Detection of an Anomalous Cluster in a Network

Ery Arias-Castro[1,*], Emmanuel J. Candès[2] and Arnaud Durand[3]

[1] *Department of Mathematics, University of California, San Diego; eariasca@ucsd.edu*
[2] *Departments of Mathematics and Statistics, Stanford University; candes@stanford.edu*
[3] *Department of Mathematics, Université Paris-Sud 11, Orsay; arnaud.durand@math.u-psud.fr*
[*] *Corresponding author*

Abstract. *We consider the model surveillance problem of detecting whether or not in a given sensor network, there is a cluster of sensors which exhibit an "unusual behavior." Formally, suppose we are given a set of nodes (sensors) and attach a time series to each node (information transmitted by the sensor). We observe a realization of this process over time and want to decide between the null, where all the variables are i.i.d. standard normal; and the alternative, where there is an emerging cluster of i.i.d. normal variables with positive mean and unit variance. The growth models used to represent the emerging cluster are quite general, and in particular include cellular automata used in modelling epidemics. We consider classes of clusters that are quite general, for which we obtain a lower bound on their respective minimax detection rate, and show that some form of scan statistic, by far the most popular method in practice, achieves that same rate within a logarithmic factor. Our results are not limited to the normal location model, but generalize to any one-parameter exponential family when the anomalous clusters are large enough. This is an extended abstract of [3].*

Keywords. *Cluster detection; Sensor network; Scan statistic; Disease outbreak detection.*

1 Introduction

We consider a spatio-temporal surveillance setting where a set of nodes transmit information over time to a central location. Under normal circumstances, the variables $X_v(t)$ behave similarly, while under abnormal circumstances, there is an emerging cluster of nodes returning slightly larger values. This models a wide-array of real-life situations, for example, the monitoring of hazardous materials [5] and target tracking [11] based on sensor networks; object tracking from video frames [12] (a digital camera may be seen as a sensor network, with CCD or CMOS pixel sensors); or the early detection of epidemics [7],

with surveillance systems now incorporating data from hospital emergency visits, ambulance dispatch calls and pharmacy sales of over-the-counter drugs.

Mathematical framework. Let $\mathbb{V}_m \subset \mathbb{R}^d$ be a set of m nodes. To each node $v \in \mathbb{V}_m$, we attach a time series, $(X_v(t) : t = 0,1,\dots,t_m)$. Our analysis is in the asymptotic setting $m \to \infty$ and $t_m \to \infty$. Let $\mathcal{K}_m$ be a class of cluster sequences of the form $(K_t : t = 0,1,\dots,t_m)$ such that $K_t \subset \mathbb{V}_m$ for all $t = 0,1,\dots,t_m$. For example, a space-time cylinder, e.g. a model used in disease outbreak detection [10], is a cluster sequence of the form $K_t = \{v \in \mathbb{V}_m : \|v - x_0\| \le r_0\}$, if $t \ge t_0$, and $K_t = \emptyset$ otherwise, so that t_0 is the time origin and x_0 the center of the emerging cluster. Another example is that of a space-time cone, of the form $K_t = \{v \in \mathbb{V}_m : \|v - x_0\| \le C(t - t_0)\}$, if $t \ge t_0$, and $K_t = \emptyset$ otherwise, where C controls how fast the cluster grows over time. The random variables $(X_v(t) : v \in \mathbb{V}_m, t = 0,1,\dots,t_m)$ are assumed independent. For concreteness, we consider a normal location model which is popular in signal and image processing to model the noise. Our analysis, however, generalizes to any exponential family under some condition on the sizes of the anomalous clusters, such as Bernoulli models which arise in sensor arrays where each sensor collects one bit (i.e. makes a binary decision) or Poisson models which come up with count data, e.g. arising in infectious disease surveillance systems. We assume the process is calibrated so that, under normal circumstances, the variables $X_v(t) \sim \mathcal{N}(0,1)$ for all $v \in \mathbb{V}$. Under abnormal circumstances, with emerging cluster $K = (K_t) \in \mathcal{K}_m$, we assume that $X_v(t) \sim \mathcal{N}(\theta_m |K|^{-1/2}, 1)$ for all $(v,t) \in K$, and $X_v(t) \sim \mathcal{N}(0,1)$ for all $(v,t) \notin K$, where $\theta_m > 0$ and $|K|$ is the size of K as a space-time cluster. The emerging cluster K is unknown. We adopt a minimax point of view, where the risk of a test is the sum of its probability of type I error and the maximum of its probability of type II error among all the specific alternatives (here, cluster sequences in the class). We say that the hypotheses are asymptotically separable if there is a sequence of tests with risk tending to zero, and that they are asymptotically inseparable if all sequences of tests have risk tending to one.

The scan statistic. The generalized likelihood-ratio test rejects for large values of

$$\max_{K \in \mathcal{K}_m} \frac{1}{\sqrt{|K|}} \sum_{v \in K} X_v.$$

Without the normalization, this is the so-called scan statistic, the prevalent method in disease outbreak detection, with many variations [9]. This is the matched filters method ubiquitous in problems of detection in a wide variety of fields, sometimes in the form of deformable templates in the engineering literature [8]. Note that the scan statistic is As advocated in [2], we will not use the scan statistic directly, but rather restrict the scanning to a subset of $\mathcal{K}_m$. More precisely, we will introduce the following metric on subsets of nodes, $K, L \subset \mathcal{K}_m$,

$$\delta(K,L) = \sqrt{2}\left(1 - |K \cap L| / \sqrt{|K||L|}\right)^{1/2},$$

and will restrict the scanning to an ε-net of $\mathcal{K}_m$ with respect to δ, i.e. a subset $\{K_j : j \in J\} \subset \mathcal{K}_m$, with the property that, for each $K \in \mathcal{K}_m$ there is $j \in J$ such that $\delta(K, K_j) \le \varepsilon$. When J is minimal, we call the resulting statistic an ε-scan statistic. The approximation precision ε will be chosen appropriately depending on the situation.

Contribution. Within the mathematical framework we obtain a lower bound on the minimax detection rate for θ_m for a large class of cluster sequences with some sort of limit, which many models for epidemics satisfy [1], and prove that an ε-scan statistic achieves that detection threshold.

2 Main results

The set of nodes. We assume that the nodes are embedded in $\Omega_d \subset \mathbb{R}^d$, a compact set with non-empty interior. Let $B(x,r)$ denote the (open) Euclidean ball with center x and radius r. We consider a finite subset $\mathbb{V}_m \subset \Omega_d$ of size m, which is evenly spread-out in the following sense: there is a constant $C \geq 1$ independent of m and a sequence $r_m^* \to 0$, such that,

$$C^{-1} m r^d \leq |B(x,r) \cap \mathbb{V}_m| \leq C m r^d, \quad \forall r \in [r_m^*, 1],\ \forall x \in \Omega_d. \tag{1}$$

In words, the number of nodes in any ball that is not too small is roughly proportional to its volume. For the regular lattice,

$$\mathbb{V}_m = \{0, m^{-1/d}, \ldots, 1 - m^{-1/d}\}^d \subset \Omega_d = [0,1]^d,$$

(assuming $m^{1/d}$ is an integer) condition (1) is satisfied for $r_m^* > \sqrt{d} m^{-1/d}$. This is the smallest possible order of magnitude. When $\mathbb{V}_m$ is obtained from sampling m points from the uniform distribution on Ω_d, (1) is satisfied with high probability for $r_m^* \geq C(\log(m)/m)^{1/d}$, when C is large enough.

Lower bound: detecting space-time cylinders. The simplest class of cluster sequences is that of space-time cylinders introduced earlier. For that class we have the following lower bound on the detection threshold.

Proposition. *Consider $\lambda_m \to 0$, with $\lambda_m \geq r_m^*$, and let $\mathcal{K}_m$ be the class of all space-time cylinders of the form $K_t = B(x, \lambda_m) \cap \mathbb{V}_m$, $\forall t = 0, \ldots, t_m$, where $x \in \Omega_d$. Then the hypotheses are asymptotically inseparable if*

$$\overline{\lim_{m \to \infty}}\ \theta_m (\log(1/\lambda_m^d))^{-1/2} < \sqrt{2}.$$

With only one possible shape and known time origin, such a model is rather limited. We now consider much larger class of cluster sequences with some sort of limit (in the sense of (2)), and show that, nevertheless, a form of scan statistic achieves that same detection rate.

Upper bound: cluster sequences with a limit. Motivated by the fact that cellular automata, which have been used to model epidemics [1], develop an asymptotic shape under some conditions [6, 4], we consider cluster sequences with some sort of (spatial) limit. We first define a class of spatial clusters with Lipschitz boundary that are not too thin. For $\kappa \geq 1$, let $\mathcal{F}_{d,d}(\kappa)$ be the subclass of bi-Lipschitz functions $f : B(0,1) \to \Omega_d$ such that $\lambda_f \lambda_{f^{-1}} \leq \kappa$, where λ_f denotes the Lipschitz constant of f. For $f \in \mathcal{F}_{d,d}(\kappa)$, the spatial cluster $f(B(0,1)) \cap \mathbb{V}_m$, which is the set of nodes that belong to the range of f, is blob-like in that it contains a ball of radius λ_f/κ and is contained within a ball of radius λ_f, so that κ control its aspect ratio. We focus here on cluster sequences obeying $K_{t_m} \neq \emptyset$, i.e. the anomalous cluster is present at the last time point. This is a standing assumption in syndromic surveillance systems [10] and any prospective surveillance setting. For a cluster sequence $K = (K_t, t \in \mathrm{T}_m)$, let $t_K = \min\{t : K_t \neq \emptyset\}$, which is the time when K originates.

Theorem. *Consider sequences $\lambda_m \to 0$ with $\lambda_m \geq r_m^*$ and $\log\log t_m = o(\log(1/\lambda_m))$, and a function $\nu(t)$ with $\lim_{t\to\infty} \nu(t) = 0$ and $\nu(t) \leq 1$ for all $t \geq 0$. Let $\mathcal{K}_m$ be a class of cluster sequences such that $t_m - \max\{t_K : K \in \mathcal{K}_m\} \to \infty$, and for each $K = (K_t, t \in \mathrm{T}_m) \in \mathcal{K}_m$ there is $f \in \mathcal{F}_{d,d}(\kappa)$ with $\lambda_f \geq \lambda_m$, such that*

$$\delta(K_t, f(B(0,1)) \cap \mathbb{V}_m) \leq \nu(t - t_K), \quad \forall t = 0, 1, \ldots, t_m. \tag{2}$$

Then there is a scan statistic over a family of space-time cylinders that asymptotically separates the hypotheses if

$$\underline{\lim_{m \to \infty}}\ \theta_m (\log(1/\lambda_m^d))^{-1/2} > \sqrt{2}.$$

If the starting time is uniformly bounded away from t_m and the convergence to the thick spatial cluster (in the sense of (2)) occurs at a uniform speed, all the cluster sequences in the class have sufficient time to develop into their 'limiting' shapes. The space-time cylinders over which we scan are based on an ε-net for the possible limiting shapes, i.e. the class of spatial clusters generated by $\mathcal{F}_{d,d}(\kappa)$. Scanning over space-time cylinders (with balls as bases) is advocated in the disease outbreak detection literature [10]. Though seemingly naive, this approach achieves, in our asymptotic setting, the minimax detection rate if the cluster sequences develop into balls, and in general falls short by a constant factor.

Acknowledgments. Work partially supported by a grant from the US National Science Foundation (DMS-06-03890) and a grant from the US Office of Naval Research.

References

[1] S. Agur, O. Diekmann, H. Heesterbeek, J. Cushing, M. Gyllenberg, M. Kimmel, F. Milner, P. Jagers, and T. Kostova, editors. *Epidemiology, cellular automata, and evolution*. Elsevier Ltd, Oxford, 1999.

[2] E. Arias-Castro, D.L. Donoho, and X. Huo. Near-optimal detection of geometric objects by fast multiscale methods. *IEEE Transactions on Information Theory*, 51(7):2402–2425, 2005.

[3] E. Arias-Castro, E.J. Candès, and A. Durand. Detection of an abnormal cluster in a network. Available from `http://arxiv.org/abs/1001.3209`, 2009.

[4] T. Bohman and J. Gravner. Random threshold growth dynamics. *Random Structures Algorithms*, 15(1):93–111, 1999.

[5] S.M. Brennan, A.M. Mielke, D.C. Torney, and A.B. Maccabe. Radiation detection with distributed sensor networks. *Computer*, 37(8):57–59, 2004.

[6] J. Gravner and D. Griffeath. Random growth models with polygonal shapes. *Annals of Probability*, 34(1):181–218, 2006.

[7] R. Heffernan, F. Mostashari, D. Das, A. Karpati, M. Kulldorff, and D. Weiss. Syndromic surveillance in public health practice, New York City. *Emerging Infectious Diseases*, 10(5):858–864, 2004.

[8] A.K. Jain, Y. Zhong, and M.P. Dubuisson-Jolly. Deformable template models: A review. *Signal Processing*, 71(2):109–129, 1998.

[9] M. Kulldorff, R. Heffernan, J. Hartman, R. Assuncao, and F. Mostashari. A space-time permutation scan statistic for disease outbreak detection. *PLOS Medicine*, 2(3):216, 2005.

[10] M. Kulldorff. Prospective time periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society Series A*, 164(1):61–72, 2001.

[11] D. Li, K.D. Wong, Yu Hen Hu, and A.M. Sayeed. Detection, classification, and tracking of targets. *Signal Processing Magazine, IEEE*, 19(2):17–29, Mar 2002.

[12] Y. Zhong, A.K. Jain, and M.P. Dubuisson-Jolly. Object tracking using deformable templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2(5):544–549, 2000.