Journal of the Korean Statistical Society (2007), **36**: *4, pp 479–488*

SIZE DISTRIBUTION OF ONE CONNECTED COMPONENT OF ELLIPTIC RANDOM FIELD

M. T. ALODAT[1]

ABSTRACT

The elliptic random field is an extension to the Gaussian random field. We proved a theorem which characterizes the elliptic random field. We proposed a heuristic approach to derive an approximation to the distribution of the size of one connected component of its excursion set above a high threshold. We used this approximation to approximate the distribution of the largest cluster size. We used simulation to compare the approximation with the exact distribution.

AMS 2000 subject classifications. Primary 60G60; Secondary 62M40.
Keywords. Connected component, elliptic random field, Euler characteristic, excursion set.

1. INTRODUCTION

Let $X(\mathbf{t}),\ \mathbf{t} \in \mathcal{C} \subset \mathbb{R}^D,\ D \geq 2$, be a real-valued random field. Adler (1981) defines the excursion set of the random field $X(\mathbf{t})$ above u in $\mathcal{C}$ as the set

$$A_u = A_u(X, \mathcal{C}) = \{\mathbf{t} \in \mathcal{C} : X(\mathbf{t}) \geq u\}. \tag{1.1}$$

Let $\mathcal{M}(A_u(Y, \mathcal{C}))$ and $\chi(A_u(Y, \mathcal{C}))$ be the number of local maxima of $X(\mathbf{t})$ in $\mathcal{C}$ above u and the Euler characteristic of $A_u(Y, \mathcal{C})$, respectively. Assuming that the excursion set does not touch the boundary of $\mathcal{C}$, Adler (2000), for large u, shows that $E\{\mathcal{M}(A_u(Y, \mathcal{C}))\} \approx E\{\chi(A_u(Y, \mathcal{C}))\}$. This gives the following accurate approximation

$$P\{\sup_{\mathbf{t} \in \mathcal{C}} X(\mathbf{t}) \geq u\} \approx E\{\mathcal{M}(A_u(Y, \mathcal{C}))\} \approx E\{\chi(A_u(Y, \mathcal{C}))\}. \tag{1.2}$$

Received October 2006; accepted June 2007.
[1]Department of Statistics, Yarmouk University, Irbid, Jordan (e-mail: malodat@yu.edu.jo)

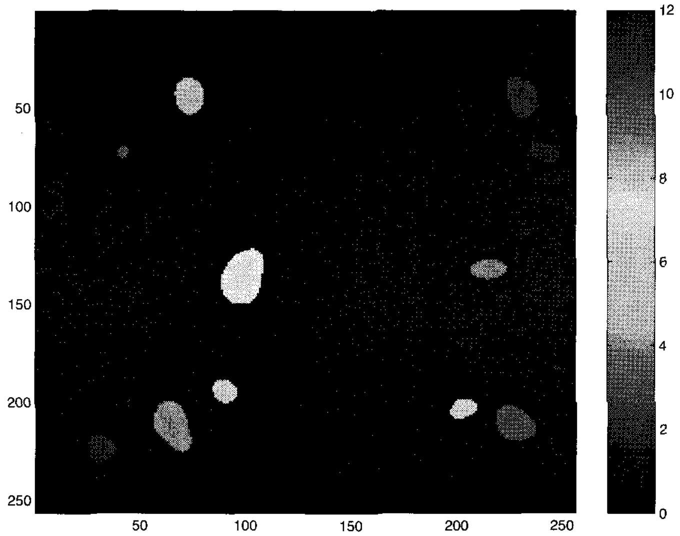

FIGURE 1.1 *Excursion set of a student random field with* $\nu = 5$ *above the threshold* $u = 3.5$.

According to Adler (1981), if $X(\mathbf{t})$ is a homogeneous and smooth Gaussian random field, then with probability tends to one as $u \to \infty$, the excursion set is a finite union of connected components (clusters) such that each connected component contains a local maximum for $X(\mathbf{t})$, see Figure 1.1. The Gaussian random field is used in the literature as a model for images obtained by FMRI (Functional Magnetic Resonance Imaging) technique about the living human brain. The hypothesis to be tested is H_0: no activation in the brain region $\mathcal{C}$. This hypothesis can be tested using

(i) **Value of** $X(\mathbf{t})$: If $\mu(\mathbf{t})$ represents the mean of the image at $\mathbf{t}$, then the above hypothesis is equivalent to $H_0 : \mu(\mathbf{t}) = 0$ for all $\mathbf{t} \in \mathcal{C}$. We reject H_0 if $\sup_{\mathbf{t}\in\mathcal{C}} X(\mathbf{t})$ is large. The p-value for this test statistic can be calculated according to (1.2).

(ii) **Components of** $A_u(X, \mathcal{C})$: This method tests the hypothesis H_0: mean of the maximum cluster size $= \mu_0$. The null hypothesis is rejected if maximum cluster size is large. To find the p-value of this test we need the distribution of the largest cluster size.

Nosko (1969) derived an approximation to the distribution of the size of one connected component of the excursion set of a homogeneous smooth zero mean and unit variance Gaussian random field above a high threshold. Friston *et al.*

(1993) used Nosko's approximation to find the p-value for the second method. Cao (1999) extends Nosko's results to χ^2, t and F random fields. In this paper, we proved a theorem which characterizes the elliptic random field. We derived the expected number of the Euler characteristic of its excursion set. Also we proposed a heuristic approach to derive an approximation to the distribution of the size of one connected component of its excursion set. We used simulation to compare the approximation with the exact distribution.

2. ELLIPTIC RANDOM FIELD

From here to the end of this paper, we will assume that all random variables and vectors used in this paper have densities. A random vector W is said to be m-dimensional multivariate elliptic with parameters μ and Σ if its *pdf* is of the form

$$f_W(w) = \kappa \times h\left((w-\mu)^T\Sigma^{-1}(w-\mu)\right), \quad w \in \mathbb{R}^m,$$

for some function h, κ a normalizing constant, and Σ is positive definite. The parameters μ, Σ are called location and scale parameters, respectively. We will use the notation $W \sim E_m(\mu, \Sigma)$ to denote a m-dimensional multivariate elliptic distribution with parameters μ and Σ. If $W \sim E_D(\mu, \Sigma)$ and

$$W^T = (W_1,\ W_2),\ \ \mu^T = (\mu_1,\ \mu_2)\ \text{ and }\ \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

then $E\{W_1|W_2\} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(W_2 - \mu_2)$.

DEFINITION 2.1. A random field $Y(\mathbf{t})$ is said to be a elliptic random field if every finite-dimensional distribution is multivariate elliptic.

The homogeneous elliptic random field can be characterized by the following theorem.

THEOREM 2.1. *Let $Y(\boldsymbol{t})$, $\boldsymbol{t} \in \mathcal{C} \subset \mathbb{R}^D$ be a homogeneous elliptic random field with zero mean and covariance function $\boldsymbol{R}_Y(\boldsymbol{t})$. Then $Y(\boldsymbol{t})$ admits the following stochastic representation*

$$Y(\boldsymbol{t}) = RX(\boldsymbol{t}),$$

where $X(\boldsymbol{t})$ is a homogeneous Gaussian random field with zero mean, unit variance, covariance function $\boldsymbol{R}_X(\boldsymbol{t}) = E\{R^2\}\boldsymbol{R}_Y(\boldsymbol{t})$ and R is a positive random variable with finite variance independent of $X(\boldsymbol{t})$.

PROOF. Let $\mathbf{t} = \mathbf{t}_1, \ldots, \mathbf{t}_n \in \mathcal{C}$. Then, by definition, the random vector, $W = (Y(\mathbf{t}_1), \ldots, Y(\mathbf{t}_n))$ has a multivariate elliptic distribution. From the multivariate theory, a random vector $W \sim E_n(\mu, \Sigma)$ if and only if $W \stackrel{d}{=} \mu + RH$, where R is a positive scalar random variable independent of $H \sim N_n(\mathbf{0}, \Sigma)$ (Muirhead, 1982). So the vector W admits the representation $W \stackrel{d}{=} R(X(\mathbf{t}_1), \ldots, X(\mathbf{t}_n))$, where $(X(\mathbf{t}_1), \ldots, X(\mathbf{t}_n)) \sim N_n(\mathbf{0}, \Sigma^*)$, and Σ^* is a $n \times n$ with elements $\mathrm{Cov}(Y(\mathbf{t}_i), Y(\mathbf{t}_j))$, $i, j = 1, \ldots, n$. This implies that $Y(\mathbf{t}) \stackrel{d}{=} RX(\mathbf{t})$ and the covariance function of $Y(\mathbf{t})$ is $E\{R^2\}\mathbf{R}_X(\mathbf{t})$. This establishes the theorem. □

The class of the elliptic random fields can be considered as an extension to the class of the Gaussian random fields. For example, if $R \stackrel{d}{=} \sqrt{\nu/S}$ where S is a chi-square random variable with ν degrees of freedom, then $Y(\mathbf{t})$ is a student random field, which has more variability than the Gaussian random field. As $\nu \to \infty$, the student random field will be a Gaussian random field. This makes the elliptic class to be more suitable for modelling than the Gaussian class. One more advantage of the elliptic field is that it is easy to simulate and its covariance function is proportional to the Gaussian one. Since the random variable R does not depend on $\mathbf{t}$,

$$P\left\{\sup_{\mathbf{t}\in\mathcal{C}} Y(\mathbf{t}) \geq u\right\} = \int_0^\infty P\left\{\sup_{\mathbf{t}\in\mathcal{C}} X(\mathbf{t}) \geq \frac{u}{r}\right\} f_R(r)dr. \tag{2.1}$$

In the literature, several good approximations are available for $P\{\sup_{\mathbf{t}\in\mathcal{C}} X(t) \geq u/r\}$, but we can not plug them in (2.1) since they are valid for large levels. So we will go to approximate the left hand side of (2.1) based on the Euler characteristic and the number of the local maxima of $Y(\mathbf{t})$.

3. ASSUMPTIONS AND REGULARITY CONDITIONS

We will assume that the random field $Y(\mathbf{t})$ has to satisfy regularity conditions given in Adler (1981). Therefore we will assume that $X(\mathbf{t})$, $\mathbf{t} \in S \subset \mathbb{R}^D$, has zero-mean, unit variance, homogeneous, ergodic and twice differentiable in the mean-square sense Gaussian random field. If $\dot{X}(\mathbf{t})$ and $\ddot{\mathbf{X}}(\mathbf{t})$ be the gradient and the matrix of the second partial derivatives of $X(\mathbf{t})$ respectively, then $\dot{Y}(\mathbf{t}) = R\dot{X}(\mathbf{t})$ and $\ddot{\mathbf{Y}}(\mathbf{t}) = R\ddot{\mathbf{X}}(\mathbf{t})$. Let $\dot{X}_j(\mathbf{t}) = \partial X(\mathbf{t})/\partial t_j$ and $\ddot{\mathbf{X}}_{ij} = \partial^2 X(\mathbf{t})/\partial t_i \partial t_j$ satisfy the following assumptions:

$$\max_{i,j} E\{|\ddot{\mathbf{X}}_{ij}(\mathbf{t}) - \ddot{\mathbf{X}}_{ij}(\mathbf{0})|\} \leq C\|\mathbf{t}\|^2,$$

where $C > 0$ and all $\mathbf{t}$ in some neighborhood of $\mathbf{0}$. It easy to show that if the field $X(\mathbf{t})$ satisfies the regularity conditions, then so the field $Y(\mathbf{t})$ does. For a Gaussian random field $X(\mathbf{t})$, the conditional distribution of $\ddot{\mathbf{X}} = \ddot{\mathbf{X}}(\mathbf{0})$ given $X = X(\mathbf{0})$ is $N_{D\times D}(-X\Lambda,\ M(\Lambda))$, where $\Lambda = \mathrm{Var}(\dot{X})$, $\dot{X} = \dot{X}(\mathbf{0})$ and $M(\Lambda)$ is a matrix depending on Λ.

4. Size of One Connected Component

Using Taylor approximation about $\mathbf{0}$, we can write the following approximation for $Y(\mathbf{t})$

$$\tilde{Y}(\mathbf{t}) = Y(\mathbf{0}) + \mathbf{t}^T\dot{Y}(\mathbf{0})\mathbf{t} + \frac{1}{2}\mathbf{t}^T\ddot{\mathbf{Y}}(\mathbf{0})\mathbf{t}.$$

Given that $Y \geq u$ and $\ddot{\mathbf{Y}}$ is negative definite, we can solve the equation $u = \tilde{Y}(\mathbf{t})$ for $\mathbf{t}$. So the size of the connected components that contains $\mathbf{0}$ will be approximated by the D-dimensional volume of the ellipsoid surrounded by $u = \tilde{Y}(\mathbf{t})$. Therefore, the volume is

$$\tilde{S} = \omega_D \frac{\left(2(Y-u) - \dot{Y}^T\ddot{\mathbf{Y}}^{-1}\dot{Y}\right)^{\frac{D}{2}}}{\det(-\ddot{\mathbf{Y}})^{\frac{1}{2}}}, \tag{4.1}$$

where ω_D is the volume of the D-dimensional unit ball. For large values of u, the second term in the numerator of equation (4.1) can be neglected. Since the curvature matrix $\ddot{\mathbf{Y}}$ is random, we will replace it by its conditional mean given that $Y = u$, *i.e.*, $\mathrm{Cov}(Y, \ddot{\mathbf{Y}}) = E\{R^2\}\Lambda$ and $\mathrm{Var}\{Y\} = E\{R^2\}$. The conditional mean of $\ddot{\mathbf{Y}}$ given $Y = u$ is given by

$$\begin{aligned} E\{\ddot{\mathbf{Y}}|Y=u\} &= E\{\ddot{\mathbf{Y}}\} + \mathrm{Cov}(Y,\ddot{\mathbf{Y}})\mathrm{Cov}(Y)^{-1}(u - E\{Y\}) \\ &= E\{\ddot{\mathbf{Y}}\} - uE\{R^2\}\Lambda E\{R^2\}^{-1}(u - E\{Y\}) \\ &= -u\Lambda. \end{aligned}$$

This reduces equation (4.1) to

$$\tilde{S} \approx \omega_D 2^{\frac{D}{2}} \det(\Lambda)^{-\frac{1}{2}} \left(\frac{Y-u}{u}\right)^{\frac{D}{2}}. \tag{4.2}$$

To find the distribution of $\tilde{S}$, we need to find the distribution of $Y - u$ given that $Y(\mathbf{t})$ has a local maximum of a height exceeding u at $\mathbf{0}$. Let $F_Y(y)$ denotes this distribution. Using the same argument as that of Adler (1981, p. 158) and noting that $E\{\mathcal{M}(A_u(Y,\mathcal{C}))\} \approx E\{\chi(A_u(Y,\mathcal{C}))\}$ we can write

$$1 - F_Y(y) \approx \frac{E\{\chi(A_{y+u}(Y,\mathcal{C}))\}}{E\{\chi(A_u(Y,\mathcal{C}))\}}. \tag{4.3}$$

For a Gaussian $X(\mathbf{t})$ with $\Lambda = \mathrm{Var}(\dot{X}(\mathbf{0}))$, Adler (1981) gives the following expression for $E\{\chi(A_u(X,\mathcal{C}))\}$ as

$$E\{\chi(A_u(X,\mathcal{C}))\} = \lambda(\mathcal{C})\frac{\exp(-\frac{u^2}{2})\det(\Lambda)^{\frac{1}{2}}}{(2\pi)^{(D+1)/2}}H_{D-1}(u), \tag{4.4}$$

where

$$H_n(x) = n!\sum_{j=0}^{[n/2]}\frac{(-1)^j x^{n-2j}}{j!(n-2j)!2^j}. \tag{4.5}$$

Using equation (4.4) and Theorem 5.2.1 of Adler (1981), we can relate the Euler characteristic of the excursion set of $X(\mathbf{t})$ and Euler characteristic of the excursion set of $Y(\mathbf{t})$ as follows:

$$\begin{aligned} E\{\chi(A_u(Y,\mathcal{C}))\} &= \int_0^\infty E\{\chi(A_{\frac{u}{r}}(X,\mathcal{C}))\}f_R(r)dr, \\ &= \frac{\lambda(\mathcal{C})\det(\Lambda)^{\frac{1}{2}}}{(2\pi)^{\frac{D+1}{2}}}\int_0^\infty \exp\left(-\frac{u^2}{2r^2}\right)H_{D-1}\left(\frac{u}{r}\right)f_R(r)dr. \end{aligned}$$

For large u, this leads to

$$1 - F_Y(y) \approx \frac{\int_0^\infty \exp\left(-\frac{(u+y)^2}{2r^2}\right)H_{D-1}(\frac{u+y}{r})f_R(r)dr}{\int_0^\infty \exp\left(-\frac{u^2}{2r^2}\right)H_{D-1}(\frac{u}{r})f_R(r)dr}. \tag{4.6}$$

For $D = 2$ and $R = \sqrt{\nu/S}$, where S is a chi-square random variable with ν degrees of freedom, the field $Y(\mathbf{t})$ is called the student random field. In this case, $H_1(x) = x$ and

$$f_R(r) = \frac{2\nu^{\frac{\nu}{2}}r^{-\nu-1}\exp\left(-\frac{\nu}{2r^2}\right)}{\Gamma(\frac{\nu}{2})2^{\frac{\nu}{2}}} \quad \text{for} \quad r > 0.$$

As a function in u, the integral

$$\int_0^\infty \exp\left(-\frac{u^2}{2r^2}\right)H_1\left(\frac{u}{r}\right)f_r(r)dr$$

is proportional to

$$u\int_0^\infty r^{-\nu-2}\exp\left(-\frac{u^2+\nu}{2r^2}\right)H_1(\frac{u}{r})dr = \frac{u}{(u^2+\nu)^{\frac{\nu+1}{2}}}.$$

So the following holds

$$\begin{aligned} 1-F_Y(y) &\approx \frac{u+y}{u}\left(\frac{u^2+\nu}{(u+y)^2+\nu}\right)^{\frac{\nu+1}{2}} && (4.7)\\ &\approx \left(\frac{u^2+\nu}{(u+y)^2+\nu}\right)^{\frac{\nu+1}{2}} && (4.8)\end{aligned}$$

for large values of u and y is small with respect to u. From the last equation we note that $1-F_Y(y)$ is equivalent to $\exp(-uy)$ for large ν, which is the same result of Nosko (1969) reported in Alder (1981). It can be noted that $1-F_Y(y)$ does not depend on D.

5. EXCURSION SET AND POISSON CLUMPING

If λ_D denotes the Lebesque measure on $\mathbb{R}^D$, then for any fixed threshold u and a homogeneous random field $Y(\mathbf{t})$, $\mathbf{t}\in\mathcal{C}$,

$$\lambda_D\{A_u(Y,\mathcal{C})\} = \int_{\mathcal{C}} \mathbf{1}_{A_u}(Y(\mathbf{t})-u)d\mathbf{t}.$$

Taking the expectation on both sides

$$\begin{aligned} E\{\lambda_D(A_u(Y,\mathcal{C}))\} &= \int_{\mathcal{C}} E\{\mathbf{1}_{[u,\ \infty)}(Y(\mathbf{t}))\}d\mathbf{t}\\ &= \int_{\mathcal{C}} P\{Y(\mathbf{0})\geq u\}d\mathbf{t}\\ &= \lambda_D\{\mathcal{C}\}P\{U(\mathbf{t})\geq u\}. && (5.1)\end{aligned}$$

Aldous (1989) introduced the Poisson clumping heuristic (PCH), which means throwing random sets (clumps) at random according to a Poisson point process, *i.e.*, the centers of the sets are generated by a Poisson random variable. Cao (1999) used the PCH to model the excursion set $A_u(Y,\mathcal{C})$, where each cluster is considered as a clump and the local maximum is considered to be the center of the cluster. Let N be the number of connected components of $A_u(Y,\mathcal{C})$ and $S_1,\ldots,S_N$ be the sizes of these clusters. So

$$\begin{aligned} E\{\lambda_D(A_u(Y,\mathcal{C}))\} &= E[NE\{S_1\}]\\ &= E\{N\}E\{S_1\}. && (5.2)\end{aligned}$$

By (5.1) and (5.2),

$$E\{\lambda_D(A_u(Y,\mathcal{C}))\} = \lambda_D(\mathcal{C})P\{Y(\mathbf{0}) > u\}.$$

If we approximate $E\{N\}$ by $E\{\chi(A_u(Y,\mathcal{C}))\}$, then

$$E\{S_1\} \approx \frac{\lambda_D(\mathcal{C})P\{Y(\mathbf{0}) > u\}}{E\{\chi(A_u(Y,\mathcal{C}))\}}. \tag{5.3}$$

6. Distribution of Maximum Cluster Size

Using the idea of PCH, Cao (1999) writes the distribution of $\tilde{S}_{\max} = \max_i \tilde{S}_i$ in terms of $E\{N\}$ and $P\{\tilde{S}_1 \geq s\}$ as follows:

$$P\{\tilde{S}_{\max} \leq s | N \geq 1\} = \frac{\exp(-E\{N\}P\{\tilde{S}_1 \geq s\}) - \exp(-E\{N\})}{1 - \exp(-E\{N\})},$$

where $P\{\tilde{S}_1 \geq s\}$ can be obtained using (4.2) and (4.7) while $E\{N\}$ will be approximated by $E\{\chi(A_u(Y,\mathcal{C}))\}$.

7. Simulation

We simulated two large samples each of size m =5,000, one from the approximate distribution of the cluster size and the other one from the exact distribution. The exact one is simulated using the equation $Y(\mathbf{t}) = RX(\mathbf{t})$ in the region $\mathcal{C} = [1,128] \times [1,128]$, where $R = \sqrt{\nu/S}$, S is a chi-square random variable with ν degrees of freedom independent of $X(\mathbf{t})$. Then we threshold the field $Y(\mathbf{t})$ by a fixed level u. We used the MatLab function BWLABEL to find the sizes of the clusters in the excursion set $\{\mathbf{t} \in [1,128] \times [1,128] : Y(\mathbf{t}) \geq u\}$. In our simulation, we considered $\nu = 10$ and the values $u = 2.5$, 3.0 and 3.5. The cumulative distribution function (*cdf*) for each sample is plotted in Figure 7.1. We note that the two *cdf*'s are very closed.

In Figure 7.2, we compare the cluster size distributions for the Gaussian random field and for the elliptic random field for $u = 2.5$, $\nu = 5$, 10 and 25. We see that there is a difference between the two distributions for small values of ν, while this difference disappears for large values of ν. This is expected result since the elliptic random field ($T(\mathbf{t}) = X(\mathbf{t})\sqrt{\nu/S}$) is equivalent to the Gaussian one as $\nu \to \infty$. So we have to be careful when applying the Gaussian approximation to the cluster size for real data.

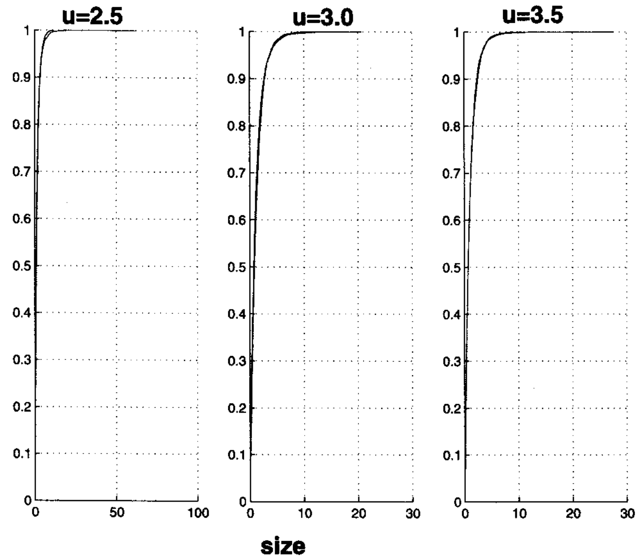

FIGURE 7.1 *The exact distribution function (dotted) of the cluster size and the approximate distribution function (smooth) of the cluster size for* $u = 2.5,\ 3.0$ *and* $u = 3.5$.

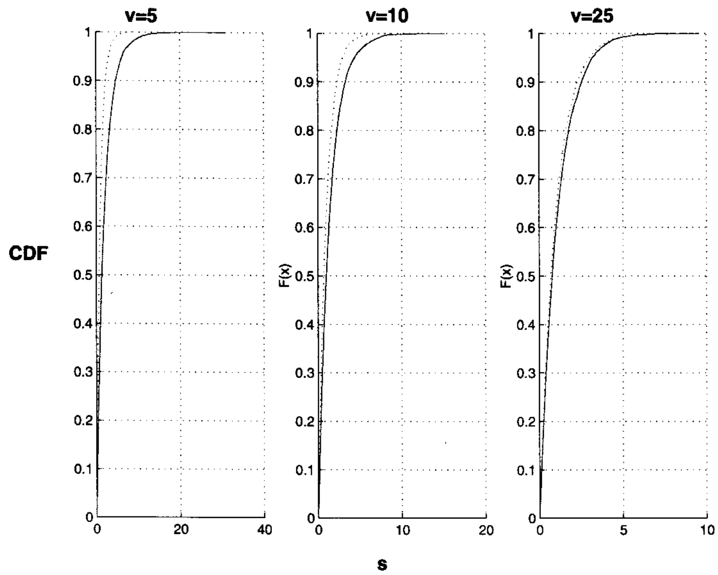

FIGURE 7.2 *The distribution functions for two large samples obtained from the distributions of the cluster size. The Gaussian random field (dotted) and the elliptic random field (smooth) for* $u = 2.5$ *and* $\nu = 5,\ 10$ *and* 25.

8. Conclusion

In this paper, we derived an approximation to the distribution of the size of one connected component of the excursion set of the elliptic random field. For the case when $D = 2$ and $R = \sqrt{\nu/S}$, where S is a chi-square random variable with ν degrees of freedom, our result match the Nosko's one when $\nu \to \infty$, *i.e.*, when the field is Gaussian. For small values of ν the simulation shows that the approximation works very well. We also note that for this case $E\{S_1\}$, obtained based on PCH, and $E\{\tilde{S}\}$ obtained based on our approximation are the same. This means that the approximate distribution derived here need not to be corrected to the mean as in Friston *et al.* (1993).

Acknowledgements

The author thanks the referee for his comments which helped me to improve the paper.

References

Adler, R. J. (1981). *The Geometry of Random Fields.* John Wiley & Sons, New York.

Adler, R. J. (2000). "On excursion sets, tube formulae, and maxima of random fields", *The Annals of Applied Probability,* **10**, 1–74.

Aldous, D. (1989). *Probability Approximations via the Poisson Clumping Heuristic.* Springer-Verlag, New York.

Cao, J. (1999). "The size of the connected components of excursion sets of χ^2, t and F fields". *Advances in Applied Probability,* **31**, 579–595.

Friston, K. J., Worsley, K. J., Frackowiak, R. S. J., Mazziotta, J. C. and Evans, A. C. (1993). "Assessing the significance of focal activations using their spatial extent", *Human Brain Mapping,* **1**, 210–220.

Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory,* John Wiley & Sons, New York.

Nosko, V. P. (1969). "Local structure of Gaussian random fields in the neighborhood of high-level shines", *Doklady Akademii Nauk SSSR,* **189**, 714–717.