Extreme value distributions for weakly correlated fitnesses in block models

Kavita Jain
Theoretical Sciences Unit and Evolutionary and Organismal Biology Unit, Jawaharlal Nehru Centre for Advanced Scientific Research, Jakkur PO, Bangalore 560064, India
E-mail: jain@jncasr.ac.in

Received 21 February 2011
Accepted 30 March 2011
Published 27 April 2011

Abstract. We study the limit distribution of the largest fitness for weakly correlated and identically distributed random fitnesses. A fitness variable is obtained by taking a linear combination of a fixed number of independent random variables drawn from the same parent distribution and two fitnesses are correlated if they have at least one common term in the respective sum. We find that for certain classes of parent distributions, the extreme value distribution for correlated random variables can be related either to one of the known limit laws for independent variables or to the parent distribution itself. For other cases, new limiting distributions appear. The conditions under which these results hold are identified.

Keywords: stochastic processes (theory), mutational and evolutionary processes (theory)
1. Introduction

Extreme value theory [1] deals with the smallest or the largest of a set of random variables \( \{x_1, \ldots, x_N\} \) and has found numerous applications in physics [2], engineering [3], biology [4] and finance [5]. If the variables \( x_i \) are independent and identically distributed (i.i.d.) according to a continuous distribution \( p(x) \), the probability of the \( k \)th maximum is given by [1]

\[
\tilde{P}_{N}^{(k)}(x) = \frac{(1 - q(x))^{k-1} p(x) q(x)^{N-k}}{B(k, N-k+1)}
\]

where the cumulative distribution \( q(x) = \int_{0}^{x} dy p(y) \) and \( B(m, n) \) is the beta function. A classic result in extreme value theory of i.i.d. random variables states that for large \( N \) (and \( k \ll N \)), the distribution \( \tilde{P}_{N}^{(k)}(x) \) is of the following scaling form [1]:

\[
\tilde{P}_{N}^{(k)}(x) \approx \frac{1}{\tilde{a}_N} \tilde{F}^{(k)} \left( \frac{x - \tilde{b}_N}{\tilde{a}_N} \right).
\]

In the above expression, while the location factor \( \tilde{b}_N \) and the scale factor \( \tilde{a}_N \) depend on the details of the parent distribution \( p(x) \), the scaling function \( \tilde{F}^{(k)} \) is determined by the large \( x \) behavior of \( p(x) \). For \( k > 1 \), the distribution \( \tilde{P}_{N}^{(k)}(x) \) can be related to \( \tilde{P}_{N}^{(1)}(x) \) by a transformation and the maximum value distribution itself can be one of the following three types [1].

1. The Fréchet distribution if \( p(x) \) decays as a power law.
2. The Weibull distribution if \( p(x) \) is bounded above.
3. The Gumbel distribution if \( p(x) \) is unbounded above and decays faster than a power law.

doi:10.1088/1742-5468/2011/04/P04020
Much less is known about the extreme value statistics when the random variables are not independent. It is interesting to investigate how the correlations affect the limiting distributions of the three i.i.d. classes mentioned above. If the correlations are strong, one may expect the extreme value distribution to be different from the i.i.d. limit laws and universal distributions may not even exist [2,6]. For weakly correlated variables, on the other hand, the i.i.d. results may still hold. In this paper, we study the extreme value statistics of random variables with weak correlations and identify the relevant parameters depending on which the extreme value distribution is seen to be one of the suitably rescaled i.i.d. limit distributions or the (rescaled) parent distribution or a completely different distribution unrelated to these.

The extreme value statistics of correlated random variables has been studied for stationary Gaussian process with correlations decaying logarithmically or faster [7] and specific physical examples such as the random energy model with correlated random potential [8,9], 1/f noise [10], the directed polymer on a Cayley tree [11], fluctuating interfaces [12] and mass transport models [13]. Here we study this problem in the context of biological evolution where the question of extreme values naturally arises as a large population evolves to maximize its fitness. During the first step in the evolution, an individual in an initially unfit population may acquire one or more mutations. If the total population is infinitely large, the subpopulation with \( D \) mutations is dominated by the fittest mutant in this subpopulation [14]. As several experiments have shown that the fitnesses are not completely random [15]–[17], we are led to study the extreme value statistics of correlated fitnesses.

In order to examine the implications of correlations amongst fitnesses, several models such as the NK model [18], the block model [17] and the rough Mount Fuji-type model [19] have been proposed. In this paper, we will work with the block model which has been employed in recent years to address questions pertaining to evolutionary dynamics [20]–[22]. The extreme value statistics in the block model has been studied for both strongly and weakly correlated fitnesses for an exponentially decaying parent distribution \( p(x) \) [23]. However the question of the universality of the extreme value distributions was not addressed. Here we extend the previous study by considering a broad class of parent distributions and focus on weakly correlated but identically distributed fitnesses.

The rest of the paper is organized as follows. The block model of correlated fitnesses is defined in section 2. The extreme value distributions are obtained when a single mutation is present in the initial fitness in section 3.1 and in the presence of multiple mutations in section 3.2. Finally a discussion and summary are given in section 4.

2. The block model of correlated fitnesses

Many biomolecules such as proteins, antibodies and enzymes have natural domains or partitions and can be modeled as a sequence of length \( L \) divided into several blocks [17]. In the simplest setting, a sequence represented by a binary string of zeros and ones is divided into \( B \) blocks of equal length \( \ell = L/B, \ 1 \leq \ell \leq L \) (see figure 1). A block configuration is assigned a random fitness value regardless of its position in the sequence. These block fitnesses are chosen independently from a common distribution \( p(f) \) which is nonzero on the interval \([a, b]\) and zero elsewhere. Then the sequence fitness is obtained by averaging over the fitnesses of the blocks in the sequence.
Figure 1. Block model: a binary sequence of length $L = 12$ is divided into two blocks of equal length $\ell = 6$. The sequences on the left and right differing by single mutation have correlated fitnesses as the first block is common to the two sequences.

If two sequences have several blocks in common, their fitnesses will also be similar and hence correlated. An attractive feature of the block model is that the fitness correlations can be tuned with the block length $\ell$. For $\ell = 1$, as two distinct sequences can have up to $L - 1$ blocks in common, sequence fitnesses are maximally correlated, while for $\ell = L$, we obtain the model with maximally uncorrelated fitnesses, as no common blocks are possible [14]. To see how fitness correlations vary with block length, consider a set of sequences carrying one mutation relative to the sequence with all zeros. Then as a single mutation in this sequence leaves $B - 1$ blocks unchanged, the fitness $w_j$ of the one mutant neighbor of the initial sequence is given by

$$ w_j = \frac{(B - 1)f_0 + f_j}{B}, \quad j = 1, \ldots, \ell $$

(3)

where $f_0$ and $f_j$ denote the fitness of the block configuration with all zeros and with $\ell - 1$ zeros and a one at the $j$th locus respectively. Using the fact that the block fitnesses are independently distributed, we find that the correlation amongst the sequence fitnesses $\{w_j\}$ is given by

$$ \langle w_i w_j \rangle - \langle w_i \rangle \langle w_j \rangle = \left[ \frac{(B - 1)^2 + \delta_{i,j}}{B^2} \right] \sigma^2 = \left[ \left( 1 - \frac{\ell}{L} \right)^2 + \delta_{i,j} \left( \frac{\ell}{L} \right)^2 \right] \sigma^2 $$

(4)

where $\sigma^2$ is the variance of the block fitness distribution. It is clear from the last equation that the correlations decay as the block length $\ell$ increases towards $L$. In this paper, we are interested in weakly correlated fitnesses which are obtained when $L \to \infty, \ell \to \infty$ with $B$ fixed.

3. Distribution of the largest fitness

Starting from an initial unfit sequence composed of identical blocks (say, all zeros), we are interested in finding the distribution of the largest fitness amongst $\binom{B}{D}$ sequences with $D$ mutations relative to the initial sequence. In this paper, we focus on the extreme value distribution for nonindependent and identically distributed (ni.i.d.) random variables (for some results on nonindependent and nonidentically distributed fitnesses, see [6, 23]). Such ni.i.d. sequence fitnesses are obtained when either the number of mutations $D = 1$ for any $B > 1$, or the number of blocks $B = 2$ and $D$ is odd. In the first case (discussed in section 3.1), the distribution of sequence fitnesses $w_j$ defined by (3) is given by

$$ \text{Prob}(w_j) = \int_0^\infty \int_0^\infty df_0 df_j p(f_0)p(f_j) \delta \left( w_j - \frac{(B - 1)f_0 + f_j}{B} \right) $$

(5)

which is independent of $j$, and thus the sequence fitnesses with a single mutation are identically distributed. In the second case when $B = 2$ and a sequence has $D$ mutations,
the sequence fitness is obtained by averaging over the fitness of the first block with \(d\) mutations and the second block with \(d' = D - d\) mutations (see section 3.2). On writing down a sequence fitness distribution similar to (5), it is readily verified that only those sequences with different block configurations in the first and second block have identically distributed fitness. Since a sequence with the same configuration in both blocks can occur only for even \(D\), it follows that n.i.i.d. fitnesses are obtained when \(D\) is odd. In the rest of the paper, we assume \(L\) to be an even integer and consider \(D \leq L/2\), as the results for \(D > L/2\) can be obtained on replacing \(D\) by \(L - D\).

### 3.1. A single mutation in the initial sequence

We first consider the extreme value statistics of the fitness set \(\{w_j\}\) defined by (3). For a given \(f_0\), the sequence fitness \(v = ((B - 1)f_0 + f)/B\) is the largest amongst the set \(\{w_j\}\) if \(f = \max\{f_1, \ldots, f_\ell\}\). But as the block fitnesses are i.i.d. random variables, this event occurs with a probability \(\hat{F}_\ell^{(1)}(f)\). Then the probability \(P_{\ell}^{(1)}(w)\) that the largest fitness in the set \(\{w_j\}\) has a value \(w\) can be written as

\[
P_{\ell}^{(1)}(w) = \int_0^\infty \int_0^\infty df_0 df p(f_0) \hat{F}_\ell^{(1)}(f) \delta(w - v).
\]

(6)

In general, the probability \(P_{\ell}^{(k)}(w)\) that the \(k\)th maximum has a value \(w\) is given by

\[
P_{\ell}^{(k)}(w) = \int_0^\infty \int_0^\infty df_0 df p(f_0) \hat{F}_\ell^{(k)}(f) \delta(w - v) = \frac{B}{B - 1} \int_{-\infty}^\infty df p \left( \frac{Bw - f}{B - 1} \right) \hat{F}_\ell^{(k)}(f).
\]

(7)

For large \(\ell\), using (2) we obtain

\[
P_{\ell}^{(k)}(w) \approx \frac{B}{B - 1} \int_{-\infty}^\infty df p \left( \frac{Bw - f}{B - 1} \right) \frac{1}{\tilde{a}_\ell} \hat{F}_\ell^{(k)} \left( \frac{f - \tilde{b}_\ell}{\tilde{a}_\ell} \right).
\]

(8)

The behavior of the distribution function in the above equation can be classified as follows:

(i) If \(\tilde{a}_\ell\) diverges with \(\ell\), it is useful to rewrite (8) as

\[
P_{\ell}^{(k)}(w) \approx B \int_{-\infty}^\infty dz p(z) \frac{1}{\tilde{a}_\ell} \hat{F}_\ell^{(k)} \left( \frac{(1 - B)z + Bw - \tilde{b}_\ell}{\tilde{a}_\ell} \right).
\]

(9)

In the limit \(\ell, w \to \infty\), the ratio \(z/\tilde{a}_\ell\) in the argument of \(\hat{F}_\ell^{(k)}\) above can be ignored and we can write

\[
P_{\ell}^{(k)}(w) \approx \frac{B}{\tilde{a}_\ell} \hat{F}_\ell^{(k)} \left( \frac{Bw - \tilde{b}_\ell}{\tilde{a}_\ell} \right).
\]

(10)

where we have used that \(p(z)\) is normalized to unity. Thus up to a rescaling, the extreme value distribution function for correlated variables falls in the same universality class as the i.i.d. ones if the scale factor \(\tilde{a}_\ell\) diverges. The scaling variable in (10) indicates that the distribution decays faster for correlated variables \((B > 1)\) as one would intuitively expect. An example of a class of block fitness distributions for which \(\tilde{a}_\ell\) diverges with \(\ell\)

doi:10.1088/1742-5468/2011/04/P04020

5
Extreme value distributions for weakly correlated fitnesses in block models

Figure 2. Scaled distribution for the first maximum when the block fitness distribution $p(f) = e^{-\sqrt{f}/(2\sqrt{f})}$ (main) and $e^{-f}$ (inset) for $B = 2$. The points refer to exact integration of (7) for $\ell = 10^2(+)\,10^9(\times)$ and $10^{10}(\bigcirc)$. The solid lines show (10) in the main part and (14) in the inset. The inset also shows the Gumbel distribution for comparison (dotted curve).

is $p(f) = \delta f^{\delta-1} e^{-f^\delta}$. In this case, the Gumbel scaling function is $\tilde{F}^{(1)}(y) = e^{-y} e^{-e^{-y}}$ with the location factor $\tilde{b}_\ell = (\ln \ell)^{1/\delta}$ and the scale factor $\tilde{a}_\ell = \delta^{-1}(\ln \ell)^{1-\delta/\delta}$, which diverges with $\ell$ for $0 < \delta < 1$ [1]. A good agreement is seen between the exact distribution (7) and the asymptotic result (10) in figure 2 for $\delta = 1/2$.

(ii) If the scale factor $\tilde{a}_\ell$ vanishes as $\ell \to \infty$, a change of variables in (8) for $B > 1$ gives

$$P^{(k)}(w) \approx \frac{B}{B - 1} \int_{-\infty}^{\infty} dz \tilde{F}(z)p\left(\frac{\tilde{a}_\ell z + \tilde{b}_\ell - Bw}{1 - B}\right)$$

$$\approx \frac{B}{B - 1} p\left(\frac{Bw - \tilde{b}_\ell}{B - 1}\right), \quad B \neq 1$$

where the last expression is obtained in the limit $\ell, w \to \infty$. Thus the probability distribution of the $k$th maximum is a rescaled parent distribution and independent of $k$ if $\tilde{a}_\ell$ goes to zero with increasing $\ell$. Figure 3 shows that for Gaussian distributed block fitnesses, the asymptotic distribution in (12) approaches the exact distribution in (7) as $\ell$ increases.

(iii) If the scale factor $\tilde{a}_\ell$ is independent of $\ell$ for large $\ell$, one may expect new limiting distributions to arise. For example, for exponentially distributed block fitnesses ($\delta = 1$), the scale factor is unity and the limiting distribution of the $k$th maximum for i.i.d. variables is given by [1]

$$\tilde{P}^{(k)}(x) \approx \tilde{F}^{(k)}(y) = \exp\left(-e^{-y} - ky\right) \sim \begin{cases} e^{-ky}, & y \to \infty \\ e^{-e^{-y}}, & y \to -\infty \end{cases}$$

doi:10.1088/1742-5468/2011/04/P04020
where $y = x - \ln \ell$. Using this result in (8) for large $\ell$, we can write

$$P^{(k)}_\ell(w) \approx \frac{B}{B - 1} \int_0^{Bw} df e^{-(Bw-f)/(B-1)} e^{-\ell - f} \frac{(\ell e^{-f})^k}{(k-1)!}$$

$$= \frac{B}{B - 1} \frac{(\ell e^{-Bw})^{1/(B-1)}}{(k-1)!} \int_0^\ell df e^{-f} f^{k/(B/(B-1))}$$

$$\approx \frac{B}{B - 1} \frac{e^{-u/(B-1)}}{(k-1)!} \Gamma \left( k - 1, \frac{1}{B-1}, e^{-u} \right)$$

(14)

where the last expression is obtained in the scaling limit $\ell, w \to \infty$ keeping $u = Bw - \ln \ell$ finite, and $\Gamma(a, z)$ is the incomplete gamma function [24]. When $u \to -\infty$, an asymptotic expansion for large $e^{-u}$ shows that $P^{(k)}_\ell(w) \sim e^{-u}$ and thus the behavior of the backward tail is similar to that in the i.i.d. case. On the other hand, when $u \to \infty$, the incomplete gamma function is well approximated by the complete gamma function $\Gamma(k - 1/(B - 1), 0)$ except for $k = 1$ and $B = 2$, and thus gives $P^{(k)}_\ell(w) \sim e^{-u/(B-1)}$. For $k = 1$ and $B = 2$, the incomplete gamma function is given by $\Gamma(0, e^{-u}) \approx u - C$ where $C$ is the Euler–Mascheroni constant but the tail of the maximum value distribution decays exponentially fast in this case also. A comparison of the distribution $P^{(1)}_\ell(w)$ with the Gumbel distribution is shown in the inset of figure 2.

3.2. Multiple mutations in the initial sequence

We now turn to the case where a sequence is composed of two blocks and carries odd $D \geq 1$ mutations. For a sequence divided into two blocks of equal length, the sequence fitness is obtained by averaging over the fitness of the first block with $d$ mutations and the second block with $d' = D - d$ mutations. Although the integer $d$ runs from 0 to $D$,
distinct sequence fitnesses are obtained if \( d \leq d_u = (D - 1)/2 \). As the first block can have \( s_d = \binom{d}{2} \) independent fitnesses and the second block can have \( s_{d'} \) fitnesses, the number of distinct sequence fitnesses is given by \( N_D = \sum_{d=0}^{d_u} s_d s_{d'} = (1/2)\binom{L}{D} \) [24] which reduces to \( \ell \) for \( D = 1 \).

We are interested in the distribution \( P_{N_D}^{(1)}(w) \) of the largest fitness \( w \) among such \( N_D \) correlated fitnesses. It is convenient to work with the cumulative distribution \( P_{N_D}^{(1)}(w) \) defined as

\[
P_{N_D}^{(1)}(w) = \int_0^w dw' P_{N_D}^{(1)}(w')
\]

which gives the probability that all the \( N_D \) fitnesses are smaller than \( w \). If \( G_{s_d}(w) \) denotes the probability that the fitness of all the sequences carrying \( d \) mutations in the first block and \( d' \) in the second block is smaller than \( w \), we can write

\[
P_{N_D}^{(1)}(w) = \prod_{d=0}^{d_u} G_{s_d}(w).
\]

Our task thus reduces to finding the distribution \( G_{s_d}(w) \) which can be calculated in a similar manner to the single-mutation case. For a given \( d \), the largest sequence fitness \( v = (f_d + f_{d'})/2 \) is obtained when the block fitnesses \( f_d \) and \( f_{d'} \) of the first and second blocks, respectively, are both the largest amongst the set of \( s_d \) and \( s_{d'} \) i.i.d. random variables. Since the probability \( G_{s_d}(w) \) equals the probability that \( v < w \), we get

\[
G_{s_d}(w) = \int_0^\infty df_d \int_0^\infty df_{d'} \tilde{P}_{s_d}^{(1)}(f_d) \tilde{P}_{s_{d'}}^{(1)}(f_{d'}) \Theta(w - v)
\]

\[
= \int_0^\infty df_d \tilde{P}_{s_d}^{(1)}(f_d) \tilde{P}_{s_d}^{(1)}(2w - f_d)
\]

\[
= \int_0^\infty df_{d'} \tilde{P}_{s_{d'}}^{(1)}(2w - f_{d'}) \tilde{P}_{s_{d'}}^{(1)}(f_{d'})
\]

where the distribution \( \tilde{P}(w) = \int_0^w dw' \tilde{P}(w') \).

In the limit \( \ell \to \infty \), the distributions \( G_{s_0}(w) \) and \( G_{s_d}(w), d > 0 \) need to be analyzed separately. In the former case, since \( s_0 = 1 \), on using (1) and (2) in (18) we have

\[
G_{s_0}(w) \approx \int_{-\infty}^\infty df \frac{1}{\bar{a}_{s_D}} \tilde{F}^{(1)} \left( \frac{f - \bar{b}_{s_D}}{\bar{a}_{s_D}} \right)
\]

which is simply the cumulative maximum value distribution for \( s_D \) random variables when there is a single mutation in the initial sequence, and therefore exhibits a behavior similar to that already discussed in section 3.1. For \( d > 0 \), as both \( s_d, s_{d'} \) are large for \( \ell \gg 1 \), using (2) in (17) we get

\[
G_{s_d}(w) \approx \int_{-\infty}^\infty df \frac{1}{\bar{a}_{s_d}} \tilde{F}^{(1)} \left( \frac{f - \bar{b}_{s_d}}{\bar{a}_{s_d}} \right) \tilde{F}^{(1)} \left( \frac{2w - f - \bar{b}_{s_{d'}}}{\bar{a}_{s_{d'}}} \right)
\]

\[
\approx \tilde{F}^{(1)} \left( \frac{2w - \bar{b}_{s_d} - \bar{b}_{s_{d'}}}{\bar{a}_{s_{d'}}} \right)
\]

doi:10.1088/1742-5468/2011/04/P04020
where $\tilde{F}(w) = \int_0^w dw' \tilde{F}(w')$ and the last expression is obtained if $\tilde{a}_{s_d}/\tilde{a}_{s_d'} \to 0$ as $\ell \to \infty$. The distribution $\tilde{G}_{s_d}(w)$ can be found in a similar manner in the opposite limit $\tilde{a}_{s_d'}/\tilde{a}_{s_d} \to 0$ and we have

$$\tilde{G}_{s_d}(w) \approx \tilde{F}^{(1)} \left( \frac{2w - \tilde{b}_{s_d} - \tilde{b}_{s_d'}}{\tilde{a}_{s_d'}} \right), \quad \text{if } \tilde{a}_{s_d}/\tilde{a}_{s_d'} \to 0 \quad (22a)$$

and

$$\tilde{G}_{s_d}(w) \approx \tilde{F}^{(1)} \left( \frac{2w - \tilde{b}_{s_d} - \tilde{b}_{s_d'}}{\tilde{a}_{s_d}} \right), \quad \text{if } \tilde{a}_{s_d'}/\tilde{a}_{s_d} \to 0. \quad (22b)$$

Thus if the appropriate scale factor ratio vanishes for all $d \leq d_u$, one can express $P_{N_D}^{(1)}(w)$ as a product of the parent distribution $q$ and the i.i.d. functions $\tilde{F}$. Otherwise a novel distribution may be expected. If each integral in the product on the right-hand side (RHS) of (16) can be calculated, a further step is required for $D > 1$ to ascertain whether there is a single scaling variable and the product is reducible to a single function. To illustrate these points, we now apply the above discussion to specific block fitness distributions.

(i) For a class of block fitness distributions $p(f) = \gamma^{-f^{-\gamma}} f^{-1-\gamma}$, $\gamma > 0$, which decay as a power law, the location factor is $\tilde{b}_N = 0$ and the scale factor is $\tilde{a}_N = N^\gamma$, so we have the ratio $\tilde{a}_{s_d}/\tilde{a}_{s_d'} = (s_d/s_d')^{1/\gamma}$. As described below, two distinct cases arise depending on whether the number of mutations $D$ is comparable to $\ell$.

Case a. When $D/\ell \to 0$ as $\ell \to \infty$, using the approximation $s_d \approx \ell d/d!$, we find that $\tilde{a}_{s_d}/\tilde{a}_{s_d'} \approx \ell^{2d-D}$ which vanishes for all $d \leq d_u$ for large $\ell$. Then due to (22a), the cumulative distribution $P_{N_D}^{(1)}(w)$ for large $\ell$ can be written as

$$P_{N_D}^{(1)}(w) \approx \prod_{d=0}^{d_u} \tilde{F}^{(1)} \left( \frac{2w}{\tilde{a}_{s_d'}} \right) = \prod_{d=d_u+1}^{D} \tilde{F}^{(1)} \left( \frac{2w}{\tilde{a}_{s_d}} \right) \quad (23)$$

where the cumulative Fréchet distribution $\tilde{F}^{(1)}(y) = e^{-y^{-\gamma}}, y > 0$, and zero otherwise [1]. In the limit $w, \ell \to \infty$ with $W_n = 2w/\tilde{a}_{s_n}$ fixed, as the Fréchet scaling function is

$$\tilde{F}^{(1)} \left( \frac{2w}{\tilde{a}_{s_m}} \right) \approx \tilde{F}^{(1)} \left( \frac{W_n \tilde{a}_{s_n}}{\tilde{a}_{s_m}} \right) \to \begin{cases} 1 & \text{if } m < n \\ 0 & \text{if } m > n, \end{cases} \quad (24)$$

a nontrivial extreme value distribution is obtained if $W_D = 2w/\tilde{a}_{s_D}$ is kept fixed. Thus we find that the maximum value distribution is given by

$$P_{N_D}^{(1)}(w) \approx \frac{2}{\tilde{a}_{s_D}} \tilde{F}^{(1)} \left( \frac{2w}{\tilde{a}_{s_D}} \right) \quad (25)$$

which is verified in figure 4 for $D = 5$ and $\gamma = 1$. As the above result holds for $D = 1$ as well (see (10)), we conclude that the universality class does not change from the i.i.d. class for $1 \leq D \ll \ell$ for parent distributions decaying as a power law.

Case b. For finite $D/\ell$ in the large $\ell$ limit, we can write

$$\frac{s_d}{s_d'} = \prod_{n=1}^{D-2d} \frac{d+d+n}{\ell - D + d + n} \approx \exp \left[ \ell \int_0^{R-2r} dx \ln \left( \frac{r+x}{1-R + r + x} \right) \right], \quad R - 2r > 0 \quad (26a)$$
Extreme value distributions for weakly correlated fitnesses in block models

Figure 4. Scaled distribution for the first maximum for $p(f) = f^{-2}e^{-1/f}$ and $D = 5$ obtained using exact integration of (16) for $\ell = 40$ ( ), $60$ ( ) and $80$ ( ). The solid line shows the Fréchet scaling function $\tilde{F}^{(1)}(y) = y^{-2}e^{-1/y}$.

and

$$s_d/s_d' = \prod_{n=1}^{D-2d} \frac{d + n}{\ell - D + d + n} \approx \left( \frac{r}{1 - r} \right)^{D-2d}, \quad R - 2r = 0 \tag{26b}$$

where $r = d/\ell$ and $R = D/\ell$. Since the integrand in the exponential on the RHS of (26a) is always negative, $s_d/s_d'$ decays to zero as $\ell$ increases for nonzero $R - 2r$ but remains finite for vanishing $R - 2r$. As a result, the integral in (17) or (18) does not reduce to i.i.d. distributions for $d \sim d_u$ and we may expect novel extreme value distributions for $D \sim \ell$.

(ii) A similar analysis can be carried out for bounded distributions $p(f) = \nu(1 - f)^{\nu-1}$, $\nu > 0$, for which $\tilde{b}_N = 1$ and $\tilde{a}_N = N^{-1/\nu}$. When the number of mutations $D = 1$, a nonuniversal parent distribution is obtained by replacing $\ell$ by $s_D$ in (12) while a rescaled universal distribution holds for $1 < D \ll \ell$. Since we have $\tilde{a}_{s_d}/\tilde{a}_{s_d'} \to 0$ as $\ell \to \infty$, due to (12) and (22b), we get

$$P_{N_0}^{(1)}(w) \approx q(2w - 1) \prod_{d=1}^{d_u} \tilde{F}^{(1)} \left( \frac{2(w - 1)}{\tilde{a}_{s_d}} \right) \tag{27}$$

where the Weibull distribution $\tilde{F}^{(1)}(y) = e^{-(y)^\nu}$, $y \leq 0$, and unity otherwise [1]. As $w \to 1$ and $\ell \to \infty$, while the cumulative distribution $q \to 1$, the distribution $\tilde{F}^{(1)}[2(w - 1)/\tilde{a}_{s_m}]$ approaches zero for $m > n$ and unity for $m < n$ if $W_n = 2(w - 1)/\tilde{a}_{s_m}$ is kept fixed. Then a nontrivial maximum value distribution is obtained if $2(w - 1)/\tilde{a}_{s_{da}}$ remains finite and we obtain

$$P_{N_0}^{(1)}(w) \approx \frac{2}{\tilde{a}_{s_{da}}} \tilde{F}^{(1)} \left( \frac{2(w - 1)}{\tilde{a}_{s_{da}}} \right), \quad 1 < D \ll \ell. \tag{28}$$

doib:10.1088/1742-5468/2011/04/P04020
However for $D \sim \ell$, new extreme value distributions are expected for the reasons mentioned above.

(iii) We finally consider block fitness distributions for which the scale factor ratio typically tends to a finite limit as $\ell \to \infty$. For $p(f) = e^{-f}$ considered in section 3.1, we have the scale factor ratio $\tilde{a}_{sd}/\tilde{a}_{sd'} = (\ln s_{d}/\ln s_{d'})^{(1-\delta)/\delta}$, $\delta \neq 1$. Using Stirling’s approximation for $d > 0$, we may write

$$\frac{\ln s_{d}}{\ln s_{d'}} \approx \frac{r \ln((1-r)/r) - \ln(1-r)}{r' \ln((1-r')/r') - \ln(1-r')} \rightarrow \begin{cases} \text{constant,} & \text{if } D/\ell \to 0 \\ 0, & \text{if } d/\ell \to 0, D/\ell \text{ finite} \\ \text{constant,} & \text{if both } d/\ell, D/\ell \text{ finite} \end{cases} \quad (29)$$

where $r = d/\ell$ and $r' = d'/\ell$. Thus we cannot express $P^{(1)}(w)$ in terms of the known functions even for $1 < D \ll \ell$ when $\delta \neq 1$. But some progress is possible for $\delta = 1$ which we discuss next.

As we have already mentioned, for $p(f) = e^{-f}$ the scale factor is independent of $\ell$ and thus the scale ratio is a constant for any $D$. However it is possible to find $G_{sd}$ for all $d \leq d_u, D \geq 1$, for exponentially distributed fitnesses. Although the maximum value distribution for this case has been studied in a previous work [23], here we present a simpler derivation. For large $\ell$, using (13) for $k = 1$ in (18), we obtain

$$P_{N_d}^{(1)}(w) \approx \int_{0}^{2w} df_0 (1 - e^{-2w+f_0}) s_{D} e^{-f_0} e^{-f_0} \prod_{d=1}^{d_u} \int_{0}^{2w} df_d e^{-s_d e^{-2w+f_d}} s_{d'} e^{-f_d} e^{-f_d} e^{-f_d} (30)$$

$$= \int_{s^D e^{-2w}}^{s_d} df_0 e^{-f_0} \left(1 - s_{D} e^{-2w}/f_0\right) \prod_{d=1}^{d_u} \int_{s_{d}' e^{-2w}}^{s_{d'}} df_d e^{-f_d} e^{-s_d e^{-2w}/f_d} \quad (31)$$

where we have used (19) for the $d = 0$ term. In the limits $w, \ell \to \infty$, the above integrals are nontrivial if $W_d = s_d s_{d'} e^{-2w}$ is finite and we get

$$P_{N_d}^{(1)}(w) \approx \int_{W_0}^{\infty} df_0 e^{-f_0} \left(1 - W_0/f_0\right) \prod_{d=1}^{d_u} \int_{0}^{\infty} df_d e^{-f_d} e^{-W_d/f_d} \quad (32)$$

$$= \left[e^{-W_0} - W_0 \Gamma(0, W_0)\right] \prod_{d=1}^{d_u} 2\sqrt{W_d} K_1(2\sqrt{W_d}) \quad (33)$$

where $K_1(x)$ is the modified Bessel function of the second kind [24]. One can check that the above equation reduces to the cumulative distribution for $D = 1$ obtained in section 3.1 on using $W_0 = e^{-u}$ in (14).\textsuperscript{1} We now consider the above cumulative distribution when $D/\ell$ is zero and when it is finite in the large $\ell$ limit.

Case a. Since the distribution $G_{sd}$ moves towards larger $w$ as $d$ increases (see figure 5), we may try $W_{d_u}$ as a scaling variable to reduce the product in (16) to a single function. For $D \ll \ell$, on using $s_d \approx \ell^d/d!$ for large $\ell$, we see that

$$\frac{W_d}{W_{d_u}} \approx d_u!(D - d_u)!/d!(D - d)! \quad (34)$$

\textsuperscript{1} The first factor in (33) was not taken into account in [23].

doi:10.1088/1742-5468/2011/04/P04020
Extreme value distributions for weakly correlated fitnesses in block models

Figure 5. Cumulative distribution $P^{(1)}_D(w)$ (bold line) for an exponential parent distribution when $D = 21$ and $\ell = 80$ obtained using (30). The other curves show the individual factors $G_{s_d}$ in the product in (30) with $d = 0, \ldots, d_u$ (left to right).

which is finite for all $d < d_u$. Thus unlike for the algebraically decaying or bounded block fitness distributions, all the factors in (33) contribute to the cumulative distribution $P^{(1)}_{N_D}(w)$.

Case b. For $D \sim \ell$, the ratio $W_d/W_{d_u}$ goes to zero for nonzero $R - 2r$ and unity for vanishing $R - 2r$:  

$$\frac{W_d}{W_{d_u}} = \prod_{n=1}^{d_u-d} \frac{d+n}{D-d_u+n} \times \frac{\ell - D + d + n}{\ell - d_u + n}$$  

$$\approx \begin{cases} 0, & R - 2r = 0 \\ \exp \left[ \ell \int_0^{R-2r/2} dx \ln \left( \frac{r + x}{(R/2) + x} \times \frac{1 - R + r + x}{1 - (R/2) + x} \right) \right], & R - 2r > 0 \end{cases}$$  

(35)

Thus although each term in (16) can be calculated (unlike for other block fitness distributions discussed above), most of the terms contribute to the distribution. As shown in figure 5, the distributions $G_{s_d}$ are quite well separated for small $d$ but get clustered at larger $d$, which is consistent with the behavior of $W_d/W_{d_u}$ discussed above.

4. Conclusions

A set of random variables obtained by summing a fixed number of i.i.d. random variables are correlated when they have at least one common term in the sum. Such nonindependent variables may describe diverse quantities such as the breeding value of an animal [25, 26], the fitness of a protein or antibody [17] and the energy of a directed polymer [11]. For two different linear combinations of i.i.d. random variables, we showed that as the sum is over a set of independently distributed variables, it is possible to write the extreme value distribution for weakly correlated random variables as an integral involving the
maximum value distribution of i.i.d. random variables which makes the problem amenable to analysis at least in some cases. Interestingly, even with weak correlations, a rich variety of extreme value distributions result: they can be highly nonuniversal parent distributions or universal extreme value distributions for i.i.d. random variables or novel distributions unrelated to these.

When a single mutation occurs in the initial sequence, the limiting extreme value distribution (up to a rescaling) is found to be one of the i.i.d. distributions when the initial distribution \( p(f) \) decays faster than an exponential and the parent distribution if \( p(f) \) decays slower than an exponential, but a novel distribution appears for exponentially decaying \( p(f) \). Such a classification has also been observed in the context of near-extreme value statistics of i.i.d. random variables \[27\]. The situation is more complex when multiple mutations are introduced in the initial sequence. When the number of mutations \( D \) does not scale with \( \ell \) and so \( D/\ell \to 0 \) for large \( \ell \), our analysis showed that \( D = 1 \) and \( 1 < D \ll \ell \) may exhibit different distributions. While the algebraically decaying parent distributions remain robust in that the universal Fréchet distribution holds for \( 1 \leq D \ll \ell \), for bounded distributions the nonuniversal parent distribution for the single-mutation problem changed to a universal Weibull distribution for \( D > 1 \). For unbounded, non-exponential parent distributions decaying faster than a power law also, the extreme value distribution for the single-mutation case fails to hold for \( D > 1 \). For all the three classes of block fitness distributions considered, novel extreme value distributions are expected if the number of mutations in both blocks is comparable to the block length. For exponentially distributed block fitnesses, we analyzed the maximum value distribution for all \( D \) and gave explicit expressions for two new extreme value distributions (see (14) and (33)). It would be very interesting to see whether they occur in other extreme value problems as well.

Acknowledgment

The author thanks S Sabhapandit for pointing out reference \[27\].

References

[1] David H A and Nagaraja H N, 2003 Order Statistics (New York: Wiley)
[2] Clusel M and Bertin E, Global fluctuations in physical systems: a subtle interplay between sum and extreme value statistics, 2008 Int. J. Mod. Phys. B 22 3311
[3] Katz R W, Parlange M B and Naveau P, Statistics of extremes in hydrology, 2002 Adv. Water Res. 25 1287
[4] Gillespie J H, 1991 The Causes of Molecular Evolution (Oxford: Oxford University Press)
[5] Embrechts P, Klüppelberg C and Mikosch T, 2005 Modelling Extremal Events: for Insurance and Finance (Berlin: Springer)
[6] Seetharaman S and Jain K, Evolutionary dynamics on strongly correlated fitness landscapes, 2010 Phys. Rev. E 82 031109
[7] Leadbetter M R, Lindgren G and Rootzén H, 1983 Extremes and Related Properties of Random Sequences and Processes (Berlin: Springer)
[8] Carpentier D and Le Doussal P, Glass transition of a particle in a random potential, front selection in nonlinear renormalization group, and entropic phenomena in Liouville and sinh–Gordon models, 2001 Phys. Rev. E 63 026110
[9] Fyodorov Y V and Bouchaud J-P, Freezing and extreme-value statistics in potential, 2008 J. Phys. A: Math. Theor. 41 372001
Fyodorov Y V, Le Doussal P and Rosso A, Statistical mechanics of logarithmic REM: duality, freezing and extreme value statistics of 1/f noises generated by Gaussian free fields, 2009 J. Stat. Mech. P10005
[10] Antal T, Droz M, Györgyi G and Rácz Z, 1/f noise and extreme value statistics, 2006 Eur. Phys. Rev. Lett. 87 240601

doi:10.1088/1742-5468/2011/04/P04020
Extreme value distributions for weakly correlated fitnesses in block models

[11] Dean D S and Majumdar S N, Extreme-value statistics of hierarchically correlated variables deviation from Gumbel statistics and anomalous persistence, 2001 Phys. Rev. E 64 046121
[12] Majumdar S N and Comtet A, Exact maximal height distribution of fluctuating interfaces, 2004 Phys. Rev. Lett. 92 225501
[13] Evans M R and Majumdar S N, Condensation and extreme value statistics, 2008 J. Stat. Mech. P05004
[14] Jain K and Krug J, Evolutionary trajectories in rugged fitness landscapes, 2005 J. Stat. Mech. P04008
[15] Carneiro C and Hartl D L, Adaptive landscapes and protein evolution, 2010 Proc. Nat. Acad. Sci. 107 1747
[16] Miller C R, Joyce P and Wichman H A, Mutational effects and population dynamics during viral adaptation challenge current models, 2011 Genetics 187 185
[17] Perelson A S and Macken C A, Protein evolution on partially correlated landscapes, 1995 Proc. Nat. Acad. Sci. 92 9657
[18] Kauffman S A, 1993 The Origins of Order (New York: Oxford University Press)
[19] Aita T, Uchiyama H, Inaoka T, Nakajima M, Kobuto T and Husimi Y, Analysis of a local fitness landscape with a model of the rough Mt. Fuji-type landscape: application to prolyl endopeptidase and thermolysin, 2000 Biopolymers 54 64
[20] Kryazhimskiy S, Tkacik G and Plotkin J B, The dynamics of adaptation on correlated fitness landscapes, 2009 Proc. Nat. Acad. Sci. 106 18638
[21] Orr H A, The population genetics of adaptation on correlated fitness landscapes: the block model, 2006 Evolution 60 1113
[22] Welch J J and Waxman D, The NK model and population genetics, 2005 J. Theor. Biol. 234 329
[23] Jain K, Dasgupta A and Das G, Exact and limit distributions of the largest fitness on correlated fitness landscapes, 2009 J. Stat. Mech. L10001
[24] Gradshteyn I S and Ryzhik I M, 1980 Table of Integrals, Series, and Products (New York: Academic)
[25] Hill W G, Order statistics for correlated variables and implications in genetic selection programmes, 1976 Biometrics 32 889
[26] Rawlings J O, Order statistics for a special class of unequally correlated multinormal variates, 1976 Biometrics 32 875
[27] Sabhapandit S and Majumdar S N, Density of near-extreme events, 2007 Phys. Rev. Lett. 98 140201

doi:10.1088/1742-5468/2011/04/P04020