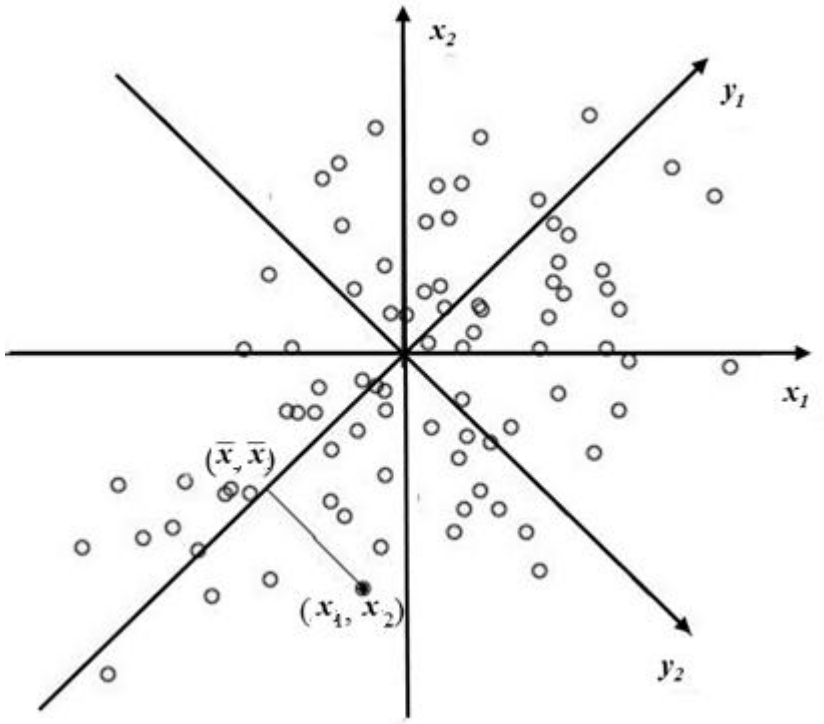
Aleksei Korneev, Anatoly Krichevets, and Dmitriy Ushakov

Large samples: Subtle Effects and Disappointing Artefacts

Appendix 1

Running Variance Estimate

The method of Running Variance Estimate (RVE) is designed for analysis of joint probability distribution of variables. The picture below shows the scatter plot of two variables. Marginal distributions (the distributions of the points projection upon axes) are very close to the normal one but the image in general differ from an ellipsis that corresponds to correlated normal distributions. Such a picture corresponds to SLODR that means the smaller correlation between variables when they have a big values (the upper right quadrant of scatter plot), and bigger correlation between variables when they have a small values (the top-left quadrant of scatter plot).



We begin our description from two dimensional example presented by the picture. Let $f(x_1, x_2)$. be two-dimensional normal distribution with standard normal marginal and correlation ρ (which could be considered as positive and not equal to 1). This distribution density is expressed by formula:

$$f(x_1,\ x_2) = \frac{1}{2\pi\sqrt{1-\rho^2}} exp(-\frac{1}{2(1-\rho^2)}(x_1^2 - 2\rho x_1 x_2 + x_2^2))$$

Changing variable $y_1 = (x_1 + x_2)/\sqrt{2}$ и $y_2 = (x_1 - x_2)/\sqrt{2}$ (this may be done by substitution in the formula given above $x_1 = (y_1 + y_2)/\sqrt{2}$ и $x_2 = (y_1 - y_2)/\sqrt{2}$), we get

$$f(y_1,\ y_2) = \frac{1}{2\pi\sqrt{1-\rho^2}} exp(-(\frac{y_1^2}{1+\rho} + \frac{y_2^2}{1-\rho})).$$

Doing so we get independent random values y_1 and y_2, having variations $1 + \rho$ and $1 — \rho$ respectively.

Mark that the sum $(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2$ due to Pythagoras' theorem is equal to the hypothenuse squared that lies between points (x_1, x_2) and $(\bar{x},\ \bar{x})$ i.e. just the y_2 squared. That is the estimate of variation of the random value $y_2 = (x_1 - x_2)/\sqrt{2}$ (it is produced by the one-point sample but

using the general mean equal to zero) that is equal to 1 - ρ. This y_2 is the RVE value for 'respondent' presented by (x_1, x_2).
Taking mean value for the set of 'respondents' we get more exact variance estimate of $1 - p$. As y_1 and y_2 are independent random values, we can sum any subset of the sample, having got different estimates of the same value. For example, we can take two subsamples: that one defined by the condition $y_1 > 0$, and other one defined by the condition $y_1 <= 0$. The proportion of these estimates has F-distribution, and we can then test the hypothesis on the equality of estimates on low and high subsamples of the sample. If the proportion is extremely big we have argument for SLODR effect in the sample.
Taking smaller partitions one can see the presence of any other dependencies RVE on g.

When there are more variables the calculation may be easily generalized. Then the sum $(x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2$ estimates the general variance of the $(n - 1)$-dimensional normal random value, defined upon the hyper-plane that is orthogonal to the vector $(\bar{x}, \bar{x} \ldots \bar{x})$. Since the correlation of all our original variables are equal as we proposed, than for any orthonormal basis in this hyper-plane the random value projections upon these coordinate axes were independent and not correlated with the random value projection upon axis collinear to the vector $(\bar{x}, \bar{x} \ldots \bar{x})$ (that may be considered as factor g). The variance estimated of the $(n - 1)$-dimensional random value is equal to $n - 1 - \rho$ (while the g-variance is $1 + \rho$). This variance as we proposed may be estimated with any subsample of the sample. We can take two subsamples as in the two-dimensional example and get the F-ratio with corresponding degrees of freedom. Or we can use any other division of sample and find more complicated dependences RVE on factor g.
If inter-correlations of original variables are not equal it is necessary to take weighted sum instead of $(x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2$ but in the case of $\rho_{12} = \rho_{34}$ and $\rho_{13} = \rho_{23} = \rho_{14} = \rho_{24}$ (that is almost so for our data) the weighting coefficients are equal due to evident symmetrical properties of the correlation matrix.