Orbit-based conditional tests. A link between permutations and Markov bases

Submitted to JSPI
arXiv: arXiv:1707.08513

Orbit-based conditional tests. A link between permutations and Markov bases

ROBERTO FONTANA [1,*]

[1]*Department of Mathematical Sciences, Politecnico di Torino, Italy.*
E-mail: [*] roberto.fontana@polito.it
and

FRANCESCA ROMANA CRUCINIO [2,**]

[2]*Department of Statistics, University of Warwick, United Kingdom.*
E-mail: [**] f.crucinio@warwick.ac.uk

Supplementary material

Appendix A: Properties of the Graph

In this appendix we focus on the description of the graph $\mathcal{G}$ induced by the Markov basis $\mathcal{B}_N$ over the fiber $\mathcal{F}_{N,t}$. First we give two formulae to compute the number of vertices $|V|$, that corresponds to the cardinality of the fiber $\mathcal{F}_{N,t}$, and the number of edges $|E|$ in $\mathcal{G}$. Then we prove that the graph is bipartite and we give a way to compute the number of orbits of permutations contained in the fiber $\pi \subseteq \mathcal{F}_{N,t}$.

The computation of the cardinality of $\mathcal{F}_{N,t}$ can be seen as the problem of distributing t resources among N users, when users may get no resources. This is a well known exercise in combinatorics which gives $|V| = \binom{t+N-1}{N-1}$ (Lovász *et al.*, 2006, Theorem 3.4.2).

The number of edges can be computed observing that every edge corresponds to the addition/subtraction of a move $\mathbf{m} \in \mathcal{B}_N$, therefore we just have to check which moves are admissible for a generic vertex $\mathbf{y} \in \mathcal{F}_{N,t}$.

The set of vertices can be divided into three subsets:

- the internal vertices, i.e. the vectors with no component equal to 0;
- the vertices corresponding to vectors with $y_1 \neq 0$ and $1 \leq z < N$ components equal to zero;
- the vertices corresponding to vectors with $y_1 = 0$ and $1 \leq z^* < N - 1$ additional components equal to zero.

Consider the first subset, i.e. the internal vertices. This set has cardinality $\binom{t-1}{N-1}$ and for each vertex in this set every move $\mathbf{m}_K$ $1 \leq K \leq N - 1$ with every sign $\varepsilon = \pm 1$ is admissible. This is a consequence of the absence of entries equal to 0, which means that we can add or subtract 1 from every entry. Thus each vertex in this set has $2(N - 1)$ edges.

Secondly, consider the set of vertices with z zero components and $y_1 \neq 0$; this set has cardinality given by the number of possible vectors with sum of entries equal to t and z zero components $\binom{t-1}{N-1-z}$ times the possible positions for the z zero components $\binom{N-1}{z}$. For the vertices in this set the $2(N-1-z)$ moves which do not involve the z zero components are admissible and within the ones which involve the zero components only the z moves with $\varepsilon = +1$ are admissible. Thus every vertex in this set has $2(N-1-z)+z = 2N-2-z$ edges.

Finally, consider the set of vertices with $y_1 = 0$ and z^* additional null components and denote the total number of zero components $z = z^*+1$. The cardinality of this set is given by the product between the number of possible vectors with sum of entries equal to t and z zero components $\binom{t-1}{N-1-z}$ and the possible positions for the z^* additional zero components $\binom{N-1}{z^*} = \binom{N-1}{z-1}$. For the vertices in this set $\varepsilon = +1$ is the only admissible sign and if $\varepsilon = +1$ the moves involving the z^* zero components are not admissible; therefore each vertex in this set has $N-1-z^* = N-z$ edges.

Thus the total number of edges is given by the sum of these three terms

$$
\begin{aligned}
&2(N-1)\binom{t-1}{N-1}+ \\
&\sum_{z=1}^{N-1}(2N-2-z)\binom{t-1}{N-1-z}\binom{N-1}{z}+ \\
&\sum_{z=1}^{N-1}(N-z)\binom{t-1}{N-1-z}\binom{N-1}{z-1}
\end{aligned}
$$

divided by two (because by counting the edges of each vertex we count the same edge twice).

To prove that $\mathcal{G}$ is bipartite we observe that it is not possible to return to the starting vector by an odd sequence of moves: consider the first component y_1 of a generic vector $\mathbf{y} \in \mathcal{F}_{N,t}$ and a generic path of moves. Every move acts on y_1 with a $+1$ or a -1. To come back to y_1 the sequence of $+1$ and -1 has to be even. This proves that $\mathcal{G}$ has no cycle of odd length, hence the graph is bipartite.

Finally, the number of orbits of permutation contained in the fiber is equivalent to the number of partitions of t into N or fewer parts. This number is $\mathrm{part}(t, N)$, where part is the partition function defined in Kunz (2006) and Wilf (2000). The values of the partition function can be computed using the recurrence $\mathrm{part}(t, N) = \mathrm{part}(t, N-1) + \mathrm{part}(t-N, N)$ and depend on both the sample size N and the sum of entries t.

Appendix B: Markov basis for the new parametrisation of $\mathcal{F}_{N,t}$

In this section we give a way to build the Markov basis for the parametrisation of $\mathcal{F}_{N,t}$ in equation (4.3), given the value of the sum of entries t. A similar construction is presented in Chen *et al.* (2005).

The linear sufficient statistic for this parametrisation of $\mathcal{F}_{N,t}$ is $a(\mathbf{f}_\pi) = A_t\mathbf{f}_\pi$ where

$$A_t = \begin{pmatrix} 0 & 1 & \dots & t \\ 1 & 1 & \dots & 1 \end{pmatrix}.$$

As recalled in Section 2.2, a Markov basis for A_t is a finite set of moves $\mathcal{B}_t^\pi = \{\mathbf{m}_1^\pi, \dots, \mathbf{m}_K^\pi\}$ which belongs to the integer kernel of A_t and induces a connected graph over the fiber associated with A_t. We are interested in a Markov Basis for the specific fiber $\mathcal{F}_{N,t} = \{\mathbf{f}_\pi : A_t\mathbf{f}_\pi = (t, N)^T\}$. With a slight abuse of notation we still denote by $\mathcal{B}_t^\pi$ such a basis. We denote as $\lfloor x \rfloor$ the floor of x, $\lfloor x \rfloor = \max(m \in \mathbb{Z} \mid m \leq x)$.

A procedure to build a basis for $\mathcal{F}_{N,t} = \{\mathbf{f}_\pi : A_t\mathbf{f}_\pi = (t, N)^T\}$ is available in Chen *et al.* (2005), however this procedure results in a basis which is not minimal. Indeed, the number of moves proposed by Chen *et al.* (2005) is $\binom{t-1}{2}$ while the number K of moves we get (see Proposition B.2) satisfies $K \leq \binom{t-1}{2}$. The following proposition gives a way to build a minimal basis.

***Proposition* B.1.** For any integer t one can build a Markov basis $\mathcal{B}_t^\pi$ for the fiber $\mathcal{F}_{N,t}$ considering the moves $\mathbf{m}_{k,i}$ which are built as follows: for every $2 \leq k \leq t$ and for every $1 \leq i \leq \lfloor k/2 \rfloor$ the $t+1$ components vector $\mathbf{m}_{k,i}$ is constructed by

1. setting all the components of $\mathbf{m}_{k,i}$ equal to zero;
2. setting $(\mathbf{m}_{k,i})_0 = -1$ and $(\mathbf{m}_{k,i})_k = -1$;
3. setting $(\mathbf{m}_{k,i})_i = 1$;
4. setting $(\mathbf{m}_{k,i})_{k-i} = (\mathbf{m}_{k,i})_{k-i} + 1$.

Proof. First we observe that for any $\mathbf{m} \in \mathcal{B}_t^\pi$ its components $(\mathbf{m})_i, i = 0, \dots, t$ are in $\{-1, 0, 1, 2\}$. It follows that $\mathbf{m} \in \mathbb{Z}^{t+1}$. We also observe that $m_i = 2$ if and only if $i = k-i$, that is if $i = k/2$.

Secondly, $A_t\mathbf{m} = \mathbf{0}$ because

$$\begin{aligned}(0, 1, \dots, t)\mathbf{m} &= 0 \cdot m_0 + i \cdot m_i + (k-i) \cdot m_{k-i} + k \cdot m_k = \\ &= 0 \cdot (-1) + i \cdot 1 + (k-i) \cdot 1 + k \cdot (-1) = 0\end{aligned}$$

and

$$\begin{aligned}(1, 1, \dots, 1)\mathbf{m} &= 1 \cdot m_0 + 1 \cdot m_i + 1 \cdot m_{k-i} + 1 \cdot m_k = \\ &= 1 \cdot (-1) + 1 \cdot 1 + 1 \cdot 1 + 1 \cdot (-1) = 0.\end{aligned}$$

Thirdly, we prove that the points of the re-parametrised $\mathcal{F}_{N,t}$ are connected by the moves of $\mathcal{B}_t^\pi$ by induction over t. Preliminary, we observe that $\mathcal{B}_t^\pi$ can be considered as the disjoint union of $\mathcal{B}_{t,0}^\pi$ and $\mathcal{B}_{t,1}^\pi$ where $\mathcal{B}_{t,0}^\pi = \{(x_0, \dots, x_t) \in \mathcal{B}_t^\pi : x_t = 0\}$ and $\mathcal{B}_{t,1}^\pi = \{(x_0, \dots, x_t) \in \mathcal{B}_t^\pi : x_t = -1\}$. $\mathcal{B}_{t,0}^{pi}$ is obtained with $2 \leq k < t$ and $\mathcal{B}_{t,1}^\pi$ is obtained with $k = t$.

By construction it holds that $\mathcal{B}_{t,0}^\pi = \{(x_0, \dots, x_{t-1}, 0) : (x_0, \dots, x_{t-1}) \in \mathcal{B}_{t-1}^\pi\}$.

- For $t = 1$, we have $\mathcal{F}_{N,1} = \{(N-1, 1)\}$ and $\mathcal{B}_1^\pi = \emptyset$.
- For $t = 2$, we have $\mathcal{F}_{N,2} = \{(N-1, 0, 1), (N-2, 2, 0)\}$ and $\mathcal{B}_2^\pi = \{(-1, 2, -1)\}$. It follows that the two orbits into $\mathcal{F}_{N,2}$ are connected by the move of $\mathcal{B}_2^\pi$.

- Let us now suppose that $\mathcal{B}_t^\pi$ connects the re-parametrised $\mathcal{F}_{N,t}$ and let us prove that $\mathcal{B}_{t+1}^\pi$ connects the remainder $\mathcal{F}_{N,t+1}$. We observe that $\mathcal{F}_{N,t+1}$ is the disjoint union of the sets $\tilde{\mathcal{F}}_{N,t+1}$ and $\{(N-1,0,\dots,0,1)\}$, where $\tilde{\mathcal{F}}_{N,t+1}$ contains the points $(x_0, x_1+1, x_2, \dots, x_t, 0)$ with $(x_0, x_1, x_2, \dots, x_t) \in \mathcal{F}_{N,t}$. It is easy to verify that $\mathcal{B}_{t+1,0}^\pi$ connects all the points of $\tilde{\mathcal{F}}_{N,t+1}$ and that $\mathcal{B}_{t+1,1}^\pi$ connects the point $\{(N-1,0,\dots,0,1)\}$ to the points of $\tilde{\mathcal{F}}_{N,t+1}$.

□

***Remark* B.1.** As for the Markov basis in equation (3), the basis $\mathcal{B}_t^\pi$ can be obtained providing to `4ti2` the matrix A_t which defines the re-parametrised fiber and its dimensions. In general `4ti2` will provide a larger number of moves than those obtained using Proposition B.1. This is due to the fact that `4ti2` gives a basis for *all* the fibers $\{\mathbf{f}_\pi : A_t\mathbf{f}_\pi = \mathbf{b}, \mathbf{b} \geq \mathbf{0}\}$, while we have built a Markov basis for the specific fiber $\mathcal{F}_{N,t}$, where $\mathbf{b} = (t, N)^T$.

***Example* B.1.** For $t = 6$, `4ti2` gives 15 moves: the nine listed in equation (4.6) plus the six below

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & -1 & 0 \\ 0 & 1 & -1 & 0 & 0 & -1 \\ 1 & -1 & 0 & -2 & 0 & -1 \\ -2 & -1 & -1 & 0 & -1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

As one can easily check none of these moves is admissible. For example to use the first move $(0,0,0,0,1,-2,1)$ we need a vector $\mathbf{f}$ such that $\sum_{i=0}^6 if_i \geq 10$, but such $\mathbf{f}$ does not belong to $\mathcal{F}_{N,6}$.

Additionally, the procedure suggested by Chen *et al.* (2005) gives $\binom{6-1}{2} = 10$ moves while only 9 are needed to get a connected graph (see Figure 2).

***Proposition* B.2.** The number K of moves in $\mathcal{B}_t^\pi$ is equal to

$$K = \begin{cases} \frac{t^2}{4} & \text{if } t \text{ is even} \\ \frac{t^2-1}{4} & \text{if } t \text{ is odd} \end{cases}.$$

Proof. From Proposition B.1 it follows that the total number of moves in $\mathcal{B}_t^\pi$, for a generic t, is given by $\sum_{k=2}^t \lfloor k/2 \rfloor$.

Thus if $2 \leq k \leq t$ we need to compute the sum of the following sequence

k	2	3	4	5	6	7	...	t
$\lfloor k/2 \rfloor$	1	1	2	2	3	3	...	$\lfloor t/2 \rfloor$

.

If t is odd then this sum is

$$\begin{aligned}\sum_{k=2}^{t}\lfloor k/2\rfloor &= 2\cdot\sum_{k=1}^{(t-1)/2} k = \\ &= 2\cdot\frac{1}{2}\cdot\left(\frac{t-1}{2}\cdot\left(\frac{t-1}{2}+1\right)\right) = \\ &= \frac{(t-1)(t+1)}{4} = \frac{t^2-1}{4}.\end{aligned}$$

If t is even then

$$\begin{aligned}\sum_{k=2}^{t}\lfloor k/2\rfloor &= 2\cdot\sum_{k=1}^{(t-2)/2} k + \frac{t}{2} = \\ &= 2\cdot\frac{1}{2}\cdot\left(\frac{t-2}{2}\cdot\left(\frac{t-2}{2}+1\right)\right) + \frac{t}{2} = \\ &= \frac{t}{2}\left(\frac{t-2}{2}+1\right) = \frac{t^2}{4}.\end{aligned}$$

□

Appendix C: Properties of Estimators

In this appendix we show the proofs of the 3 properties of the estimators $\mathbb{I}_{(U(\mathbf{y})\leq u)}$ and $F_U(u \mid \pi)$ presented in Section 5. First we prove that both estimators are unbiased, then the relation between the variances of the two estimators and finally, thanks to Lemma C.1, we prove a similar result for the mean absolute deviation.

Proposition. Unbiasedness of both estimators (Proposition 5.1)
$\mathrm{E}_{\mathrm{p}}\left(\mathbb{I}_{(U(\mathbf{y})\leq u)}(\mathbf{y})\right) = \mathrm{E}_{\mathrm{p}_\pi}\left(F_U(u \mid \pi)\right) = F_U(u \mid \mathcal{F}_{N,t})$.

Proof. If we compute the expectation of $F_U(u \mid \pi)$ using p_π we get

$$\begin{aligned}
\mathrm{E}_{\mathrm{p}_\pi}\left(F_U(u \mid \pi)\right) &= \sum_{\pi \subseteq \mathcal{F}_{N,t}} \mathrm{p}_\pi(\pi) F_U(u \mid \pi) \\
&= \sum_{\pi \subseteq \mathcal{F}_{N,t}} \mathrm{p}_\pi(\pi) \frac{1}{\#\pi} \sum_{\mathbf{y} \in \pi} \mathbb{I}_{(U(\mathbf{y}) \leq u)}(\mathbf{y}) \\
&= \sum_{\pi \subseteq \mathcal{F}_{N,t}} \mathrm{p}_\pi(\pi) \sum_{\mathbf{y} \in \pi} \mathbb{I}_{(U(\mathbf{y}) \leq u)}(\mathbf{y})\, \mathrm{p}(\mathbf{y}|\pi) \\
&= \sum_{\pi \subseteq \mathcal{F}_{N,t}} \sum_{\mathbf{y} \in \pi} \mathbb{I}_{(U(\mathbf{y}) \leq u)}(\mathbf{y})\, \mathrm{p}(\mathbf{y}) \\
&= \sum_{\mathbf{y} \in \mathcal{F}_{N,t}} \mathrm{p}(\mathbf{y}) \mathbb{I}_{(U(\mathbf{y}) \leq u)}(\mathbf{y}) \\
&= \mathrm{E}_{\mathrm{p}}\left(\mathbb{I}_{(U(\mathbf{y}) \leq u)}(\mathbf{y})\right) \\
&= F_U(u \mid \mathcal{F}_{N,t}).
\end{aligned}$$

□

Proposition. Comparison of Variances (Proposition 5.2)
$\mathrm{var}_{\mathrm{p}}\left(\mathbb{I}_{(U(\mathbf{y}) \leq u)}\right) \geq \mathrm{var}_{\mathrm{p}_\pi}\left(F_U(u \mid \pi)\right)$.

Proof. From Proposition 5.1 both $\mathbb{I}_{(U(\mathbf{y}) \leq u)}$ and $F_U(u \mid \pi)$ are unbiased estimator of the distribution of U over the fiber $\mathcal{F}_{N,t}$, $F_U(u \mid \mathcal{F}_{N,t})$. Then it is enough to show that

$$\mathrm{E}_{\mathrm{p}}\left(\left(\mathbb{I}_{(U(\mathbf{y}) \leq u)}\right)^2\right) \geq \mathrm{E}_{\mathrm{p}_\pi}\left(\left(F_U(u \mid \pi)^2\right)\right)$$

From $\left(\mathbb{I}_{(U(\mathbf{y}) \leq u)}\right)^2 = \mathbb{I}_{(U(\mathbf{y}) \leq u)}$ we have

$$\begin{aligned}
\mathrm{E}_{\mathrm{p}}\left(\left(\mathbb{I}_{(U(\mathbf{y}) \leq u)}\right)^2\right) &= \mathrm{E}_{\mathrm{p}}\left(\mathbb{I}_{(U(\mathbf{y}) \leq u)}\right) \\
&= \sum_{\mathbf{y} \in \mathcal{F}_{N,t}} \mathrm{p}(\mathbf{y}) \mathbb{I}_{(U(\mathbf{y}) \leq u)} \\
&= \sum_{\pi \subseteq \mathcal{F}_{N,t}} \sum_{\mathbf{y} \in \pi} \mathrm{p}(\mathbf{y}) \mathbb{I}_{(U(\mathbf{y}) \leq u)} \\
&= \sum_{\pi \subseteq \mathcal{F}_{N,t}} \sum_{\mathbf{y} \in \pi} \mathrm{p}(\mathbf{y} \mid \pi)\, \mathrm{p}_\pi(\pi) \mathbb{I}_{(U(\mathbf{y}) \leq u)} \\
&= \sum_{\pi \subseteq \mathcal{F}_{N,t}} \mathrm{p}_\pi(\pi) \sum_{\mathbf{y} \in \pi} \frac{1}{\#\pi} \mathbb{I}_{(U(\mathbf{y}) \leq u)} \\
&\geq \sum_{\pi \subseteq \mathcal{F}_{N,t}} \mathrm{p}_\pi(\pi) \left(\sum_{\mathbf{y} \in \pi} \frac{1}{\#\pi} \mathbb{I}_{(U(\mathbf{y}) \leq u)} \right)^2 \\
&= \mathrm{E}_{\mathrm{p}_\pi}\left(\left(F_U(u \mid \pi)^2\right)\right).
\end{aligned}$$

The $\geq$ sign comes from

$$0 \leq \left(\sum_{\mathbf{y} \in \pi} \frac{\mathbb{I}_{(U(\mathbf{y}) \leq u)}}{\#\pi} \right) \leq 1 \quad \Rightarrow \quad \left(\sum_{\mathbf{y} \in \pi} \frac{\mathbb{I}_{(U(\mathbf{y}) \leq u)}}{\#\pi} \right)^2 \leq \left(\sum_{\mathbf{y} \in \pi} \frac{\mathbb{I}_{(U(\mathbf{y}) \leq u)}}{\#\pi} \right).$$

□

To prove the result in Proposition 5.3 for the mean absolute deviation (MAD), we need the following Lemma.

***Lemma* C.1.** Let $x, y \in [0,1]$, then the following facts hold:

1. $x - 2xy + y \geq 0$;
2. $|x - y| \leq x - 2xy + y$.

Proposition. Comparison of MAD (Proposition 5.3)
$\mathrm{MAD}_{\mathrm{p}} \left[\mathbb{I}_{(U(\mathbf{y}) \leq u)} \right] \geq \mathrm{MAD}_{\mathrm{p}_\pi} \left[F_U(u \mid \pi) \right]$.

Proof.

$$\begin{aligned}
&\mathrm{MAD}_{\mathrm{p}} \left[\mathbb{I}_{(U(\mathbf{y}) \leq u)} \right] = \mathrm{E} \left(|\mathbb{I}_{(U(\mathbf{y}) \leq u)} - F_U(u \mid \mathcal{F}_{N,t})| \right) \\
&= \sum_{\mathbf{y} \in \mathcal{F}_{N,t}} \mathrm{p}(\mathbf{y}) |\mathbb{I}_{(U(\mathbf{y}) \leq u)} - F_U(u \mid \mathcal{F}_{N,t})| \\
&= \sum_{\pi \subseteq \mathcal{F}_{N,t}} \sum_{\mathbf{y} \in \pi} \mathrm{p}(\mathbf{y}) |\mathbb{I}_{(U(\mathbf{y}) \leq u)} - F_U(u \mid \mathcal{F}_{N,t})| \\
&= \sum_{\pi \subseteq \mathcal{F}_{N,t}} \mathrm{p}_\pi(\pi) \cdot \frac{1}{\#\pi} \sum_{\mathbf{y} \in \pi} |\mathbb{I}_{(U(\mathbf{y}) \leq u)} - F_U(u \mid \mathcal{F}_{N,t})|.
\end{aligned}$$

We divide the vectors $\mathbf{y} \in \pi$ into two classes C_0 and $C_1 = \bar{C}_0$, such that $C_0 = \left\{ \mathbf{y} \in \pi : \mathbb{I}_{(U(\mathbf{y}) \leq u)} = 0 \right\}$ and $C_1 = \left\{ \mathbf{y} \in \pi : \mathbb{I}_{(U(\mathbf{y}) \leq u)} = 1 \right\}$, then

$$\begin{aligned}
&\mathrm{MAD}_p \left[\mathbb{I}_{(U(\mathbf{y}) \leq u)} \right] = \\
&= \sum_{\pi \subseteq \mathcal{F}_{N,t}} \mathrm{p}_\pi(\pi) \cdot \frac{1}{\#\pi} \left(\sum_{C_0} F_U(u \mid \mathcal{F}_{N,t}) + \sum_{C_1} (1 - F_U(u \mid \mathcal{F}_{N,t})) \right) \\
&= \sum_{\pi \subseteq \mathcal{F}_{N,t}} \mathrm{p}_\pi(\pi) \cdot \frac{1}{\#\pi} \left(\#C_0 \cdot F_U(u \mid \mathcal{F}_{N,t}) + \#C_1 \cdot (1 - F_U(u \mid \mathcal{F}_{N,t})) \right).
\end{aligned}$$

By looking at the definition of $F_U(u \mid \pi)$ in equation (4) we observe that $\#C_1/\#\pi = F_U(u \mid \pi)$ and $\#C_0/\#\pi = 1 - \#C_1/\#\pi = 1 - F_U(u \mid \pi)$, thus

$$\begin{aligned}
&\mathrm{MAD}_{\mathrm{p}} \left[\mathbb{I}_{(U(\mathbf{y}) \leq u)} \right] = \\
&= \sum_{\pi \subseteq \mathcal{F}_{N,t}} \mathrm{p}_\pi(\pi) \left(F_U(u \mid \mathcal{F}_{N,t}) - 2 F_U(u \mid \pi) F_U(u \mid \mathcal{F}_{N,t}) + F_U(u \mid \pi) \right).
\end{aligned}$$

Now, by Lemma C.1, it holds that

$$|F_U(u \mid \mathcal{F}_{N,t}) - F_U(u \mid \pi)| \leq F_U(u \mid \mathcal{F}_{N,t}) - 2F_U(u \mid \pi)F_U(u \mid \mathcal{F}_{N,t}) + F_U(u \mid \pi)$$

Therefore

$$\begin{aligned}
&\mathrm{MAD}_{\mathrm{p}}\left[\mathbb{I}_{(U(\mathbf{y})\leq u)}\right] = \\
&= \sum_{\pi \subseteq \mathcal{F}_{N,t}} \mathrm{p}_\pi(\pi)\left(F_U(u \mid \mathcal{F}_{N,t}) - 2F_U(u \mid \pi)F_U(u \mid \mathcal{F}_{N,t}) + F_U(u \mid \pi)\right) \\
&\geq \sum_{\pi \subseteq \mathcal{F}_{N,t}} \mathrm{p}_\pi(\pi)|F_U(u \mid \mathcal{F}_{N,t}) - F_U(u \mid \pi)| \\
&= \mathrm{MAD}_{\mathrm{p}_\pi}\left[F_U(u \mid \pi)\right].
\end{aligned}$$

□

Appendix D: Permutation and MCMC sampling

In this section we carry out a brief analysis of the limit case which we get when we sample just one orbit π and we carry out a standard Monte Carlo sampling over π. If the sampled orbit is $\pi_{\mathbf{y}_{obs}}$, i.e. the one which contains the observed vector $\mathbf{y}_{obs}$, the sampling procedure proposed in Section 4 corresponds to the standard permutation sampling (Pesarin and Salmaso, 2010).

We observe that $\mathbf{y}_{obs}$ is an observation sampled from the distribution p and that the corresponding orbit $\pi_{\mathbf{y}_{obs}}$ is an observation sampled from the distribution p_π, where p and p_π are the probability distributions in Proposition 5.1.

Two well-known remarkable properties of the permutation sampling immediately follows from Proposition 5.1. First $F_U(u \mid \pi_{\mathbf{y}_{obs}})$, the cumulative distribution function conditional to the orbit of the observed sample, is an unbiased estimator of $F_U(u \mid \mathcal{F}_{N,t})$, the cumulative distribution function over the fiber $\mathcal{F}_{N,t}$. Secondly, it is an unbiased estimator of $F_U(u \mid \mathcal{F}_{N,t})$ for *any* distribution function p, that does not need to be specified. In fact the estimator $F_U(u \mid \pi_{\mathbf{y}_{obs}})$ does not need any expression of p to be computed.

***Example* D.1.** As a simple example consider again the fiber $\mathcal{F}_{3,6}$ in Figure 1. We select $n_1 = 2$ and $n_2 = 1$ and we compare the exact cumulative distribution over the fiber $F_U(u \mid \mathcal{F}_{N,t})$ and the cumulative distribution over $\pi = \pi_{(1,2,3)}$, the orbit with highest probability, $F_U(u \mid \pi)$. We get two distributions (Table 1) which are considerably close, even if the cardinality of the selected orbit is low ($\#\pi_{(1,2,3)} = 6$) compared to the the cardinality of $\mathcal{F}_{3,6}$, which is 28. However, it is easy to see that some orbits do not give a good approximation of the distribution over $\mathcal{F}_{N,t}$. If we refer again to the fiber $\mathcal{F}_{3,6}$ and we consider $\pi_{(2,2,2)}$, we get a cumulative distribution which has only two values, 0 and 1 (Table 1). This difference is due to the unequal probabilities of the orbits in $\mathcal{F}_{N,t}$ (these probabilities are reported in Table 1).

Table 1. Cumulative distributions of U

u	0	1	2	3	4	5	6
$\mathcal{F}_{3,6}$	0.001	0.018	0.100	0.320	0.649	0.912	1
$\pi_{(1,2,3)}$	0	0	0	0.333	0.667	1	1
$\pi_{(2,2,2)}$	0	0	0	0	1	1	1

Table 2. UMPU tests. Scenario definition

n_1	6	10	30
n_2	4	15	20

(a) Sample sizes

μ_1	1	1	1
μ_2	1	1.5	2

(b) Population means

Appendix E: Simulation study

E.1. UMPU Tests

We start by considering the uniformly most powerful unbiased test in equation (2.1). We compare the *exact* conditional cumulative distribution function $F(u \mid \mathcal{F}_{N,t})$ in the case of Poisson data with

- the *approximated* conditional cumulative distribution function obtained using Algorithm 1;
- the *approximated* conditional cumulative distribution function obtained using Algorithm 2;
- the standard permutation cumulative distribution function $\hat{F}(u \mid \pi_{\mathbf{y}_{obs}})$ (this is the limit case of Algorithm 2 when only the first step is performed, i.e. $N_{sim} = 0$).

A preliminary simulation study is presented in Crucinio and Fontana (2017).

We consider Poisson distributed data $\mathbf{Y}_1 = (Y_1, \dots, Y_{n_1})$ of size n_1 from $\text{Poisson}(\mu_1)$ and $\mathbf{Y}_2 = (Y_{n_1+1}, \dots, Y_{n_1+n_2})$ of size n_2 from $\text{Poisson}(\mu_2)$. In this case the exact distribution (2.2) under $H_0 : \mu_1 = \mu_2 = \mu$ is known to be a binomial distribution with t trials and probability of success $n_1/(n_1 + n_2)$. Then the exact value of $F_U(u \mid \mathcal{F}_{N,t})$ is given by

$$F_U(u \mid \mathcal{F}_{N,t}) = \text{p}(\text{Binomial}(t, \theta_0) \leq u) = \sum_{k=0}^{u} \binom{t}{k} \theta_0^k (1-\theta_0)^{t-k}, \tag{E.1}$$

with $\theta_0 = n_1/(n_1 + n_2)$.

We consider 9 scenarios built taking three different sample sizes (n_1, n_2) (Table 3a) and, for each sample size, three different population means (μ_1, μ_2) (Table 3b).

First we compare how fast the two MCMC procedures converge to the true distribution $F_U(u \mid \mathcal{F}_{N,t})$. We draw one random sample $\mathbf{y}_{obs}$ for each scenario above, we run both Algorithm 1 and 2 for 15 seconds and at every step we compute the corresponding estimate of $F_U(u \mid \mathcal{F}_{N,t})$ (which is based on the indicator function $\mathbb{I}_{(U(\mathbf{y}) \leq u)}$ in the first case and on the permutation distribution $F_U(u \mid \pi)$ in the second one). The number of Monte Carlo permutations for every sampled orbit π is kept fixed and equal to $5,000$. In both Algorithm 1 and 2 we do not include burn-in steps.

Table 3. UMPU tests. Error analysis: standard deviation and MAD

Scenario				Std Dev			MAD		
n_1	n_2	μ_1	μ_2	Algorithm 2	Algorithm 1	Permutation	Algorithm 2	Algorithm 1	Permutation
6	4	1	1	0.0112	0.0237	0.0523	0.0094	0.0197	0.0368
6	4	1	1.5	0.0109	0.0239	0.0451	0.0094	0.0246	0.0367
6	4	1	2	0.0116	0.0202	0.0450	0.0102	0.0311	0.0371
10	15	1	1	0.0181	0.0248	0.0278	0.0140	0.0205	0.0202
10	15	1	1.5	0.0096	0.0213	0.0270	0.0143	0.0289	0.0205
10	15	1	2	0.0067	0.0116	0.0232	0.0139	0.0342	0.0209
20	30	1	1	0.0261	0.0307	0.0189	0.0213	0.0252	0.0143
20	30	1	1.5	0.0072	0.0223	0.0163	0.0221	0.0320	0.0143
20	30	1	2	0.0025	0.0091	0.0093	0.0215	0.0320	0.0144
Overall				0.0133	0.0218	0.0326	0.0151	0.0276	0.0239

Figure 1 shows four examples of the behaviour of the two MCMC procedures which are representative of the 9 scenarios. In all the 9 scenarios the convergence of Algorithm 2 to (E.1) is much faster than that of Algorithm 1. In fact Algorithm 2 achieves satisfactory convergence in ~ 0.1 seconds while Algorithm 1 takes more than 10 seconds.

To further explore the convergence of the three sampling algorithms, we consider again the 9 scenarios above and for every scenario we draw 1,000 samples. For each sample we run a standard permutation sampling and both Algorithm 1 and Algorithm 2 with the settings described above (i.e. 15 seconds, no burn-in and 5,000 permutations per orbit).

For each sample we compute the errors of the three estimated distributions. We denote by $\hat{F}_U^{\pi}$, $\hat{F}_U^{\mathbf{y}}$ and $\hat{F}_U(u \mid \pi_{\mathbf{y}_{obs}})$ the estimated distributions obtained by Algorithm 2, Algorithm 1 and the standard permutation method, respectively. The errors $\mathrm{E}_{\mathbf{y}}$, E_{π} and E_{perm} are defined as

$$\mathrm{E}_{\mathbf{y}} = \hat{F}_U^{\mathbf{y}}(u \mid \mathcal{F}_{N,t}) - \sum_{k=0}^{u} \binom{t}{k} \theta_0^k (1-\theta_0)^{t-k}, \tag{E.2}$$

$$\mathrm{E}_{\pi} = \hat{F}_U^{\pi}(u \mid \mathcal{F}_{N,t}) - \sum_{k=0}^{u} \binom{t}{k} \theta_0^k (1-\theta_0)^{t-k}, \tag{E.3}$$

$$\mathrm{E}_{\text{perm}} = \hat{F}_U(u \mid \pi_{\mathbf{y}_{obs}}) - \sum_{k=0}^{u} \binom{t}{k} \theta_0^k (1-\theta_0)^{t-k}. \tag{E.4}$$

We know from Proposition 5.1 that the expected value of $\mathrm{E}_{\mathbf{y}}$, E_{π} and E_{perm} is 0. On average the observed errors are very close to 0 for every sampling procedure. The overall sample means of the errors are $+0.0017$, -0.0093 and $+0.0052$ for Algorithm 2, Algorithm 1 and the standard permutation method respectively (the overall sample means have been computed considering all the 9,000 simulations).

In Table 3 we report the standard deviation and mean absolute deviation (MAD) of the errors. To analyse the difference between Algorithm 2, 1 and standard permutation we compute the ratio between the standard deviations of the first two and the standard deviation of Algorithm 2. We repeat the same computation for the MAD. The results are in Table 4. Algorithm 1 gives a

Table 4. UMPU tests. Error analysis: ratios of standard deviation and MAD

Scenario				Std Dev		MAD	
n_1	n_2	μ_1	μ_2	Algorithm 1	Permutation	Algorithm 1	Permutation
6	4	1	1	2.12	4.68	2.09	3.90
6	4	1	1.5	2.19	4.13	2.60	3.89
6	4	1	2	1.73	3.87	3.06	3.65
10	15	1	1	1.37	1.54	1.47	1.45
10	15	1	1.5	2.23	2.83	2.02	1.43
10	15	1	2	1.73	3.44	2.47	1.50
20	30	1	1	1.18	0.72	1.18	0.67
20	30	1	1.5	3.09	2.25	1.45	0.65
20	30	1	2	3.58	3.66	1.48	0.67
Overall				1.64	2.45	1.82	1.58

Table 5. UMPU tests. Number of iterations

Scenario				N. iterations		ratio
n_1	n_2	μ_1	μ_2	Orbit	Fiber	Fiber/Orbit
6	4	1	1	23977	53842	2.25
6	4	1	1.5	24169	53210	2.20
6	4	1	2	24560	57382	2.34
10	15	1	1	11950	52504	4.39
10	15	1	1.5	11564	54836	4.74
10	15	1	2	7326	53492	7.30
20	30	1	1	4675	45576	9.75
20	30	1	1.5	3174	44817	14.12
20	30	1	2	2572	48003	18.66

standard deviation that is often double (the overall value is 1.64) than that of Algorithm 2. This overall ratio becomes 2.45 when we compare the standard deviation of Algorithm 2 with that of the standard permutation. The corresponding overall values of the ratios of the MADs are 1.82 and 1.58, respectively.

These results are consistent with Propositions 5.2 and 5.3 in Section 5 and are confirmed by Figure 2, in which the histograms of the absolute errors for four scenarios are shown. These plots are representative of the behaviour in any of the 9 scenarios.

The number of iterations made by Algorithm 2 and Algorithm 1 in the allocated 15 seconds are reported in Table 5. Algorithm 2 performs better than Algorithm 1 even if the number of iterations made is lower: in 15 seconds, Algorithm 1 makes from twice to almost 19 times more iterations than Algorithm 2. Despite this difference in the number of iterations, Algorithm 2 always achieves lower variance and MAD.

***Remark* E.1.** It would be possible to further reduce the computational time required by Algorithm 2 by exploiting one of the key features of this new approach, namely the possibility of sampling from each orbit independently. The computations of the Monte Carlo cdf $\hat{F}_U(u_{obs} \mid \pi)$ (step 5 of Algorithm 2) could be made in *parallel*: once the chain reaches an orbit π the Monte Carlo sampling over π can be performed while the chain keeps on moving on the set of orbits.

Table 6. Linear Regression. Scenario definition

n_1	3	7	15
n_2	4	8	15
n_3	3	8	20

(a) Sample sizes

μ_1	1	1	1
μ_2	1	1	1
μ_3	1	1.5	2

(b) Population means

E.2. Simple Linear Regression

We carry on our simulation study by considering the hypothesis test described in Section 6.2.

If the samples we are interested in are Poisson distributed, we know from McCullagh and Nelder (1989) that the vector $(Y_1, \dots, Y_N)$ given the sum $T = \sum_{i=1}^N Y_i$ follows a multinomial distribution

$$\text{Multinomial}\left(t, \frac{n_1}{n_1 + \ldots + n_k}, \dots, \frac{n_k}{n_1 + \ldots + n_k}\right),$$

where $n_1, \dots, n_k$ are the number of observations Y_i in each of the k groups.

We set the number of groups $k = 3$ and we consider the test statistic $U = \sum_{i=1}^{n_1} Y_i + 2\sum_{i=n_1+1}^{n_2+n_1} Y_i + 3\sum_{i=n_1+n_2+1}^{N} Y_i$. The distribution of U given T can be computed from the Multinomial distribution above as

$$\mathrm{p}(U = u \mid T = t) = \sum_{x_1 = 2t-u}^{(3t-u)/2} \mathrm{p}(\text{Multinomial} = (x_1, 3t - 2x_1 - u, u - 2t + x_1)). \tag{E.5}$$

We compare the *exact* values obtained using equation (E.5) with

- the *approximated* conditional cumulative distribution function obtained using Algorithm 1;
- the *approximated* conditional cumulative distribution function obtained using Algorithm 2;
- the standard permutation cumulative distribution function $\hat{F}(u \mid \pi_{\mathbf{y}_{obs}})$ (this is the limit case of Algorithm 2 when $N_{\text{iter}} = 0$).

As we did earlier, we consider 9 scenarios built taking three different sample sizes (n_1, n_2, n_3) (Table 7a) and, for each sample size, three different population means (μ_1, μ_2, μ_3) (Table 7b).

Figure 3 shows two examples of convergence for Algorithm 1 and Algorithm 2 in 10 seconds. As usual, we report the exact value computed with equation (E.5) and the value given by a standard permutation sampling. The behaviour of the two Algorithm is similar to the one observed in Appendix E.1. Algorithm 2 converges in the least time to the exact value.

As a further comparison, we compute the errors for the 3 sampling procedures as shown in equations (E.2), (E.3) and (E.4), where we substitute the binomial cdf with the cdf we obtain by summing up the probabilities in equation (E.5). The overall means of the errors are +0.0027, +0.0032, -0.0005 for Algorithm 1, 2 and standard permutation sampling. As expected, all the estimates are unbiased (Proposition 5.1 and Remark 5.1).

Table 7. Linear Regression. Error analysis: standard deviation and MAD

Scenario						Std Dev			MAD		
n_1	n_2	n_3	μ_1	μ_2	μ_3	Algorithm 2	Algorithm 1	Permutation	Algorithm 2	Algorithm 1	Permutation
3	4	3	1	1	1	0.010	0.010	0.045	0.009	0.007	0.033
3	4	3	1	1	1.5	0.008	0.009	0.042	0.006	0.006	0.030
3	4	3	1	1	2	0.006	0.008	0.043	0.005	0.005	0.030
7	8	8	1	1	1	0.006	0.016	0.050	0.006	0.012	0.030
7	8	8	1	1	1.5	0.005	0.014	0.089	0.004	0.010	0.040
7	8	8	1	1	2	0.003	0.009	0.081	0.003	0.006	0.036
15	15	20	1	1	1	0.007	0.024	0.200	0.006	0.018	0.121
15	15	20	1	1	1.5	0.008	0.017	0.225	0.006	0.012	0.126
15	15	20	1	1	2	0.007	0.009	0.130	0.004	0.004	0.040
Overall						0.007	0.014	0.120	0.005	0.009	0.054

Table 8. Linear Regression. Error analysis: ratios of standard deviation and MAD

Scenario						Std Dev		MAD	
n_1	n_2	n_3	μ_1	μ_2	μ_3	Algorithm 1	Permutation	Algorithm 1	Permutation
3	4	3	1	1	1	0.99	4.61	0.86	3.86
3	4	3	1	1	1.5	1.13	5.20	1.01	4.80
3	4	3	1	1	2	1.21	6.68	1.05	5.98
7	8	8	1	1	1	2.48	7.85	2.17	5.36
7	8	8	1	1	1.5	2.85	18.43	2.60	10.61
7	8	8	1	1	2	2.73	24.11	2.04	13.13
15	15	20	1	1	1	3.28	27.00	3.25	21.28
15	15	20	1	1	1.5	2.18	28.26	2.10	22.94
15	15	20	1	1	2	1.29	19.41	1.12	10.61
Overall						1.98	17.20	1.72	10.39

We report the standard deviations and MADs in Table 7 and the corresponding ratios in Table 8. As observed above, Algorithm 1 gives estimates whose variability measures are on average twice as big than those obtained with Algorithm 2.

In conclusion, the observations made for the simulation study on the UMPU test (Appendix E.1) apply in this case, too: Algorithm 2 outperforms Algorithm 1 in terms of speed and accuracy.

E.3. Computational Details

The simulation study presented in this section was implemented in SAS/IML®. The software code is available upon request. We performed the analysis using a standard laptop (CPU Intel core 2 Duo T6570 CPU 2.10GHz 2.10GHz, RAM 4GB).

Acknowledgments

R. Fontana acknowledges that the present research has been partially supported by MIUR grant Dipartimenti di Eccellenza 2018-2022 (E11G18000350001). The authors thank the anonymous referee for his/her helpful comments.

References

Chen, Y., Dinwoodie, I., Dobra, A., and Huber, M. (2005). Lattice points, contingency tables, and sampling. *Contemporary Mathematics*, **374**, 65–78.

Crucinio, F. R. and Fontana, R. (2017). Comparison of conditional tests on Poisson data. In *Statistics and Data Science: proceedings of the Conference of the Italian Statistical Society*, pages 333–338. Firenze University Press.

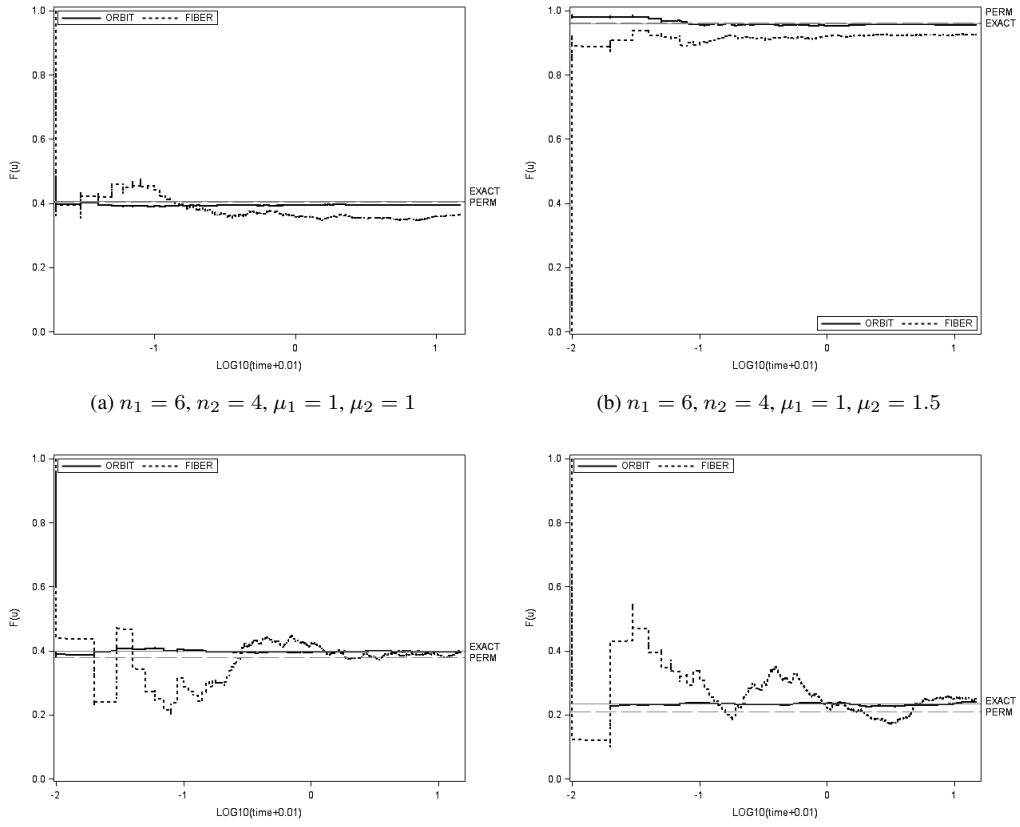
Kunz, M. (2006). Partitions and their lattices. *arXiv preprint arXiv:math/0604203*.

Lovász, L., Pelikán, J., and Vesztergombi, K. (2006). *Discrete Mathematics: Elementary and Beyond*. Undergraduate Texts in Mathematics. Springer New York.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models, Second Edition*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

Pesarin, F. and Salmaso, L. (2010). *Permutation tests for complex data: theory, applications and software*. John Wiley & Sons.

Wilf, H. S. (2000). Lectures on integer partitions. Available at https://www.math.upenn.edu/ wilf/PIMS/PIMSLectures.pdf. Retrieved 20 Apr 2017.



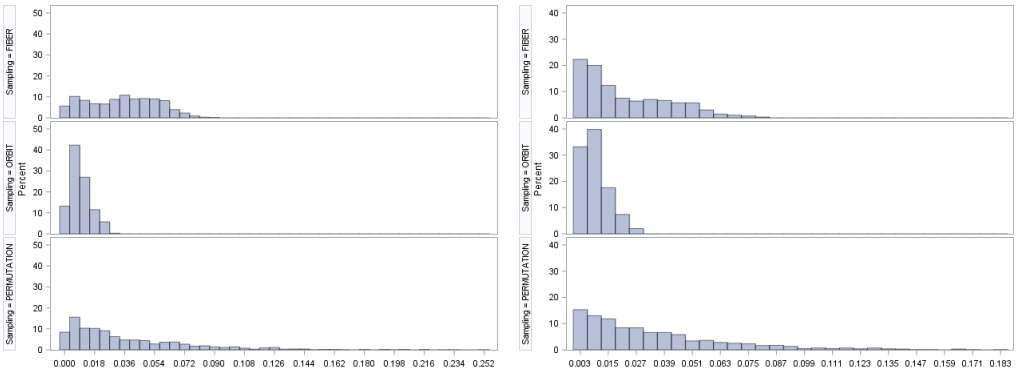
(a) $n_1 = 6, n_2 = 4, \mu_1 = 1, \mu_2 = 1$

(b) $n_1 = 6, n_2 = 4, \mu_1 = 1, \mu_2 = 1.5$

(c) $n_1 = 10, n_2 = 15, \mu_1 = 1, \mu_2 = 1$

(d) $n_1 = 30, n_2 = 20, \mu_1 = 1, \mu_2 = 1.5$

Figure 1. **UMPU tests**. Comparison of the convergence to the exact value (solid horizontal line) in 15 seconds of Algorithm 2 (solid line) and Algorithm 1 (dashed line). The Monte Carlo permutation estimate of $F_U(u \mid \mathcal{F}_{N,t})$ (dashed horizontal line) is reported too. The number of Monte Carlo permutations for $\pi_{\mathbf{y}_{obs}}$ is $5,000$.



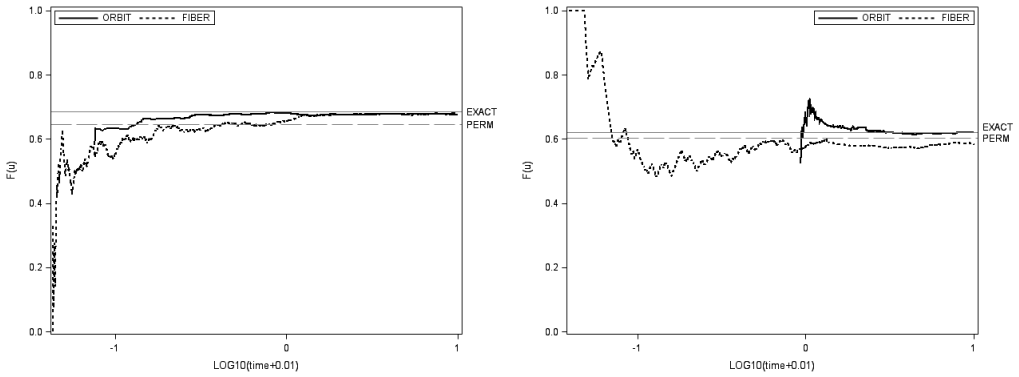
(a) $n_1 = 6, n_2 = 4, \mu_1 = 1, \mu_2 = 1$

(b) $n_1 = 6, n_2 = 4, \mu_1 = 1, \mu_2 = 1.5$

(c) $n_1 = 10, n_2 = 15, \mu_1 = 1, \mu_2 = 1$

(d) $n_1 = 10, n_2 = 15, \mu_1 = 1, \mu_2 = 1.5$

Figure 2. **UMPU tests**. Observed absolute values of the approximation error for Algorithm 1, E_y, the approximation error for Algorithm 2, E_π, and the approximation error for the standard permutation sampling, $\mathrm{E}_{\mathrm{perm}}$.



(a) $n_1 = 3, n_2 = 4, n_3 = 3, \mu_1 = 1, \mu_1 = 1, \mu_3 = 2$ (b) $n_1 = 15, n_2 = 15, n_3 = 20, \mu_1 = 1, \mu_2 = 1, \mu_3 = 1$

Figure 3. **Linear Regression**. Comparison of the convergence to the exact value (solid horizontal line) in 10 seconds of Algorithm 2 (solid line) and Algorithm 1 (dashed line). The Monte Carlo permutation estimate of $F_U(u \mid \mathcal{F}_{N,t})$ (dashed horizontal line) is reported too. The number of Monte Carlo permutations for $\pi_{\mathbf{y}_{obs}}$ is $5,000$.