Spatial Inference in Sensor Networks using Multiple Hypothesis Testing and Bayesian Clustering

Martin Gölz*†, Michael Muma†, Topi Halme*, Abdelhak Zoubir† and Visa Koivunen*
Aalto University*, Finland and TU Darmstadt†, Germany
Email: {goelz, muma, zoubir}@spg.tu-darmstadt.de, {topi.halme, visa.koivunen@aalto}.fi

***Abstract*—The problem of statistical inference in large-scale sensor networks observing spatially varying fields is addressed. A method based on multiple hypothesis testing and Bayesian clustering is proposed. The method identifies homogeneous regions in a field based on similarity in decision statistics and locations of the sensors. High detection power is achieved while keeping false positives at a tolerable level. A variant of the EM-algorithm is employed to associate sensors with clusters. The performance of the method is studied in simulation using different detection theoretic criteria.**

***Index Terms*—IoT, p-values, Distributed Inference, Statistical Signal Processing, Large-Scale Sensor Networks, BIC**

I. Introduction

Observing and monitoring phenomena that occur within a spatial field is essential to a variety of applications [1]. This includes tasks such as detecting occupied radio spectrum in shared spectrum environments, identifying regions of poor air quality in environmental monitoring, smart buildings and different Internet of Things (IoT) applications. Many of these practical problems can be modeled using a multiple hypothesis testing framework, with the goal of identifying homogeneous spatial regions within which a defined null hypothesis H_0 (e.g. pollution remaining at tolerable level, radio spectrum being unoccupied) is in place, and regions where alternative hypotheses are true. These regions can be formed by assessing observations made by multiple sensors placed at distinct locations. In an IoT setup with a massive amount of available sensors connected to a common Fusion Center (FC), one might consider communicating all the acquired data to the FC, which then segments the field into sub-regions corresponding to null and alternate hypotheses. However, transmitting all observed data leads to serious communication overhead and significantly reduced lifespan of the sensor network due to power consumption. It is thus favorable to conduct distributed inference, where each sensor produces a single test statistic condensing its observed evidence on hypothesis H_0, for example, in form of a p-value. A p-value quantifies evidence against H_0 independent of the observation data model, allowing different probability models at different sensors. This is desirable since the properties of the observed phenomenon or field typically vary locally but smoothly.

Assuming that an occurring phenomenon influences not just a single point but is continuous and varies smoothly in space, we can increase the detection performance and reduce error levels in a network with high sensor density by forming clusters of sensors that are close to each other and exhibit similar decision statistics, in our case p-values. This helps if the evidence collected at a single sensor is not be reliable enough to reject H_0. Accounting for similar evidence accumulated over multiple sensors might allow rejecting H_0 reliably. In fact, we will demonstrate that the algorithm proposed in Sec. III detects an event associated with alternative hypotheses even if individual sensors within the affected area would in majority decide in favor of the null.

Any distinct location in the surveillance area is coupled with probabilities of H_0 or alternative hypotheses H_m. By observing the field, we will find out how likely H_0 or H_m hold at this sensor location. Each sensor computes a p-value from its observations and communicates it to the FC, which combines p-values from multiple sensors with known sensor location information to identify the homogeneous sub-regions. Observing p-values at different sensor locations is equivalent to sampling their field. Finally, each sensor is associated with membership of its most probable cluster that corresponds to either the null or alternative hypotheses. The field consists of an unknown number of alternative regions caused by phenomena letting the null hypothesis fail and a null region, that comprises the areas within which H_0 holds. The number of occurring events and thus resulting alternative areas, their positions and sizes are assumed to be unknown. We derive a novel Bayesian Cluster Enumeration algorithm for our data model to select the best fitting spatial field model from a family of candidate models. The method stems from the idea in [2], where a decision criterion for Gaussian data is provided.

In multiple comparisons problems, classic single sensor type I and II error levels are not sufficiently meaningful in the sense that they do not encourage statements about confidence in discoveries. As the number of tested hypotheses grows, the probability of committing at least one false alarm reaches unity quickly [3]. Moreover, the proportion of false discoveries may become intolerably high. The first False Discovery Rate (FDR) method that controls the expected proportion of false positives among all discoveries was introduced to alleviate these problems [4]. Efforts on controlling the FDR in a spatial context were made in [5], where the sensors had to be clustered before the actual inference. Typically no such prior clustering information is available. Therefore, dynamic estimation of clusters from observed test statistics should improve the overall performance. In [6], FDR is controlled by estimating parameters of a priori probabilities for H_0 and H_m spatially varying across the grid.

978-9-0827-9703-9/19/$31.00 ©2019 IEEE

The paper is structured as follows. In Sec. II, we introduce the data model and derive the Bayesian Information Criterion (BIC) that is later used to select the best fitting spatial field from a number of candidate models that the hypothesis testing is based upon. The proposed spatial inference algorithm is derived in Sec. III. In Sec. IV, we demonstrate the validity of the approach by comparing it to other potentially applicable algorithms and outline a methodology to use it for FDR-control. Conclusions are drawn in Sec. V.

II. Bayesian Cluster Enumeration

A. Data Model and Problem Formulation

Let us consider a sensor network of N sensors and denote by p_n the p-value of the nth sensor in known location (x_n, y_n), where $n = 1, \dots, N$. Feature vectors from all sensors, i.e., $\mathcal{S} \triangleq \{\mathbf{s}_1, \dots, \mathbf{s}_N\}$ are available at a FC, where $\mathbf{s}_n = [p_n, x_n, y_n]^\top$. An unknown number, K, of events associated with alternative hypotheses occur at unknown locations within the network, each of them causing H_0 to not hold within a spatially homogeneous region described by parameter vector $\boldsymbol{\zeta}_k$. All N_k sensors within the same region $\boldsymbol{\zeta}_k$ belong to independent, mutually exclusive and non-empty clusters $\mathcal{C}_k$, $k = 1, \dots, K$. The remaining nodes form the cluster $\mathcal{C}_0$, for which H_0 holds. The quantities K, $\boldsymbol{\zeta}_k$, N_k, and the amount of evidence against H_0 acquired at each affected sensor are unknown. Assuming statistically independent p-values, the p_n of the N_0 sensors in $\mathcal{C}_0$ are distributed uniformly [7], i.e., $P \sim \mathcal{U}(0,1)$. According to [8], under H_m, $P \sim \mathcal{B}(a)$, where $\mathcal{B}(a) = ap^{(a-1)}$ is the beta distribution with a single free shape parameter a. The p-value probability density functions (PDF) are assumed to be identical within the clusters but might differ between clusters, as expressed by the associated beta distribution shape parameters a_k, $k = 1, \dots, K$. Hence, the p-value PDF conditioned on a sensor belonging to an alternative cluster $\mathcal{C}_k,\ k = 1, \dots, K$ is

$$f_{p\text{-val}}(p|\mathcal{C}_k) = \mathcal{B}(a_k) = \begin{cases} a_k p^{(a_k-1)} & p \in (0,1) \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

while the PDF conditioned on the sensor being in $\mathcal{C}_0$ is

$$f_{p\text{-val}}(p|\mathcal{C}_0) = \mathcal{B}(1) = \mathcal{U}(0,1) = \begin{cases} 1 & p \in (0,1) \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

The PDF representing the probability that a sensor is placed at (x, y) given that it belongs to $\mathcal{C}_k$ becomes

$$f_{\text{coord}}(x, y|\mathcal{C}_k) = \frac{\tau^{(k)}(x, y)}{A(\mathcal{C}_k)}, \tag{3}$$

with $\tau^{(k)}(x, y) = 1$ if $(x, y) \in A(\mathcal{C}_k)$ and $\tau^{(k)}(x, y) = 0$ otherwise, $A(\mathcal{C}_k)$ denoting the area covered by cluster $\mathcal{C}_k$, ensuring that the PDFs integrate to 1. With (1)-(3), the conditional PDF of the feature vector $\mathbf{s}_n$ results in

$$f(\mathbf{s}_n|\mathcal{C}_k) = f_{p\text{-val}}(p_n|\mathcal{C}_k) f_{\text{coord}}(x_n, y_n|\mathcal{C}_k). \tag{4}$$

The clusters $\mathcal{C}_k, k = 0, \dots, K$ are thus fully described by the parameter matrix $\boldsymbol{\Theta}_K \triangleq [\boldsymbol{\theta}_0, \dots, \boldsymbol{\theta}_K]$, with $\boldsymbol{\theta}_k = [a_k, \boldsymbol{\zeta}_k]^\top$.

The goal of this work is to detect the regions for which H_0 does not hold by partitioning $\mathcal{S}$ into $K + 1$ mutually exclusive clusters $\mathcal{C}_0, \dots, \mathcal{C}_K$. In Sec. III, we propose an algorithm for identifying K, the clusters $\mathcal{C}_k$ and assigning cluster memberships for each sensor. The proposed method for estimating K stems from the two-stage Bayesian cluster enumeration method in [2]. From Eq. (18) in [2] onward, a multivariate Gaussian model is assumed, which is in contrast to our uniform and beta distributed data. In the following subsection, we will derive a new Bayesian cluster enumeration criterion for the data model introduced above.

B. Proposed Bayesian Cluster Enumeration Criterion

To estimate K, we select the candidate model that clusters the available test statistics given candidate models M_l of orders $l = 1, \dots, L_{\max}$, with maximum posterior probability

$$M_{\hat{K}} = \arg\max_{M_l} p\left(M_l|\mathcal{S}\right). \tag{5}$$

From (5), the authors of [2] derive a general Bayesian cluster enumeration criterion in which both, the network log-likelihood function (LLF) and penalty term, depend on the data model. In its general form, for a large number of sensors ($N \to \infty$), their BIC becomes

$$\text{BIC}(M_l) = \log \mathcal{L}_N\left(\boldsymbol{\tau}_l, \boldsymbol{\Theta}_l\right) - \frac{1}{2}\sum_{k=1}^{l} \log\left|\hat{\mathbf{J}}_k\right|, \tag{6}$$

where $\log \mathcal{L}_N\left(\boldsymbol{\tau}_l, \boldsymbol{\Theta}_l\right)$ is the LLF and $\left|\hat{\mathbf{J}}_k\right|$ is the determinant of the estimated Fisher Information Matrix (FIM). We derive the FIM for the data model of this paper in Eq. (20). Note that the LLF depends on the parameter matrix $\boldsymbol{\Theta}_l$ for model order l and the matrix of cluster association coefficients $\boldsymbol{\tau}_l = [\boldsymbol{\tau}_l^{(1)}, \dots, \boldsymbol{\tau}_l^{(N)}]$, summarizing the association vectors $\boldsymbol{\tau}_l^{(n)} = [\tau_l^{(n,0)}, \dots, \tau_l^{(n,K)}]$ of the individual sensors. When $\tau_l^{(k,n)} := \tau_l^{(k)}(x_n, y_n) = 1$, sensor n at (x_n, y_n) belongs to cluster $\mathcal{C}_k$. With (3), and assuming mutual exclusiveness of the non-overlapping clusters, only one coordinate PDF provides a non-zero value for a sensor at (x_n, y_n), i.e., $\tau_l^{(k,n)} = 1$ for the kth coordinate PDF providing a non-zero value and all other entries of $\boldsymbol{\tau}_l^{(n)}$ being zero.

For the remainder of this paper, we assume an initial circular shape for the alternative clusters $\mathcal{C}_k,\ k = 1, \dots, K$. This models, for example, a point source causing an event that affects all sensors within a certain distance to the source (e.g. point source and radio wave attenuation) similarly. In Sec. IV, we show that even if the true cluster shapes differ significantly from circular shape, the algorithm can identify different cluster shapes reliably. We initially assume circular clusters, $\boldsymbol{\zeta}_k = [x_{\text{c},k}, y_{\text{c},k}, r_k]^\top$, with radius r_k, and cluster center coordinates $x_{\text{c},k}$ and $y_{\text{c},k}$, respectively. Further, $A\left(\mathcal{C}_k\right) = r_k^2\pi$, and $A\left(\mathcal{C}_0\right) = 1 - \sum_{k=1}^{K} r_k^2\pi$. The network LLF becomes

$$\begin{aligned} \log \mathcal{L}_N(\boldsymbol{\tau}_l, \boldsymbol{\Theta}_l) &= \sum_{n=1}^{N} \log \sum_{k=0}^{l} f\left(\mathbf{s}_n|\mathcal{C}_k\right) \cdot P\left(\mathcal{C}_k\right) \\ &= \sum_{n=1}^{N} \log \Bigg(\tau_l^{(0,n)} \frac{1}{1 - \sum_{i=k}^{l} \pi r_i^2} \frac{N_0}{N} \\ &\qquad + \sum_{k=1}^{l} \tau_l^{(k,n)} \frac{1}{\pi r_k^2} \mathcal{B}\left(p_n; a_k\right) \frac{N_k}{N} \Bigg), \end{aligned} \tag{7}$$

where $P\left(\mathcal{C}_k\right) = N_k/N$ is the probability of cluster $\mathcal{C}_k$ with $N_k = \sum_{n=1}^{N} \tau_l^{(k,n)}$. Using the definition of the FIM, its entry at index $[i,j]$ is given by

$$\hat{\mathbf{J}}_{k,[i,j]} = E\left[\left(\frac{\partial}{\partial \theta_{k,i}} \log \mathcal{L}\left(\boldsymbol{\theta}_k|\mathcal{C}_k\right)\right)\left(\frac{\partial}{\partial \theta_{k,j}} \log \mathcal{L}\left(\boldsymbol{\theta}_k|\mathcal{C}_k\right)\right)\right], \tag{8}$$

where the LLF conditioned on being in cluster $\mathcal{C}_k$ for $k = 1, \ldots, K$ and using parameters r_k and a_k is

$$\begin{aligned}\log \mathcal{L}(\boldsymbol{\theta}_k|\mathcal{C}_k) &= \sum_{m=1}^{N_k} \log\left(\frac{1}{\pi r_k^2} a_k p_m^{(a_k-1)} \frac{N_k}{N}\right) \qquad (9)\\ &= -N_k \log\left(\pi\right) - 2N_k \log\left(r_k\right) + N_k \log\left(a_k\right)\\ &\quad + \left(a_k - 1\right) \sum_{m=1}^{N_k} \log\left(p_m\right) + N_k \log\left(\frac{N_k}{N}\right).\end{aligned}$$

III. Proposed Algorithm

This section describes the proposed algorithm to assign the sensors to clusters $\mathcal{C}_1, \ldots, \mathcal{C}_K$ associated with H_m, or to the null cluster representing H_0. The proposed method stems from an Expectation-Maximization (EM) algorithm with cluster association coefficients as latent variables. The model parameter estimates $\hat{\boldsymbol{\theta}}_k$ are updated based on the feature vectors $\mathbf{s}_n$, weighted by the estimated probability $\hat{v}_{n,k}$ of belonging to respective cluster $\mathcal{C}_k$. To be consistent with the EM-framework, the uniform coordinate PDF in (3) must be replaced by a smooth approximation thereof, i.e.,

$$f_{\text{EM,coord}}(x, y|\mathcal{C}_k) = \frac{1}{\nu}\left(1 - \tanh\left(\frac{r_{x,y} - r_k}{c}\right)\right). \tag{10}$$

Here, $r_{x,y} = \sqrt{(x - x_{c,k})^2 + (y - y_{c,k})^2}$ is the distance from the cluster center, ν a normalization constant to ensure $\int_{\mathcal{R}} f_{\text{EM,coord}}(x, y) = 1$, and $\mathcal{R}$ is the set of all possible $r_{x,y}$. For constant $c \to 0$, Eq. (10) converges to the uniform PDF and clustering is based only on closeness in (x, y)-coordinates. For large values of c, the PDF becomes uniform over the entire spatial domain and the p-values dominate the clustering. In the numerical experiments, we select a sharp but continuous PDF by letting $c = 0.01$. Eq. (4) is approximated with Eq. (10) by

$$f_{\text{EM}}\left(\mathbf{s}_n|\mathcal{C}_k\right) = f_{\text{EM, coord}}(x, y|\mathcal{C}_k) f_{p\text{-val}}\left(p|\mathcal{C}_k\right). \tag{11}$$

The objective is to maximize the network likelihood, that is,

$$\mathcal{L}_{\text{EM}} \triangleq \mathcal{L}_{\text{EM}}\left(\boldsymbol{\tau}_l, \boldsymbol{\Theta}_l\right) = \sum_{n=1}^{N} \log \sum_{k=0}^{l} \frac{N_k}{N} f_{\text{EM}}\left(\mathbf{s}_n|\mathcal{C}_k\right). \tag{12}$$

At the ith iteration, the E-step provides the cluster association probabilities as

$$\hat{v}_{n,k}^{(i)} = \frac{\hat{N}_k^{(i-1)} f_{\text{EM}}(\mathbf{s}_n|\hat{\mathcal{C}}_k^{(i-1)})}{\sum_{j=0}^{l} \hat{N}_j^{(i-1)} f_{\text{EM}}(\mathbf{s}_n|\hat{\mathcal{C}}_j^{(i-1)})}, \tag{13}$$

with $\hat{\mathcal{C}}_k^{(i-1)}, k = 1, \ldots, l$ denoting the clusters with $\hat{N}_k^{(i-1)}$ elements defined by parameter estimates $\hat{\boldsymbol{\theta}}_k^{(i-1)}$. The M-step provides updates $\hat{\boldsymbol{\theta}}_k^{(i)} = [\hat{a}_k^{(i)}, \hat{\boldsymbol{\zeta}}_k^{(i)}]^\top$, maximizing the LLF. $\mathcal{C}_0$ is completely specified by the geometrical alternative cluster features $\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_K$.

The shape parameters a_k of the beta distribution and the cluster centers are determined using MLEs

$$\hat{a}_k^{(i)} = -\frac{\sum_{n=1}^{N} \hat{v}_{n,k}^{(i)}}{\sum_{n=1}^{N} \hat{v}_{n,k}^{(i)} \log p_n} \tag{14}$$

$$\hat{x}_{\text{c},k}^{(i)} = \frac{\sum_{n=1}^{N} \hat{v}_{n,k}^{(i)} x_n}{\sum_{n=1}^{N} \hat{v}_{n,k}^{(i)}} \qquad \hat{y}_{\text{c},k}^{(i)} = \frac{\sum_{n=1}^{N} \hat{v}_{n,k}^{(i)} y_n}{\sum_{n=1}^{N} \hat{v}_{n,k}^{(i)}}. \tag{15}$$

However, Eq. (10) does not allow the computation of an MLE for the coordinate cluster boundary r_k since its derivative is never equal to zero. An estimate of r_k is found by

$$\hat{r}_k^{(i)} = \sqrt{2\frac{\sum_{n=1}^{N} \hat{v}_{n,k}^{(i)} \left(\left(x_n - \hat{x}_{\text{c},k}^{(i)}\right)^2 + \left(y_n - \hat{y}_{\text{c},k}^{(i)}\right)^2\right)}{\sum_{n=1}^{N} \hat{v}_{k,n}^{(i)}}}. \tag{16}$$

Because (16) is not an MLE, a modification of the updating scheme is required to guarantee that the parameter update $\hat{r}_k^{(i)}, k = 1, \ldots, l$ does not decrease the overall network LLF, see [9]. In particular, after updating $\hat{N}_k^{(i)}$, and $\hat{\boldsymbol{\zeta}}_k^{(i)}$, we define the spatially updated parameter matrix $\hat{\boldsymbol{\Theta}}_{l,\text{su}}^{(i)}$ whose kth column is $\hat{\boldsymbol{\theta}}_{k,\text{su}} = [\hat{a}_k^{(i-1)}, \hat{\boldsymbol{\zeta}}_k^{(i)}]^\top$, and resulting LLF $\mathcal{L}_{\text{EM,su}}^{(i)} \triangleq \mathcal{L}_{\text{EM}}(\hat{\boldsymbol{\tau}}_l^{(i)}, \hat{\boldsymbol{\Theta}}_{l,\text{su}}^{(i)})$. If we increase the LLF by the spatial update, we keep these updated parameters. Otherwise, we decide for convergence of LLF in spatial domain and update the shape parameters of the beta distribution. The complete Spatial Inference based on Clustering of p-values (SPACE-COP) algorithm, including initialization, model order selection, and post processing is summarized in Alg. 1.

IV. Numerical Experiments

We present Monte Carlo (MC) simulation results for three different scenarios: two setups are composed of different numbers of randomly generated true clusters of circular geometry in the spatial domain (Sc. 1 and Sc. 2) and one simulation with fixed clusters of differing geometries and topology associated with alternative hypotheses (Sc. 3), to illustrate the capability of SPACE-COP to identify clusters and sub-regions of arbitrary shape. We assume the sensors to be uniformly distributed across the normalized map, i.e., we do not make assumptions on particularly beneficial sensor locations. Clusters are non-overlapping and lie fully within the defined spatial domain of interest. We use 100 MC runs, a network comprising $N = 10,000$ uniformly distributed sensors transmitting p-values and positions to the FC. The initialization shape parameter is $a_{\text{init}} = 0.0005$, and the shape parameter of the beta distribution associated with the p-values under the alternative hypotheses is set to $a_{\text{sim,max}} = 0.35$, which provides at least 35% of p-values with $p < 0.05$. Thus, when conducting single sensor (SS) thresholding at level $\alpha_{\text{ss}} = 0.05$, a vast majority of the sensors within the area where an alternative hypothesis holds accepts H_0.

We consider four quantitative performance measures: The overall probability of sensors correctly assigned to regions of null and alternatives p_{ca}, the probabilities of missed detection p_{md} and false alarm p_{fa} (false discoveries) and the false discovery proportion (FDP), whose expectation is the FDR.

Algorithm 1 SPACE-COP

```
Define $a_{\text{init}}$ and obtain $\mathcal{K} = \{\mathbf{s}_n \in \mathcal{S} | \mathcal{B}(p_n; a_{\text{init}}) > 1\}$
for $l = 1 : L_{\max}$ do
    Apply $k$-means on $\mathcal{K}$ to find $l$ cluster centers $(\hat{x}_k^{(0)}, \hat{y}_k^{(0)})$ associated with alt. hypotheses and clusters $\mathcal{C}_k^{(0)}$ for $k = 1, \dots, l$ based on $(\tilde{x}_n, \tilde{y}_n)$ of sensors in $\mathcal{K}$
    Initialize $\hat{r}_k^{(0)} = \sqrt{2 \cdot \text{med}((\tilde{x}_n - \hat{x}_k^{(0)})^2 + (\tilde{y}_n - \hat{y}_k^{(0)})^2)}$
    Set $\hat{N}_0^{(0)} = N(1 - \sum_{k=1}^{l} (\hat{r}_k^{(0)})^2 \pi)$, $\hat{N}_k^{(0)} = N(\hat{r}_k^{(0)})^2 \pi$
    Set $\hat{a}_k^{(0)} = 0.1$, $k = 1, \dots, l$ and $i = 0$
    procedure EXPECTATION-MAXIMIZATION
        while $(\mathcal{L}_{\text{EM}}^{(i)} - \mathcal{L}_{\text{EM}}^{(i-1)} > \delta) \vee (\mathcal{L}_{\text{EM}}^{(i-1)} - \mathcal{L}_{\text{EM}}^{(i-2)} > \delta)$ do
            Increment $i = i + 1$
            Compute $\hat{v}_{n,k}^{(i)}$ from Eq. (13)
            Update $\hat{N}_k^{(i)} = \sum_{n=1}^{N} \hat{v}_{n,k}^{(i)}$
            Update $\hat{\boldsymbol{\zeta}}_k^{(i)}$ from (15), (16)
            if $\mathcal{L}_{\text{EM,su}}^{(i)} - \mathcal{L}_{\text{EM}}^{(i-1)} < \delta$ then
                $\hat{\mathbf{a}}^{(i)} = \hat{\mathbf{a}}^{(i-1)}$
            else
                Set $\hat{\boldsymbol{\zeta}}_k^{(i)} = \hat{\boldsymbol{\zeta}}_k^{(i-1)}$ and update $\hat{\mathbf{a}}^{(i)}$ from (14)
            end if
        end while
    end procedure
    Hard decision about cluster memberships by $\hat{\mathcal{C}}_k = \left\{ \mathbf{s}_n \in \mathcal{S} | k = \arg\max \hat{v}_{n,k}^{(i-1)} \right\}$, $k = 0, \dots, l$
    Compute BIC$(M_l)$ according to (6)
end for
Model Order Selection by $\hat{K} = \arg\max_l \text{BIC}(M_l)$
Eliminate clusters with $\hat{a}_k^{(i-1)} > 0.8$
Merge clusters that overlap in $(x, y)$ domain
```

We aim to maximize p_{ca} while keeping the other criteria low. Consider the trade-off between p_{fa} and p_{md}: A procedure that requires only little alternative evidence to reject H_0 generally provides low p_{md} but large p_{fa} and vice-versa. As FDR control procedures such as BH control the proportion of false positives among all positives and are less conservative in terms of Type I error level control, they are favorable in multiple hypothesis problems [4]. The Bayesian FDR (BFDR) [10] approximates the FDR when used with appropriate hypothesis probabilities for each sensor. An upcoming journal paper will introduce a BFDR-control mechanism by rank-ordering and thresholding the converged zero-cluster association probabilities that SPACE-COP provides similar to [6].

We compare the performance of our suggested method to a classic SS Neyman-Pearson detector with a predefined level α_{ss}, the BH procedure [4] that controls the FDR below a given threshold value q_{BH} and the two-step cluster testing and trimming (BTS) procedure [5] that controls the FDR asymptotically at q_{BTS}. For simulation, the parameters of the different procedures are chosen such that they provide performance measures at a comparable scale. The first two competitors consider individual sensors only, hence, they lack power $p_{\text{D}} = 1 - p_{\text{md}}$ as compared to the clustering approaches. BTS controls the FDR using clusters defined beforehand. We simply divide the grid into a uniform constellation of M_{BTS} squared clusters to compare it to SPACE-COP. We use $M_{\text{BTS}} = 1024$, the value providing the best results for the considered scenarios. The results are displayed in Table I.

TABLE I: Simulation Results (in %). Our method associates most sensors correctly with true hypotheses.

	Measure	SPACE-COP	BTS	BH	SS
Sc. 1	p_{ca}	97.96	95.51	93.18	93.1
	p_{md}	5.98	33.46	55.78	48.91
	p_{fa}	1.53	0.83	1.5	0.56
	FDP	10.07	8.15	8.95	19.41
Sc. 2	p_{ca}	98.63	92.98	85.41	85.54
	p_{md}	3.19	17.49	47.91	48.28
	p_{fa}	0.68	2.97	1.57	1.51
	FDP	1.71	8.16	7.29	7.44
Sc. 3	p_{ca}	93.61	92.07	86.91	86.7
	p_{md}	9.47	21.31	44.89	37.95
	p_{fa}	5.34	3.43	2.34	4.97
	FDP	14.42	10.82	11.19	19.86

A. Scenario 1

We first consider a setting with $K_1 = 7$ small regions where alternative hypotheses are in place and define $L_{\max} = 15$. On average over all MC runs, 11.37% of network sensors were observing one of the alternative hypotheses. BH (with $q_{\text{BH}} = 10\%$) and SS ($\alpha_{\text{ss}} = 1.5\%$) both assign approximately 93% of sensors correctly and both significantly lack detection power since sensors are correctly associated with alternative hypotheses in only about 50% of cases. As stated earlier, the power increases significantly when considering multiple sensors in the vicinity. When applying BTS with $q_1 = 0.05$ and $q_{\text{BTS}} = 0.25$, as suggested in [5], we observe a significant improvement in detection power $p_{\text{D}} = 1 - p_{\text{md}}$ and a resulting increase in p_{ca}. Using our spatial model, we approximately double the gain in p_{ca} from BTS compared to the single sensor schemes and reduce p_{md} from 33% to 6%, whilst keeping p_{fa} and FDP at tolerable levels.

B. Scenario 2

The second setting consists of $K_2 = 3$ larger regions where alternative hypotheses hold with an additional constraint to the previous scenario: The alternative clusters cover at least 25% of the total network coverage area. This resulted in on average over all MC runs 27.87% true alternative sensors.

Compared to Sc. 1, SPACE-COP has further increased its performance gain over the competitors. This was expected, since the geometrical structure of the problem is more pronounced due to the larger size of the clusters and our algorithm allows to adapt the cluster size to the recorded data.

C. Scenario 3

Finally, we illustrate that our algorithm produces reliable results also for a setting in which the initial assumption on cluster shapes is clearly violated for the regions associated with alternative hypotheses. For illustration purposes, we consider a setting where the alternative hypotheses hold in a non-convex half moon shaped region, a torus and a triangular region, as depicted in Fig. 2. The observed phenomena along their parameters are still generated independently for each run.

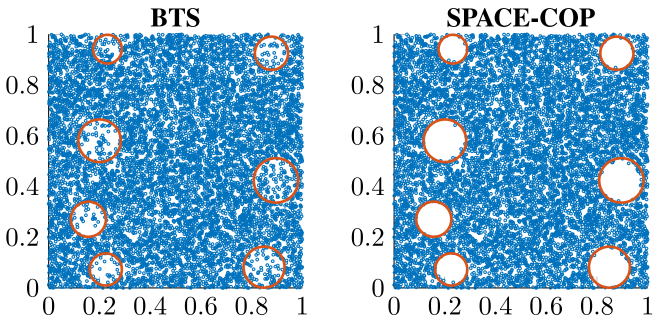

Fig. 1: Exemplary run (best competitor vs. proposed), with sensors detected to belong to the H_0 area as blue dots and true alternative areas as red circles as according to Sc. 1.

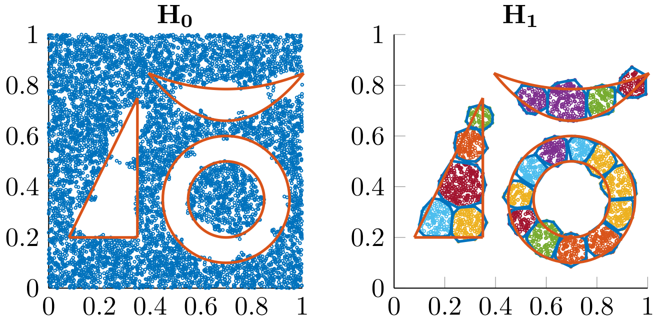

Fig. 2: Exemplary run using SPACE-COP on Sc. 3, red boundaries indicating ground truth, small blue dots being H_0 on the left; small circles on the right being detected alternative hypotheses sensors and blue inferred cluster hulls.

We use an increased candidate model order $L_{\max} = 20$. In Fig. 2, we show the clustering results straight after application of the BIC, before the merging of overlapping clusters. SPACE-COP cuts the alternative regions into smaller circles that usually overlap and are hence merged after the application of the BIC. The final resulting clusters may have an arbitrary shape due to the trade-off between spatial coordinates and observed p-values in the objective of the EM-algorithm: If a sensor lies outside the estimated circular area of a cluster but still has similar decision statistics (small p-value) to the cluster members, the sensor will be associated with the cluster where the alternative hypothesis is in place. Consequently, the identified clusters may take an arbitrary shape that differs from the initially assumed circular shape.

The results in Table I demonstrate that, overall, we still outperform the competitors, despite a reduced gain w.r.t to BTS. During simulation, the estimated model order by the BIC $\hat{K}$ was either $L_{\max}$, or at very close to it. Hence, a further increase in allowed candidate model order should also increase the performance gain, because smaller circles will be able to approximate the shape of the alternative clusters even better. The threshold for the SS was set to $\alpha_{\text{ss}} = 0.05$ in this experiment, to adjust it to the resulting p_{fa} of its competitors.

V. Conclusion

We have derived a novel algorithm applicable to large-scale sensor networks to perform statistical inference in a distributed manner and identify homogeneous regions in an observed phenomenon or field where the null hypothesis does not hold. The algorithm finds clusters by considering both observed p-values at each sensor and proximity among the sensors. The approach uses distribution of p-values under null and alternative hypotheses, Expectation-Maximization and BIC to associate sensors with clusters. Simulation results demonstrate its validity also for cases in which the assumption on underlying shape of alternative areas was clearly violated and true alternative areas followed arbitrary shapes.

Appendix

$$\begin{aligned}\hat{\mathbf{J}}_{k,[2,1]} = \hat{\mathbf{J}}_{k,[1,2]} &= E\left[\frac{\partial \log \mathcal{L}\left(\boldsymbol{\theta}_k|\mathcal{C}_k\right)}{\partial a_k}\frac{\partial \log \mathcal{L}\left(\boldsymbol{\theta}_k|\mathcal{C}_k\right)}{\partial r_k}\right]\\ &= E\left[\left(\frac{N_k}{a_k} + \sum_{m=1}^{N_k}\log\left(p_m\right)\right)\left(-\frac{2N_k}{r_k}\right)\right]\\ &= -\frac{2N_k^2}{r_k}\left(\frac{1}{a_k} + \Psi\left(a_k\right) - \Psi\left(a_k+1\right)\right) \qquad (17)\end{aligned}$$

$$\begin{aligned}\hat{\mathbf{J}}_{k,[1,1]} &= E\left[\frac{\partial \log \mathcal{L}\left(\boldsymbol{\theta}_k|\mathcal{C}_k\right)}{\partial a_k}\frac{\partial \log \mathcal{L}\left(\boldsymbol{\theta}_k|\mathcal{C}_k\right)}{\partial a_k}\right] \qquad (18)\\ &= \frac{N_k^2}{a_k}\left(\frac{1}{a_k} + 2\left(\Psi\left(a_k\right) - \Psi\left(a_k+1\right)\right)\right) + N_k^2\left(\Psi(a_k)\right.\\ &\quad \left. - \Psi\left(a_k+1\right)\right)^2 + N_k\left(\Psi_1(a_1) - \Psi_1\left(a_k+1\right)\right)\end{aligned}$$

$$\hat{\mathbf{J}}_{k,[2,2]} = E\left[\frac{\partial \log \mathcal{L}(\boldsymbol{\theta}_k|\mathcal{C}_k)}{\partial r_k}\frac{\partial \log \mathcal{L}(\boldsymbol{\theta}_k|\mathcal{C}_k)}{\partial r_k}\right] = \frac{4N_k^2}{r_k^2} \qquad (19)$$

$$\left|\hat{\mathbf{J}}_k\right| = \frac{4N_k^3}{r_k^2}\left(\Psi_1\left(a_k\right) - \Psi_1\left(a_k+1\right)\right) \qquad (20)$$

- Ψ, Ψ_1 The digamma and trigramma functions

References

[1] E. Arias-de-Reyna *et al.*, “Crowd-based learning of spatial fields for the internet of things: From harvesting of data to inference,” *IEEE Signal Process. Mag.*, vol. 35, no. 5, pp. 130–139, Sep. 2018.

[2] F. Teklehaymanot *et al.*, “Bayesian cluster enumeration criterion for unsupervised learning,” *IEEE Trans. Signal Process.*, vol. 66, pp. 5392 – 5406, 10 2018.

[3] B. Efron, *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, ser. Institute of Mathematical Statistics Monographs. Cambridge University Press, 2010.

[4] Y. Benjamini and Y. Hochberg, “Controlling the false discovery rate: A practical and powerful approach to multiple testing,” *J. Royal Stat. Soc., Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[5] Y. Benjamini and R. Heller, “False discovery rates for spatial signals,” *J. Am. Stat. Assoc.*, vol. 102, no. 480, pp. 1272–1281, 2007.

[6] T. Halme *et al.*, “Bayesian multiple hypothesis testing for distributed detection in sensor networks,” in *2019 IEEE Data Sci. Workshop*, June 2019.

[7] G. Casella and R. Berger, *Statistical Inference*. Brooks/Cole Publishing Company, 1990.

[8] S. Pounds and S. Morris, “Estimating the occurence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values,” *Bioinformatics*, vol. 19, pp. 1236 – 1242, 01 2008.

[9] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006.

[10] B. Efron *et al.*, “Empirical bayes analysis of a microarray experiment,” *J. Am. Stat. Assoc.*, vol. 96, no. 456, pp. 1151–1160, 2001.