

Testing the significance of the RV coefficient

Application to napping data

Julie Josse, François Husson and Jérôme Pagès
Applied Mathematics Department
Agrocampus Rennes, IRMAR CNRS UMR 6625

Agrostat

January, the 25th 2008

The data

- Sensory evaluation
- 8 wines, 12 panelists, 2 sessions
- Direct collection of sensory distances: napping (*Pagès 2003*)

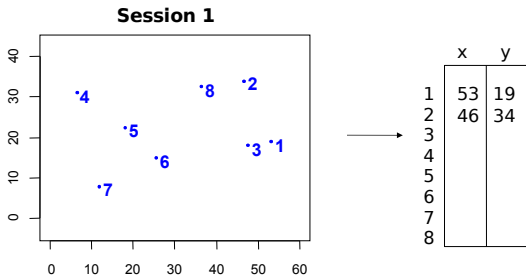


Figure: A napping configuration.

Problems

Repeatability: is the product configuration given by a taster roughly the same from one session to the other?

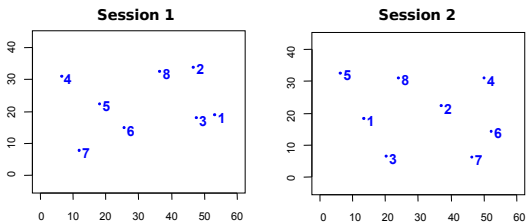


Figure: Panelist 10's configurations.

⇒ Test : H_0 "the two configurations are not correlated" versus H_1 "the two configurations are correlated".

The *RV* coefficient, *Escoufier 1973* (1)

⇒ A measure of relationship between two sets of variables.

- Let $X_{n \times p}$ and $Y_{n \times q}$, if X and Y are centered by columns, the *RV* coefficient is defined by (with $\|A\| = \sqrt{\text{tr}(A'A)}$):

$$RV(X, Y) = \frac{\text{tr}(XX'YY')}{\sqrt{\text{tr}(X'X)^2 \text{tr}(Y'Y)^2}} = \frac{\langle W_X, W_Y \rangle}{\|W_X\| \|W_Y\|}.$$

- Distance between data matrices:

$$\begin{aligned} d(X, Y) &= \left\| \frac{XX'}{(\text{tr}(XX')^2)^{1/2}} - \frac{YY'}{(\text{tr}(YY')^2)^{1/2}} \right\|, \\ &= \sqrt{2} \sqrt{1 - RV(X, Y)}. \end{aligned}$$



The RV coefficient, *Escoufier 1973* (2)

Properties:

- $0 \leq RV(X, Y) \leq 1$
- $RV(X, Y) = 0$, if and only if $X^T Y = 0$
- $RV(X, BX + c) = 1$, B is an orthogonal matrix ($B'B = I$) and c is a constant vector
- if $p = q = 1$, $RV(X, Y) = r^2(X, Y)$

The asymptotic distribution

- Robert *et al* (1985): joint parent distribution belongs to the class of normal distributions
- Cléroux and Ducharme (1989): elliptical distributions
- Cléroux (1995): tests based on rank

⇒ The tests derived are very sensitive to the departure from the distribution hypothesis and to the sample size.

Permutation tests

- Compute the RV coefficient between the two configurations X and Y
- Permute the rows of one matrix (Y for example) and the RV coefficient is computed for each of the $n!$ permutations
- The p -value is the proportion of the values greater to the observed one

⇒ When n is important, it is not possible to perform the $n!$ permutations in term of computational cost.

To approximate the RV permutation distribution

Two approaches:

- random sampling from all possible permutations
- approximation by a continuous distribution using the analytical moments of the exact permutation distribution under the null hypothesis
- Several types of moments-based approximations:
 - Transformations: Log transformation (Heo & Gabriel, 1998)
 - The Pearson family
 - Edgeworth expansion

Calculating the first moments

- The first three moments are obtained (without doing any permutations) under H_0 (Kazi-Aoual *et al.*, 1995).

$$\mathbb{E}_{H_0}(RV) = \frac{\sqrt{\beta_x \times \beta_y}}{n-1},$$

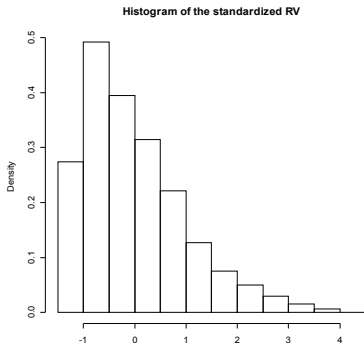
with,

$$\beta_x = \frac{(\text{tr}(X'X))^2}{\text{tr}((X'X)^2)} = \frac{(\sum \lambda_i)^2}{\sum \lambda_i^2}.$$

- β_x can be seen as a measure of complexity (or dimensionality or an equivalent number of variables).
 $1 \leq \beta_x \leq p$.

A normal approximation

- The RV converges weakly to a normal distribution.



Test based on the standardized RV :

$$RV_{\text{std}} = \frac{RV - \mathbb{E}_{H_0}(RV)}{\sqrt{\mathbb{V}_{H_0}(RV)}}.$$

⇒ Problem: the exact distribution of the standardized RV distribution is often skewed.

Pearson type III approximation (Johnson *et al*, 1994)

- The standardized *RV* distribution is approximated by:

$$f(x) = \frac{(2/\gamma)^{4/\gamma^2}}{\Gamma(4/\gamma^2)} \left(\frac{2 + \gamma x}{\gamma} \right)^{(4-\gamma^2)/\gamma^2} e^{-2(2+\gamma x)/\gamma^2}.$$

This Pearson type III distribution has zero mean, unit variance and skewness equal to γ .

- It includes several frequently encountered distributions (exponential, chi square,...)
- ⇒ Provides adequate approximations in many cases.

Edgeworth expansion (Johnson *et al*, 1994)

- Edgeworth expansion approximates the distribution around the limit distribution (often the normal distribution) by a combination of Hermite polynomials with coefficients defined in terms of cumulants (which depend on the moments).

$$f(x) \approx \phi(x) \left(1 + \frac{1}{6} k_3 H_3(x) + \frac{1}{24} (k_4 - 3) H_4(x) + \dots \right).$$

- Truncated to the first order:

$$f(x) \approx \phi(x) \left(1 + \frac{1}{6} \gamma (x^3 - 3x) \right).$$

- This first order term corrects the basic normal approximation for the main effect of skewness.

Approximation of the distribution

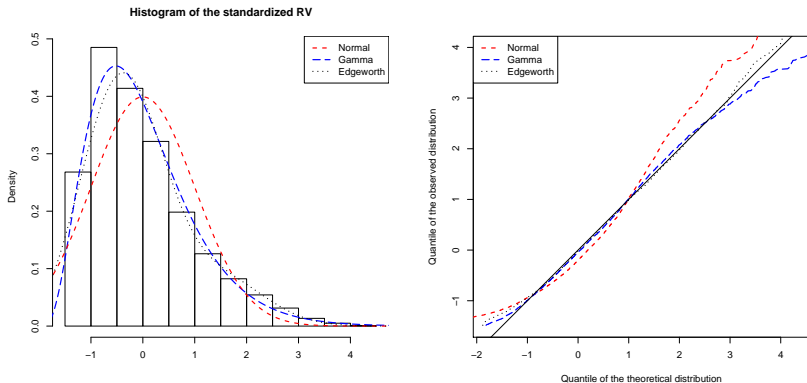


Figure: Normal, Edgeworth and Pearson approximations of the standardized RV.

Simulation study (1)

- Vary the number of individuals and the number of variables
- Two underlying distributions: Normal and Uniform
- 10000 simulations are drawn, for each parameters set (n and p) and for both distribution, under the null hypothesis

⇒ Number of null hypothesis rejected (a value of 5 per cent is expected).

Simulation study (2)

Normale	$n = 6$	$n = 10$	$n = 30$	$n = 100$	$n = 1000$
$p = q = 2$	0.161	0.104	0.067	0.056	0.047
$p = q = 5$	0.209	0.131	0.071	0.054	0.052
$p = q = 10$	0.323	0.178	0.092	0.053	0.051
$p = q = 30$	0.748	0.445	0.169	0.082	0.052

Table: Empirical significant level for the asymptotic test

$p = q$	$\mathbb{E}_{H_0}(RV)$	$\mathbb{V}_{H_0}(RV)$	γ	Normal	Pears	Rand	Edge	logRV
$n = 6$								
2	0.307	0.034	0.682	0.078	0.061	0.050	0.057	0.054
5	0.502	0.018	0.287	0.063	0.055	0.051	0.053	0.045
10	0.657	0.008	0.146	0.055	0.051	0.049	0.051	0.045
$n = 30$								
2	0.065	0.002	1.313	0.075	0.052	0.053	0.042	0.056
5	0.145	0.002	0.554	0.061	0.050	0.051	0.048	0.048
10	0.252	0.001	0.282	0.058	0.052	0.053	0.050	0.048
$n = 100$								
2	0.020	2e-04	1.389	0.070	0.051	0.051	0.039	0.054
5	0.048	2e-04	0.566	0.062	0.051	0.053	0.048	0.048
10	0.091	2e-04	0.285	0.056	0.049	0.051	0.048	0.047

Table: Empirical significant level for the different approximations

Application on napping data

Taster	RV	RV_{std}	$E_{H_0}(RV)$	γ	Norm	Pears	Rand	Edge	Log	Exact
1	0.552	2.078	0.245	0.790	0.019	0.035	0.034	0.039	0.041	0.039
2	0.222	-0.335	0.263	0.467	0.631	0.605	0.581	0.605	0.564	0.580
3	0.357	0.944	0.216	1.066	0.173	0.160	0.160	0.168	0.132	0.162
4	0.127	-0.579	0.217	0.944	0.719	0.684	0.648	0.683	0.695	0.642
5	0.640	2.727	0.222	1.141	0.003	0.017	0.023	0.015	0.022	0.020
6	0.144	-0.320	0.190	1.127	0.625	0.559	0.535	0.561	0.530	0.528
7	0.794	3.841	0.157	1.571	6.00E-05	0.006	0.004	0.001	0.011	0.004
8	0.058	-0.873	0.195	1.238	0.810	0.817	0.828	0.798	0.914	0.830
9	0.492	2.080	0.144	1.609	0.019	0.044	0.057	0.060	0.037	0.055
10	0.280	0.373	0.221	1.312	0.354	0.287	0.270	0.285	0.245	0.276
11	0.224	0.046	0.217	1.221	0.482	0.401	0.380	0.401	0.352	0.376
12	0.190	-0.262	0.231	1.039	0.603	0.539	0.523	0.541	0.505	0.520
Jury	0.850	3.27	0.597	0.044	5e-04	7e-04	4e-04	7e-04	0.002	3e-04

- Only three tasters yield linked configurations (1, 5, 7)
- The panel is "repeatable"

Conclusion

⇒ Two solutions:

- Random approximation ⇒ problem to perform plenty of tests
- The pearson approximation

- The normal approximation is not accurate
- The log transformation improves the normal one
- Pearson and Edgeworth perform quite well, but Edgeworth presents shortcomings

Other applications

- In many fields, the problem of relating data from different sources is usually faced
- To compare two factorial maps (such as PCA)
- To impute only with informative data set in the framework of missing values in multiple multivariate dataset

FactoMineR

The function `coeffRV` is implemented in the FactoMineR package (R)



<http://factominer.free.fr>