multivariate OLS - comments on correlation

Steven Brawer
12/2017

introduction

This document sketches two aspects of correlation.

The first part illustrates a method of creating new features as a linear combination of existing features, applicable in the case that the original features are highly correlated. One advantage of doinbg this is that the estimated variance in slope of the most important of the most significant new feature is smaller than the variance of the original slopes. In addition, the problem is reduced in that, in terms of the new features, only one feature is significant.

The second treats the case where all correlation coefficients are small. Then the OLS problem is solved to lowest order in the correlation by doing an analytic expansion of the inverse of the correlation matrix. This may be useful by showing in detail how correlation works for OLS.

The treatment here is not exhaustive, but it should be useful as a start, and provides insight into the interpretation of OLS.

notation

The notation is the same as in the document *OLS in two dimensions - correlation.pdf,* referred to as **OLS2**.

strong correlation

We have **m** features and **N** samples. Most correlation coefficients among the features are assumed to be large - greater than 0.7. Perhaps a few correlation coefficients can be as low as 0.5, but most should be larger.

This could be considered as part of an approach where one creates a linear combination of those features which have high mutual correlation. Defining a correlation cutoff ρ_c, the features i,j would both be included if $|\rho_{i,j}| \geq \rho_c$.

In this document, we simply assume all features are sufficiently strongly correlated, to keep things simple.

Assume all features x^i are standardized. That is, they have exactly mean 0 and variance 1 and the values are the same in all ensembles. The output vector is not standardized. Data is generated from

$$y_\mu = \sum_{i=0}^{m-1} b_i x_\mu^i + \varepsilon_\mu \qquad (1)$$

where i labels the feature and μ labels the sample. It is also assumed that all the slopes b_i are comparable in magnitude (otherwise this procedure may not make sense). Note that the **magnitude of the slopes is relative to standardized features**. Later we show how to combine the true slope b with correlation to give a more useful criterion for using the cutoff.

In this document we will mainly use vector notation, so eq (1) is

$$y = \sum_{i=0}^{m-1} b_i x^i + \varepsilon \tag{2}$$

where the residual ε is an N-dimensional vector, with mean 0 and variance σ^2 *in the ensemble average only*. That is, in a given ensemble, the ε are not standardized (in order that they be independent). In addition, the residual for each sample is independent of the residual in a different sample (in the ensemble average). See OLS2.

Because the variance of each feature is exactly 1, and the mean is exactly 0, the correlation coefficient between features i and j is

$$\rho_{i,j} =< x^i x^j > \tag{3}$$

where for two N-dimensional vectors u and v we define <...>

$$< uv >= \frac{1}{N} \sum_{\mu=1}^{N} u_i v_i \tag{4}$$

change of variable

Define

$$A_k \equiv \sum_{j=0}^{m-1} \rho_{k,j} \tag{5}$$

We switch to variables $p^0, p^1, p^2, ..., p^{m-1}$ where p^0 is normal to all the other p's, but they are not normal to each other. That is

$$\begin{aligned} &< p^0 p^j >= 0 \quad j = 1,2,..,m-1 \\ &< p^i p^j > \neq 0 \quad i, j \neq 0 \end{aligned} \tag{6}$$

Generally the non-zero correlations, while not 0, are small (see the numerical example below). Now **assume m is odd**. Define

$$\begin{aligned} &\theta_k \equiv \frac{2\pi k}{m} \quad k = 0,1,...,\frac{m-1}{2} \\ &p^{2k} = \sum_{i=0}^{m-1} \frac{x^i}{A_i} \cos(i\theta_k) \quad k = 1,2,,...,\frac{m-1}{2} \\ &p^{2k-1} = \sum_{i=0}^{m-1} \frac{x^i}{A_i} \sin(i\theta_k) \quad k = 1,2,...,\frac{m-1}{2} \end{aligned} \tag{7}$$

and, **for the most important new feature,**

$$p^0 = \sum_i x^i \tag{8}$$

So each p is linear in x, and is an N-dimensional vector which is independent of ensemble. It is easy to see that the first of eqs (6) holds. Letting

$$t_{k,i} = one\ of\ \sin(i\theta_k)\ or\ \cos(i\theta_k) \tag{9}$$

as appropriate, we see that, for k > 0,

$$\begin{aligned} < p^0 p^k > &= \sum_i \sum_j < x^i x^j > t_{k,j} \frac{1}{A_j} \\ &= \sum_j \left(\frac{\sum_i \rho_{j,i}}{A_j} \right) t_{k,j} \\ &= \sum_j t_{k,j} = 0 \end{aligned} \tag{10}$$

where the last step uses eq (5).

If m is even, then (m-1)/2 in the above equations is replaced by m/2 for the cosine and (m/2)-1 for the sine.

We now do the OLS in terms of the new variables p. Assume first that the p's have been standardized. In particular, it means that, instead of (8), we have

$$p^0 = \frac{1}{C0} \sum_i x^i \quad where\ C0 = \sqrt{\sum_{i,j} \rho_{i,j}} \tag{11}$$

The estimated output in terms of the new variables is

$$\hat{y} = \sum_{i=0}^{m-1} \gamma_j p^j \tag{12}$$

The slope γ_0 is easy to find because of the first of eqs (6). We find

$$\gamma_0 = \frac{1}{C0} \sum_{i,j} b_j \rho_{i,j} + < \varepsilon p^0 > \tag{13}$$

The expectation is the first term on the RHS of (13). To get an estimate of this term, assume all b's and all $\rho's$ are the same. Then

$$E[\gamma_0] \sim O(mb\sqrt{\rho}) \tag{14}$$

alternatives

An alternative to the equations of the previous section is:

$$p^0 = \frac{1}{C0}\sum_i h_i x^i \quad where\ C0 = \sqrt{\sum_{i,j} h_i h_j \rho_{i,j}} \tag{15}$$

$$\theta_k \equiv \frac{2\pi k}{m} \quad k = 0,1,...,\frac{m-1}{2}$$
$$p^{2k} = \sum_{i=0}^{m-1} \frac{h_i x^i}{A_i}\cos(i\theta_k) \quad k = 1,2,,...,\frac{m-1}{2} \tag{16}$$
$$p^{2k-1} = \sum_{i=0}^{m-1} \frac{h_i x^i}{A_i}\sin(i\theta_k) \quad k = 1,2,...,\frac{m-1}{2}$$

$$A_k \equiv \sum_{j=0}^{m-1} h_k h_j \rho_{k,j} \tag{17}$$

where the h_i are constants. Eqs (6) are still valid.

Eq (15) describes a **portfolio**, where the features are weighted.

One useful estimate of the h_i are as estimates for the slopes b_i, eq (1). In this case, those features associated with smaller slope have less effect. This approach, which could probably be useful for a relatively small number of samples, would allow replacement of features with a combination of correlations and slopes. In this approach, features i,j would be included if $|h_i h_j \rho_{i,j}| \leq \rho_c$.

numerical results

We look at numerical results for eqs (7) and (8). N = number of samples = 150 and m = number of features = 7. The true slope b, eq (1), for a given feature is generated randomly to be between 1 and 2. The standard deviation of ε is 1.0.

The correlation matrtix for the features x, computed numerically from the x's themselves, is

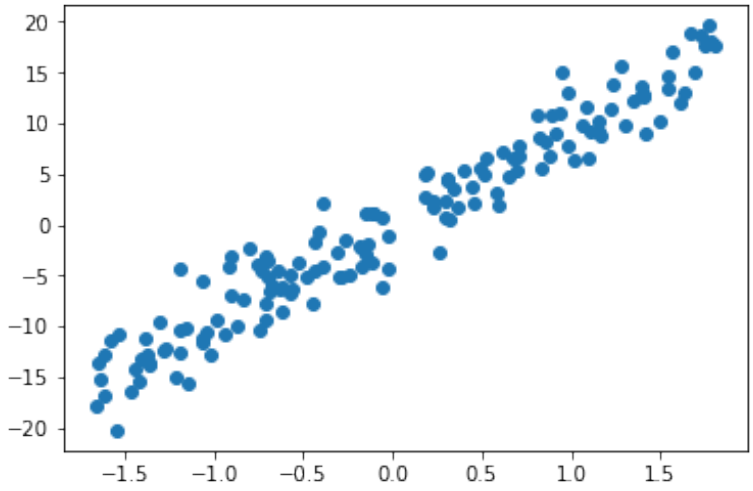
```
[[ 1.     0.895  0.91   0.89   0.858  0.916  0.963]
 [ 0.895  1.     0.788  0.833  0.862  0.856  0.878]
 [ 0.91   0.788  1.     0.857  0.811  0.869  0.877]
 [ 0.89   0.833  0.857  1.     0.805  0.854  0.896]
 [ 0.858  0.862  0.811  0.805  1.     0.873  0.826]
 [ 0.916  0.856  0.869  0.854  0.873  1.     0.868]
 [ 0.963  0.878  0.877  0.896  0.826  0.868  1.   ]]
```

so the correlations are large.

Here are the estimated slopes for the features x, from OLS. The first is the bias term.;

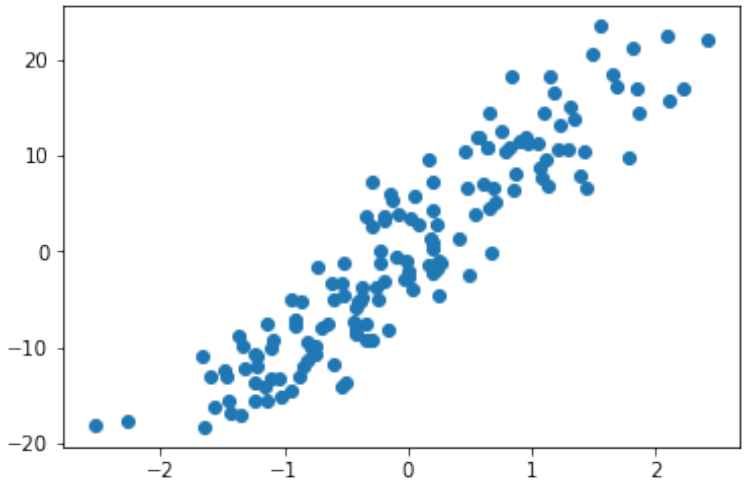
[-0. 1.55 1.618 1.292 1.93 1.406 2.047 1.314]

The following graphs show y, eq (1), vs x for several features x. All 7 graphs look pretty much the same.



Because of the large correlation coefficients, a plot of y vs x for one x really mixes in all the other x's. Thus the slope of the above graph is ~40/3 ~13.3, whereas the estimated slope for a given x is between 1 and 2.

Here is a plot for a different feature in a different ensemble.

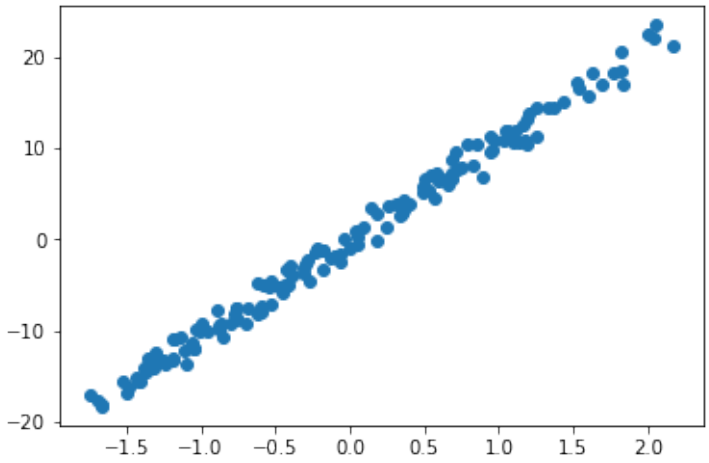


Here the slope is about 10.

We change variables to the p's, as given in eqs (7) and (8). The following is the correlation matrix of the p's, computed numerically from the p's.

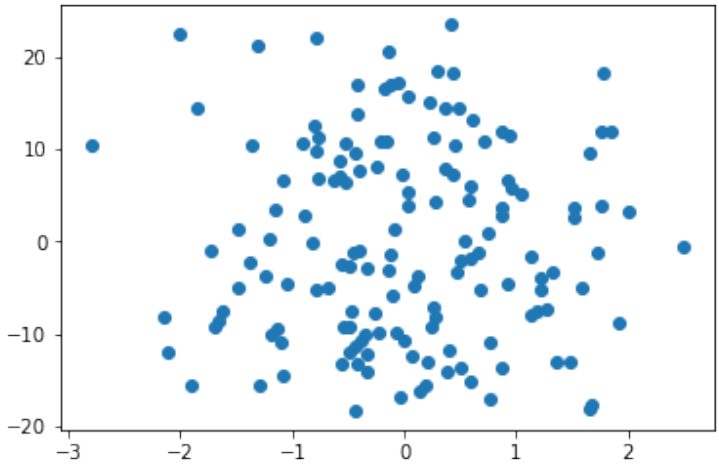
```
[[ 1.     0.    -0.    -0.     0.     0.     0.   ]
 [ 0.     1.    -0.074 -0.229 -0.094  0.26  -0.03 ]
 [-0.    -0.074  1.    -0.04  -0.347 -0.018  0.03 ]
 [-0.    -0.229 -0.04   1.    -0.022  0.113 -0.012]
 [ 0.    -0.094 -0.347 -0.022  1.    -0.043 -0.53 ]
 [ 0.     0.26  -0.018  0.113 -0.043  1.     0.022]
 [ 0.    -0.03   0.03  -0.012 -0.53   0.022  1.   ]]
```

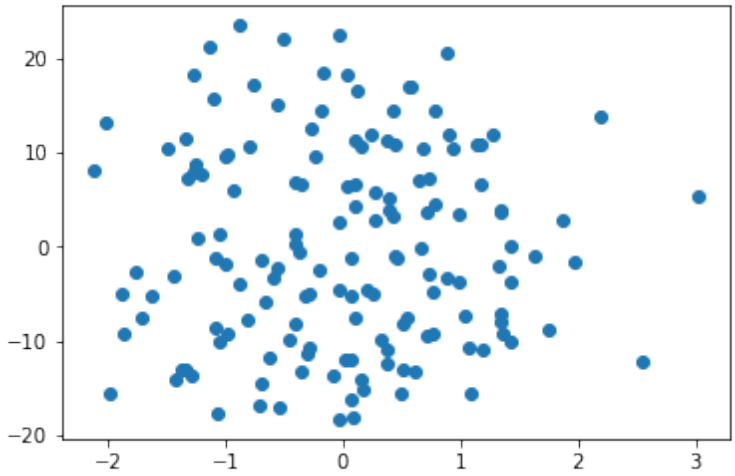
We see that the correlations are quite small, and that eqs (6) are valid. The following is a graph of y vs p^0.



The slope is significant and the variance is seen to be small.

The graphs for y vs the other p's are very similar to each other, and here are two of them.





Just based on these graphs, we can see that the other p's are barely significant.

The following are the estimated slopes for the p's. The first one is the bias.

[-0. 10.555 -0.053 -0.063 0.04 -0.007 0.265 0.038]

We see that, in agreement with the graphs above, the slope for p^0 is large (10.6), while small for the other p's.

weak correlation

We now assume that all correlation coefficients are small, generally less than 0.2. The estimated slopes for are

$$\beta = \rho^{-1} X^T y \tag{18}$$

X is a matrix of the features, as described in OLS2. In this section, we derive an approximation for ρ^{-1} for small correlation coefficients. We assume ρ is well-behaved, so everything exists. In all of this, if m is the number of features and ρ is a typical magnitude of a correlation coefficient, then it is assumed that

$$m\rho \ll 1 \tag{19}$$

Define the **symmetric matrix a**

$$a \equiv \rho^{-1} \tag{20}$$

Each element of a is

$$\begin{aligned} a_{i,i} &= 1 \\ a_{i,j} &= a_{i,j}^{(1)} + a_{i,j}^{(2)} + ...i \neq j \end{aligned} \tag{21}$$

where $a_{i,j}^{(k)}$ is of order ρ^k in some correlation coefficient.

From (20) we find

$$\rho a = I \tag{22}$$

where I is the unit matrix.

Writing out eq (22), using the first of eqs (21), we have

$$\sum_{i(\neq k)} \left(a_{i,k} + \rho_{i,k} \right) + \sum_{\substack{i(\neq j) \\ j(\neq k)}} \rho_{i,j} a_{j,k} = 0 \tag{23}$$

where we also have used $\rho_{i,i} = 1$. Now using the second of eqs (21), and collecting terms of the same order of ρ

$$\sum_{i(\neq k)} \left(a_{i,k}^{(1)} + \rho_{i,k} \right) = 0$$
$$\sum_{i(\neq k)} a_{i,k}^{(2)} + \sum_{\substack{i(\neq j) \\ j(\neq k)}} \rho_{i,j} a_{j,k}^{(1)} = 0 \tag{24}$$

and so forth. We see that

$$a_{i,k} = -\rho_{i,k} \quad i \neq k \tag{25}$$

The second order term is (though we will not use it)

$$a_{i,k}^{(2)} = -\sum_{j(\neq i,k)} \rho_{i,j} \rho_{j,k} = 0 \tag{26}$$

Putting (21) and (25) into (18), we get the estimated slope

$$E[\beta_i] = < x^i y > - \sum_{j(\neq i)} \rho_{i,j} < x^j y > + O(\rho^2) = b_i + O(\rho^2) \tag{27}$$

We can define a set of variables p which, to $O(\rho^2)$, are orthogonal, as

$$p^i = x^i - \frac{1}{2} \sum_{j(\neq i)} \rho_{i,j} x^j \tag{28}$$

$$\text{so } < p^i p^j > = O(\rho^2) \quad i \neq j \tag{29}$$