Extreme value statistics of smooth Gaussian random fields

Stéphane Colombi, Olaf Davis, Julien Devriendt, Simon Prunet, Joe Silk

Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

Mon. Not. R. Astron. Soc. **414,** 2436–2445 (2011) doi:10.1111/j.1365-2966.2011.18563.x

Extreme value statistics of smooth Gaussian random fields

Stéphane Colombi,[1⋆] Olaf Davis,[2] Julien Devriendt,[2] Simon Prunet[1] and Joe Silk[1,2]

[1]*Institut d'Astrophysique de Paris, CNRS UMR 7095 and UPMC, 98 bis, bd Arago, F-75014 Paris, France*
[2]*Department of Astrophysics, University of Oxford, Keble Road, Oxford OX1 3RH*

Accepted 2011 February 18. Received 2011 February 17; in original form 2010 December 22

ABSTRACT

We consider the Gumbel or extreme value statistics describing the distribution function $p_{\rm G}(\nu_{\rm max})$ of the maximum values of a random field ν within patches of fixed size. We present, for smooth Gaussian random fields in two and three dimensions, an analytical estimate of $p_{\rm G}$ which is expected to hold in a regime where local maxima of the field are moderately high and weakly clustered.

When the patch size becomes sufficiently large, the negative of the logarithm of the cumulative extreme value distribution is simply equal to the average of the Euler characteristic of the field in the excursion $\nu \geq \nu_{\rm max}$ inside the patches. The Gumbel statistics therefore represents an interesting alternative probe of the genus as a test of non-Gaussianity, e.g. in cosmic microwave background temperature maps or in 3D galaxy catalogues. It can be approximated, except in the remote positive tail, by a negative Weibull-type form, converging slowly to the expected Gumbel-type form for infinitely large patch size. Convergence is facilitated when large-scale correlations are weaker.

We compare the analytic predictions to numerical experiments for the case of a scale-free Gaussian field in two dimensions, achieving impressive agreement between approximate theory and measurements. We also discuss the generalization of our formalism to non-Gaussian fields.

Key words: methods: analytical – methods: statistical – large-scale structure of Universe.

1 INTRODUCTION

Gumbel or extreme value statistics are concerned with the extrema of samples drawn from random distributions (Gumbel 1958). In the case of sample means, the Central Limit Theorem states that the means of many samples of size N drawn from some distribution will be normally distributed in the large-N limit; analogously, it can be shown that in the same limit the cumulative distribution of the sample maximum or minimum ν will tend to one of the family of the following functions:

$$G_{\gamma_{\rm G}}(\nu) = \exp[-(1+\gamma_{\rm G} y)^{-1/\gamma_{\rm G}}], \quad (1)$$

with

$$y = \frac{\nu - a}{b}, \quad (2)$$

where a and b are location and scale parameters (see e.g. Coles 2001). The shape parameter $\gamma_{\rm G}$ characterizes the distribution: a distribution with $\gamma_{\rm G} = 0$ is known as having the 'Gumbel type',

$$G_0 = \exp[-\exp(-y)], \quad (3)$$

while $\gamma_{\rm G} < 0$ and $\gamma_{\rm G} > 0$ correspond, respectively, to forms of the 'negative Weibull type' and the 'Fréchet type'.

Distributions given by equation (1) have seen application to time-series data in many fields such as climate (see e.g. Katz & Brown 1992), hydrology (see e.g. Katz, Parlange & Naveau 2002), seismology (see e.g. Cornell 1968), insurance and finance (see e.g. Embrechts & Schmidli 1994), etc., in predicting the incidence of extreme events from knowledge of past data. Here we consider applications to 2D and 3D random fields relevant to cosmology, but our approach is sufficiently general so that extension to other fields should not prove difficult.

In three dimensions, one is naturally interested in the occurrence of the most massive clusters in galaxy surveys (Bhavsar & Barrow 1985; Cayón, Gordon & Silk 2010; Holz & Perlmutter 2010; Davis et al. 2011) of large-scale mass concentrations (Yamila Yaryura, Baugh & Angulo 2010) such as the Sloan Great Wall (Gott et al. 2005) or the largest voids observed in the spatial distribution of galaxies. In two dimensions, the most obvious application concerns the temperature fluctuations in the cosmic microwave background (CMB; Mikelsosn, Silk & Zuntz 2009), in particular the analysis of the hottest hotspots (Coles 1988) and the coldest cold spots. There are several works that suggest the existence of anomalies in current CMB experiments (see e.g. Larson & Wandelt 2004; Ayaita et al. 2010), the most prominent one being the cold spot discovered

⋆E-mail: colombi@iap.fr

in the temperature maps measured by the *Wilkinson Microwave Anisotropy Probe* (*WMAP*; Vielva et al. 2004; Cruz et al. 2005).

In this work, we consider a random field in two or three dimensions, smoothed on some scale ℓ, and take large 'patches' of size $L \gg \ell$. The values of the field at all points inside a given patch constitute one sample, and the extreme value of the field in the patch is our statistic of interest (henceforth we restrict ourselves in particular to the maxima, though the case of patch minima is exactly analogous).

Although some of the results present in this paper also apply to the non-Gaussian case, we restrict our practical calculations to Gaussian fields of known power spectrum and ask how to derive analytically the distribution of patch maxima and how to explicitly relate the results to equations (1) and (2). This effort is not new: the calculation of the extreme value distribution of Gaussian fields has been paid a lot of attention by mathematicians, not only for time-series (see e.g. Leadbetter, Lindgren & Rootzén 1983) but also in larger number of dimensions (see e.g. Bickel & Rosenblatt 1973; Rosenblatt 1976; Adler 1981; Aldous 1989; Piterbarg 1996; Adler & Taylor 2007), leading to a number of important results. In particular, in the very large patch limit, convergence to the Gumbel-type distribution (3) was rigorously demonstrated (e.g. Bickel & Rosenblatt 1973). Thus, many of the results derived in this paper can be found in the mathematical literature, but we recall them for the sake of completeness as they are needed to understand how the statistics behave in various regimes.

The method we employ to estimate the extreme value statistics relies on a local maxima approach and was already utilized in a more rigorous mathematical set-up (e.g. Adler & Taylor 2007). The central point is the observation that the probability of the patch maximum being below some specified density threshold is exactly the probability of encountering zero-points *above* that threshold. If we also identify the highest point in the patch with the field highest *peak* there (an assumption which is in fact non-trivial in general; see Section 2.2.1) then the problem is reduced to that of finding the void probability for peaks as a function of threshold.

The distribution and clustering statistics of peaks has achieved a good deal of attention in the astrophysical literature, in part due to their role as the nucleation points of rich clusters of galaxies when the field in question is that of matter overdensities. Combining results obtained previously for the peak abundances and their correlation functions allows us to predict the void probability for both 2D and 3D fields, and hence the extreme value statistics. Once again, although most of the results can be found in the mathematical literature, the key novelty of the present paper lies in the derivation and test of approximate predictions in an intermediate regime where the patch size is large enough compared to the coherence length of the field, but *not so large* that either the asymptotic limit expected for Gaussian random fields (Bickel & Rosenblatt 1973; equation 3) or the Poisson regime (Aldous 1989) has been reached.

This paper is organized as follows. Section 2 sets up the general framework, with a few definitions followed by general results. In particular, the Gumbel statistics is related to the void probability, which is expressed in terms of average number density of peaks above some threshold and their N-point correlation functions. Section 3 focuses on the Gaussian field case, where a general estimate taking into account full clustering of the peaks is performed and explicit asymptotic formulae are derived and related to equation (1). Convergence to equation (3) for Gaussian random fields is recovered when the patch size tends to infinity. An explicit link to the Euler characteristic is also established, in agreement with the literature. In Section 4, we test the theoretical predictions against numerical experiments in the 2D case. Finally, Section 5 summarizes the results obtained in this paper and discusses their generalization to non-Gaussian fields.

2 THEORY

2.1 Definitions

We consider, in a D-dimensional space with $D = 2$ or 3, a random field $\delta(x)$ of zero average. We suppose that this field is statistically stationary (invariance of the N-point correlation functions by translation) and isotropic (invariance of the N-point correlation functions by rotation).

2.1.1 Smoothing window

This field is smoothed with a window of size ℓ:

$$\delta_\ell = \delta * W_\ell(x). \tag{4}$$

For instance, the Gaussian smoothing window $W_\ell(x) \propto \exp(-x^2/2\ell^2)$, which we shall choose for all practical calculations, reads in Fourier space

$$W_\ell(k) = \exp[-(k\ell)^2/2]. \tag{5}$$

The top-hat smoothing window will be needed as well, on scales $L \gg \ell$. In three dimensions, it is a sphere of radius L which reads in Fourier space

$$W_L(k) = 3[\sin(kL) - kL\cos(kL)]/(kL)^3. \tag{6}$$

In two dimensions, it is a disc of radius L which reads in Fourier space

$$W_L(k) = \frac{2J_1(kL)}{kL}, \tag{7}$$

where J_1 is the Bessel function of the first kind and of first order:

$$J_1(x) = \frac{1}{\pi}\int_0^\pi \cos[y - x\sin(y)]\mathrm{d}y. \tag{8}$$

2.1.2 Gumbel statistics

From now on we measure the height of the field using the density contrast in units of its standard deviation,

$$\nu \equiv \frac{\delta_\ell}{\sigma_0}, \tag{9}$$

with

$$\sigma_0^2 \equiv \langle\delta_\ell^2\rangle. \tag{10}$$

We consider a large spherical patch of size L at random position x_0, $L \gg \ell$, and measure in that patch the maximum value of the smoothed density field:

$$\nu_{\max} \equiv \max\{\nu(x); |x - x_0| \le L\}. \tag{11}$$

The goal is to study the Gumbel statistics, i.e. the probability distribution function $p_{\rm G}(\nu_{\max})\mathrm{d}\nu_{\max}$ of the values of $\nu_{\max}$ when we choose an infinite number of random realizations of x_0. This distribution contains a dependence on the choice of smoothing and on the size of the patch,

$$p_{\rm G}(\nu_{\max}) = p_{\rm G}(\nu_{\max}, \ell, L), \tag{12}$$

which we leave implicit in the remainder of the paper.

2.2 General results

In a sufficiently non-degenerate field δ_ℓ, the set of local maxima – the peaks of the density field – is a discrete ensemble of points of positions p_i and density $\nu_i = \delta_i/\sigma_0$. We shall assume that this property is valid in all the subsequent calculations.

2.2.1 Fundamental assumption

The fundamental assumption we make is that the maximum of the density in a patch can be approximated by the density at the highest peak contained in the patch:

$$\nu_{\rm max} \simeq \bar{\nu}_{\rm max} \equiv \max\{\nu_i, |p_i - x_0| \leq L\}. \tag{13}$$

This assumption is valid only in the regime where $L \gg \ell$. Indeed, there can be local maxima outside the patch but sufficiently close to its edge such that the density measured at a point on the edge of the patch is larger than the maximum density measured in the set of peaks contained in the patch. In other words, if we consider the population of local maxima of the density field defined inside the $(D-1)$-dimensional manifold given by the border of the patch of densities $\hat{\nu}_j$, then we have, in reality,[1]

$$\nu_{\rm max} = \max(\bar{\nu}_{\rm max}, \hat{\nu}_j) \geq \bar{\nu}_{\rm max}. \tag{14}$$

Obviously, one expects $\nu_{\rm max}$ to approach $\bar{\nu}_{\rm max}$ as the ratio L/ℓ increases and the ratio of the patch volume to area near its edge decreases.

2.2.2 General expression of the Gumbel statistics

Let us define the cumulative Gumbel distribution by

$$P_{\rm G}(\nu) \equiv {\rm Prob.}(\nu_{\rm max} \leq \nu) \equiv \int_{-\infty}^{\nu} p_{\rm G}(\nu_{\rm max}){\rm d}\nu_{\rm max}. \tag{15}$$

Such a probability, given the assumptions of Section 2.2.1, is also the probability that none of the local maxima contained in the patch is above the threshold. In other words, if we consider the population of local maxima satisfying $\nu_i > \nu$, none of them belongs to the patch. This happens with a probability $P_0(\nu)$, where P_0 is the probability of finding no maxima with normalized density larger than ν inside a spherical cell or a disc of radius L:

$$P_{\rm G}(\nu) = P_0(\nu); \tag{16}$$

hence

$$p_{\rm G}(\nu) = \frac{{\rm d}P_0}{{\rm d}\nu}. \tag{17}$$

The calculation of such a void probability can be performed using standard count-in-cell formalism if the number density $n(\nu_i > \nu)$ and the connected N-point correlation functions,[2] $\xi_N^{\rm p}(x_1, \ldots, x_N)$, of local maxima above the threshold, are known (White 1979; Fry 1985; Balian & Schaeffer 1989; Szapudi & Szalay 1993).

In particular, one can define the averaged correlations over a cell of size L and volume $V = (4\pi/3)L^3$ or area $V = \pi L^2$,

$$\bar{\xi}_N^{\rm p}(L) \equiv \frac{1}{V^N}\int_V {\rm d}^D x_1 \cdots {\rm d}^D x_N \xi_N^{\rm p}(x_1, \ldots, x_N), \tag{18}$$

the normalized cumulants,

$$S_N^{\rm p}(L) \equiv \frac{\bar{\xi}_N^{\rm p}(L)}{\bar{\xi}_2^{\rm p}(L)^{N-1}}, \quad S_1^{\rm p} \equiv S_2^{\rm p} \equiv 1, \tag{19}$$

and

$$N_{\rm c} \equiv nV\bar{\xi}_2^{\rm p}(L). \tag{20}$$

Each of these expressions contains an implicit ν-dependence. The number $N_{\rm c}$ represents the typical number of peaks above the threshold per overdense patch in excess to the average. It measures the deviation from a pure Poisson distribution due to clustering.

With these definitions, the void probability can be written as

$$P_0(\nu_{\rm max}) = \exp\left[-nV\sigma(N_{\rm c})\right], \tag{21}$$

with

$$\sigma(y) = \sum_{N\geq 1}(-1)^{N-1}\frac{S_N^{\rm p}}{N!}y^{N-1}. \tag{22}$$

The challenge is now to relate the statistical properties of the local maxima to that of the underlying density field. This is made difficult by the fact that the void probability depends on the full hierarchy of correlations up to any order: in particular one has to relate the N-point correlation functions of the peaks to the N-point correlation functions of the underlying density field. We denote the latter by $\xi_N(x_1, \ldots, x_N)$, and similarly the averaged N-point correlation functions of the density field by $\bar{\xi}_N(L)$.

This exercise has been performed in detail for random Gaussian fields by Bardeen et al. (1986) (hereafter BBKS) and Bond & Efstathiou (1987) (hereafter BE) in the 3D and 2D cases, respectively, extending earlier calculations of Kaiser (1984) and Politzer & Wise (1984). Note that these latter calculations did not consider statistics of peaks, but more generally of regions of δ_ℓ above the density threshold. However, in the rare event regime considered here, the two approaches should become equivalent (this is discussed in detail in BBKS).

The non-Gaussian case has been examined as well for a quite general class of hierarchical models by (Bernardeau & Schaeffer 1999, hereafter BS). The statistics under consideration in that work was that of overdense cells of size ℓ and not of peaks of the density field smoothed with a Gaussian window of size ℓ. Again, the approach of BS should give the same results as those obtained for peaks in the rare event regime.

2.2.3 Asymptotic expression of $\sigma(y)$

A fundamental result of the calculations of BS is that in the high-peak limit, i.e. $\nu \gg 1$, and at large enough separations, i.e.

$$\frac{\xi_2(x_i, x_j)}{\sigma_0^2} \ll 1, \tag{23}$$

$$\xi_N^{\rm p}(x_1, \ldots, x_N) \simeq \sum_{\rm trees}\sum_{\rm labels}\prod_{\rm links}\xi_2^{\rm p}(x_i, x_j), \tag{24}$$

in the notation of these authors. This expression is valid at least in the framework of the minimal hierarchical-tree model. The trees refer to ensembles of distinct pair associations of elements in the ensemble $\{1, \ldots, N\}$ such that a fully connected structure containing exactly the N elements is constructed without any loop. The labels take into account all the possible combinations of elements in $\{1, \ldots, N\}$ that lead to the same tree topology. In each tree topology, there are always $N-1$ links, by definition. The total number of combinations of all the trees and the labels yields N^{N-2} possibilities. Fig. 1 of

[1] See Adler & Taylor (2007) for a rigourous formulation corresponding to a more general patch shape than just a sphere.

[2] The connected N-point correlation functions are equal to zero for a Gaussian field if $N \geq 3$.

BS can be examined to understand the process. For instance, for the four-point correlation function there are two tree topologies: (i) the 'star' where one point is connected to all the others and (ii) the 'line' where one point is connected to one or two others depending on its position (at the end or in the middle). There are four possible labellings of the star and 12 possible labellings for the line.

Equation (24) also applies to Gaussian fields, independently of the shape of the smoothing window, at least if $\nu \gg 1$ and the following condition, more restrictive than (23) holds:

$$\nu^2 \frac{\xi_2(x_i, x_j)}{\sigma_0^2} \ll 1. \tag{25}$$

Indeed, the unconnected part of the N-point correlation function (the moment), μ_n, is given by

$$\mu_n = \prod_{i>j} \left[\xi_2^{\rm p}(x_i, x_j) + 1\right] \tag{26}$$

in the high-threshold regime (Politzer & Wise 1984; Cline et al. 1987). Extracting the connected part from this expression consists exactly in extracting the ensemble of distinct pair associations in $\{1, \ldots, N\}$ such that the corresponding topology is fully connected. In the large-separation limit, i.e. at leading order in $\xi_2^{\rm p}(x_i, x_j)$, or equivalently if condition (25) is verified, only the tree topologies remain (because they correspond to the minimum power in $\xi_2^{\rm p}$ while being fully connected), and each label for each tree is given the same weight in equation (26), hence leading to equation (24) in that regime.

Equation (24) reads, after volume averaging in a sphere of radius L (BS),

$$S_N^{\rm p}(L) \simeq N^{N-2}. \tag{27}$$

This result applies as well to the general tree-hierarchical model (BS). This means that the function σ defined in equation (22) reads (BS)

$$\sigma(y) = \left(1 + \frac{1}{2}\theta\right) {\rm e}^{-\theta}, \quad \theta {\rm e}^{\theta} = y. \tag{28}$$

Note that when $N_{\rm c} \ll 1$, which occurs at some point for a large enough value of ν for which there are very few peaks above the threshold in average per patch,

$$\sigma(N_{\rm c}) \simeq 1 - N_{\rm c}/2 \tag{29}$$

by definition (equation 22). Therefore, even though we expect the low-end tail of the Gumbel statistics to be affected by the potential crudeness of our approximation of the function $\sigma(y)$, the high-end tail should still be quite well described.

3 THE GAUSSIAN FIELD CASE

Once the function $\sigma(y)$ is specified, one needs to carry out the calculation of the number density of peaks above the threshold as well as their averaged two-point correlation function. The detailed expressions can be found for a Gaussian field in BBKS and BE.

3.1 Shape parameters of the power spectrum: γ and $R_\star$

The important parameters that control the number density of peaks above threshold ν and their two-point correlation function are the moments

$$\sigma_j^2 \equiv \int \frac{k^{D-1}{\rm d}k}{2\pi^{D-1}} P(k) W_\ell^2(k) k^{2j}, \tag{30}$$

where $D = 2$ or 3 is the dimension of the space considered. Then BBKS and BE define the coherence parameter γ as

$$\gamma \equiv \frac{\sigma_1^2}{\sigma_0 \sigma_2} \tag{31}$$

and the scalelength $R_\star$ as

$$R_\star = \sqrt{D} \frac{\sigma_1}{\sigma_2}. \tag{32}$$

For a Gaussian smoothing window and a scale-free power spectrum $P(k)$ given by

$$P(k) = Ak^n, \tag{33}$$

the integrals in equation (30) can be performed analytically, yielding the simple expressions

$$\sigma_0^2 = \left(\frac{\ell}{\ell_0}\right)^{-(n+D)}, \quad \ell_0 = \left[\frac{A}{4\pi^{D-1}} \Gamma\left(\frac{n+D}{2}\right)\right]^{1/(n+D)}, \tag{34}$$

$$\gamma = \sqrt{\frac{n+D}{n+D+2}}, \quad R_\star = \sqrt{\frac{2D}{n+D+2}}\ell, \tag{35}$$

valid for $n > -D$.[3]

For a top-hat smoothing the scaling law (34) remains valid with a different correlation length, which for the 3D case can be written (see e.g. Lokas et al. 1996) as

$$\ell_0 = \left\{\frac{9A\ \Gamma[(n+3)/2]\ \Gamma[(1-n)/2]}{8\pi^{3/2}\ \Gamma[1-n/2]\ \Gamma[(5-n)/2]}\right\}^{1/(n+3)}. \tag{36}$$

On the other hand, we did not find any simple analytic expression of the correlation length for a top-hat smoothing in two dimensions.

3.2 Number density of peaks

The number density of peaks above the threshold is

$$n(\nu) = \int_\nu^\infty {\rm d}\nu' \mathcal{N}(\nu'). \tag{37}$$

This integral is easily performed numerically, using the fact that the function $\mathcal{N}(\nu)$ is given by

$$\mathcal{N}(\nu) = \frac{1}{(2\pi)^{(D+1)/2} R_\star^D} {\rm e}^{-\nu^2/2} G_D(\gamma, \gamma\nu) \tag{38}$$

in D dimensions, with G_3 approximated by equation (4.4) of BBKS and G_2 given by equation (A1.9) of BE. For completeness, we rewrite these equations here:

$$G_3(\gamma, w) \simeq \frac{w^3 - 3\gamma^2 w + (Bw^2 + C_1)\exp(-Aw^2)}{1 + C_2 \exp(-C_3 w)}, \tag{39}$$

with

$$A = \frac{5}{2(9 - 5\gamma^2)}, \tag{40}$$

$$B = \frac{432}{\sqrt{10\pi}(9 - 5\gamma^2)^{5/2}}, \tag{41}$$

$$C_1 = 1.84 + 1.13(1 - \gamma^2)^{5.72}, \tag{42}$$

$$C_2 = 8.91 + 1.27 \exp(6.51\gamma^2), \tag{43}$$

[3] The case $D = 3$ was derived in BBKS, with the expression for the correlation length ℓ_0 given in e.g. Lokas et al. (1996).

$$C_3 = 2.58\exp(1.05\gamma^2) \tag{44}$$

and

$$G_2(\gamma, w) = (w^2-\gamma^2)\left\{1-\frac{1}{2}\mathrm{erfc}\left[\frac{w}{\sqrt{2(1-\gamma^2)}}\right]\right\} + \frac{w(1-\gamma^2)}{\sqrt{2\pi(1-\gamma^2)}}\exp\left[-\frac{w^2}{2(1-\gamma^2)}\right] + \frac{1}{\sqrt{3-2\gamma^2}}\exp\left(-\frac{w^2}{3-2\gamma^2}\right) \times\left\{1-\frac{1}{2}\mathrm{erfc}\left[\frac{w}{\sqrt{2(1-\gamma^2)(3-2\gamma^2)}}\right]\right\}. \tag{45}$$

3.3 Correlation function of peaks

In the large-separation regime the two-point correlation function of the peaks (25) reads, if one neglects contributions from higher order derivatives of $\xi_2(r)$, which is a fair approximation according to BBKS and BE if ξ_2 is a power law of negative index,

$$\xi_2^{\rm p} = \frac{\langle\tilde{\nu}\rangle^2}{\sigma_0^2}\xi_2, \tag{46}$$

where

$$\langle\tilde{\nu}\rangle = \frac{\int_\nu^\infty \tilde{\nu}(\nu')\mathcal{N}(\nu')\mathrm{d}\nu'}{\int_\nu^\infty \mathcal{N}(\nu')\mathrm{d}\nu'}, \tag{47}$$

and the effective threshold $\tilde{\nu}(\nu)$ writes

$$\tilde{\nu} = \nu - \frac{\gamma\theta}{1-\gamma^2}, \tag{48}$$

with θ approximated by equation (6.14) of BBKS in three dimensions:

$$\theta \simeq \frac{3(1-\gamma^2)+(1.216-0.9\gamma^4)\exp[-\gamma/2(\gamma\nu/2)^2]}{\sqrt{3(1-\gamma^2)+0.45+(\gamma\nu/2)^2}+\gamma\nu/2}, \tag{49}$$

and given in two dimensions by

$$\theta = (1-\gamma^2)\frac{H(\gamma,\gamma\nu)}{G_2(\gamma,\gamma\nu)}, \tag{50}$$

where $H(\gamma, w) \equiv \partial G_2/\partial w$ is given by equation (A4.7a) of BE. For completeness,

$$H(\gamma, w) = 2w\left\{1-\frac{1}{2}\mathrm{erfc}\left[\frac{w}{\sqrt{2(1-\gamma^2)}}\right]\right\} + \frac{4(1-\gamma^2)^2}{(3-2\gamma^2)\sqrt{2\pi(1-\gamma^2)}}\exp\left[-\frac{w^2}{2(1-\gamma^2)}\right] - \frac{2w}{(3-2\gamma^2)^{3/2}}\exp\left(-\frac{w^2}{3-2\gamma^2}\right) \times\left\{1-\frac{1}{2}\mathrm{erfc}\left[\frac{w}{\sqrt{2(1-\gamma^2)(3-2\gamma^2)}}\right]\right\}. \tag{51}$$

In the large-threshold limit, we simply have

$$\langle\tilde{\nu}\rangle \to \nu, \quad \nu\to\infty, \tag{52}$$

as derived by Kaiser (1984). After averaging over volume V, the expression of $\bar{\xi}_2^{\rm p}(L)$ is thus simply given by

$$\bar{\xi}_2^{\rm p}(L) = \frac{\langle\tilde{\nu}\rangle^2}{\sigma_0^2}\bar{\xi}_2(L), \tag{53}$$

where $\bar{\xi}_2(L)$ is the averaged two-point correlation function of the underlying density field. It can be derived easily from the power spectrum of the underlying (smoothed) density field, δ_ℓ, using equation (30) with the top-hat window and replacing ℓ with L. The largeness of the patch size, L, compared to the smoothing scale, ℓ, should guarantee that the large-separation approximation (25) is verified in practice. This can be checked a posteriori by examining the range of values of ν where $p_{\rm G}(\nu)$ is significant.

Note that Heavens & Sheth (1999) and Desjacques (2008) (the latter in the large-separation limit) performed the exact calculation of the two-point correlation function of peaks in two and three dimensions, respectively, by taking into account corrections depending on second and fourth derivative of the two-point correlation function of the underlying field. These corrections can be significant on large scales, where the power spectrum of the underlying field significantly deviates from a power law. However, they get progressively smaller with increasing threshold ν, and in the rare event limit in which we are working here, they are probably irrelevant, except perhaps for the low-ν tail of the Gumbel statistics. Still, this assumption should be checked explicitly for the cold dark matter case by assessing the differences in the distribution introduced by computing the two-point correlation function of peaks with and without them. Such an investigation is beyond the scope of this paper, but it should be kept in mind when comparing analytic estimates of the Gumbel statistics to real data. Also note that in this more rigorous context, the simple proportionality relation (53) does not apply anymore.

3.4 Asymptotic regime

A particularly interesting case corresponds to the regime $\nu \gg 1$ and the Poisson limit $N_{\rm c} \ll 1$, where $\sigma(N_{\rm c}) \simeq 1$. Such a regime is expected to be reached if L/ℓ is sufficiently large and has been studied previously (Aldous 1989). However, since they will prove to be very useful to understand the intermediate (as opposed to asymptotic) regime that we discuss later, we recast the main results using our formalism. Here, the calculations will be facilitated by examining the cumulative Gumbel distribution, $P_{\rm G}(\nu)$.

In the large-ν regime, the number density of peaks is proportional to the Euler characteristic $\mathcal{E}$ of the underlying density field[4] (e.g. BBKS, BE). Just how large a value of ν is required for this to be true depends on the level of accuracy one aims to reach in the description of the function $P_{\rm G}(\nu)$. For instance, BBKS suggest $\gamma\nu > 2.5$ in three dimensions for a 10 per cent accuracy on approximating $n(\nu)$ by the Euler characteristic.

With the additional assumption that $N_{\rm c} \ll 1$, equations (4.14) of BBKS and equation (3.3) of BE read, respectively, in three and two dimensions,

$$P_{\rm G,3}(\nu) \simeq \exp(-\mathcal{E}_3 V) = \exp\left[-U_3(\nu^2-1)\exp\left(-\frac{\nu^2}{2}\right)\right], \tag{54}$$

$$P_{\rm G,2}(\nu) \simeq \exp(-\mathcal{E}_2 V) = \exp\left[-U_2\,\nu\exp\left(-\frac{\nu^2}{2}\right)\right], \tag{55}$$

with

$$U_D = \frac{\gamma^D V}{(2\pi)^{(D+1)/2}R_\star^D}. \tag{56}$$

[4] The Euler characteristic is seen here as an alternate count of critical point number densities of various kinds included in the excursion in regions with normalized density larger than ν, see e.g. Colombi, Pogosyan & Souradeep (2000) and Adler & Taylor (2007).

For scale-free power spectra we have

$$U_D = \left(\frac{4}{3}\right)^{D-2} \frac{\pi}{(2\pi)^{(D+1)/2}} \left(\frac{n+D}{2D}\right)^{D/2} \left(\frac{L}{\ell}\right)^D. \quad (57)$$

Note that in the original derivation (Aldous 1989), the right-hand part of equation (54) contains a term scaling like ν^2 and not $\nu^2 - 1$. Recall however that these expressions still assume ν to be sufficiently large compared to unity. Note also that in the very large threshold limit, one recovers the classical result (see Adler 1981; Adler & Taylor 2007)

$$1 - P_{\rm G,D}(\nu \gg 1) \simeq \mathcal{E}_D V. \quad (58)$$

In the calculations presented in Adler & Taylor (2007), though, the edge effects discussed in Section 2.2.1 are not neglected, i.e. $\mathcal{E}_D V$ must be in fact viewed as the ensemble average of the Euler characteristic of the intersection between the excursion and the patch.[5]

An interesting value of ν corresponds to

$$n(\nu_\star)V = 1, \quad (59)$$

or $P_{\rm G} = 1/{\rm e}$. Obviously we must have $\nu_\star$ sufficiently large compared to unity for equations (54) and (55) to hold, as well as

$$N_{\rm c}(\nu_\star) = \frac{\nu_\star^2}{\sigma_0^2}\bar{\xi}_2(L) \ll 1, \quad (60)$$

to remain in the Poisson limit. The last condition imposes a constraint on the size L of the patch, which must be large enough compared to the smoothing scale ℓ. This obviously depends on spectral index: the ratio L/ℓ should be larger if n is small since there is more power on large scale.

Asymptotically, $\nu_\star$ reads

$$\nu_\star \simeq \sqrt{2\ln U_D}\left[1 + \frac{(D-1)\ln(2\ln U_D)}{4\ln U_D}\right]. \quad (61)$$

This equation shows that $\nu_\star$ grows rather slowly with L/ℓ.

Now we compare the expressions (54) and (55) to the standard cumulative form (1). To determine the parameters a, b and $\gamma_{\rm G}$ of equations (1) and (2), we perform a second-order Taylor expansion near $y = 0$, where $G_{\gamma_{\rm G}}(y = 0) = 1/{\rm e}$.

At second order in $\gamma_{\rm G} y$ we have

$$\ln(-\ln G_{\gamma_{\rm G}}) \simeq -y + \gamma_{\rm G}\frac{y^2}{2}. \quad (62)$$

Similarly we have

$$\ln(-\ln P_{G,3}) \simeq -\frac{a^2}{2} + \ln[U(a^2-1)] - \frac{ab(a^2-3)}{a^2-1}y - \frac{b^2(a^4+3)}{(a^2-1)^2}\frac{y^2}{2}, \quad (63)$$

$$\ln(-\ln P_{G,2}) \simeq -\frac{a^2}{2} + \ln(Ua) - \frac{b(a^2-1)}{a}y - \frac{b^2(a^2+1)}{a^2}\frac{y^2}{2}. \quad (64)$$

Our particular choice of the expansion is convenient since it implies

$$a = \nu_\star. \quad (65)$$

[5] Note that having a very accurate determination of the high-ν tail of the Gumbel distribution has been paid a lot of attention by mathematicians and numerous methods have been employed to do so, estimating the Euler characteristic being one amongst them (see the introduction of Azaïs & Delmas 2002, for a panorama on various methods).

Then

$$b_3 = \frac{1}{\nu_\star}\frac{\nu_\star^2 - 1}{\nu_\star^2 - 3}, \quad (66)$$

$$b_2 = \frac{\nu_\star}{\nu_\star^2 - 1}, \quad (67)$$

$$\gamma_{\rm G,3} = -\frac{\nu_\star^4 + 3}{\nu_\star^2(\nu_\star^2-3)^2} < 0, \quad (68)$$

$$\gamma_{\rm G,2} = -\frac{\nu_\star^2 + 1}{(\nu_\star^2-1)^2} < 0, \quad (69)$$

where the labels 2 and 3 refer to the number of dimensions considered, D. It is interesting to study the asymptotic values of these parameters when $\ln U_D \gg 1$; hence when $\nu_\star$ is very large,

$$b \sim 1/\nu_\star, \quad \gamma_{\rm G} \sim -1/\nu_\star^2, \quad (70)$$

thus

$$\gamma_{\rm G} y \sim \frac{\nu}{\nu_\star} - 1. \quad (71)$$

We can thus understand that the range of validity of the Taylor expansion translated in terms of the ν variable is $\nu \in [\nu_\star(1-\varepsilon), \nu_\star(1+\varepsilon)]$, where ε is a fraction of unity, for both $P_{\rm G}(\nu)$ and $G_{\gamma_{\rm G}}(\nu)$. While $\nu_\star$ does not in general correspond exactly to the position of the maximum of $p_{\rm G}(\nu)$ and

$$g_{\gamma_{\rm G}}(\nu) \equiv \frac{{\rm d}G_{\gamma_{\rm G}}}{{\rm d}\nu}, \quad (72)$$

it is rather close to it, and increasingly so as L/ℓ becomes larger. This means in practice that the functions $p_{\rm G}(\nu)$ and $g_{\gamma_{\rm G}}(\nu)$ should be well described close to their maximum by our second-order Taylor expansion in an interval corresponding to confidence levels up to ∼90–95 per cent. Hence, the functions $p_{\rm G}(\nu)$ and $g_{\gamma_{\rm G}}(\nu)$ should match each other quite well in that interval, but not in the tails, especially in the large-ν one.

Note as well that $\gamma_{\rm G}$ is negative, as measured experimentally by e.g. Mikelson et al. (2009) and that it converges to zero, so that the function $P_{\rm G}(\nu)$ converges to equation (3) as demonstrated more rigorously by e.g. Bickel & Rosenblatt (1973) (see also Rosenblatt 1976).

This asymptotic result was first exploited in cosmology by Coles (1988) to analyse the hottest hotspots in temperature fluctuations of the CMB. However, it is important to note that convergence to form (3) is rather slow. On the other hand, form (1) with values of a, b and $\gamma_{\rm G}$ given by equations (65), (66), (67), (68) and (69) along with implicit equation (59) to determine $\nu_\star$ remains always a good fit of equations (54) and (55) in the 90–95 per cent confidence region, but again, not in the tails. These analytical results will be illustrated explicitly in Section 4.

4 MEASUREMENTS

To check the validity of the theoretical calculations, we performed numerical experiments in the 2D case. We generated scale-free random Gaussian fields on sets of 100 realizations on a grid of size 4096^2 for each value of the spectral index we considered, $n = 0$, -0.5, -1 and -1.5. Smoothing was performed with a Gaussian window of size $\ell = 5$ pixels and 400 non-overlapping circular patches of radius $L = 100$ pixels were extracted from each realization, amounting to a grand total of 40 000 patches to measure $p_{\rm G}(\nu)$ for each value of n.

The results are displayed in Fig. 1 for $n = 0$, -0.5 and in Fig. 2 for $n = -1.0$ and -1.5. Agreement between the measurements and

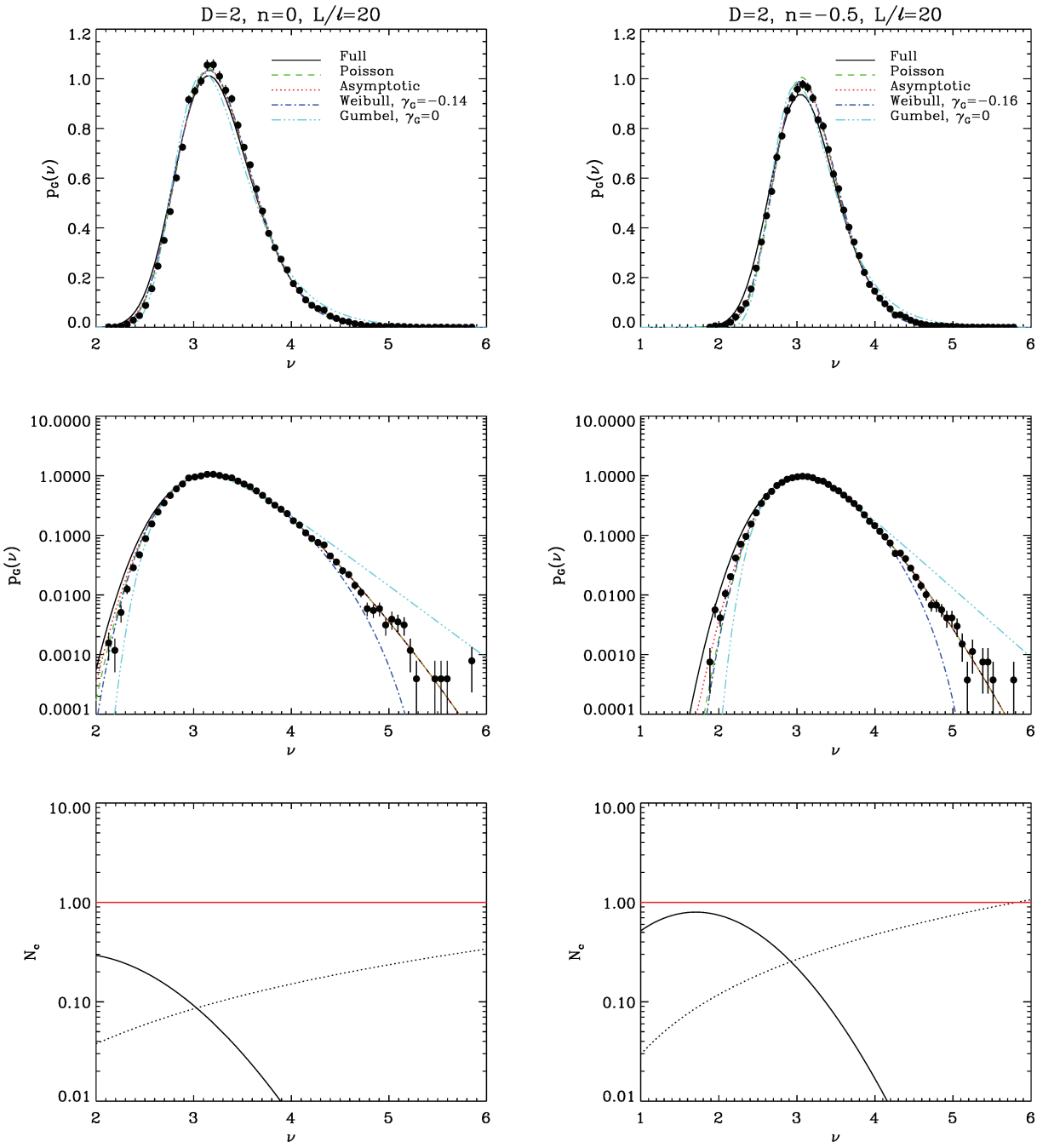

Figure 1. The Gumbel statistics measured in the case $L/\ell = 20$ for 2D random Gaussian fields with power spectra $P(k) \propto k^n$. The spectral index, $n = 0$ and -0.5, is indicated on each panel. The values $n = -1$ and -1.5 are examined in Fig. 2. First and second row of panels: $p_G(\nu)$ and its logarithm as a function of ν. The symbols correspond to the measurements in simulated data as described in the text. Vertical error bars show 1σ errors calculated from 100 independent realizations of the field. As indicated in the top panels, the solid curves correspond to our theoretical prediction (equations 16, 21, 28, 38, 45, 47, 48, 50, 51); the short-dashed ones are the same but assume that peaks are unclustered (Poisson limit, or $N_c = 0$, equivalently $\sigma = 1$ in equation 21, but still use equations 38 and 45 to determine the peak abundance); the dotted ones further assume that the number density of peaks in the excursion is approximated by the Euler characteristic (equation 55); the dot–dashed curves give form (1) fitted on the dotted curves, with the value of γ_G obtained from matching the Taylor expansion discussed in Section 3.4 (equations 59, 65, 67, 69); finally, the last curves (three dots–one dash repeated) correspond to the asymptotic behaviour (3) expected when the ratio L/ℓ approaches infinity: they are the same as the dot–dashed curves but with $\gamma_G = 0$. Third row of panels: N_c (solid curves) and $\bar{\xi}_2^p \equiv \nu^2 \bar{\xi}_2(L)/\sigma_0^2$ (dotted curves) as functions of ν. When $N_c \gtrsim 1$, one expects the effect of the clustering of peaks to become significant. On the other hand, when $\bar{\xi}_2^p \gtrsim 1$, our description of the N-point correlation functions of peaks becomes inaccurate, but this only has a significant impact on the analytical calculation of $p_G(\nu)$ if $N_c \gtrsim 1$. Note that the intersection of the solid and the dotted curves is expected to be in the vicinity of the maximum of $p_G(\nu)$.

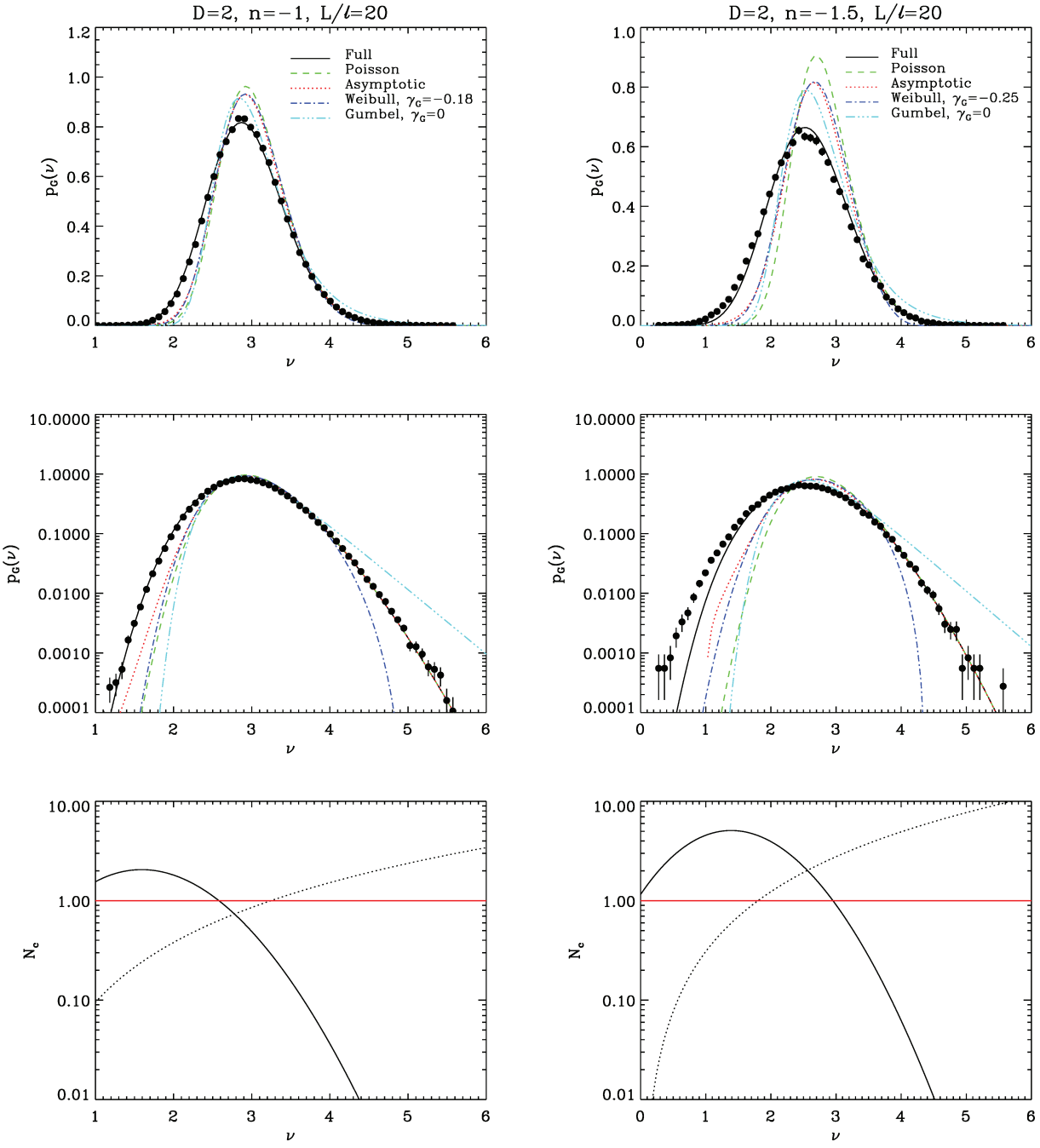

Figure 2. Same as Fig. 1, but for $n = -1$ and -1.5.

theory is spectacular – with the best results in the high-ν region, as expected. Even the case $n = -1.5$ is well described by theoretical predictions despite the fact that condition (25) is broken while $N_c \gtrsim 1$. Except for $n = -1$, the low-ν tail is slightly off the theory, which overestimates a bit the measurements for $n > -1$ and significantly underestimates them for $n = -1.5$. Still, our analytic calculations are sufficiently accurate to define confidence regions with an accuracy level of a few per cent. As expected, the effect of clustering between peaks decreases with increasing n and becomes rather small for $n > -1$, given the choice of $L/\ell = 20$. In that regime, the number density of peaks is in fact well approximated by the Euler characteristics, and the function $p_G(\nu)$ is well fitted by a negative Weibull-type form (equation 1) with the parameters derived in Section 3.4, except in the high-ν tail, as expected. In agreement with the predictions of Section 3.4, $|\gamma_G|$ decreases with n and the asymptotic regime (3) is approached slowly although not reached yet, especially in the tails. It was indeed argued in Section 3.4 that convergence to it is rather slow and requires an increasingly large value of L/ℓ as $-n$ becomes larger. Note that the data points or the solid curves can also be fitted easily by form (1) with appropriate choice of a, b and γ_G (Mikelson et al. 2009), except of course for the high-ν tail. For simplicity we did not show that fit because it is purely phenomenological and there is no simple analytic expression for a, b and γ_G except in the asymptotic regime studied in Section 3.4.

5 SUMMARY AND DISCUSSION

We have computed analytically the Gumbel statistics for random Gaussian fields smoothed with a Gaussian window of size ℓ. The Gumbel statistics, $p_{\rm G}(\nu_{\rm max}){\rm d}\nu_{\rm max}$, represents the probability distribution of the maximum $\nu_{\rm max}$ of the field estimated in a patch of size L thrown at random. Our important results can be summarized as follows.

(i) For L sufficiently large in front of ℓ, $\nu_{\rm max}$ can be approximated by the maximum value of the density estimated at the positions of the peaks included in the patch. As a result, the cumulative Gumbel distribution, $P_{\rm G}(\nu) = \int_{-\infty}^{\nu} p_{\rm G}(\nu_{\rm max}){\rm d}\nu_{\rm max}$, can be seen as the void probability P_0 of finding no peak with density larger than ν in the patch.

(ii) We have made use of the standard count-in-cell formalism (White 1979; Fry 1985; Balian & Schaeffer 1989; Szapudi & Szalay 1993) to compute this void probability as a function of the average number of peaks above the threshold in the patch, $n(\nu)V$, and their correlation functions averaged over the patch, $\bar{\xi}_N^{\rm p}(L)$. These quantities were themselves calculated using results of the literature: BBKS and BE to estimate $n(\nu)$ and $\bar{\xi}_2^{\rm p}$; BS and Politzer & Wise (1984) to evaluate higher order correlations through a hierarchy of normalized cumulants given by $S_N^{\rm p} \equiv \bar{\xi}_N^{\rm p}/(\bar{\xi}_2^{\rm p})^{N-1} = N^{N-2}$. Rigorously speaking, these calculations are only valid in the large-separation limit, $\bar{\xi}_2^{\rm p} \ll 1$ and $\nu \gg 1$. They also neglect contributions from higher order derivatives of the correlation function of the density field, which can in principle be taken into account following Heavens & Sheth (1999) and Desjacques (2008).

(iii) In the regime $\nu \gg 1$ and in the Poisson limit, $N_{\rm c} \equiv nV\bar{\xi}_2^{\rm p} \ll 1$, the quantity $-\ln(P_{\rm G})$ is simply proportional to the Euler characteristic of the excursion (Aldous 1989). This allows one to derive tractable analytical expressions for the Gumbel statistics (equations 54, 55, 56). We have shown that in this case $P_{\rm G}(\nu)$ is well fitted by a negative Weibull-type distribution (1) (with $\gamma_{\rm G} < 0$), except in the high-ν tail. As expected, $\gamma_{\rm G} \to 0$ when $L/\ell \to \infty$ and one converges slowly to the Gumbel-type distribution, equation (3), as shown long ago (Bickel & Rosenblatt 1973).

(iv) Our analytical calculations were successfully tested against numerical experiments of 2D scale-free Gaussian random fields, in particular in a regime where both $N_{\rm c} \gtrsim 1$ and $\bar{\xi}_2^{\rm p} \gtrsim 1$, i.e. where the validity of the 'exact' calculations mentioned in point (ii) remains questionable.

Note that our calculations can be easily extended to non-Gaussian fields, using e.g. the formalism of Pogosyan, Gay & Pichon (2009) to estimate the number density of peaks, $n(\nu)$, and modifications of the Press & Schechter formalism (Press & Schechter 1974) to compute $\bar{\xi}_2^{\rm p}$ in the high-ν regime (see e.g. Desjacques & Seljak 2010; Valageas 2010, and references therein). The hierarchical relation $S_N^{\rm p} \simeq N^{N-2}$ should still hold if $\nu \gg 1$ (BS), as extensively discussed at the end of Section 2.2.2 and in the beginning of Section 2.2.3. In the Poisson limit and for $\nu \gg 1$, the result obtained in point (iii) above should still hold: $-\ln(P_{\rm G})$ should be simply given by the Euler characteristic, which itself can be easily estimated in the non-Gaussian case (Pogosyan et al. 2009; Matsubara 2010). In fact, one expects that $P_{\rm G}$ should still be well fitted by the family of distributions (1) but with a different value of $\gamma_{\rm G}$ (Mikelson et al. 2009). However, convergence to the asymptotic form (3) in the limit $L/\ell \to \infty$ remains to be proven. In the intermediary regime probed by the Euler characteristic, the Gumbel statistics provides an interesting test of non-Gaussianity, as shown experimentally by Mikelson et al. (2009) on the simulated temperature maps of the CMB. In a companion paper (Davis et al. 2011), we have studied applications of the Gumbel statistics to clusters of galaxies, where the quantity of interest is the probability distribution function of the mass of the most massive cluster in the patch (see also Cayón et al. 2010; Holz & Perlmutter 2010). We plan to apply 2D Gumbel statistics to the analysis of CMB data, including non-Gaussian corrections, in the near future.

ACKNOWLEDGMENTS

The authors thank F. Bernardeau for very useful discussions. OD acknowledges the support of an STFC studentship and JD's research is supported by Adrian Beecroft, the Oxford Martin School and STFC.

REFERENCES

Adler R. J., 1981, The Geometry of Random Fields, Wiley Series in Probability and Mathematical Statistics. Wiley, Chichester
Adler R. J., Taylor J. E., 2007, Random Fields and Geometry, Springer Monographs in Mathematics. Springer, New York
Aldous D., 1989, Probability Approximations via the Poisson Clumping Heuristic, Applied Mathematical Sciences, Vol. 77. Springer-Verlag, New York
Ayaita Y., Weber M., Wetterich C., 2010, Phys. Rev. D, 81, 3507
Azaïs J.-M., Delmas C., 2002, Extremes, 5, 181
Balian R., Schaeffer R., 1989, A&A, 220, 1
Bardeen J. M., Bond J. R., Kaiser N., Szalay A. S., 1986, ApJ, 304, 15 (BBKS)
Bernardeau F., Schaeffer R., 1999, A&A, 349, 697 (BS)
Bhavsar S. P., Barrow J. D., 1985, MNRAS, 213, 857
Bickel P., Rosenblatt M., 1973, in Krishnaiah P. R., ed., Multivariate Analysis III. Academic Press, New York, p. 3
Bond J. R., Efstathiou G., 1987, MNRAS, 226, 655 (BE)
Cayón L., Gordon C., Silk J., 2010, preprint (arXiv:1006.1950)
Coles P., 1988, MNRAS, 231, 125
Coles S., 2001, An Introduction to Statistical Modeling of Extreme Values. Springer, New York
Colombi S., Pogosyan D., Souradeep T., 2000, Phys. Rev. Lett., 85, 5515
Cornell C. A., 1968, Bull. Seismological Soc. America, 58, 1583
Cruz M., Martínez-González E., Vielva P., Cayón, L., 2005, MNRAS, 356, 29
Davis O., Devriendt J. E. G., Colombi S., Silk J., 2011, MNRAS, doi:10.1111/j.1365.2966.2011.18286.x
Desjacques V., 2008, Phys. Rev. D, 78, 103503
Desjacques V., Seljak U., 2010, Class. Quantum Grav., 27, 124011
Embrechts P., Schmidli H., 1994, Math. Methods Operations Res., 39, 1
Fry J. N., 1985, ApJ, 289, 10
Gott J. R., III, Juríc M., Schlegel D., Hoyle F., Vogeley M., Tegmark M., Bahcall N., Brinkmann J., 2005, ApJ, 624, 463
Gumbel E. J., 1958, Statistics of Extremes. Columbia Univ. Press, New York (reprinted in 2004 by Dover Press, New York)
Heavens A. F., Sheth R., 1999, MNRAS, 310, 1062
Holz D. E., Perlmutter S., 2010, preprint (arXiv:1004.5349)
Kaiser N., 1984, ApJ, 284, L9
Katz R. W., Brown B. G., 1992, Climatic Change, 21, 289
Katz R. W., Parlange M. B., Naveau P., 2002, Advances Water Resources, 25, 1287
Larson D. L., Wandelt B. D., 2004, ApJ, 613, L85
Leadbetter M. R., Lindgren G., Rootzén H., 1983, Extremes and Related Properties of Random Sequences and Processes. Springer-Verlag, New York
Lokas E. L., Juszkiewicz R., Bouchet F. R., Hivon E., 1996, ApJ, 467, 1

Matsubara T., 2010, Phys. Rev. D, 81, 3505
Mikelsons G., Silk J., Zuntz J., 2009, MNRAS, 400, 898
Piterbarg V. I., 1996, Asymptotic Methods in the Theory of Gaussian Processes and Fields, Translations of Mathematical Monographs, Vol. 148. American Mathematical Society, Providence, RI
Pogosyan D., Gay C., Pichon C., 2009, Phys. Rev. D, 80, 1301
Politzer H. D., Wise M. B., 1984, ApJ, 285, L1
Press W. H., Schechter P., 1974, ApJ, 187, 425
Rosenblatt M., 1976, Ann. Probability, 4, 1009
Szapudi I., Szalay A. S., 1993, ApJ, 408, 43
Valageas P., 2010, A&A, 514, 46
Vielva P., Martínez-González E., Barreiro R. B., Sanz J. L., Cayón L., 2004, ApJ, 609, 22
White S. D. M., 1979, MNRAS, 186, 145
Yamila Yaryura C., Baugh C. M., Angulo R. E., 2011, MNRAS, doi:10.1111/j.1365-2966.2011.18233.x

This paper has been typeset from a TEX/LATEX file prepared by the author.