Metrika (2007) 65: 69–82
DOI 10.1007/s00184-006-0060-0

ORIGINAL ARTICLE

Fernando López-Blázquez · Jacek Wesołowski

Top-k-lists

Received: 14 March 2005 / Published online: 3 May 2006

Abstract Top-k-lists are introduced as sequences of k-dimensional random vectors with ordered components being k largest observations from a sequence of independent identically distributed random variables. Such lists changing in time are natural stochastic models of ranking tables which appear in many situations in real life, when one wants to keep a track of several best results in a given field. Here we study basic properties of top-k-lists as joint distributions, conditional structures, representations, driving examples of top-k-lists from exponential and uniform distributions, asymptotics and a relation to generalized order statistics.

Keywords Top-k-list · kth records · Record representations · Limit theorems · Generalized order statistics

Mathematics Subject Classification (2000) 62G30 · 62E15 · 62E20 · 60E05 · 60F05

1 Introduction

In many applications it is usual to keep track of the best k outcomes of an experiment up to a given moment. Throughout this paper we will assume that an outcome is better than another if its value is larger. For instance, in sports it is usual to deal with the list of the k best performances up to the moment, like in the table below, which gives the $k = 5$ best performances in long-jump at a certain moment:

F. López-Blázquez
Departamento de Estadistica e Investigación Operativa,
Universidad de Sevilla, Sevilla, Spain
E-mail: lopez@us.es

J. Wesołowski (✉)
Wydział Matematyki i Nauk Informacyjnych,
Politechnika Warszawska, Warszawa, Poland
E-mail: wesolo@mini.pw.edu.pl

Mike Powell	8.95 m
Bob Beamon	8.90 m
Carl Lewis	8.87 m
Robert Emmiyan	8.86 m
Carl Lewis	8.79 m

Such tables changed many times in the history of this kind of sport competition. A sequence of these tables, changing in time, will be called top-k-lists. The aim of this paper is to provide a basic mathematical model for such lists under the assumption that consecutive results are outcomes of random experiments.

First, we introduce notation and recall some basic results about order statistics (see, for instance, David and Nagaraja (2003)) and kth records (see, for instance, Arnold et al. (1998)). Given a list of k real numbers, $x_1, \ldots, x_k$, they can be arranged in an increasing order to obtain $x_{1:k} \leq \cdots \leq x_{k:k}$. Then we denote

$$\operatorname{ord}(x_1, \ldots, x_k) = (x_{1:k}, \ldots, x_{k:k}) .$$

If F is a distribution function (df), the quantile function of F (also called generalized inverse) is defined as

$$Q_F(u) = \inf\{x : \ F(x) \geq u\}, \quad \text{for } u \in (0, 1].$$

Let $X_1, \ldots, X_k$ be independent and identically distributed (iid) random variables (rvs) from an absolutely continuous df F with a density f. The ith component of $\operatorname{ord}(X_1, \ldots, X_k)$, denoted $X_{i:k}$, is called the ith order statistic from the sample of size k, $i = 1, \ldots, k$. The joint density g of order statistics is

$$g(x_1, \ldots, x_k) = k! \prod_{i=1}^{k} f(x_i), \quad \text{for } x_1 \leq x_2 \leq \cdots \leq x_k. \tag{1}$$

We recall now the concept of kth records. Let $(X_i)_{i\geq 1}$ be a sequence of iid rvs with a common df F. Consider the kth record times defined recurrently as

$$\begin{aligned} T_0^{(k)} &= k, \\ T_{n+1}^{(k)} &= \min\left\{ j : \ j > T_n^{(k)} \text{ and } X_j > X_{T_n^{(k)}-k+1:T_n^{(k)}} \right\}, \quad n \geq 0, \end{aligned}$$

then the rv $R_n^{(k)} = X_{T_n^{(k)}-k+1:T_n^{(k)}}$, $n \geq 0$, is the nth kth record from the sequence $(X_i)_{i\geq 1}$. If F is absolutely continuous with density f, then the density of $R_n^{(k)}$ is

$$f_{R_n^{(k)}}(x) = \frac{k^{n+1}}{n!} \{-\log(1 - F(x))\}^n (1 - F(x))^{k-1} f(x), \tag{2}$$

for $x \in \left(Q_F(0^+), Q_F(1)\right)$.

We will denote $X \sim \text{EXP}(\lambda)$ if the rv X follows the exponential distribution with the mean $1/\lambda$, $\lambda > 0$. Also we will write $X \sim G(p, \lambda)$ if X follows a gamma distribution with convolution parameter $p > 0$ and scale parameter $1/\lambda > 0$. If P_X and P_Y are distributions of rvs X and Y then $P_X \otimes P_Y$ denotes the product distribution, i.e. if $(X, Y) \sim P_X \otimes P_Y$ then X and Y are independent. The symbol $\mathbf{1}_k$ denotes a vector with k components, all of them equal 1.

2 Distribution of top-k-lists

A new concept of top-k-lists that we introduce in this paper is given in the next definition:

Definition 1 *Let $(X_i)_{i\geq 1}$ be a sequence of iid rvs with a common df F. The nth top-k-list from $(X_i)_{i\geq 1}$ is defined as follows:*

$$L_n^{(k)} = \left(Y_{1,n},\ Y_{2,n},\ \ldots,\ Y_{k,n}\right), \quad n \geq 0, \tag{3}$$

where $Y_{j,n} = X_{T_n^{(k)}-k+j:T_n^{(k)}}$, $j = 1, \ldots, k$, $n = 0, 1, \ldots$.

Note that $Y_{1,n} = R_n^{(k)}$. The index i of the sequence (X_i) in the previous definition can be viewed as a discrete time. The evolution of the list is as follows. At time k, the 0th top-k-list, $L_0^{(k)} = (X_{1:k}, \ldots, X_{k:k})$, is available. The list remains unaltered until time $T_1^{(k)}$. At this moment, the first element of the list $L_0^{(k)}$ is removed and the rv $X_{T_1^{(k)}}$ enters the list. Then $L_1^{(k)} = \text{ord}\left(X_{T_1^{(k)}},\ X_{2:k}, \ldots, X_{k:k}\right)$. From now on, the process behaves in a similar way: an $(n-1)$th top-k-list $L_{n-1}^{(k)}$ remains unaltered until the nth kth record time $T_n^{(k)}$ occurs, and then

$$L_n^{(k)} = \text{ord}\left(X_{T_n^{(k)}},\ Y_{2,n-1},\ \ldots,\ Y_{k,n-1}\right), \quad n \geq 1. \tag{4}$$

Note that the model of top-k-lists covers at least three important models for ordered statistical data:

- Order statistics as the 0th top-k-list $L_0^{(k)}$
- Records as the sequence $(L_n^{(1)})$
- kth records as the sequence of the first components of $L_n^{(k)}$s

From now on, we will assume that F is absolutely continuous with a density f. Our main aim now is to obtain the joint density of the random vector $L_n^{(k)}$.

Theorem 2 *For any $w \in \left(Q_F(0^+), Q_F(1)\right)$ let $Z_1^{(w)}, \ldots, Z_k^{(w)}$ be iid rvs from the truncated density*

$$f_w(z) = \frac{f(z)}{1-F(w)} I(z \geq w). \tag{5}$$

Then

(a) For all $n \geq 1$, the rv $X_{T_n^{(k)}}$ and the vector $\left(Y_{2,n-1}, \ldots, Y_{k,n-1}\right)$ are conditionally independent given $Y_{1,n-1}$. Moreover,

$$\left(X_{T_n^{(k)}} \mid Y_{1,n-1} = w\right) \stackrel{d}{=} Z_1^{(w)}. \tag{6}$$

(b) For all $n \geq 0$,

$$\left(Y_{2,n}, \ldots, Y_{k,n} \mid Y_{1,n} = w\right) \stackrel{d}{=} ord\left(Z_2^{(w)}, \ldots, Z_k^{(w)}\right). \tag{7}$$

(c) The density of $L_n^{(k)}$ *is*

$$f_{L_n^{(k)}}(y_1, \ldots, y_k) = \frac{k^n k!}{n!}[-\log(1 - F(y_1))]^n \prod_{j=1}^{k} f(y_j), \tag{8}$$

for $Q_F(0^+) < y_1 < \cdots < y_k < Q_F(1)$.

Proof (a): Note first that for all $n \geq 1$ the rv $T_{n-1}^{(k)}$ is a stopping time with respect to the natural filtration of the sequence (X_i). Since the components of the sequence are independent then the random elements

$$\left(X_1, \ldots, X_{T_{n-1}^{(k)}}\right) \quad \text{and} \quad \left(X_{T_{n-1}^{(k)}+1}, \ldots\right)$$

are also independent. Observe also that $(Y_{1,n-1}, \ldots, Y_{k,n-1})$ is $\sigma\left(X_1, \ldots, X_{T_{n-1}^{(k)}}\right)$ measurable, and consequently $(Y_{1,n-1}, \ldots, Y_{k,n-1})$ and $\left(X_{T_{n-1}^{(k)}+1}, \ldots\right)$ are independent as well. Thus for $n \geq 1$, and $y \in \mathbb{R}$, we have

$$\begin{aligned}
&P\left(X_{T_n^{(k)}} > y \mid Y_{1,n-1}, \ldots, Y_{k,n-1}\right) \\
&\quad = I(y \leq Y_{1,n-1}) + I(y > Y_{1,n-1}) P\left(X_{T_n^{(k)}} > y \mid Y_{1,n-1}, \ldots, Y_{k,n-1}\right) \\
&\quad = I(y \leq Y_{1,n-1}) + I(y > Y_{1,n-1}) \\
&\qquad \times \sum_{m=1}^{\infty} P\Bigg(X_{T_{n-1}^{(k)}+r} \leq Y_{1,n-1}, r = 1, \ldots, m-1,\ X_{T_{n-1}^{(k)}+m} \\
&\qquad\qquad > y \mid Y_{1,n-1}, \ldots, Y_{k,n-1}\Bigg) \\
&\quad = I(y \leq Y_{1,n-1}) + I(y > Y_{1,n-1}) \sum_{m=1}^{\infty} F^{m-1}(Y_{1,n-1})(1 - F(y)) \\
&\quad = I(y \leq Y_{1,n-1}) + I(y > Y_{1,n-1}) \frac{1 - F(y)}{1 - F(Y_{1,n-1})}.
\end{aligned} \tag{9}$$

Thus (6) is proved. Note that the conditional probability given in (9) does not depend on the values $Y_{2,n-1}, \ldots, Y_{k,n-1}$. This fact implies that $X_{T_n^{(k)}}$ and $\left(Y_{2,n-1}, \ldots, Y_{k,n-1}\right)$ are conditionally independent given $Y_{1,n-1}$.

(b) and (c): We will prove (b) by induction and as a by-product we will obtain (c). Recall that $Y_{j,0} = X_{j:k}$ and that the conditional distribution of $(X_{2:k} \ldots, X_{k:k})$ given $X_{1:k} = w$, $w \in \left(Q_F(0^+), Q_F(1)\right)$, is the same as the joint distribution of the $k-1$ order statistics of a sample of size $k-1$ from the truncated density given in (5), see for instance Theorem 2.5 in David and Nagaraja (2003). Then (7) holds for $n = 0$. Suppose now that (7) is true for $n - 1 \geq 0$, that is to say

$$\left(Y_{2,n-1}, \ldots, Y_{k,n-1} \mid Y_{1,n-1} = w\right) \stackrel{d}{=} \text{ord}\left(Z_2^{(w)}, \ldots, Z_k^{(w)}\right). \tag{10}$$

Now using both assertions of (a) and (10) we conclude that

$$\left(L_n^{(k)} \mid Y_{1,n-1} = w\right) \stackrel{d}{=} \operatorname{ord}\left(Z_1^{(w)}, \ldots, Z_k^{(w)}\right),$$

or equivalently, by (1),

$$f_{L_n^{(k)}|Y_{1,n-1}=w}(y_1, \ldots, y_k) = k! \prod_{i=1}^{k} \frac{f(y_i)}{1 - F(w)}, \quad \text{for } w < y_1 < \cdots < y_k.$$

Now, as $Y_{1,n-1} = R_{n-1}^{(k)}$, for $y_1 < \cdots < y_k$ we have by (2)

$$\begin{aligned} f_{L_n^{(k)}}(y_1, \ldots, y_k) &= \int_{-\infty}^{y_1} f_{L_n^{(k)}|Y_{1,n-1}=w}(y_1, \ldots, y_k) f_{R_{n-1}^{(k)}}(w) \mathrm{d}w \\ &= \frac{k! k^n}{(n-1)!} \left(\int_{-\infty}^{y_1} \frac{[-\log(1 - F(w))]^{n-1}}{1 - F(w)} f(w) \mathrm{d}w \right) \prod_{i=1}^{k} f(y_i) \\ &= \frac{k! k^n}{n!} [-\log(1 - F(y_1))]^n \prod_{i=1}^{k} f(y_i), \end{aligned}$$

and this proves (c). Once we know the joint density of $L_n^{(k)} = \left(Y_{1,n}, \ldots, Y_{k,n}\right)$ and the density of $Y_{1,n} = R_n^{(k)}$ is given in (2) one immediately gets

$$\begin{aligned} &f_{Y_{2,n},\ldots,Y_{k,n}|Y_{1,n}=w}(y_2, \ldots, y_n) \\ &\quad = \frac{f_{L_n^{(k)}}(w, y_2, \ldots, y_k)}{f_{R_n^{(k)}}(w)} \\ &\quad = (k-1)! \prod_{i=2}^{k} \frac{f(y_i)}{1 - F(w)}, \quad w < y_2 < \cdots < y_k, \end{aligned}$$

and thus (7) is proved. □

Remark 3 One of the referees pointed out to the following connection of top-k-lists with the sequence $(M_n^{(k)})$ of k largest observation up to the moment n. If one considers such a sequence of vectors $M_n^{(k)} = (X_{n-k+1:n}, \ldots, X_{n:n})$, $n = k, k+1, \ldots$, then, it is known that the conditional distribution of largest $k-1$ order statistics given $X_{n-k+1:n} = w$ has the density given by (5) for any $n \geq k$. Moreover, $(L_n^{(k)}(\omega))$ can be obtained from $(M_n^{(k)}(\omega))$ by deleting vectors which are repeated in the second sequence for any $\omega \in \Omega$, the sample space. Therefore it is not strange that conditionally both the sequences behave in a similar way, what is exhibited in Theorem 2 (b). On the other hand marginal distributions of $X_{n-k+1:n}$ and $X_{T_n^{(k)}-k+1:T_n^{(k)}}$ are different reflecting the difference between a fixed sampling scheme of $n - k + 1$ lists of k largest elements, and the inverse sampling scheme of $n - k + 1$ distinct lists of k largest elements.

3 Top-k-lists from some specific distributions

Let $L_n^{(k)} = \left(Y_{1,n}, \ldots, Y_{k,n}\right)$, $n \geq 0$, be a top-k-list from a continuous df F in the interval $\left(Q_F(0^+), Q_F(1)\right)$. Then for the transformation $G(x) = -\log(1 - F(x))$, $x \in \left(Q_F(0^+), Q_F(1)\right)$, we have

$$\left(G(Y_{1,n}), \ldots, G(Y_{k,n})\right) \stackrel{d}{=} L_n^{*(k)},$$

where $L_n^{*(k)}$ is the nth top-k-list from a standard exponential distribution, i.e. with the expectation equal 1.

Similarly,

$$\left(F(Y_{1,n}), \ldots, F(Y_{k,n})\right) \stackrel{d}{=} \left(Z_{1,n}, \ldots, Z_{k,n}\right),$$

with $\left(Z_{1,n}, \ldots, Z_{k,n}\right)$ the nth top-k-list from a uniform $\mathcal{U}(0, 1)$ distribution.

Then, there exist transformations that put in relation the distribution of a top-k-list from an arbitrary continuous df F with the distribution of top-k-lists from the standard exponential or uniform distributions. So these two special top-k-lists deserve a particular attention. Also, the exact analytic formulas for the densities of top-k-lists, we obtain below, can be helpful for simulations.

3.1 Top-k-lists from exponential distributions

A rv W has the (standard) exponential distribution if its density function is $f_W(w) = \exp(-w)I(w > 0)$. By (8) the density function of $L_n^{*(k)} = \left(Y_{1,n}^*, \ldots, Y_{k,n}^*\right)$, the nth top-k-list from an exponential distribution, is

$$f_{L_n^{*(k)}}(y_1, \ldots, y_k) = \frac{k!k^n}{n!} y_1^n \exp\left(-\sum_{i=1}^{k} y_i\right), \quad \text{for } 0 < y_1 < \cdots < y_k. \tag{11}$$

We have the following representation for top-k-lists from exponential distributions:

Theorem 4 *If $L_n^{*(k)}$ is the nth top-k-list from the standard exponential distribution, then*

$$L_n^{*(k)} \stackrel{d}{=} \left(\xi_{1,n},\ \xi_{1,n} + \xi_2,\ \xi_{1,n} + \xi_2 + \xi_3,\ \ldots,\ \xi_{1,n} + \xi_2 + \cdots + \xi_k\right),$$

where $\xi_{1,n} \sim G(n + 1, k)$ and $\xi_j \sim \mathrm{EXP}(k - j + 1)$, $j = 2, \ldots, k$, are all independent.

Proof In the density of $L_n^{*(k)}$ given in (11), consider the change of variables

$$u_1 = y_1,\ u_2 = y_2 - y_1, \ldots,\ u_k = y_k - y_{k-1}.$$

The transformed density so obtained is

$$g(u_1, \ldots, u_k) = \frac{k^{n+1}}{n!} u_1^n \exp\left(-ku_1\right) \prod_{i=2}^{k} (k - i + 1) \exp\left(-(k - i + 1)u_i\right), \tag{12}$$

for $u_i > 0$, $i = 1, \ldots, k$. Clearly, (12) is the joint density of a vector $(\xi_{1,n},\ \xi_2,\ \xi_3,\ \ldots,\ \xi_k)$ of independent components with $\xi_{1,n} \sim G(n+1, k)$ and $\xi_j \sim \mathrm{EXP}(k-j+1)$, $j = 2, \ldots, k$. □

Recall now that $R_n^{*(k)}$, the nth kth record from an EXP(1) distribution, follows a $G(n+1, k)$ distribution and that the order statistics of an iid sample, $W_1, \ldots, W_m$, from the standard exponential distribution satisfy

$$(W_{1:m},\ \ldots,\ W_{m:m}) \stackrel{d}{=} \left(\frac{\eta_1}{m},\ \frac{\eta_1}{m} + \frac{\eta_2}{m-1},\ \ldots,\ \frac{\eta_1}{m} + \frac{\eta_2}{m-1} + \cdots + \eta_m\right),$$

with independent $\eta_i \sim \mathrm{EXP}(1)$, $i = 1, \ldots, m$ – see for instance Nevzorov (2001), Representation 3.4. Using these results and Theorem 4, we can obtain the following representation for the distribution of standard exponential top-k-lists:

$$L_n^{*(k)} \stackrel{d}{=} R_n^{*(k)}\mathbf{1}_k + (0,\ W_{1:k-1},\ \ldots,\ W_{k-1:k-1})$$

or

$$L_n^{*(k)} \stackrel{d}{=} R_{n-1}^{*(k)}\mathbf{1}_k + (W_{1:k},\ \ldots,\ W_{k:k}), \tag{13}$$

where $R_n^{*(k)}$ and $R_{n-1}^{*(k)}$ are, respectively, nth and $(n-1)$th kth records from a standard exponential sequence of observations independent of $(W_1, \ldots, W_k)$. These representations are useful to obtain marginal distributions and to establish limit results as we will see later. For instance, by (13), $Y_{j,n}^* \stackrel{d}{=} R_{n-1}^{*(k)} + W_{j:k}$, $j = 1, \ldots, k$. Thus the density of $Y_{j,n}^*$ can be obtained by convolution of the respective densities of $R_{n-1}^{*(k)}$ and $W_{j:k}$ as

$$f_{Y_{j,n}^*}(y) = \frac{k^n k!}{(n-1)!(j-1)!(k-j)!} \mathrm{e}^{-(k-j+1)y} \int_0^y x^{n-1} \left(\mathrm{e}^{-x} - \mathrm{e}^{-y}\right)^{j-1} \mathrm{d}x, \quad y > 0.$$

Also the representation from Theorem 3 is quite useful for computing means, variances and covariances of elements of the top-k-list for exponential distribution. In this way we easily get

$$E\left(Y_{i,n}^*\right) = \frac{n}{k} + \sum_{l=1}^{i} \frac{1}{k-l+1}, \quad i = 1, 2, \ldots, k,$$

$$\mathrm{Var}\left(Y_{i,n}^*\right) = \frac{n}{k^2} + \sum_{l=1}^{i} \frac{1}{(k-l+1)^2}, \quad i = 1, 2, \ldots, k,$$

$$\mathrm{Cov}\left(Y_{i,n}, Y_{j,n}\right) = \frac{n}{k^2} + \sum_{l=1}^{i \wedge j} \frac{1}{(k-l+1)^2}, \quad i, j = 1, 2, \ldots, k,$$

where $i \wedge j$ denotes minimum of two numbers i and j.

For numerical purposes a more explicit formula can be derived expanding the above integral or using Laplace transform techniques:

$$f_{Y^*_{j,n}}(y) = \sum_{r=1}^{n+1} \gamma_r^{(j)} g(y;\ r, k) + \sum_{m=1}^{j-1} \delta_m^{(j)} g(y;\ 1, k-m),$$

where g is the gamma density

$$g(y;\ p, a) = \frac{a^p}{\Gamma(p)} \exp(-ay) y^{p-1}, \quad \text{for } y,\ p,\ a > 0,$$
$$\delta_m^{(j)} = \frac{(-1)^{j-m-1}}{m} \binom{k}{j} \binom{j-1}{m} j \left(\frac{k}{m}\right)^n,$$

and the $\gamma_r^{(j)}$s satisfy the recurrent formula:

$$\gamma_r^{(j+1)} = -\frac{k-j}{j} \sum_{s=r}^{n+1} \left(\frac{k}{j}\right)^{s-r} \gamma_s^{(j)},\ r = 1, \ldots, n+1,\ j \in \{1, \ldots, k\}$$

with $\gamma_{n+1}^{(1)} = k^{n+1}$ and $\gamma_s^{(1)} = 0$, for $s = 1, \ldots, n$.

3.2 Top-k-lists from uniform distributions

Let us consider the case in which the parent distribution is uniform in the interval $(0, 1)$, $\mathcal{U}(0, 1)$. Let $U_{i:n}$ be the ith order statistic from an iid sample of size n from $\mathcal{U}(0, 1)$, $i = 1, \ldots, n$. The nth top-k-list will be denoted $\hat{L}_n^{(k)} = \left(Z_{1,n}, \ldots, Z_{k,n}\right)$. From Theorem 2(c),

$$f_{\hat{L}_n^{(k)}}(z_1, \ldots, z_k) = \frac{k! k^n}{n!} \left(-\log(1 - z_1)\right)^n, \quad \text{for } 0 < z_1 < \cdots < z_k < 1. \quad (14)$$

In the next theorem we present different representations related to the distribution of top-k-lists from uniform distributions:

Theorem 5 *Let $L_n^{(k)} = \left(Z_{1,n}, \ldots, Z_{k,n}\right)$ be the nth top-k-list from a uniform distribution $\mathcal{U}(0, 1)$. Then*

(i)

$$\left(-k \log(1 - Z_{1,n}),\ \frac{Z_{2,n} - Z_{1,n}}{1 - Z_{1,n}},\ \frac{Z_{3,n} - Z_{2,n}}{1 - Z_{1,n}}, \ldots,\ \frac{Z_{k,n} - Z_{k-1,n}}{1 - Z_{1,n}}\right)$$
$$\sim G(n+1, 1) \otimes \mathcal{U}(S_{k-1}),$$

where $\mathcal{U}(S_{k-1})$ (a special example of the Dirichlet distribution) is a uniform distribution on the $(k-1)$-dimensional unit simplex

$$S_{k-1} = \{x = (x_1, \ldots, x_{k-1}) \in (0, \infty)^{k-1} :\ x_1 + \cdots + x_{k-1} < 1\}.$$

Thus the first component of the above random vector and the subvector of all remaining components are independent.

(ii)

$$\left(-k\log\left(1-Z_{1,n}\right), \left(\frac{1-Z_{2,n}}{1-Z_{1,n}}\right)^{k-1}, \left(\frac{1-Z_{3,n}}{1-Z_{2,n}}\right)^{k-2}, \ldots, \frac{1-Z_{k,n}}{1-Z_{k-1,n}}\right) \sim G(n+1,1)\otimes\mathcal{U}(0,1)\otimes\cdots\otimes\mathcal{U}(0,1).$$

(iii)

$$Z_{j,n} \stackrel{d}{=} \begin{cases} 1-\exp\left(-\frac{G_{1,n}}{k}\right), & j=1 \\ 1-\exp\left(-\frac{G_{1,n}}{k}\right)\prod_{i=1}^{j-1} U_i^{1/k-i} & j=2,\ldots,k \end{cases},$$

where $G_{1,n}\sim G(n+1,1)$ and $U_i\sim\mathcal{U}(0,1)$, $i=1,\ldots,k-1$, are independent.

(iv)

$$\left(\frac{Z_{2,n}-Z_{1,n}}{1-Z_{1,n}}, \frac{Z_{3,n}-Z_{1,n}}{1-Z_{1,n}}, \ldots, \frac{Z_{k,n}-Z_{1,n}}{1-Z_{1,n}}\right) \stackrel{d}{=} (U_{1:k-1},\ldots,U_{k-1:k-1}).$$

Proof Observe that *(i)* follows from (14) by a change of variables

$$(z_1,\ldots,z_k)\to(y_1,\ldots,y_k)=\left(-k\log(1-z_1), \frac{z_2-z_1}{1-z_1}, \ldots, \frac{z_k-z_1}{1-z_1}\right)\in S_{k-1},$$

with the jacobian of the inverse equal to e^{-y_1}/k. Also *(ii)* follows similarly by taking a proper transformation or can be deduced from properties of the Dirichlet distribution. Now *(iii)* is an immediate consequence of *(ii)*. Finally *(iv)* follows by transforming a subvector of all components except the first one of the k-variate vector from either *(i)* or *(ii)*. □

The product representation given in *(iii)* of Theorem 4 is useful for computation of expectations, variances and covariances. We get

$$E\left(Z_{i,n}\right) = 1-\frac{k^n(k-i+1)}{(k+1)^{n+1}}, \quad i=1,2,\ldots,k,$$

$$\mathrm{Var}\left(Z_{i,n}\right) = \frac{k^n(k-i+2)(k-i+1)}{(k+2)^{n+1}(k+1)} - \left(\frac{k^n(k-i+1)}{(k+1)^{n+1}}\right)^2, \quad i=1,2,\ldots,k,$$

$$\mathrm{Cov}\left(Z_{i,n},Z_{j,n}\right) = \left[\frac{k^n(k-(i\wedge j)+2)(k-(i\wedge j)+1)}{(k+2)^{n+1}(k+1)} - \left(\frac{k^n(k-(i\wedge j)+1)}{(k+1)^{n+1}}\right)^2\right] \times\frac{k-(i\vee j)+1}{k-(i\wedge j)},$$

where $i,j=1,2,\ldots,k,,i\neq j$ and $i\vee j$ denotes the maximum of the numbers i and j.

4 Limit results for top-k-lists

Resnick (1973) showed that there are only three types of distributions that can arise as non-degenerate limits of suitably normalized record values (below Φ denotes the distribution function of a standard normal distribution $\mathcal{N}(0, 1)$):

(1) Log-normal type with the df

$$\Phi_\alpha(x) = \begin{cases} 0 & x < 0 \\ \Phi\left(\log x^\alpha\right) & x \geq 0 \end{cases}, \quad \alpha > 0.$$

(2) Negative log-normal type with the df

$$\widetilde{\Phi}_\alpha(x) = \begin{cases} 0 & x < 0 \\ \Phi\left(\log\left(-x\right)^{-\alpha}\right) & x \geq 0 \end{cases}, \quad \alpha > 0.$$

(3) Normal type with the df

$$\Phi(x).$$

If G is one of the possible limit dfs described above, we say that a df F is in the domain of record attraction of G, and write $F \in D_R(G)$, if there exist normalizing sequences (a_n) and (b_n) of real numbers such that

$$\lim_n P\left(\frac{R_n - a_n}{b_n} \leq x\right) = G(x), \quad x \in \mathbb{R},$$

where $(R_n)_{n\geq 0}$ denotes the sequence of ordinary records (1st records) from the parent distribution F.

The next result, proved in Resnick (1973), gives a characterization of the distributions in each one of the domains of record attraction. We reproduce it in the next lemma since it is of crucial importance for the proof of asymptotic behaviour of top-k-lists.

Lemma 6 (Resnick) *Given a continuous df F consider the function*

$$\psi_F(u) = Q_F\left(1 - \exp(-u)\right), \quad u > 0.$$

Then for $\alpha > 0$

(1) $F \in D_R(\Phi_\alpha)$ if and only if for all $x > 0$

$$\lim_{s\to\infty} \frac{\psi_F^{-1}(sx) - \psi_F^{-1}(s)}{\sqrt{\psi_F^{-1}(s)}} = \alpha \log(x).$$

(2) $F \in D_R(\widetilde{\Phi}_\alpha)$ if and only if for all $x > 0$

$$\lim_{\epsilon\to 0^+} \frac{\psi_F^{-1}(x_0 - \epsilon x) - \psi_F^{-1}(x_0 - \epsilon)}{\sqrt{\psi_F^{-1}(x_0 - \epsilon)}} = -\alpha \log(x),$$

where $x_0 = Q_F(1)$ is finite.

(3) $F \in D_R(\Phi)$ *if and only if for all real* x

$$\lim_{s\to\infty} \frac{\psi_F\left(s + x\sqrt{s}\right) - \psi_F(s)}{\psi_F\left(s + s\sqrt{s}\right) - \psi_F(s)} = x.$$

The main result of this section, given in the theorem below, shows that these three limiting types of univariate distribution govern also the asymptotic laws of k-variate random vectors of top-k-lists.

Theorem 7 *Let* $L_n^{(k)}$ *be the nth top-k-list from the iid sequence with a df* F.

(i) If $F \in D_R(\Phi_\alpha)$ *then*

$$\frac{L_n^{(k)}}{\psi_F(n/k)} \xrightarrow{d} Z_1 \mathbf{1}_k, \quad as\ n \to \infty,$$

where Z_1 *has the df* $\Phi_{\alpha\sqrt{k}}$.

(ii) If $F \in D_R(\widetilde{\Phi}_\alpha)$, *then*

$$\frac{L_n^{(k)} - x_0\mathbf{1}_k}{x_0 - \psi_F(n/k)} \xrightarrow{d} Z_2 \mathbf{1}_k, \quad as\ n \to \infty,$$

where Z_2 *has the df* $\widetilde{\Phi}_{\alpha\sqrt{k}}$.

(iii) If $F \in D_R(\Phi)$, *then*

$$\frac{\sqrt{k}\left(L_n^{(k)} - \psi_F(n/k)\mathbf{1}_k\right)}{\psi_F\left(n/k + \sqrt{n/k}\right) - \psi_F(n/k)} \xrightarrow{d} Z_3 \mathbf{1}_k, \quad as\ n \to \infty,$$

where Z_3 *is standard normal.*

Proof First, we will establish the limit result for top-k-lists from the standard exponential distributions. Let $L_n^{*(k)}$ be the nth top-k-list from EXP(1) distribution. Then according to the representation given in Theorem 4

$$L_n^{*(k)} \stackrel{d}{=} \xi_{1,n}\mathbf{1}_k + \mathbf{V}, \tag{15}$$

with $\mathbf{V} = (0,\ \xi_2,\ \xi_2 + \xi_3, \ldots, \xi_2 + \cdots + \xi_k)$, where $\xi_{1,n} \sim G(n+1, k)$ and $\xi_j \sim \text{EXP}(k - j + 1)$, $j = 2, \ldots, k$, are independent. By the classical central limit theorem [note that $\xi_{1,n}$ is a sum of $n+1$ iid exponential EXP(k) rvs]

$$\frac{k}{\sqrt{n}}\left(\xi_{1,n} - \frac{n}{k}\right) \xrightarrow{d} Z \sim \mathcal{N}(0,1), \quad \text{as } n \to \infty. \tag{16}$$

Then, from (15), we have

$$J_n^{(k)} := \frac{k}{\sqrt{n}}\left(L_n^{*(k)} - \frac{n}{k}\mathbf{1}_k\right) \stackrel{d}{=} \frac{k}{\sqrt{n}}\left(\xi_{1,n} - \frac{n}{k}\right)\mathbf{1}_k + \frac{k}{\sqrt{n}}\mathbf{V}.$$

Since $\mathbf{V}$ does not depend on n, $kn^{-1/2}\mathbf{V} \xrightarrow{P} \mathbf{0}_k$. Consequently (16) implies

$$J_n^{(k)} \xrightarrow{d} Z\mathbf{1}_k \text{ as } n \to \infty, \quad \text{with } Z \sim \mathcal{N}(0,1).$$

Note that if $L_n^{(k)}$ is the nth top-k-list with a parent df F then

$$L_n^{(k)} \stackrel{d}{=} \psi_F\left(L_n^{*(k)}\right) = \psi_F\left(\frac{\sqrt{n}}{k} J_n^{(k)} + \frac{n}{k}\mathbf{1}_k\right). \tag{17}$$

Here and later on any function $h : \mathbb{R} \to \mathbb{R}$, is naturally extended to the function $h : \mathbb{R}^d \to \mathbb{R}^d$ (we use the same symbol assuming it will not be confusing for readers) by assigning $h(y_1, \ldots, y_d) = (h(y_1), \ldots, h(y_d))$.

In the following, we will prove (i) and (iii). The proof of (ii) is analogous to that of (i) and therefore is skipped.

Proof of (i): We start with an analytical observation which will be used later on in the proof: Let $g, g_n, n = 1, 2, \ldots$ be real continuous functions on $\mathbb{R}^k$ such that $g_n \to g$ pointwise, and $(\mathbf{y}_n)$ is a sequence of points in $\mathbb{R}^k$ which converges to $\mathbf{y} \in \mathbb{R}^k$. Then $g_n(\mathbf{y}_n) \to g(\mathbf{y})$.

This observation is justified by the following argument: For any $\varepsilon > 0$ let $\delta > 0$ be such that $\mathbf{x} \in B(\mathbf{y}, \delta) = \{\mathbf{x} : ||\mathbf{x} - \mathbf{y}|| \le \delta\}$ implies $|g(\mathbf{x}) - g(\mathbf{y})| < \varepsilon/2$. Take now N_1 large enough to have $||\mathbf{y}_n - \mathbf{y}|| \le \delta$ for any $n > N_1$. Consider now the difference $g_n - g$ on the closed ball $B(\mathbf{y}, \delta)$, which is a compact set. Since the functions involved are continuous the convergence of $g_n - g \to 0$ on $B(\mathbf{y}, \delta)$ is uniform, thus one can choose N_2 large enough to have $|g_n(\mathbf{x}) - g(\mathbf{x})| < \varepsilon/2$ for any $n > N_2$ and for any $\mathbf{x} \in B(\mathbf{y}, \delta)$. Now for $n > \max\{N_1, N_2\}$ we get

$$|g_n(\mathbf{y}_n) - g(\mathbf{y})| \le |g_n(\mathbf{y}_n) - g(\mathbf{y}_n)| + |g(\mathbf{y}_n) - g(\mathbf{y})| < \varepsilon.$$

Since $\varepsilon > 0$ was taken arbitrarily the proof of the observation is complete.

Let $\mathbf{x} \in \mathbb{R}^k$. Then, using (17), we get

$$\begin{aligned} P\left(\frac{L_n^{(k)}}{\psi_F(n/k)} \le \mathbf{x}\right) &= P\left(\frac{\psi_F\left(\frac{\sqrt{n}}{k} J_n^{(k)} + \frac{n}{k}\mathbf{1}_k\right)}{\psi_F(n/k)} \le \mathbf{x}\right) \\ &= P\left(J_n^{(k)} \le \frac{k}{\sqrt{n}}\left(\psi_F^{-1}\left(\psi_F(n/k)\mathbf{x}\right) - n/k\mathbf{1}_k\right)\right) = F_{J_n^{(k)}}(\mathbf{z}_{n,k}), \end{aligned} \tag{18}$$

where

$$\mathbf{z}_{n,k} = \sqrt{k}\frac{\psi_F^{-1}\left(\psi_F(n/k)\mathbf{x}\right) - \psi_F^{-1}\left(\psi_F(n/k)\mathbf{1}_k\right)}{\sqrt{\psi_F^{-1}\left(\psi_F(n/k)\right)}}.$$

Using Lemma 6(i), we get $\lim_n \mathbf{z}_{n,k} = \alpha\sqrt{k}\log(\mathbf{x})$. Then, taking limits in both sides of (18) and using the observation given in the beginning of the proof of (i) with $g_n = F_{J_n^{(k)}}$, $g = \Phi\mathbf{1}_k = F_{Z\mathbf{1}_k}$ (here $Z \sim \mathcal{N}(0, 1)$), $y_n = \mathbf{z}_{n,k}$ and $y = \alpha\sqrt{k}\log\,\mathbf{x}$ we obtain:

$$\lim_n P\left(\frac{L_n^{(k)}}{\psi_F\,(n/k)} \le \mathbf{x}\right) = F_{Z\mathbf{1}_k}(\alpha\sqrt{k}\log\,\mathbf{x}).$$

To finish the proof of (i), note that $F_{Z\mathbf{1}_k}(\alpha\sqrt{k}\log\,\mathbf{x})$ is the df of the random vector $Z_1\mathbf{1}_k$, where the df of Z_1 is $\Phi_{\alpha\sqrt{k}}$.

Proof of (iii): We will use in the proof the following observation: Let $(Y_n)_{n\geq 1}$ be a sequence of $k$dimensional random vectors such $Y_n \xrightarrow{d} Y$ and let $(g_n)_{n\geq 1}$ be a sequence of continuous functions such that $\lim_n g_n(\mathbf{x}) = g(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^k$, where g is also continuous. Then $g_n(Y_n) \xrightarrow{d} g(Y)$. It follows via the Skorokhod theorem in view of the analytic fact given in the beginning of the proof of (i).

So, note that relying on the representation (17), we have

$$\frac{\sqrt{k}\left(L_n^{(k)} - \psi_F(n/k)\mathbf{1}_k\right)}{\psi_F\left(n/k + \sqrt{n/k}\right) - \psi_F(n/k)} = \frac{\sqrt{k}\left[\psi_F\left(\sqrt{n/k}\frac{J_n^{(k)}}{\sqrt{k}} + n/k\mathbf{1}_k\right) - \psi_F(n/k)\mathbf{1}_k\right]}{\psi_F\left(n/k + \sqrt{n/k}\right) - \psi_F(n/k)}.$$

Now, to get the final result, we use the observation given in the beginning of the proof of (iii) with

$$Y_n = J_n^{(k)} \xrightarrow{d} Y = Z_3 \sim \mathcal{N}(0, 1)$$

and [use Lemma 6(iii)]

$$g_n(\mathbf{x}) = \frac{\sqrt{k}\left[\psi_F\left(\sqrt{n/k}\frac{\mathbf{x}}{\sqrt{k}} + n/k\mathbf{1}_k\right) - \psi_F(n/k)\mathbf{1}_k\right]}{\psi_F\left(n/k + \sqrt{n/k}\right) - \psi_F(n/k)} \longrightarrow g(\mathbf{x}) = \mathbf{x} \in \mathbb{R}^k.$$

□

5 Top-k-lists and generalized order statistics

As it has already been explained at nth kth record time, $T_n^{(k)}$, a new top-k-list is created in the following way: the first element of the $(n-1)$th list, $Y_{1,n-1} = R_{n-1}^{(k)}$ is removed from the list and the observation $X_{T_n^{(k)}}$ enters the list at the appropriate position. So, at time $T_n^{(k)}$ the rvs in the vector

$$B_n^{(k)} = (R_0^{(k)}, R_1^{(k)}, \ldots, R_{n-1}^{(k)}, Y_{1,n}, Y_{2,n}, \ldots, Y_{k,n})$$

are the ordered elements that have belonged to any of top-k-lists which appeared up to the instant $T_n^{(k)}$. That is, $B_n^{(k)}$ represents the ordered largest $n+k$ observations form the sequence (X_i) up to $T_n^{(k)}$.

The joint density of $B_n^{(k)}$ can be written as

$$\begin{aligned} &f_{B_n^{(k)}}(x_0, \ldots, x_{n-1}, y_1, \ldots, y_k) \\ &\quad = g_1(y_2, \ldots, y_k | x_0, \ldots, x_{n-1}, y_1) g_2(x_0, \ldots, x_{n-1}, y_1) \end{aligned} \tag{19}$$

for $x_0 < \cdots < x_{n-1} < y_1 < \cdots < y_k$, where g_2 is the joint density of $(R_0^{(k)}, R_1^{(k)}, \ldots, R_{n-1}^{(k)}, R_n^{(k)})$ given by

$$g_2(x_0, x_1, \ldots, x_{n-1}, y_1) = k^n [1 - F(y_1)]^{k-1} f(y_1) \prod_{i=0}^{n-1} \frac{f(x_i)}{1 - F(x_i)} \tag{20}$$

[the formula follows immediately from the joint density of ordinary records taken from the iid sequence with the parent df $1-(1-F)^k$ instead of F] and g_1 is the conditional density of $(Y_{2,n},\ldots,Y_{k,n})$ given $(R_0^{(k)},R_1^{(k)},\ldots,R_{n-1}^{(k)},Y_{1,n})$. By the obvious extension of the argument used in the proof of Theorem 2 it follows that

$$g_1(y_2,\ldots,y_k|x_0,\ldots,x_{n-1},y_1)=(k-1)!\prod_{j=2}^{k}\frac{f(y_j)}{1-F(y_1)}. \tag{21}$$

Then plugging (20) and (21) into (19) we get

$$f_{B_n^{(k)}}(x_0,\ldots,x_{n-1},y_1,\ldots,y_k)=k^n k!\left(\prod_{i=0}^{n-1}\frac{f(x_i)}{1-F(x_i)}\right)\prod_{j=1}^{k}f(y_j) \tag{22}$$

for $x_0<\cdots<x_{n-1}<y_1<\cdots<y_k$. A direct inspection of the density (22) indicates that $B_n^{(k)}$ is distributed according to the generalized order statistics law of $(X(j,n+k,\tilde{m},1),\ j=1,\ldots,n+k)$, with $\tilde{m}$ being an $(n+k-1)$-dimensional vector with the first n components equal -1 and the last $k-1$ components equal 0 – see Kamps (1995). Under this perspective, the distribution of a top-k-list can be viewed as a marginal distribution of particular generalized order statistics:

$$L_n^{(k)}\stackrel{d}{=}(X(j,n+k,\tilde{m},1),\ j=n+1,\ldots,n+k).$$

Acknowledgements The authors are grateful for the referees' remarks which were helpful in preparing the final version of the paper. In particular the remark at the end of Section 2 is due to one of the referees. This research has been partially financed by Grants FQM-331 and MTM-2004-0909.

References

Arnold BC, Balakrishnan N, Nagaraja HN (1998) Records. Wiley, New York
David HA, Nagaraja HN (2003) Order statistics. Wiley, New York
Kamps U (1995) A concept of generalized order statistics. Teubner, Stuttgart
Nevzorov VB (2001) Records: a mathematical theory. American Mathematical Society, Providence
Resnick SI (1973) Limit laws for record values. Stoch Process Appl 1:67–87