0  0
1  1  0  0      4  4  0  0      7  2  0  0
1  1  1  0      7  0  10 0      8  6  10 0
```

```
Map[({SquareMatrixNormedDet[#], CramersV[#]} // N) &, %, {2}]
```

```
({Indeterminate, 0.346944}  {Indeterminate, 0.485071}  {Indeterminate, 0.291258})
```

And this is a possible aggregation of the categories such that diagonal elements arise. Yet, while the determinant measure recovers, Cramer's V collapses, meeting on the common ground of the 2 × 2 case again.

```
agg = {{1, 1, 0, 0}, {0, 0, 1, 1}};
TableForm[(%%%/. (x_?SquareMatrixQ :→ agg . x . Transpose[agg]))]
```

```
1  0      5   0      2   0
4  1      15  10     23  10
```

```
Map[({SquareMatrixNormedDet[#], CramersV[#]} // N) &, %, {2}]
```

```
( {0.2, 0.2}   {0.316228, 0.316228}   {0.1557, 0.1557} )
```

What next

By way of arriving at a sub-conclusion, we see the impact both of the diagonals and the confounding by the 2 × 2 case. In the latter case the two philosophies meet but the room is too small to let them bloom in full. This brings us at the next subsection, where we replace numerics with analytics.

■ A parametric 3 × 3 case

To get more grip on the problem we can consider a parametric example. The following case has some theoretical value. (1) We use the fact that both measures are insensitive to the total number of observations *N*. Thus we can normalize, say, and also taking a bit more flexibility, to one row with unit values. (2) We can reduce the size of the problem by taking a subblock of zeros such that the determinant of the 3 × 3 case is the same as the determinant of the lower 2 × 2 submatrix.

```
matpar = {{a, b, 0}, {c, d, 0}, {1, 1, 1}}
```

$$\begin{pmatrix} a & b & 0 \\ c & d & 0 \\ 1 & 1 & 1 \end{pmatrix}$$

We find that the denominators of the two measures have the same products of the border sums. But they differ strongly in their numerators.

```
dt = SquareMatrixNormedDet[matpar]
```

$$\frac{a\,d - b\,c}{\sqrt{3}\,\sqrt{(a+b)\,(a+c+1)\,(b+d+1)\,(c+d)}}$$

```
cv = CramersV[matpar] /. x_?NumericQ :> Rationalize[x] // Simplify;
```

```
Denominator[cv // PowerExpand // Simplify]
```

$$6\sqrt{a+b}\ \sqrt{a+c+1}\ \sqrt{b+d+1}\ \sqrt{c+d}$$

```
Numerator[cv // PowerExpand // Simplify]
```

$$\sqrt{6}\ \sqrt{((c(b+d+2)+d(b+4d+2))a^2+((c+d)b^2+(c^2-2(2d+1)c+(d-2)d)b+2d^2+c(d-2)d+c^2(d+2))a+b((d+2)c^2+(d-2)dc+2d^2+b(4c^2+(d+2)c+2d)))}$$

For reference, we may also look at what happens in the 2 × 2 sub-case and include it in the list of our formulas. Also, it appears to be most illuminating to consider the case that $c = 0$. The determinant measure then appears to be a decreasing function of b only, with a limit of 0. Cramer's V first has a dip but then is an increasing function with limit 1/2. We can show the formulas, a table of values, and a plot.

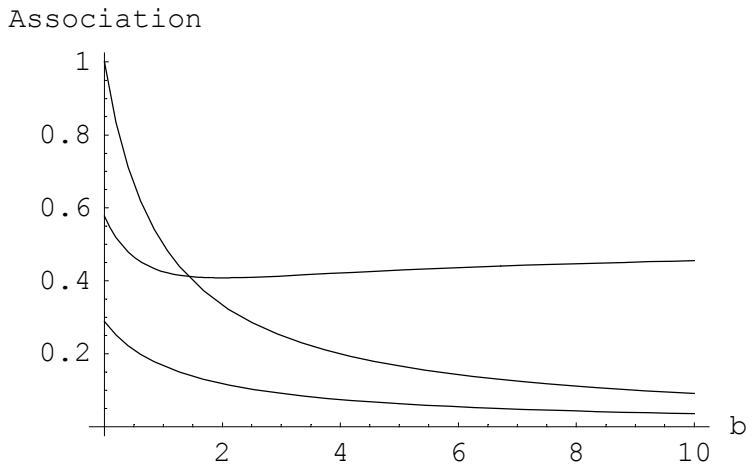
```
measures[b_] =
  {CorrelationPr2By2[{{a, b}, {c, d}}], dt, cv} /. {a → 1, c → 0, d → 1} // Simplify
```

$$\left\{\frac{1}{\sqrt{(b+1)^2}},\ \frac{1}{\sqrt{6}\ \sqrt{(b+1)(b+2)}},\ \frac{1}{2}\sqrt{\frac{3b^2+2b+8}{3b^2+9b+6}}\right\}$$

```
TableForm[measures /@ {0, 1, 10, 10^10} // N // Chop,
  TableHeadings → {{0, 1, 10, Infinity}, {"2 × 2 subcase", Det, CramersV}}]
```

	2 × 2 subcase	Det	CramersV
0	1.	0.288675	0.57735
1	0.5	0.166667	0.424918
10	0.0909091	0.0355335	0.45505
∞	0	0	0.5

```
Plot[Evaluate[measures[b]], {b, 0, 10}, AxesLabel → {b, "Association"}];
```



Given that we might multiply the matrix with *N,* the change in scores from 0 to 1 is important. For the 2 × 2 reference case all measures have the same formula and {{1, 0}, {*b*, 1}} generates the correlation of 50% for $b = 1$. For the 3 × 3 case the Cramer's V 15 percentage points drop between 0 and 1 can best be interpreted as an approximation of the change from 58% at 0 to the limit value of 50%. The basic analysis is that Cramer's V gives a score of about 50% to a triangular matrix, give or take some percentage points around it. The Det measure on the other hand behaves the same for 2 × 2 and 3 × 3. As the weight in the lower triangular matrix rises in comparison to the diagonal, the correlation drops. It is not quite right to consider {{1, 0}, {*b*, 1}} as a lower triangular matrix since *b* is just a cell, yet from this angle it makes some sense.

A question might be whether we want the correlation score for triangular matrices to fall with their weight or to be stable and consistently around 50% ? This question might be a bit misleading as we might usefully employ both measures, yet it is a good question to clarify what the measures do. The best way to answer this question seems to be to consider the 2 × 2 reference case and in particular the matrix {(1, 0}, {1, 1}}. Though the determinant measure and Cramer's V have the same formula, they still have entirely different philosophies. Cramer's V assigns a value of 50% because the diagonal gets a score of 1 but the cross-diagonal gets 0, thus the average is 50%. (When there are two diagonals, then there are also two columns and two rows, and then the result is 0%.) The determinant measure assigns 50% because the main diagonal is offset by off-diagonal elements.

PM 1. We may also say that the result of the volume ratio measure is merely from the effect that the formula works that way. The formula gives the ratio of the areas created

by the vectors, and 1 in that cell is just one value in the range between 0 and infinity. Since the determinant measure has such an interpretation it is easier to understand what the score expresses. To say something like that is less feasible with the Pearson statistic. It is a formula too but less tractable. PM 2. One might possibly argue that Cramer's V is inconsistent in that it does not assign 50% to the lower diagonal of {{1, 0}, {*b*, 1}} for any $b \neq 0$. However, a cell is not a lower triangular matrix if it comes to that, whatever said before, and thus the major issue are the (main, sub or cross) diagonals. The 2×2 case may just be a sandbox that allows too many philosophies and we need the perspective of more variables.

■ A higher dimensional case

Cramer's V is only defined for $m \times n$ matrices. But the χ^2 test can handle higher dimensions. We can run the test on the hat shops case, for some level of significance and some *p*-value. The hypothesis of independence is rejected at the 5% significance level. And then ? This rejection tells us little about the amount of association. The χ^2 *p*-value does not impress as a good measure for association either.

test = Test[Chi2, Pearson, mat]

Observed

5	8
1	2
2	1
8	5

Theoretical

4.	4.
4.	4.
4.	4.
4.	4.

Pearson (o − t)^2 / t

0.25	4.
2.25	1.
1.	2.25
4.	0.25

$$\left\{\text{ArrayDepth} \to 3,\ \text{BorderSums} \to \begin{pmatrix} 16 & 16 \\ 16 & 16 \\ 16 & 16 \end{pmatrix},\ \text{Chi2PValue} \to 0.000107511,\right.$$

DegreesOfFreedom → 1, Dimensions → {2, 2, 2}, Do → Reject, MarginalPr →
{MarginalPr(1) → {0.5, 0.5}, MarginalPr(2) → {0.5, 0.5}, MarginalPr(3) → {0.5, 0.5}},
NumberOfObservations → 32, Partition → {1, 2, 3},

$$\text{ProbabilityMatrix} \to \begin{pmatrix} \{0.125, 0.125\} & \{0.125, 0.125\} \\ \{0.125, 0.125\} & \{0.125, 0.125\} \end{pmatrix},$$

$$\left.\text{SignificanceLevel} \to 0.05,\ \text{Test} \to \text{Pearson},\ \text{TestStatistic} \to 15.\right\}$$

■ Generalizing Cramer's V

There are two ways to generalize Cramer's V to the higher $n_1 \times ... \times n_k$ dimensions: (1) to adjust the χ^2 score for those higher dimensions, (2) to follow the scheme as designed for the volume ratio, i.e. collect the pairwise results into an association or correlation matrix and then summarize that matrix in an overall correlation. Let us just do the latter. For the hat shops example (that is $2 \times 2 \times 2$) this procedure generates the same result for Cramer' s V as for the volume ratio, since the 2×2 formulas are the same, except for the sign of course. PM 1. We thus use a generalized PairwiseMeasureMatrix routine, that recognizes a PairwiseMeasure, either VolumeRatio or CramerV. PM 2. Rourtinewise speaking, it is not smart to allow just any pairwise measure, since it might be ill-defined, and generate a lot of intractable output. Thus it is useful to have a test on routines that are recognized. PM 3. This paragraph neglects the technical issue of the difference between "association" and "correlation".

```
TableForm[{{N[PairwiseMeasureMatrix[mat, PairwiseMeasure →
VolumeRatio]],
    N[PairwiseMeasureMatrix[mat, PairwiseMeasure → CramersV]]}},
TableHeadings -> {{}, {VolumeRatio, CramersV}}]
```

VolumeRatio			CramersV		
1.	−0.221917	0.61807	1.	0.221917	0.61807
−0.221917	1.	0.0413449	0.221917	1.	0.0413449
0.61807	0.0413449	1.	0.61807	0.0413449	1.

Differences arise for the higher dimensions. When we use the triangular matrices above then it appears that these contain some zero columns so that pairwise volume ratios are indeterminate, while Cramer's V still produces values.

```
mattr = triangular[[1]];

TableForm[{{N[PairwiseMeasureMatrix[mattr, PairwiseMeasure →
VolumeRatio]],
    N[PairwiseMeasureMatrix[mattr, PairwiseMeasure →
CramersV]]}}, TableHeadings -> {{}, {VolumeRatio, CramersV}}]
```

VolumeRatio			CramersV		
1.	Indeterminate	Indeterminate	1.	0.264402	0.2968
Indeterminate	1.	0.055802	0.264402	1.	0.442297
Indeterminate	0.055802	1.	0.2968	0.442297	1.

The above result of course depends upon the parameters of the selected (default) methods. When we restrict our attention to the border matrices then we get non-indeterminate results (in this case).

```
bm = BorderMatrices[mattr]
```

$$\left\{\{1, 2\} \to \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 6 & 9 & 18 \\ 6 & 10 & 18 & 34 \end{pmatrix}, \{1, 3\} \to \begin{pmatrix} 4 & 3 & 2 & 1 \\ 17 & 5 & 11 & 1 \\ 23 & 16 & 19 & 10 \end{pmatrix}, \{2, 3\} \to \begin{pmatrix} 8 & 0 & 0 & 0 \\ 8 & 10 & 0 & 0 \\ 12 & 7 & 11 & 0 \\ 16 & 7 & 21 & 12 \end{pmatrix}\right\}$$

```
TableForm[{{N[PairwiseMeasureMatrix[mattr, PairwiseMeasure →
VolumeRatio, BordersOrAll →  BorderMatrices]],
N[PairwiseMeasureMatrix[mattr, PairwiseMeasure → CramersV,
BordersOrAll →  BorderMatrices]]}}, TableHeadings -> {{},
{VolumeRatio, CramersV}}]
```

VolumeRatio			CramersV		
1.	−0.00629941	0.0160951	1.	0.0884428	0.164769
−0.00629941	1.	0.0337156	0.0884428	1.	0.364248
0.0160951	0.0337156	1.	0.164769	0.364248	1.

The transformation of a matrix of such correlation coefficients into a PSD correlation matrix is another issue. This is discussed in this paper elsewhere in the context of the volume ratio. The routine NominalCorrelationMatrix has been implemented such that this applies the volume ratios and constructs a PSD correlation matrix. This might be generalized for Cramer's V as well, yet, this can best be done in relation to theory (approach (1)), which has not been implemented at this moment of writing.

Appendix C: Overall correlation between real variables

■ The notion of overall correlation

Consider a block of data $X = \{x_1, ..., x_k\}$ and $x_i \in \mathbb{R}^n$ real data vectors, $i = 1, ..., k$. Let the correlation matrix **R** contain the pairwise correlations between x_i and x_j. The correlation matrix **R** is by definition a symmetric matrix with 1 on its diagonal and $0 \leq \text{Det}[\mathbf{R}] \leq 1$. It is only a secondary issue that it is positive semi-definite (PSD) and that the method of OLS creates such a matrix.

It is not customary of statistical textbooks and reference guides to discuss "overall correlation". The books and guides concentrate on both the pairwise correlations and the multiple correlation coefficient, where the latter arises when a particular $y = x_i$ for some i is selected as the dependent variable so that the others are the explanatory

variables. It appears that it is not common for real data to summarize that correlation matrix into one single number. The textbooks and guides will, at some point in the discussion, express that this **R** is a positive semi-definite (PSD) matrix. For a PSD matrix it holds that for any vector $v \neq 0$, v' **R** $v \geq 0$. This may seem mysterious at first. With OLS expression $y = X\beta$, then the variance-coveriance matrix is based upon $X'X$, which is a PSD matrix since $\beta' X' X \beta = y' y$ is a sum of squares. The latter also holds when the dependent variable y is included in the data X, and we define some arbitrary $z = X b$. Let $\mathcal{R}$ = Det[**R**] the determinant of **R** and let $\mathcal{R}_{i,j}$ be that specific co-factor. The multiple correlation coefficient for the OLS regression of the first variable on the others is $(R^2)_{1.2,..,k} = 1 - \mathcal{R} / \mathcal{R}_{1,1}$, see Johnston (1972:134). This is where statistical textbooks and reference guides tend to stop. It is not customary for statistical textbooks and reference guides to state that $0 \leq \mathcal{R} \leq 1$. At least, this holds before March 15 2007, and perhaps developments on the internet are fast and the situation has changed by the time that you read this. An improvement with respect to March 15 2007 would be: (i) to discuss overall association or correlation between a set of variables irrespective of order, (ii) define overall correlation as $R^2 = 1 - \mathcal{R}$ or $R = \sqrt{1 - \mathcal{R}}$, require $0 \leq R^2 \leq 1$ and deduce that $0 \leq \mathcal{R} \leq 1$, (iii) only then derive that **R** must be PSD for the latter to be true. Note the shift in presentation content and order. The condition of PSD-ness is just a consequence of requiring that $0 \leq \mathcal{R} \leq 1$. We only impose that condition since it is useful for something. The key point why it is useful is overall correlation and the possibility to derive arbitrary $(R^2)_i = (R^2)_{i.2,..,k} = 1 - \mathcal{R} / \mathcal{R}_{i,i}$, $i = 1, ..., k$. The fact that we can estimate the correlation matrix with $X' X$ is only secondary to these considerations, since if we would not be able to use that estimator then we would design another estimator. Precisely since it follows from $0 \leq \mathcal{R} \leq 1$ that **R** must be PSD brings us to using $X' X$ since it is PSD. The current practice in the textbooks and guides to slip in PSD-ness without much justification and with neglect of the concept of overall correlation turns the true argument upside down and deletes important information.

For nominal data, the steps (i) to (iii) above should be extended with: (iv) show for nominal data that PSD-ness may require an approximation. In that way (i) to (iv) make for a logical structure that is transparant in its simplicity. The situation for both real and nominal data can be summarized in the general point of the choice of some f such that overall correlation within the data is $R = f$[**R**] with **R** now defined on pairwise correlations only.

■ Derivation of 0 ≤ Det[R] ≤ 1

With proper theory, we impose $0 \leq \text{Det}[\mathbf{R}] \leq 1$ (next to $-1 \leq \mathbf{R}_{i,j} \leq 1$ and the unit diagonal) and then derive PSD-ness. However, in current practice PSD-ness is imposed and now we need to derive that $0 \leq \text{Det}[\mathbf{R}] \leq 1$.

As said, a possible measure for overall correlation is $R^2 = 1 - \mathcal{R}$ where $\mathcal{R} = \text{Det}[\mathbf{R}]$. Whatever we said about taking the maximal multiple correlation as the superior measure, there can still be value in just this determinant measure. The measure is valid if we can show that $0 \leq R^2 = 1 - \mathcal{R} \leq 1$ or $0 \leq \text{Det}[\mathbf{R}] \leq 1$. From the PSD property already $\text{Det}[\mathbf{R}] \geq 0$. For the proof on the upper bound, such that $R^2 = 1 - \mathcal{R} \geq 0$ or $\text{Det}[\mathbf{R}] \leq 1$, we now use the property that we have 1 on the diagonal, so that the trace is k. (PM. The diagonal is 1 since the correlation of a variable with itself is 1. It might suffice that the trace is k, allowing other values on the diagonal, but we also want to take submatrices, so the only consistent overall appearance is 1 on the diagonal.)

We can look at $k = 2$ and $k > 2$. Since **R** is a symmetric and positive semi-definite matrix, its determinant is given by the product of all its (nonnegative) eigenvalues λ_i, while the sum of those is also given by the trace of the diagonal, i.e. the dimension, see Johnston (1972:105-109). Thus we have $\text{Det}[\mathbf{R}] = \lambda_1 ... \lambda_k$ while $k = \lambda_1 + ... + \lambda_k$ and $\lambda_i \geq 0$. For singular **R** we trivially have $\mathcal{R} = 0$. For regular **R** and $k = 2$ we trivially have $0 \leq \mathcal{R} \leq 1$. Namely, for $k = 2$, we know that the correlation r between two variables satisfies $-1 \leq r \leq 1$ and the determinant is:

```
Det[{{1, r}, {r, 1}}]
```

$$1 - r^2$$

Using the overall correlation destroys information about the direction of the correlation:

```
OverallCorrelation[{{1, r}, {r, 1}}, Mode → Det]
```

$$\sqrt{r^2}$$

At issue remains only regular **R** and $k > 2$. A step in the proof is that the maximal determinant is given by $\lambda_i = \lambda$ for all i. We might show this by calculus, maximizing $\mathcal{R}$ subject to the given constraints, but it is simpler to assume that all eigenvalues are at their maximal value except for the first two. Then we get $\mathcal{R} = \lambda_1 \lambda_2 \lambda^{k-2}$. Since the sum is given, the difference between λ and λ_1 is reflected in λ_2, thus $\mathcal{R} = (\lambda - x)(\lambda + x)\lambda^{k-2} = (\lambda^2 - x^2)\lambda^{k-2} = \lambda^k - X$, where X is a nonnegative amount. Thus the determinant is

maximal when $x = 0$, and then all $\lambda_i = \lambda$. Then also $k = k\,\lambda$, or $\lambda = 1$, and $\mathcal{R} = 1$. This proof can be supported by a drawing in the 2D plane where $2 = \lambda_1 + \lambda_2$ and their product finds a maximum in $\lambda_1 = \lambda_2 = 1$, meaning that $r = 0$.

Hence, an option for overall correlation is to use *R* = Sqrt[1 - Det[**R**]]. Yet its usefulness must show in practice. PM 1. Det[**R**] is also used in tests on (multi-) collinearity (Farrar-Glauber). PM 2. Testing that all correlations are zero gives a test statistic - *n* Log[Det[**R**]]. PM 3. It can also be mentioned that there are other measures for overall correlation, notably in the area of entropy. This however seems to lead too far from the correlation paradigm.

■ Matrices with theoretical values

In theoretical exercises one may be tempted to fill in arbitrary parameters that however violate some key assumptions. The following is an example when we assume arbitrary values and set one correlation to -1.

```
rmat = {{1, a, b}, {a, 1, c}, {b, c, 1}}
```

$$\begin{pmatrix} 1 & a & b \\ a & 1 & c \\ b & c & 1 \end{pmatrix}$$

In the standard definition, the overall correlation takes the maximum of the multiple correlation coefficients.

```
oc = OverallCorrelation[%]
```

$$\sqrt{\text{Max}\left(1 - \frac{-a^2 + 2\,b\,c\,a - b^2 - c^2 + 1}{1 - a^2},\right.}$$
$$\left. 1 - \frac{-a^2 + 2\,b\,c\,a - b^2 - c^2 + 1}{1 - b^2}, 1 - \frac{-a^2 + 2\,b\,c\,a - b^2 - c^2 + 1}{1 - c^2}\right)$$

Arbitrarily setting *a* to -1 generates error messages.

```
% /. a → −1
```

Power::infy : Infinite expression $\frac{1}{0}$ *encountered. More…*

Max::nord : Invalid comparison with ComplexInfinity attempted. More…

$$\sqrt{\text{Max}\left(1 - \frac{-b^2 - 2\,c\,b - c^2}{1 - b^2}, 1 - \frac{-b^2 - 2\,c\,b - c^2}{1 - c^2}, \text{ComplexInfinity}\right)}$$

The point to note is that the correlations are mutually dependent. If we set *a* to -1 then this has consequences for *b* and *c*.

```
InequalitySolve[{-1 ≤ b ≤ 1, -1 ≤ c ≤ 1, Det[rmat /. a → -1] ≥ 0}, {b, c}]
```

$$-1 \le b \le 1 \wedge c = -b$$

```
oc /. c → -b // Simplify
```

$$\sqrt{\text{Max}\left(-\frac{2\,b^2}{a-1}, \frac{a^2+2\,b^2\,a+b^2}{1-b^2}\right)}$$

Only now it is proper to set a to its value, so that we don't get error messages.

```
% /. a → -1
```

$$\sqrt{\text{Max}(1, b^2)}$$

Working with theoretical matrices, the determinant approach is more robust, while the PSD condition is included in the condition that the outcome must between 0 and 1.

```
OverallCorrelation[rmat, Mode → Det]
```

$$\sqrt{a^2 - 2\,b\,c\,a + b^2 + c^2}$$

■ PSD for *k* = 3

The interdependence depends upon the PSD-ness of the matrix and less upon the data that are behind it. These data generate PSD-ness, but once it has been set, PSD-ness takes over control. The following gives a more general solution for $k = 3$. Consider the situation that $0 < r < 1$ is the correlation of Y and Z and that a new variable X is introduced. Instead of arbitray numbers *a, b* and *c* we might prefer to express that we are using correlations that satisfy PSD-ness.

```
rmat = {{1, r_{x,y}, r_{x,z}}, {r_{x,y}, 1, r}, {r_{x,z}, r, 1}}
```

$$\begin{pmatrix} 1 & r_{x,y} & r_{x,z} \\ r_{x,y} & 1 & r \\ r_{x,z} & r & 1 \end{pmatrix}$$

```
cond = {0 < r < 1, 0 ≤ r_{x,y} ≤ 1, 0 ≤ r_{x,z} ≤ 1, Det[rmat] ≥ 0}
```

$$\{0 < r < 1,\ 0 \le r_{x,y} \le 1,\ 0 \le r_{x,z} \le 1,\ -r^2 + 2\,r_{x,y}\,r_{x,z}\,r - r_{x,y}^2 - r_{x,z}^2 + 1 \ge 0\}$$

An overall solution can be found as follows (that apparently requires a trick of replacing variables). This generates a rather long expression so it is simpler to take a particular numerical example.

```
rul = {r_x,y → rxy, r_x,z → rxz};
Solution[] = InequalitySolve[cond /. rul, {rxy, rxz, r}] /. eluR[rul] /.
      Inequality[0, Less, r, Less, 1] → True;

rul = {r_x,y → rxy, r_x,z → rxz};
cond /. rul /. r → 3/10 /. eluR[rul];

res = InequalitySolve[%, {r_x,y, r_x,z}]
```

$$\left(0 \le r_{x,y} \le \frac{\sqrt{91}}{10} \bigwedge 0 \le r_{x,z} \le \frac{3\, r_{x,y}}{10} + \frac{1}{10}\sqrt{91}\sqrt{1-r_{x,y}^2}\right) \bigvee$$
$$\left(\frac{\sqrt{91}}{10} < r_{x,y} < 1 \bigwedge \frac{3\, r_{x,y}}{10} - \frac{1}{10}\sqrt{91}\sqrt{1-r_{x,y}^2} \le r_{x,z} \le \frac{3\, r_{x,y}}{10} + \frac{1}{10}\sqrt{91}\sqrt{1-r_{x,y}^2}\right) \bigvee$$
$$\left(r_{x,y} = 1 \bigwedge r_{x,z} = \frac{3}{10}\right)$$

Setting to values e.g. $r_{x,y} = 90\%$

```
res /. r_x,y → 9/10 // N
```

$$0. \le r_{x,z} \le 0.685812$$

■ Numerical example: Klein I model

It can be useful to support this discussion with a numerical example of correlation in real data. The Klein I model is a good example, see Theil (1971:432). Data and regressions can be found at UCLA ATS (2007).

```
lis = Partition[TextToMatrix["
1920 39.8 12.7 28.8 2.7 180.1 44.9 2.2 2.4 3.4
1921 41.9 12.4 25.5−0.2 182.8 45.6 2.7 3.9 7.7
1922 45.0 16.9 29.3 1.9 182.6 50.1 2.9 3.2 3.9
1923 49.2 18.4 34.1 5.2 184.5 57.2 2.9 2.8 4.7
1924 50.6 19.4 33.9 3.0 189.7 57.1 3.1 3.5 3.8
1925 52.6 20.1 35.4 5.1 192.7 61.0 3.2 3.3 5.5
1926 55.1 19.6 37.4 5.6 197.8 64.0 3.3 3.3 7.0
1927 56.2 19.8 37.9 4.2 203.4 64.4 3.6 4.0 6.7
1928 57.3 21.1 39.2 3.0 207.6 64.5 3.7 4.2 4.2
1929 57.8 21.7 41.3 5.1 210.6 67.0 4.0 4.1 4.0
1930 55.0 15.6 37.9 1.0 215.7 61.2 4.2 5.2 7.7
1931 50.9 11.4 34.5−3.4 216.7 53.4 4.8 5.9 7.5
1932 45.6 7.0 29.0−6.2 213.3 44.3 5.3 4.9 8.3
1933 46.5 11.2 28.5−5.1 207.1 45.1 5.6 3.7 5.4
1934 48.7 12.3 30.6−3.0 202.0 49.7 6.0 4.0 6.8
1935 51.3 14.0 33.2−1.3 199.0 54.4 6.1 4.4 7.2
1936 57.7 17.6 36.8 2.1 197.7 62.7 7.4 2.9 8.3
1937 58.7 17.3 41.0 2.0 199.8 65.0 6.7 4.3 6.7
1938 57.5 15.3 38.2−1.9 201.8 60.9 7.7 5.3 7.4
1939 61.6 19.0 41.6 1.3 199.9 69.5 7.8 6.6 8.9
1940 65.0 21.1 45.0 3.3 201.2 75.7 8.0 7.4 9.6
1941 69.7 23.5 53.3 4.9 204.5 88.4 8.5 13.8 11.6", Number], 10];

{year, cons, profit, wpriv, invest, klag, xprod, wgov, govt, taxes} =
    Transpose[lis] ;
```

Currently, we just take the example of the regression of consumption on profits, lagged profits and the total wage sum (in the reference: table 16.4). According to the references, the OLS estimate gives an R-Squared of 0.9810. We can verify this as follows:

```
dat = {cons, profit, Lag[profit], wpriv + wgov} // Transpose // Rest;

cm = CorrelationMatrix[dat]
```

$$\begin{pmatrix} 1. & 0.715338 & 0.65205 & 0.982703 \\ 0.715338 & 1. & 0.769128 & 0.634156 \\ 0.65205 & 0.769128 & 1. & 0.579332 \\ 0.982703 & 0.634156 & 0.579332 & 1. \end{pmatrix}$$

The routine MultipleRSquared[**R**, i] calculates RSquared[i] = 1 - $\mathcal{R}$ / $\mathcal{R}_{i,i}$. Taking $i = 1$ gives us the squared multiple correlation coefficient of the OLS regression of the first variable, consumption, on the other three variables including a constant. The value we get fits the one reported in the literature.

MultipleRSquared[cm, 1]

0.981008

If we want to summarize the correlation matrix into one number, we might consider 1 - Det[cm]. Note that the following still is a squared value.

1 – Det[cm]

0.995522

However, since we are more used to consider multiple correlations, we might better work with these. The routine OverallCorrelation[mat] has two options: Mode to select either Det or MultipleRSquared (default), and in the latter mode the option Function to select the function (default Max). Note that we now have a Sqrt[RSquared] outcome.

OverallCorrelation[cm]

0.990459

After this evaluation, the Results[...] allow us to identify the dependent variable in the multiple correlation that has been selected.

Results[OverallCorrelation]

{Mode → MultipleRSquared, Function → Max, List → {0.981008, 0.719033, 0.627152, 0.976314}, Take → 0.981008, Position → (1), Out → 0.990459}

Appendix D: Other relations for the 2 × 2 case

■ An analytical stepping stone

The 2 × 2 case is a crucial stepping stone. One can arrive at various measures using different philosophies, test the adequacy, and if that test is passed, see whether it can be generalized. Sometimes different philosophies give the same result for the simplest case, which provides a base for agreement and a source for confusion. The following lists 5 approaches that generate the same correlation measure for the 2 × 2 contingency table, using different approaches.

■ 1. Just recall

This is just to recall that the 2 × 2 case is generated both by the Volume Ratio measure and Cramer's V.

■ 2. Epidemiology - pooled test

However, in epidemiology, the following approach is possible for the risk ratio or risk difference. Let $p_0 = a / (a + c)$, $p_1 = b / (b + d)$ and p the pooled probability found in the sum column. Then a test statistic for $p_0 = p_1$ differs from the VolumeRatio measure only in $\sqrt{N}$ with $N = a + b + c + d$.

```
p0 − p1
──────────────────────────────── /.
√(p (1 − p) (1/(a + c) + 1/(b + d)))

  {p0 → a/(a + c), p1 → b/(b + d), p → (a + b)/(a + b + c + d)};

FullSimplify[%, Assumptions → {a ≥ 0, b ≥ 0, c ≥ 0, d ≥ 0}]
```

$$\frac{a\,d - b\,c}{\sqrt{\frac{(a+b)\,(a+c)\,(b+d)\,(c+d)}{a+b+c+d}}}$$

■ 3. Epidemiology - Matthew and Mantel-Haenszel

The literature contains a reference to a Matthew 1975 measure of correlation for disease test matrices, which appears to be the same formula as the Volume Ratio measure and Cramer's V.

The Mantel-Haenszel test (Kleinbaum et al. (2003:348)) also contains the determinant expression, a large sample approximation of Fisher's exact test.

■ 4. Assigning values {0, 1} or {-1, 1} or {i, j}

For a 2 × 2 table, we can assign values {0, 1} as in logic, see Colignatus (2007a), and then calculate the Pearson coefficient of correlation for real-valued data. In fact, let us assign arbitrary values {i, j}. It appears that all values drop out, as long as $i \neq j$.

```
FullSimplify[CorrelationPr2By2[Definition, {{a, b}, {c, d}}, {i, j}],
  Assumptions → {Thread[{a, b, c, d} ≥ 0]}]
```

$$\frac{(a\,d - b\,c)\,(i-j)^2}{\sqrt{(a+b)\,(a+c)\,(b+d)\,(c+d)\,(i-j)^4}}$$

(i) Originally, this author assigned the values {1, -1} rather than {1, 0} or True | False as in logic. A reason to avoid zero is that it might needlessly destroy information. A reason to use 1 versus -1 is that equal numbers of observations might be thought to balance each other, with an average outcome of zero. But above general rule shows that it does not matter.

(ii) The data points {1, 1}, {1, -1}, {-1, 1}, {-1, -1} generate x = {1, 1, -1, -1} and y = {1, -1, 1, -1} with frequencies {*a, c, b, d*}. Indeed, we might as well give a lists of {1, 1}, b lists of {1, -1} etcetera.

(iii) Thus we get the normal correlation between x = {1, 1, -1, -1} and y = {1, -1, 1, -1}with frequencies {*a, c, b, d*}.

(iv) Using formal parameters {*a, c, b, d*} and simplification shows that the determinant is used in the numerator and the row sums and column sums in the denominator. The values 1 and -1 assigned to the nominal data do not occur any more.

It may be mentioned that it was this relation that caused the author to investigate the issue into the direction that resulted into the general measure suggested above. This case was generalized first to $n \times n$, then $m \times n$, then $n_1 \times n_2 \times ... \times n_k$. The latter was done first for bordermatrices and then for the inner submatrices. Subsequently, PSD-ness and the comparison with the χ^2 measure were included.

Appendix E: An example in a higher dimension

The following shows an example of the Nominal Correlation measure in a higher dimension. The example shows that the algorithm is straightforward. It also indicates that, especially considering higher dimensions, a weighted average is a sensible choice.

```
mat2 = Table[i + 10 j + 100 k + 1000 m, {i, 2}, {j, 3}, {k, 4}, {m, 2}]
```

$$\begin{pmatrix} \begin{pmatrix} 1111 & 2111 \\ 1211 & 2211 \\ 1311 & 2311 \\ 1411 & 2411 \end{pmatrix} & \begin{pmatrix} 1121 & 2121 \\ 1221 & 2221 \\ 1321 & 2321 \\ 1421 & 2421 \end{pmatrix} & \begin{pmatrix} 1131 & 2131 \\ 1231 & 2231 \\ 1331 & 2331 \\ 1431 & 2431 \end{pmatrix} \\ \begin{pmatrix} 1112 & 2112 \\ 1212 & 2212 \\ 1312 & 2312 \\ 1412 & 2412 \end{pmatrix} & \begin{pmatrix} 1122 & 2122 \\ 1222 & 2222 \\ 1322 & 2322 \\ 1422 & 2422 \end{pmatrix} & \begin{pmatrix} 1132 & 2132 \\ 1232 & 2232 \\ 1332 & 2332 \\ 1432 & 2432 \end{pmatrix} \end{pmatrix}$$

This is the border matrix for dimension 3 and 4 (that should have sizes $m = 4$ and $n = 2$).

```
BorderMatrix[mat2, {3, 4}]
```

$$\begin{pmatrix} 6729 & 12729 \\ 7329 & 13329 \\ 7929 & 13929 \\ 8529 & 14529 \end{pmatrix}$$

These are the 4 by 2 submatrices used in the summation of the border matrix.

```
TabledBorderMatrix[mat2, {3, 4}]
```

$$\begin{pmatrix} \text{Mat}\left(\begin{pmatrix} 1111 & 2111 \\ 1211 & 2211 \\ 1311 & 2311 \\ 1411 & 2411 \end{pmatrix}\right) & \text{Mat}\left(\begin{pmatrix} 1121 & 2121 \\ 1221 & 2221 \\ 1321 & 2321 \\ 1421 & 2421 \end{pmatrix}\right) \\ \text{Mat}\left(\begin{pmatrix} 1131 & 2131 \\ 1231 & 2231 \\ 1331 & 2331 \\ 1431 & 2431 \end{pmatrix}\right) & \text{Mat}\left(\begin{pmatrix} 1112 & 2112 \\ 1212 & 2212 \\ 1312 & 2312 \\ 1412 & 2412 \end{pmatrix}\right) \\ \text{Mat}\left(\begin{pmatrix} 1122 & 2122 \\ 1222 & 2222 \\ 1322 & 2322 \\ 1422 & 2422 \end{pmatrix}\right) & \text{Mat}\left(\begin{pmatrix} 1132 & 2132 \\ 1232 & 2232 \\ 1332 & 2332 \\ 1432 & 2432 \end{pmatrix}\right) \end{pmatrix}$$

Check that it fits.

```
(% // Add) /. Mat → Identity
```

$$\begin{pmatrix} 6729 & 12729 \\ 7329 & 13329 \\ 7929 & 13929 \\ 8529 & 14529 \end{pmatrix}$$

These are all the submatrices.

```
VolumeRatioMatrix[mat2, BordersOrAll → Show]
```

$$1$$

$$\text{Add}\left[\begin{pmatrix} \text{VR}\left[\begin{pmatrix} 1111 & 1121 & 1131 \\ 1112 & 1122 & 1132 \end{pmatrix}\right] & \text{VR}\left[\begin{pmatrix} 2111 & 2121 & 2131 \\ 2112 & 2122 & 2132 \end{pmatrix}\right] & \text{VR}\left[\begin{pmatrix} 1211 & 1221 & 1231 \\ 1212 & 1222 & 1232 \end{pmatrix}\right] \\ \text{VR}\left[\begin{pmatrix} 1311 & 1321 & 1331 \\ 1312 & 1322 & 1332 \end{pmatrix}\right] & \text{VR}\left[\begin{pmatrix} 2311 & 2321 & 2331 \\ 2312 & 2322 & 2332 \end{pmatrix}\right] & \text{VR}\left[\begin{pmatrix} 1411 & 1421 & 1431 \\ 1412 & 1422 & 1432 \end{pmatrix}\right] \end{pmatrix}\right.$$

$$\text{Add}\left[\begin{pmatrix} \text{VR}\left[\begin{pmatrix} 1111 & 1211 & 1311 & 1411 \\ 1112 & 1212 & 1312 & 1412 \end{pmatrix}\right] & \text{VR}\left[\begin{pmatrix} 2111 & 2211 & 2311 & 2411 \\ 2112 & 2212 & 2312 & 2412 \end{pmatrix}\right] & \text{VR}\left[\left(\begin{matrix} 1121 & 12 \\ 1122 & 12 \end{matrix}\right.\right. \\ \text{VR}\left[\begin{pmatrix} 2121 & 2221 & 2321 & 2421 \\ 2122 & 2222 & 2322 & 2422 \end{pmatrix}\right] & \text{VR}\left[\begin{pmatrix} 1131 & 1231 & 1331 & 1431 \\ 1132 & 1232 & 1332 & 1432 \end{pmatrix}\right] & \text{VR}\left[\left(\begin{matrix} 2131 & 22 \\ 2132 & 22 \end{matrix}\right.\right. \end{matrix}\right.\right.$$

$$\text{Add}\left[\begin{pmatrix} \text{VR}\left[\begin{pmatrix} 1111 & 2111 \\ 1112 & 2112 \end{pmatrix}\right] & \text{VR}\left[\begin{pmatrix} 1211 & 2211 \\ 1212 & 2212 \end{pmatrix}\right] & \text{VR}\left[\begin{pmatrix} 1311 & 2311 \\ 1312 & 2312 \end{pmatrix}\right] \\ \text{VR}\left[\begin{pmatrix} 1411 & 2411 \\ 1412 & 2412 \end{pmatrix}\right] & \text{VR}\left[\begin{pmatrix} 1121 & 2121 \\ 1122 & 2122 \end{pmatrix}\right] & \text{VR}\left[\begin{pmatrix} 1221 & 2221 \\ 1222 & 2222 \end{pmatrix}\right] \\ \text{VR}\left[\begin{pmatrix} 1321 & 2321 \\ 1322 & 2322 \end{pmatrix}\right] & \text{VR}\left[\begin{pmatrix} 1421 & 2421 \\ 1422 & 2422 \end{pmatrix}\right] & \text{VR}\left[\begin{pmatrix} 1131 & 2131 \\ 1132 & 2132 \end{pmatrix}\right] \\ \text{VR}\left[\begin{pmatrix} 1231 & 2231 \\ 1232 & 2232 \end{pmatrix}\right] & \text{VR}\left[\begin{pmatrix} 1331 & 2331 \\ 1332 & 2332 \end{pmatrix}\right] & \text{VR}\left[\begin{pmatrix} 1431 & 2431 \\ 1432 & 2432 \end{pmatrix}\right] \end{pmatrix}\right]$$

NominalCorrelationMatrix[mat2] // N

$$\begin{pmatrix} 1. & -1.42186\times10^{-6} & -0.0000194491 & -0.0000834515 \\ -1.42186\times10^{-6} & 1. & 0. & -0.00136274 \\ -0.0000194491 & 0. & 1. & -0.0186297 \\ -0.0000834515 & -0.00136274 & -0.0186297 & 1. \end{pmatrix}$$

And the total measure of association can be applied again to this correlation matrix.

OverallCorrelation[%]

0.0186796

Or directly:

NominalCorrelation[mat2] // N

0.0186796

Appendix F: A note on the Frobenius theorems

These are just some notes, reflecting an angle for future prospection.

A contingency matrix is a nonnegative matrix, so that the Frobenius theorems apply. It is not quite clear how the eigenvectors come into play. There might be a relation here. The Frobenius eigenvalue $\lambda \geq 0$ is a real value that is at least as large as the absolute value of any other (possibly complex) eigenvalue. (Takayama (1974: 375)). Thus $\det(A) = \lambda_1 ... \lambda_n \leq \lambda^n$ and thus, if $\lambda > 0$ (which is definitely the case when A is indecomposable) then the implied Frobenius ratio measure is $\text{FrobRatio}(A) = \det(A) / \lambda^n \leq 1$. It is not clear however how this relates to the notion of the volume ratio.

With square matrix A, if $A\,x = \lambda\,x$ for vector $x \neq 0$ and scalar λ, then we say that λ is an eigenvalue and x its eigenvector. Let $r = A\ 1$ the row sums and $c = 1'\,A$ the column sums. Then $1'\,A\,x = \lambda\ 1'\,x$ or $\lambda = c'x\ /\ 1'x = c'x^*$, when $1'\,x \neq 0$, and $x^* = x\ /\ 1'x$ a normalized vector. Since A and A' have the same eigenvalues, there is also a $y'\,A = \lambda\,y'$ or $\lambda = y'r\ /\ y'\ 1$. When $x \neq 0$ (everywhere) and in particular $x > 0$ then there is also the possibility of a change in dimensions such that $D^{-1} = \text{diag}(x)$ and $B = D\ A\ D^{-1}$ such that $B\ 1 = \lambda\ 1$.

In general det(A) = $\lambda_1 ... \lambda_n$. Collecting all eigenvalues on the diagonal in Λ, zero everywhere else, and all eigenvectors in X (also using spanning vectors for higher multiplicity) then $A X = X \Lambda$.

Appendix G: On inference and causality

This discussion originated from considering the links between logic (Colignatus (2007a)) and causality (Pearl (2000)). Suppose that rain is a cause and wetness of streets is an effect. Is the observation that the streets are wet a good predictor of what was the cause ?

```
mat = {{"Observation count", "It rains", "It doesn't rain", "Total"},
       {"The streets are wet", 25, 3, □},
       {"The streets are not wet", 0, □, □},
       {"Total", □, □, 100}};
```

```
mat = Headed2DTableSolve[mat]
```

Observation count	It rains	It doesn't rain	Total
The streets are wet	25	3	28
The streets are not wet	0	72	72
Total	25	75	100

Instead of remembering all these 100 cases either individually or by frequency distribution, the memory processing unit might save on storage and retrieval costs by adopting a general rule (induction) that "If it rains then the streets are wet". This can become a general rule for which we can use a truthtable. The truthtable tests the condition whether the frequency is zero or non-zero, with entries True or False in the table.

In the course of considering these issues, the author noted that the common notion "correlation doesn't mean causation" has little use for such nominal data - since there is no standard measure of correlation. But now there is:

```
CorrelationPr2By2[Take[mat, -3, -3]] // N
```

0.92582

Given this measure of correlation, we must say (and we can do it now, with some relief that we can do so) that the normal caveats apply, i.e. that correlation is no causation, and that correlation itself doesn't say much about the actual model.

PM 1. The author may now continue considering the links between logic and causality. His hypothesis is that causality cannot be inferred from common statistics and has to do with the model and the order of calculation (time's arrow). In logic, there is a difference between implication ⇒ and inference ⊢. In causal models, there is a difference between equality == and assignment = (using *Mathematica*'s notation; other notations use = for equality and := for assignment). What causality is in Nature, inference is in the Mind. The overall umbrella would be the difference between statics and dynamics.

PM 2. A result of this continued work is Colignatus (2007f). A useful link there is that if s is the marginal probability of a success, c the marginal probability of the cause, R the risk of a success (e.g. death) and B background risk, then $s = R\,c + B\,(1 - c) = B + (R - B)\,c$, where B can be seen as the constant in a regression and $R - B$ gives the regression coefficient for the cause. But R would also be a conditional probability P[S, C] / P[C]. This might be a paradigmatic example that conditional probabilities could be seen as regression coefficients (or differences of them).

As said, the pairwise correlations of nominal variables are based upon the volume ratio measure, which is based upon using $\rho_{x,y}{}^2 = \beta_{y,x}\ \beta_{x,y} = n_{i,j}{}^2 / (n_{+j}\ \ n_{i+})$. Possibly, $\beta_{\mathrm{yi,xj}}$ might not be defined as the conditional probability P[y_i| x_j] but rather as P[y_i| x_j] - P[y_i| Not[x_j]], like above paradigmatic example. This kind of regression coefficient keeps the total N constant. However, as it looks now, the use of the determinant on the normalized matrix seems sufficient to handle that overall condition. See also **Appendix M**.

PM 3. **Appendix K** on the variances below discusses the variances of the variables. This differs from the variances of a category, say column x_j in a $m \times n$ contingency matrix. We have no need for the variances of the categories since we use the conditional probabilities to directly move to correlation. Only if one would use a different approach then the selection of such variance could enter the discussion.

Appendix H: Positive semidefiniteness for nominal correlation

■ On the sign

This appendix uses no signs for the associations and correlations, to eliminate a possible arbitrariness from the implementation of its determination. For now, we simulate only 2 × 2 × 2 matrices so that the sign would be unambiguous (based upon square matrices), but it is useful to mention the issue here for the case of larger sizes. Note though that disregarding the sign may also negatively affect the results.

```
SetOptions[PairwiseSign, ForceSign → 1];
```

■ In general

Appendix C discussed positive semidefiniteness (PSD) in the context of real data. For nominal data and the $n \times n$ contingency table we have allowed pairwise correlations $-1 \leq r \leq 1$, where the negative values arise from the order of presentation. The square matrix however is less general and hence we proceed with the more general case. For the more general $m \times n$ case, we take the measure of pairwise nominal correlation r as the square root of an expression of squares, neglect the sign, thus with $0 \leq r \leq 1$. For the most general case of k dimensions, the matrix $\mathbf{R}$ is created from these pairwise correlations, thus with $0 \leq \mathbf{R} \leq 1.1'$. This matrix is only an "association matrix" since we don't know yet that the overall correlation is in the proper range between 0 and 1. For proper overall correlation we require that $0 \leq \text{Det}[\mathbf{R}] \leq 1$. This can be translated into the equivalent condition of positive semidefiniteness (PSD), i.e. that for any $x \neq 0$ that $x'.\mathbf{R}.x \geq 0$. Outcomes outside of those ranges however cannot yet be excluded on a priori grounds. Thus let us see what we can say about that analytically.

Note the issue of terminology. We collected the outcomes of the pairwise "nominal correlations" into $\mathbf{R}$, call it an "assocation matrix" and sometimes call it a "correlation matrix". We can only do the latter for the cases when $0 \leq \text{Det}[\mathbf{R}] \leq 1$. The key questions are: (a) is this always the case, (b) is this mostly the case, (c) or are these situations infrequent, (d) and what to do when it is not true ? One example question to ask: with

VR[*x, y*] the volume ratio between *x* and *y*, how does VR[*x, y*] and VR[*x, z*] affect VR[*y, z*] ? Given the definitions of the volume ratios there is some reason to expect that this works *like* correlations in *many* cases. Yet, is there some proof that a matrix that is constructed in this manner is a PSD matrix ? The key questions thus might be translated as: (i) in what cases are symmetry, $0 \le r \le 1$ and diagonal 1 sufficient for PSD, and (ii) does the Volume Ratio measure fall in that category ?

Below we consider the issue formally, symbolically and by simulation. The first two approaches fail, given the capacities of this author, but the third one gives some clarity. PSD-ness tends to be assured when we take border matrices only and fails in 10% of the cases when we use the method of averaging results from inner matrices. The simulation also produced examples that can further be investigated in detail.

■ Formally

For $k = 2$ we have $\mathbf{R} = \{\{1, r\}, \{r, 1\}\}$ and we know that this is PSD when $0 \le r \le 1$.

To extend $k = 2$ to higher dimensions: let **R** hold for k variables, and let r be the correlation vector of some k+1st variable with the other k. The new **R*** is:

{{1, *r* '}, {*r*, R}}

$$\begin{pmatrix} 1 & r' \\ r & R \end{pmatrix}$$

We only know $0 \le r \le 1$, call this "correlation", but this will only be true iff $0 \le$ Det[**R***] ≤ 1. A determinant of matrix A can be decomposed in that of its submatrices as Det[A] = Det[A_{22}] Det[A_{11} - A_{12} $A_{22}^{-1} A_{21}$] if the inverse exists.

(a) When **R** is not regular then the sufficient and necessary condition for PSD-ness of **R*** is $0 \le$ Det[**R** - *r.r'*] ≤ 1. Here the problem is on the left hand side, since **R** - *r.r* need not be PSD.

(b) When **R** is regular and thus PD then the sufficient and necessary condition for PSD-ness of **R*** is $0 \le (r' \mathbf{R}^{-1} r) \le 1$.

Note that we can find this result also from the definition of multiple correlation: $(R^2)_i = (R^2)_{i.2,..,k} = 1 - \mathcal{R} / \mathcal{R}_{i,i}$. Applying this to the larger matrix, $\mathcal{R}$ = Det[**R***] and $\mathcal{R}_{1,1}$ = Det[**R**], thus we have $(0 \le R^2 \le 1) \Leftrightarrow (0 \le 1 -$ Det[**R***] / Det[**R**] $\le 1) \Leftrightarrow (0 \le 1 -$ Det[**R**] $(1 - r' \mathbf{R}^{-1} r)$ / Det[**R**] $\le 1) \Leftrightarrow (0 \le r' \mathbf{R}^{-1} r \le 1)$ again.

The left hand side holds since $\mathbf{R}^{-1}$ is also PD. It is not easy to satisfy the right hand side. The following gives a sufficient case but that may not be interesting. The row sums are $s = \mathbf{R}\ 1 = \mathbf{R}'\ 1$. Denote $\mu = \text{Max}[r]$ so that actually $0 \le r \le \mu\ 1$ and this becomes $0 \le \mathbf{R}\ \mathbf{R}^{-1}\ r \le \mu\ 1$. Premultiplication with 1' gives $0 \le s'\ \mathbf{R}^{-1}\ r \le \mu\ k$ or $0 \le \sigma / \mu\ \mathbf{R}^{-1}\ r \le 1$. Thus a sufficient condition for $(r'\ \mathbf{R}^{-1} r) \le 1$ is that $r \le \sigma / \mu = \mathbf{R}\ 1 / (k\ \text{Max}[r])$. Thus if r is about the the row averages then the result will be PSD. This case is not too interesting. (PM. Suppose that $R\ v = r$, or $v = \mathbf{R}^{-1}\ r$. We are tempted to premultiply $0 \le \mathbf{R}\ \mathbf{R}^{-1}\ r \le \mu\ 1$ with v but cannot do so since we don't know its signs.)

Presently, there seems little scope to extend on this.

PM. Additional notes

The following are some notes on additional derivations that show no additional results, and it might be useful for others to see this in order to avoid them.

When we use scalar x and vector y then PSD requires $x^2 + 2\ x\ r'.y + y'.R.y \ge 0$ with normalization $x^2 + y'.y = 1$. Using alternative normalized vector $\{1, y\}$, the condition for $\mathbf{R}^*$ is that for any $y \ne 0$ that $1 + 2\ r'.y + y'.\mathbf{R}.y \ge 0$. This does not seem to lead far.

For a general proof we might use the decomposition of a PD matrix $\mathbf{R} = P'P$. For $\mathbf{R}^*$ we get:

$$\mathbf{R}^* = \begin{pmatrix} 1 & r' \\ r & \mathbf{R} \end{pmatrix} = Q'.Q = \begin{pmatrix} a & b' \\ c & S' \end{pmatrix}\begin{pmatrix} a & c' \\ b & S \end{pmatrix} = \begin{pmatrix} a^2 + b'b & a\,c' + b'S \\ a\,c + S'b & S'S + c.c' \end{pmatrix}$$

This is a bit forbidding. Let us first try the easy decomposition:

$$\mathbf{R}^* = \begin{pmatrix} 1 & r' \\ r & \mathbf{R} \end{pmatrix} = Q'.Q = \begin{pmatrix} 1 & 0 \\ r & S' \end{pmatrix}\begin{pmatrix} 1 & r' \\ 0 & S \end{pmatrix} = \begin{pmatrix} 1 & r' \\ r & S'S + r.r' \end{pmatrix}$$

So that $\mathbf{R} = P'P = S'.S + r.r'$. Note that by form $S'.S$ and $r.r'$ are PSD. However we must keep track of the proper calculation order. We define $M = R - r.r'$, then determine when this M is PSD, only then decompose $M = S'.S$, and then construct Q. For $M = R - r.r'$ to be PSD, the necessary and sufficient condition is that for all $y \ne 0$ that $y'.\mathbf{R}.y \ge (r'.y)^2$. An approach for $\mathbf{R}$ PD is as follows. For M we find $M = P'P - r.r' = P'\ (\text{I} - W)\ P$, where $W = P'^{-1} r.r'\ P^{-1}$. First, W is symmetric. Secondly, $W'\ W = W\ W = P'^{-1} r.r'\ P^{-1}\ P'^{-1} r.r'\ P^{-1} = P'^{-1} r.r'\ (P'\ P)^{-1} r.r'\ P^{-1} = P'^{-1} r.r'\ \mathbf{R}^{-1} r.r'\ P^{-1} = P'^{-1} r.(r'\ \mathbf{R}^{-1} r)\ .r'\ P^{-1}$ where the expression in brackets $\lambda = (r'\ \mathbf{R}^{-1} r) \ge 0$ is a scalar, and nonnegative since the inverse of the correlation matrix is PSD too. Thus $W\ W = \lambda\ W$. Hence we take $V = f[\lambda] W$ and $S = (\text{I}$

- V) P, and find $M = S'S = P'$ (I - V) (I - V) $P = P'$ (I - V - V + V V) $P = P'$ (I - 2 $f[\lambda]$ W + $f[\lambda]^2 W$ W) P, that reduces to $M = P'$ (I - W) P if and only if - 2 $f[\lambda]$ W + $f[\lambda]^2 \lambda$ W = -W. Thus:

```
Solve[-2 f[λ] + λ f[λ]^2 == -1, f[λ]]
```

$$\left\{\left\{f(\lambda) \to \frac{1-\sqrt{1-\lambda}}{\lambda}\right\}, \left\{f(\lambda) \to \frac{\sqrt{1-\lambda}+1}{\lambda}\right\}\right\}$$

Which concludes this easy decomposition. The point to note is that the solution requires $\lambda = (r' \mathbf{R}^{-1} r) \leq 1$ in order for the easy decomposition to be real-valued. This is the same as what we already knew from the determinant.

PM. Weaker properties on **R** follow from this. Let ρ = Max[**R** - I], thus the maximal off-diagonal element. Then **R** $\leq$ (1 - ρ) I + ρ 1.1' and for vector $0 \leq r \leq 1$ we find r' **R** $r \leq$ (1 - ρ) $r' r + \rho$ $(1'r)^2 \leq (1'r)^2$ since $\rho = 1$ is the highest value. For the row sums $0 \leq s \leq$ ((1 - ρ) + ρ k) 1 = (1 + (k - 1) ρ) 1.

■ A symbolic approach for *k* = 3

For k = 3 we may take a symbolic nonnegative matrix, determine the formal pairwise nominal correlations, construct the matrix, take the determinant and determine its possible range.

```
TableForm[mat = {{{a, b}, {c, d}}, {{e, f}, {g, h}}},
  TableHeadings → {{"Shop1", "Shop2"}, {"Green", "Blue"}, {"Fit", "No fit"}}]
```

		Green	Blue
Shop1	Fit	*a*	*c*
	No fit	*b*	*d*
Shop2	Fit	*e*	*g*
	No fit	*f*	*h*

```
case = BorderMatrices;
```

```
ncmat[case] =
  NominalCorrelationMatrix[mat, BordersOrAll → BorderMatrices] // Simplify
```

NominalCorrelationMatrix::num : No PSD test since non−numeric result

$$\begin{pmatrix} 1 & \sqrt{\frac{(c(e+f)+d(e+f)-(a+b)(g+h))^2}{(a+b+c+d)(a+b+e+f)(c+d+g+h)(e+f+g+h)}} & \sqrt{\frac{(b(e+g)+d(e} \\ \sqrt{\frac{(c(e+f)+d(e+f)-(a+b)(g+h))^2}{(a+b+c+d)(a+b+e+f)(c+d+g+h)(e+f+g+h)}} & 1 & \sqrt{\frac{(d e+h e-c f-f} \\ \sqrt{\frac{(b(e+g)+d(e+g)-(a+c)(f+h))^2}{(a+b+c+d)(a+c+e+g)(b+d+f+h)(e+f+g+h)}} & \sqrt{\frac{(d e+h e-c f-f g-b(c+g)+a(d+h))^2}{(a+b+e+f)(a+c+e+g)(b+d+f+h)(c+d+g+h)}} & 1 \end{pmatrix}$$

One can imagine that this does not give an insightful result. The evaluation had to be aborted and the following cell thus is locked.

```
FullSimplify[dt = Det[%], Assumptions → Thread[{a, b, c, d, e, f, g, h} ≥ 0]]
```

$Aborted

We may however recall the $k = 3$ discussion in **Appendix C**, where Solution[] contained the solution of the condition on the $r_{i,j}$ (using $0 < r < 1$). This evaluates easily, but gives another forbidding expression and thus is not shown. The subsequent simplification again is aborted and now locked.

```
sol = Solution[] /. {r_{x,y} → ncmat[case][[1, 2]],
      r_{x,z} → ncmat[case][[1, 3]], r → ncmat[case][[2, 3]]}

FullSimplify[%, Assumptions → Thread[{a, b, c, d, e, f, g, h} ≥ 0]]
```

$Aborted

■ Simulation for *k* = 3

Introduction

For $k = 3$ we may assign arbitrary numbers and determine whether $0 \le \text{Det}[\mathbf{R}] \le 1$. There are four methods to consider, using only the border matrices or all inner submatrices, and for the latter using the standard weighted average, the minimum or the maximum of the inner correlations.

```
TableForm[mat = {{{a, b}, {c, d}}, {{e, f}, {g, h}}},
  TableHeadings → {{"Shop1", "Shop2"}, {"Green", "Blue"}, {"Fit", "No fit"}}]
```

		Green	Blue
Shop1	Fit	*a*	*c*
	No fit	*b*	*d*
Shop2	Fit	*e*	*g*
	No fit	*f*	*h*

We may consider cell values from 1 to a million. There will be some duplication in the calculations due to symmetry. Using 4 values for 8 variables we investigate $4^8 = 65536$ matrices.

```
possibilities = {1, 100, 10000, 10^6};

tab = Table[possibilities, {i, 1, 8}];

combi = Outer[List, Sequence @@ tab];
```

For these values, we will find that the method of using bordermatrices only still gives PSD correlation matrices while the method of using inner correlation matrices for 10% of the cases gives a determinant out off the range.

We don't need to show the details of these simulations and can directly proceed to the review of the results.

For *k* = 3, method → All, Inner → Automatic

For *k* = 3, method → All, Inner → Min

For *k* = 3, method → All, Inner → Max

For *k* = 3, method → BorderMatrices

Review

We checked $0 \leq \text{Det}[\mathbf{R}] \leq 1$ where the $\mathbf{R}$ was based upon nominal correlations using either the border matrices or all submatrices. The following summarizes the results. This is a contingency table on issues pertaining to contingency tables.

```
heading = TableHeadings → {{True, False, Sum}, {BorderMatrices,
All, AllMin, AllMax}};

InsideTable[Set, ncposdef, heading]
```

```
InsideTable[Show, Length]
```

	BorderMatrices	All	AllMin	AllMax
True	65536	58576	65536	40384
False	0	6960	0	25152
Sum	65536	65536	65536	65536

It appears that the measure is more sensitive to the use of the inner submatrices (and the weighted sums used), while it is not affected when correlations are based upon border matrices only or the strict Min condition. It is not clear whether the latter is only a result of this particular numerical example, where elements in the contingency table range from 1 to a million. For normal ranges it might be less of a problem. Another overall impression is that PSD-ness exists when correlations have lower values and comes into problem when there are some high correlations (that might conflict).

■ Be sure to reset

This should be reset now to its original default value.

```
ResetOptions[PairwiseSign]
```

{ForceSign → {Automatic, False}}

```
VolumeRatio[{{a, b}, {c, d}}] // Simplify
```

$$\sqrt{\frac{(b\,c-a\,d)^2}{(a+b)\,(a+c)\,(b+d)\,(c+d)}}\ \text{sgn}(a\,d-b\,c)$$

Appendix I: Overview of variants

■ Introduction

We have these variants: *mode* {MultipleRSquared, Det}, *method* {*B* = BorderMatrices, *A* = All inner matrices}, for the latter the *manner* {Automatic is the weighted average, Min, Max}, and subsequently whether we test for the bounds and if necessary adjust by maximizing λ such that $0 \le \lambda \le 1$ and $0 \le \text{Det}[\mathbf{R} = (1 - \lambda)\ \mathbf{R}_B + \lambda\ \mathbf{R}_A\ ] \le 1$, and possibly even a PSD-approximation on top. **Appendix H** gave a simulation of outcomes without any such PSD-adjustment. It will be instructive to see concrete examples of the cases with and without adjustment.

A review will present the various matrices of association and correlation. Since these can be equal, i.e. when there is no adjustment with some λ, we will include a test in the review on equality and only show both when there is difference. The size of the output already indicates whether there is a difference or not. In this review, we might consider print all the λ but do this standardly only for the suggested default algorithm and by exception only for the case when there is value different from 0.

PM. We run these cases with the default setting that the sign of the correlations matter. For the submatrices considered here this would be unambiguous since these are 2×2 (square). The signs obviously are important for the Min and Max manners, while they also mitigate the outcome of averaging.

■ 1. The example from the introduction

This contingency table is innerly fully symmetric and (thus) all variants generate about the same outcome, with a slight difference in the mode MultipleRSquared or Det. The results fall into the bounds.

```
TableForm[mat = {{{5, 1}, {8, 2}}, {{2, 8}, {1, 5}}},
  TableHeadings → {{"Shop1", "Shop2"}, {"Green", "Blue"}, {"Fit", "No fit"}}]
```

		Green	Blue
Shop1	Fit	5	8
	No fit	1	2
Shop2	Fit	2	1
	No fit	8	5

```
BorderMatrices[mat]
```

$$\left\{\{1, 2\} \to \begin{pmatrix} 6 & 10 \\ 10 & 6 \end{pmatrix}, \{1, 3\} \to \begin{pmatrix} 13 & 3 \\ 3 & 13 \end{pmatrix}, \{2, 3\} \to \begin{pmatrix} 7 & 9 \\ 9 & 7 \end{pmatrix}\right\}$$

```
NominalCorrelationReview[mat] // N
```

$$\left\{\text{Matrix(NominalCorrelation, Equal, VolumeRatio)} \to \{\text{True, True, True, True}\},\right.$$

$$\text{Matrix(NominalCorrelation)} \to \left\{\text{BorderMatrices} \to \begin{pmatrix} 1. & -0.25 & 0.625 \\ -0.25 & 1. & -0.125 \\ 0.625 & -0.125 & 1. \end{pmatrix}, \text{Automatic} \to \right.$$

$$\begin{pmatrix} 1. & -0.221917 & 0.61807 \\ -0.221917 & 1. & 0.0413449 \\ 0.61807 & 0.0413449 & 1. \end{pmatrix}, \text{Min} \to \begin{pmatrix} 1. & -0.221917 & 0.61807 \\ -0.221917 & 1. & 0.0413449 \\ 0.61807 & 0.0413449 & 1. \end{pmatrix},$$

$$\text{Max} \to \left.\begin{pmatrix} 1. & -0.221917 & 0.61807 \\ -0.221917 & 1. & 0.0413449 \\ 0.61807 & 0.0413449 & 1. \end{pmatrix}\right\}, \text{Table} \to \begin{matrix} \text{VolumeRatio(MultipleRSquared)} \\ \text{VolumeRatio(Det)} \\ \text{NominalCorrelation(MultipleRSq} \\ \text{NominalCorrelation(Det)} \end{matrix}$$

```
Factor /. Results[NominalCorrelationMatrix, {All, Automatic}, True]
```

```
0
```

■ 2. A case of conditional independence

In conditional independence, the problem is essentially a 2 × 2 case so that all outcomes are the same.

```
TableForm[mat = {{{a, b}, {b, a}}, {{a, b}, {b, a}}} /. {a → 5, b → 8},
  TableHeadings → {{"Shop1", "Shop2"}, {"Green", "Blue"}, {"Fit", "No fit"}}]
```

		Green	Blue
Shop1	Fit	5	8
	No fit	8	5
Shop2	Fit	5	8
	No fit	8	5

BorderMatrices[mat]

$$\left\{\{1, 2\} \to \begin{pmatrix} 13 & 13 \\ 13 & 13 \end{pmatrix}, \{1, 3\} \to \begin{pmatrix} 13 & 13 \\ 13 & 13 \end{pmatrix}, \{2, 3\} \to \begin{pmatrix} 10 & 16 \\ 16 & 10 \end{pmatrix}\right\}$$

NominalCorrelationReview[mat] // N

$$\left\{\text{Matrix(NominalCorrelation, Equal, VolumeRatio)} \to \{\text{True, True, True, True}\},\right.$$

$$\text{Matrix(NominalCorrelation)} \to \left\{\text{BorderMatrices} \to \begin{pmatrix} 1. & 0. & 0. \\ 0. & 1. & -0.230769 \\ 0. & -0.230769 & 1. \end{pmatrix},\right.$$

$$\text{Automatic} \to \begin{pmatrix} 1. & 0. & 0. \\ 0. & 1. & -0.230769 \\ 0. & -0.230769 & 1. \end{pmatrix}, \text{Min} \to \begin{pmatrix} 1. & 0. & 0. \\ 0. & 1. & -0.230769 \\ 0. & -0.230769 & 1. \end{pmatrix},$$

$$\left.\text{Max} \to \begin{pmatrix} 1. & 0. & 0. \\ 0. & 1. & -0.230769 \\ 0. & -0.230769 & 1. \end{pmatrix}\right\}, \text{Table} \to \begin{matrix} \text{VolumeRatio(MultipleRSquared)} \\ \text{VolumeRatio(Det)} \\ \text{NominalCorrelation(MultipleRSquared)} \\ \text{NominalCorrelation(Det)} \end{matrix}$$

Factor /. Results[NominalCorrelationMatrix, {All, Automatic}, True]

0

■ 3. An example where Det[R] is not in the required range

This example has been taken from the simulation. The suggested standard algorithm comes into problems but this may have to do with the wide range of input values. PSD-adjustment still saves the day.

TableForm[mat = {{{1, 1}, {1, 1}}, {{1, 10000}, {1000000, 100}}},
TableHeadings → {{"Shop1", "Shop2"}, {"Green", "Blue"}, {"Fit", "No fit"}}]

		Green	Blue
Shop1	Fit	1	1
	No fit	1	1
Shop2	Fit	1	1000000
	No fit	10000	100

BorderMatrices[mat]

$$\left\{\{1, 2\} \to \begin{pmatrix} 2 & 2 \\ 10001 & 1000100 \end{pmatrix}, \{1, 3\} \to \begin{pmatrix} 2 & 2 \\ 1000001 & 10100 \end{pmatrix}, \{2, 3\} \to \begin{pmatrix} 2 & 10001 \\ 1000001 & 101 \end{pmatrix}\right\}$$

The default mode uses MultipleRSquared, such that the Automatic application of weighted averages of inner submatrices generates a correlation that is 4. The cause is that the "correlation matrix" is not really a correlation matrix since the determinant is not between 0 and 1. We have to adjust to get a result within the bounds. Also the Min case gives a problem.

NominalCorrelationReview[mat] // N

CorrelationMatrixQ::psd : Not PSD, Det = −0.174775 outside [*0, 1*]

CorrelationMatrixQ::psd : Not PSD, Det = −0.00935449 outside [*0, 1*]

CorrelationMatrixQ::psd : Not PSD, Det = −0.174775 outside [*0, 1*]

General::stop :
Further output of CorrelationMatrixQ::psd will be suppressed during this calculation. More…

$$\{\text{Matrix(NominalCorrelation, Equal, VolumeRatio)} \to \{\text{True, False, False, True}\},$$

$$\text{Matrix(VolumeRatio)} \to \left\{\text{BorderMatrices} \to \begin{pmatrix} 1. & 0.00984938 & -0.00979953 \\ 0.00984938 & 1. & -0.994838 \\ -0.00979953 & -0.994838 & 1. \end{pmatrix},\right.$$

$$\text{Automatic} \to \begin{pmatrix} 1. & 0.494305 & -0.064702 \\ 0.494305 & 1. & -0.994933 \\ -0.064702 & -0.994933 & 1. \end{pmatrix},$$

$$\text{Min} \to \begin{pmatrix} 1. & -0.069307 & -0.0703492 \\ -0.069307 & 1. & -0.994937 \\ -0.0703492 & -0.994937 & 1. \end{pmatrix},$$

$$\left.\text{Max} \to \begin{pmatrix} 1. & 0.499999 & 0.4999 \\ 0.499999 & 1. & 0. \\ 0.4999 & 0. & 1. \end{pmatrix}\right\}, \text{Matrix(NominalCorrelation)} \to$$

$$\left\{\text{BorderMatrices} \to \begin{pmatrix} 1. & 0.00984938 & -0.00979953 \\ 0.00984938 & 1. & -0.994838 \\ -0.00979953 & -0.994838 & 1. \end{pmatrix},\right.$$

$$\text{Automatic} \to \begin{pmatrix} 1. & 0.123777 & -0.0227107 \\ 0.123777 & 1. & -0.99486 \\ -0.0227107 & -0.99486 & 1. \end{pmatrix},$$

$$\text{Min} \to \begin{pmatrix} 1. & -0.0473198 & -0.0535304 \\ -0.0473198 & 1. & -0.99491 \\ -0.0535304 & -0.99491 & 1. \end{pmatrix},$$

$$\left.\text{Max} \to \begin{pmatrix} 1. & 0.499999 & 0.4999 \\ 0.499999 & 1. & 0. \\ 0.4999 & 0. & 1. \end{pmatrix}\right\}, \text{Table} \to \begin{matrix} \text{VolumeRatio(MultipleRSquared)} \\ \text{VolumeRatio(Det)} \\ \text{NominalCorrelation(MultipleRSquared)} \\ \text{NominalCorrelation(Det)} \end{matrix}$$

Factor /. Results[NominalCorrelationMatrix, {All, Automatic}, True]

0.235165

■ 4. Another example where Det[R] is not in the required range

This example has been taken from the simulation too. Here the values range only between 1 and 100.

```
TableForm[mat = {{{1, 1}, {1, 100}}, {{100, 100}, {100, 1}}},
  TableHeadings → {{"Shop1", "Shop2"}, {"Green", "Blue"}, {"Fit", "No fit"}}]
```

		Green	Blue
Shop1	Fit	1	1
	No fit	1	100
Shop2	Fit	100	100
	No fit	100	1

```
BorderMatrices[mat]
```

$$\left\{\{1, 2\} \to \begin{pmatrix} 2 & 101 \\ 200 & 101 \end{pmatrix}, \{1, 3\} \to \begin{pmatrix} 2 & 101 \\ 200 & 101 \end{pmatrix}, \{2, 3\} \to \begin{pmatrix} 101 & 101 \\ 101 & 101 \end{pmatrix}\right\}$$

The default mode uses MultipleRSquared, such that the Min application on inner submatrices generates a correlation of 7.39. The cause is that the matrix of pairwise correlation coefficients is not a PSD correlation matrix since the determinant is not between 0 and 1. Using the function Min increases the absolute size of the pairwise correlations, making it more difficult for them to be consistent.

NominalCorrelationReview[mat] // N

CorrelationMatrixQ::psd : Not PSD, Det = −2.10354 outside [*0, 1*]

CorrelationMatrixQ::psd : Not PSD, Det = −2.10354 outside [*0, 1*]

$$\{\text{Matrix(NominalCorrelation, Equal, VolumeRatio)} \to \{\text{True, True, False, True}\},$$

$$\text{Matrix(VolumeRatio)} \to \left\{\text{BorderMatrices} \to \begin{pmatrix} 1. & -0.562255 & -0.562255 \\ -0.562255 & 1. & 0. \\ -0.562255 & 0. & 1. \end{pmatrix},\right.$$

$$\text{Automatic} \to \begin{pmatrix} 1. & -0.490099 & -0.490099 \\ -0.490099 & 1. & -0.240197 \\ -0.490099 & -0.240197 & 1. \end{pmatrix},$$

$$\left.\text{Min} \to \begin{pmatrix} 1. & -0.980198 & -0.980198 \\ -0.980198 & 1. & -0.490099 \\ -0.980198 & -0.490099 & 1. \end{pmatrix}, \text{Max} \to \begin{pmatrix} 1. & 0. & 0. \\ 0. & 1. & 0.490099 \\ 0. & 0.490099 & 1. \end{pmatrix}\right\},$$

$$\text{Matrix(NominalCorrelation)} \to \left\{\text{BorderMatrices} \to \begin{pmatrix} 1. & -0.562255 & -0.562255 \\ -0.562255 & 1. & 0. \\ -0.562255 & 0. & 1. \end{pmatrix},\right.$$

$$\text{Automatic} \to \begin{pmatrix} 1. & -0.490099 & -0.490099 \\ -0.490099 & 1. & -0.240197 \\ -0.490099 & -0.240197 & 1. \end{pmatrix}, \text{Min} \to$$

$$\left.\begin{pmatrix} 1. & -0.66371 & -0.66371 \\ -0.66371 & 1. & -0.11897 \\ -0.66371 & -0.11897 & 1. \end{pmatrix}, \text{Max} \to \begin{pmatrix} 1. & 0. & 0. \\ 0. & 1. & 0.490099 \\ 0. & 0.490099 & 1. \end{pmatrix}\right\}, \text{Table} \to \begin{matrix} \text{Volu} \\ \text{Volu} \\ \text{Nomi} \\ \text{Nomi} \end{matrix}$$

Factor /. Results[NominalCorrelationMatrix, {All, Automatic}, True]

0

Factor /. Results[NominalCorrelationMatrix, {All, Min}, True]

0.242747

■ Conclusion

This appendix considered contingency tables of higher dimensions (more than $m \times n$). The suggested algorithm seems to work fairly well: (i) use the weighted average of the volume ratio measures of the inner matrices (the method called All with the manner Automatic), (ii) if this is not PSD then take the maximal λ such that $0 \le \lambda \le 1$ and $0 \le \text{Det}[\mathbf{R} = (1 - \lambda)\, \mathbf{R}_B + \lambda\, \mathbf{R}_A] \le 1$, (iii) if this still is not PSD (e.g. when $\mathbf{R}_B$ itself is not), take the PSD approximation (by default making the eigenvalues nonnegative). An adjustment was only required in case 3 considered above, where the range in values might be seen as extreme.

Appendix J: Finding the closest PSD correlation matrix

■ Introduction

Consider the non-PSD association matrix (symmetric, diagonal 1, elements between 0 and 1) but not $0 \leq \text{Det}[R] \leq 1$, and consider the question to find a closest PSD matrix. The solution of course depends upon the notion of "closeness". Higham (1989) uses some measures of which we might use the Frobenius norm here, i.e. the Euclidean norm on all elements. The result that he presents however concerns a general PSD matrix and not one with diagonal 1. Another approach is to set all eigenvalues to a nonnegative value. Even when we make sure that the sum of these new nonnegative eigenvalues is equal to the dimension of matrix, we find that in practice a matrix may result that has no 1 on the diagonal, just as a result of numerical rounding off. In both cases, a practical solution might be to divide rows and columns by the square root of the appropriate diagonal element. The latter reflects the way that a correlation matrix is created from a variance-covariance matrix. Next to these more theoretical approaches (though with the practical diagonal adjustment) there are also the straightforward numerical approaches by adjusting the off-diagonal elements in some uniform manner. These approaches have little theory but do have the advantage of working directly with the elements and not with some hidden matrices or eigenvalues. In that manner one can aspire at some adjustment that maintains some "grid". A very attractive approach is, for off-diagonal element *r,* to take the adjusted value $r^x x$ for $0 < x < 1$, so that one includes both a proportional effect and a power effect.

These issues can be clarified with the following examples (taken from the simulations).

```
mat = {{{1, 1}, {1, 1}}, {{1, 10000}, {1000000, 100}}};

associationMat = VolumeRatioMatrix[mat] // N
```

$$\begin{pmatrix} 1. & 0.494305 & -0.064702 \\ 0.494305 & 1. & -0.994933 \\ -0.064702 & -0.994933 & 1. \end{pmatrix}$$

CorrelationMatrixQ[%]

CorrelationMatrixQ::psd : Not PSD, Det = −0.174775 outside [*0, 1*]

False

■ Higham (1989)

Higham's solution for the Frobenius norm generates a positive semi definite matrix yet the diagonal may not be unity.

hf = ToPSDMatrix["HighamF", associationMat]

$$\begin{pmatrix} 1.00711 & 0.476604 & -0.0804922 \\ 0.476604 & 1.04404 & -0.955648 \\ -0.0804922 & -0.955648 & 1.03504 \end{pmatrix}$$

Definiteness[hf]

{Positive, Semi, Definiteness}

CorrelationMatrixQ[hf]

CorrelationMatrixQ::diag : Matrix doesn't have a diagonal of 1

False

However, when we divide by the diagonal elements then we find a proper PSD correlation matrix, and the differences from the original input are not too large.

newnc =
ToPSDCorrelationMatrix[associationMat, Do → {"HighamF", UnitDiagonalMatrix}]

$$\begin{pmatrix} 1. & 0.464793 & -0.0788379 \\ 0.464793 & 1. & -0.919307 \\ -0.0788379 & -0.919307 & 1. \end{pmatrix}$$

difhf = newnc − associationMat

$$\begin{pmatrix} 2.22045 \times 10^{-16} & -0.0295122 & -0.0141359 \\ -0.0295122 & 0. & 0.0756266 \\ -0.0141359 & 0.0756266 & -2.22045 \times 10^{-16} \end{pmatrix}$$

FrobeniusDistance → Sqrt[Add[%^2]]

FrobeniusDistance → 0.116535

■ Nonnegative eigenvalues

Setting the eigenvalues to nonnegative values doesn't guarantee a proper diagonal. But the test might already break down due to symmetry problems due to numerical precision. We neglect that issue just for now.

```
nonnegev = ToPSDMatrix["NonnegEV", associationMat]
```

$$\begin{pmatrix} 0.978986 & 0.463292 & -0.0782441 \\ 0.463292 & 1.01488 & -0.928957 \\ -0.0782441 & -0.928957 & 1.00613 \end{pmatrix}$$

```
Definiteness[nonnegev ]
```

{Positive, Semi, Definiteness}

```
CorrelationMatrixQ[nonnegev ]
```

CorrelationMatrixQ::diag : Matrix doesn't have a diagonal of 1

False

When we re-diagonalize then we find a proper PSD correlation matrix, and the differences from the original input are not too large. (Neglecting possibly symmetry issues due to numerical precision.)

```
newnc = ToPSDCorrelationMatrix[
    associationMat, Do → {"NonnegEV", UnitDiagonalMatrix}]
```

$$\begin{pmatrix} 1. & 0.464793 & -0.0788379 \\ 0.464793 & 1. & -0.919307 \\ -0.0788379 & -0.919307 & 1. \end{pmatrix}$$

The problems of the diagonal and symmetry are solved to an error of 10^{-15}. The differences between the original non-PSD correlations and the approximated PSD correlations are not too large.

```
newnc – associationMat
```

$$\begin{pmatrix} 0. & -0.0295122 & -0.0141359 \\ -0.0295122 & 2.22045 \times 10^{-16} & 0.0756266 \\ -0.0141359 & 0.0756266 & -1.11022 \times 10^{-16} \end{pmatrix}$$

In fact, the differences are the same as the earlier solution with Higham's Frobenius norm.

difhf − % // Chop

$$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

■ A numerical approach

A plain non-theoretical but numerical approach is to gradually adjust the off-diagonal correlations till the determinant is in the appropriate range. The following is a general expression, using parameters α and β, while minimizing Det[**R**[*x; α, β*]] over *x*. Only parameter values 0 or 1 have been programmed. An attractive method with $\alpha = \beta = 1$ can be called "gradual".

res = ToPSDCorrelationMatrix[{α, β, x}, associationMat]

$$\begin{pmatrix} 1 & 0.494305^{x\alpha-\alpha+1}(x\beta-\beta+1) & -0.064702^{x\alpha-\alpha+1}(x\beta-\beta+1) \\ 0.494305^{x\alpha-\alpha+1}(x\beta-\beta+1) & 1 & -0.994933^{x\alpha-\alpha+1}(x\beta-\beta+1) \\ -0.064702^{x\alpha-\alpha+1}(x\beta-\beta+1) & -0.994933^{x\alpha-\alpha+1}(x\beta-\beta+1) & 1 \end{pmatrix}$$

% /. {α → 1, β → 1}

$$\begin{pmatrix} 1 & 0.494305^{x}\,x & -0.064702^{x}\,x \\ 0.494305^{x}\,x & 1 & -0.994933^{x}\,x \\ -0.064702^{x}\,x & -0.994933^{x}\,x & 1 \end{pmatrix}$$

Since the original determinant is negative we solve for the lower value close to zero.

SolveDetIn0To1[−1, %, x, 1]

$$\begin{pmatrix} 1 & 0.480176 & -0.0747351 \\ 0.480176 & 1 & -0.9106 \\ -0.0747351 & -0.9106 & 1 \end{pmatrix}$$

Results[SolveDetIn0To1]

$$\{x \to 0.914841\}$$

The following routine combines these steps.

newnc = ToPSDCorrelationMatrix[associationMat, Do → "Gradual"]

$$\begin{pmatrix} 1 & 0.480176 & -0.0747351 \\ 0.480176 & 1 & -0.9106 \\ -0.0747351 & -0.9106 & 1 \end{pmatrix}$$

CorrelationMatrixQ[newnc]

True

In this general routine the output value of *x* is available as a text variable (since we didn' t provide a symbolic variable).

"x" /. Results[ToPSDCorrelationMatrix, {1, 1}]

0.914841

The differences between the original non-PSD correlations and the approximated PSD correlations are small. The Frobenius distance is a bit larger than the theoretical distance yet the distribution over the cells is different.

newnc – associationMat

$$\begin{pmatrix} 0. & -0.0141299 & -0.0100331 \\ -0.0141299 & 0. & 0.0843337 \\ -0.0100331 & 0.0843337 & 0. \end{pmatrix}$$

FrobeniusDistance → Sqrt[Add[%^2]]

FrobeniusDistance → 0.121758

difhf – %% // Chop

$$\begin{pmatrix} 0 & -0.0153823 & -0.00410285 \\ -0.0153823 & 0 & -0.00870711 \\ -0.00410285 & -0.00870711 & 0 \end{pmatrix}$$

■ Review

When programming these routines, the method of nonnegative eigenvalues was selected as the default, since this both has a theoretical base and is relatively easy to understand and program. Precisely this method might still have problems due to numerical rounding off. In that case one might use the routine EigenChop and reapply the procedure.

associationMat

$$\begin{pmatrix} 1. & 0.494305 & -0.064702 \\ 0.494305 & 1. & -0.994933 \\ -0.064702 & -0.994933 & 1. \end{pmatrix}$$

```
ToPSDCorrelationMatrix[associationMat, Do → {"HighamF", UnitDiagonalMatrix}]
```

$$\begin{pmatrix} 1. & 0.464793 & -0.0788379 \\ 0.464793 & 1. & -0.919307 \\ -0.0788379 & -0.919307 & 1. \end{pmatrix}$$

```
newnc = ToPSDCorrelationMatrix[
    associationMat, Do → {"NonnegEV", UnitDiagonalMatrix}]
```

$$\begin{pmatrix} 1. & 0.464793 & -0.0788379 \\ 0.464793 & 1. & -0.919307 \\ -0.0788379 & -0.919307 & 1. \end{pmatrix}$$

```
ToPSDCorrelationMatrix[associationMat, Do → "Gradual"]
```

$$\begin{pmatrix} 1 & 0.480176 & -0.0747351 \\ 0.480176 & 1 & -0.9106 \\ -0.0747351 & -0.9106 & 1 \end{pmatrix}$$

Note that these adjustment approaches have been developed just for "back up". When we determine a nominal correlation matrix such that we make sure from the outset that it is PSD then there is no need to approximate PSD-ness. For the volume ratio measure we apparently can first see of a choice of some λ is sufficient.

Appendix K: Variances and regression coefficients

■ Introduction

With **R** a correlation matrix containing the pairwise simple regressions, for either real or nominal data, then the multiple regression coefficients can be found by the cofactors of **R** and the standard deviations of the variables. The "variance-covariance-regression matrix" C contains on each row a different explained variable, indicated by 0, and the relevant regression coefficients of the other variables, such that $x == C\,x + \epsilon$.

```
CovarRegress[All, {s1, s2, s3}]
```

$$\begin{pmatrix} 0 & -\frac{s1\,R[1,2]}{s2\,R[1,1]} & -\frac{s1\,R[1,3]}{s3\,R[1,1]} \\ -\frac{s2\,R[2,1]}{s1\,R[2,2]} & 0 & -\frac{s2\,R[2,3]}{s3\,R[2,2]} \\ -\frac{s3\,R[3,1]}{s1\,R[3,3]} & -\frac{s3\,R[3,2]}{s2\,R[3,3]} & 0 \end{pmatrix}$$

To apply above scheme for nominal data, we already have a correlation matrix. We only need to define variance to get regression coefficients. Variance is usefully defined on the border sum totals of a contingency matrix. For a single variable we thus get a list of frequencies per category, $f = \{f_1, ..., f_n\}$. The discussion below will discuss the choice of a variance measure for such a list.

■ Variance for a single category

Since this subject has a certain complexity, it might be useful to state what this appendix is *not* about. This appendix is about variables that have frequencies over categories, and this appendix is not about a single category itself. This distinction can best be expressed in a short manner so that one sees the difference. In the introductory pages we expressed that a correlation between categories might be expressed as $\rho_{\text{yi,xj}} = n_{i,j} / n_{+j} * \sigma_{\text{xj}} / \sigma_{\text{yi}}$ and thus some idea about those variances of the category elements might be useful.

When we assume a distribution for variable x then the distribution of x_j is an element in that, and the distribution of y_i follows from $y_i = c_i\, x$, where c_i is the row of conditional probabilities. Instead of thinking of y as a weighted sum it may also help to remember that y is only a re-allocation of the outcomes of x, via the individual observations in the C_t matrix. There are two obvious candidates for distributions for x, with consequences for the standard deviations:

1. When x is multinomial, then x_j may be taken as binomial, and y_i is a weighted sum of binomials that might be approximated again with a binomial. With the bernoulli distribution we get $\sqrt{p_j\,(1-p_j)}$ and $\sqrt{q_i\,(1-q_i)}$ and for the binomial we get an influence of n_{++} that however drops out of the ratio. Then $\rho_{\text{yi,xj}} = n_{i,j} / n_{+j}\, \sqrt{n_{+j}\,(n_{++}-n_{+j})} \,/\, \sqrt{n_{i+}\,(n_{++}-n_{i+})} = n_{i,j} / \sqrt{n_{+j}\; n_{i+}} * \sqrt{(n_{++}-n_{+j})/(n_{++}-n_{i+})}$. Then also $\rho_{\text{xj, yi}} = n_{i,j} / n_{i+}\, \sqrt{n_{i+}\,(n_{++}-n_{i+})} \,/\, \sqrt{n_{+j}\,(n_{++}-n_{+j})} = n_{i,j} / \sqrt{n_{+j}\; n_{i+}} * \sqrt{(n_{++}-n_{i+})/(n_{++}-n_{+j})}$. For symmetry the bigger square roots ought to be the same. Since they are not, one either takes a geometric average of these correlation coefficients and forgets about the standard deviations, or adopts the following model.
2. Alternatively, x is a result of n Poisson processes that are only dependent in that they don't occur at the same time. Then y is a weighted sum of such Poissons, that might be approximated by the Poisson distribution again, so that $\sigma_{\text{yi}} = \sqrt{n_{i+}}$ and $\sigma_{\text{xj}} = \sqrt{n_{+j}}$. We then readily find $\rho_{\text{yi,xj}} = n_{i,j} / \sqrt{n_{+j}\; n_{i+}}$. Now symmetry holds without a problem. This model hinges on the approach that

either there are such Poisson processes or that, while perhaps the overall model might be a multinomial, the pairwise outcome while neglecting the rest might still be modelled as Poisson.

As said, the above only holds for the categories. This appendix however is targetted at the variables.

■ Variance for nominal variables

Minimal and maximal dispersion

Intuitively, we can choose between normal Variance and the Pearson measure, that are orderless measures, as fitting for nominal data. Of these, the Pearson measure might be best since it allows a quick connection to Cramer's V since both have the χ^2 distribution so that their ratio is an F-distribution.

However, there arises a point that is counter-intuitive and that is both Variance and PearsonVariance are zero when all frequencies are equal, $f_1 = f_i$, while, for contingency tables, this actually means *maximal* dispersion. Variance and PearsonVariance are maximal when all weight is in one cell, but for contingency tables this means *minimal* dispersion.

Hence, proper measures should at least be someting like 1 / (1 + Variance) or 1 / (1 + Pearson). To better understand the issue it is useful to link up to ordinal data. We would like a smooth transition when nominal data suddenly can be seen as ordinal (when some hidden order is exposed).

Ordered or orderless

We can see the frequencies as orderless numbers as would be appropriate for nominal variables or adopt the alternative option to include the order of presentation. There always is an order of presentation, and we might caputilate for this and accept other ordinal methods. The two approaches are:

1. Use techniques from ordinal data, but maintain the interpretation of nominal data.
2. Use a measure of orderless variance, but interprete results while keeping in mind the order of presentation.

The correlation coefficients have been constructed without any notion of order. Yet, the determinant is sensitive to the presentation order. For example, Det[{{*a, b*}, {*c, d*}}] = *ad* - *bc* while if we rearrange the rows then Det[{{*c, d*}, {*a, b*}}] = *bc* - *ad.* It may also be observed that the notion of regression has some suggestion of order. In the hat shops example, using fitness as the variable to be explained and shops and colour as the explanatory variables, such that Fitness = $\alpha + \beta$ Shop + γ Colour, then the regression coefficient of β of the shops would have the interpretation that a move from shop1 to shop2 would shift β hats from fit to no-fit. The hat shops example only uses dichotomous variables and when we use the multinomial model then the interpretation would concern steps moving along the categories in their order of presentation. We can reason about this while maintaining strict nominality, yet why make life difficult, and why not allow for a variance measure that uses the order of presentation.

The routine VariancePr is defined such that if we have a list of frequencies (e.g. the column sums) of length *n* then it assigns values {0, ..., *n* -1} to the categories and then takes the normal variance of those values weighted by those frequencies. The technique is that of ordinal data but the interpretation remains nominal, since the only order used is the order of presentation.

VariancePr is an important routine, since it helps us to see a key difference with orderless Variance.

i. Variance and Pearson take equal frequencies as the lowest dispersion.

```
Variance /@ {{1, 3, 5, 10}, {4, 4, 4, 4}} // N
```

{14.9167, 0.}

ii. VariancePr takes equal frequencies as the case of highest dispersion.
VariancePr still checks the mass distribution over the category values.

```
VariancePr /@ {{1, 3, 5, 10}, {4, 4, 4, 4}} // N
```

{0.825485, 1.25}

One might want to check the formal implementation of VariancePr:

```
VariancePr[{x1, x2}, {p1, p2}]
```

$$p1\,(-p1\,x1 + x1 - p2\,x2)^2 + p2\,(-p1\,x1 - p2\,x2 + x2)^2$$

```
VariancePr[{0, 1}, {f1, f2} / (f1 + f2)] // Simplify
```

$$\frac{f1\ f2}{(f1 + f2)^2}$$

```
VariancePr[{f1, f2}] // Simplify
```

$$\frac{f1\ f2}{(f1 + f2)^2}$$

Since we would like that the transition from nominal to ordinal data is as smooth as possible, let us stop to consider the argument why equal frequencies might mean maximal dispersion, also for nominal data. The rationale has already been expressed above. Indeed, with all data in one cell, if the data are just meaningless numbers, then the variance is maximal. But if the data represent observations on some nominal variable, then only that category is observed, and we have no dispersion at all.

Thus, the ordinal measure was very helpfull to enlighted the issue of dispersion for contingency tables. The next point however is whether we should use it or not. We can compare the ordinal measure with the measure for NominalVariance, that we shall discuss shortly.

- The ordinal measure generates different values for the various arrangements of the data.

```
VariancePr /@ Permutations[{1, 3, 20, 100}] // N // Union
```

```
{0.194264, 0.203109, 0.225481, 0.255138, 0.27595,
  0.34879, 0.562695, 0.583507, 0.604318, 0.708377, 1.23329, 1.2645}
```

- Instead, given that we are discussing nominal data, it would be better to maintain orderlessness.

```
NominalVariance /@ Permutations[{1, 3, 20, 100}] // N // Union
```

```
{40.0484}
```

Orderless

The design choice becomes clearer when we consider the variance measures that assume orderless data. For reference we keep including the plain variance of the frequencies, both division by n - 1 or by n (with the latter called Maximum Likelihood Estimation (MLE)), and the Pearson or χ^2 measure on the frequencies. The possibly less familiar options can best be explained using their formulas. Let f be the list of frequency data and let n be the number of categories, then $N = \sum_{i=1}^{n} f_i$ and $p = f / N$, so that $\overline{f} = N / n$ is the average frequency and $\hat{p} = 1/ n$ would be theoretical probabilities under a uniform distribution. In the following "." stands for the improduct. The effective number of categories is included in this list though it is not a dispersion measure but it might be used to adjust n.

- EffectiveNumberOfCategories[f] = $(\sum_{i=1}^{n} f_i)^2 / (\sum_{i=1}^{n} f_i^2)$
- VarianceFreq[Power, f] = $(\prod_{i=1}^{n} f_i)^{1/n} / \overline{f}$ = the ratio between the geometric and the arithmetic average. It ranges between 0 and 1.
- VarianceFreq[Times, f] = $N (\prod_{i=1}^{n} p_i)$ = the product of the probabilities. For n = 2 it gives the variance of the binomial distribution. See below for a discussion of the multinomial model. Note that the denominator of NominalCorrelation uses the product of the row and column sums. If those can be taken as the standarddeviations (of row-variable and column-variable separately) then the numerator becomes the covariance. See the discussion below.
- VarianceFreq[Border, f] = $n^n (\prod_{i=1}^{n} p_i)$ = the product of the probabilities, with a better correction. (The default.)
- VarianceFreq[Sum, f] = $N\ p\ .\ (1 - p)$ = the sum of the pairwise variances.
- VarianceFreq[Average, f] = $N\ p\ .\ (1 - p) / n$ = the average of the latter
- VarianceFreq[N, f] = $\left(p.\sqrt{f\ (1 - p)}\right)^2$ = the weighted average, weighing the pairwise standard deviations. It might be multiplied with n to scale it up to the level of the Sum measure.
- VarianceFreq["Pearson", f] = $\sum_{i=1}^{n} (f_i - \overline{f})^2 \Big/ \overline{f} = N (n\ p\ .\ p - 1)$ = PearsonVariance (for levels). Divide by N for the p versus their theoretical value $1 / n$. The measure has been mentioned already but it is useful to see these two formats.

In some cases a variance measure will generate a zero outcome. This is awkward when theoretically transforming a correlation matrix into a variance-covariance matrix. To

allow for theoretical results, the routine NominalVariance puts out a symbolic value "Var" if the variance would be zero, otherwise it takes, by default, VarianceFreq[Border].

PM. The chosen default variance VarianceFreq[Border] ranges from 0 to 1 and thus lacks a scaling factor related to *N*. This is not material since it affects all variables and thus does not affect the ratio of the standard deviations required for the regression coefficients.

PM. If a frequency is zero, then some measures find that the category should be dropped while Variance still generates a value. A zero value for one frequency of two categories however is a pathological case, for which neither a correlation coefficient is defined, so that this result is not problematic.

Comparison

The various possibilities are listed below using some examples that show the key properties. Values {0, ..., n - 1} are assigned to the categories for the ordered measure only, and its effect can be seen from rearranging the frequencies $\{f_1, ..., f_n\}$. The data-entries thus concern frequencies, not to be confused with the values assigned to the categories. The most important point to observe in these tables is that Variance and Pearson have opposite reactions to other measures when we compare the situations with equal frequencies to those with inequality. Under equality, Variance drops to zero while VariancePr takes its highest value.

- For the $n = 2$ situation, the assignment of numerical values $\{c_1, c_2\}$ to the categories is symmetric. The measures are distinguished by their reaction to equal frequencies $f_1 = f_2$.

NominalVariance[Table, {{1, 0}, {0, 1}, {1, 2}, {2, 1}, {0, 4}, {1, 4}, {4, 4}}] // N

	{1., 0.}	{0., 1.}	{1., 2.}	{2., 1.}	{0., 4.}	{1., 4.}	{4., 4.
EffectiveNumberOfCategories	1.	1.	1.8	1.8	1.	1.47059	2.
NominalVariance	Var	Var	0.888889	0.888889	Var	0.64	1.
Order	–	–	–	–	–	–	–
VariancePr	0.	0.	0.222222	0.222222	0.	0.16	0.25
Orderless	–	–	–	–	–	–	–
VarianceMLE	0.25	0.25	0.25	0.25	4.	2.25	0.
Variance	0.5	0.5	0.5	0.5	8.	4.5	0.
PearsonVariance	1.	1.	0.333333	0.333333	4.	1.8	0.
VarianceFreq(Power)	0.	0.	0.942809	0.942809	0.	0.8	1.
VarianceFreq(Times)	0.	0.	0.666667	0.666667	0.	0.8	2.
VarianceFreq(Border)	0.	0.	0.888889	0.888889	0.	0.64	1.
VarianceFreq(Sum)	0.	0.	1.33333	1.33333	0.	1.6	4.
VarianceFreq(Average)	0.	0.	0.666667	0.666667	0.	0.8	2.
VarianceFreq(*N*)	0.	0.	0.666667	0.666667	0.	0.8	2.

- For the $n = 4$ situation, the assignment of numerical values $\{0, ..., 3\}$ to the categories no longer is symmetric for VariancePr. VariancePr no longer generates 0 when a (single) frequency is zero. The ordered and orderless measures again are distinguished by their reaction to equal frequencies $f_1 = f_i$. The VarianceFreq[Times] measure generates very small numbers.

NominalVariance[Table, {{1, 3, 5, 10}, {10, 3, 5, 1}, {0, 4, 4, 4}, {4, 4, 4, 4}}] // N

	{1., 3., 5., 10.}	{10., 3., 5., 1.}	{0., 4., 4., 4.}	{4., 4., 4., 4.}
EffectiveNumberOfCategories	2.67407	2.67407	3.	4.
NominalVariance	0.294657	0.294657	Var	1.
Order	–	–	–	–
VariancePr	0.825485	0.975069	0.666667	1.25
Orderless	–	–	–	–
VarianceMLE	11.1875	11.1875	3.	0.
Variance	14.9167	14.9167	4.	0.
PearsonVariance	9.42105	9.42105	4.	0.
VarianceFreq(Power)	0.736765	0.736765	0.	1.
VarianceFreq(Times)	0.0218691	0.0218691	0.	0.0625
VarianceFreq(Border)	0.294657	0.294657	0.	1.
VarianceFreq(Sum)	11.8947	11.8947	8.	12.
VarianceFreq(Average)	2.97368	2.97368	2.	3.
VarianceFreq(*N*)	3.8134	3.8134	2.66667	3.

Conclusions

Conclusions are:

1. For contingency tables orderless measures Variance and Pearson cannot be used (directly) since they measure the inverse of dispersion for contingency tables.
2. For contingency tables it is often important that all frequencies are equal, as this is sometimes a requirement for balanced observations. Orderless measures Variance and Pearson then break down with zero values as well. Other VarianceFreq measures then would be more appropriate. This adds to the problem of the direction of measuring dispersion.
3. The latter thus holds in general. Henceforth, we better not use Variance or Pearson. Of the alternative measures, Power and Times have low values and generate zeros when a category has zero frequency. Only VarianceFreq[Sum] generates values such that the square root is in the range of the frequencies themselves. It is definitely an interesting measure when choosing for orderlessness. It may be noted that it is most sensitive to the length of the list and a bit less sensitive to its contents, but it does show sufficient variation. However, from alternative considerations the measure VarianceFreq[Border] has been chosen as the default nominal variance.
4. If we didn't have VarianceFreq[Sum] or VarianceFreq[Border] then the ordered approach with VariancePr would be preferable above using Variance or Pearson, for the mentioned reasons.
5. In anticipation of the discussion below: the VarianceFreq[Border] has the advantages: (a) it links up to the volume ratio measure, (b) there may be a scaling factor related to N, but this holds for all variances in a contingency table, and hence is immaterial for the transformation of correlation coefficients to regression coefficients.

PM. Perhaps we need not worry much about zero values. We already have a correlation matrix, so we don't have to divide by zero but we multiply with it in order to create the variance-covariance matrix. Zero's are only a problem for the calculation of the regression coefficients, when we divide by the standard deviations. Here it holds: (a) When one row or column has no observations then it might be deleted. (b) When we have balanced observations (equal frequencies), we could use symbolic values and see how these work out. It turns out, however, that such cases occur in a non-negligible number of situation, and it is awkward to judge each individual case. Thus when we can avoid zero values then this is advisable.

■ Correction of the product of the marginals

The routine VariancePr2By2 copies the approach of CorrelationPr2By2, i.e. assigns values {i, j} to the categories (both row and column), and uses VariancePr[*x, p*] to calculate the implied variances. There are two main conclusions: (i) It appears that the values of *i* and *j* only disappear when $i = j$ +/- 1, for example {0, 1}. Thus an assignment {-1, 1} would introduce a perhaps arbitrary value of 2^2. (ii) The variance would be the product of the marginal probabilities. This gives the idea to see nominal correlation or the volume ratio measure as a fraction with numerator and denominator, with the numerator (determinant) as the covariance, and the denominator as the product of the standard deviations. This implies that a closer look at the denominator would generate a concept of variance.

- This routine gives separately the variances of the column sums and the row sums of a 2 × 2 matrix, when the categories get values *i* and *j*.

```
FullSimplify[VariancePr2By2[Definition, {{a, b}, {c, d}}, {i, j}],
  Assumptions → {Thread[{a, b, c, d} ≥ 0]}]
```

$$\left\{\frac{(a+c)\,(b+d)\,(i-j)^2}{(a+b+c+d)^2},\ \frac{(a+b)\,(c+d)\,(i-j)^2}{(a+b+c+d)^2}\right\}$$

In above discussion, this idea has been implemented with VarianceFreq[Times], such that for example the variance of the column variable is given as the product of the marginal probabilities of the columns (times *N*). However, a product of probabilities causes small numbers and possibly problems with numerical conditioning. To reduce this, the measure already includes a multiplication by *N*. For a 2 × 2 the measure then is equal to the variance of the binomial distribution, which was the inspiration for this adjustment, and indeed reasonable numerical values are gotten. However, for larger dimensions this does not help.

In itself, a measure with a product of the frequencies is attractive, given it relation to the volume ratio. The following is an exercise to look for a way to condition the outcome to reasonable values. Let N be the total number of observations and n the number of categories. An approach might be to compare a marginal probability with the theoretical value 1 / n, and in particular take the ratio. While the Pearson measure takes the sum of the squared deviations, or weighs the differences from 1 for the ratios, we might take: $p_i = f_i / N$, $P = \prod_{i=1}^{n} p_i$, $q_i = p_i / (1 / n)$, Var = $q_1 \ldots q_n = n^n \; p_1 \ldots p_n$. Indeed, the factor 2^2 mentioned above would have a rationale, and might be a better correction factor than the implemented multiplication with N. This has been implemented in VarianceFreq[Border].

- This compares $N\,P$, $n^n \; N\,P$, $n^n \; P$ and the inverse 1 / ($n^n \; P$).

```
testcor[lis_List] := Module[{n, v, vn, w}, n = Length[lis];
    v = VarianceFreq[Times, lis]; vn = v n^n; w = vn / Add[lis];
    {v, vn, w, 1/w} // N]

testcor /@ {{1, 3, 5, 10}, {4, 4, 4, 4}, {100, 100, 100, 100}}
```

$$\begin{pmatrix} 0.0218691 & 5.59848 & 0.294657 & 3.39378 \\ 0.0625 & 16. & 1. & 1. \\ 1.5625 & 400. & 1. & 1. \end{pmatrix}$$

The difference between 16 and 400 observations pays itself back in a more accurate estimate of the p_i but if the distribution does not change then P would not change. We again must choose between P and 1 / P. The latter 1 / P (or 1 / P - 1) has the advantage that it suits the normal idea of the variance, that equality of frequencies gives the lowest dispersion. (The choice of 1 / P - 1 would cause awkward zero values.) But as discussed, the variance and thus 1 / P give the inverse dispersion for contingency data. But sticking with P, these examples show that a proper conditioning is not found yet. Perhaps $n^n \; \sqrt{N}$ is better, but, this begs the question why. Quite possibly the correction n^n is sufficient, since variance values of 1 mean that the regression coefficients would only depend upon the correlation coefficients, which might be very acceptable.

The volume ratio is based upon a normalized matrix, whence the appearance of the row and column sums. It does not quite follow that such sums “thus” should appear in a measure for the variance.

This discussion is no longer pursued. However, given all consideration, the conclusion that the VarianceFreq[Border] measure serves as the best measure for nominal correlation seems warranted.

■ PearsonVariance

The Pearson variance is $\Sigma\,(f-t)^2\,/\,t$, where the theoretical frequency now is the mean. It is attractive to use this measure. For association matrices made on NominalCorrelation and similar matrices made with Cramer's V we would want to use the same variance, and since the Pearson variance and Cramer's V both have a χ^2 distribution then their ratio is F.

This leaves some questions to ask, though.

- Cramer's V had a limit value if 1, and thus was interpreted as a correlation and not as a covariance. Thus, in this case we would *multiply* with the standarddeviations and not *divide.* Conversely, Cramer's V uses the degrees of freedom but rather we should use an alternative measure with a standarddeviation, that would have a limit in the degrees of freedom.
- We would multiply (or divide) with a product of standarddeviations - getting products of χ^2 or square roots of those.
- The Pearson measure has an inverted sense of dispersion.

We need not resolve those questions and can maintain the earlier conclusions.

■ Multinomial

Taking the border sums for a variable in a contingency table gives us in fact a multinomial distribution. Standardly available are variances and covariances of a such a distribution. The "Generalized Variance" has been defined as the determinant of the variance-covariance matrix. This approach runs dead since that determinant for the multinomial distribution appears to be zero. It should be. The events are fully correlated, in the sense that if something happens in some category then this can only be since nothing happens in the other ones. For $n = 3$, for example, observations are {1, 0, 0}, {0, 1, 0}, {0, 0, 1} and then repeated in arbitrary number.

```
mnd = MultinomialDistribution[n, {p1, p2, p3}]
```

MultinomialDistribution(n, {p1, p2, p3})

```
Variance[%]
```

$\{n\,(1-\text{p1})\,\text{p1},\ n\,(1-\text{p2})\,\text{p2},\ n\,(1-\text{p3})\,\text{p3}\}$

cvm = CovarianceMatrix[mnd]

$$\begin{pmatrix} n\,(\text{p1}-\text{p1}^2) & -n\,\text{p1}\,\text{p2} & -n\,\text{p1}\,\text{p3} \\ -n\,\text{p1}\,\text{p2} & n\,(\text{p2}-\text{p2}^2) & -n\,\text{p2}\,\text{p3} \\ -n\,\text{p1}\,\text{p3} & -n\,\text{p2}\,\text{p3} & n\,(\text{p3}-\text{p3}^2) \end{pmatrix}$$

? GeneralizedVariance

GeneralizedVariance[{{x11, ..., x1p}, ..., {xn1, ..., xnp}}] gives the generalized variance of the n p−dimensional vectors. This is equivalent to the determinant of the covariance matrix, or the product of the variances of the principal components of the data. More…

Det[cvm]

$$-\text{p1}\,\text{p2}\,\text{p3}^2\,n^3-\text{p1}\,\text{p2}^2\,\text{p3}\,n^3-\text{p1}^2\,\text{p2}\,\text{p3}\,n^3+\text{p1}\,\text{p2}\,\text{p3}\,n^3$$

% // Simplify

$$-n^3\,\text{p1}\,\text{p2}\,\text{p3}\,(\text{p1}+\text{p2}+\text{p3}-1)$$

Appendix L: Sign of a determinant of a non-square matrix

Introduction

For a non-square matrix A (with $m > n$) the standard determinant is Sqrt[Det[$A'A$]]. This measure does not allow for a sign. Since we would require a general measure, we would also lose the sign of a square matrix. All this is unfortunate, and hence we want a way to give a sign to the determinant of a non-square matrix.

An example

The following square matrix has a negative determinant. Thus, the correlation measure defined especially for a square matrix generates that sign.

```
examat = {{10, 2, 5, 7}, {1, 5, 1, 0}, {2, 8, 8, 8}, {4, 8, 4, 3}}
```

$$\begin{pmatrix} 10 & 2 & 5 & 7 \\ 1 & 5 & 1 & 0 \\ 2 & 8 & 8 & 8 \\ 4 & 8 & 4 & 3 \end{pmatrix}$$

Det[examat]

-90

And thus a negative correlation

SquareMatrixNormedDet[examat]

$$-\frac{5}{4\sqrt{2028117}}$$

A first make-shift approach

In the first version(s) of this paper we used a make-shift solution by cutting up a non-square materix into four parts, by splitting it half-ways over rows and columns. We aggregated those four parts separately, and then took the sign of the resulting 2×2 matrix.

When we apply this method then the negative sign is not found.

NominalCorrelation[examat, ForceSign → {False, True}]

$$\frac{5}{4\sqrt{2028117}}$$

The new appraoch finds the sign

The following approach finds the sign. It uses by default a new method to find the sign.

NominalCorrelation[examat]

$$-\frac{5}{4\sqrt{2028117}}$$

Extending the definition of determinant to non-square matrices

The solution consists of extending the notion of determinant to non-square matrices:

$$\begin{aligned}
\text{NsqDet}(A) &= a_{11} \quad \text{if } n = 1 \text{ and } m = 1 \\
\text{NsqDet}(A) &= f(a) \text{ if } A = a \text{ is a vector, by default } f = \text{sum} \\
\text{NsqDet}(A) &= \textstyle\sum_{i=1}^{n} (-1)^{1+i} a_{1\,i}\, \text{NsqDet}(A_{1\,i})
\end{aligned}$$

Obviously, NsqDet = Det when the matrix is square.

There are two options:

1. Use this NsqDet as the general method for the volume ratio
2. Keep using the level Sqrt[Det[$A'A$]] and use NsqDet only for the sign

Above NsqDet is only a new suggestion that comes to mind after considering some alternatives. The suggestion seems to be the closest that we can get in regarding the coefficients of the matrix as weights. But as such, the suggestion is new and untested. Hence, approach no. 2 is the most sensible one. It makes for a somewhat curious circumstance, though, that a determinant is calculated twice, once for its level and once for its sign. Perhaps the theory on using Sqrt[Det[$A'A$]] needs reconsideration too. So it would help efficiency if there would be more clarity on this choice. As it is, though, the method seems to work.

Another example

There will be consistency with the square matrices since the non-square determinant gives the very same result for square matrices.

```
NonSquareDet[examat]
```

-90

The following is an example of a non-square matrix where a change in a single element causes a switch of sign.

```
mat1 = {{1, 2, 3}, {5, 6, 7}};
mat2 = {{1, 2, 3}, {5, 0, 7}};

NonSquareDet[mat1]
```

22

```
NonSquareDet[mat2]
```

-2

For nominal correlation, the determinant consists of the sign and the level:

```
Sign[NonSquareDet[mat2]] * Sqrt[Det[mat2. Transpose[mat2]]]
```

$-6\sqrt{10}$

Nominal correlation of course also normalizes for the border sums.

```
NominalCorrelation /@ {mat1, mat2} // N
```

{0.121716, −0.516398}

PM 1

Next to the function Det that is standardly available in *Mathematica*, we now have programmed:

(1) InDet, since it appears that Det does not evaluate when one element is Indeterminate

(2) SquareMatrixNormedDet that applies InDet to a normalized square matrix.

(3) NonSquareDet that is like Det but for non-square matrices

(4) VolumeRatio for *mat* that uses sgn[*mat*] * Sqrt[Det[$A'A$]] where A is the normalized *mat.*

PM 2

The method of splitting up a matrix - now the rejected method - works as follows.

```
exmp = Table[Random[Integer, {0, 10}], {i, 4}, {j, 9}]
```

$$\begin{pmatrix} 1 & 8 & 2 & 1 & 0 & 0 & 1 & 0 & 5 \\ 0 & 4 & 6 & 5 & 0 & 0 & 7 & 7 & 3 \\ 10 & 3 & 1 & 10 & 1 & 10 & 0 & 9 & 0 \\ 1 & 3 & 2 & 2 & 2 & 2 & 0 & 7 & 9 \end{pmatrix}$$

This is the sign using above method with the non-square determinant.

```
PairwiseSign[exmp]
```

−1

This is the sign using the method of splitting up.

```
PairwiseSign[exmp, ForceSign → {False, True}]
```

1

The rejected approach consists of splitting the rows and columns down the middle and aggregating the matrix into a 2 × 2 matrix and then use that sign. If some diagonal dominates some cross-diagonal then this property would seem to be preserved by such

aggregation. If aggregation creates a diagonal where none existed before then this changes the sign from 0 to 1, and this would have no effect on a correlation of zero.

```
{forRows = SplitAggregator[4], forCols = SplitAggregator[9]} // Transpose
```

$$\begin{pmatrix} \{1, 1, 0, 0\} & \{1, 1, 1, 1, 1, 0, 0, 0, 0\} \\ \{0, 0, 1, 1\} & \{0, 0, 0, 0, 0, 1, 1, 1, 1\} \end{pmatrix}$$

```
forRows . exmp . Transpose[forCols]
```

$$\begin{pmatrix} 27 & 23 \\ 35 & 37 \end{pmatrix}$$

```
Sign[Det[%]]
```

1

Appendix M: Notes on regression

Introduction

Colignatus (2007g) compares nominal correlation and regression with logistic regression, using the risk difference as a bridge for understanding.

More on: a contingency matrix as its own correlation matrix

The main body of the text showed that a contingency matrix can be seen as its own correlation matrix.

Som PMs the notation of that section:

PM 1. We might also determine standard deviations of the marginal probabilities p_j and q_i. Correlation coefficients then follow from the reverse of the above as $\rho_{i,j} = \rho_{\text{yi,xj}} = \beta_{\text{yi,xj}}\ \sigma_{\text{xj}} / \sigma_{\text{yi}} = c_{i,j}\ \ \sigma_{\text{xj}} / \sigma_{\text{yi}} = \ n_{i,j} / n_{+j}$ * $\ \sigma_{\text{xj}} / \sigma_{\text{yi}}$. The earlier approach however already gives correlation coefficients without requiring those standard deviations for the categories. The transform is useful, though, to note that the standard deviations of y_i and x_j would be related by the square roots of the numbers of observations.

PM 2. Above correlations will all be nonnegative, since $n_{i,j} \geq 0$. Possibly, $\beta_{\text{yi,xj}}$ should not be defined as the conditional probability P[y_i| x_j] but rather as $\overline{c_{i,j}}$ = P[y_i| x_j] - P[y_i|

Not[x_j]]. This pairwise regression coefficient keeps the total N constant. It comes about by using #[Not[x_j]] = N - n_{+j} and considering that y_i = P[y_i| x_j] n_{+j} + P[y_i| Not[x_j]] #[Not[x_j]] = $\overline{c_{i,j}}$ n_{+j} + P[y_i| Not[x_j]] N. In this pairwise regression between the categories, P[y_i| Not[x_j]] N is ithe constant and $\overline{c_{i,j}}$ the marginal contribution of one unit of the column sum, keeping the total constant. In that case the regression coefficients conceivably can have negative values too. This approach creates a linear dependence in the matrix of pairwise (adjusted) correlation coefficients. However, to impose some constraint on the overall effect seems also to be handled by taking the determinant (not proven though).

PM 3. Note this distinction:

1. For real valued data, a correlation matrix **R**, that contains the pairwise correlations between k variables, allows a measure of "overall correlation" as $R_O = \sqrt{1 - \text{Det}[\mathbf{R}]}$. For example, when all off-diagonal pairwise correlations are zero so that only the diagonal of 1 of the correlations of the variables with themselves remains, then the block of data shows no overall correlation and R_O = 0. Alternatively, when the data are multicollinear then there is full overall correlation and R_O = 1. This **R** concern variables and is square and symmetric.
2. The matrix C_ρ contains correlation coefficients but differs from such an **R**. C_ρ concerns categories, has size $m \times n$, and if it would be square then it would not necessarily be symmetric. If C_ρ would happen to be filled with zeros except for something that looks like a diagonal of 1 (as far as is possible in a rectangle), then we would say that the *variables* x and y are highly correlated since the categories of x directly translate into those of y.

The proposed measure for nominal correlation between variables uses the C_ρ on the categories to determine a pairwise correlation between x and y. The measure first determines the symmetric square C'_ρ C_ρ (assuming $m \geq$ n) and then applies the determinant to aggregate the information. The taking of the determinant reminds of overall correlation R_O. However, in this case there is no overall correlation but just pairwise correlation between two variables. The dimensions of the rows and columns would differ. The procedure stops at taking the determinant and there is no substraction from 1, since there still is the difference between points 1 and 2 above.

The procedure perfectly follows the geometry of volume ratio's, and this may help to see the difference between points 1 and 2. The notion of the volume ratio only applies for the transformation of C_ρ for the categories into the pairwise correlation of variables x and y. The other aspects have all their own other reasons.

Comparison to the risk difference in treatment control studies

The main body of the text uses the hat shop case, that is rather symmetrical in order to give all room to the Simpson paradox.

To get rid of the symmetries in this example, let us consider a random example. Here we find a difference between correlation and regression coefficient, and the regression coefficient is equal to the differences in probabilities.

```
matran = Table[Random[Integer, {0, 10}], {i, 2}, {j, 2}];
```

```
Move1FromRow1To2[matran] ~Join~ NominalStatistics[matran] // N
```

$$\left\{\text{Mat(In)} \to \begin{pmatrix} 4. & 0. & 4. \\ 4. & 9. & 13. \\ 8. & 9. & 17. \end{pmatrix}, \text{Mat(Out)} \to \begin{pmatrix} 3. & 0. & 3. \\ 4.30769 & 9.69231 & 14. \\ 7.30769 & 9.69231 & 17. \end{pmatrix},\right.$$

$$\text{Row[1.]} \to \{-1., 0.\}, \text{Row[2.]} \to \{0.307692, 0.692308\}, \text{Dif} \to \{-0.692308, 0.692308\},$$
$$\text{ContingencyTableQ} \to \text{True}, \text{OverallCorrelation} \to 0.588348, \text{Length} \to \{2., 2.\},$$
$$\text{EffectiveNumberOfCategories} \to \{1.56216, 1.9931\}, \text{Variance} \to \{0.719723, 0.99654\},$$
$$\text{Spread} \to \{0.848365, 0.998268\}, \text{BorderTotals} \to \begin{pmatrix} 4. & 13. \\ 8. & 9. \end{pmatrix},$$
$$\text{BorderMatrices} \to \left\{\{1., 2.\} \to \begin{pmatrix} 4. & 0. \\ 4. & 9. \end{pmatrix}\right\}, \text{NominalCorrelationMatrix} \to \begin{pmatrix} 1. & 0.588348 \\ 0.588348 & 1. \end{pmatrix},$$
$$\left.\text{CovarMat} \to \begin{pmatrix} 0.719723 & 0.49827 \\ 0.49827 & 0.99654 \end{pmatrix}, \text{CovarRegress} \to \begin{pmatrix} 1.11022 \times 10^{-16} & 0.5 \\ 0.692308 & 0. \end{pmatrix}\right\}$$

Again, this remains a the special case of the 2 × 2 table. See Colignatus (2007g) for fuller discussion.

Appendix N: Two more practical examples

An example from Losh (2004a)

The following example for nominal data concerning (1) sex, (2) email use and (3) education have been taken from Losh (2004a):

"Consider the following example from the August 2000 Current Population Survey that examines the bivariate association between education and the use of email at home for job or money-making purposes (for example, someone who has an online business or

who places auction items on eBay). Here I am *only* considering the 31,576 valid (weighted) cases who are persons with online access at home. The more education a person has, the more likely they are to use email for commercial purposes at home. However, because men are more likely to be self-employed than women (and more likely to be employed at all), men may use home email for commercial purposes more than women do, so both education and gender may be independent variables. In fact, men are about 3 percent more likely than women to use home email for commercial purposes in this sample."

CT[Set, "Email study"]

		Yes	No
Male	< Highschool	130	1510
	Highschool	344	2605
	Some college	736	3849
	Bachelor	790	3049
	Advanced	533	1788
Female	< Highschool	99	1477
	Highschool	338	3246
	Some college	725	4515
	Bachelor	673	3358
	Advanced	330	1481

The key result of this paper is that correlation and regression coefficients are made available, based upon a determinant or volume ratio measure of association. Regression coefficiencts follow from both the correlation matrix and the variances. Though the nominal data have no order, there is an order of presentation. For these tables, "not using email", "becoming female" and "getting more education" are positive steps in the right direction.

NominalStatistics[CT["Email study", Data]] // Chop

$$\Big\{\text{ContingencyTableQ} \to \text{True},\ \text{OverallCorrelation} \to 0.133618,$$
$$\text{Length} \to \{2, 2, 5\},\ \text{EffectiveNumberOfCategories} \to \{1.99835, 1.33922, 4.36221\},$$
$$\text{Variance} \to \{0.999173, 0.506589, 0.668284\},\ \text{Spread} \to \{0.999586, 0.711751, 0.817486\},$$
$$\text{BorderTotals} \to \{\{15334, 16242\}, \{4698, 26878\}, \{3216, 6533, 9825, 7870, 4132\}\},$$
$$\text{BorderMatrices} \to \left\{\{1, 2\} \to \begin{pmatrix} 2533 & 12801 \\ 2165 & 14077 \end{pmatrix}, \{1, 3\} \to \begin{pmatrix} 1640 & 2949 & 4585 & 3839 & 2321 \\ 1576 & 3584 & 5240 & 4031 & 1811 \end{pmatrix},\right.$$
$$\left.\{2, 3\} \to \begin{pmatrix} 229 & 682 & 1461 & 1463 & 863 \\ 2987 & 5851 & 8364 & 6407 & 3269 \end{pmatrix}\right\},$$
$$\text{NominalCorrelationMatrix} \to \begin{pmatrix} 1. & 0.0404591 & 0.0661467 \\ 0.0404591 & 1. & 0.118678 \\ 0.0661467 & 0.118678 & 1. \end{pmatrix},$$
$$\text{CovarMat} \to \begin{pmatrix} 0.999173 & 0.0287849 & 0.0540517 \\ 0.0287849 & 0.506589 & 0.0690521 \\ 0.0540517 & 0.0690521 & 0.668284 \end{pmatrix},$$
$$\text{CovarRegress} \to \begin{pmatrix} 0 & 0.0464504 & 0.0760817 \\ 0.0233211 & 0 & 0.101441 \\ 0.0502518 & 0.133453 & 0 \end{pmatrix}\Big\}$$

For the matrix of regression coefficients ("CovarRegress") C and variables $x = \{x_1, x_2, x_3\}$, the relation is $x = C.x + \epsilon.$. The matrix C is not symmetric since it matters in regression what the explained variable is.

For these data, some of the conclusions are as follows: Sex and email use have a correlation coefficient of 4%, and if one male in the study were replaced by one woman then the practice of "not using email" (second row) would rise by 0.02 (the {2, 1} cell). Education and email use have a correlation coefficient of -11.9%. If education rises one step then not using email rises by -0.10 (the {2, 3} cell). The latter might be negative due to the larger percentage of women in the study who use less email.

(These coefficients would gain in interpretative value if they could be transformed into percentages and elasticities.)

An example by Linacre (2005)

Linacre (2005) gives some data referring to Uebersax (2000), with a Pearson product moment correlation of 61% and a Polychoric correlation of 67%. The important point of these data are that they are *real* data, i.e. that these are counts from a grid where the values 1, 2, etcetera are the mean values of the cells in the grid. For such data the Pearson product moment measure would be allowable. If we treat these data as mere counts then there is no correlation. PM. The CT routine requires that all categories are uniquely labelled. The labels A to E actually would be values 1 to 5 as well.

CT[Set, "Uebersax (2000)"]

	A	B	C	D	E
1	0	0	12	32	40
2	0	4	23	66	23
3	1	10	67	77	15
4	1	22	133	40	3
5	8	71	125	21	2

NominalStatistics[%]

$$\{\text{ContingencyTableQ} \to \text{True},\ \text{OverallCorrelation} \to 0.000445384,\ \text{Length} \to \{5, 5\},$$
$$\text{EffectiveNumberOfCategories} \to \{4.50837, 3.11002\},\ \text{Variance} \to \{0.731728, 0.0737838\},$$
$$\text{Spread} \to \{0.855411, 0.271632\},\ \text{BorderTotals} \to \begin{pmatrix} 84 & 116 & 170 & 199 & 227 \\ 10 & 107 & 360 & 236 & 83 \end{pmatrix},$$
$$\text{BorderMatrices} \to \left\{\{1, 2\} \to \begin{pmatrix} 0 & 0 & 12 & 32 & 40 \\ 0 & 4 & 23 & 66 & 23 \\ 1 & 10 & 67 & 77 & 15 \\ 1 & 22 & 133 & 40 & 3 \\ 8 & 71 & 125 & 21 & 2 \end{pmatrix}\right\},$$
$$\text{NominalCorrelationMatrix} \to \begin{pmatrix} 1. & 0.000445384 \\ 0.000445384 & 1. \end{pmatrix},$$
$$\text{CovarMat} \to \begin{pmatrix} 0.731728 & 0.000103488 \\ 0.000103488 & 0.0737838 \end{pmatrix},\ \text{CovarRegress} \to \begin{pmatrix} 0. & 0.00140258 \\ 0.000141429 & 0. \end{pmatrix}\}$$

CT["Uebersax (2000)", Source]

http://www.rasch.org/rmt/rmt193c.htm
Pearson correlation = 0.61
Polychoric correlation = 0.67

PM. For completeness we mention Cramer's V.

```
CramersV[CT["Uebersax (2000)", Data]]
```

0.352573

Appendix O: Manipulating a contingency table

Introduction

Above discussion occasionally uses the routine CT ("contingency table") for input. The following clarifies its use. The routine CT allows you to order, sum and take elements from a contingency matrix with a bit more ease. The routine requires that you give labels for the variables and their categories, and then it gives you the benefit to communicate in terms of those labels instead of numbers. For example, when you would sum a variable, so that it effectively disappears from the table, the numbers would change but the labels remain the same. Obviously, typing some numbers, say the 8 numbers in the $2 \times 2 \times 2$ case, is probably quicker, but some control appears useful, also for subsequent routines, and also for presentation.

The routine dbCT is a databank that allows to to store contingency tables and use labels to find them. The dimensions of a table contain the labels for the variables and their categories. The databank also contains a source label that allows you to record where the data came from.

The routine CT allows you to take a contingency table from the database and set it as a default, meaning that its identifying label is available as CT[] and that routines will call this CT[] if no other label is specified.

- The options of CT already contain an example crosstable. The Dimensions specify the variable and its categories. In this example it so happens that the Dimensions have all equal length but this need not be so in general. The options of CT also specify what the default Databank is.

Options[CT]

$$\left\{\text{Label} \to \text{Arthritis},\ \text{Dimensions} \to \begin{pmatrix} \text{Effect} & \text{Some} & \text{None} \\ \text{Treatment} & \text{Active} & \text{Placebo} \\ \text{Sex} & \text{F} & \text{M} \end{pmatrix},\right.$$

$$\text{Data} \to \begin{pmatrix} \{21, 7\} & \{13, 1\} \\ \{6, 7\} & \{19, 10\} \end{pmatrix},\ \text{Source} \to$$

http://www.math.yorku.ca/SCS/Courses/grcat/grc6.html and Koch & Stokes (1991), Reordered,

$$\left.\text{Databank} \to \text{dbCT}\right\}$$

When the routine CT sets a database up for use, it first checks whether the dimensions fit the datastructure. When all checks are OK then the table components are allocated to standard locations such as CT[*label*, Data], CT[*label*, Dimensions] and so on. This means that you and the routines can trust that these locations form a well-defined working space.

- This sets the default crosstable using the options. This also prints the table using TableForm. As one can see in the options, the Effect variable is the second one, and due to TableForm printing it appears as the main column, which is easiest for our understanding. Sex and Treatment are the explanatory variables around that main column.

CT[Set, Default]

		Active	Placebo
Some	F	21	13
	M	7	1
None	F	6	19
	M	7	10

Basic handling

- The table is available as default.

CT[]

Arthritis

CT[Data]

$$\begin{pmatrix} \{21, 7\} & \{13, 1\} \\ \{6, 7\} & \{19, 10\} \end{pmatrix}$$

CT[Variables]

{Effect, Treatment, Sex}

CT[TableHeadings]

$$\begin{pmatrix} \text{Some} & \text{None} \\ \text{Active} & \text{Placebo} \\ \text{F} & \text{M} \end{pmatrix}$$

- Elements in the table now are available by naming them. Admittedly, it is quicker to simply type 13 than this long identifier, yet, for programming it can be handy, it may reduce typing errors, and it may be more convenient when the table has more dimensions and larger numbers, where the layout on the screen might sometimes be confusing. In this case, the variables are defined in the order *effect, cause, confounder*, and selecting values should be in that order.

CT[Take, "F", "Some", "Placebo"]

CT::mis : Unknown query keys or wrong order: {*F, Some, Placebo*}

CT[Take, "Some", "Placebo", "F"]

13

Summing

- When we sum Treatment then it disappears from the table. The original table remains the default unless we also specify that the new result should become the default.

CT[Sum, "Treatment"]

	F	M
Some	34	8
None	25	17

CT[]

Arthritis

CT[Sum, Default, "Treatment"]

	F	M
Some	34	8
None	25	17

CT[]

Arthritis−Treatment

- CT[Set, ...] keeps a list of the labels that are set.

CT[List]

{Arthritis, Arthritis−Treatment}

- This gets us back to the original table. Use CT[Default, ...] for contingency tables that have already been Set.

CT[Default, "Arthritis"]

		Active	Placebo
Some	F	21	13
	M	7	1
None	F	6	19
	M	7	10

Ordering

- Sometimes another look at the table can be helpful.

CT[Order, "Treatment"]

		Some	None
Active	F	21	6
	M	7	7
Placebo	F	13	19
	M	1	10

CT[Order, {"Sex", "Treatment"}]

		Active	Placebo
F	Some	21	13
	None	6	19
M	Some	7	1
	None	7	10

- The list of contingency tables that have been set now also includes these reordered tables. The dash - means a summation (or a minus, because of removal) while -1- means a reordering.

CT[List]

{Arthritis, Arthritis−Treatment, Arthritis−1−Treatment, Arthritis−1−Treatment−1−Sex}

Working with the databank

- Since other packages may have been loaded that also set the SetDatabank default databank, we set it now. Once we do this, the various Databank routines apply to the default dbCT.

SetOptions[SetDatabank, Databank → dbCT]

{Databank → dbCT}

- This query tells us what contingency tables are available, identified by their short labels.

Databank[Query, Label]

```
{Hat shops
 Arthritis−Original
 Arthritis2
 Arthritis
 Email study
 Uebersax (2000)
 Cornfield
 Fisher Male Twins
 Fisher Female Twins
 Fisher Twins
 Arthritis−Original−Sex−1−Treatment}
```

- The following gives an explanation of the contents of the first contingency table. Note that the Dimensions and Data elements are put in Hold. The reason is that Databank only accepts lists of one level.

Explain[dbCT[Data][[1]]]

$$\Big\{\text{Label} \to \text{Hat shops},\ \text{Dimensions} \to \text{Hold}\left[\left[\begin{pmatrix} \text{Shop} & \text{Shop1} & \text{Shop2} \\ \text{Colour} & \text{Green} & \text{Blue} \\ \text{Fitness} & \text{Fit} & \text{No fit} \end{pmatrix}\right]\right],\ \text{Data} \to \text{Hold}\left[\left[\begin{pmatrix} \{5,\ 1\} & \{8,\ 2\} \\ \{2,\ 8\} & \{1,\ 5\} \end{pmatrix}\right]\right],$$

$$\text{Source} \to \text{Kleinbaum et al. (2003), "ActivEpi Companion texbook", Springer}\Big\}$$

- This gives all records of the databank.

ShowData[Transpose]

	Label	
1	Hat shops	
2	Arthritis−Original	H
3	Arthritis2	Hold[{{Sex, F, M}, {Effe
4	Arthritis	H
5	Email study	Hold[{{Sex, Male, Female}, {Email use, Yes,
6	Uebersax (2000)	
7	Cornfield	Hold[
8	Fisher Male Twins	Hold
9	Fisher Female Twins	Hold[
10	Fisher Twins	Hold
11	Arthritis−Original−Sex−1−Treatment	H

- CT has access to the databank and we might set the default to the Arthritis2 table.

CT[Set, Default, "Arthritis2"]

		None	OnlySome	Marked
F	Active	6	5	16
	Placebo	19	7	6
M	Active	7	2	5
	Placebo	10	0	1

Thus, in setting or declaring something to become the default, CT[Set, (Default,) Label → ...., Dimensions → ...., ...] works from the options, CT[Set, (Default,) *label*] works from the databank, and CT[Default, *label*] works from tables that have already been set. Having set a crosstable, the following key data are available, for example for above new default:

CT[Information]

Length	3
Levels	{2, 3, 2}
Dimensions	{{Sex, F, M}, {Effect, None, OnlySome, Marked}, {Treatment, Active, Placebo}}
Variables	{Sex, Effect, Treatment}
TableHeadings	{{F, M}, {None, OnlySome, Marked}, {Active, Placebo}}
Data	({6, 19} {5, 7} {16, 6} {7, 10} {2, 0} {5, 1})
Source	http://www.math.yorku.ca/SCS/Courses/grcat/grc6.html and Koch & Stokes (199

Dumping to the databank

When a contingency table has been created via various elaborations, then it can be dumped to the databank, that hence might be stored on a medium and retrieved for later use.

- This creates a new default table. The entry number refers to the database. These are the original Arthritis data, in a different order.

CT[Set, Default, 2]

		None	Some
F	Active	6	21
	Placebo	19	13
M	Active	7	7
	Placebo	10	1

CT[Sum, Default, "Sex"]

	Active	Placebo
None	13	29
Some	28	14

CT[Order, Default, "Treatment"]

	None	Some
Active	13	28
Placebo	29	14

CT[]

Arthritis−Original−Sex−1−Treatment

- The CT[AppendTo] command appends the default table to the default databank. If you want to adapt some information, you can first directly edit them such as CT[CT[], Source] = ... The label can be adjusted by CT[CopyData, new (, old)].

CT[AppendTo]

Hat shops	Hold[{{Shop, Shop1, Shop2}, {Colour, Green, Blue}, {Fitness, Fit, No fit}}]
Arthritis−Original	Hold[{{Sex, F, M}, {Effect, None, Some}, {Treatment, Active, Placebo}}]
Arthritis2	Hold[{{Sex, F, M}, {Effect, None, OnlySome, Marked}, {T
Arthritis	Hold[{{Effect, Some, None}, {Treatment, Active, Placebo}, {Sex, F, M}}]
Email study	Hold[{{Sex, Male, Female}, {Email use, Yes, No}, {Educa
Uebersax (2000)	Hold[{{Row, 1, 2, 3, 4, 5}, {Column, A, B, C, D, E}}]
Cornfield	Hold[{{Effect, Disease, ¬ Disease}, {Truth, Cause, ¬ Cause}, {Confounding, Confounder, ¬ Confouder}}]
Fisher Male Twins	Hold[{{Smoking habits, Alike, Unlike}, {Twins, Fraternal, Identical}}]
Fisher Female Twins	Hold[{{Smoking habits, Alike, Unlike}, {Twins, Fraternal, Identical}, {Separated at birth, Separated, ¬ Separated}}]
Fisher Twins	Hold[{{Smoking habits, Alike, Unlike}, {Twins, Fraternal, Identical}, {Sex, Male, Female}}]
Arthritis−Original−Sex−1−Treatment	Hold[{{Treatment, Active, Placebo}, {Effect, None, Some}}]
Arthritis−Original−Sex−1−Treatment	Hold[{{Treatment, Active, Placebo}, {Effect, None, Some}}]

More information

The Databank package may have some standard routines available that might be of use as well.

Economics[Databank]

Cool`Databank`

AfterSum BeforeSum Databank DataMold Explain SetDatabank ShowData Upd

Appendix P: Routines

■ Introduction

This discussion uses The Economics Pack, Cool (2001). The basic routines are all defined with the more technical term "VolumeRatio" and the routine "NominalCorrelation" builds on those.

Note the distinction between OverallCorrelation and NominalCorrelation. I have considered to replace the subtitle of this paper "A measure of association or correlation" with the more extensive "Measures of association and correlation", to indicate that this paper discusses more angles. Yet, the focus is on the proposal for nominal correlation.

This analysis started out with a small case, clarified and simplified by *Mathematica*. The subsequent steps in generalization benefitted from that testing and prototyping environment as well. The availability of linear algebra routines and matrix manipulations was essential. The routines rely on symbolic operations. The environment also allowed the quick creation of user friendly routines, that not only have a clear logical structure but also come with help support and all. Finally, these routines also form building blocks that can be used immediately within other routines. All in all, there is yet another reason to thank the makers of *Mathematica*.

■ For real or nominal data

? CovarRegress

CovarRegress[vcm] with vcm a variance−covariance matrix gives
the coefficients of OLS (of the first variable on the other ones) as −
s1 R1i / (si R11), where si is the standarddeviation of variable i (Sqrt
diagonal) and where Rij is the (i, j) cofactor of the correlation matrix
CovarRegress[cor, s] with cor a correlation matrix and s a
vector of standard deviations, gives the same
CovarRegress[All, cor, s] gives the matrix of OLS regression coefficients C, where each row has a 0 for the explained variable, while the off−diagonal coefficients are taken on the right hand side. Thus if x is the list of variables {x1, ..., xn} then x == C.x + eps
CovarRegress[All, s] gives an explanation in terms of the
cofactors of a correlation matrix, use a symbolic s

? AssociationMatrixQ

AssociationMatrixQ[m] is CORAMatrixQ[m, Full −> False]

? CorrelationMatrixQ

CorrelationMatrixQ[m] is CORAMatrixQ[
m, Full −> True], which is AssociationMatrixQ[m] && PSD

? CORAMatrixQ

CORAMatrixQ[m] is True if m is a square symmetric matrix with 1 on the diagonal and elements −1 <= # <= 1, and otherwise False. If this is satisfied then the elements can be called pairwise associations. When the option Full −> True, then the routine also tests 0 <= Det[m] <= 1. If true the matrix is positive semi−definite (PSD) and is a full correlation matrix. Default Full −> Indeterminate since a call should be done with AssociationMatrixQ or CorrelationMatrixQ that set this option. Option MessagesQ −> (default True) controls printing of diagnostic messages

? OverallCorrelation

OverallCorrelation[mat] gives the overall correlation in correlation matrix mat. Options are:
(1) Mode −> MultipleRSquared (default) | Det
(1a) If Det is selected then the routine returns Sqrt[1 − Det[mat]]
(1b) If MultipleRSquared is selected, then this routine is applied for i = 1, ..., Length[m]; such results can be found in lis = MultipleRSquared[List]. Then option Function −> f (default Max) is applied to that lis; output is Sqrt[f[lis]]. Results[OverallCorrelation] helps to identify the variable with the highest correlation
(2) Check −> True | Return | False. False is default, does no test, since you would only apply this routine to a true positive semi−definite correlation matrix. Values True | Return tests mat on being a correlation matrix, while Return quits the routine if the test fails. See also the options of CorrelationMatrixQ. PM 1. This option would be relevant if you run the mistake to apply this routine to a square contingency table, where you should use NominalCorrelation. PM 2. Since Det[mat] is always computed, there is a default test on 0 <= Det[Mat] <= 1. Suppress a message by MessagesQ −> False, e.g. in CorrelationMatrixQ

? MultipleRSquared

MultipleRSquared[mat, i] = 1 − Det[mat] / CoFactor[mat, i, i], and gives R^ 2[i; 2, ..., i−1, i+1, ..., n] or the squared multiple correlation coefficient from the OLS regression of ith (dependent) variable upon the other (independent) variables, provided that mat is the CorrelationMatrix between all variables
Option Range with value False (default) does no test, with values True | Return tests mat on being a correlation matrix, while Return quits the routine when the test fails
MultipleRSquared[mat] takes i = 1
See OverallCorrelation, and note the difference between R and R^2
See the Estimate package where the MultipleRSquared is simply called RSquared

? ToPSDCorrelationMatrix

ToPSDCorrelationMatrix[m] for association matrix m gives the closest approximation R such that R can be called a correlation matrix (in the conventional strong sense): symmetric, diagonal ones, −1 <= R[i, j] <= 1, and 0 <= Det[R] <= 1. Given the first conditions the last implies positive semi−definiteness (PSD−ness). The approximation works best when m is already close to being a correlation matrix. Option Do −> ... gives the manner how this is done:
(1) Default {"NonnegEV", UnitDiagonalMatrix} first adjusts the eigenvalues (all nonnegative) and then applies UnitDiagonalMatrix, see ToPSDMatrix
(2) {"HighamF", UnitDiagonalMatrix} uses Higham's method, see ToPSDMatrix
(3) "Proportional" uses m x with uniform scalar x for off−diagonal elements
(4) "Power" uses m^x for off−diagonal elements
(5) "Gradual" uses m^x x for off−diagonal elements (works best)
ToPSDCorrelationMatrix[{a, b, x}, m] for a and b Symbols or 1 or 0 and x a Symbol explains these numerical adjustments (3), (4) and (5)

? ToPSDMatrix

ToPSDMatrix[m] gives the closest positive semi−definite PSD
approximation to square matrix m. Option Apply −> ... gives the manner.
(1) Default "NonnegEV" sets negative eigenvalues to zero and updates the
positive values proportionally so that sum again gives the dimension
(2) "HighamF" uses the Frobenius norm for "closeness" and applies the formula from
N. J. Higham (1989), "Matrix nearness problems and applications". In M. J. C.
Gover and S. Barnett (eds), Applications of Matrix Theory, pages 1–27. Oxford
University Press. Real values only. Thus A = (m + Transpose[m])/2; {U, H} =
PolarDecomposition[A], i.e. A = U.H or U = A.Inverse[H]; higham = (A + H) /
2; but warrant symmetry by S = (H + Transpose[H])/2; higham2 = (A + S) / 2
NB. These manners do not guarantee that the diagonal is 1. If you are working on a
correlation matrix then also consider UnitDiagonalMatrix after applying ToPSDMatrix

? EigenChop

EigenChop[m] chops the square matrix m's eigensystem with delta, default option Chop −>
10^(−6). Value Automatic gives the system delta, False doesn't chop and thus allows you
to check that, with mat = Transpose[ev] we get m = mat . DiagonalMatrix[v] . Inverse[
mat]. For a symmetric matrix the Inverse is the same as the Transpose. The eigensystem
is in Results[EigenChop], with Mat = Transpose[ev] the eigenvectors in the columns.
Output is the reassembled matrix. You may check the difference m − EigenChop[m]

■ For nominal data

? NominalStatistics

NominalStatistics[c] with c a contingency table gives various statistics. From
the NominalCorrelationMatrix and Variance, a CovarMat is determined,
which allows a call to CovarRegress. Option N controls application of N[.]
to results, values N (default) or Identity, and NominalVariance −> Automatic
uses the default setting of NominalVariance with formal replacement of zeros
NominalStatistics[Results, f_Symbol, heading_List (, heading2)] can be run
afterwards with appropriate tableheadings (only once for square matrices

? EffectiveNumberOfCategories

EffectiveNumberOfCategories[f] for f a
list of frequencies per category gives Add[f]^2/ Add[f^2]

? ContingencyTableQ

ContingencyTableQ[x] is True when x is a contingency table and False otherwise. A vector is not a contingency table; this at least requires a matrix with one row. All sublists must be of equal length, see ArrayQ. Option Element −> ... has values (1) Automatic, means that all elements must be natural numbers (nonnegative integers) or Symbol. An alternative value is e.g. IntegerQ. (2) None, no test done on the elements, input as a whole must only be an Array, (3) your own test on the elements. E.g. the None effect is also got with Element −> ((#; True) &)

? CorrelationPr2By2

CorrelationPr2By2[{{n11, n12}, {n21, n22}}] gives the measure of correlation for a contingency table of two binary nominal variables (correlation and not just association)
CorrelationPr2By2[mat] may also take a 3x3 matrix but then the borders are seen as sum totals, and dropped
CorrelationPr2By2[Definition, mat (, {i, j})] gives the original definition without simplification based upon the nonnegative values (using binary values i and j)
Let C (cause) be the column variable and E (effect) the row variable. In logic, the variables take values {1, 0}. Here it is better to take {1, −1} so that equal numbers of observations give a zero mean. Output then is the normal Pearson CorrelationPr[{1, 1, −1, −1}, {1, −1, 1, −1}, {n11, n21, n12, n22}].
See SquareMatrixNormedDet for larger {n, n} and VolumeRatio and CramersV for {n, m} contingency tables
See NominalCorrelation for n1 by n2 by ... by nm in general
*** AddedUsage by Cool`Survival`Epidemiology`: ***
CorrelationPr2By2[] takes default DiseaseTestMatrix[]. Note that this measure has already been presented by Matthews 1975 but see NominalCorrelation too

? SquareMatrixNormedDet

SquareMatrixNormedDet[mat] gives InDet[NormalizedMatrix[mat]] for square contingency table mat. This routine retains the sign of the direction of the association (ordinal or presentation order). See VolumeRatio and NominalCorrelation. For 2 by 2 matrices, see CorrelationPr2By2. InDet returns Indeterminate if its input has such element

? NormalizedMatrix

NormalizedMatrix[mat] divides the elements of a n by m matrix by the square roots of their appropriate row and column sums. Note that repeated application doesn't generate the same result. Is primarily used in VolumeRatio and NominalCorrelation

? NominalCorrelation

NominalCorrelation[mat] translates into VolumeRatio[mat] for n bij m dimensional contingency table mat (just a pairwise correlation number), and OverallCorrelation[NominalCorrelationMatrix[mat]] for n1 x n2 x .... x nk dimensions and k > 2 (a correlation score based upon a full correlation matrix). Options are passed on to the appropriate subroutines. PM. For square mat you might add the option NaturalNumberQ −> True to test on the use of natural numbers (nonnegative integers) for contingency tables (for possible confusion with OverallCorrelation on a square correlation matrix). The latter gives a warning test only, and the option is not used in VolumeRatio itself

? NominalCorrelationMatrix

NominalCorrelationMatrix[mat] basically gives f[R] = R[w] = (1 − w) VR[B] + w VR[meth], where VR = VolumeRatioMatrix, B = BorderMatrices and meth is the chosen BordersOrAll −> meth option. Selected is the w such that the desired method is retained as much as possible but with a positive semi−definite (PSD) outcome. There are the following aspects to consider:

(1) If the intermediate result R[w] appears to be non−PSD, then ToPSDCorrelationMatrix[R[w]] is put out
(2) If meth = BorderMatrices, then of course w = 0
(3) If meth = Show then this only shows like in VR
(4) See Options[NominalCorrelationMatrix] for the allowed BordersOrAll options in VR
(5) Factor w is determined when Bounds −> True (default). Use Bounds −> False if you just want the results of VR without any tampering. Note that this also means that no PSD approximation is used
(6) Factor w is found with SolveDetIn0To1, using FindRoot. A possible call to ToPSDCorrelationMatrix may cause a repeated call of SolveDetIn0To1 e.g. if the Gradual way is used
(7) When both VR[B] and VR[meth] are non−PSD, then VR[meth] is directly adjusted towards PSD−ness, instead of first moving to VR[B] and only then adjusting; find the latter by directly calling BordersOrAll −> BorderMatrices
(8) Results[NominalCorrelationMatrix, ...] contain the relevant details, check these by ??Results

? VolumeRatio

VolumeRatio[mat] gives a measure of association or correlation in a contingency matrix of size n by m. There are three approaches, indicated by the option VolumeRatioMethod:
(i) "Weak" (default): (1) M = NormalizedMatrix[mat], (2) ma = M.M' or ma = M'.M whichever has smaller dimensions (potentially nonzero determinant), (3) output Sqrt[Det[ma]]
(ii) Automatic (default): the same as (i) but using a faster algorithm which can be important for large multidimensional matrices and simulations
(iii) "Strong": (1) ma = mat.mat' or ma = mat'.mat whichever has smaller dimensions (potentially nonzero determinant), (2) output Sqrt[SquareMatrixNormedDet[ma]]. This method is not preferable since (a) the 2 x 2 case differs from the theoretical notion discussed below, (b) in fact, VolumeRatio[mat] and SquareMatrixNormedDet[mat] should gives similar outcomes for any square matrix, (c) in the standard case still 5% cases with a determinant out of range, (d) also the method of only using border matrices still gives 1% of cases out of the range.
Theory: The absolute value of the determinant of real vectors gives the volume of the parallelepipedum created by those vectors. Let f: Rn −> Rm be defined by matrix A, so that f[x] = A x, and let S be a subset of Rn, then volume[f[S]] = Sqrt[Det[A'.A]]∗ volume[S]. Hence Sqrt[Det[m]] gives a normalized volume ratio.
PM. If mat is a contingency table with nonnegative numbers, for nominal variables, then VolumeRatio gives a measure of association comparable to a correlation coefficient
PM. Since the squares destroy the sign it is recovered with PairwiseSign
PM. The routine uses InDet for Det
VolumeRatio[mat] gives OverallCorrelation[VolumeRatioMatrix[mat]] if mat is more−dimensional (MatrixQ[mat] is false). In that latter case the BordersOrAll option is passed on to VolumeRatioMatrix, and the Mode, Function and Check options passed on to OverallCorrelation. This call thus basically generates a number of association since it is not warranted that the matrix is positive semi−definite (PSD). Use NominalCorrelation[mat] for that

? PairwiseSign

PairwiseSign[mat] gives 1 or −1 as the sign of association between rows and columns. ForceSign −> (default {Automatic, False}) uses Sign[NonSquareDet[mat]]. If True then symbolic values other than 1 or −1 are set to 1. With value {False, True | False} mat is aggregated to a 2 by 2 matrix m, then Sign[Det[m]]. If the value is 1 then the sign is forced to be 1

? ForceSign

ForceSign is an option of PairwiseSign, see there

? VolumeRatioMatrix

VolumeRatioMatrix[mat] for a {n1, ..., nk} matrix gives a k by k assocation matrix, containing the VolumeRatio[m[ni, nj]] measures of association. There are three major ways to obtain that latter individual measure:
BordersOrAll −> BorderMatrices sums the other dimensions
BordersOrAll −> All (default) determines all individual m[p][ni, nj] matrices in the lower dimensions, determines their association, and (in default) gives the sum, weighted by total numbers in those submatrices
BordersOrAll −> Show shows that latter method
NB. The BordersOrAll −> All may be refined by non−default application by the option Inner −> Automatic | Min | Max | {f1, f2}, where the default is Automatic and uses {f1, f2} = VolumeRatioMatrix[Split], VolumeRatioMatrix[Average]} (weighted average), where choice Min uses {f1, f2} = {VolumeRatio, Min}, Max uses {VolumeRatio, Max}, and where f1 and f2 in general are functions that work on lists

? PairwiseMeasure

PairwiseMeasure is an option to indicate a measure for nominal variables that can be used to create a matrix of pairwise correlations. Examples are PairwiseMeasure −> VolumeRatio | CramersV

? PairwiseMeasureMatrix

PairwiseMeasureMatrix[mat] for a {n1, ..., nk} contingency table gives a k by k assocation matrix. Option PairwiseMeasure −> f gives the measure such that f[m[ni, nj]] are the measures of association, that then become the elements of the matrix. The default is VolumeRatio, and this routine is modelled after VolumeRatioMatrix (so that it should have the same outcomes). There are three major ways to obtain that latter individual measure:
BordersOrAll −> BorderMatrices sums the other dimensions
BordersOrAll −> All (default) determines all individual m[p][ni, nj] matrices in the lower dimensions, determines their association, and (in default) gives the sum, weighted by total numbers in those submatrices
BordersOrAll −> Show shows that latter method
NB. The BordersOrAll −> All may be refined by non−default application by the option Inner −> Automatic | Min | Max | {f1, f2}, where the default is Automatic and uses {f1, f2} = PairwiseMeasureMatrix[Split], PairwiseMeasureMatrix[Average]} (weighted average), where choice Min uses {f1, f2} = {VolumeRatio, Min}, Max uses {VolumeRatio, Max}, and where f1 and f2 in general are functions that work on lists
NB. UsedPairwiseMeasure −> {VolumeRatio, CramersV} determines the pairwise measures that are allowed for this routine. The reason for being so strict is that a pairwise measure must be properly defined, and you would not want to see lots of confusing output when you would enter a insufficiently defined measure, or just a typing error. Adjust this option only when you feel confident about it

? NominalVariance

NominalVariance[f] with f a list of frequency per category, uses default (
VarianceFreq[Border, #]&). Option NominalVariance −> ... also accepts
Variance, VarianceMLE, PearsonVariance, (VarianceFreq[Sum, #]&) and
VariancePr without message; puts out a message if another routine is used
NominalVariance[Table, {f1, ..., fn}] puts out all measures for lists f1, ..., fn
Option Replace −> ... replaces a zero outcome by
a symbolic expression, default "Var", but no replacement if False
Options "KnownVarMeasure" states the variance measures
that the routine tests upon (spelling errors)

? VariancePr

VariancePr[x_List, p_List] gives the variance
of a list of values x using the list of associated probabilities p
VariancePr[f] with p = f / Add[f] for frequencies f, assigns values 0, ...,
k−1 to the categories. For Length k = 2, this is the binomial variance.
See VarianceFreq if k > 2 and you want averages of pairwise variances
∗∗∗ AddedUsage by Cool`Probability`: ∗∗∗
VariancePr[q_Prospect] gives the variance

? VarianceFreq

VarianceFreq[method, f] for a list of frequencies f, neglects the influence of the
categories and just considers dispersion in those frequencies seen as orderless
numbers. Here n = Add[f], p = f / n, and k = Length[f]. See VariancePr when
the scores of the categories matter. See EffectiveNumberOfCategories as well.
VarianceFreq[Times, f] gives n (Times @@ p). For Length
k = 2, this is the binomial variance as well
VarianceFreq[Border, f] gives k^k (Times @@ p)
VarianceFreq[Sum, f] gives n p.(1 − p), the sum of pairwise variances
VarianceFreq[Average, f] gives the arithmetic average (n p.(1 − p)) / k
VarianceFreq[N, f] gives the weighted average (p . Sqrt[
x (1−p)])^2 where the pairwise standard deviations are weighted
VarianceFreq[Power, f] is (Times @@ f)^(1/k) / Average[f], thus the ratio between
the geometric and arithmetic average; its outcome lies between 0 and 1
VarianceFreq["Pearson", f] gives n (k ∗ p.p − 1); see the Chi2 package

? TabledBorderMatrix

TabledBorderMatrix[mat, {i, j}] decomposes the {i, j} border matrix into the submatrices
that cause its sum value. These submatrices are indicated with label Mat. PM
Check the same outcome as BorderMatrix by using (% // Add) /. Mat −> Identity

? NominalCorrelationReview

NominalCorrelationReview[mat] applies the various standardly defined volume ratio measures to mat and determines the overall correlation. Defined are: (1) Option BordersOrAll takes values BorderMatrices and All, with the latter split to the inner methods Automatic, Min, Max. (2) Option Mode has values MultipleRSquared or Det. (3) The overall correlations are calculated either with or without adjustment for PSD−ness. PM. This is only defined for higher dimensional matrices since a m x n matrix just has the VolumeRatio[mat] for which all these distinctions don't apply

Literature

Colignatus is the name of Thomas Cool in science.

Becker, L.A. (1999), "Measures of Effect Size (Strength of Association)", http://web.uccs.edu/lbecker/SPSS/glm_effectsize.htm, Retrieved from source

Cool, Th. (1999, 2001), "The Economics Pack, Applications for *Mathematica*", http://www.dataweb.nl/~cool, ISBN 90-804774-1-9, JEL-99-0820

Colignatus, Th. (2006), "On the sample distribution of the adjusted coefficient of determination (R2Adj) in OLS", http://library.wolfram.com/infocenter/MathSource/6269/

Colignatus, Th. (2007a), "A logic of exceptions", http://www.dataweb.nl/~cool, ISBN 978-90-804774-4-5

Colignatus, Th. (2007b), "Voting theory for democracy", 2nd edition, http://www.dataweb.nl/~cool, ISBN 978-90-804774-5-2

Colignatus, Th. (2007c), "A measure of association (correlation) in nominal data (contingency tables), using determinants", a earlier version of this paper (3rd publishable draft), http://ideas.repec.org/p/pra/mprapa/2662.html

Colignatus, Th. (2007d), "Correlation and regression in contingency tables. A measure of association or correlation in nominal data (contingency tables), using determinants", this paper, to be put at MPRA as well, as the improved version of Colignatus, Th. (2007c), but useful to mention in this list of references if only an abridged version of this paper is eventually published, http://mpra.ub.uni-muenchen.de/3394/

Colignatus, Th. (2007e), "Elementary statistics and causality", work in progress, http://www.dataweb.nl/~cool

Colignatus, Th. (2007f), "The 2 × 2 × 2 case in causality, of an effect, a cause and a confounder", http://mpra.ub.uni-muenchen.de/3614/, Retrieved from source

Colignatus, Th, (2007g), "A comparison of nominal regression and logistic regression for contingency tables, including the 2 × 2 × 2 case in causality", http://mpra.ub.uni-muenchen.de/3615/, Retrieved from source

Friendly, M. (2007), "Categorical Data Analysis with Graphics", Retrieved from http://www.math.yorku.ca/SCS/Courses/grcat/grc6.html (citing the data from Koch & Stokes (1991))

Garson, D. (2007), "Nominal Association: Phi, Contingency Coefficient, Tschuprow's T, Cramer's V, Lambda, Uncertainty Coefficient", http://www2.chass.ncsu.edu/garson/pa765/assocnominal.htm, Retrieved from source

Higham, N. J. (1989), "Matrix nearness problems and applications". In M. J. C. Gover and S. Barnett (eds), "Applications of Matrix Theory", pages 1–27. Oxford University Press

Johnston J. (1972), "Econometric methods", McGraw-Hill

Kleinbaum, D.G., K.M. Sullivan and N.D. Barker (2003), "ActivEpi Companion texbook", Springer

Linacre J. M, (2005), "Correlation Coefficients: Describing Relationships", Rasch Measurement Transactions, 19:3 p. 1028-9, retrieved from http://www.rasch.org/rmt/rmt193c.htm (citing the data from Uebersax (2000))

Losh, S.C. (2004), "Guide 5: Bivariate Associations and Correlation Coefficient Properties", http://edf5400-01.fa04.fsu.edu/Guide5.html, Retrieved from Source

Losh, S.C. (2004a), "Guide 6: Multivariate Crosstabulations and Causal Issues", http://edf5400-01.fa04.fsu.edu/Guide6.html, Retrieved from Source

Mood, A.M. and F.A. Graybill (1963), "Introduction to the theory of statistics", McGraw-Hill

Pearl, J. (2000), “Causality. Models, reasoning and inference”, Cambridge

Schield, M. (1999), “Simpson’s paradox and Cornfield’s conditions”, Augsburg College ASA-JSM

Simon, R. (2007), “Lecture Notes and Exercises 2006/07”, http://www.maths.lse.ac.uk/Courses/MA201/, Retrieved from source

Takayama A. (1974), “Mathematical economics”, The Dryden Press

Theil H. (1971), “Principles of econometrics”, North-Holland

UCLA ATS (2007), “SAS Textbook Examples. Econometric Analysis, Fourth Edition by Greene. Chapter 16: Simultaneous Equations Models”, http://www.ats.ucla.edu/stat/SAS/examples/greene/chapter16.htm, Retrieved from source

(Other) websites

http://en.wikipedia.org/wiki/Contingency_table

http://post.queensu.ca:8080/SASDoc/getDoc/en/procstat.hlp/corr_sect26.htm

http://en.wikipedia.org/wiki/Fisher%27s_exact_test