Package `subrank`: estimating copula density using ranks and sub-sampling

Jérôme Collet

March 31, 2016

1 Introduction

This package estimates the copula density of a d-dimensional random variable, without parametric assumptions, using ranks and subsampling. The main feature of this method is that simulation studies show a low sensitivity to dimension, on realistic cases.
This vignette provides:

- A description of the estimation method, in different ways.
- A description of the regression derived from the estimation method. In the case of regression, it is possible to estimate density on-the-fly, so computational burden is largely reduced.
- A convergence proof.
- Some hints on convergence speed. In the simple useless case of independence, we prove the convergence speed is the same as for kernel density estimation. In more structured cases, some numerical simulations show the impact of dimension is much smaller than for a kernel estimation.

It is usual to use ranks to estimate copulas [Fermanian, 2005], or test uniformity of the copula [Hoeffding, 1948, Kojadinovic and Holmes, 2009], since it gives straightforwardly uniform marginals. Nevertheless, with usual methods, we do not use all the features of ranks. They do not only have uniform marginals: their marginals are **exactly** uniform on $\{1, \cdots, n\}$ for **any** sample, and **any** dimension of a multidimensional variable. We have to use this regularization to reduce sampling errors. Furthermore, the smaller the sample is, the stronger the regularization is, so it is compulsory to sub-sample to adjust the strength of the regularization.
Sub-sampling is now widely used and studied, so we only cite the pioneering work [Breiman, 1996] and the survey [Bühlmann, 2012]. These methods are applied to **estimators**, so the theoretical results in these papers do not exactly fit our goal. Nevertheless, the main conclusion is that sub-sampling related methods are smoothing methods: their efficiency increases when applied to discontinuous, non-linear functions. Yet, the ranking of the observations is highly discontinuous and non-linear, so sub-sampling is relevant.

1.1 Implementation

This package is partly written in `C` (parallelized with `Open-MP` [OpenMP Architecture Review Board, 2010]), for reasons of efficiency. The random number generator is described in [Roy, 2006].

2 Description of the estimation method

Given a sample of size n, for a given sub-sample size $m < n$, one draws many sub-samples, without replacement. For each sub-sample and each observation, one obtains a vector of ranks (in

set $\{1,\cdots,m\}^d$). For each point $\mathbf{r}$ of $\{1,\cdots,m\}^d$, we propose to count the sub-samples where $\mathbf{r}$ appears (where there is an observation of which d ranks are the point $\mathbf{r}$). This counting is the estimator we propose, it converges in some ways to the copula density. We give in the following a formal description.

2.1 Mathematical formulation

We study a random variable $\mathbf{X} = (X_1,\cdots,X_d)$ in $\mathbb{R}^d$, with copula $c(\mathbf{u}) = c(u_1,\cdots,u_d)$; its marginals are assumed to be continuous.
We define

- $\mathcal{S}$ a sample of $\mathbf{X}$, $\mathbf{x}$ an element of $\mathcal{S}$,
- $\mathbf{R}(\mathbf{x},\mathcal{S})$ the d-uple of the ranks of the components of $\mathbf{x}$ in sample $\mathcal{S}$.

This allows defining a random variable $\mathbf{B}_\bullet(\mathcal{S})$: it is an array, filled with 0s and 1s, d dimensional, each dimension being indexed from 1 to $|\mathcal{S}|$ (where $||$ stands for cardinal). For a d-uple of ranks $\mathbf{r} \in \{1,\cdots,|\mathcal{S}|\}^d$, we define

$$\mathbf{B}_\mathbf{r}(\mathcal{S}) = \mathbf{1}\{\exists \mathbf{x} \in \mathcal{S} \mid \mathbf{R}(\mathbf{x},\mathcal{S}) = \mathbf{r}\}.$$

In other words, $\mathbf{B}_\mathbf{r}(\mathcal{S})$ is equal to 1 if and only if $\mathbf{r}$ is equal to $\mathbf{R}(\mathbf{x},\mathcal{S})$ for a $\mathbf{x}$ in $\mathcal{S}$.
The object we want to estimate is

$$P(|\mathcal{S}|,\mathbf{r},c) = \frac{1}{|\mathcal{S}|}\mathbb{E}\left(\mathbf{B}_\mathbf{r}(\mathcal{S})\right),$$

because we will show that this discrete array tends to the density copula c. Let us note that dividing by $|\mathcal{S}|$ makes the sum of P equal to 1, as $\forall \mathbf{r}, \forall \mathbf{X}, \sum_\mathbf{r} \mathbf{B}_\mathbf{r}(\mathcal{S}) = |\mathcal{S}|$.
In a practical setting, we have an n-sample $\mathcal{T}$, we choose a sub-sample size $m < n$, and we estimate $P(m,\mathbf{r},c)$ using the following U-statistic:

$$\hat{P}_n(m,\mathbf{r},\mathbf{X}) = \frac{1}{m\binom{n}{m}} \sum_{\mathcal{S}\subset\mathcal{T},|\mathcal{S}|=m} \mathbf{B}_\mathbf{r}(\mathcal{S}).$$

Remarks

- If $m = d = 2$, counting the sub-samples reaching a vector of ranks is very similar to the Kendall's τ computation. In such a case, $\{1,\cdots,m\}^d = \{(1,1),(1,2),(2,1),(2,2)\}$. A pair of concordant (resp. discordant) observations generate points $(1,1)$ and $(2,2)$ (resp. $(1,2)$ and $(2,1)$), so we have $\hat{\tau} = 2(\hat{P}(2,(1,1),\mathbf{X}) - \hat{P}(2,(1,2),\mathbf{X}))$.
- A simple example, in 2 dimensions, is the case $X_2 = f(X_1)$ with f strictly increasing. Then the only weighted points of $\{1,\cdots,m\}^2$ are on the diagonal. On the other hand, if all components are independent, a symmetry argument gives a uniform distribution on $\{1,\cdots,m\}^d$ (for n infinite, the case of a finite n is studied in the following).
- Most of the times $\binom{n}{m}$ will be far too large, so we will not be able to draw all sub-samples. We will use a random sub-sampling to obtain an approximation of $\hat{P}_n$.

2.2 Graphical description of the estimation method

The estimation method is described in figure 1.

2.3 Small completely detailed example

We propose a small completely detailed example. Table 1 is a sample in $\mathbb{R}^2$ (each observation is identified by a lowercase letter), we choose $n = 4$ and $m = 3$. Table 2 summarizes the computations.

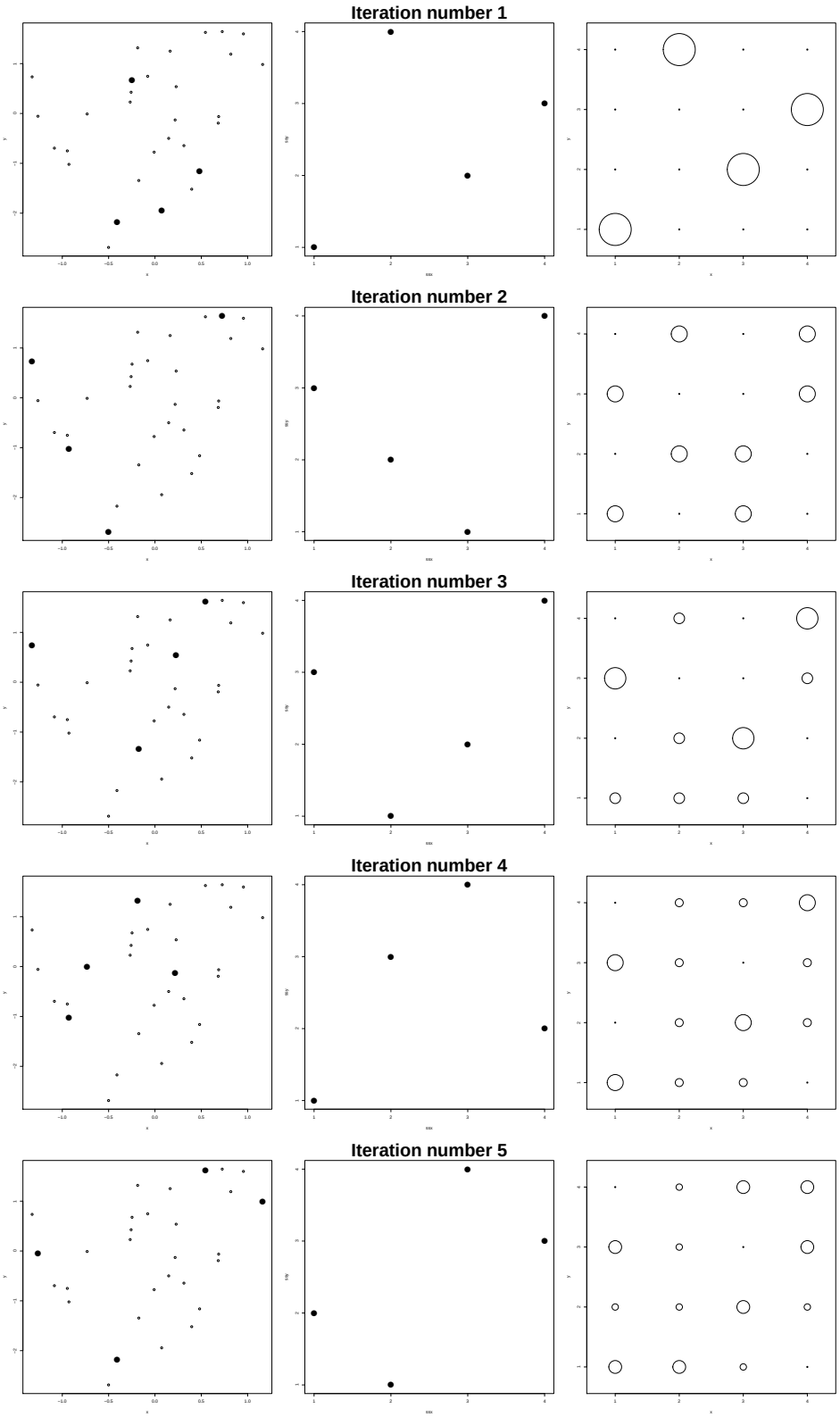

Figure 1: **First steps of the estimation**: pictures on the left represent original data with points of the sub-sample in black; pictures in the center represent the sub-sample in the rank space; radii of circles on the right represent, for each rank vector, current sum of the hits.

Table 1: **Example data**: R_X (resp R_Y) stands for the rank in X (resp. Y)

Observation	X	Y	R_X	R_Y
a	2.29	-0.97	4	1
b	-1.2	-0.95	1	2
c	-0.69	0.75	2	4
d	-0.41	-0.12	3	3

Table 2: **Computation of $\hat{P}_4(3, \mathbf{r}, \mathbf{X})$ for the data of Table 1**: the sub-samples are named: $A = \{a, b, c\}, B = \{a, b, d\}, C = \{a, c, d\}, D = \{b, c, d\}$. For example, bD in the first cell of the first line means that observation b in sub-sample B is the first one in X and in Y, and 1/12 is the value of $\hat{P}_4(3, (1, 1), \mathbf{X})$, since $12 = 3 \times \binom{4}{3}$

		Rank in X		
		1	2	3
Rank in Y	1	{bD}; 1/12	∅; 0/12	{aA,aB,aC}; 3/12
	2	{bA,bB}; 2/12	{dC}; 1/12	{dD}; 1/12
	3	{cC}; 1/12	{cA,dB,cD}; 3/12	∅; 0/12

2.4 Use for regression

Using Theorem 1, it is possible to use the estimation $\hat{P}_n$ to build a regression model. For a given value of some components of $\mathbf{x}$, we know the conditional copula. Using this, we may simulate values of the unknown components.

More precisely, we have a training set of n observations in dimension d, and we use it to estimate $\hat{P}_n$ (given a sub-sample size m), and d cumulative distribution functions. We also have a new observation $\mathbf{x}$, with only d' components known, we want to know the remaining components. The first step is to use the estimated CDFs to compute the $F_i(x_i)$'s (for the d' known components). Multiplying by m and taking the integer part gives a vector of ranks in $\{1, \cdots, m\}^{d'}$. Conditionally to this vector, we choose randomly a vector of ranks in $\{1, \cdots, m\}^{d-d'}$. Dividing by m gives the $F_i(x_i)$'s for the $d - d'$ unknown components of $\mathbf{x}$. We may smooth this prediction, adding a beta-distributed noise, or use another distribution. Finally, we use estimated CDFs, and their inverse, to compute the unknown components of $\mathbf{x}$. This result is obviously random, since the vector of ranks in $\{1, \cdots, m\}^{d-d'}$ is randomly chosen. So, it is a simulation method to provide a probabilistic forecast.

Prediction on-the-fly Doing so, computation time is very important, and memory requirements too. That is why we prefer to estimate the copula density on-the-fly, around the value of the known components. It reduces drastically the computation time and the memory requirements.

We consider the same training set, and the same incomplete observation. We draw a sub-sample with size m in the training set, we add the incomplete observation, and we replace each observation by its ranks. We look at whether, for each dimension, the rank of the new observation is the neighbor (differ by 1) of the rank of an observation of the sub-sample of the training set. In such a case, we use the value of this observation as a prediction (for the unknown components of the incomplete observation).

This operation is repeated many times, which gives a probabilistic forecast.

3 Convergence to the copula

Definition 1 *Let m be a strictly positive integer, and r an integer between 1 and $m+1$. Given m random variables U_i with distribution $U[0,1]$, the density $K_{r,m}$ is the density of the random variable $U \rightsquigarrow U[U_{(r-1)}, U_{(r)}]$, where $U_{(r)}$ denotes the value with rank r, $U_{(0)} = 0$ and $U_{(m+1)} = 1$.*

It is possible to express $K_{r,m}$:

$$K_{r,m}(k) = \frac{m!}{(r-2)^+!(m-r)^+!} \times \int_{0<u<k<v<1} \frac{u^{(r-2)^+}(1-v)^{(m-r)^+}}{(v-u)} dudv,$$

where $(r-2)^+$ stands for the positive part of $r-2$, in order to encompass all cases in one formula. These integrals seem to be difficult to write in a simpler form; even if $r = 1$, the integration is not easy. Nevertheless, it is possible to derive moments of this distribution:

$$\mathbb{E}_{K_{r,m}}(X) = \frac{2r-1}{2(m+1)} \simeq \frac{r}{m+1}$$
$$\mathbb{V}_{K_{r,m}}(X) = \frac{12r(m-r)+24r-7m-10}{12(m+1)^2(m+2)} \simeq \frac{r(m-r)}{(m+1)^2(m+2)}.$$

Theorem 1 *Assuming that $\mathbf{X} = (X_1, \cdots, X_d)$ in $\mathbb{R}^d$ has continuous marginals, and copula density c is bounded by M, we have:*

$$m^d \times |P(m, \mathbf{r}, c) - P_K(m, \mathbf{r}, c)| \leq \frac{2M^2d^2}{m} + O(m^{-2}),$$

with:

$$P_K(m, \mathbf{r}, c) = \int_{[0,1]^d} \left(\prod_{l=1}^{d} K_{r_l, m-1}(u_l) \right) c(\mathbf{u}) d\mathbf{u}.$$

4 Convergence speed

In the independent case, one shows that the AMISE is the same as for a kernel estimation. For more details, see B.

4.1 Use for independence test

We propose here to use independence testing to measure the accuracy of our estimation technique. To build a test, we need a test statistic: we use the Kullback–Leibler divergence, between $\hat{P}$ and the asymptotic value in case of independence, it means $P \equiv m^{-d}$. Then, we have to simulate many samples with independent components, which gives the distribution of the test statistic. Finally, we simulate samples with dependent components, and count the number of independence rejections.
In order to study the behavior of this testing method when the dimension increases, we will use as a test case a given dependence between a given number of components, all other components being independent. We studied three cases.
The first two are described by the same equation, with a parameter p switching from a monotonic dependence if equal to 1 to a non-monotonic if equal to 2:

$$\begin{array}{ll} X_2 = a \cdot X_1^p + \epsilon & \\ X_i \sim \mathcal{N}(0,1) & \text{if} \quad i \neq 2 \ . \\ \epsilon \sim \mathcal{N}(0,1) & \end{array}$$

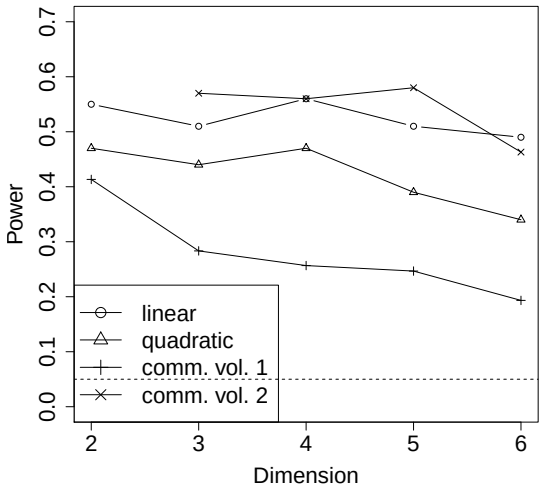

Figure 2: **Example of use for independence testing**: the sample size increases linearly with the dimension, so is equal to $15 \times d$. The sub-sample size is 8. The thin interrupted line represents the 5% level. The linear increasing of sample size is enough to obtain good power for each dimension.

The third dependence is the following model:

$$\begin{array}{lll} V \sim \mathcal{LN}(0, a) & & \\ X_i \sim \mathcal{N}(0, V) & \text{if} & i \leq d_d \\ X_i \sim \mathcal{N}(0, 1) & \text{if} & d_d < i \leq d \end{array}$$

where $\mathcal{LN}$ denotes the log-normal distribution. In other words, the first d_d components have the same random volatility. If a is large, the dependence is strong, and null if $a = 0$.
The main insight of this simulation study is: in order to maintain the same test power, it is enough to increase the sample size linearly w.r.t. dimension. Indeed, the sample size is here $15 \times d$, so it is 30 for dimension 2, ..., and 90 for dimension 6: this increase is enough to obtain good power for each dimension.
An additional point is the test power is never smaller than 1.1 times the power of the Deheuvels test, which is a widely used independence test. More precisely, The Deheuvels test shows a good power on the linear dependence, because it is a monotonic dependence. On others, its power is around 0.05, so the test is useless.

4.2 Forecasting competition

We used this method to participate in a forecasting competition Hong et al. [2014], with good results: the model, yet simple, finished tenth in the competition, among 250 competitors, with a forecast error 10% larger than the winner.
The topic of the probabilistic wind power forecasting track was to forecast the probabilistic distribution (in quantiles) of the wind power generation for 10 wind farms on a rolling basis. The target variable was power generation, normalized here by the respective nominal capacities of each wind farm. The explanatory variables that can be used are the past power measurements (if relevant), and input weather forecasts, given as u and v components (zonal and meridional). These forecasts were given at two heights, 10 m and 100 m above ground level, in order to give a rough idea of wind profiles.
Since it is possible to estimate directly the density for dimensions at around 6, we only had to

estimate the joint density of the load factor LF, and the 4 wind speeds u_{10}, v_{10}, u_{100}, v_{100}. In a second step, we simulated 1000 values of fc, knowing the values of the other variables.
This simple first model was a bit modified, because some temporal smoothing improved the results. In the final model, the 4 wind speeds are smoothed, with a window size equal to 5 time steps, and a set of weights derived from the triweight kernel. It appeared also useful to take into account the variability of the wind. To do so, a first step is to compute the absolute speed of the wind at each level. Then, one computes the local sum of squares of the speed, and it is added to the regressors of the load factor.

References

D.J. Best and P.G. Gipps. Algorithm AS 71: The Upper Tail Probabilities of Kendall's Tau. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 23(1):98–100, 1974.

Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

Peter Bühlmann. Bagging, boosting and ensemble methods. In *Handbook of Computational Statistics*, pages 985–1022. Springer-Verlag, 2012.

el Dairi, Khalil Antoine. Bornes sur des transformées de couples de variables aléatoires, 2005. URL `http://blogperso.univ-rennes1.fr/arthur.charpentier/public/maths/stage2A-2004.pdf`.

J.-D. Fermanian. Goodness-of-fit tests for copulas. *Journal of Multivariate Analysis*, 95(1):119–152, 2005. doi: DOI:10.1016/j.jmva.2004.07.004.

W. Hoeffding. A non-parametric test of independence. *Ann. Math. Statist.*, 19(4):546–557, 1948.

M. Hollander and D.A. Wolfe. *Nonparametric Statistical Methods*. Wiley, 1999.

Tao Hong, Shu Fan, Hamidreza Zareipour, Pierre Pinson, Alberto Troccoli, and Rob J Hyndman. Global Energy Forecasting Competition 2014, 2014. URL `http://www.drhongtao.com/gefcom`.

I. Kojadinovic and M. Holmes. Tests of independence among continuous random vectors based on Cramér-von Mises functionals of the empirical copula process. *Journal of Multivariate Analysis*, 100(6):1137–1154, 2009. ISSN 0047-259X. doi: DOI:10.1016/j.jmva.2008.10.013. URL `http://www.sciencedirect.com/science/article/B6WK9-4TTHW8V-1/2/3430550964db005a83d88a0094b588d3`.

OpenMP Architecture Review Board. The OpenMP API specification for parallel programming, 2010. URL `http://openmp.org/wp/`.

M. Petkovšek, H.S. Wilf, and D. Zeilberger. $A = B$. A K Peters, 1996. URL `http://www.math.upenn.edu/%7Ewilf/AeqB.pdf`.

J.-S. Roy. Randomkit: A library to generate random numbers, 2006. URL `http://js2007.free.fr/code/index.html#RandomKit`.

Paul D Valz and Mary E Thompson. Exact inference for Kendall's s and Spearman's ρ with extension to Fisher's exact test in r× c contingency tables. *Journal of Computational and Graphical Statistics*, 3(4):459–472, 1994.

K. Wegschaider. MultiSum, 1997. URL `http://www.risc.jku.at/research/combinat/risc/software/MultiSum/`.

H.S. Wilf. *generatingfunctionology*. Academic Press, 1994. URL `http://www.math.upenn.edu/~wilf/gfologyLinked2.pdf`.

A Proof of theorem 1

Without loss of generality, we assume the random variable $\mathbf{X}$ has uniform marginals. We will compute $P(m, \mathbf{r}, c)$, conditioning by: the rank $\mathbf{r}$ is reached by the last observation. Using symmetry, it would be the same for the m other observations. In order for the last observation to reach rank $\mathbf{r}$, one needs its first coordinate to be between $X_{1,(r_1-1,m-1)}$ and $X_{1,(r_1,m-1)}$, where $X_{1,(r_1-1,m-1)}$ denotes the value with rank $r_1 - 1$ among $m - 1$, for the first coordinate; and it is the same for other coordinates, so we get

$$P(m, \mathbf{r}, c) = \mathbb{E}\left(\int_{\bigcap [X_{l,(r_l-1,m-1)}; X_{l,(r_l,m-1)}]} c(\mathbf{u}) d\mathbf{u}\right). \tag{1}$$

We know that on each dimension, order statistics are Beta distributed. Then, if we have independence, the equality with P_K is obvious. If, for any couples (l, l') and $(\epsilon_l, \epsilon_{l'})$ $(\epsilon_l \in \{0, 1\})$, the values $X_{1,(r_1-\epsilon_l,m-1)}$ and $X_{1',(r_{l'}-\epsilon_{l'},m-1)}$ derive from different observations, they are independent and equality with P_K is true.
So we study the probability of coincidence: some of the values $X_{1,(r_1-\epsilon_l,m-1)}$ and $X_{1',(r_{l'}-\epsilon_{l'},m-1)}$ derive from the same observation. For a given couple (l, l'), we assume that $X_{l,(r_l-1,m-1)}$ is the value of observation 1 (without loss of generality). We want to know the probability that $X_{l',(r'_l-1,m-1)}$ is the value of observation 1 as well. To compute this probability, we have to consider the $m - 2$ remaining observations:

$$\mathbb{P}\left(X_{l',1} \in [X_{l',(r_{l'}-1,m-2)}, X_{l',(r_{l'},m-2)}]\right) \leq M \times \mathbb{E}\left(X_{l',(r_{l'},m-2)} - X_{l',(r_{l'}-1,m-1)}\right) = \frac{M}{m-2}.$$

If the couple (l, l') is not given, we know that, for each dimension, $X_{1,(r_1,m-1)}$ and $X_{1,(r_1-1,m-1)}$ are obviously from different observations. So we have to choose 2 dimensions, and for each the value of ϵ_l, so the probability of a coincidence is less than $2d(d-1) \times M/(m-2)$.
We need the following result, inspired from el Dairi, Khalil Antoine [2005]:

Proposition 1 *Let ϕ be a superadditive function, $\mathbf{X}$ a d-dimensional random variables with marginals F_i, Γ_F the set of the distributions with marginals F_i. Assuming $\phi(\mathbf{X})$ is uniformly integrable on Γ_F, we know, for $H \in \Gamma_F$*

$$\mathbb{E}_{\phi(\mathbf{X})}(H) \leq \mathbb{E}_{\phi(\mathbf{X})}\left(\bar{H}\right),$$

where $\bar{H}$ is the distribution with marginals F_i, and copula equal to the upper Fréchet-Hoeffding bound.

If we have e coincidences, for example on dimensions $1, \cdots, e$, we have to bracket

$$\mathbb{E}\left(\prod_{l=1}^{d}\left(X_{l,(r_l,m-1)} - X_{l,(r_l-1,m-1)}\right)\right).$$

A lower-bound is obviously 0, so we focus on upper-bound. We know:

$$\mathbb{E}\left(\prod_{l=1}^{d}\left(X_{l,(r_l,m-1)} - X_{l,(r_l-1,m-1)}\right)\right) = m^{-(d-e)}\mathbb{E}\left(\prod_{l=1}^{e}\left(X_{l,(r_l,m-1)} - X_{l,(r_l-1,m-1)}\right)\right).$$

In the product in the expectation of r.h.s., we know the distribution of each term, since it is the difference of 2 consecutive order statistics, but we do not know the joint distribution. Using 1, we know the worst case is the perfect correlation. So we have

$$\mathbb{E}\left(\prod_{l=1}^{e}\left(X_{l,(r_l,m-1)} - X_{l,(r_l-1,m-1)}\right)\right) \leq \mathbb{E}\left(\left(X_{1,(r_1,m-1)} - X_{1,(r_1-1,m-1)}\right)^e\right) =$$
$$\frac{e!m!}{(m+e)!} \leq e!m^{-e},$$

which gives

$$\mathbb{E}\left(\prod_{l=1}^{d}\left(X_{l,(r_l,m-1)}-X_{l,(r_l-1,m-1)}\right)\right)\leq e!m^{-d}.$$

To provide a global bound on $|P(m,\mathbf{r},c)-P_K(m,\mathbf{r},c)|$, we address only two cases: 1 coincidence, and more than 1. The probability of the first case has been computed previously.
For the latter, we know the number of choices of dimension l, interval side ϵ_l is finite, and does not depend on m. The probability of more than 1 coincidence, given the choice of dimensions l and interval sides ϵ_l is less than $(M/(m-2))^2$. For each coincidence, the contribution to the value of $\mathbb{E}\left(\prod_{l=1}^{d}\left(X_{l,(r_l,m-1)}-X_{l,(r_l-1,m-1)}\right)\right)$ is less than $d!m^{-d}$. So the total contribution is of order $O(m^{-2})$.
Then last step is taking into account the value of copula c, which is bounded by M. We finally get:

$$m^d\times|P(m,\mathbf{r},c)-P_K(m,\mathbf{r},c)|\leq\frac{2M^2d^2}{m}+O(m^{-2}).$$

B Independent case study

In the following, we will study

$$\hat{T}_n(m,\mathbf{X})=m^d\times\sum_{\mathbf{r}}\left(\hat{P}_n(m,\mathbf{r},\mathbf{X})-m^{-d}\right)^2,$$

where components of $\mathbf{X}$ are globally independent. More precisely, we study $\mathbb{V}\left(\hat{T}_n(m,\mathbf{X})\right)$.

Theorem 2 *Assuming that $\mathbf{X}=(X_1,\cdots,X_d)$ in $\mathbb{R}^d$ has continuous marginals, if the components are globally independent and if the sample size n tends to infinity,*

$$\begin{aligned}
n\mathbb{E}\left(\hat{T}_n(m,\mathbf{X})\right) &\rightarrow S_1^d-m^2+(m-1)^2\left(\tfrac{m^2-2m+S_1}{(m-1)^2}\right)^d+2(m-1)\left(\tfrac{m-S_1}{m-1}\right)^d\\
n^2\mathbb{V}\left(\hat{T}_n(m,\mathbf{X})\right) &\rightarrow 2S_2^d-2m^4+\\
&\quad 2(m-1)^4\left(\tfrac{m^4-4m^3+6m^2-4m+S_2}{(m-1)^4}\right)^d+12(m-1)^2\left(\tfrac{m^2-2m+S_2}{(m-1)^2}\right)^d+\\
&\quad 8(m-1)\left(\tfrac{m-S_2}{m-1}\right)^d+8(m-1)^3\left(\tfrac{m^3-3m^2+3m-S_2}{(m-1)^3}\right)^d
\end{aligned}$$

with

$$S_1=\frac{m4^{m-1}}{(2m-1)\binom{2m-2}{m-1}}\quad S_2=\frac{m^2\binom{4m-3}{2m-2}}{\left((2m-1)\binom{2m-2}{m-1}\right)^2}\ .$$

A numerical study shows that the first terms are the most important. Furthermore, it is possible to approximate these first terms using Stirling's formula. So we have

$$\mathbb{V}\left(\hat{T}_n(m,\mathbf{X})\right)\simeq 2n^{-2}\left(\sqrt{\tfrac{\pi m}{8}}\right)^d\quad\mathbb{E}\left(\hat{T}_n(m,\mathbf{X})\right)\simeq n^{-1}\left(\tfrac{\sqrt{\pi m}}{2}\right)^d\ .$$

We can check again the equivalence with Kendall's τ. We show $\hat{T}_n(2,\mathbf{X})=\tau^2$, so we check that $\mathbb{E}\left(\hat{T}_n(2,\mathbf{X})\right)\simeq\mathbb{V}(\tau)$ when n tends to infinity.

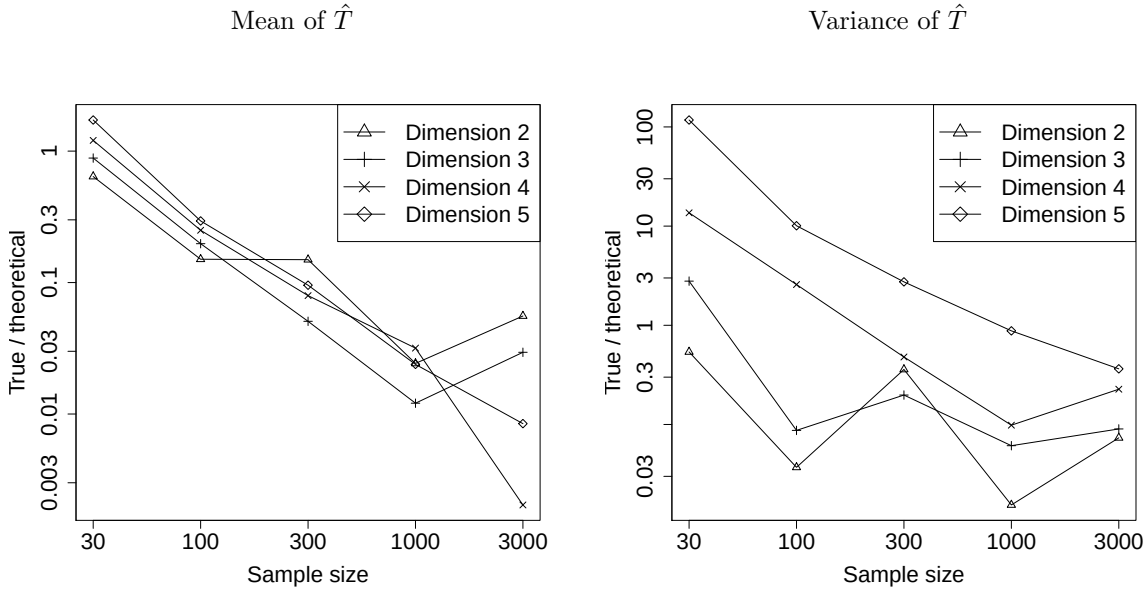

Figure 3: **Convergence of true value towards theoretical value**: for mean and variance of $\hat{T}(m, \mathbf{X})$, where components of $\mathbf{X}$ are independent, we study the convergence to 0 of absolute relative error $|\text{true/theoretical} - 1|$. Sub-sample size is 10, scale is logarithmic on both axes. The convergence becomes slower when dimension increases.

B.1 Numerical verification

We study by simulation the convergence of true value towards theoretical value, for mean and variance of $\hat{T}_n(m, \mathbf{X})$, where components of $\mathbf{X}$ are independent. We choose a sub-sample size equal to 10 and for each dimension in $\{2, 3, 4, 5\}$, we plot in figure 3 the absolute relative error of the theoretical $|\text{true/theoretical} - 1|$: it decreases rapidly when the sample size is between 30 and 3000. For all simulations, we choose a number of sub-samples equal to $n \times m^d$, and a number of samples equal to 100.

B.2 Evaluation of AMISE in the independent case

In the asymptotic case, this estimation behaves as a kernel estimation, with bandwidth of order $m^{-1/2}$. In the independent case, the variance is of order $m^{d/2}$.
So, **in the independent case**, the AMISE is the same as for a kernel estimation.
We think, without proof, that this case is the worst case for the method we propose. Indeed, with independence, the exclusion constraints (exactly 1 observation on each hyperplane of $\{1, \cdots, m\}^d$), are not very important. That is why we think in other cases it would show better performances than the kernel estimation. The simulation study corroborates this intuition.
Furthermore, if we use the optimal value of m, to compute T and its variance, then the convergence speed is of order $n^{(-d-8)/(d+4)}$, which makes it useful for independence testing.

B.3 Proof of Theorem 2

Addressing the behaviour of $\hat{P}_n(m, \mathbf{r}, \mathbf{X})$ is obviously a bit more difficult than addressing Kendall's τ, which is the case $m = d = 2$. Now, as far as we know, there is no simple formula for the probability distribution of Kendall's τ: neither Hollander and Wolfe [1999], nor the `cor.test` function documentation (in the package `stats`) in `R`, nor in the package `Kendall` of the same software, nor

the SAS documentation mention such a formula. There exist recursive ones in Best and Gipps [1974], Valz and Thompson [1994], which can be used only if the sample is small.
That is why we will derive here equivalents of the first two moments of T, for large values of the sample size.

B.3.1 The U-statistics: reminders and notations

Let h be a measurable function, symmetric in its n arguments. Then if we have a sample $X_1, \cdots, X_N$ with $n > m$, we define the U-statistic U_m:

$$U_n = \frac{1}{\binom{n}{m}} \sum_{S \subset \{1,\cdots,N\}}^{|S|=m} h\left(X_{S(1)}, \cdots, X_{S(m)}\right).$$

where $S(i)$ denotes the i^{th} element of S.
Furthermore, we define

$$h_c = \mathbb{E}\left(h\left(x_1, \cdots, x_c, X_{c+1}, \cdots, X_m\right)\right),$$

and

$$\sigma_c = \mathbb{V}\left(h_c\left(X_1, \cdots, X_c\right)\right).$$

Then $\mathbb{E}\left(U_m\right) = h_0$ and, when $n \to +\infty$, $n(U_m - h_0)$ converges in distribution to $\mathcal{N}\left(0, m^2\sigma_1\right)$ if $\sigma_1 \neq 0$.
Furthermore, if we have another U-statistic V_n defined by a kernel g, we may also define g_c and $\sigma_{c,c}$:

$$\sigma_{c,c} = \mathbb{C}ov\left(h_c\left(X_1, \cdots, X_c\right), g_c\left(X_1, \cdots, X_c\right)\right).$$

The covariance between U_n and V_n converges to $\sigma_{1,1}$ when $N \to +\infty$.
In the following, for each $\mathbf{r}$, we have a U-statistic $\hat{P}_n(m, \mathbf{r}, \mathbf{X})$, whose normal convergence we will use.
As we are in a slightly special case of U-statistics (we are only interested in the case $c = 1$, but we study a large number of U-statistics at the same time), one has to adapt the notations. We note:

$$\begin{aligned} h(\mathbf{r}, x_1) &= \mathbb{E}\left(B_{\mathbf{r}}\left(x_1, X_2 \cdots, X_n\right)\right) \\ \sigma(\mathbf{r}, \mathbf{s}) &= \mathbb{C}ov\left(h(\mathbf{r}, X_1), h(\mathbf{s}, X_1)\right). \end{aligned}$$

So we obtain

Proposition 2 *If $n \to \infty$, one has*

- $n\left(\hat{P}_n(m, \mathbf{r}, \mathbf{X}) - P(m, \mathbf{r}, c)\right)\Big)$ *converges in distribution to $\mathcal{N}\left(0, m^2\sigma(\mathbf{r}, \mathbf{r})\right)$; if $\sigma(\mathbf{r}, \mathbf{r}) \neq 0$,*
- $n\mathbb{C}ov\left(\hat{P}_n(m, \mathbf{r}), \hat{P}_n(m, \mathbf{s}, \mathbf{X})\right) \to m^2\sigma(\mathbf{r}, \mathbf{s})$.

This Central Limit Theorem allows us to use the following computation:

Proposition 3 *Let $\mathbf{X}$ be a vector such that$\mathbf{X} \rightsquigarrow \mathcal{N}(0, V)$, where the coefficients of V are denoted $\sigma(\mathbf{r}, \mathbf{s})$. We study $D = \sum_{\mathbf{r}} X_{\mathbf{r}}^2$. Then*

$$\mathbb{E}\left(D\right) = \sum_{\mathbf{r}} \sigma(\mathbf{r}, \mathbf{r}) \qquad \mathbb{V}\left(D\right) = 2\sum_{\mathbf{r},\mathbf{s}} \sigma(\mathbf{r}, \mathbf{s})^2.$$

B.3.2 Calculation of the individual covariances

We calculate in the same way the variances and covariances. We consider the first observation, its vector of ranks, and a given rank $\mathbf{r}$. There are 3 cases:

- Other observations are such that the vector of ranks of the first observation is $\mathbf{r}$,
- Other observations are such that the vector of ranks of the first observation is equal to $\mathbf{r}$ for some dimensions,
- Other observations are such that the vector of ranks of the first observation is different from $\mathbf{r}$ for all dimensions.

We know the probability that the l^{th} coordinate of the first observation reaches rank r_l:

$$P = \binom{m-1}{r_l-1} x_{1,l}^{r_l-1}(1-x_{1,l})^{m-r_l} = b_{m-1,r_l-1}(x_{1,l}),$$

where b_{m-1,r_l-1} is a Bernstein polynomial, with well-known properties, for example

$$\int_0^1 b_{m,r}(x)dx = \frac{1}{m+1}.$$

We use this to calculate the probability of each one of the three cases:

$$\mathbb{P}\left(B_{\mathbf{r}}\left(x_1, X_2 \cdots, X_n\right) = \frac{1}{m}\right) =$$
$$1 \times \prod_{l=1}^{d} b_{m-1,r_l-1}(x_{1,l}) + 0 \times \left(1 - \prod_{l=1}^{d}\left(1 - b_{m-1,r_l-1}(x_{1,l})\right) - \prod_{l=1}^{d} b_{m-1,r_l-1}(x_{1,l})\right) +$$
$$\frac{m-1}{(m-1)^d} \times \prod_{l=1}^{d}\left(1 - b_{m-1,r_l-1}(x_{1,l})\right).$$

One can remark that integrating this probability over $\mathbf{x}_1$ gives back the unconditional probability

$$\mathbb{P}\left(B_{\mathbf{r}}\left(X_1, X_2 \cdots, X_n\right) = \frac{1}{m}\right) =$$
$$1 \times \int_{[0,1]^d} \prod_{l=1}^{d} b_{m-1,r_l-1}(x_{1,l})d\mathbf{x}_1 + \frac{1}{(m-1)^{d-1}} \times \int_{[0,1]^d} \prod_{l=1}^{d}\left(1 - b_{m-1,r_l-1}(x_{1,l})\right)d\mathbf{x}_1 =$$
$$1 \times \frac{1}{m^d} + \frac{1}{(m-1)^{d-1}} \times \left(1 - \frac{1}{m}\right)^d = \frac{1}{m^{d-1}}.$$

The conditional mean is

$$h(\mathbf{r}, \mathbf{x}_1) = \frac{1}{m}\mathbb{E}\left(B_{\mathbf{r}}\left(\mathbf{x}_1, X_2 \cdots, X_n\right)\right) =$$
$$\frac{1}{m} \times \left(\prod_{l=1}^{d} b_{m-1,r_l-1}(x_{1,l}) + \frac{1}{m(m-1)^{d-1}} \times \prod_{l=1}^{d}\left(1 - b_{m-1,r_l-1}(x_{1,l})\right)\right).$$

So, we have to calculate

$$\sigma(\mathbf{r}, \mathbf{s}) = \mathbb{E}\left(\left[h(\mathbf{r}, \mathbf{X}_1) - \mathbb{E}\left(h(\mathbf{r}, \mathbf{X}_1)\right)\right] \times \left[h(\mathbf{s}, \mathbf{X}_1) - \mathbb{E}\left(h(\mathbf{s}, \mathbf{X}_1)\right)\right]\right) =$$
$$\int_{[0,1]^d} \begin{matrix}\left[\frac{1}{m}\prod_{l=1}^{d} b_{m-1,r_l-1}(x_{1,l}) + \frac{1}{m(m-1)^{d-1}}\prod_{l=1}^{d}\left(1 - b_{m-1,r_l-1}(x_{1,l})\right) - \frac{1}{m^d}\right] \times \\ \left[\frac{1}{m}\prod_{l=1}^{d} b_{m-1,s_l-1}(x_{1,l}) + \frac{1}{m(m-1)^{d-1}}\prod_{l=1}^{d}\left(1 - b_{m-1,s_l-1}(x_{1,l})\right) - \frac{1}{m^d}\right]\end{matrix} d\mathbf{x}_1.$$

For these covariance computations, we will calculate expressions such as

$$\int_0^1 b_{m,r}(x)b_{m,s}(x)dx = \frac{\binom{m}{r}\binom{m}{s}}{\binom{2m}{r+s}} \times \frac{1}{2m+1}.$$

We note

$$\mathcal{A}(m,r,s) = \frac{\binom{m}{r}\binom{m}{s}}{\binom{2m}{r+s}}.$$

We multiply two sums of three terms. We denote by $D_{i,j}$ the product of terms numbered i in the first sum and j in the second one.

$$\begin{array}{rcl}
D_{1,1} & = & \int_{[0,1]^d} \left[\frac{1}{m} \times \prod_{l=1}^d b_{m-1,r_l-1}(x_{1,l})b_{m-1,s_l-1}(x_{1,l})\right] dx \\
& = & \frac{1}{m^2(2m-1)^d} \times \prod_{l=1}^d \mathcal{A}(m-1,r_l-1,s_l-1) \\
D_{2,2} & = & \int_{[0,1]^d} \left[\frac{1}{m^2(m-1)^{2d-2}} \times \prod_{l=1}^d \left((1-b_{m-1,r_l-1}(x_{1,l}))(1-b_{m-1,s_l-1}(x_{1,l}))\right)\right] dx \\
& = & \frac{1}{m^2(m-1)^{2d-2}} \times \prod_{l=1}^d \left(1 - \frac{2}{m} + \frac{\mathcal{A}(m-1,r_l-1,s_l-1)}{2m-1}\right) \\
D_{3,3} & = & \int_{[0,1]^d} \left[\frac{1}{m^d}\right]^2 dx \\
& = & \frac{1}{m^{2d}} \\
D_{1,2} & = & \int_{[0,1]^d} \frac{1}{m^2(m-1)^{d-1}} \times \prod_{l=1}^d \left(b_{m-1,r_l-1}(x_{1,l}) - b_{m-1,r_l-1}(x_{1,l})b_{m-1,s_l-1}(x_{1,l})\right)dx \\
& = & \frac{1}{m^2(m-1)^{d-1}} \times \prod_{l=1}^d \left(\frac{1}{m} - \frac{\mathcal{A}(m-1,r_l-1,s_l-1)}{2m-1}\right) \\
D_{1,3} & = & \int_{[0,1]^d} \left[\frac{1}{m} \times \prod_{l=1}^d b_{m-1,r_l-1}(x_{1,l}) \times \frac{1}{m^d}\right] dx \\
& = & \frac{1}{m^{2d+1}} \\
D_{2,3} & = & 2\int_{[0,1]^d} \left[\frac{1}{m^{d+1}(m-1)^{d-1}} \times \prod_{l=1}^d \left(1 - b_{m-1,r_l-1}(x_{1,l})\right)\right] dx \\
& = & \frac{m-1}{m^{2d+1}}
\end{array}$$

It is clear that $D_{1,2} = D_{2,1}$, $D_{1,3} = D_{3,1}$ and $D_{2,3} = D_{3,2}$, so we write simply $2D_{1,2}$, etc. We remark that

$$D_{3,3} - 2D_{1,3} - 2D_{2,3} = -m^{-2d}.$$

B.3.3 Calculation of the sum of the covariances

We know

$$\sigma(\mathbf{r},\mathbf{s}) = D_{1,1} + D_{2,2} + D_{3,3} + 2D_{1,2} - 2D_{1,3} - 2D_{2,3} = D_{1,1} + D_{2,2} + 2D_{1,2} - m^{-2d},$$

so

$$\mathbb{E}\left(\hat{T}_n(m,\mathbf{X})\right) = \sum_{\mathbf{r}} \sigma(\mathbf{r},\mathbf{r}) = \sum_{\mathbf{r}} \left(D_{1,1}(\mathbf{r},\mathbf{r}) + D_{2,2}(\mathbf{r},\mathbf{r}) + 2D_{1,2}(\mathbf{r},\mathbf{r})\right) - m^{-d}.$$

In a similar way

$$\mathbb{V}\left(\hat{T}_n(m,\mathbf{X})\right) = 2\sum_{\mathbf{r},\mathbf{s}} \sigma^2(\mathbf{r},\mathbf{s}) = 2\sum_{\mathbf{r},\mathbf{s}} \left(D_{1,1}(\mathbf{r},\mathbf{s}) + D_{2,2}(\mathbf{r},\mathbf{s}) + 2D_{1,2}(\mathbf{r},\mathbf{s}) - m^{-2d}\right)^2 =$$

$$2\sum_{\mathbf{r},\mathbf{s}} \left(\begin{array}{l} D_{1,1}^2(\mathbf{r},\mathbf{s}) + D_{2,2}^2(\mathbf{r},\mathbf{s}) + 4D_{1,2}^2(\mathbf{r},\mathbf{s}) + m^{-4d} \\ +2D_{1,1}D_{2,2} + 4D_{1,1}D_{1,2} + 4D_{2,2}D_{1,2} \\ -2m^{-2d}\left[D_{1,1}(\mathbf{r},\mathbf{s}) + D_{2,2}(\mathbf{r},\mathbf{s}) + 2D_{1,2}(\mathbf{r},\mathbf{s})\right] \end{array}\right) =$$

$$2\sum_{\mathbf{r},\mathbf{s}} \left(D_{1,1}^2(\mathbf{r},\mathbf{s}) + D_{2,2}^2(\mathbf{r},\mathbf{s}) + 6D_{1,2}^2(\mathbf{r},\mathbf{s}) + 4D_{1,1}D_{1,2} + 4D_{2,2}D_{1,2}\right) - 2m^{-2d}.$$

We will need to know some sums involving $\mathcal{A}(m,r,s)$, they are proved in B.4.1.

$$\sum_r \mathcal{A}(m,r,s) = \frac{2m+1}{m+1}, \quad \sum_r \mathcal{A}(m,r,r) = \frac{4^m}{\binom{2m}{m}}, \quad \sum_{r,s} (\mathcal{A}(m,r,s))^2 = \frac{\binom{4m+1}{2m}}{(\binom{2m}{m})^2}$$

The sums of the terms $D_{i,j}$ are sums of products, we transform them easily into products of sums. For example

$$\sum_{\mathbf{r},\mathbf{s}} D_{1,1}^2 = \frac{1}{m^4(2m-1)^{2d}} \times \sum_{\mathbf{r},\mathbf{s}} \left(\prod_{l=1}^{d} \mathcal{A}^2(m-1, r_l-1, s_l-1) \right) =$$

$$\frac{1}{m^4(2m-1)^{2d}} \times \left(\sum_{r,s} \mathcal{A}^2(m-1, r-1, s-1) \right)^d .$$

All of these sums (3 sums of degree 1 with $\mathbf{r} = \mathbf{s}$, 3 sums of degree 1, 3 sums of squares, and 3 sums of double products) are calculated in B.4.2. Using these sums, the proof of Theorem 2 is obvious.

B.4 Tools for the proof of Theorem 2

B.4.1 Combinatorial computations

We need to compute some sums involving $\mathcal{A}(m,r,s)$. We first remark:

$$\mathcal{A}(m,r,s) = \frac{\binom{m}{r}\binom{m}{s}}{\binom{2m}{r+s}} = \frac{\binom{r+s}{r}\binom{2m-r-s}{m-r}}{\binom{2m}{m}}.$$

We show

Proposition 4 *If $m > 1$ and $0 \le s \le m$:*

$$\sum_r \binom{r+s}{r}\binom{2m-r-s}{m-r} = \binom{2m+1}{m}$$

$$\sum_r \binom{2r}{r}\binom{2m-2r}{m-r} = 4^m$$

$$\sum_{r,s} \left(\binom{r+s}{r}\binom{2m-r-s}{m-r} \right)^2 = \binom{4m+1}{2m}.$$

These sums are very similar to convolutions, so it is interesting to use generating functions [Wilf, 1994]. They are quite simple for the first series:

$$\sum_n \binom{n+k}{n} x^n = \frac{1}{(1-x)^{k+1}}$$

$$\sum_n \binom{2n}{n} x^n = \frac{1}{\sqrt{1-4x}}.$$

It is also possible to demonstrate the first identity using combinatorial arguments, counting the number of paths joining two opposite corners of a rectangle with sides m and $m+1$. The last identity is more difficult: one needs to use the powerful tools developed in Petkovšek et al. [1996]. As the sum is over 2 variables, one needs to use the package `multisum` [Wegschaider, 1997]. The code proving identity is

```
Get["C:\Users\Jerome\Desktop\Celine\MultiSum.m"]
FindRecurrence[ (Binomial[r+s,r]*Binomial[2*m-r-s,m-r])^2,
  m, {r,s}, 4 ]
SumCertificate[%]
CheckRecurrence[ %, Binomial[4*m+1,2*m] ]
```

B.4.2 Sums of covariance terms

We summarize here the sums of terms $D_{i,j}$. These sums are all computed in the same way, and they are compulsory to check the other computations.

$$\begin{array}{lcl}
\sum_{\mathbf{r}} D_{1,1}(\mathbf{r},\mathbf{r}) & = & \frac{1}{m^2} \times R_1^d \\
\sum_{\mathbf{r}} D_{2,2}(\mathbf{r},\mathbf{r}) & = & \left(\frac{m-1}{m}\right)^2 \times \left(\frac{m-2+R_1}{(m-1)^2}\right)^d \\
\sum_{\mathbf{r}} D_{1,2}(\mathbf{r},\mathbf{r}) & = & \frac{m-1}{m^2} \times \left(\frac{1-R_1}{m-1}\right)^d \\
\sum_{\mathbf{r},\mathbf{s}} D_{1,1} & = & \frac{1}{m^2} \\
\sum_{\mathbf{r},\mathbf{s}} D_{2,2} & = & \left(\frac{m-1}{m}\right)^2 \\
\sum_{\mathbf{r},\mathbf{s}} D_{1,2} & = & \frac{m-1}{m^2} \\
\sum_{\mathbf{r},\mathbf{s}} D_{1,1}^2 & = & \frac{1}{m^4} \times R_2^d \\
\sum_{\mathbf{r},\mathbf{s}} D_{2,2}^2 & = & \left(\frac{m-1}{m}\right)^4 \times \left(\frac{m^2-4m+6-\frac{4}{m}+R_2}{(m-1)^4}\right)^d \\
\sum_{\mathbf{r},\mathbf{s}} D_{1,2}^2 & = & \frac{(m-1)^2}{m^4} \times \left(\frac{1-\frac{2}{m}+R_2}{(m-1)^2}\right)^d \\
\sum_{\mathbf{r},\mathbf{s}} D_{1,1}D_{2,2} & = & \sum_{\mathbf{r},\mathbf{s}} D_{1,2}^2 \\
\sum_{\mathbf{r},\mathbf{s}} D_{1,1}D_{1,2} & = & \frac{m-1}{m^4} \times \left(\frac{\frac{1}{m}-R_2}{m-1}\right)^d \\
\sum_{\mathbf{r},\mathbf{s}} D_{2,2}D_{1,2} & = & \frac{(m-1)^3}{m^4} \times \left(\frac{m-3+\frac{3}{m}-R_2}{(m-1)^3}\right)^d
\end{array},$$

with

$$R_2 = \frac{\binom{4m-3}{2m-2}}{\left((2m-1)\binom{2m-2}{m-1}\right)^2} \quad R_1 = \frac{4^{m-1}}{(2m-1)\binom{2m-2}{m-1}}.$$