Testing the significance of the \textit{RV} coefficient

Application to napping data

Julie Josse, François Husson and Jérôme Pagès

\textit{Applied Mathematics Department}
\textit{Agrocampus Rennes, IRMAR CNRS UMR 6625}

\textit{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 \textit{RV} coefficient, \textit{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 \textit{RV} coefficient is defined by (with $\|A\| = \sqrt{\text{tr}(A'A)}$):

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

- Distance between data matrices:

\[
d(X, Y) = \| \frac{XX'}{(\text{tr}(XX')^2)^{1/2}} - \frac{YY'}{(\text{tr}(YY')^2)^{1/2}} \|,
\]

\[
= \sqrt{2} \sqrt{1 - RV(X, Y)}.
\]
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,

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

- $\beta_x$ 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 $RV$:

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

⇒ 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:

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

This Pearson type III distribution has zero mean, unit variance and skewness equal to \( \gamma \).

- It includes several frequently encountered distributions (exponential, chi square,...)

\[ \Rightarrow \text{Provides adequate approximations in many cases.} \]
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).

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

- 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 \((n\text{ and } p)\) and for both distribution, under the null hypothesis

\[\Rightarrow\text{ Number of null hypothesis rejected (a value of 5 per cent is expected).}\]
Simulation study (2)

<table>
<thead>
<tr>
<th>$n$</th>
<th>$n = 6$</th>
<th>$n = 10$</th>
<th>$n = 30$</th>
<th>$n = 100$</th>
<th>$n = 1000$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p = q = 2$</td>
<td>0.161</td>
<td>0.104</td>
<td>0.067</td>
<td>0.056</td>
<td>0.047</td>
</tr>
<tr>
<td>$p = q = 5$</td>
<td>0.209</td>
<td>0.131</td>
<td>0.071</td>
<td>0.054</td>
<td>0.052</td>
</tr>
<tr>
<td>$p = q = 10$</td>
<td>0.323</td>
<td>0.178</td>
<td>0.092</td>
<td>0.053</td>
<td>0.051</td>
</tr>
<tr>
<td>$p = q = 30$</td>
<td>0.748</td>
<td>0.445</td>
<td>0.169</td>
<td>0.082</td>
<td>0.052</td>
</tr>
</tbody>
</table>

**Table:** Empirical significant level for the asymptotic test

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

**Table:** Empirical significant level for the different approximations
Application on napping data

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

- 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://factomineur.free.fr