

# Nonparametric tests of independence between random vectors

R. Beran<sup>1</sup>   M. Bilodeau<sup>2</sup>   **P. Lafaye de Micheaux<sup>3</sup>**

<sup>1</sup>Department of Statistics  
University of California Davis

<sup>2</sup>Département de Mathématiques et de Statistique  
Université de Montréal

<sup>3</sup>Laboratoire Jean Kuntzman  
Université Pierre Mendès France, Grenoble

Swiss Statistics Meeting, Nov 16 (2006)

# Outline of the talk

- 1 Goals, tools and notations
- 2 Test of independence : non serial case
- 3 The Bootstrap
- 4 Bootstrap Validity
- 5 Examples

# Goals and notations

## Goal 1

Construct a test of mutual independence between the random vectors  $X^{(1)} \in \mathbb{R}^{d_1}, \dots, X^{(p)} \in \mathbb{R}^{d_p}$ .

Let  $P$  be the joint law of  $\mathbf{X} = (X^{(j)})_{j=1}^p$  and  $P^{(j)}$  be the marginal law of  $X^{(j)}$ .

Sample = the data :  $\mathbf{X}_1, \dots, \mathbf{X}_n$ .

## Example : the asthma case

Families  $i$  ( $i = 1, \dots, n$ ) with 3 persons with asthma disease.

$$\mathbf{x}_i = \begin{pmatrix} X^{(1)} \\ X^{(2)} \\ X^{(3)} \end{pmatrix} = \begin{pmatrix} \text{Phenotype for the father} \\ \text{Phenotype for the mother} \\ \text{Phenotype for the child} \end{pmatrix}$$

- Dependence between (father,child) and/or (mother,child) solely implies genetic factors.
- Dependence between (father,mother) implies environmental factors.

# Goals and notations

## Goal 2

Let  $Y_1, Y_2, \dots$  be a stationary sequence of random vectors in  $\mathbb{R}^q$ . Construct a serial test of mutual independence between the  $Y_i$ 's.

This problem, with the overlapping difficulty, is treated similarly as the previous case by letting  $X_i = (Y_i, \dots, Y_{i+p-1}) \in \mathbb{R}^{pq}$  and  $X_i^{(j)} = Y_{i+j-1}$ ;  $i = 1, \dots, n - p + 1$ ;  $j = 1, \dots, p$ .

# Notations

- For all  $(\mathbf{s}^{(j)}, t^{(j)}) \in \mathcal{S}_{d_j} \times \mathbb{R}$ , define the half-space  $H$  by

$$H(\mathbf{s}^{(j)}, t^{(j)}) = \left\{ \mathbf{x}^{(j)} \in \mathbb{R}^{d_j} : \langle \mathbf{s}^{(j)}, \mathbf{x}^{(j)} \rangle \leq t^{(j)} \right\}.$$

- The collection of half-spaces in  $\mathbb{R}^{d_j}$  separating probabilities is

$$\mathcal{F}^{(d_j)} = \left\{ H(\mathbf{s}^{(j)}, t^{(j)}) : (\mathbf{s}^{(j)}, t^{(j)}) \in \mathcal{S}_{d_j} \times \mathbb{R} \right\}$$

- $\mathcal{F} = \mathcal{F}^{(d_1)} \times \dots \times \mathcal{F}^{(d_p)}$ .

# How to characterize independence

Define, for all  $(s^{(j)}, t^{(j)}) \in \mathcal{S}_{d_j} \times \mathbb{R}$ ,

$$\begin{aligned} \nu_A((s^{(j)}, t^{(j)})_{j=1}^p) &= \sum_{B \subset A} (-1)^{|A \setminus B|} P(\times_{j=1}^p H^B(s^{(j)}, t^{(j)})) \\ &\quad \cdot \prod_{j \in A \setminus B} P^{(j)}(H(s^{(j)}, t^{(j)})), \end{aligned}$$

and

$$H^B(s^{(j)}, t^{(j)}) = \begin{cases} H(s^{(j)}, t^{(j)}), & j \in B; \\ \mathbb{R}^{d_j}, & j \notin B. \end{cases}$$

## Proposition

*The marginals  $X^{(1)}, \dots, X^{(p)}$  are independent if and only if  $\nu_A((s^{(j)}, t^{(j)})_{j=1}^p) = 0$ , for all  $(H(s^{(j)}, t^{(j)}))_{j=1}^p \in \mathcal{F}$  and all  $A \subset \{1, \dots, p\}$ ,  $|A| > 1$ .*

# To give an idea

Case  $p = 3$ .

$$\nu_{\{1,2\}} = P^{(1,2)} - P^{(1)}P^{(2)}.$$

Then,  $\nu_{\{i,j\}} = 0 \Rightarrow X^{(i)} \perp X^{(j)}, i, j = 1, 2, 3; i < j$ .

$$\begin{aligned} \nu_{\{1,2,3\}} &= P^{(1,2,3)} + 3P^{(1)}P^{(2)}P^{(3)} \\ &\quad - P^{(1,2)}P^{(3)} - P^{(1,3)}P^{(2)} - P^{(2,3)}P^{(1)} - P^{(1)}P^{(2)}P^{(3)}. \end{aligned}$$

Then,  $\nu_{\{1,2,3\}} = 0 \Rightarrow \{X^{(1)}, X^{(2)}, X^{(3)}\}$  independent.

## Notations and processes used

The processes considered, for  $A \subset \{1, \dots, p\}$ , are

$$R_{n,A}((s^{(j)}, t^{(j)})_{j=1}^p) = \sqrt{n} \sum_{B \subset A} (-1)^{|A \setminus B|} \mathbb{P}_n \left( \times_{j=1}^p H^B(s^{(j)}, t^{(j)}) \right) \\ \cdot \prod_{j \in A \setminus B} \mathbb{P}_n^{(j)}(H(s^{(j)}, t^{(j)})),$$

where  $\mathbb{P}_n$  is the empirical law of  $\mathbf{X}_1, \dots, \mathbf{X}_n$  iid and  $\mathbb{P}_n^{(j)}$  is the empirical law of  $X_1^{(j)}, \dots, X_n^{(j)}$  iid.

## Notations and processes used

Another process used is

$$\begin{aligned}
 \check{R}_{n,A}((s^{(j)}, t^{(j)})_{j=1}^p) &= \sqrt{n} \sum_{BCA} (-1)^{|A \setminus B|} \mathbb{P}_n \left( \times_{j=1}^p H^B(s^{(j)}, t^{(j)}) \right) \\
 &\quad \cdot \prod_{j \in A \setminus B} P^{(j)}(H(s^{(j)}, t^{(j)})) \\
 &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \prod_{k \in A} \left[ \mathbb{1}\{X_i^{(k)} \in H(s^{(k)}, t^{(k)})\} \right. \\
 &\quad \left. - P^{(k)}(H(s^{(k)}, t^{(k)})) \right],
 \end{aligned}$$

from the [multinomial formula](#).

# The first theorem

## Theorem

If  $X^{(1)}, \dots, X^{(p)}$  are independent, then

$$\{\check{R}_{n,A} : A \in \mathcal{I}_p\} \rightsquigarrow \{R_A : A \in \mathcal{I}_p\},$$

where  $\rightsquigarrow$  is the weak convergence of Hoffmann-Jørgensen.

The processes  $R_A$  are independent Gaussian of mean 0 and autocovariance function  $C_A((s^{(j)}, t^{(j)})_{j=1}^p, (\check{s}^{(j)}, \check{t}^{(j)})_{j=1}^p)$

$$= \prod_{k \in A} \left[ P^{(k)}(H(s^{(k)}, t^{(k)}) \cap H(\check{s}^{(k)}, \check{t}^{(k)})) \right. \\ \left. - P^{(k)}(H(s^{(k)}, t^{(k)}))P^{(k)}(H(\check{s}^{(k)}, \check{t}^{(k)})) \right].$$

## and the second

The processes  $R_{n,A}$  et  $\check{R}_{n,A}$  are asymptotically equivalent.

### Theorem

*For all  $A \in \mathcal{I}_p$ ,  $\|R_{n,A} - \check{R}_{n,A}\|_{\mathcal{F}} \rightarrow 0$ , where convergence is in outer probability.*

A critical region for an independence test is obtained by combining Kolmogorov type statistics :

$$\cup_{A \in \mathcal{I}_p} \{ \|R_{n,A}\|_{\mathcal{F}} > r_A \}.$$

The asymptotic significance level of the test is

$$\alpha = 1 - \prod_{A \in \mathcal{I}_p} P\{\|R_A\|_{\mathcal{F}} \leq r_A\}.$$

The critical values  $r_A$  can be chosen as the  $\beta$ -quantiles of the law of  $\|R_A\|_{\mathcal{F}}$ , where  $\beta = (1 - \alpha)^{1/(2^p - p - 1)}$ .

However, the law of  $R_A$  depends on the unknown marginal laws  $P^{(k)}$ . The critical values are thus derived from the Bootstrap.

# The Bootstrap technique

- The data :  $X_1, \dots, X_n$
- The statistic  $T = T(X_1, \dots, X_n)$  : only one observed value based on the initial sample
- Intensive computer simulation technique
- Several drawing with replacement in the sample
- For  $b = 1, \dots, B$ 
  - Draw with replacement :  $X_1^*(b), \dots, X_n^*(b)$
  - Compute  $T(X_1^*(b), \dots, X_n^*(b))$
- Estimate the law of  $T$

## Bootstrap of $R_{n,A}$

Define the quarter-space semi-metric  $d_R$  between two finite collections of probabilities by

$$\begin{aligned} d_R \left( (P^{(j)})_{j=1}^p, (Q^{(j)})_{j=1}^p \right) \\ = \sum_{j=1}^p \sup_{H_1, H_2 \in \mathcal{F}^{(d_j)}} |P^{(j)}(H_1 \cap H_2) - Q^{(j)}(H_1 \cap H_2)|. \end{aligned}$$

With this semi-metric, the marginal empirical probabilities converge

$$d_R \left( (\mathbb{P}_n^{(j)})_{j=1}^p, (P^{(j)})_{j=1}^p \right) \xrightarrow{P} 0.$$

## Theorem (Bootstrap validity)

Let  $(P_n^{(j)})_{j=1}^p$ ,  $n = 1, 2, \dots$  be any sequence satisfying

$$d_R \left( (P_n^{(j)})_{j=1}^p, (P^{(j)})_{j=1}^p \right) \rightarrow 0. \quad (1)$$

If  $X_{n1}, \dots, X_{nn}$  are iid from  $P_n^{(1)} \times \dots \times P_n^{(p)}$ ,  $\hat{\mathbb{P}}_n$  is the empirical distribution of  $X_{n1}, \dots, X_{nn}$  and

$$\begin{aligned} R_{n,A}^*((s^{(j)}, t^{(j)})_{j=1}^p) &= \sqrt{n} \sum_{B \subset A} (-1)^{|A \setminus B|} \hat{\mathbb{P}}_n(\times_{j=1}^p H^B(s^{(j)}, t^{(j)})) \\ &\quad \cdot \prod_{j \in A \setminus B} \hat{\mathbb{P}}_n^{(j)}(H(s^{(j)}, t^{(j)})), \end{aligned}$$

then  $\{R_{n,A}^* : A \in \mathcal{I}_p\} \rightsquigarrow \{R_A : A \in \mathcal{I}_p\}$ .

# The dependogram

- Graphical display giving the  $\|R_{n,A}\|_{\mathcal{F}}$  values (vertical bar)
- Star at the height given by the bootstrap approximation to the  $\beta$ -quantile,  $\beta = (1 - \alpha)^{1/(2^p - p - 1)}$ , of  $\|R_A\|_{\mathcal{F}}$
- Subsets such that the vertical bar exceeds this quantile can be flagged for dependent vectors

Subsets							
1	{1,2}	4	{2,3}	7	{1,2,3}	10	{2,3,4}
2	{1,3}	5	{2,4}	8	{1,2,4}	11	{1,2,3,4}
3	{1,4}	6	{3,4}	9	{1,3,4}		

TAB.: Lexicographic order of the subsets for  $p = 4$  in the non serial dependogram.

## The dependogram

- Graphical display giving the  $\|R_{n,A}\|_{\mathcal{F}}$  values (vertical bar)
- Star at the height given by the bootstrap approximation to the  $\beta$ -quantile,  $\beta = (1 - \alpha)^{1/(2^p - p - 1)}$ , of  $\|R_A\|_{\mathcal{F}}$
- Subsets such that the vertical bar exceeds this quantile can be flagged for dependent vectors

Subsets							
1	{1,2}	4	{2,3}	7	{1,2,3}	10	{2,3,4}
2	{1,3}	5	{2,4}	8	{1,2,4}	11	{1,2,3,4}
3	{1,4}	6	{3,4}	9	{1,3,4}		

TAB.: Lexicographic order of the subsets for  $p = 4$  in the non serial dependogram.

## The dependogram

- Graphical display giving the  $\|R_{n,A}\|_{\mathcal{F}}$  values (vertical bar)
- Star at the height given by the bootstrap approximation to the  $\beta$ -quantile,  $\beta = (1 - \alpha)^{1/(2^p - p - 1)}$ , of  $\|R_A\|_{\mathcal{F}}$
- Subsets such that the vertical bar exceeds this quantile can be flagged for dependent vectors

Subsets							
1	{1,2}	4	{2,3}	7	{1,2,3}	10	{2,3,4}
2	{1,3}	5	{2,4}	8	{1,2,4}	11	{1,2,3,4}
3	{1,4}	6	{3,4}	9	{1,3,4}		

TAB.: Lexicographic order of the subsets for  $p = 4$  in the non serial dependogram.

## The dependogram

- Graphical display giving the  $\|R_{n,A}\|_{\mathcal{F}}$  values (vertical bar)
- Star at the height given by the bootstrap approximation to the  $\beta$ -quantile,  $\beta = (1 - \alpha)^{1/(2^p - p - 1)}$ , of  $\|R_A\|_{\mathcal{F}}$
- Subsets such that the vertical bar exceeds this quantile can be flagged for dependent vectors

Subsets							
1	{1,2}	4	{2,3}	7	{1,2,3}	10	{2,3,4}
2	{1,3}	5	{2,4}	8	{1,2,4}	11	{1,2,3,4}
3	{1,4}	6	{3,4}	9	{1,3,4}		

**TAB.:** Lexicographic order of the subsets for  $p = 4$  in the non serial dependogram.

## Dependence among 4 discrete variables

- $W_1, \dots, W_6$  iid with  $W_i, i \in \{1, 3, 4, 6\} \sim \text{Poisson}(1)$  and  $W_i, i \in \{2, 5\} \sim \text{Poisson}(3)$
- $X^{(1)} = W_1 + W_2, X^{(2)} = W_2 + W_3, X^{(3)} = W_4 + W_5,$  and  $X^{(4)} = W_5 + W_6$
- $(X^{(1)}, X^{(2)})$  independent of the pair  $(X^{(3)}, X^{(4)})$  with each pair having a correlation of  $\frac{3}{4}$

### Remark

$$\nu_{\{1,2,3,4\}} = \nu_{\{1,2\}} \cdot \nu_{\{3,4\}} = (P^{(1,2)} - P^{(1)}P^{(2)})(P^{(3,4)} - P^{(3)}P^{(4)})$$

## Dependence among 4 discrete variables

- $W_1, \dots, W_6$  iid with  $W_i, i \in \{1, 3, 4, 6\} \sim \text{Poisson}(1)$  and  $W_i, i \in \{2, 5\} \sim \text{Poisson}(3)$
- $X^{(1)} = W_1 + W_2, X^{(2)} = W_2 + W_3, X^{(3)} = W_4 + W_5,$  and  $X^{(4)} = W_5 + W_6$
- $(X^{(1)}, X^{(2)})$  independent of the pair  $(X^{(3)}, X^{(4)})$  with each pair having a correlation of  $\frac{3}{4}$

### Remark

$$\nu_{\{1,2,3,4\}} = \nu_{\{1,2\}} \cdot \nu_{\{3,4\}} = (P^{(1,2)} - P^{(1)}P^{(2)})(P^{(3,4)} - P^{(3)}P^{(4)})$$

## Dependence among 4 discrete variables

- $W_1, \dots, W_6$  iid with  $W_i, i \in \{1, 3, 4, 6\} \sim \text{Poisson}(1)$  and  $W_i, i \in \{2, 5\} \sim \text{Poisson}(3)$
- $X^{(1)} = W_1 + W_2, X^{(2)} = W_2 + W_3, X^{(3)} = W_4 + W_5,$  and  $X^{(4)} = W_5 + W_6$
- $(X^{(1)}, X^{(2)})$  independent of the pair  $(X^{(3)}, X^{(4)})$  with each pair having a correlation of  $\frac{3}{4}$

### Remark

$$\nu_{\{1,2,3,4\}} = \nu_{\{1,2\}} \cdot \nu_{\{3,4\}} = (P^{(1,2)} - P^{(1)}P^{(2)})(P^{(3,4)} - P^{(3)}P^{(4)})$$

## Dependence among 4 discrete variables

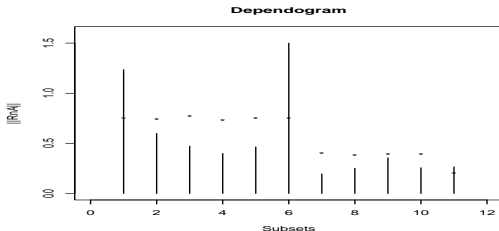
- $W_1, \dots, W_6$  iid with  $W_i, i \in \{1, 3, 4, 6\} \sim \text{Poisson}(1)$  and  $W_i, i \in \{2, 5\} \sim \text{Poisson}(3)$
- $X^{(1)} = W_1 + W_2, X^{(2)} = W_2 + W_3, X^{(3)} = W_4 + W_5,$  and  $X^{(4)} = W_5 + W_6$
- $(X^{(1)}, X^{(2)})$  independent of the pair  $(X^{(3)}, X^{(4)})$  with each pair having a correlation of  $\frac{3}{4}$

### Remark

$$\nu_{\{1,2,3,4\}} = \nu_{\{1,2\}} \cdot \nu_{\{3,4\}} = (P^{(1,2)} - P^{(1)}P^{(2)})(P^{(3,4)} - P^{(3)}P^{(4)})$$

## Dependence among 4 discrete variables

**FIG.:** The two structures of dependence are evident in subsets 1 and 6 which correspond, respectively, to the two subsets  $A = \{1, 2\}$  and  $A = \{3, 4\}$ .  $n = 100$ .



## Dependence between three bivariate vectors

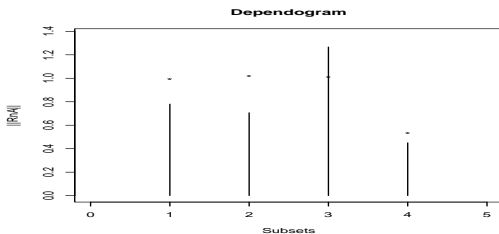
Consider  $n = 50$  observations on six variables  $W_i$ ,  $i = 1, \dots, 6$ , jointly distributed as a multivariate normal with mean vector 0 and covariance matrix

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 & .4 & .5 \\ 0 & 0 & 0 & 1 & .1 & .2 \\ \hline 0 & 0 & .4 & .1 & 1 & 0 \\ 0 & 0 & .5 & .2 & 0 & 1 \end{pmatrix}.$$

Partition into 3 sub-vectors  $X^{(1)} = (W_1, W_2)$ ,  $X^{(2)} = (W_3, W_4)$  and  $X^{(3)} = (W_5, W_6)$ .

## Dependence between three bivariate vectors

**FIG.:** The dependence between the last two subvectors shows up in the third subset  $A = \{2, 3\}$ .  $n = 50$ .



## 4-dependent variables which are 2-independent and 3-independent

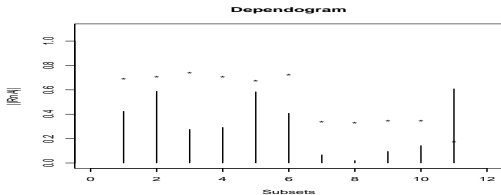
- $W$  is discrete uniform on the set  $\{1, 2, 3, 4, 5, 6, 7, 8\}$
- 

$$\begin{aligned}X^{(1)} &= \mathbb{I}\{W \in \{1, 2, 3, 5\}\}, & X^{(2)} &= \mathbb{I}\{W \in \{1, 2, 4, 6\}\} \\X^{(3)} &= \mathbb{I}\{W \in \{1, 3, 4, 7\}\}, & X^{(4)} &= \mathbb{I}\{W \in \{2, 3, 4, 8\}\}.\end{aligned}$$

- These four dependent binary variables are 2-independent or pairwise independent ; they are also 3-independent

## 4-dependent variables which are 2-independent and 3-independent

**FIG.:** This dependogram identifies the 4-dependence in the last subset  $A = \{1, 2, 3, 4\}$ . No other dependencies were declared significant.  $n = 100$

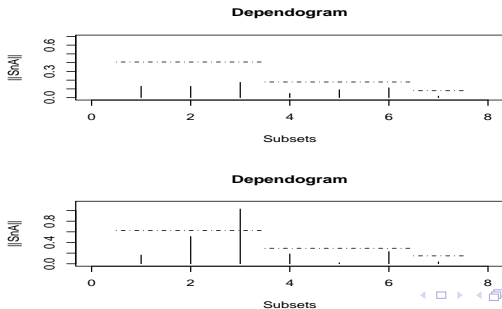


## Serial independence in a binary sequence (0 and 1)

- $W_i = \begin{cases} 0 & \text{with probability } 0.2 \\ 1 & \text{with probability } 0.8 \end{cases}$  iid
- $Y_i = W_i W_{i+3}, i = 1, \dots, n - 3$
- $Y_i$  which is dependent at lag 3

## Serial independence in a binary sequence (0 and 1)

**FIG.:** The upper dependogram does not declare any serial dependence in the i.i.d. sequence  $W_j$ . The lower dependogram for the sequence  $Y_j$  exhibits a serial dependence at lag 3 through the subset 3 corresponding to  $A = \{1, 4\}$ . The minimal value of  $p = 4$  was used.  $n = 100$ .

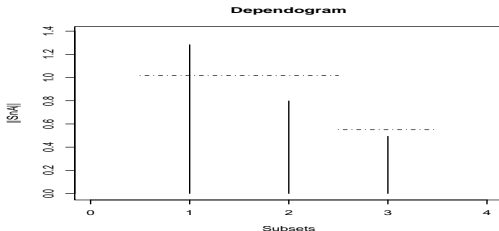


## Serial independence in directionnal data

- $U_j$  i.i.d.  $N_2(0, I_2)$
- $W_i = U_i + \sqrt{2}U_{i+1}, i = 1, \dots, n - 1$  with serial dependence at lag 1.
- $Y_i = W_i/|W_i|$  with serial dependence at lag 1 on the circle.

## Serial independence in directionnal data

**FIG.:** The dependogram for the angular gaussian sequence  $Y_i$  on the circle exhibits a serial dependence at lag 1 through the first subset corresponding to  $A = \{1, 2\}$ .  $n = 75$ .







## Multinomial formula

Let  $A$  be a non empty subset of  $\{1, 2, \dots, p\}$ . Then,

$$\sum_{B \subset A} \left( \prod_{i \in B} u^{(i)} \right) \left( \prod_{j \in A \setminus B} v^{(j)} \right) = \prod_{i \in A} (u^{(i)} + v^{(i)}).$$

# Bibliography

-  Blum, Kiefer, Rosenblatt, *The Annals of Mathematical Statistics* **32** (1961) 485–498.
-  Deheuvels, *Journal of Multivariate Analysis* **11** (1981) 102–113.
-  Genest, Rémillard, *Test* **13** (2004) 335–369.
-  Ghoudi, Kulperger, Rémillard, *Journal of Multivariate Analysis* **79** (2001) 191–218.