Fifth Colloquium on Mathematics and Computer Science DMTCS proc. **AI**, 2008, 59–70

Point process stabilization methods and dimension estimation

J. E. Yukich[†]

Department of Mathematics, Lehigh University, 14 East Packer Ave, Christmas-Saucon, Bethlehem, USA

We provide an overview of stabilization methods for point processes and apply these methods to deduce a central limit theorem for statistical estimators of dimension.

Keywords: Dimension estimation, laws of large numbers, central limit theorems, stabilization

1 Introduction

This section provides background on stabilization methods for functionals of point processes, anticipating the subsequent application to statistical estimation of dimension in the next section.

Sums of spatially dependent terms. Fundamental questions pertaining to large, complex geometric structures often involve sums of spatially dependent terms having short range interactions, but complicated long range dependence. This phenomenon arises in problems across a wide range of fields, including random geometric graphs and networks, discrete stochastic geometry, statistical mechanics, statistics of random samples, and percolation models.

Functionals of large complex geometric structures are often represented as sums of spatially dependent terms

$$\sum_{x \in \mathcal{X}} \xi(x, \mathcal{X}), \tag{1.1}$$

where $\mathcal{X} \subset \mathbb{R}^d$ is locally finite and where the measurable function ξ, defined on all pairs $(x, \mathcal{X})$, with $x \in \mathcal{X}$, represents the interaction of x with respect to $\mathcal{X}$.

When $\mathcal{X}$ is a random n point set, laws of large numbers for (1.1) can sometimes be obtained via subadditive or ergodic theoretic methods, whereas central limit theorems may be deduced via M-dependent or mixing methods. There are examples of interest where these classical methods are either not directly applicable, or if they are applicable, they may fail to produce rates of convergence or explicit asymptotics in terms of the underlying geometry and point densities. Stabilization methods for functionals of point processes provide another approach for handling sums of spatially dependent terms. This approach has proved useful in both refining existing asymptotic results and establishing new results in situations where

[†]Research supported in part by NSF grant DMS-0805570

the classical methods are not directly applicable. Stabilization methods, described below, have been used in problem areas as disparate as random packing [3, 17], convex hulls [23], ballistic deposition models [3, 15, 17], quantization [24], loss networks [24], continuum and lattice percolation [12], and geometric graphs in Euclidean combinatorial optimization [16, 18, 26]. Rather than review these applications, our goal here is twofold: (i) review the essential theory underpinning stabilization and (ii) employ the techniques to describe the limit theory for statistical estimators of dimension.

Stabilization of functionals of point processes. When $\mathcal{X}$ is random the range of spatial dependence of ξ at a given $x \in \mathcal{X}$ is random and the purpose of stabilization is to quantify this range in a way useful for asymptotic analysis. There are several similar notions of stabilization, but the essence is captured by the notion of stabilization of ξ with respect to homogeneous Poisson points on $\mathbb{R}^d$, defined as follows. Given $\mathcal{X} \subset \mathbb{R}^d$, $a > 0$ and $y \in \mathbb{R}^d$, we let $a\mathcal{X} := \{ax : x \in \mathcal{X}\}$ and $y + \mathcal{X} := \{y + x : x \in \mathcal{X}\}$. For all $\lambda > 0$ we define the re-scaled version of ξ by

$$\xi_\lambda(x, \mathcal{X}) := \xi(\lambda^{1/d}x, \lambda^{1/d}\mathcal{X}). \tag{1.2}$$

When $x \in \mathbb{R}^d \setminus \mathcal{X}$, we abbreviate notation and write $\xi(x, \mathcal{X})$ instead of $\xi(x, \mathcal{X} \cup \{x\})$.

It will be useful to work with point processes more general than homogeneous Poisson point processes. Let κ be a probability density function on $\mathbb{R}^d$ with support $A \subseteq \mathbb{R}^d$. For all $\lambda > 0$, let $\mathcal{P}_\lambda$ denote a Poisson point process in $\mathbb{R}^d$ with intensity measure $\lambda\kappa(x)dx$. We shall assume throughout that κ is bounded with supremum denoted $\|\kappa\|_\infty$.

Say that ξ is translation invariant if $\xi(x, \mathcal{X}) = \xi(x + z, \mathcal{X} + z)$ for all $z \in \mathbb{R}^d$. Let $B_r(x)$ denote the Euclidean ball centered at x with radius r and let $\mathbf{0}$ denote a point at the origin of $\mathbb{R}^d$. Letting $\mathcal{H}_\tau$ be a rate τ homogenous Poisson point process on $\mathbb{R}^d$, we say that a translation invariant ξ is *homogeneously stabilizing* if for all $\tau > 0$ there exists an almost surely finite random variable $R := R(\tau)$ such that

$$\xi(\mathbf{0}, (\mathcal{H}_\tau \cap B_R(\mathbf{0})) \cup \mathcal{A}) = \xi(\mathbf{0}, \mathcal{P}_\tau \cap B_R(\mathbf{0})) \tag{1.3}$$

for all locally finite $\mathcal{A} \subset \mathbb{R}^d \setminus B_R(\mathbf{0})$. Thus ξ stabilizes if the value of ξ at $\mathbf{0}$ is unaffected by changes in the configuration outside $B_R(\mathbf{0})$.

An elementary example of a homogeneously stabilizing functional goes as follows. Let $\xi(x, \mathcal{X})$ denote the distance between x and its nearest neighbor in $\mathcal{X}$. Clearly the value of $\xi(\mathbf{0}, \mathcal{H}_\tau)$ depends only $\mathcal{H}_\tau \cap B_D(\mathbf{0})$, where $D := D(\mathcal{H}_\tau)$ is the distance between $\mathbf{0}$ and its nearest neighbor in $\mathcal{H}_\tau$. In other words D is a radius of stabilization and clearly it is almost surely finite.

An important feature of stabilization is that it yields weak convergence results for the scaled functional ξ_λ, as illustrated by the following lemma, proved in [14]. Recall that almost every $x \in \mathbb{R}^d$ is a *Lebesgue point* of κ, that is to say for almost all $x \in \mathbb{R}^d$ we have that $\epsilon^{-d} \int_{B_\epsilon(x)} |\kappa(y) - \kappa(x)|dy$ tends to zero as ϵ tends to zero.

Lemma 1.1 *Let x_0 be a Lebesgue point for κ. If ξ is homogeneously stabilizing then we have*

$$\xi_\lambda(x_0, \mathcal{P}_\lambda) \xrightarrow{\mathcal{D}} \xi(\mathbf{0}, \mathcal{H}_{\kappa(x_0)}). \tag{1.4}$$

To see (1.4), we have by translation invariance $\xi_\lambda(x_0, \mathcal{P}_\lambda) = \xi(\mathbf{0}, \lambda^{1/d}(\mathcal{P}_\lambda - x_0))$. By the stabilization of ξ, we have that $(\mathbf{0}, \mathcal{H}_{\kappa(x_0)})$ is a continuity point for ξ ([14]). The result follows by the weak convergence $\lambda^{1/d}(\mathcal{P}_\lambda - x_0) \xrightarrow{\mathcal{D}} \mathcal{H}_{\kappa(x_0)}$ and the continuous mapping theorem.

To establish limit theorems for stabilizing functionals we will need one further notion of stabilization, one quantifying the local spatial dependence of ξ_λ uniformly over both $x \in A$ and $\mathcal{P}_\lambda,\ \lambda \geq 1$. This notion of stabilization, originating in [3], implies the exponential decay of the spatial dependencies of ξ_λ. It turns out that many functionals in geometric probability which are homogeneously stabilizing are also exponentially stabilizing with respect to κ and A in the following sense.

Definition 1.1 *ξ is* exponentially stabilizing *with respect to κ and A if for all $\lambda \geq 1$ and all $x \in A$, there exists an almost surely finite random variable $R := R(x, \lambda)$ (*a radius of stabilization *for ξ_λ at x) such that for all finite $\mathcal{A} \subset (\mathbb{R}^d \setminus B_{\lambda^{-1/d}R}(x))$, we have*

$$\xi_\lambda\left(x, [\mathcal{P}_\lambda \cap (B_{\lambda^{-1/d}R}(x))] \cup \mathcal{A}\right) = \xi_\lambda\left(x, \mathcal{P}_\lambda \cap (B_{\lambda^{-1/d}R}(x))\right), \tag{1.5}$$

and moreover the tail probability $\tau(t)$ defined for $t > 0$ by $\tau(t) := \sup_{\lambda \geq 1,\ x \in A} P[R(x, \lambda) > t]$ satisfies $\limsup_{t\to\infty} t^{-1} \log \tau(t) < 0$.

Roughly speaking, $R := R(x, \lambda)$ is a radius of stabilization if the value of $\xi_\lambda(x, \mathcal{P}_\lambda)$ is unaffected by changes to the points outside $B_{\lambda^{-1/d}R}(x)$.

Limit theorems for sums $\sum_{x\in\mathcal{P}_\lambda\cap A} \xi_\lambda(x, \mathcal{P}_\lambda)$ naturally require moment conditions on the summands, thus motivating the next definition.

Definition 1.2 *ξ has a moment of order $p > 0$ (with respect to κ and A) if*

$$\sup_{\lambda \geq 1,\ x \in A} \mathbb{E}\left[|\xi_\lambda(x, \mathcal{P}_\lambda)|^p\right] < \infty. \tag{1.6}$$

Limit theory for sums of stabilizing terms. Let X_i be i.i.d. with density κ, and put $\mathcal{X}_n := \{X_1, ..., X_n\}$. Consider the point measures

$$\mu_\lambda := \mu_\lambda^\xi := \sum_{x\in\mathcal{P}_\lambda} \xi_\lambda(x, \mathcal{P}_\lambda)\delta_x \quad \text{and} \quad \rho_n := \rho_n^\xi := \sum_{i=1}^n \xi_n(X_i, \mathcal{X}_n)\delta_{X_i}, \tag{1.7}$$

where δ_x denotes the unit point mass at x. We consider in (1.7) sums involving ξ_λ and ξ_n, rather than just ξ, since the former dilate the point sets in question, so that there are roughly a constant number of points per unit volume. Let $B(A)$ denote the class of all bounded $f : A \to \mathbb{R}$ and for all measures μ on $\mathbb{R}^d$ let $\langle f, \mu\rangle := \int f d\mu$. Put $\bar{\mu} := \mu - \mathbb{E}[\mu]$.

It is sometimes the case that thermodynamic limits for sums of spatially dependent terms can be established by either subadditive methods or ergodic theoretic methods [25, 27]. The first method has the drawback that it does not easily yield explicit limiting constants, this even in the case when κ is uniform. The latter method yields such constants, but only when the sums represent the restriction of a globally defined process to expanding windows, in contrast to the present set-up which considers processes on a fixed window and, after multiplication by $\lambda^{1/d}$, then yields limits for processes on expanding volume λ windows.

For all $f \in B(A)$ we have that

$$\mathbb{E}\left[\langle f, \mu_\lambda\rangle\right] = \lambda \int_A f(x)\mathbb{E}\left[\xi_\lambda(x, \mathcal{P}_\lambda)\right]\kappa(x)dx. \tag{1.8}$$

Assuming that (1.6) holds for some $p > 1$, Lemma 1.1 gives for all Lebesgue points x of κ the convergence $\mathbb{E}\,[\xi_\lambda(x, \mathcal{P}_\lambda)] \to \mathbb{E}\,[\xi(\mathbf{0}, \mathcal{H}_{\kappa(x)})]$. The set of points failing to be Lebesgue points has measure zero and by the bounded convergence theorem it follows that

$$\lim_{\lambda \to \infty} \lambda^{-1} \mathbb{E}\,[\langle f, \mu_\lambda \rangle] = \int_A f(x) \mathbb{E}\,[\xi(\mathbf{0}, \mathcal{H}_{\kappa(x)})] \kappa(x) dx.$$

This simple convergence of means of $\langle f, \mu_\lambda \rangle$ can be improved as follows to one providing convergence in L^q, where $q = 1$ or $q = 2$.

Theorem 1.1 *(WLLN [14, 18]) Put $q = 1$ or $q = 2$. Let ξ be a homogeneously stabilizing (1.3) translation invariant functional satisfying the moment condition (1.6) for some $p > q$. Then for all $f \in B(A)$ we have*

$$\lim_{n \to \infty} n^{-1} \langle f, \rho_n \rangle = \lim_{\lambda \to \infty} \lambda^{-1} \langle f, \mu_\lambda \rangle = \int_A f(x) \mathbb{E}\,[\xi(\mathbf{0}, \mathcal{H}_{\kappa(x)})] \kappa(x) dx \ \text{ in } \ L^q. \tag{1.9}$$

Suppose that $f \equiv 1$. If $\mathbb{E}\,[\xi(\mathbf{0}, \mathcal{H}_\tau)] = \mathbb{E}\,[\xi(\mathbf{0}, \mathcal{H}_1)]$, which would be the case for functionals which satisfy the scaling relation $\xi(\alpha x, \alpha \mathcal{X}) = \xi(x, \mathcal{X})$ for any scalar $\alpha > 0$ (see section 2 for an example), then since κ is a probability density the limit (1.9) simplifies to

$$\lim_{\lambda \to \infty} \lambda^{-1} \sum_{x \in \mathcal{P}_\lambda} \xi_\lambda(x, \mathcal{P}_\lambda) = \mathbb{E}\,[\xi(\mathbf{0}, \mathcal{H}_1)] \ \text{ in } \ L^q. \tag{1.10}$$

On the other hand, if there is an $\alpha > 0$ such that ξ satisfies for all $\tau > 0$ the scaling relation $\mathbb{E}\,[\xi(\mathbf{0}, \mathcal{H}_\tau)] = \tau^{-\alpha} \mathbb{E}\,[\xi(\mathbf{0}, \mathcal{H}_1)]$, which, for example is the case when ξ measures length of edges of a graph on $\mathcal{H}_\tau$ ($\alpha = 1/d$), then (1.9) becomes

$$\lim_{\lambda \to \infty} \lambda^{-1} \sum_{x \in \mathcal{P}_\lambda} \xi_\lambda(x, \mathcal{P}_\lambda) = \mathbb{E}\,[\xi(\mathbf{0}, \mathcal{H}_1)] \int_A (\kappa(x))^{(d-\alpha)/d} dx \ \text{ in } \ L^q,$$

a limit appearing regularly in problems in Euclidean combinatorial optimization [25, 27].

To state central limit theorems we next put for all $\tau > 0$,

$$V^\xi(\tau) := \mathbb{E}\,[\xi(\mathbf{0}, \mathcal{H}_\tau)^2] + \tau \int_{\mathbb{R}^d} [\mathbb{E}\,\xi(\mathbf{0}, \mathcal{H}_\tau \cup \{z\}) \xi(z, \mathcal{H}_\tau \cup \{\mathbf{0}\})] - (\mathbb{E}\,[\xi(\mathbf{0}, \mathcal{H}_\tau)])^2 dz. \tag{1.11}$$

The scalars $V^\xi(\tau)$ should be interpreted as mean pair correlation functions for the functional ξ on homogenous Poisson points $\mathcal{H}_\tau$. By extending Lemma 1.1 to an analogous result giving the weak convergence of the joint distribution of $\xi_\lambda(x, \mathcal{P}_\lambda)$ and $\xi_\lambda(x + \lambda^{-1/d} z, \mathcal{P}_\lambda)$ for all pairs of points x and z in $\mathbb{R}^d$, we may show for exponentially stabilizing ξ and for bounded A that $\lambda^{-1} \mathrm{Var}[\langle f, \mu_\lambda \rangle]$ converges as $\lambda \to \infty$ to a weighted average of the mean pair correlation functions. Furthermore, by using either Stein's method [13, 19] or the cumulant method [3], we may establish the asymptotic normality of $\langle f, \lambda^{-1/2} \mu_\lambda \rangle$, $f \in B(A)$, as shown by:

Theorem 1.2 *(CLT [3, 13]) Let ξ be a homogeneously stabilizing (1.3) translation invariant functional satisfying the moment condition (1.6) for some $p > 2$. Suppose further that A is bounded and that ξ is* exponentially stabilizing *with respect to κ and A as in (1.5). Then for all $f \in B(A)$ we have*

$$\lim_{\lambda \to \infty} \lambda^{-1} \mathrm{Var}[\langle f, \mu_\lambda \rangle] = \int_A f(x) V^\xi(\kappa(x)) \kappa(x) dx \tag{1.12}$$

as well as convergence of the finite-dimensional distributions $(\langle f_1, \lambda^{-1/2}\overline{\mu}_\lambda\rangle, \ldots, \langle f_k, \lambda^{-1/2}\overline{\mu}_\lambda\rangle)$, $f_1, \ldots, f_k \in B(A)$, *to a Gaussian field with covariance kernel*

$$(f, g) \mapsto \int_A f(x)g(x)V^\xi(\kappa(x))\kappa(x)dx. \tag{1.13}$$

Stabilization methods yield analogous convergence results for the de-Poissonized finite-dimensional distributions $(\langle f_1, n^{-1/2}\overline{\rho}_n\rangle, \ldots, \langle f_k, n^{-1/2}\overline{\rho}_n\rangle)$, $f_1, \ldots, f_k \in B(A)$. In order to state our results we need one more definition. Put for all $\tau > 0$,

$$\delta(\tau) := \delta^\xi(\tau) := \mathbb{E}\left[\xi(\mathbf{0}, \mathcal{H}_\tau)\right] + \tau \int_{\mathbb{R}^d} [\mathbb{E}\left[\xi(\mathbf{0}, \mathcal{H}_\tau \cup \{y\}) - \xi(\mathbf{0}, \mathcal{H}_\tau)\right] dy. \tag{1.14}$$

Suppose that ξ is a homogeneously stabilizing (1.3) translation invariant functional satisfying the moment condition (1.6) for some $p > 2$. Suppose further that ξ is *exponentially stabilizing* as in (1.5) and *binomially exponentially stabilizing*, which, loosely speaking, means that (1.5) is satisfied when the Poisson point set $\mathcal{P}_\lambda$ is replaced by a binomial point set (see section 2 of [13] for details). Roughly speaking, if the moment condition (1.6) is also satisfied when $\mathcal{P}_\lambda$ is replaced by a binomial point set (see [13] for details) then for all $f \in B(A)$ we have

$$\lim_{n\to\infty} n^{-1}\mathrm{Var}[\langle f, \rho_n\rangle] = \int_A f(x)V^\xi(\kappa(x))\kappa(x)dx - \left(\int_A \delta(\kappa(x))\kappa(x)dx\right)^2 \tag{1.15}$$

as well as convergence of the finite-dimensional distributions $(\langle f_1, n^{-1/2}\overline{\rho}_n\rangle, \ldots, \langle f_k, n^{-1/2}\overline{\rho}_n\rangle)$, $f_1, \ldots, f_k \in B(A)$, to a Gaussian field with covariance kernel

$$(f, g) \mapsto \int_A f(x)g(x)V^\xi(\kappa(x))\kappa(x)dx - \left(\int_A \delta(\kappa(x))f(x)\kappa(x)dx\right)\left(\int_A \delta(\kappa(x))g(x)\kappa(x)dx\right). \tag{1.16}$$

Remarks.

(i) *Applicability*. The task of determining whether a given functional ξ is exponentially stabilizing is sometimes no more difficult than deciding whether it is homogenously stabilizing. This is the case with functionals involving nearest neighbor graphs, the subject of this note. On the other hand, there are situations where establishing stabilization is technically complicated, as is the case with random sequential packing [3, 17] and the Euclidean minimal spanning tree graph [9, 10, 18] and its variants [1].

(ii) *Exact constants*. Stabilization produces an explicit identification of limiting means (1.9), variances (1.12), and covariances (1.16) for the random measures (1.7). This yields exact constants in quantization problems [24], the total edge length of the nearest neighbors graph on i.i.d. samples [26, 18], as well as statistics of random samples including high dimensional versions of information gain and log-likelihood [4].

(iii) *Translation invariance*. For ease of exposition, the above results assume translation invariance of ξ. This assumption is not necessary and may be removed (see [3, 13, 14]), provided that we put $\xi_\lambda(x, \mathcal{X}) := \xi(x, x + \lambda^{1/d}(-x + \mathcal{X}))$.

(iv) *Extensions*. Analogous limit results hold for stabilizing functionals of Gibbsian input, provided that the potential decays fast enough [24]. Stabilization theory also extends to treat marked point processes [3, 13, 16, 17]. Large and moderate deviations for the sums (1.1) are studied in [2, 22].

Rates of convergence to the normal. Stabilization for point processes provides rates of convergence in both the weak law of large numbers (see Theorem 1.1 of [21]) and central limit theorems (see Corollary 2.1 of [19]). We assume that κ has compact support $A \subset \mathbb{R}^d$. For $\lambda > 0$, define the functional $H^\xi_\lambda := \sum_{x \in \mathcal{P}_\lambda \cap A} \xi_\lambda(x, \mathcal{P}_\lambda)$ and the centered version $\overline{H}^\xi_\lambda := H^\xi_\lambda - \mathbb{E}\,[H^\xi_\lambda]$. Suppose that there is a constant $\sigma^2(\xi, \kappa) \in [0, \infty)$ such that

$$\liminf_{\lambda \to \infty} \lambda^{-1} \mathrm{Var}[H^\xi_\lambda] = \sigma^2(\xi, \kappa). \tag{1.17}$$

The following is a special instance of Corollary 2.1 of [19], which provides rates of normal approximation for integrals of the point measures μ_λ against bounded test functions, whenever ξ is exponentially stabilizing in the sense of Definition 1.1. Let Φ be the cumulative distribution function for the standard normal. Put $\xi_\lambda(x, \mathcal{X}) := \xi(x, x + \lambda^{1/d}(-x + \mathcal{X}))$.

Theorem 1.3 *(Corollary 2.1 of [19]) Suppose $\|\kappa\|_\infty < \infty$. Suppose that ξ is exponentially stabilizing as in (1.5) and that ξ satisfies the moments condition (1.6) for some $p > 3$. If (1.17) holds with $\sigma^2(\xi, \kappa) > 0$, then there exists a finite constant C depending on d, ξ, κ, and p such that for all $\lambda \geq 2$,*

$$\sup_{t \in \mathbb{R}} \left| P\left[\frac{H^\xi_\lambda - \mathbb{E}\, H^\xi_\lambda}{\sqrt{\mathrm{Var}[H^\xi_\lambda]}} \leq t \right] - \Phi(t) \right| \leq C (\log \lambda)^{3d} \lambda^{-1/2}. \tag{1.18}$$

This result is a consequence of dependency graphs methods and a normal approximation result of Chen and Shao [8] using the Stein method. For the proof, we refer to [19] for complete details. The proof of Theorem 1.3 does not depend on the representation $\xi_\lambda(x, \mathcal{X}) := \xi(x, x + \lambda^{1/d}(-x + \mathcal{X}))$, but only upon the exponential stabilization of ξ and the moment bound (1.6) for some $p > 3$.

2 Statistical estimators of dimension

2.1 The Levina and Bickel dimension estimator

Estimating the intrinsic dimension of a high dimensional data set is a central problem in statistical analysis. Levina and Bickel [11] propose a dimension estimator making use of nearest neighbor statistics. The goal is to estimate the dimension of random variables lying on a manifold of unknown dimension m embedded in a higher dimensional space $\mathbb{R}^d$, $d \geq m$. For all $k = 3, 4, ...$, the Levina and Bickel estimator of the dimension m of a data cloud $\mathcal{X}$ in $\mathbb{R}^d$ is given by

$$\hat{m}_k := \hat{m}_k(\mathcal{X}) := (\mathrm{card}(\mathcal{X}))^{-1} \sum_{x \in \mathcal{X}} \hat{m}_k(x, \mathcal{X}), \tag{2.1}$$

where for all $x \in \mathcal{X}$ we have

$$\hat{m}_k(x, \mathcal{X}) := (k-2) \left(\sum_{j=1}^{k-1} \log \frac{D_k(x)}{D_j(x)} \right)^{-1}, \tag{2.2}$$

where $D_j(x) := D_j(x, \mathcal{X}), 1 \le j \le k$, are the distances between x and its jth nearest neighbor in $\mathcal{X}$. Notice that $\hat{m}_k$ is *scale invariant* in the sense that for all $\alpha > 0$ we have

$$\hat{m}_k(\alpha x, \alpha \mathcal{X}) = \hat{m}_k(x, \mathcal{X}). \tag{2.3}$$

If $\mathcal{X}_n := \{X_1, ..., X_n\}$, where the X_i are i.i.d. random variables having support on a submanifold $A \subset \mathbb{R}^d$, then Levina and Bickel [11] argue that the statistic $\hat{m}_k(\mathcal{X}_n)$ estimates the intrinsic dimension of $\mathcal{X}_n$, i.e., the dimension of A. Their arguments rest on the observation that if $\mathcal{H}$ is a rate one homogeneous Poisson point process on $\mathbb{R}^m$, then for all $y \in \mathbb{R}^m$ the sum $U := m \sum_{j=1}^{k-1} \log(D_k(y, \mathcal{H})/D_j(y, \mathcal{H}))$ has a Gamma$(k-1, 1)$ distribution so that $\mathbb{E}\, U^{-1} = (k-2)^{-1}$, i.e., for all $y \in \mathbb{R}^m$ we have $\mathbb{E}\,[\hat{m}_k(y, \mathcal{H})] = m$. Chatterjee [7] provides a rate of normal approximation for the statistic $\hat{m}_k(X_1, ..., X_n)$. For $k > 9$ he obtains rates of convergence with respect to the Kantorovich Wasserstein distance of the order $n^{-(k-9)/(2k-1)}$ under minimal assumptions on the distribution of X_i. Under appropriate conditions on k and X_i, Bickel and Yan [6] establish a central limit theorem for a centered and scaled version of $\hat{m}_k(X_1, ..., X_n)$ when the submanifold $A \subset \mathbb{R}^d$ is flat.

Here we consider another approach to establishing a rate of normal approximation. Letting $N := N(\lambda)$ be an independent Poisson random variable with mean $\lambda > 0$, consider i.i.d. random variables $X_1, X_2, ..., X_{N(\lambda)}$ whose distribution has support A. *We assume that* $X_1, X_2, ..., X_{N(\lambda)}$ *represents an embedding into* $\mathbb{R}^d$ *of a lower dimensional sample*, that is we assume there is some $m \in \mathbb{N}$ and some one to one and smooth $g : \mathbb{R}^m \to \mathbb{R}^d$ such that $X_i = g(Y_i)$, where Y_i are i.i.d. with a density κ' having bounded convex support $A' \subset \mathbb{R}^m$, and where κ' is bounded away from zero and infinity. We assume without loss of generality that $\{Y_i\}_{i=1}^{N(\lambda)}$ is the realization of a Poisson point process $\mathcal{P}'_\lambda$ on A' having intensity measure $\lambda \kappa'(x)dx$. Put $\mathcal{P}_\lambda := \{X_1, X_2, ..., X_{N(\lambda)}\}$. As in [11], we assume that $X_i,\ 1 \le i \le N(\lambda)$, are close iff $Y_i,\ 1 \le i \le N(\lambda)$, are close. More precisely, we assume there are positive constants K_1 and K_2 such that for all $y_1, y_2 \in A'$

$$K_1 ||y_1 - y_2|| \le ||g(y_1) - g(y_2)|| \le K_2 ||y_1 - y_2||. \tag{2.4}$$

Stabilization methods provide a rate of normal approximation in the sup norm distance for $\hat{m}_k(\mathcal{P}_\lambda)$, which may be seen as follows. Although $g : \mathbb{R}^m \to \mathbb{R}^d$ is unknown, it yields for each $k \in \mathbb{N}$ a function m_k^g defined on pairs $(y, \mathcal{Y})$, where $y \in \mathbb{R}^m$ and where $\mathcal{Y}$ is a finite point set in $\mathbb{R}^m$. Define for all $y \in \mathbb{R}^m$, all finite point sets $\mathcal{Y} \subset \mathbb{R}^m$, and all $\lambda > 0$

$$m_k^g(y, \mathcal{Y}) := \hat{m}_k(g(y), g(\mathcal{Y}))$$

and

$$m_{k,\lambda}^g(y, \mathcal{Y}) := \hat{m}_{k,\lambda}(g(y), g(\mathcal{Y})) := \hat{m}_k(\lambda^{1/m} g(y), \lambda^{1/m} g(\mathcal{Y})).$$

The scale invariance (2.3) of $\hat{m}_k$ implies that $m_{k,\lambda}^g(y, \mathcal{Y}) = m_k^g(y, \mathcal{Y})$ holds for all $\lambda > 0$.

With these definitions it follows that

$$\hat{m}_k(\mathcal{P}_\lambda) = \sum_{y \in \mathcal{P}'_\lambda} m_k^g(y, \mathcal{P}'_\lambda) = \sum_{y \in \mathcal{P}'_\lambda} m_{k,\lambda}^g(y, \mathcal{P}'_\lambda). \tag{2.5}$$

The functionals $m_{k,\lambda}^g$ act upon the input by the non-linear function g and thus they do not share the same representation as the functionals ξ_λ defined at (1.2). Still, the $m_{k,\lambda}^g$ are locally determined by points in neighborhoods of $\mathbb{R}^m$ having a diameter decaying exponentially fast. Using this observation, coupled with moment bounds for $m^g_{k,\lambda}$ and a modification of the dependency graph arguments of [19], it is possible to show that $\hat{m}_k(\mathcal{P}_\lambda)$ satisfies the rate of normal approximation (1.18). This approach to proving asymptotic normality holds for other functionals of nearest neighbor distances on manifolds and we shall give the details elsewhere.

2.2 The 'reach' statistic

Here we consider a second way to estimate the dimension of random data. As above, we let $\mathcal{X}$ be a locally finite subset in $\mathbb{R}^d$. For each $k \in \mathbb{N}$, let $G := G_k(\mathcal{X})$ be the directed k-nearest-neighbors graph over $\mathcal{X}$. Given vertices x and y in $\mathcal{X}$, and following [5], we say that y can be 'reached' in j steps from x, if there exists a path $v_0, v_1, \ldots, v_j$ in G, with $v_0 = x$ and $v_j = y$. The reach in j steps of vertex $x \in \mathcal{X}$, here denoted by $r_{j,k}(x, \mathcal{X})$, is the total number of vertices that can be reached from x in at most j steps using edges of G, that is

$$r_{j,k}(x, V) := \text{card}\{y \in V : \ y \neq x, \ \ y \ \text{ is reached in } \ l \ \text{ steps from } \ x; \ l \leq j\}.$$

When $\mathcal{X}$ is finite, we define the average reach statistic as

$$\bar{r}_{j,k}(\mathcal{X}) := (\text{card}(\mathcal{X}))^{-1} \sum_{x \in \mathcal{X}} r_{j,k}(x, \mathcal{X}). \tag{2.6}$$

As shown by the next result, the statistic (2.6) has the property that $\bar{r}_{j,k}(\{X_1, ..., X_n\})$ a.s. converges to a constant depending only on j, k, and d, whenever the X_i are i.i.d. with a density on $\mathbb{R}^d$. The significance of this in dimension estimation is explored in [5]. Recall that given $X_1, X_2, ..., X_n$ i.i.d. random variables we let $\mathcal{X}_n := \{X_1, ..., X_n\}$.

Theorem 2.1 *Fix $k \in \mathbb{N}$. Let $X_1, X_2, ..., X_n$ be i.i.d. random variables from a distribution on $\mathbb{R}^d$ having a density κ. Then for all $1 \leq j \leq k$, as $n \to \infty$ we have*

$$\bar{r}_{j,k}(\mathcal{X}_n) := n^{-1} \sum_{i=1}^{n} r_{j,k}(X_i, \mathcal{X}_n) \to \beta(j, k, d) \ \ a.s. \ \ and \ \ \ in \ \ L^2$$

where $\beta(j, k, d) := \mathbb{E}\left[r_{j,k}(\mathbf{0}, \mathcal{H})\right]$ and where $\mathcal{H}$ is a rate one homogeneous Poisson point process on $\mathbb{R}^d$.

Theorem 2.1 extends the analogous a.s. asymptotics of [5], which assumes continuity of κ. To prove Theorem 2.1 we proceed as follows. Notice that $r_{j,k}$ satisfies scale invariance in the sense that for all $\alpha > 0$ we have $r_{j,k}(x, \mathcal{X}) = r_{j,k}(\alpha x, \alpha \mathcal{X})$ and thus $r_{j,k}(X_i, \mathcal{X}_n) = r_{j,k}(n^{1/d} X_i, n^{1/d} \mathcal{X}_n)$. The stated convergence in L^2 is now an easy consequence of Theorem 1.1 and almost sure convergence follows from Azuma's inequality as in [5]. Indeed, the conditions of Theorem 1.1 are easily satisfied since $r_{j,k}$ are functions of nearest neighbor distances and are therefore homogeneously stabilizing. The $r_{j,k}$ are bounded by a constant depending only on j, k, and d thus they satisfy the moment bound (1.6) for all p.

Our last result provides conditions under which $\bar{r}_{j,k}(\mathcal{X}_n)$ satisfies a central limit theorem. It extends the analogous result of [5] to non-uniform random variables and provides explicit variance asymptotics. Given $\bar{r}_{j,k}$ and $\tau > 0$, let $V^{\bar{r}_{j,k}}(\tau)$ and $\delta^{\bar{r}_{j,k}}(\tau)$ be as in (1.11) and (1.14), respectively.

Theorem 2.2 *Fix $k \in \mathbb{N}$. Let $X_1, X_2, ..., X_n$ be i.i.d. random variables from a distribution on $\mathbb{R}^d$ having a density κ which is bounded away from zero and infinity on a bounded set A. Then*

$$\lim_{n\to\infty} n^{-1}\mathrm{Var}[\bar{r}_{j,k}(\mathcal{X}_n)] = \sigma^2_{j,k},$$

where

$$\sigma^2_{j,k} := \int_A V^{\bar{r}_{j,k}}(\kappa(x))\kappa(x)dx - \left(\int_A \delta^{\bar{r}_{j,k}}(\kappa(x))\kappa(x)dx\right)^2.$$

Also, as $n \to \infty$ it is the case that

$$n^{-1/2}(\bar{r}_{j,k}(\mathcal{X}_n) - \mathbb{E}\,\bar{r}_{j,k}(\mathcal{X}_n)) \xrightarrow{\mathcal{D}} N(0, \sigma^2_{j,k}).$$

The proof Theorem 2.2 is an immediate consequence of Theorem 1.2 applied to the reach functionals $r_{j,k}$, where for all $n \in \mathbb{N}$ we recall that $r_{j,k}(X_i, \mathcal{X}_n) = r_{j,k}(n^{1/d}X_i, n^{1/d}\mathcal{X}_n)$. Indeed, it is easy to show that the reach functional $r_{j,k}$ is exponentially stabilizing under the given conditions on κ (see Lemma 6.1 of [16] and Theorem 2.4 of [18]). Since $r_{j,k}$ are bounded all of the conditions of Theorem 1.2 are satisfied and Theorem 2.2 follows.

Acknowledgements. It is a pleasure to thank Persi Diaconis for discussions related to the Levina and Bickel statistic and for communicating Peter Bickel's question. I thank Sourav Chatterjee and Mathew Penrose for comments leading to an improved exposition and I thank Peter Bickel for sharing the preprint [6].

References

[1] F. Baccelli and C. Bordenave (2007), The radial minimal spanning tree of a Poisson point process, *Ann. Appl. Probab.*, **17**, 305-359.

[2] Y. Baryshnikov, P. Eichelsbacher, T. Schreiber, and J. E. Yukich (2008), Moderate deviations for some point measures in geometric probability, *Annales de l'Institut Henri Poincaré - Probabilités et Statistiques*, **44**, 3, 422-446.

[3] Y. Baryshnikov and J. E. Yukich (2005), Gaussian limits for random measures in geometric probability, *Ann. Appl. Probab.*, **15**, 1A, 213-253.

[4] Y. Baryshnikov, M. Penrose, and J. E. Yukich (2008), Gaussian limits for generalized spacings, *Ann. Appl. Probab.*, to appear, Electronically available via `http://www.lehigh.edu/~jey0/publications.html`.

[5] M. Brito, A. Quiroz, and J. E. Yukich (2002), Graph-theoretic procedures for dimension identification *Journal of Multivariate Analysis*, **81**, 67-84.

[6] Bickel, P. and D. Yan (2008), Sparsity and the possibility of inference, *Sankhya*, to appear.

[7] S. Chatterjee (2008), A new method of normal approximation, *Ann. Probab.*, **36**, 4, 1584-1610.

[8] L. Chen and Q.-M. Shao (2004), Normal approximation under local dependence, *Ann. Probab.* **32**, 1985-2028.

[9] H. Kesten and S. Lee (1996), The central limit theorem for weighted minimal spanning trees on random points, *Ann. Appl. Probab.*, **6**, 495-527.

[10] S. Lee (1997), The central limit theorem for Euclidean minimal spanning trees I, *Ann. Appl. Probab.*, **7**, 996-1020.

[11] E. Levina and P. J. Bickel (2005), Maximum likelihood estimation of intrinsic dimension, in *Advances in NIPS*, **17**, Eds. L. K. Saul, Y. Weiss, L. Bottou.

[12] M. D. Penrose (2005), Multivariate spatial central limit theorems with applications to percolation and spatial graphs, *Ann. Probab.*, **33**, 1945- 1991.

[13] M. D. Penrose (2007), Gaussian limits for random geometric measures, *Electron. J. Probab.* **12**, 989-1035.

[14] M. D. Penrose (2007), Laws of large numbers in stochastic geometry with statistical applications, *Bernoulli*, **13**, 4, 1124-1150.

[15] M. D. Penrose (2008), Existence and spatial limit theorems for lattice and continuum particle systems, *Probability Surveys* **5**, 1-36.

[16] M. D. Penrose and J. E. Yukich (2001), Central limit theorems for some graphs in computational geometry, *Ann. Appl. Probab.* **11**, 1005-1041.

[17] M. D. Penrose and J. E. Yukich (2002), Limit theory for random sequential packing and deposition, *Ann. Appl. Probab.* **12**, 272-301.

[18] M.D. Penrose and J.E. Yukich (2003), Weak laws of large numbers in geometric probability, *Ann. Appl. Probab.*, **13**, pp. 277-303.

[19] M. D. Penrose and J. E. Yukich (2005), Normal approximation in geometric probability, in Stein's Method and Applications, Lecture Note Series, Institute for Mathematical Sciences, National University of Singapore, **5**, A. D. Barbour and Louis H. Y. Chen, Eds., 37-58.

[20] T. Schreiber (2008), Limit theorems in stochastic geometry, New Perspectives in Stochastic Geometry, Oxford University Press, to appear.

[21] T. Schreiber, M. D. Penrose, and J. E. Yukich (2007), Gaussian limits for multidimensional random sequential packing at saturation, *Comm. Math. Physics*, 272, 167-183.

[22] T. Schreiber and J. E. Yukich (2005), Large deviations for functionals of spatial point processes with applications to random packing and spatial graphs, *Stochastic Processes and Their Applications*, **115**, 1332-1356.

[23] T. Schreiber and J. E. Yukich (2008), Variance asymptotics and central limit theorems for generalized growth processes with applications to convex hulls and maximal points, *Ann. Probab.*, **36**, 363-396.

[24] T. Schreiber and J. E. Yukich (2008), Stabilization and limit theorems for geometric functionals of Gibbs point processes, preprint.

[25] J. M. Steele (1997), Probability Theory and Combinatorial Optimization, SIAM.

[26] A. Wade (2007), Explicit laws of large numbers for random nearest neighbor type graphs, *Adv. in Appl. Probab.*, **39**, 326-342.

[27] J. E. Yukich (1998), Probability Theory of Classical Euclidean Optimization Problems, *Lecture Notes in Mathematics*, **1675**, Springer, Berlin.