On Looking at Large Correlation Matrices
Author(s): M. Hills
Source: *Biometrika*, Vol. 56, No. 2 (Aug., 1969), pp. 249-253
Published by: Oxford University Press on behalf of Biometrika Trust
Stable URL: http://www.jstor.org/stable/2334418
Accessed: 18-01-2016 04:25 UTC

*Biometrika* (1969), **56**, 2, *p.* 249
*With* 3 *text-figures*
*Printed in Great Britain*

# On looking at large correlation matrices

By M. HILLS
*London School of Hygiene and Tropical Medicine*

## Summary

Two graphical techniques, familiar in other contexts, are applied to a correlation matrix. The method of half-normal plotting is used to determine which coefficients are numerically too large to have come from zero population values. A visual clustering method is used to select clusters of variables which have high positive correlations with each other.

The first and sometimes only impression gained from looking at a large correlation matrix is its largeness. This note describes the application of two graphical techniques to the problem of spotting some structure in the matrix. They are illustrated by a matrix taken from a study of the physiological effects of examination strain on 48 medical students. The variables of interest were differences in levels between a normal period of time and the period of the examination. Differences were obtained for 13 physiological measurements including blood pressure, pulse rate and various substances secreted in the urine and blood. The matrix of correlations between the 13 differences based on 48 subjects is shown in Table 1.

The first technique is an aid to answering the question: Which of these coefficients is large enough, either positive or negative, to be considered as arising from a population coefficient different from zero? It is convenient to use the $z$ transform of the correlations, $r$, given by

$$z = \tfrac{1}{2}\log\{(1+r)/(1-r)\}.$$

The standard error of any $z$ is then approximately $1/\sqrt{45}$ which is $1/6{\cdot}708 = 0{\cdot}1491$. The $\frac{1}{2}\times 13\times 12 = 78$ values of $z$ in half the symmetric matrix are not statistically independent, but if this fact is ignored the values of $z$ may be plotted in a half-normal plot to see which are too large numerically to have come from a random sample from a normal distribution with mean zero and standard deviation 0·1491. The half-normal plot is familiar in the analysis of factorial experiments (Daniel, 1959) and more recently has been applied to multidimensional contingency tables (Cox & Lauh, 1966), an application which suggested that it might also be useful with correlation matrices.

The plot is very simple to do using tables of the transformations $r\to z$ and

$$G(c) = \frac{1}{\sqrt{(2\pi)}}\int_{-c}^{c} e^{-\frac{1}{2}u^2}\,du \to c.$$

The method is to transform the correlations using the $z$ transform and to rank the $z$'s (ignoring sign) as $|z_1| \leqslant |z_2| \leqslant \ldots \leqslant |z_{78}|$. Then the proportion of $z$'s numerically less than $|z_i|$ is approximately $(i-\frac{1}{2})/78$. The $\frac{1}{2}$ in $(i-\frac{1}{2})/78$ is to avoid 0 and 1 at the ends. If the $z$'s have the distribution $N(0,\sigma^2)$ then each $z_i$ should be approximately equal to $\sigma z_i'$, where $G(z_i') = (i-\frac{1}{2})/78$. Thus if $|z_i|$ is plotted against $z_i'$ the points should lie about a straight line through the origin with slope $\sigma$.

Figure 1 shows the plot for the correlations in Table 1. The points lie very closely about a straight line up to $|z| = 0{\cdot}3$ after which they deviate considerably. The slope of this line differs from the theoretical value of 0·1491 because the 78 $z$'s are clearly not a random sample from $N(0, 0{\cdot}1491^2)$. When the 62 values of $|z|$ up to 0·3 are re-plotted using $(i-\frac{1}{2})/62$ in place of $(i-\frac{1}{2})/78$, this discrepancy disappears; see Fig. 2.

The two figures provide a striking visual demonstration of the fact that correlations with $z$ values numerically larger than 0·3 are too large for it to be reasonable to assume that they have come from zero population values. The value 0·3 is actually very close to the $z$ point at which a correlation coefficient would be declared significantly different from zero at the 5 % level $(0+1{\cdot}96\times 0{\cdot}1491 = 0{\cdot}29)$ but this will not always be the case.

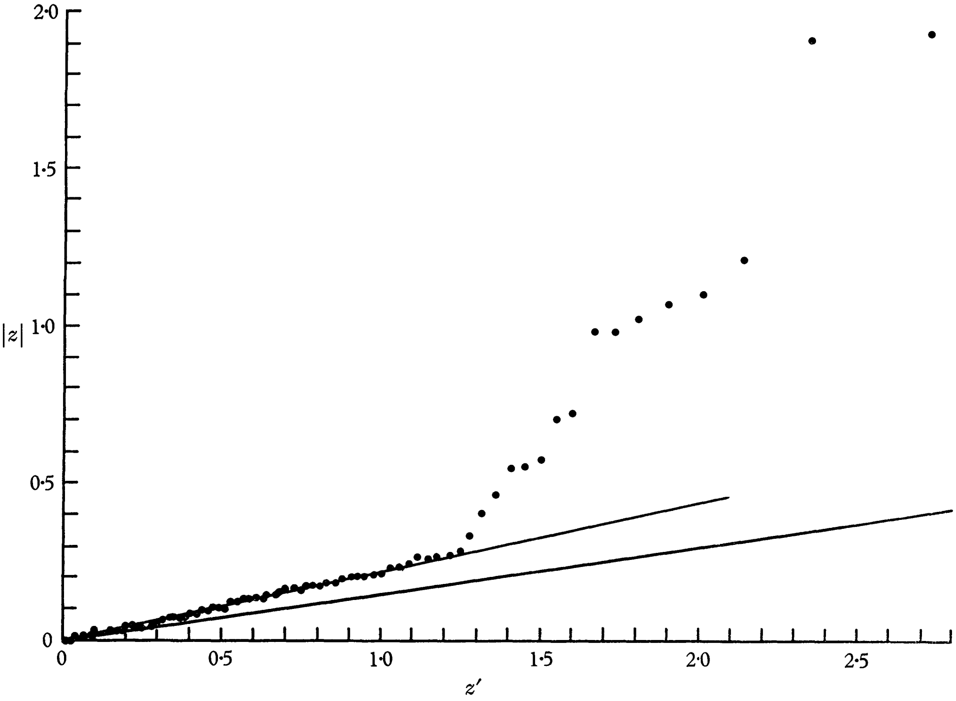

Fig. 1. Half-normal plot of all 78 correlation coefficients. Thicker line has theoretical slope, 0·1491.

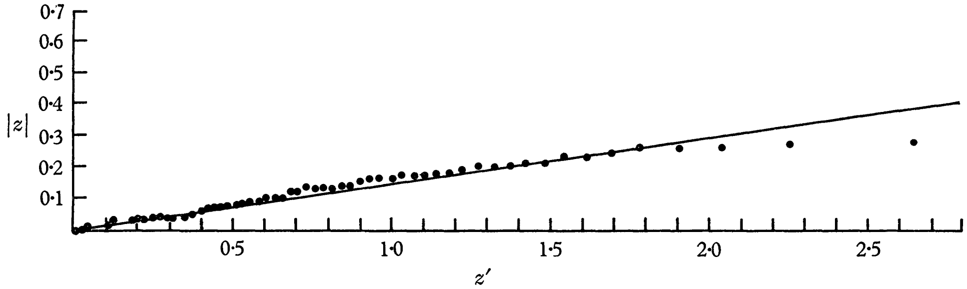

Fig. 2. Half-normal plot omitting the 16 numerically largest correlation coefficients. Line has theoretical slope, 0·1491.

The effect of the dependence between the correlation coefficients in Table 1 is difficult to predict. The half normal plot is a way of detecting contamination in either tail of a normal distribution so the interesting question is whether the dependence causes one to ignore or suspect contamination incorrectly. Cox & Lauh (1966) suggested that, in the null case with

Table 1. *Matrix of correlations between differences in 13 physiological measurements. The differences are between normal values for the subject and values at the time of the examination*

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SYS | DIA | P.P. | PUL | CORT | U.V. | TOT/100 | ADR/100 | NOR/100 | ADR/TOT | TOT/HR | ADR/HR | NOR/HR |
| 1·00 | 0·60 | 0·80 | 0·20 | 0·12 | −0·20 | 0·07 | 0·32 | −0·03 | −0·03 | −0·09 | 0·14 | −0·13 |
| — | 1·00 | ·00 | ·28 | ·13 | − ·22 | ·07 | ·19 | ·01 | − ·05 | ·04 | ·10 | ·01 |
| — | — | 1·00 | ·04 | ·06 | − ·09 | ·04 | ·26 | − ·04 | ·00 | − ·14 | ·10 | − ·17 |
| — | — | — | 1·00 | ·15 | − ·16 | − ·07 | ·20 | − ·13 | − ·12 | − ·16 | ·08 | − ·18 |
| — | — | — | — | 1·00 | − ·22 | − ·01 | ·07 | − ·03 | − ·01 | − ·17 | − ·03 | − ·16 |
| — | — | — | — | — | 1·00 | − ·13 | − ·17 | − ·08 | ·04 | ·24 | ·20 | ·18 |
| — | — | — | — | — | — | 1·00 | ·19 | ·96 | − ·38 | ·77 | − ·03 | ·75 |
| — | — | — | — | — | — | — | 1·00 | − ·10 | ·43 | − ·04 | ·75 | − ·25 |
| — | — | — | — | — | — | — | — | 1·00 | − ·51 | ·79 | − ·25 | ·84 |
| — | — | — | — | — | — | — | — | — | 1·00 | − ·49 | ·50 | − ·62 |
| — | — | — | — | — | — | — | — | — | — | 1·00 | ·03 | ·96 |
| — | — | — | — | — | — | — | — | — | — | — | 1·00 | − ·26 |
| — | — | — | — | — | — | — | — | — | — | — | — | 1·00 |

dependence arising from the non-orthogonal factorial contrasts in a multidimensional contingency table, the half-normal plot of the contrasts was still close enough to a straight line with slope $\sigma$ for the dependence to be ignored. Bearing in mind that the method is graphical and in no way 'exact', it seems unlikely that its useful properties are disturbed by the sort of dependence found in a correlation matrix.

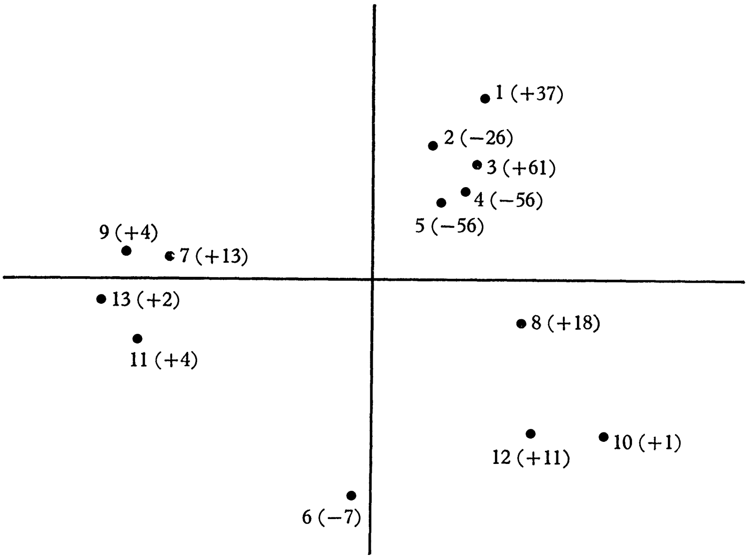

Fig. 3. Representation of the correlation matrix in two dimensions. Numbers in brackets are the co-ordinates in a third dimension.

The next question which was asked was 'Are there any groups of variables such that members of the same group are all fairly highly positively correlated with each other and behave similarly in their correlations with other variables?' The graphical technique which was used is a simple application of a result described by Gower (1966), in which the variables are represented by points, $P_i$, in $m$-dimensional Euclidean space, $E_m$. The points are chosen so that the distances, $P_iP_j$, are as close as possible to $2(1-r_{ij})$, where $r_{ij}$ is the correlation between the $i$th and $j$th variables. Thus if $P_iP_j$ is small then $r_{ij}$ must be close to 1 and the largest value of $P_iP_j$ occurs when $r_{ij} = -1$. The co-ordinates of the points in $E_m$ are given by the elements in the first $m$ eigenvectors of the matrix with $(i,j)$th element equal to $(r_{ij}-\bar{r}_{i.}-\bar{r}_{.j}+\bar{r}_{..})$.

The results for $E_2$ are shown in Fig. 3. There are three main clusters plus variable 6 which seems remote from the others. A measure of the lack of distortion in pushing the data into $E_2$ is the sum of the first two eigenvalues expressed as a percentage of the trace of the matrix and was 53·43 %. This is rather low and suggests that one should also look at the third dimension. It is sufficient for our purpose to add the third co-ordinate in brackets at the side of each point in Fig. 3. It is now clear that the upper right-hand cluster includes some very divergent points. The two clusters selected were therefore

(1) Variables 7, 9, 13, 11;

(2) Variables 8, 12, 10.

These make quite good physiological sense. The variables in the clusters were:

7. Total catecholomines (= adrenaline + noradrenaline) secreted in 100 ml of urine (TOT/100);

9. Noradrenaline secreted in 100 ml. urine (NOR/100);
11. Total catecholomines secreted in 1 hr (TOT/HR);
13. Noradrenaline secreted in 1 hr (NOR/HR);
8. Adrenaline secreted in 100 ml of urine (ADR/100);
10. Adrenaline as a percentage of total catecholomines (ADR/TOT);
12. Adrenaline secreted in 1 hr (ADR/HR).

This suggests that the change in total catecholomines is primarily due to the change in noradrenaline and that the same information is obtained from measurements per 100 ml. of urine as per hour.

Table 2. *Correlation coefficients within the two clusters picked from Fig.* 3

| Cluster 1 | | | | Cluster 2 | | |
|---|---|---|---|---|---|---|
| 7 | 9 | 11 | 13 | 8 | 10 | 12 |
| TOT/100 | NOR/100 | TOT/HR | NOR/HR | ADR/100 | ADR/TOT | ADR/HR |
| 1·000 | 0·957 | 0·771 | 0·754 | 1·000 | 0·429 | 0·753 |
| — | 1·000 | 0·791 | 0·836 | — | 1·000 | 0·501 |
| — | — | 1·000 | 0·959 | — | — | 1·000 |

Table 3. *Correlation coefficients including only one varlable from each cluster*

| 1 | 2 | 3 | 4 | 5 | 6 | 8 | 9 |
|---|---|---|---|---|---|---|---|
| SYS | DIA | P.P. | PUL | CORT | U.V. | ADR/100 | NOR/100 |
| 1·000 | 0·602 | 0·800 | 0·197 | 0·123 | −0·123 | 0·322 | −0·026 |
| — | 1·000 | ·002 | ·277 | ·128 | − ·224 | ·189 | ·012 |
| — | — | 1·000 | ·039 | ·058 | − ·087 | ·261 | − ·042 |
| — | — | — | 1·000 | ·145 | − ·161 | ·198 | − ·131 |
| — | — | — | — | 1·000 | − ·222 | ·073 | − ·034 |
| — | — | — | — | — | 1·000 | − ·172 | − ·077 |
| — | — | — | — | — | — | 1·000 | − ·098 |
| — | — | — | — | — | — | — | 1·000 |

The correlation matrices for the two clusters are shown in Table 2. If only variable 9 from the first cluster and variable 8 from the second are included in the whole correlation matrix, it begins to look more like a diagonal matrix; see Table 3. A half-normal plot verifies that apart from the correlations of variable 1 with variables 2 and 3, the others may all be taken as coming from zero population values.

I am grateful to Professor D. R. Cox for some helpful comments about the half-normal plot, and to Dr P. Bridgeman for permission to use the data.

## References

Cox, D. R. & Lauh, E. (1966). A note on the graphical analysis of multidimensional contingency tables. *Technometrics* **9**, 481–8.

Daniel, C. (1959). Use of half-normal plots in interpreting factorial two-level experiments. *Technometrics* **1**, 311–41.

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* **53**, 325–38.

[*Received December* 1968. *Revised January* 1969]