Aleksei Korneev, Anatoly Krichevets, and Dmitriy Ushakov

Large samples: Subtle Effects and Disappointing Artefacts

Table S1

Goodness of fit of the model obtained to different datasets (see section 3.3).

Dataset	Type of data	Group	$\chi 2$ (df=1)	CFI	RMSEA
Real data	Raw	Low	10.959	0.996	0.042
		High	11.937	0.995	0.044
	Normalized	Low	6.264	0.998	0.031
		High	21.761	0.990	0.061
Selection of cases	Raw	Low	0.573	> 0.999	< 0.001
		High	0.053	> 0.999	< 0.001
	Normalized	Low	0.743	> 0.999	< 0.001
		High	0.073	> 0.999	< 0.001
Different density of tasks	Raw	Low	0.016	> 0.999	< 0.001
		High	2.136	> 0.999	0.014
	Normalized	Low	0.157	> 0.999	< 0.001
		High	1.016	> 0.999	0.002
Simulation of SLODR	Raw	Low	0.010	> 0.999	< 0.001
		High	1.662	> 0.999	0.012

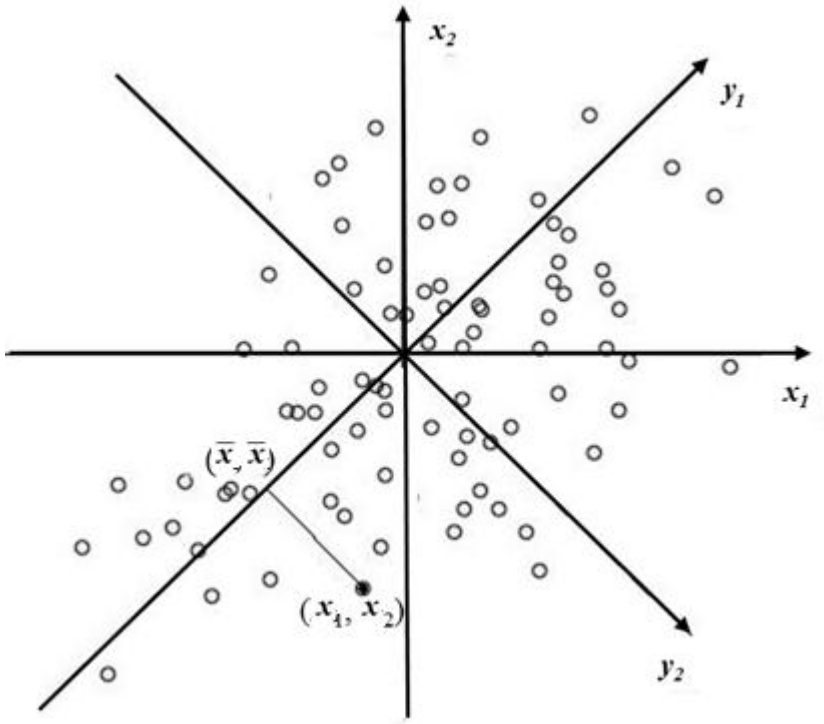
Aleksei Korneev, Anatoly Krichevets, and Dmitriy Ushakov

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 $f(x_1, x_2)$. ИНДЕКСЫ ВНИЗ be two-dimensional normal distribution with standard normal marginal and correlation ρ (witch 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

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

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).ИНДЕКСЫ
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.

Aleksei Korneev, Anatoly Krichevets, and Dmitriy Ushakov

Large samples: Subtle Effects and Disappointing Artefacts

Appendix 2

2 A) SPSS syntax for simulation of the selection of respondents according to the external criterion.

```
/* a*a*b*c - correlation between variables belonging to different groups
/* b*b - correlation within first group
/* c*c - correlation within second group

COMPUTE a=.915.
COMPUTE b=0.838.
COMPUTE c=0.809.
EXECUTE.


COMPUTE Ra = RV.Normal(0, 1) .
EXECUTE.
COMPUTE Rb = Ra * a + SQRT(1 - a*a) * RV.Normal(0, 1).
EXECUTE.
COMPUTE Rc = Ra * a + SQRT(1 - a*a) * RV.Normal(0, 1).
EXECUTE.

/* V11, V12 – first group of variables, V21, V22 – second group of variables

COMPUTE V11 = Rb * b + SQRT(1 - b*b) * RV.Normal(0, 1).
COMPUTE V12 = Rb * b + SQRT(1 - b*b) * RV.Normal(0, 1).
EXECUTE.



COMPUTE V21 = Rc * c + SQRT(1 - c*c) * RV.Normal(0, 1).
COMPUTE V22 = Rc * c + SQRT(1 - c*c) * RV.Normal(0, 1).
EXECUTE.

************************************************************

/* correlation matrix:

CORRELATIONS
  /VARIABLES=V11 V12 V21 V22
  /PRINT=TWOTAIL NOSIG
  /MISSING=PAIRWISE.

/* computing of “g-factor”]

COMPUTE fsc=V11 + V12 + V21 + V22.
EXECUTE.

/* standardization

DESCRIPTIVES VARIABLES=fsc
  /SAVE
  /STATISTICS=MEAN STDDEV MIN MAX.

/* random addition  to control skewness

COMPUTE fsc_corr=Zfsc * 1. + RV.NORMAL(0,1) * .005.
EXECUTE.

/* selection of “respondents”
```

```
USE ALL.
COMPUTE filter_$=(fsc_corr <= .5).
VARIABLE LABELS filter_$ 'fsc_corr <= .3 (FILTER)'.
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
FORMATS filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE.

/* checking result

CORRELATIONS
  /VARIABLES=V11 V12 V21 V22
  /PRINT=TWOTAIL NOSIG
  /MISSING=PAIRWISE.

DESCRIPTIVES VARIABLES=V11 V12 V21 V22
/STATISTICS=MEAN STDDEV SKEWNESS.

/* PCA

FACTOR
  /VARIABLES V11 V12 V21 V22
  /MISSING LISTWISE
  /ANALYSIS V11 V12 V21 V22
  /PRINT INITIAL EXTRACTION
  /CRITERIA FACTORS(1) ITERATE(25)
  /EXTRACTION PC
  /ROTATION NOROTATE
  /SAVE REG(ALL)
  /METHOD=CORRELATION.

/* variable for division to subsamples

COMPUTE fscFA_corr=0.71 * FAC1_1 + 0.71 * RV.NORMAL(0,1).
EXECUTE.

/* division to subsamples

RANK VARIABLES=fscFA_corr (A)
  /NTILES(2)
  /PRINT=YES
  /TIES=MEAN.
```

2 B) SPSS syntax for Simulation of the unequal distribution of tasks according to their difficulty.

```
/* a*a*b*c  correlztion between variables belonging to different groupd
/* b*b correlation within first group
/* c*c correlation within second group

COMPUTE a=.8.
COMPUTE b=0.73.
COMPUTE c=0.69.
EXECUTE.

/* scratching parameter

COMPUTE as=14/15.
EXECUTE.

COMPUTE Ra = RV.Normal(0, 1) .
```

```
EXECUTE.
COMPUTE Rb = Ra * a + SQRT(1 - a*a) * RV.Normal(0, 1).
EXECUTE.
COMPUTE Rc = Ra * a + SQRT(1 - a*a) * RV.Normal(0, 1).
EXECUTE.

/* V11, V12 – first group of variables, V21, V22 – second group of variables

COMPUTE V11 = Rb * b + SQRT(1 - b*b) * RV.Normal(0, 1).
EXECUTE.
COMPUTE V12 = Rb * b + SQRT(1 - b*b) * RV.Normal(0, 1).
EXECUTE.



COMPUTE V21 = Rc * c + SQRT(1 - c*c) * RV.Normal(0, 1).
EXECUTE.
COMPUTE V22 = Rc * c + SQRT(1 - c*c) * RV.Normal(0, 1).
EXECUTE.

*********************************************************

/* scretching of the left half

IF  (V11 < 0) Vn11= - Abs(V11-1) ** (1/as) + 1.
EXECUTE.

.* compressing of the right e

IF  (V11 >= 0) Vn11=(V11+1) ** (as) - 1.
EXECUTE.


IF  (V12 < 0) Vn12= - Abs(V12-1) ** (1/as) + 1.
IF  (V12 >= 0) Vn12=(V12+1) ** (as) - 1.
EXECUTE.

IF  (V21 < 0) Vn21= - Abs(V21-1) ** (1/as) + 1.
EXECUTE.

IF  (V21 >= 0) Vn21=(V21+1) ** (as) - 1.
EXECUTE.

IF  (V22 < 0) Vn22= - Abs(V22-1) ** (1/as) + 1.
EXECUTE.

IF  (V22 >= 0) Vn22=(V22+1) ** (as) - 1.
EXECUTE.

/* division to subsamples

FACTOR
  /VARIABLES Vn11 Vn12 Vn21 Vn22
  /MISSING LISTWISE
  /ANALYSIS Vn11 Vn12 Vn21 Vn22
  /PRINT INITIAL EXTRACTION
  /CRITERIA FACTORS(1) ITERATE(25)
  /EXTRACTION PC
  /ROTATION NOROTATE
  /SAVE REG(ALL)
  /METHOD=CORRELATION.

COMPUTE fsc_corr=0.71 * FAC1_1 + 0.71 * RV.NORMAL(0,1).
EXECUTE.
```

```
RANK VARIABLES=fsc_corr (A)
  /NTILES(2)
  /PRINT=YES
  /TIES=MEAN.
```

2 C) SPSS syntax for simulation of the "right" SLODR

```
COMPUTE filter=RV.BINOM(1,0.5).
EXECUTE.

COMPUTE a1=RV.NORMAL(0,1).
EXECUTE.
COMPUTE a2=RV.NORMAL(0,1).
EXECUTE.


IF (filter) g0=RV.NORMAL(-0.5,1).
EXECUTE.

IF (filter) v11=RV.NORMAL(0,1) * 0.6 + 0.8 * (g0 * 0.8 + a1 * 0.6).
IF (filter) v12=RV.NORMAL(0,1) * 0.6 + 0.8 * (g0 * 0.8 + a1 * 0.6).
IF (filter) v21=RV.NORMAL(0,1) * 0.6 + 0.8 * (g0 * 0.8 + a2 * 0.6).
IF (filter) v22=RV.NORMAL(0,1) * 0.6 + 0.8 * (g0 * 0.8 + a2 * 0.6).
EXECUTE.


IF ( ~ filter) g0=RV.NORMAL(0.5,1).
EXECUTE.

IF ( ~ filter) v11=RV.NORMAL(0,1) * 0.8 + 0.6 * (g0 * 0.6 + a1 * 0.8).
IF ( ~ filter) v12=RV.NORMAL(0,1) * 0.8 + 0.6 * (g0 * 0.6 + a1 * 0.8).
IF ( ~ filter) v21=RV.NORMAL(0,1) * 0.8 + 0.6 * (g0 * 0.6 + a2 * 0.8).
IF ( ~ filter) v22=RV.NORMAL(0,1) * 0.8 + 0.6 * (g0 * 0.6 + a2 * 0.8).
EXECUTE.



/* standardization with Zv11, Zv12, Zv21, Zv22 as a result

DESCRIPTIVES VARIABLES= Zv11 Zv12 Zv21 Zv22
  /SAVE
  /STATISTICS=MEAN STDDEV MIN MAX.


COMPUTE MeanV=(Zv11 + Zv12 + Zv21 + Zv22) / 4.
EXECUTE.

/* standardization of ZMeanV as a result

DESCRIPTIVES VARIABLES= MeanV
  /SAVE
  /STATISTICS=MEAN STDDEV MIN MAX.

/* division to subsamples

COMPUTE fsc_corr =.71 * ZMeanV + .71 * RV.NORMAL(0,1).
EXECUTE.

RANK VARIABLES=fsc_corr (A)
  /NTILES(2)
  /PRINT=YES
  /TIES=MEAN.
```