Tina Memo No. 2016-014
Internal.

De-correlating a Pair of Variables via Linear Combination.

Neil Thacker and Scott Notley.

Last updated
10 / 8 / 2016

Imaging Science and Biomedical Engineering Division,
Medical School, University of Manchester,
Stopford Building, Oxford Road,
Manchester, M13 9PT.

De-correlating a Pair of Variables via Linear Combination.

N.A.Thacker, S. Notley. 10/9/2016

Abstract

We have observed that many clinical studies ofen generate alternative summary variables from imaging data which are then analysed for their potential for group discrimination or detection of treatment response. Such data has complementary information and this is lost when simply selecting the 'best'. However, when trying to assess the use of joint variables, correlation between noise distributions (non-independance) complicates the construction of statistical similarity scores and hypothesis tests.

Decorrelation of variables is achieved by a rotation to the eigen-vectors of the error covariance matrix, which normally involves a matrix inverse. This document derives 2D solutions for the direct construction of convenient independent variables and so avoids the calculation of a matrix inverse and eigen vectors. The derivations are corroborated using results for simulated distributions. We also show that these insights do not extend to 3D or beyond. However, even in 2D these ideas have potential for application in clinical studies as a simple way to improve the efficiency of null hypothesis tests via variable combination.

Introduction

When attempting to summarise the main characteristics of a variable distribution there is a wide variety of possible choices, exemplified by the use of medians, percentiles and inter-quartile ranges. Such variables constructed following clinical studies are often treated as alternatives, when in fact they may contain complementary information. The main obstacle to extracting this information is the correlation between measurements which prevents simple use, for example in T-tests or Chi-square calculations.

We will show below that the effects of correlation can be easily removed via a process of linear combination. By this we mean that the ususal complications involved in calculation of eigen values, eigen vectors and a matrix inverse, can be avoided. There are a variety of ways of doing this which have different advantages depending upon the circumstances. In the most general case, a variable can be constructed which not only combines the information in two variables but can be optimised to maximise its use in discrimination between groups. Whilst solutions for the use of more variables is mathematically tractable (rotating along the eigen vectors), we show that these simplifications are specific to 2D, and do not extend beyond.

We assume that we start from a position where the noise characteristics of our variables have already been determined, in the form of the covariance matrix A. This can be done from repeatability data, error propagation or an analysis of variation. Also we assume that they have been appropriately defined (perhaps following a Bland-Altman assessment and non-linear transformation) to have homogenous noise.

Correlations in 2 Dimensions

Defining a general matrix A and eigen vector v as

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \qquad and \qquad v = (x, y)^T$$

Then from

$$(\lambda I - A)v = 0$$

we deduce

$$y = \frac{\lambda - a}{b} x = \frac{c}{\lambda - d} x \qquad (1)$$

so that

$$v = \frac{(b, \lambda - a)^T}{|(b, \lambda - a)|} \qquad or \qquad v = \frac{(\lambda - d, c)^T}{|(\lambda - d, c)|} \qquad (2)$$

Therefore

$$\frac{\lambda - a}{b} \;=\; \frac{c}{\lambda - d} \qquad (3)$$

rearranging and solving for λ[1]

$$bc \;=\; (\lambda - a)(\lambda - d) \;=\; \lambda^2 - (a+d)\lambda + ad$$

$$\lambda_{\pm} \;=\; \frac{(a+d) \pm \sqrt{(a+d)^2 \;-\; 4(ad - bc)}}{2}$$

Defining now the correlation matrix C for the Gaussian noise process associated with a two dimensional vector $(m,\ n)^T$, $a \;=\; var(m)$, $d \;=\; var(n)$ and observing that for a covariance matrix $b \;=\; c$ we have

$$C \;=\; \begin{bmatrix} \frac{a}{\sqrt{a^2}} & \frac{b}{\sqrt{ad}} \\ \frac{c}{\sqrt{ad}} & \frac{d}{\sqrt{d^2}} \end{bmatrix} \;=\; \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

where ρ is the correlation factor for m and n.

So that the eigen values are

$$\lambda_{\pm} \;=\; \frac{(2) \pm \sqrt{(2)^2 \;-\; 4(1-\rho^2)}}{2} \;=\; 1 \pm \rho$$

and its eigen vectors are

$$v_+ \;=\; (\rho,\ \rho)^T / |(\rho,\ \rho)| \;=\; (1/\sqrt{2},\ 1/\sqrt{2})^T$$

$$v_- \;=\; (\rho, -\rho)^T / |(\rho, -\rho)| \;=\; (1/\sqrt{2}, -1/\sqrt{2})^T$$

i.e. **the eigenvectors of a 2D correlation matrix are independent of ρ and always lie along the diagonals.**

We can therefore rotate the original correlated variables (m, n) to generate two uncorrelated ones (m', n'), by first scaling to their standard deviations and then re-projecting along the eigen vectors of the correlation matrix

$$m' \;=\; \frac{m}{\sqrt{2a}} \;+\; \frac{n}{\sqrt{2d}} \qquad (4)$$

and

$$n' \;=\; \frac{m}{\sqrt{2a}} \;-\; \frac{n}{\sqrt{2d}} \qquad (5)$$

where m' and n' have a Gaussian noise process described by the covariance matrix

$$\begin{bmatrix} \lambda_+ & 0 \\ 0 & \lambda_- \end{bmatrix} \;=\; \begin{bmatrix} (1 \;+\; \rho) & 0 \\ 0 & (1 \;-\; \rho) \end{bmatrix}$$

For situations where the original correlation matrix is determined from repeated samples, the variances λ_+ and λ_- can be estimated directly from corresponding repeated estimates of m' and n', making the calculation of ρ (or λ_+, λ_-) unnecessary. This gives us a completely deterministic process which can be applied to two variable which is guaranteed to output two uncorrrelated variables, regardless of the level of correlation. Note that for variables which are already uncorrelated it would be unnecessary to apply this process, but equally rotating homogenous uncorrelated variables does not cause any harm (beyond redefining the variables) as it does not introduce correlations. This observation allows us to define alternative (rotated) variables which may be more convenient (see below).

Maintaining a Primary Variable in 2D

Rotating our initial variables might lead to complication when trying to interpret data, it may be more convenient to maintain one of our variables (e.g. m) and construct a new variable by correcting the second variable (n) for correlation (n'').

[1] This same result is more easily obtained by solving for λ starting with the identities $Tr(A) \;=\; a + b \;=\; \lambda_- + \lambda_+$ and $Det(A) \;=\; ad - bc \;=\; \lambda_- \lambda +$. Then

$$\lambda_{\pm} \;=\; \frac{Tr(A) \pm \sqrt{Tr(A)^2 \;-\; 4\ Det(A)}}{2}$$

but the above approach also makes explicit the eigen vectors.

Given two independent and homogenous variables, g and h, derived from m and n (as above)

$$g = \frac{m/\sigma_m + n/\sigma_n}{\sqrt{2\lambda_+}} \qquad h = \frac{m/\sigma_m \; - \; n/\sigma_n}{\sqrt{2\lambda_-}}$$

where $\sigma_m \; = \; \sqrt{a}$ and $\sigma_n \; = \; \sqrt{d}$. As the distribution of g vs h is isotropic with unit variance in all directions, and rotation of this data can be selected as a valid reprojection. We require a rotated variable which aligns with the first variable m, i.e.

$$g' \; = \; g \; sin \; \phi \; + \; h \; cos \; \phi \; = \; f \; m \qquad (6)$$

For this to be true the contribution from n to equation (6) must be zero, so

$$\frac{sin \; \phi}{\sigma_n \sqrt{\lambda_+}} \; - \; \frac{cos \; \phi}{\sigma_n \sqrt{\lambda_-}} \; = \; 0$$

We deduce that

$$\phi = atan(\sqrt{\lambda_+/\lambda_-}) \; = \; atan(\sqrt{(1+\rho)/(1-\rho)})$$

and therefore

$$cos \; \phi \; = \; \frac{\sqrt{1-\rho}}{\sqrt{2}} \qquad sin \; \phi \; = \; \frac{\sqrt{1+\rho}}{\sqrt{2}}$$

substituting into (14) we find

$$f \; = \; \frac{1}{\sigma_m}$$

A second independent variable can now be constructed perpendicular to m, i.e.

$$h' \; = \; g \; cos \; \phi \; - \; h \; sin \; \phi \qquad (7)$$

This is given by

$$\frac{\sqrt{\lambda_-}(m/\sigma_m + n/\sigma_n)}{2\sqrt{\lambda_+}} - \frac{\sqrt{\lambda_+}(m/\sigma_m - n/\sigma_n)}{2\sqrt{\lambda_-}}$$

$$= \; \frac{n}{\sigma_n} \frac{1}{\sqrt{1-\rho^2}} \; - \; \frac{m}{\sigma_m} \frac{\rho}{(\sqrt{1-\rho^2})}$$

dividing both variables ((6) and (7)) by f we have two independent variables

$$(m, n'') \; = \; (\; m \; , \; n \; \frac{\sigma_m}{\sigma_n \sqrt{(1-\rho^2)}} \; - \; m \; \frac{\rho}{\sqrt{(1-\rho^2)}} \;) \qquad (8)$$

i.e. **it is possible to correct a second variable for correlation with the first via a weighted subtraction.**

As this is a rotation of the homogenous g, h variables this new space must also have homogenous noise characteristics. Given that the first variable has a variance of $\sigma_m^2 \; = \; a$, this must also be true of the second variable.

Correlations in 3 Dimensions

We show here that the useful diagonal form of the eigen vectors of a 2D correlation matrix does not extend to 3D and beyond. For a 3D correlation matrix

$$C \; = \; \begin{bmatrix} 1 & \rho_1 & \rho_3 \\ \rho_1 & 1 & \rho_2 \\ \rho_3 & \rho_2 & 1 \end{bmatrix} \qquad and \qquad v \; = \; (x, y, z)^T$$

Then

$$(1-\lambda)x \; + \; \rho_1 y \; + \; \rho_3 z \; = \; 0 \qquad (9)$$

$$\rho_1 x \; + \; (1-\lambda)y \; + \; \rho_2 x \; = \; 0 \qquad (10)$$

$$\rho_3 x \; + \; \rho_2 y \; + \; (1-\lambda)z \; = \; 0 \qquad (11)$$

From (11)

$$z \; = \; \frac{\rho_3 x + \rho_2 y}{\lambda - 1} \qquad (12)$$

From (9)

$$y \;=\; \frac{[(1-\lambda)^2 \;-\; \rho_3^2]x}{\rho_1(\lambda-1) \;+\; \rho_2\rho_3} \qquad (13)$$

From (10)

$$y \;=\; \frac{[\rho_1(\lambda-1) \;+\; \rho_3\rho_2]x}{(1-\lambda)^2 \;-\; \rho_2^2} \qquad (14)$$

Equating (13) and (14), putting $(1-\lambda) \;=\; s$, and rearranging

$$s^3 \;-\; s[\rho_1^2 \;+\; \rho_2^2 \;+\; \rho_3^2] \;-\; 2\rho_1\rho_2\rho_3 \;=\; 0$$

Now putting $A \;=\; \rho_1^2 \;+\; \rho_2^2 \;+\; \rho_3^2$ and $B \;=\; 2\rho_1\rho_2\rho_3$, gives the depressed cubic

$$s^3 \;- As \;-\; B \;=\; 0 \qquad (15)$$

Making the usual association with $4cos^3\theta \;-\; 3cos\theta - cos3\theta \;=\; 0$, and substituting $s = 2\sqrt{A/3}\; cos(\theta)$ into (15), we get

$$4cos^3\theta \;-\; 3cos\theta - \frac{3B}{2A}\sqrt{\frac{3}{A}} \;=\; 0$$

i.e.

$$cos(3\theta) \;=\; \frac{3B}{2A}\sqrt{\frac{3}{A}} \quad \rightarrow \quad \theta(A,B) \;=\; acos(\frac{3B}{2A}\sqrt{\frac{3}{A}}) \;/\; 3$$

now defining

$$s(A,B) \;=\; 2\sqrt{A/3}\;\; cos[\theta(A,B)] \qquad (16)$$

the three solutions are obtained from (16) using

$$s_0 \;=\; s(A,B), \quad s_1 \;=\; -\,s(A,-B), \quad s_1 \;=\; -\,s_0 - s_2$$

The unit eigen vectors can now be constructed by computing $\lambda_i \;=\; 1 - s_i$, substituting back into equations (9) and (10) (or (11)) and re-normalising.

In n dimensions, solution for the eigen values will involve solution of an nth order polynomial. In 3D the solution already has two degrees of freedom, so unlike the 2D case, the eigen vectors can point in any direction and therefore are not forced to align with the diagonals of the data space.

Monté-Carlo Simulation

As an illustration of these ideas, the decorrelation method (equation (8) above), was tested on median and percenile values for a Gaussian random variable generated by Monté-Carlo. Simulated data was produced using Matlab 2014a. 500 realisations of a Gaussian distributed signal were generated, each of 1000 samples. For each realisation the correlation co-efficients between the median of the signal distribution and each percentile was calculated. figure 1 shows the average correlation co-efficient (ρ) between the median and percentiles, over 100 runs. The symmetry of the correlation coefficient implies that a difference of symmetrical percentiles (for example the inter-quartile range) will be independant of the median.

Figure 2a shows scatter plots of the 70th percentile vs the median values. Similarly, figure 3a shows scatter plots of the 95th percentile vs the median values. In both cases the correlation between the variables may be observed. Figure 2b and figure 3b show the median vs the de-correlated 70th and 95th percentiles respectively. In both cases the estimated correlation co-efficients for the new variables we found to be zero (to the limts of available numerical precision). These results corroborate the preceeding mathematical theory.

Conclusions

Many research projects generate correlated measurement variables. Sometimes these correlations are inherrent to data, other times they are introduced by poor methodology (such as use of cumulative distributions). The conventional method for combining two independent measurements **of the same thing** would weight estimates by their variance and not the standard deviation, but this is only applicable when we are trying to improve an

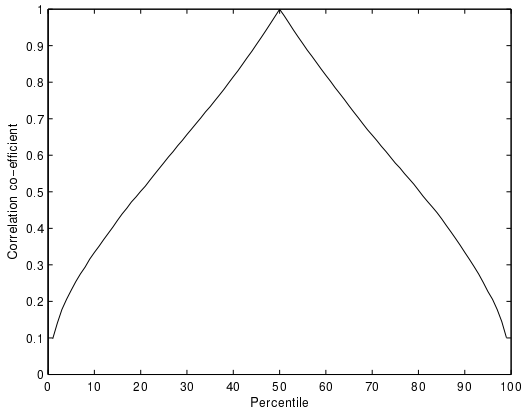

Figure 1: Correlations co-efficients between the median and percentiles

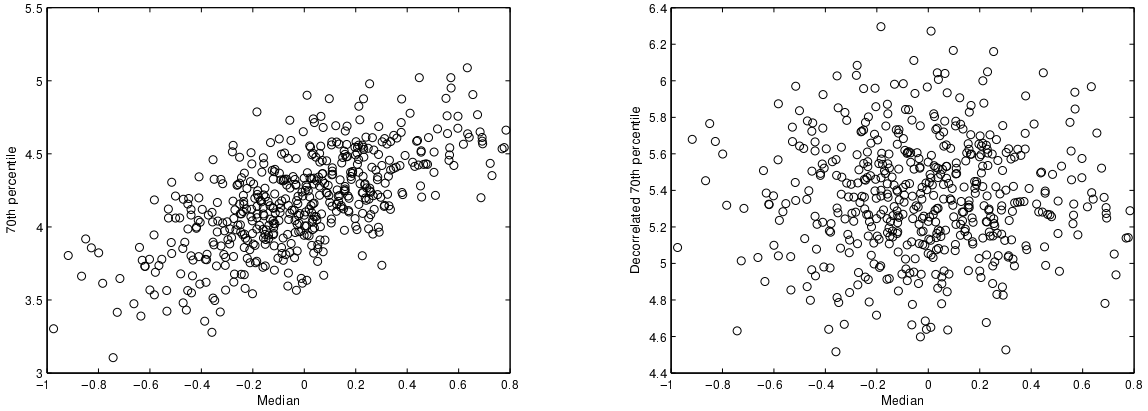

Figure 2: Scatter plots showing the correlations between the median and (a) the 70th percentile (correlation co-efficient 0.6) & (b) the de-correlated 70th percentile (zero correlation to within error).

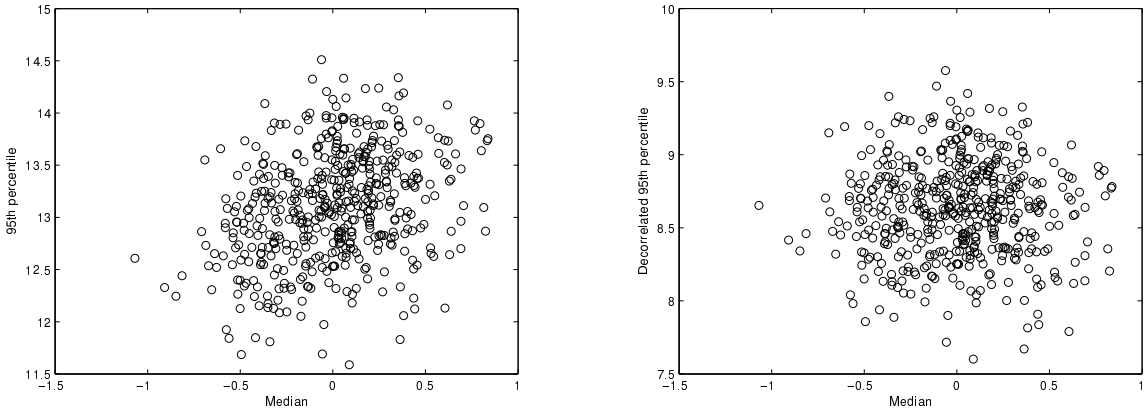

Figure 3: Scatter plots showing the correlations between the median and (a) the 95th percentile (correlation co-efficient 0.31) & (b) the de-correlated 95th percentile (zero correlation to within error).

estimate of a single quantity. The intended application here is to construct two variables with independent noise processes and these variables may represent different forms of information.

As a consequence of our analysis we can say that a statistical difference (chi-square) between two measurements $\Delta m, \Delta n$ can be expressed in four equivalent ways

$$\begin{aligned}
\chi^2 &= \frac{(\Delta m')^2}{\lambda_+^2} + \frac{(\Delta n')^2}{\lambda_-^2} \\
&= \frac{(\Delta m)^2}{\sigma_m^2} + \frac{(\Delta n'')^2}{\sigma_m^2} \\
&= \Delta g'(\phi)^2 + \Delta h'(\phi)^2 \\
&= (\Delta m, \Delta n)^T C^{-1} (\Delta m, \Delta n)
\end{aligned}$$

Each of these have potential for utility depending upon the application.

The variables (m', n') need neither knowledge of ρ, the eigen vectors nor a matrix inverse (C^{-1}), simplifying their use in spreadsheet calculations. The variables (m, n'') require additional knowledge of ρ, but does not need the eigen values and has the advantage of maintaining one of the variables in it's original form (e.g. to investigate the information added to a chi-square distance metric from a second variable). The variables g' and h' (equations (6) and (7)) are the more general form, and can be optimised over ϕ in order to find the combination of the original variables which gives the best discriminative power in a null hypothesis test.

We have shown however that, the mathematical property of a correlation matrix which makes these new variables possible (diagonality of the eigen vectors) does not apply beyond two dimensions.