Random permutations and their discrepancy process

Guillaume Chapuy

LIX,École polytechnique, Palaiseau, France

received 26 Fev 2006, revised 23rd January 2008, accepted .

Let σ be a random permutation chosen uniformly over the symmetric group $\mathfrak{S}_n$. We study a new ”process-valued” statistic of σ, which appears in the domain of computational biology to construct tests of similarity between ordered lists of genes. More precisely, we consider the following ”partial sums”:

$$Y^{(n)}_{p,q} = \mathrm{card}\{1 \le i \le p \ : \ \sigma_i \le q\} \ \text{ for } 0 \le p, q \le n$$

We show that a suitable normalization of $Y^{(n)}$ converges weakly to a bivariate tied down brownian bridge on $[0,1]^2$, i.e. a continuous centered gaussian process $X^\infty_{s,t}$ of covariance:

$$\mathbb{E}\left[X^\infty_{s,t} X^\infty_{s',t'}\right] = (min(s,s') - ss')(min(t,t') - tt')$$

Keywords: Weak convergence, Bivariate Brownian bridge, Random permutations

1 Introduction

On the one hand, random permutations appear as a natural model for data in various topics, such as analysis of algorithms, statistical mechanics, or genomic statistics. On the other hand, from the mathematical point of view, various parameters of random permutations have been studied in combinatorics and probability, like the distribution of the lengths of cycles, the repartition of the eigenvalues on the unit circle([Wie00]), or the length of large monotonic subsequences (see [BDJ99, AD99]). The description of the typical behaviour of these statistics provide powerful tools for applications, like the construction of statistical tests, or precise analysis of the execution time of algorithms. A lot of examples of permutations statistics and their application to analysis of algorithms can be found all along the book [FS].

The purpose of this paper is to describe a new ”process-valued” statistic of random permutations, recently introduced in the context of computational biology ([YBSS06]) to extract information from the large amounts of data produced by the microarray technology. Roughly speaking, a DNA microarray provides, for a given list of genes (say $1, \dots n$), an ordering (i.e., a permutation) $\tau_1, \dots, \tau_n$, of the genes by their expression level in a given experimental condition. Here, the typical value of n is 1000. In [YBSS06], the following problem is addressed: being given two such orderings τ and τ' corresponding to two classes of experiences (say on healthy and sick patients), is it possible to quantify the similarity between them? To measure this similarity, the following numbers are introduced: for each $p \leq n$, let O_p be the number of genes $i \in [\![1, n]\!]$ that appear in the first p ranks of both lists. Then, an heuristic which appears to be efficient on applications is the following: measuring "how far" the list $(O_p)_{p=1..n}$ is from its expected value on independant permutations allows to compute a relevant "similarity score" between the lists. We will not discuss what "how far" means here: complex heuristics arguments are used in [YBSS06] to compute the final score from the values (O_p). Instead we will give a mathematical description of the expected behaviour of the process (O_p) (which in [YBSS06] was obtained by simulations) when n is large, under the assumption that both lists are uniform independant random permutations.

More precisely, let n be a positive integer, and let $\mathfrak{S}_n$ denote the symmetric group on $[\![1, n]\!]$. Let $\sigma = (\sigma_1, \ldots, \sigma_n)$ be a random permutation chosen uniformly on $\mathfrak{S}_n$ (i.e. such that for all ν in $\mathfrak{S}_n$, $\mathbb{P}(\sigma = \nu) = 1/n!$). From σ, we contruct the following random variable $Y^{(n)}$:

$$Y^{(n)}_{p,q} = \mathrm{card}\{i \leq p \ : \ \sigma_i \leq q\} = \mathrm{card}\left(\sigma([\![1,p]\!]) \cap [\![1,q]\!]\right) \quad (0 \leq p, q \leq n)$$

Remark that if $\sigma = \tau' \circ \tau^{-1}$, $Y^{(n)}_{p,p}$ is exactly the number O_p. More, if $I = [\![p, p']\!]$ and $J = [\![q, q']\!]$ are two integer sub-intervals of $[\![1, n]\!]$, one deduces from $Y^{(n)}$ the cardinality of $\sigma(I) \cap J$ as follows :

$$|\sigma(I) \cap J| = Y^{(n)}_{p',q'} - Y^{(n)}_{p',q-1} - Y^{(n)}_{p-1,q'} + Y^{(n)}_{p-1,q-1}$$

The maximum deviation of the previous quantity from its typical value over all possible choices of intervals I and J is called the *discrepancy* of the permutation σ:

$$\mathrm{disc}(\sigma) = \max_{I,J} \left\{ \left| |\sigma(I) \cap J| - \frac{|I||J|}{n} \right| \right\}$$

This quantity and its applications to pseudo-random permutations have been studied in detail by Cooper (see [Coo04]). Our purpose here is to study a related but rather different object. We investigate the asymptotic behaviour of the *discrepancy process of the permutation* σ, which is the process defined over $[\![0, n]\!]^2$ as:

$$Z^{(n)}_{p,q} = Y^{(n)}_{p,q} - \mathbb{E}\left[Y^{(n)}_{p,q}\right]$$

where we will see that $\mathbb{E}\left[Y^{(n)}_{p,q}\right] = \frac{pq}{n}$.

In this purpose, we define the *normalized discrepancy process* $(X^{(n)}_{s,t})_{0 \leq s,t \leq 1}$ as follows:

- for $s, t \in [0, 1]$ such that sn and st are integers, put $X^{(n)}_{s,t} = \dfrac{Z^{(n)}_{sn,tn}}{\sqrt{n}} = \dfrac{Y^{(n)}_{sn,tn} - stn}{\sqrt{n}}$
- complete the process $(X^{(n)}_{s,t})$ in such a way that it is continuous and affine on each closed "lattice triangle" of vertices $\left\{(\frac{k}{n}, \frac{l}{n}), (\frac{k+1}{n}, \frac{l}{n}), (\frac{k+1}{n}, \frac{l+1}{n})\right\}$ or $\left\{(\frac{k}{n}, \frac{l}{n}), (\frac{k}{n}, \frac{l+1}{n}), (\frac{k+1}{n}, \frac{l+1}{n})\right\}$

Our main result is the following theorem:

Theorem 1 *The normalized discrepancy process $(X^{(n)}_{s,t})_{s,t \in [0,1]^2}$ converges, in the sense of weak convergence on $\mathcal{C}([0, 1]^2)$ to a process $X^{\infty}_{s,t}$. This process has the law of a bivariate tied down brownian bridge, which is a centered continuous gaussian process on $[0, 1]^2$ of covariance:*

$$\mathbb{E}\left[X^{\infty}_{s,t} X^{\infty}_{s',t'}\right] = (min(s, s') - ss')(min(t, t') - tt')$$

Remark that this limit is natural, in view of the ”permutational symmetry” of the model: the bivariate tied down Brownian Bridge is one of the simplest cases of continuous bivariate processes with exchangeable increments, according to the complete classification of the paper [Kal88]. However, our approach to identify the limit and show the convergence is elementary and does not involve the theory of exchangeable random variables.

The rest of the paper is organised as follows: in section 2, we compute the first moments of Y, in section 3, we recall the some facts about the theory of bivariate continuous processes, in section 4, we give a key theorem (Theorem 4) on product moments from which we deduce tightness, and in section 5, we compute the finite-dimensional limit laws.

2 First moments.

We begin with the first order moments of $Y^{(n)}$: expectation and variance.

Since $Y_{p,q}^{(n)} = \sum_{i=1}^{p} \mathbb{1}_{\sigma_i \le q}$, the linearity of expectation gives:

$$\mathbb{E}[Y_{p,q}] = \sum_{i=1}^{p} \mathbb{E}\left[\mathbb{1}_{\sigma_i \le q}\right] = \sum_{i=1}^{p} \mathbb{P}\left(\sigma_i \le q\right) = \frac{pq}{n} \tag{1}$$

Indeed, for any fixed i, the uniform measure on $\mathfrak{S}_n$ for σ induces the uniform measure on $\{1, \dots, n\}$ for σ_i, so $\mathbb{P}\left(\sigma_i \le q\right) = \frac{q}{n}$.

In the same spirit, we can get the second order moment:

$$\begin{aligned}
\mathbb{E}[{Y_{p,q}^{(n)}}^2] &= \sum_{i=1}^{p}\sum_{j=1}^{p} \mathbb{E}\left[\mathbb{1}_{\sigma_i \le q\,,\,\sigma_j \le q}\right] \\
&= \sum_{i \ne j} \mathbb{P}\left(\sigma_i \le q \text{ and } \sigma_j \le q\right) + \sum_{i} \mathbb{P}\left(\sigma_i \le q\right) \\
&= p(p-1) \times \frac{q(q-1)}{n(n-1)} + p \times \frac{q}{n}
\end{aligned}$$

The variance follows:

$$\begin{aligned}
\mathbb{VAR}[Y_{p,q}^{(n)}] &= \mathbb{E}[{Y_{p,q}^{(n)}}^2] - \mathbb{E}[Y_{p,q}^{(n)}]^2 \\
&= \frac{pq(n-p)(n-q)}{n^2(n-1)}
\end{aligned}$$

Remark that if $p = sn$ and $q = tn$, the variance grows linearly with n, which justifies the normalization factor $1/\sqrt{n}$ we chose for $X^{(n)}$.

3 Continuous random processes

In this section, we recall some facts about the theory of continuous random processes on $[0,1]^2$. The material of this section is taken from [BW71], wich generalises to a multidimensional time the classical tightness criterion of [Bil68].

Let $\mathcal{C} = \mathcal{C}\left([0,1]^2, \mathbb{R}\right)$ be the space of continous functions on the unit square, equipped with the uniform norm $\|f\|_\infty = \max_{s,t} |f(s,t)|$, and the induced $Borel\ \sigma - field$.

We recall that a family $(\mu_i)_{i\in I}$ of probability measures on $\mathcal{C}$ is said to be *tight* if for every $\epsilon > 0$ there exists a compact set $K \subset \mathcal{C}$ such that $\mu_i(K) \geq 1 - \epsilon$ for all i. We say that a family of processes is tight if the family of their laws is tight.

The following theorem is a classical result:

Theorem 2 (Prohorov) *Let P, P_n be probability measures on $\mathcal{C}$. If the finite-dimensional distributions of P_n converge weakly to those of P, and if $(P_n)_{n\geq 0}$ is tight, then (P_n) converges weakly to P.*

In order to formulate the next theorem, we need some definitions. A *bloc* B is a subset of $[0,1]^2$ of the form $(s,s'] \times (t,t']$. Its area $(s'-s)(t'-t)$ will be denoted $\mathcal{A}(B)$. Given a bloc B and an element X of $\mathcal{C}$ let

$$X(B) = X(s',t') + X(s,t) - X(s,t') - X(s',t)$$

be the *increment of X around B*.

Then we have the following:

Theorem 3 (Tightness criterion, [BW71]) *Let $W^{(n)}_{s,t}$ be a sequence of random processes in $\mathcal{C}$ such that for all $n \geq 0$ on has: $\mathbb{P}\left(W^{(n)}_{s,0} = 0 \, for\ all\ s \in [0,1]\right) = 1$.*

Suppose that there exist real numbers $\beta_1, \beta_2, \gamma_1, \gamma_2$ and $c > 0$ such that for all blocs B and C one has for all $n \geq 0$:

$$\mathbb{E}\left[|W^{(n)}(B)|^{\gamma_1} |W^{(n)}(C)|^{\gamma_2}\right] \leq c\mathcal{A}(B)^{\beta_1}\mathcal{A}(C)^{\beta_2}$$

with $\beta_1 + \beta_2 > 1$ and $\gamma_1 + \gamma_2 > 0$.

Then the sequence of processes $X^{(n)}$ is tight.

4 Product moments

In this section, we generalize the computation we made for the second order moment to any product moment. We begin with a lemma.

Lemma 1 *Let $0 \leq a_1 \leq a_2 \leq \ldots \leq a_k \leq n$ be k integers, and $b_1, b_2, \ldots b_k$ be k* **distinct** *values in $[\![1,n]\!]$. Then:*

$$\mathbb{P}\left(\sigma_{b_1} \leq a_1,\, \sigma_{b_2} \leq a_2, \ldots, \sigma_{b_k} \leq a_k\right) = \frac{a_1(a_2-1)(a_3-2)\ldots(a_k-k+1)}{n(n-1)(n-2)\ldots(n-k+1)}$$

Proof: The law of $\sigma_{b_{i+1}}$ given the values $\sigma_{b_1}, \ldots \sigma_{b_i}$ is the uniform law on $[\![1,n]\!] \setminus \{\sigma_{b_1}, \ldots \sigma_{b_i}\}$. Thus we have:

$$\mathbb{P}\left(\sigma_{b_{i+1}} \leq a_{i+1} \,|\, \sigma_{b_1} \leq a_1,\, \sigma_{b_2} \leq a_2, \ldots, \sigma_{b_i} \leq a_i\right) = \frac{a_{i+1} - i}{n - i}$$

and the lemma follows. □

Theorem 4 *Let* $0 \leq p_1 \leq p_2 \leq \ldots \leq p_k \leq n$*,* $0 \leq q_1 \leq q_2 \leq \ldots \leq q_k \leq n$ *and* $0 \leq \alpha_1, \ldots \alpha_k$ *be integers. Let* $\pi \in \mathfrak{S}_k$ *be a permutation of* $[\![1,k]\!]$*. For* $1 \leq i \leq k$*, define:* $I_i = \{i\} \times [\![1, \alpha_i]\!]$ *and put* $I = \bigcup_{i=1}^{k} I_i$*.*

To any partition P *of* I*, associate the following numbers* $r_i(P)$ *and* $s_i(P)$ *defined for* $1 \leq i \leq k$*:*

$$\begin{aligned} r_i(P) &= card\{S \in P \,:\, S \cap (I_1 \cup \ldots \cup I_{i-1}) \neq \emptyset\} && (2) \\ s_i(P) &= card\{S \in P \,:\, S \cap (I_{\pi^{-1}_1} \cup \ldots \cup I_{\pi^{-1}_{i-1}}) \neq \emptyset\} \;\; (\textit{note that } r_1(P) = s_1(P) = 0\,) && (3) \end{aligned}$$

Then we have:

$$\mathbb{E}\left[{Y^{(n)}_{p_1,q_{\pi_1}}}^{\alpha_1} {Y^{(n)}_{p_2,q_{\pi_2}}}^{\alpha_2} \ldots {Y^{(n)}_{p_k,q_{\pi_k}}}^{\alpha_k}\right] = \sum_P \prod_{i=1}^{k} \prod_{j=s_i(P)}^{s_{i+1}(P)-1} \frac{(q_i - j)}{(n-j)} \prod_{j=r_i(P)}^{r_{i+1}(P)-1} (p_i - j) \qquad (4)$$

where the sum is taken over all partitions P *of* I*.*

The meaning of this theorem is the following: for each generic position of the points (i.e., for each permutation π), and each $\alpha_1, \ldots \alpha_k$, the associated product moment is a computable rational function in n and the p_i, q_i. For example, let us compute $\mathbb{E}\left[Y^{(n)}_{p,q} Y^{(n)}_{p',q'}\right]$ for $p < p', q < q'$. In this case, I has cardinality two, and there are two partitions to consider. The partition with one equivalency class gives the term $\frac{pq}{n}$, whereas the one with two classes gives $\frac{pq(p'-1)(q'-1)}{n(n-1)}$. Hence we have:

$$\mathbb{E}\left[Y^{(n)}_{p,q} Y^{(n)}_{p',q'}\right] = \frac{pq}{n} + \frac{pq(p'-1)(q'-1)}{n(n-1)} \qquad (5)$$

In the case where $p < p'$ and $q' < q$, one gets:

$$\mathbb{E}\left[Y^{(n)}_{p,q} Y^{(n)}_{p',q'}\right] = \frac{pq'}{n} + \frac{pq'(p'-1)(q-1)}{n(n-1)} \qquad (6)$$

We now prove the theorem. **Proof Proof of Theorem 4:** Since $Y^{(n)}_{p,q} = \sum_{i \leq p} \mathbb{1}_{\sigma_i \leq q}$ we have:

$$\begin{aligned} &\mathbb{E}\left[{Y^{(n)}_{p_1,q_{\pi_1}}}^{\alpha_1} {Y^{(n)}_{p_2,q_{\pi_2}}}^{\alpha_2} \ldots {Y^{(n)}_{p_k,q_{\pi_k}}}^{\alpha_k}\right] \\ = &\underbrace{\sum_{i^{(1)}_1=1}^{p_1} \cdots \sum_{i^{(1)}_{\alpha_1}=1}^{p_1}}_{\alpha_1 \text{ sums}} \underbrace{\sum_{i^{(2)}_1=1}^{p_2} \cdots \sum_{i^{(2)}_{\alpha_2}=1}^{p_2}}_{\alpha_2 \text{ sums}} \cdots \underbrace{\sum_{i^{(k)}_1=1}^{p_k} \cdots \sum_{i^{(k)}_{\alpha_k}=1}^{p_k}}_{\alpha_k \text{ sums}} \mathbb{P}\left(\sigma_{i^{(l)}_j} \leq q_{\pi_l} \text{ for all } l \text{ and } j \leq \alpha_l\right) \end{aligned}$$

Now, fix a choice of indices $(i^{(l)}_j)_{l \leq k, j \leq \alpha_l}$. This induces a partition P of I: (l,j) and (k,m) are in the same class if and only if $i^{(l)}_j = i^{(k)}_m$. Let $r = \mathrm{card}(P)$ be the number of classes in P. We define the following quantities:

- For each class $U \in P$, let

  $$c(U) = \min\{i,\ U \cap I_i \neq \emptyset\}.$$

  Define $(\theta(i),\ i = 1..r)$ as the only non-decreasing sequence of numbers such that the following equality between multisets holds :

  $$\{\{c(U),\ u \in P\}\} = \{\{\theta(1), \dots, \theta(r)\}\}.$$

  In the same way, define

  $$d(U) = \min\{i,\ U \cap I_{\pi_i^{-1}} \neq \emptyset\}$$

  and let $(\mu(i),\ i = 1..r)$ be the only non-decreasing sequence of numbers such that

  $$\{\{d(U),\ u \in P\}\} = \{\{\mu(1), \dots, \mu(r)\}\}$$

- For $U \in P$, let v_U be the common value of the indices on the class U.

Since the q_i form an increasing sequence, the strongest condition imposed on the value of σ on a class U is to be $\leq q_{d(U)}$. Hence we have:

$$\mathbb{P}\left(\sigma_{i_j^{(l)}} \leq q_l \text{ for all } l \text{ and } j \leq \alpha_l\right) = \mathbb{P}\left(\sigma_{v_U} \leq q_{d(U)} \text{ for all } U\right)$$

From Lemma 1 and the definition of the $\mu(i)$, we have:

$$\mathbb{P}\left(\sigma_{v_U} \leq q_{d(U)} \text{ for all } U\right) = \frac{q_{\mu(1)}(q_{\mu(2)} - 1) \dots (q_{\mu(r)} - r + 1)}{n(n-1)\dots(n-r)} \tag{7}$$

Now, if the partition P is fixed, the number of possible choices of indices corresponding to P is $p_{\theta(1)}(p_{\theta(2)} - 1) \dots (p_{\theta(r)} - r + 1)$. Indeed, this corresponds to the choice of the values v_U, with the constraint $v_U \leq p_{c(U)}$. Thus the quantity (7) appears exactly $p_{\theta(1)}(p_{\theta(2)} - 1) \dots (p_{\theta(r)} - r + 1)$ times, and we have:

$$\mathbb{E}\left[{Y_{p_1,q_1}^{(n)}}^{\alpha_1} {Y_{p_2,q_2}^{(n)}}^{\alpha_2} \dots {Y_{p_k,q_k}^{(n)}}^{\alpha_k}\right] = \sum_P \prod_{j=1}^{r} (p_{\theta(j)} - j + 1) \frac{(q_{\mu(j)} - j + 1)}{n - j + 1} \tag{8}$$

In this last formula, we have to identify the values of the $\theta(j)$ and $\mu(j)$. But, from the definition of $r_i(P)$ and $c(U)$, we have:

$$\text{card}\{U \in P,\ c(U) = j - 1\} = r_j(P) - r_{j-1}(P)$$

from which we deduce:

$$\{i,\ \theta(i) = j\} = \left[\!\left[\sum_{i=1}^{j} r_i(P), \sum_{i=1}^{j+1} r_i(P) - 1 \right]\!\right] \tag{9}$$

A similar argument shows that

$$\{i,\ \mu(i) = j\} = \left[\!\left[\sum_{i=1}^{j} s_i(P), \sum_{i=1}^{j+1} s_i(P) - 1 \right]\!\right] \tag{10}$$

Putting together Equations (8), (9), and (10) gives Equation (4). □

In order to apply the tightness criterion (Theorem 3), we now prove the following theorem:

Theorem 5 *There exists a constant $c > 0$ such that for every n, and for every blocs B and C, one has:*

$$\mathbb{E}\left[Z^{(n)}(B)^2 Z^{(n)}(C)^2\right] \leq cn^2 \mathcal{A}(B)\mathcal{A}(C)$$

The proof of the theorem needs the following lemma, which will be proved in section 5.

Lemma 2 *For every $\epsilon \in \left(0, \frac{1}{2}\right)$ and every $r \geq 0$, there exists a constant c such that for all $n \geq 0$ and for all bloc B, one has:*

$$\mathbb{E}\left[Z^{(n)}(B)^{2r}\right] < cn^{r+\epsilon}$$

Proof Proof of Theorem 5:

Up to lower error terms, we can assume that B and C are "lattice blocs", i.e. $B = (\frac{p_1}{n}, \frac{q_1}{n}] \times (\frac{p_2}{n}, \frac{q_2}{n}]$ and $C = (\frac{r_1}{n}, \frac{s_1}{n}] \times (\frac{r_2}{n}, \frac{s_2}{n}]$, where p_i, q_i, r_i, s_i are integers.

Then, the quantity $\mathbb{E}\left[Z^{(n)}(B)^2 Z^{(n)}(C)^2\right]$ can be expanded as a finite sum involving only product moments of blocs of order less than four. From here, we have to distinguish a finite number of cases, depending on the respective positions of the blocs (corresponding to different values of the permutation π in Theorem 4). For each of these cases, we obtain from Theorem 4:

$$\mathbb{E}\left[Z^{(n)}(B)^2 Z^{(n)}(C)^2\right] = \frac{P(p_1, q_1, p_2, q_2, r_1, r_2, s_1, s_2)}{Q(n)}$$

were P and Q are polynomials.

Since the expectancy vanishes for $p_i = q_i$ or $r_i = s_i$, there exists a polynomial R such that:

$$\mathbb{E}\left[Z^{(n)}(B)^2 Z^{(n)}(C)^2\right] = (q_1 - p_1)(q_2 - p_1)(s_1 - r_1)(s_2 - t_2)\frac{R(p_1, q_1, p_2, q_2, r_1, r_2, s_1, s_2)}{Q(n)} \quad (11)$$

$$= n^4 \mathcal{A}(B)\mathcal{A}(C)\frac{R(p_1, q_1, p_2, q_2, r_1, r_2, s_1, s_2)}{Q(n)} \quad (12)$$

Now, by the Cauchy-Schwartz inequality and the lemma, we have, for some fixed $\epsilon < \frac{1}{2}$ and $c > 0$:

$$\mathbb{E}\left[Z^{(n)}(B)^2 Z^{(n)}(C)^2\right] \leq \left(\mathbb{E}\left[Z^{(n)}(B)^4\right]\mathbb{E}\left[Z^{(n)}(C)^4\right]\right)^{1/2} \leq c^2 n^{2+\epsilon} \quad (13)$$

Equations (12) and (13) implie that:

$$n^2 \mathcal{A}(B)\mathcal{A}(C)R(p_1, q_1, p_2, q_2, r_1, r_2, s_1, s_2) \leq c^2 Q(n) n^{\epsilon} \quad (14)$$

Thus, if $deg(R)$ denotes the total degree of the polynomial R, we have: $deg(R) + 2 \leq deg(Q)$.

This implies that for $p_i, q_i, r_i, s_i \in [\![0, n]\!]$ the quantity

$$\frac{R(p_1, q_1, p_2, q_2, r_1, r_2, s_1, s_2)n^2}{Q(n)}$$

is uniformly bounded. This and Equation (12) complete the proof of the theorem. □

From the last theorem and the tightness criterion (Theorem 3), we have:

Corollary 1 *The sequence of processes $X^{(n)}_{(s,t)\in[0,1]^2}$ is tight.*

5 Finite-dimensional laws.

In this section, we compute the finite dimensional laws, and we prove Lemma 2. Our approach is the following: compute discrete probabilities and prove a local limit law by the use of Stirling's formula.

Let $1 \leq p \leq q \leq n$ and $1 \leq l \leq m \leq n$. Let us consider the random vector $A^{(n)} = (A_i^{(n)})_{1 \leq i \leq 4} \in \mathbb{R}^4$ defined by:

$$\begin{array}{rcl} A_1^{(n)} & = & Y^{(n)}\left([1,p] \times [1,l]\right) \\ A_2^{(n)} & = & Y^{(n)}\left([1,p] \times [l+1,m]\right) \\ A_3^{(n)} & = & Y^{(n)}\left([p+1,q] \times [1,l]\right) \\ A_4^{(n)} & = & Y^{(n)}\left([p+1,q] \times [l+1,m]\right) \end{array}$$

The vector $A^{(n)}$ counts the number of points in the four regions described in Figure (1).

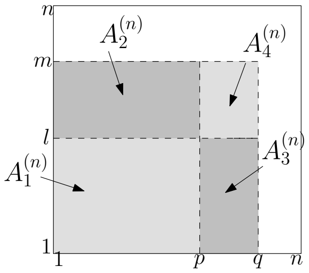

Fig. 1: The four regions involved in the definition of $A^{(n)}$

We have the following lemma.

Lemma 3 *Let k_1, k_2, k_3, k_4 be integers. Then:*

$$\mathbb{P}\left(A^{(n)} = (k_1, k_2, k_3, k_4)\right) = \binom{p}{k_1, k_2}\binom{q-p}{k_3, k_4}\binom{n-q}{l-k_1-k_3, m-l-k_2-k_4}\binom{n}{l, m-l}^{-1} \tag{15}$$

Proof: All the permutations such that $A = (k_1, k_2, k_3, k_4)$ can be constructed as follows:

- choose the k_1 numbers in $[\![1,p]\!]$ whose images will be in $[\![1,l]\!]$, and the k_2 ones whose images will be in $[\![l+1,m]\!]$. This corresponds to the first multinomial coefficient in Equation (15).

- do the analogue for the intevals $[\![p+1,q]\!]$ and $[\![q+1,n]\!]$. This corresponds to the next two multinomial coefficients.

- choose the effective values of the images among the $l!(m-l)!(n-m-l)!$ possible choices.

Since the probability of any given permutation is $\frac{1}{n!}$, the lemma is proved. □

Fix $0 \leq a \leq a+b \leq 1$, and $0 \leq c \leq c+d \leq 1$, and put:

$$\begin{aligned} p &= an \\ q &= (a+b)n \\ l &= cn \\ m &= (c+d)n \end{aligned}$$

Then one has $\mathbb{E}\left[A\right] = (acn, adn, bcn, bdn)$. Let us put:

$$\begin{aligned} k_1 &= acn + \alpha_1\sqrt{n} \\ k_2 &= adn + \alpha_2\sqrt{n} \\ k_3 &= bcn + \alpha_3\sqrt{n} \\ k_4 &= bdn + \alpha_4\sqrt{n} \end{aligned}$$

where $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ are fixed in a given compact set K of $\mathbb{R}^4$. From now on, we will use the notation O_K instead of O when the involved constant depends only on K.

Then one has the following lemma:

Lemma 4

$$\mathbb{P}\left(A = (k_1, k_2, k_3, k_4)\right) = \frac{(2\pi n)^{-2}}{\Sigma} e^{-Q(\alpha_1, \alpha_2, \alpha_2, \alpha_4)} \left(1 + O_K\left(\frac{1}{\sqrt{n}}\right)\right) \tag{16}$$

where Q is a positive quadratic form of the α_i, and $\Sigma = abcd(1-a-b)(1-c-d)$.

Proof: We recall Stirling's formula:

$$n! = \left(\frac{n}{e}\right)^n \sqrt{2\pi n}\left(1 + O\left(\frac{1}{n}\right)\right)$$

Then an easy calculation leads to:

$$\mathbb{P}\left(A = (k_1, k_2, k_3, k_4)\right) = \frac{(2\pi n)^{-2}}{\Sigma} \prod_{(X,Y)\in J_+} \left(\frac{X}{Y}\right)^Y \prod_{(X,Y)\in J_-} \left(\frac{X}{Y}\right)^{-Y} \left(1 + O_K\left(\frac{1}{\sqrt{n}}\right)\right)$$

where J_+ designates the set of all pairs of variables (X, Y) such that $\dfrac{X!}{Y!}$ appears in one of the first three multinomials of Equation (15), and J_- is the analogue for the fourth one.

Now, if $X = An$ and $Y = Bn + \epsilon\sqrt{n}$, one has:

$$\begin{aligned} \left(\frac{X}{Y}\right)^Y &= \left(\frac{A}{B}\right)^Y \left(\frac{1}{1 + \epsilon B^{-1} n^{-1/2}}\right)^{Bn + \epsilon\sqrt{n}} \\ &= \left(\frac{A}{B}\right)^Y \exp\left(\epsilon\sqrt{n} + \frac{\epsilon^2 B^{-1}}{2}\right)\left(1 + O_K\left(\frac{1}{\sqrt{n}}\right)\right) \end{aligned}$$

Now the lemma follows from the following observations:

- The product of all terms of the form $\left(\frac{A}{B}\right)^Y$ equals 1. Indeed, among the 12 elements in $J_+ \cup J_-$, A/B takes each of the value $c^{-1}, d^{-1}, (1-c-d)^{-1}$ exactly four times, and the sum of contributions vanishes in each case (for example, the value $A/B = c^{-1}$ appears with the exponent $k_1 + k_3 + (l - k_1 - k_3) - l = 0$).
- The product of the terms of the form $\exp(\epsilon\sqrt{n})$ equals 1. Indeed, the sum of the values of ϵ over all pairs is 0.

□

A slight modification of the last proof leads to the proof of Lemma 2, that we sketch now.

Proof Proof of Lemma 2 (sketch): Let us fix $\epsilon < 1/2$. We first prove that there exist positive constants D, D' such that for all $\alpha, \beta \in [0,1]$, one has:

$$\mathbb{P}\left(Z^{(n)}{}_{\alpha n, \beta n} = n^{\frac{1}{2}+\epsilon}\right) \leq De^{-D'n^{2\epsilon}}$$

This follows from these two obervations:

- One can assume that $\min(\alpha n, \beta n, (1-\alpha)n, (1-\beta)n) \geq n^{\frac{1}{2}+\epsilon}$ (otherwise the probability is zero).
- It is easily checked from the proof of the last lemma that if the α_i go to infinity with n slower than n^ϵ, and if $\min(an, bn, cn, dn, (1-a-b)n, (1-c-d)n) \geq n^{\frac{1}{2}}$, then Equation (16) is still valid, with uniform error term.

Now, fix n, α, and β and let $p(k) = \mathbb{P}\left(Z^{(n)}{}_{\alpha n, \beta n} = k\right)$. Then, Equation (15) and the study of the difference $p(k+1) - p(k)$ show that the function $k \mapsto p(k)$ is decreasing on $[\![k_0, n]\!]$, for some $k_0 = k_0(n) = O(1)$.

Hence, for n large enough, one has, for every $k \geq n^{\frac{1}{2}+\epsilon} : p(k) \leq p(n^{\frac{1}{2}+\epsilon})$ This implies:

$$\begin{aligned}\mathbb{E}\left[Z^{(n)}{}_{\alpha n, \beta n}{}^{2r}\right] &\leq \sum_{k \leq n^{\frac{1}{2}+\epsilon}} p(k)k^{2r} + \sum_{k > n^{\frac{1}{2}+\epsilon}} p(k)k^{2r} \\ &\leq n^{r+2\epsilon} + n^{2r+1}e^{-D'n^{2\epsilon}} \\ &\leq D''n^{r+2\epsilon}\end{aligned}$$

where D'' only depends on r and ϵ. This proves the lemma. □

Since the error term in Lemma 4 is uniform on compact sets, the limit density is the density of a probability measure, and the local limit law implies the limit in distribution. Thus we have proved:

Theorem 6 *The random vector $\frac{1}{\sqrt{n}}(A - \mathbb{E}[A])$ converges in distribution to a quadrivariate gaussian vector of covariance Q.*

Corollary 2 *The vector $(X^{(n)}_{s,t}, X^{(n)}_{s',t'})$ converges in distribution to a centered gaussian vector of covariance:*

$$\mathbb{E}\left[X^\infty_{s,t} X^\infty_{s',t'}\right] = (min(s,s') - ss')(min(t,t') - tt')$$

Proof: Here again, we have to distinguish cases, depending on the respective positions of $s, s't, t'$. In each case, $(X^{(n)}_{s,t}, X^{(n)}_{s',t'})$ is a linear combination of A_1, A_2, A_3, A_4, so Theorem 6 implies that it converges to a gaussian vector.

Then all we have to do is determine the covariance. From Equations (5) and(1), one has, for $p < p'$ and $q < q'$:

$$\mathbb{E}\left[X^{(n)}_{p/n,q/n}, X^{(n)}_{p'/n,q'/n}\right] = \frac{pq(n-p')(n-q')}{n^3(n-1)}$$

and taking a limit concludes the proof in the case $s < s'$, $t < t'$.

The other case is analogue, using Equations (6) and(1). □

The next corollary will be our final step:

Corollary 3 (Finite-dimensional laws) *Let $k, l \geq 2$ be integers.*

Then for all $0 = s_0 < s_1 < \ldots < s_k < 1$ and $0 = t_0 < t_1 < \ldots < t_l < 1$, the vector $\left(X^{(n)}_{t_i,s_j}\right)_{1\leq i\leq k\ ,\ 1\leq j\leq l}$ converges in distribution to a centered gaussian vector T of covariance

$$\mathbb{E}[T_{i,i'}T_{j,j'}] = (min(t_i, t_{i'}) - t_i t_{i'})(min(s_j, s_{j'}) - s_j s_{j'})$$

Proof: We prove by induction on $k + l$ that the corollary holds, with a local limit law (i.e. convergence of the discrete probability function to the density, uniformly on compact sets).

The case $k + l \leq 4$ is known from Lemma 4.

Let k and l such that $k + l > 4$, and assume that the result is true for every k' and l' such that $k' + l' < k + l$. By symmetry, we can assume that $k > 2$.

Let us define:

$$\begin{aligned} M_{i,j} &= Z^{(n)}\left([s_i, s_{i+1}] \times [t_j, t_{j+1}]\right) \text{ for } 0 \leq i < k, 0 \leq j < l \\ L_{1,j} &= M_{0,j} + M_{1,j} + \ldots + M_{k-2,j} \text{ for } 0 \leq j < l \\ L_{2,j} &= M_{k-1,j} \end{aligned}$$

Then we have the two following facts, the second being a consequence of the induction hypothesis:

- the law of $(L_{2,j})_{0\leq j<l}$ given $(M_{i,j})_{0\leq i<k-1,0\leq j<l}$ depends only on $(L_{1,j})_{0\leq j<l}$.
- both vectors $(L_{i,j})_{1\leq i\leq 2,0\leq j<l}$ and $(M_{i,j})_{0\leq i<k-1,0\leq j<l}$ have a gaussian local limit law.

Expressing the density function of $(M_{i,j})_{0\leq i<k,0\leq j<l}$ from these two facts easily shows that it has a gaussian local limit law, and concludes the induction. □

From Corollaries 1, 3 and the Prohorov theorem (Theorem 2), we have proved Theorem 1.

Acknowledgements

The author is grateful to Pierre Nicodeme and Rainer Spang for communicating and discussing this problem and to Gilles Schaeffer and Eric Fusy for interesting discussions and comments.

References

[AD99] David Aldous and Persi Diaconis. Longest increasing subsequences: from patience sorting to the Baik-Deift-Johansson theorem. *Bull. Amer. Math. Soc. (N.S.)*, 36(4):413–432, 1999.

[BDJ99] Jinho Baik, Percy Deift, and Kurt Johansson. On the distribution of the length of the longest increasing subsequence of random permutations. *J. Amer. Math. Soc.*, 12(4):1119–1178, 1999.

[Bil68] Patrick Billingsley. *Convergence of probability measures*. John Wiley & Sons Inc., New York, 1968.

[BW71] P. J. Bickel and M. J. Wichura. Convergence criteria for multiparameter stochastic processes and some applications. *Ann. Math. Statist.*, 42:1656–1670, 1971.

[Coo04] Joshua N. Cooper. Quasirandom permutations. *J. Combin. Theory Ser. A*, 106(1):123–143, 2004.

[FS] Philippe Flajolet and Robert Sedgewick. *Analytic Combinatorics*. preliminary version available on the web : http://algo.inria.fr/flajolet/Publications.

[Kal88] Olav Kallenberg. Some new representations in bivariate exchangeability. *Probab. Theory Related Fields*, 77(3):415–455, 1988.

[Wie00] Kelly Wieand. Eigenvalue distributions of random permutation matrices. *Ann. Probab.*, 28(4):1563–1587, 2000.

[YBSS06] Xinan Yang, Stefan Bentink, Stefanie Scheid, and Rainer Spang. Similarities of ordered gene lists. *J. Bioinformatics and Computational Biology*, 4(3):693–708, 2006.