IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 12 December 2022, accepted 19 December 2022, date of publication 23 December 2022,
date of current version 2 January 2023.

Digital Object Identifier 10.1109/ACCESS.2022.3231909

RESEARCH ARTICLE

A Unified Bayesian Framework for Joint Estimation and Anomaly Detection in Environmental Sensor Networks

ALESSIO FASCISTA[1], **(Member, IEEE), ANGELO COLUCCIA**[1], **(Senior Member, IEEE), AND CHIARA RAVAZZI**[2], **(Member, IEEE)**

[1]Department of Innovation Engineering, University of Salento, 73100 Lecce, Italy
[2]Institute of Electronics, Computers and Telecommunication Engineering (IEIIT), National Research Council (CNR), 10129 Turin, Italy

Corresponding author: Alessio Fascista (alessio.fascista@unisalento.it)

ABSTRACT Advanced large-scale environmental monitoring systems relying on the emerging aerial/terrestrial technologies of wireless sensor networks (WSNs), unmanned aerial vehicles (UAVs), and mobile crowdsensing, impose strong requirements on the reliability of the collected data. Unfortunately, sensing units can suddenly suffer unexpected anomalies due to accidental faults or malicious causes. Outlier detection methods have been widely employed to identify and discard unreliable measurements from large data sets, but further improvements in the sensing processes can be obtained by adopting advanced signal processing algorithms that take full advantage of all the collected information without rejecting the measurements. In this paper, we propose a novel unified Bayesian framework that enable simultaneous estimation of a common parameter of interest and identification of multiple and possibly different types of anomalies that can affect sensors in environmental sensor networks. Specifically, we consider two rather general error models based on Gaussian mixtures able to capture different variations affecting the quality of the collected measurements. For each model, we illustrate the optimal joint maximum-likelihood and maximum a-posteriori (ML-MAP) estimation method, which represents the benchmark for the problem at hand, and propose novel reduced-complexity two-step algorithms able to achieve almost the same performance of the joint ML-MAP, but at a fraction of its computational cost. The derivations of all the algorithms are also extended to handle the more general case in which the probability of occurrence of anomalies is unknown and should be inferred from the data using an Empirical Bayes approach. Extensive performance analyses using both synthetic and real experimental data acquired in a network of environmental monitoring stations deployed in the Apulia region, south of Italy, demonstrate the effectiveness of the proposed framework.

INDEX TERMS Anomaly detection, Gaussian mixture models, maximum-likelihood and Bayesian estimation, outlier detection, sensor networks.

I. INTRODUCTION

Enabling large-scale, continuous, and pervasive monitoring of the environment in all its different dimensions (air, land, and water) requires that a number of fundamental parameters such as temperature, humidity, pressure, water turbidity, PMx and VOCs concentrations are accurately estimated and analyzed over time [1], [2], [3], [4]. To complement the limited capabilities of traditional monitoring systems [5], [6], the emerging technologies of wireless sensor networks (WSNs) [7], unmanned aerial vehicles (UAVs) [8], and mobile crowdsensing [9] can be exploited to achieve a *pervasive* and *fine-grained* monitoring, by means of a larger number of low-cost sensing units. In fact, by providing a more capillary coverage of the target areas and increased sensing rates, such technologies are able to correctly capture

The associate editor coordinating the review of this manuscript and approving it for publication was Mauro Gaggero.

the spatio-temporal variations of the physical phenomena of interest even over very small scales, so acting as enablers of integrated and large-scale environmental monitoring [10].

Given the expected increased complexity of next-generation monitoring systems, guaranteeing a strong reliability of measurements collected by the heterogeneous sensor nodes is of utmost importance. In particular, fixed monitoring stations or WSN nodes can suddenly experience failures that may severely compromise their ability to provide accurate measurements, especially when they operate in hostile environments (e.g., in presence of high-temperatures or wildfires, under flooding conditions, etc.) [11]. UAV platforms typically experience significant geometric and spectro/radiometric limitations, which result in frequent miscalibrations of the onboard sensors [12]. On the other hand, crowdsensing nodes (such as smartphones, smartwatches, common transportation systems) represent a valuable source of additional environmental perception, but at the same time pose severe risks related to the possibility that users contribute with unreliable data and potentially jeopardize the sensing campaign. Generally, two main possible cases are distinguished: in a first case, similarly to WSNs and UAVs, data unreliability is mainly due to faults/defects or miscalibrations in the users devices, which unintentionally provide corrupted data. In other cases, malicious users may contribute with fake sensing data (e.g., fake GPS readings, alterated measurements) just to earn the rewards associated to the crowdsensing tasks, affecting in turn the integrity of the data collected by the monitoring system [13], [14].

To handle the different nature of anomalies that can impair the sensing capabilities of sensor nodes involved in next-generation monitoring systems, advanced signal processing algorithms need to be conceived. The main goal consists in estimating the relevant parameters of interest, while at the same time identifying the possible presence of different types of faults (either accidental or intentional) and weighting the corresponding contributions accordingly.

A. RELATED WORK

In estimation theory, whenever measurements are used to estimate a quantity of interest, measurement errors must be adequately accounted for, and the statistical properties of these errors identified to enable robust estimation by either *discarding* the unreliable measurements or by *weighting* them differently in the estimation algorithms. In the specific scenario we are considering, we need to process data acquired from different monitoring systems and sensors and the noise affecting the measurements can be drastically heterogeneous and, more importantly, its distribution may be unknown. In addition, as is the case with most large data sets, some data may be farther from the sample mean than it might be expected. Outliers in the data can be due to various reasons: a) they could simply be due to chance, i.e., some measurements produced data that are far from the mean values; or b) systematic errors occurred in the data collection.

1) OVERVIEW ON OUTLIER/ANOMALY DETECTION

The general literature on outlier/anomaly detection is quite vast and contains several different algorithms whose main objective is to identify and discard spurious measurements. One of the most popular techniques is the Random sample and consensus, also known as RANSAC [15]. This is an iterative algorithm able to estimate parameters of a mathematical model with a certain fidelity degree also when the number of outliers is significant. The tricky point of this method is the choice of the number of iterations: if the number of iterations is limited, there are no guarantees on the optimality of the solution, as the latter can depend on the specific considered scenario such as type of sensor and number of parameters to estimate in the model. Methods mainly based on non-convex optimization that alternate consensus steps with minimization of convex norms of the residuals have been also widely considered. We refer the interested reader to [16] for more details on robust estimation. These methods exhibit local convergence, but the robustness of the solution depends on the quality of the initial guess. Moreover, their computational cost often becomes unaffordable as the number of data increases.

More advanced solutions for outliers' identification and mitigation have been devised in literature, as for instance those based on *clustering* techniques. One of the most popular clustering algorithm used for outlier detection is the density-based clustering non-parametric algorithm, also known as DBSCAN [17]. The algorithm starts with a given set of points in some space and groups points with many nearby neighbors, and reject points that lie in low-density regions. Another interesting family of techniques consists in a modern approach to outlier rejection based on linear programs. The idea is to avoid hard binary classifications ("outlier" or "inlier"), and look for the largest set of measurements that are internally "coherent". The problem, posed as a linear program, can be solved via convex relaxation. The simulations provided in [18] in a different context show promising results in selecting good measurements, allowing to obtain a global solution by not relying on the availability of an initial guess.

Anomaly/outlier detection methods based on the outlier *probability* have been also considered in literature to improve the process of excluding unreliable measurements from the subsequent elaboration steps. These techniques have been successfully applied, for instance, for online distributed structural identification following a hierarchical approach in [19]. The same idea has been then used to enhance the stability of dynamic filtering approaches such as the extended Kalman Filter (EKF), when applied for online structural monitoring and damage detection [20], [21]. To solve the instability problems of traditional EKF implementations, these algorithms carefully assign the noise covariance matrices at each filter update step by using real-time estimates of the noise parameters, followed by suitable mechanisms to remove abnormal measurements. Very similar principles have been adopted to deal with the presence of outliers in marine robotics, mainly for navigation and model identification tasks. More specifically, robust state estimation algorithms based on modified versions of the KF, Luenberger observer, and Rauch-Tung-Striebel smoother have been proposed and validated under different operational scenarios [22], [23].

2) OVERVIEW ON JOINT ESTIMATION AND ANOMALY DETECTION

Although outlier/anomaly detection methods as those discussed above provide satisfactory performance in a number of different application scenarios, by discarding the unreliable measurements they do not take full advantage of all the information available in the collected data. To fill this gap, techniques that simultaneously identify anomalies and estimate the parameters of interest without rejecting the measurements have been conceived, known as *joint estimation and anomaly detection*. Far less works fall in this category compared to the huge literature on outlier/anomaly detection, despite the existing references point out very promising performance improvements. For instance, in [24] the joint problem of hypothesis testing and parameter estimation, which typically arises in radar and cognitive radio contexts, is addressed. Using a Bayesian estimation cost function that depends on both the detection result and the estimation scheme, a novel optimal joint detector and estimator is devised that, taking into account the coupling nature of the two subproblems, achieves superior performance compared to methods that treat detection and estimation separately. A similar problem is also considered in [25], where the additional presence of uncertainties in the guessed prior probability is explicitly taken into account at the algorithms design stage. Joint reconstruction and anomaly detection methods have been also successfully applied in remote sensing applications to reconstruct hyperspectral images from compressed observations [26]. Interestingly, experimental analyses conducted on real data confirm that joint approaches provide superior performance compared to standalone reconstruction and anomaly detection algorithms.

Focusing on the specific topic of monitoring using sensor networks, joint estimation and detection algorithms have been developed using mixtures probability models, such as Laplacian and Gaussian mixtures models. Under this framework, the existing approaches have been designed mainly for distributed contexts where estimation of common parameters and classification of sensor states are collaboratively performed by all the nodes in the network. Such methods consider distributed versions of joint maximum likelihood (ML) and maximum a-posteriori (MAP) estimators, and include additional communication constraints imposed by the network. The interested reader is referred to [27] and [28] for more details on the theory behind joint ML-MAP estimation, which is completely general and can be applied to a numerous engineering contexts involving joint decision and estimation processes, where qualities of decision and estimation affect each other (e.g., target detection and tracking).

The constrained maximisation of the log-likelihood function for a mixture model, due to its combinatorial nature, is an NP-hard problem and there is no closed form solution for the model parameters. In order to overcome this issue, in [29], [30], [31], and [32] different iterative techniques are proposed that try to approximate the optimal (centralized) solution of the joint ML-MAP estimator. A common point of the strategies is to consider the complete log-likelihood function based on the missing data. After choosing some initial values for the mixture parameters, the following updates are alternated: in the first step, current values are used for the parameters to estimate the signal and to evaluate the posterior distribution type of measurement (inlier or outlier); in the second step these probabilities are used to re-estimate the mixture parameters. Besides the design of the algorithms, the contribution in [29] and [32] includes rigorous proofs of convergence that make the distributed techniques well-suited to work when nodes can cooperate among each other and the communication to a central processing unit is not allowed.

B. MAIN CONTRIBUTION AND OUTLINE OF THE PAPER

From the analysis of the above literature, it emerges that there is a lack of a general framework able to properly handle the joint estimation of common parameters of interest and identification of *multiple* and possibly *different types* of sensor anomalies in environmental sensor networks. More specifically, [29] mainly deals with anomalies that introduce only a deterministic bias in the measurement errors, modeled using a Gaussian mixture model. For this model, an iterative, distributed, consensus-like algorithm based on ML estimation is proposed, which tries to approximate the optimal centralized ML. Authors in [30] extend the previous model to the case in which the additive bias can take on two different values based on a (possibly unknown) Bayesian prior. The key contribution is a distributed estimator based on gossip-like communications that approximates the optimal centralized joint ML-MAP. On the other hand, in [31] sensor anomalies that impact only on the variance of the measurement errors are considered, assuming that the probability of occurrence of an anomaly (i.e., the hyperparameter of the Bayesian prior) is perfectly known a priori. The main goal is the development of a distributed, iterative procedure which copes with the communication constraints imposed by the network, and tries to approximate the optimal solution of the joint ML-MAP using input driven consensus algorithms. Considering the same model for the anomalies, [32] deals with the problem of joint estimation and anomaly detection starting from relative measurements (expressed as difference between two individual measurements) given as inputs. The optimal centralized joint ML-MAP estimator is illustrated and an approximated distributed version with some convergence guarantees is also provided.

In this work, we take a different path and investigate the applicability of techniques based on Gaussian mixtures probability models under the framework of a (possibly large-scale) environmental monitoring system where sensing nodes may not communicate and/or interact among each other (as is the case of crowdsensing nodes, but also of WSNs and UAVs nodes when energy-saving is a priority) and all the collected measurements need to be processed at a centralized unit. Compared to the most closely related works discussed above [29], [30], [31], and [32], we aim at advancing the literature by proposing a novel theoretical framework that simultaneously handles anomalies of different nature affecting either the mean or the standard deviation of the measurement errors, and additionally deals with the most general cases in which the hyperparameter of the Bayesian prior describing the probability of occurrence of anomalies is completely unknown. More precisely, the following contributions are provided:

- a unified Bayesian framework for joint estimation of common parameters and detection of sensors anomalies that considers two rather general error models based on Gaussian mixtures able to capture variations affecting the quality of the measurements due to different anomalous operational conditions (accidental or intentional);
- for both the considered error models, we start by illustrating the optimal joint ML-MAP estimators already discussed in [29] and [30] for the case of anomalies introducing an additive bias and in [31] for the case of anomalies affecting the variance of the errors, which represent the *benchmark* for the specific problem at hand. Then, we propose novel reduced-complexity two-step methods able to achieve almost the same performance of the joint ML-MAP but at a fraction of its computational cost, as demonstrated by means of a theoretical cost analysis. We also extend the derivations of all the approaches to the case in which the hyperparameter of the Gaussian mixture is unknown and should be inferred from the data using an Empirical Bayes approach;
- an exhaustive performance analysis is conducted to test the algorithms effectiveness in terms of performance, robustness to potential mismatches and to increasing percentages of anomalous nodes in the network, scalability, and computational cost. The assessment is conducted under different sensors anomalous conditions, both on simulated data as well as on experimental data collected by a monitoring network deployed in the Apulia region, south of Italy.

The paper is organized as follows. In Sec. II we formally present general observations models for acquisition of a common physical quantity in an environmental monitoring network and we distinguish two separate cases for modeling anomalies (multiplicative and additive models). Sec. III is devoted to the algorithms designed for the multiplicative error model, whereas Sec. IV presents the algorithms based on the additive error model. In Sec. V we conduct a theoretical cost analysis to discuss the complexity of the considered algorithms. Section VI contains an extensive performance assessment both on simulated and real scenarios. Finally, summary and some concluding remarks completes the paper in Section VII.

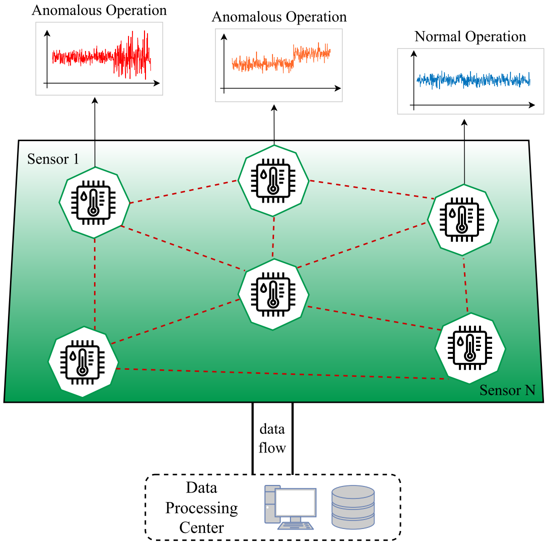


FIGURE 1. General scenario of an environmental monitoring network including sensors with different anomalous operational conditions.

II. GENERAL SCENARIO AND OBSERVATION MODELS

We consider an environmental monitoring network (EMN) comprising N sensing devices (e.g., fixed monitoring stations, WSN nodes, UAV platforms, crowdsensing nodes), available at different locations in a surveyed area and each measuring an unknown global environmental parameter θ. Individual sensors make local, noisy measurements of the phenomenon of interest, as shown in Fig. 1. Accordingly, each measurement y_i carried out by the i-th sensor is expressed as

$$y_i = \theta + x_i, \qquad i = 1, \ldots, N \tag{1}$$

where x_i denotes the local measurement error, modeled as a Gaussian random variable with mean a_i and standard deviation b_i, namely $x_i \sim \mathcal{N}(a_i, b_i^2)$. In this respect, the parameter θ can be interpreted as a global value capturing the average magnitude of a given phenomenon of interest (e.g., temperature, pressure, pH, PMx concentration, ...). Small local deviations from such a mean value will be treated as fluctuations and thus included in the additive noise term x_i.

In this work, we propose a unified estimation and anomaly detection framework that includes two different and rather general error models, able to capture possible variations (associated to a sensor abnormal condition) either in the local parameter a_i or b_i. More specifically, in the first model the mean of the error term x_i, namely a_i, is zero while the standard deviation can take on two possible values, i.e., $b_i \in \{\alpha, \beta\}$, with $\beta > \alpha$ (with α typically small, related to the sensor precision). The lower value of the standard deviation α represents the normal operational behavior of the sensor, whereas β is indicative of an anomalous condition, e.g. a misfunctioning or degradation of the sensor, or a deliberate attack. On the other hand, in the second model the mean of the measurement error can take on two possible values, $a_i \in \{\gamma, \nu\}$ with $\nu > \gamma$

(with γ typically zero in sensors with no bias), while the standard deviation b_i is fixed to σ, the latter representing the intrinsic sensor precision. Accordingly, the additive error terms x_i, $i = 1, \ldots, N$ can be expressed as

$$
\begin{aligned}
x_i^{\text{MUL}} &= b_i n_i, b_i \in \{\alpha, \beta\} \\
x_i^{\text{ADD}} &= a_i + \sigma n_i, a_i \in \{\gamma, \nu\}
\end{aligned}
\tag{2}
$$

with $n_i \sim \mathcal{N}(0,1)$. From (2), it can be easily observed that the term b_i acts as a multiplicative error factor, whereas a_i introduces an additive shift in the measurement error. Given their different nature, such models will be denoted as the *multiplicative* and *additive* error model, respectively.

To account for the inherent randomness of failures that can occur on sensors, we assume that a_i in the additive model and b_i in the multiplicative model follow a discrete probability distribution $\mathcal{B}(q, s, p)$ such that, denoting by χ a random variable distributed according to $\mathcal{B}(q, s, p)$, we have that

$$
\chi = \begin{cases} q \text{ with probability } 1-p \\ s \text{ with probability } p \end{cases} . \tag{3}
$$

The above probability distribution represents a *prior information* that will be exploited within a Bayesian framework to perform estimation and anomalous sensors detection tasks. Accordingly, the two models can be summarized as

$$
\begin{array}{lll}
\text{Multiplicative Model:} & a_i = 0 & b_i \sim \mathcal{B}(\alpha, \beta, p) \\
\text{Additive Model:} & a_i \sim \mathcal{B}(\gamma, \nu, p) & b_i = \sigma
\end{array}
$$

The considered models are very versatile and can represent measurement errors due to sensor failures (e.g., systematic or stochastic errors) as well as cyber attacks (e.g., byzantine attack) or malicious alterations of data. As in typical environmental monitoring systems, each sensor sends the measurements to a fusion center as soon as they are ready. In this work, the fusion center has a twofold goal: recovering the value of the global parameter θ from sensor's measurement streams while detecting, at the same time, the possible presence of faulty sensors. In the following, we separately deal with the two measurement error models. More specifically, we start from the multiplicative model, which has been also considered in a different distributed context in [31] and [32], and illustrate both a joint ML-MAP approach and a novel two-step algorithm able to attain the same estimation and detection performances of the joint ML-MAP, but at a reduced computational cost. We then extend the two approaches to the case in which the hyperparameter p of the prior distribution is unknown and should be inferred from the data using an empirical Bayes approach. Then, we focus on the additive error model and, similarly, illustrate both the joint ML-MAP estimator and a reduced-complexity two-step algorithm, for both the cases of known (fixed as a design parameter) and unknown hyperparameter p.

III. BAYESIAN ALGORITHMS FOR THE MULTIPLICATIVE ERROR MODEL

Starting from the model described in the previous section, each sensor measurement y_i is a Gaussian mixture distributed according to the probability density function (pdf)

$$
\begin{aligned}
f^{\text{MUL}}(y_i) &= (1-p) f^{\text{MUL}}(y_i|\theta, \alpha) + p\, f^{\text{MUL}}(y_i|\theta, \beta) \\
&= \frac{1-p}{\sqrt{2\pi}\alpha} e^{-\frac{(y_i-\theta)^2}{2\alpha^2}} + \frac{p}{\sqrt{2\pi}\beta} e^{-\frac{(y_i-\theta)^2}{2\beta^2}}
\end{aligned}
\tag{4}
$$

being each conditional pdf given by

$$
f^{\text{MUL}}(y_i|\theta, b_i) = \frac{1}{\sqrt{2\pi} b_i} e^{-\frac{(y_i-\theta)^2}{2b_i^2}}, \quad b_i \in \{\alpha, \beta\}. \tag{5}
$$

A. JOINT ML-MAP ESTIMATION

The goal is to estimate the global parameter θ and the specific state of each sensor b_i using a Bayesian inference framework. In [31], the value of the hyperparameter p used to characterize the multiplicative error model is assumed perfectly known a priori. However, in most practical cases the value of p, which determines the probability of having a faulty sensor, is generally unknown. To deal with such a more realistic condition, in our framework we investigate two different scenarios: in the first case, p is considered a design parameter to be tuned according to some a-priori (coarse) information about the quantity of faulty sensors extracted, e.g., from experimental data on the network. In the second case, p is treated as a completely unknown parameter that should be inferred from the data.

1) CASE OF p AS A DESIGN PARAMETER

Let us start with the former case and denote with $f^{\text{MUL}}(\boldsymbol{y}, \boldsymbol{b}|\theta)$ the joint distribution of the sensors measurements $\boldsymbol{y} = [y_1 \cdots y_N]^{\mathsf{T}}$ and of the sensors state $\boldsymbol{b} = [b_1 \cdots b_N]^{\mathsf{T}}$ (interpreted as a density in $\boldsymbol{y}$ and probability in $\boldsymbol{b}$), given the parameter θ and considering a fixed design parameter $p = p_d$. Notice that the latter can be arbitrarily different from the true value of p. The optimal solution for the problem at hand would be to consider a joint ML-MAP approach [27], which consists in maximizing $f^{\text{MUL}}(\boldsymbol{y}, \boldsymbol{b}|\theta)$ with respect to both θ and $\boldsymbol{b}$. More specifically, the whole maximization is partly ML in the deterministic parameter θ, and partly MAP in the discrete random vector $\boldsymbol{b}$. The underlying idea behind this joint estimation procedure consists in finding $\boldsymbol{b}$ which has the maximum posterior joint probability with θ, and the corresponding θ which maximizes that probability. In doing so, the mutual dependency between estimation and decision processes and the way they affect each other are explicitly taken into account. In formulas, the joint ML-MAP estimation problem can be expressed as

$$
(\hat{\theta}^{\text{MUL}}_{\text{JML},p_d}, \hat{\boldsymbol{b}}^{\text{MUL}}_{\text{JML},p_d}) = \arg\max_{\theta, \boldsymbol{b}} \ \ell^{\text{MUL}}(\theta, \boldsymbol{b}) \tag{6}
$$

where $\ell^{\text{MUL}}(\theta, \boldsymbol{b})$ denotes the log-likelihood of $f^{\text{MUL}}(\boldsymbol{y}, \boldsymbol{b}|\theta)$ and

$$
\begin{aligned}
\ell^{\text{MUL}}(\theta, \boldsymbol{b}) \propto \sum_{i \in \mathcal{N}_\alpha} \log \left\{ \frac{(1-p_d)}{\alpha} e^{-\frac{(y_i-\theta)^2}{2\alpha^2}} \right\} \\
+ \sum_{i \in \mathcal{N} \setminus \mathcal{N}_\alpha} \log \left\{ \frac{p_d}{\beta} e^{-\frac{(y_i-\theta)^2}{2\beta^2}} \right\}
\end{aligned}
\tag{7}
$$

with $\mathcal{N} = \{1, 2, \ldots, N\}$ the set containing all the indexes of sensors in the network, while $\mathcal{N}_\alpha = \{i \in \mathcal{N} | b_i = \alpha\}$. It is not difficult to show that the inner maximization wrt $\boldsymbol{b}$ in (6) can be carried out separately for each b_i, $i = 1, \ldots, N$, and is solved by directly comparing the log-likelihood function evaluated in the two possible values $\{\alpha, \beta\}$ each b_i can takes on. After some calculations, we then obtain

$$\hat{b}^{\text{MUL}}_{i,\text{JML}}(\theta) = \begin{cases} \alpha \text{ if } |y_i - \theta| < \delta^{\text{MUL}} \\ \beta \text{ otherwise} \end{cases} \tag{8}$$

where

$$\delta^{\text{MUL}} = \sqrt{2\frac{\log\left(\frac{1-p_d}{p_d}\frac{\beta}{\alpha}\right)}{\alpha^{-2} - \beta^{-2}}}. \tag{9}$$

Accordingly, the final joint ML-MAP solution can be derived by plugging the closed-form estimate of the vector $\hat{\boldsymbol{b}}^{\text{MUL}}_{\text{JML},p_d}(\theta) = [\hat{b}^{\text{MUL}}_{1,\text{JML}}(\theta) \cdots \hat{b}^{\text{MUL}}_{N,\text{JML}}(\theta)]$ obtained using (8) back into $\ell^{\text{MUL}}(\theta, \hat{\boldsymbol{b}}^{\text{MUL}}_{\text{JML},p_d})$ and by solving the outer maximization wrt the remaining parameter θ as

$$\begin{aligned} \hat{\theta}^{\text{MUL}}_{\text{JML},p_d} &= \arg\max_{\theta} \ \ell^{\text{MUL}}(\theta, \hat{\boldsymbol{b}}^{\text{MUL}}_{\text{JML},p_d}(\theta)), \\ \hat{\boldsymbol{b}}^{\text{MUL}}_{\text{JML},p_d} &= \hat{\boldsymbol{b}}^{\text{MUL}}_{\text{JML},p_d}(\hat{\theta}^{\text{MUL}}_{\text{JML},p_d}). \end{aligned} \tag{10}$$

Some important aspects are now discussed in order. First, it is worth observing that though the problem is generally cast as an *estimation* problem, the nature of the involved optimization variables, that is, the global environmental parameter θ and the sensors states b_i, $i = 1, \ldots, N$, is different being the former a continuous variable and the latter binary variables. In this regard, estimation of θ is carried out following the ML rationale encoded through the log-likelihood function in (7), which depends on the unknown sensors state b_i's. However, since the latter variables take on binary values (either α or β in this case), their "estimation" practically coincide with a *decision/detection* among two possible states, which is performed through the MAP classifier reported in (8), using as a threshold for the decision the expression in (9). It can be noticed that also the MAP classifier depends in turn on the unknown value of θ, justifying the joint ML-MAP optimal approach where the way estimation and decision processes affect each other is explicitly taken into account in solving the problem.

Second, it should be noted that $\ell^{\text{MUL}}(\theta, \hat{\boldsymbol{b}}^{\text{MUL}}_{\text{JML},p_d}(\theta))$ is differentiable except at a finite number of points, and between two successive non-differentiable points the function is concave. Therefore, the local maxima of the function coincide with its critical points. Given its nature, this strategy can be considered a joint decision and estimation approach [28].

2) CASE OF UNKNOWN HYPERPARAMETER p

The above approach can be naturally extended to the case in which the hyperparameter p is completely unknown and should be inferred from the collected measurements. To solve the problem in such a case, we can adopt a joint parameter-hyperparameter ML-MAP Bayesian estimation approach

$$(\hat{\theta}^{\text{MUL}}_{\text{JML}}, \ \hat{\boldsymbol{b}}^{\text{MUL}}_{\text{JML}}, \ \hat{p}^{\text{MUL}}_{\text{JML}}) = \arg\max_{\theta, \boldsymbol{b}, p} \ \ell^{\text{MUL}}(\theta, \boldsymbol{b}, p). \tag{11}$$

The maximization of the log-likelihood $\ell^{\text{MUL}}(\theta, \boldsymbol{b}, p)$ follows the same rationale of the approach devised for the case of fixed $p = p_d$, except for the last maximization wrt θ in (10) that is instead replaced by a maximization over the joint space (θ, p). Again, the computational cost increases from a 1D search in case of fixed $p = p_d$ to a 2D search for the case of unknown p. It is worth noting that it is not evident *ex ante* which of the two approaches generally provides the best performance, therefore the two variants of the algorithm will be thoroughly compared later in the performance evaluation provided in Sec. VI.

B. REDUCED-COMPLEXITY TWO-STEP ESTIMATION AND SENSORS CLASSIFICATION

In this section, we derive a novel two-step algorithm able to perform estimation of the global parameter θ and classification of the sensors state $\{b_i\}_{i=1}^N$, but at a reduced cost compared to the optimal joint ML-MAP algorithm illustrated in the previous section. The idea behind the proposed two-step algorithm is to first obtain an estimate of the global parameter θ, and then to use its estimated value to perform a classification of the sensors state $\{b_i\}_{i=1}^N$ in a second step. Given the processing chain underlying this strategy, the proposed algorithm will be denoted as *estimation-then-classification (EC)* approach [28].

1) ESTIMATION STEP

To obtain an estimate of the global parameter θ, we propose a maximum likelihood (ML) approach based on the unconditional distribution of the whole collected sensors measurements, i.e.,

$$\hat{\theta}^{\text{MUL}}_{\text{EC},p_d} = \arg\max_{\theta \in \mathbb{R}} \prod_{i=1}^{N} f^{\text{MUL}}(y_i) = \prod_{i=1}^{N} \sum_{b_i \in \{\alpha, \beta\}} f^{\text{MUL}}(y_i | b_i) f(b_i) \tag{12}$$

with $f(b_i)$ denoting the probability mass function of the random variable b_i (which we recall is distributed according to the discrete distribution $\mathcal{B}(\alpha, \beta, p)$) and $p = p_d$ fixed as a design parameter. It is worth noting that the estimator in (12) performs a marginalization wrt to the distribution of the sensors state parameters $\{b_i\}_{i=1}^N$, so as to get rid of their unknown values. It is not difficult to show that the above problem is equivalent to the following optimization problem

$$\begin{aligned} \hat{\theta}^{\text{MUL}}_{\text{EC},p_d} &= \arg\max_{\theta \in \mathbb{R}} \sum_{i=1}^{N} \ell^{\text{MUL}}(\theta; y_i) \\ &= \sum_{i=1}^{N} \log \left\{ \frac{1 - p_d}{\alpha} e^{-\frac{(y_i - \theta)^2}{2\alpha^2}} + \frac{p_d}{\beta} e^{-\frac{(y_i - \theta)^2}{2\beta^2}} \right\} \end{aligned} \tag{13}$$

where $\ell^{\text{MUL}}(\theta; y_i)$ denotes the log-likelihood of the unconditional distribution $f^{\text{MUL}}(y_i)$.

The above approach can be extended also to the case in which the hyperparameter p of the $\mathcal{B}$ distribution is completely unknown. In this case, a natural solution would be to infer its value from all the collected measurements by using a joint parameter-hyperparameter ML estimation approach

$$(\hat{\theta}^{\text{MUL}}_{\text{EC}},\ \hat{p}^{\text{MUL}}_{\text{EC}}) = \underset{\theta\in\mathbb{R},\ 0<p<1}{\arg\max} \sum_{i=1}^{N} \ell^{\text{MUL}}(\theta, p; y_i). \tag{14}$$

Solving the above problem requires that a two dimensional (2D) search over the joint (θ, p) space is performed, while the optimization problem in (13) for the case of fixed p involves only a one dimensional (1D) search over the space of θ.

In practical scenarios, one can choose which of the two variants of the proposed Bayesian algorithms can be more convenient based on the possible availability of some a-priori information about the percentage of faulty sensors. For the case of WSN nodes, for instance, the value of p_d can be coarsely inferred from experimental fault tests on the network. Nonetheless, we anticipate that the value of p_d is not very critical for the proposed algorithms: in Sec. VI we will show that the algorithms indeed possess an inherent robustness against erroneous values of p_d. Clearly, when the chosen p_d is very close to the true proportion of anomalies, the proposed algorithms will disclose their best performance. On the other hand, when no a-priori information on the percentage of faulty nodes is available, the novel extensions of the algorithms that infer p from data (following an empirical Bayes approach) allow to overcome the need of choosing a specific value for p_d, though at the price of an increased computational cost.

2) CLASSIFICATION STEP

Once the global parameter θ (and possibly the hyperparameter p) has been estimated, in a second step we can retrieve the state of each sensor — which for the multiplicative model is represented by the two possible values of b_i (either normal or faulty) — by adopting a Bayesian maximum a-posteriori (MAP) classifier. To this aim, we first derive the corresponding MAP distribution

$$\begin{aligned} f(b_i|y_i) &= \frac{f^{\text{MUL}}(y_i|\hat{\theta}, b_i)f(b_i)}{f^{\text{MUL}}(y_i)} \\ &= \frac{\frac{1}{b_i} e^{-\frac{(y_i-\hat{\theta})^2}{2b_i^2}} \left[\frac{p}{\beta}\frac{b_i-\alpha}{\beta-\alpha} - \frac{1-p}{\alpha}\frac{b_i-\beta}{\beta-\alpha}\right]}{\frac{1-p}{\alpha} e^{-\frac{(y_i-\hat{\theta})^2}{2\alpha^2}} + \frac{p}{\beta} e^{-\frac{(y_i-\hat{\theta})^2}{2\beta^2}}} \end{aligned} \tag{15}$$

using the previously estimated value of θ. Then, by comparing the MAP distribution $f(b_i|y_i)$ evaluated in the two possible values of b_i, we obtain the MAP classifier for the state condition of the i-th sensor node, which has exactly the same form of (8), but uses the value of the global parameter $\hat{\theta}$ estimated in the previous step. Notice that, given the order of the two processing steps discussed above, the proposed algorithm first tries to make the best estimation $\hat{\theta}$ of the global parameter θ, and then do decision based on the estimation as if it was the true value. This is tantamount to replacing the original composite distribution $f^{\text{MUL}}(y_i|\theta, b_i)$, for $\theta \in \mathbb{R}$, with its single (simple) most likely version $f^{\text{MUL}}(y_i|\hat{\theta}, b_i) = \max_\theta f^{\text{MUL}}(y_i|\theta, b_i)$. The latter is then used as a surrogate to construct the MAP distribution and perform the classification task in the second step. In this respect, thus, the novel two-step EC approach *decides* on the sensors state using a MAP classifier with posterior distribution given in (15), but using as input an estimate of θ (either $\hat{\theta}^{\text{MUL}}_{\text{EC},p_d}$ or $\hat{\theta}^{\text{MUL}}_{\text{EC}}$) obtained in the first step through an unconditional ML estimation process with log-likelihood function expressed by (13).

IV. BAYESIAN ALGORITHMS FOR THE ADDITIVE ERROR MODEL

With the same rationale of Sec. III, in this section we illustrate the optimal joint ML-MAP estimator and propose an alternative reduced-complexity two-step algorithm for the additive error model discussed in Sec. II. According to this error model, each sensor measurement y_i is a Gaussian mixture distributed according to the pdf

$$\begin{aligned} f^{\text{ADD}}(y_i) &= (1-p)f^{\text{ADD}}(y_i|\theta, \gamma) + p\, f^{\text{ADD}}(y_i|\theta, \nu) \\ &= \frac{1-p}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i-\gamma-\theta)^2}{2\sigma^2}} + \frac{p}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i-\nu-\theta)^2}{2\sigma^2}} \end{aligned} \tag{16}$$

where

$$f^{\text{ADD}}(y_i|\theta, a_i) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i-a_i-\theta)^2}{2\sigma^2}}, \quad a_i \in \{\gamma, \nu\}. \tag{17}$$

A. JOINT ML-MAP ESTIMATION

Also in this case, the goal is to estimate the global parameter θ and the specific state of each sensor a_i, in the two different cases of p fixed as a design parameter ($p = p_d$) or p unknown.

1) CASE OF p AS A DESIGN PARAMETER

Let us consider the former case and denote with $f^{\text{ADD}}(\boldsymbol{y}, \boldsymbol{a}|\theta)$ the joint distribution of the sensors measurements $\boldsymbol{y}$ and of the sensors state $\boldsymbol{a} = [a_1 \ \cdots \ a_N]^{\mathsf{T}}$ (interpreted as a density in $\boldsymbol{y}$ and probability in $\boldsymbol{a}$), given the parameter θ, with $p = p_d$. The joint ML-MAP estimation problem can be formulated as

$$(\hat{\theta}^{\text{ADD}}_{\text{JML},p_d},\ \hat{\boldsymbol{a}}^{\text{ADD}}_{\text{JML},p_d}) = \underset{\theta,\boldsymbol{a}}{\arg\max}\ \ell^{\text{ADD}}(\theta, \boldsymbol{a}) \tag{18}$$

where $\ell^{\text{ADD}}(\theta, \boldsymbol{a})$ denotes the log-likelihood of $f^{\text{ADD}}(\boldsymbol{y}, \boldsymbol{a}|\theta)$ and

$$\begin{aligned} \ell^{\text{ADD}}(\theta, \boldsymbol{a}) \propto & \sum_{i\in\mathcal{N}_\gamma} \log\left\{(1-p_d)e^{-\frac{(y_i-\gamma-\theta)^2}{2\sigma^2}}\right\} \\ & + \sum_{i\in\mathcal{N}\setminus\mathcal{N}_\gamma} \log\left\{p_d e^{-\frac{(y_i-\nu-\theta)^2}{2\sigma^2}}\right\} \end{aligned} \tag{19}$$

with $\mathcal{N}_\gamma = \{i \in \mathcal{N}|a_i = \gamma\}$. The inner maximization wrt $\boldsymbol{a}$ in (18) can be carried out separately for each a_i, $i = 1, \ldots, N$, and is solved by directly comparing the log-likelihood function evaluated in the two possible values $\{\gamma, \nu\}$ each a_i can takes on. After some calculations, we then obtain

$$\hat{a}_{i,\mathrm{JML}}^{\mathrm{ADD}}(\theta) = \begin{cases} \gamma \text{ if } \theta - y_i > \delta^{\mathrm{ADD}} \\ \nu \text{ otherwise} \end{cases} \tag{20}$$

where

$$\delta^{\mathrm{ADD}} = \frac{\sigma^2}{\nu - \gamma} \log\left(\frac{p_d}{1-p_d}\right) - \frac{\nu+\gamma}{2}. \tag{21}$$

Accordingly, the final joint ML-MAP solution can be derived by plugging the closed-form estimate of the vector $\hat{\boldsymbol{a}}_{\mathrm{JML},p_d}^{\mathrm{ADD}}(\theta) = [\hat{a}_1(\theta) \cdots \hat{a}_N(\theta)]$ obtained using (20) back into $\ell^{\mathrm{ADD}}(\theta, \hat{\boldsymbol{a}})$ and by solving the outer maximization wrt the remaining parameter θ as

$$\begin{aligned} \hat{\theta}_{\mathrm{JML},p_d}^{\mathrm{ADD}} &= \arg\max_{\theta} \ \ell^{\mathrm{ADD}}(\theta, \hat{\boldsymbol{a}}_{\mathrm{JML},p_d}^{\mathrm{ADD}}(\theta)), \\ \hat{\boldsymbol{a}}_{\mathrm{JML},p_d}^{\mathrm{ADD}} &= \hat{\boldsymbol{a}}_{\mathrm{JML},p_d}^{\mathrm{ADD}}(\hat{\theta}_{\mathrm{JML},p_d}^{\mathrm{ADD}}). \end{aligned} \tag{22}$$

Following a reasoning similar to that in Sec. III-A, also in this case estimation of θ is carried out following the ML rationale encoded through the log-likelihood function in (19), which depends on the unknown sensors state a_i's. The latter variables take on binary values (either γ or ν in this case), hence their estimation coincide with a *decision* among two possible states, which is performed through the MAP classifier reported in (20), using the threshold given in (21). From a quick inspection, it emerges that also $\ell^{\mathrm{ADD}}(\theta, \hat{\boldsymbol{a}}(\theta))$ is differentiable except at a finite number of points, and between two successive non-differentiable points the function is concave. Therefore, the local maxima of the function coincide with its critical points.

2) CASE OF UNKNOWN HYPERPARAMETER p

The above approach can be analogously extended to the case in which the hyperparameter p is completely unknown. Using a joint parameter-hyperparameter ML-MAP estimation approach, it follows that

$$(\hat{\theta}_{\mathrm{JML}}^{\mathrm{ADD}}, \ \hat{\boldsymbol{a}}_{\mathrm{JML}}^{\mathrm{ADD}}, \ \hat{p}_{\mathrm{JML}}^{\mathrm{ADD}}) = \arg\max_{\theta, \boldsymbol{a}, p} \ \ell^{\mathrm{ADD}}(\theta, \boldsymbol{a}, p). \tag{23}$$

Again, the maximization of $\ell^{\mathrm{ADD}}(\theta, \boldsymbol{a}, p)$ follows the same rationale of the approach devised for the case of fixed p, except for the last maximization wrt θ in (22) that is instead replaced by a maximization over the joint space (θ, p).

B. REDUCED-COMPLEXITY TWO-STEP ESTIMATION AND SENSORS CLASSIFICATION

We now illustrate a two-step algorithm able to perform estimation of the global parameter θ and classification of the sensors state $\{a_i\}_{i=1}^N$, but at a reduced cost compared to the optimal joint ML-MAP algorithm derived in the previous section.

1) ESTIMATION STEP

Assuming the hyperparameter $p = p_d$ to be a fixed design parameter, we first obtain an estimate of the θ parameter using a ML approach based on the unconditional distribution of the whole data as

$$\hat{\theta}_{\mathrm{EC},p_d}^{\mathrm{ADD}} = \arg\max_{\theta \in \mathbb{R}} \prod_{i=1}^{N} f^{\mathrm{ADD}}(y_i) = \prod_{i=1}^{N} \sum_{a_i \in \{\gamma,\nu\}} f^{\mathrm{ADD}}(y_i|a_i) f(a_i) \tag{24}$$

with $f^{\mathrm{ADD}}(y_i|a_i)$ given in (17) and $f(a_i)$ the probability mass function of the random variable a_i (which we recall is distributed according to the discrete distribution $\mathcal{B}(\gamma, \nu, p)$). After simple calculations, the above optimization problem can be more conveniently expressed as

$$\begin{aligned} \hat{\theta}_{\mathrm{EC},p_d}^{\mathrm{ADD}} &= \arg\max_{\theta \in \mathbb{R}} \sum_{i=1}^{N} \ell^{\mathrm{ADD}}(\theta; y_i) \\ &= \sum_{i=1}^{N} \log\left\{(1-p_d) e^{-\frac{(y_i-\gamma-\theta)^2}{2\sigma^2}} + p_d e^{-\frac{(y_i-\nu-\theta)^2}{2\sigma^2}}\right\} \end{aligned} \tag{25}$$

where $\ell^{\mathrm{ADD}}(\theta; y_i)$ denotes the log-likelihood of the unconditional distribution $f^{\mathrm{ADD}}(y_i)$. Also in this case, the proposed estimator lends itself to be naturally extended to the case in which p is unknown, using a joint parameter-hyperparameter ML estimation approach

$$(\hat{\theta}_{\mathrm{EC}}^{\mathrm{ADD}}, \ \hat{p}_{\mathrm{EC}}^{\mathrm{ADD}}) = \arg\max_{\theta \in \mathbb{R}, \ 0<p<1} \sum_{i=1}^{N} \ell^{\mathrm{ADD}}(\theta, p; y_i). \tag{26}$$

2) CLASSIFICATION STEP

In the second step, we use the estimate $\hat{\theta}$ of the global parameter (and possibly the estimate of p) to derive a MAP classifier able to discriminate the state of each sensor. To this aim, we construct the corresponding MAP distribution

$$\begin{aligned} f(a_i|y_i) &= \frac{f^{\mathrm{ADD}}(y_i|\hat{\theta}, a_i) f(a_i)}{f^{\mathrm{ADD}}(y_i)} \\ &= \frac{e^{-\frac{(y_i-a_i-\theta)^2}{2\sigma^2}} \left[(a_i-\gamma)\frac{p}{\nu-\gamma} - (a_i-\nu)\frac{1-p}{\nu-\gamma}\right]}{(1-p) e^{-\frac{(y_i-\gamma-\theta)^2}{2\sigma^2}} + p e^{-\frac{(y_i-\nu-\theta)^2}{2\sigma^2}}}. \end{aligned} \tag{27}$$

By comparing the MAP distribution $f(a_i|y_i)$ evaluated in the two possible values of a_i, we obtain the final MAP classifier whose expression is the same as that in (20) using (21), except for the value of θ that is replaced by its estimate $\hat{\theta}$. The two-step EC approach thus performs a *decision* on the sensors state using a MAP classifier with posterior given in (27), using as input an estimate of θ (either $\hat{\theta}_{\mathrm{EC},p_d}^{\mathrm{ADD}}$ or $\hat{\theta}_{\mathrm{EC}}^{\mathrm{ADD}}$) obtained in the first step through an unconditional ML estimation process with log-likelihood given in (25).

V. COST ANALYSIS

In this section, we investigate in detail the computational complexity of the novel two-step EC algorithms proposed in Secs. III-B (multiplicative model) and IV-B (additive model), in comparison to the optimal joint ML-MAP estimators illustrated in Secs. III-A (multiplicative model) and IV-A (additive model). Let us start by considering the case in which the hyperparameter of the Gaussian mixtures is set to a fixed design value $p = p_d$. Asymptotically speaking, the complexity in performing the optimization required by the optimal joint ML-MAP estimators (please refer to (6) and (18)) can be expressed as the sum of the following terms

$$\mathcal{O}(PN + PN + P \log P) \tag{28}$$

where P denotes the number of evaluation points used to perform the 1D search over the space of θ. The first two addends represent the cost required to construct and evaluate the log-likelihood function, and to jointly perform a classification of the N sensors states (as either regular or anomalous) for each of the P individual trial values. The third term instead indicates the complexity required to select the best candidate trial value (among the P) corresponding to the maximum of the log-likelihood function.

On the other hand, by analyzing the different steps involved in the proposed two-step EC algorithms, it emerges that the overall asymptotic complexity is given by $\mathcal{O}(PN + P \log P + N)$. The first term represents the cost associated to the *estimation step*, which involves P-times an evaluation of the log-likelihood function, followed again by a search for the trial value corresponding to its maximum. The third addend represents instead the complexity required by the subsequent *classification* step, which in this case is performed only once using the value of the global parameter $\hat{\theta}$ estimated in the previous step as it if was the true one. In doing so, the two-step EC algorithms are able to reduce the complexity required by the joint ML-MAP algorithms, which instead perform P different classifications, one for each individual trial value.

The complexities of the extended versions of the joint ML-MAP (please refer to (11) and (23)) and two-step EC algorithms (please refer to (14) and (26)) that also estimate the hyperparameter p from the data can be obtained by incorporating in the above expressions the cost related to an additional 1D search over the space of p. Assuming for the sake of the exposition that the number of evaluation points used to perform such an additional search is also P, we end up with a complexity in the order of $\mathcal{O}(P^2N + P^2N + P^2 \log P^2)$ for the joint ML-MAP algorithms, and of $\mathcal{O}(P^2N + P^2 \log P^2 + N)$ for the two-step EC algorithms. Clearly, the big-O notation may hide constants that can impact onto the actual computational cost. Therefore, to corroborate the above asymptotic analysis, in the next section we will also compare the average runtimes of all the algorithms when executed on the same hardware platform.

VI. PERFORMANCE ASSESSMENT AND RESULTS

In this section, we assess the performance of the unified Bayesian framework by testing all the presented algorithms on synthetic data, as well as on timeseries of real sensors data (temperature measured by a network of fixed monitoring stations deployed in the south of Italy) accounting for the presence of an anomalous stream coming from a faulty sensor.

A. STATE-OF-THE-ART COMPETITOR ALGORITHMS

We describe below additional state-of-the-art algorithms that will be compared against the Bayesian algorithms. In this respect, it is worth remarking that the joint ML-MAP approach, which has been already illustrated in other similar works dealing with joint estimation and anomaly detection but in different distributed contexts [29], [30], [31], [32], represents the *benchmark* for the considered problem. Indeed, it provides the optimal solution by seeking for the maximum of the log-likelihood function (please refer to (7) for the multiplicative model and (19) for the additive model) with respect to both the global parameter (estimation) and sensors states (classification). More specifically, the whole maximization is partly ML in the global parameter θ, and partly MAP in the discrete random vector containing the sensors states (either $\boldsymbol{b}$ or $\boldsymbol{a}$ according to the error model). In doing so, the joint ML-MAP approach explicitly takes into account the mutual dependency between estimation and decision processes and the way they affect each other, thus representing the best possible algorithm (i.e., any other possible algorithm could only perform worse than it).

1) DBSCAN-BASED APPROACH

A natural competitor can be identified by considering a dual approach in which classification of sensor nodes (as either normal or anomalous) is performed in a first step, and then an estimate of the global parameter θ is obtained in a second step using the decisions made in the classification step as if they were the true ones. To this aim, the sensors classification problem can be interpreted as a clustering problem, where the goal is to group together measurements y_i's coming from the same distribution. A rather common clustering algorithm is the k-means, which for the specific problem at hand would have $k = 2$. However, k-means and other techniques such as hierarchical clustering are not suitable for the considered models for two main reasons: i) both the multiplicative and additive model distributions have a non-convex shape; ii) the typical low values[1] of the hyperparameter p would make it difficult to distinguish two different clusters, being in that cases more appropriate to consider a single cluster with some individual outliers.

To make a more fair comparison, for the classification step we consider the well-known density-based spatial clustering of applications with noise (DBSCAN, [17]). Compared to k-means, DBSCAN does not need to assume a specific number of clusters and is compatible with non-convex shapes in general. Moreover, it is suitable for making decisions even with low values of p, being its derivation inclusive of a notion of ‘‘noise’’, and is also very robust to outliers. DBSCAN only

[1] Apart from extraordinary cases, it is reasonable to expect that in real applications the probability of having faulty sensors is usually quite low.

requires the definition of two design parameters, which can be set according to the specific error model at hand:

- the neighborhood search radius ϵ, which in the following is set to $\beta/2$ and $|\nu|/2$ (i.e., half the entity of the anomaly) for the multiplicative and additive error model, respectively;
- the minimum number of neighboring points `minPts` to elect a measure as a core point, which is set to $N/2$ (i.e., half the dimension of the network) for both models.

Once the classification step has been performed using DBSCAN, it is possible to retrieve an estimate of θ using a weighted least square approach for the multiplicative model, and a corrected arithmetic mean for the additive model, namely

$$\hat{\theta}^{\text{MUL}}_{\text{DBSCAN}} = \frac{\sum_{i=1}^{N} y_i/\hat{b}^2_{i,\text{DBSCAN}}}{\sum_{i=1}^{N} 1/\hat{b}^2_{i,\text{DBSCAN}}} \tag{29}$$

$$\hat{\theta}^{\text{ADD}}_{\text{DBSCAN}} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{a}_{i,\text{DBSCAN}}) \tag{30}$$

where $\hat{b}_{i,\text{DBSCAN}}$ and $\hat{a}_{i,\text{DBSCAN}}$ are the outcomes of the classification step performed by DBSCAN based on the multiplicative and additive model, which reflect the estimated operational state of the i-th sensor in the network. The specific order of the processing steps discussed above explains the rationale for choosing the DBSCAN-based approach as one of the considered competitors: indeed, it can be naturally seen as the "dual" approach compared to the proposed two-step EC algorithms, which we recall first perform estimation of the global parameter, and then do sensor state classification based on the estimation as if it was the true value.

2) SIMPLE ESTIMATION AND CLASSIFICATION (SEC)

With the aim of understanding the potential advantages of the Bayesian algorithms, we also consider a very simple and lightweight estimation and sensor classification (SEC) algorithm for both models:

- under the multiplicative model, the estimate of θ is chosen as the median value of the measurements y_i's, whereas a naive 3α threshold is adopted to make decision about the i-th sensor state: $b_i = \alpha$ if $|y_i - \hat{\theta}| < 3\alpha$; $b_i = \beta$ otherwise;
- under the additive model, we consider the sample mean as estimate of θ, and always decide for the normal operational behavior of sensors, namely $a_i = \gamma$, $\forall i = 1, \ldots, N$.

Both these approaches should be interpreted as the upper bound for estimation (in terms of error) and the lower bound for classification (in terms of accuracy), respectively. Methods providing lower performance than these simple algorithms cannot be considered effective for the problem at hand.

B. SIMULATION ANALYSIS

In this section, we conduct a simulation analysis to assess the algorithms performance when operating on synthetically-generated measurements. The numerical assessment is performed under different sensors anomalous conditions (in terms of both entity of the anomaly and number of anomalous sensors), also taking into account possible mismatches between the assumed and actual model parameters.

1) SIMULATION SETUP

The considered scenario consists of a network of $N = 20$ sensor nodes, which aims at estimating the average temperature of a given region having a true value of $\theta = 10$ °C. Under regular operating conditions of all sensors, the values of the standard deviations α and σ in the multiplicative and additive error models are set to 1 (which means that the intrinsic sensors precision is about ±3 °C), whereas the mean of the measurement error in the additive model is set to $\gamma = 0$ (we assume sensors with no bias). Such values are compatible with the typical errors (in Celsius degree scale) experienced by commercial temperature/humidity sensors. It is also worth noting that we opted to conduct the performance evaluation using a relatively small number of sensors ($N = 20$) for two simple reasons: i) we want to consider conservative scenarios where the number of active sensors that can effectively carry out measurements over a given time window may be less than the total number of nodes potentially available in the network. Indeed, it is not infrequent to have situations where part of the involved nodes are temporarily unavailable (e.g., set in power-saving mode as happens in WSNs, or voluntarily switched-off as is the case of crowdsensing nodes). Moreover, ii) having a larger number of sensor nodes in general would be more beneficial (given the same amount of faulty nodes) for all the algorithms, being the inference process (estimation and classification) performed on a higher volume of measurements. The latter point will be confirmed by some of the results discussed in the following.

Some of the sensors may suddenly experience an undesired anomalous condition, which in turn affects the quality of the measurements they provide. In order to test the capability of the algorithms to cope with different fault conditions, we first introduce a quantitative metric that represents how severe is an anomalous condition compared to the normal operating condition of a sensor. To keep the analysis general with respect to the value assumed by θ as well as to the specific error model at hand, we adopt the ratio between the root mean square (RMS) value of the observables y_i's under anomalous and regular operating conditions (which has the meaning of a "signal-to-noise" ratio), denoted in the following as *anomalous-to-regular condition ratio (ARR)* and defined (in dB units) as

$$\text{ARR} = 10\log_{10}\left(\frac{\mu_1^2 + \sigma_1^2}{\mu_0^2 + \sigma_0^2}\right) \tag{31}$$

where $\mu_1 = \text{E}[y_i|H_1]$ (with $\mu_1 = \nu$ for the additive model and $\mu_1 = 0$ for the multiplicative model) and

$\sigma_1^2 = \mathrm{VAR}[y_i|H_1]$ (with $\sigma_1 = \sigma$ for the additive model and $\sigma_1 = \beta$ for the multiplicative model), with $\mathrm{E}[\cdot]$ and $\mathrm{VAR}[\cdot]$ denoting the statistical expectation and variance operators. Similarly, $\mu_0 = \mathrm{E}[y_i|H_0]$ (with $\mu_0 = \gamma$ for the additive model and $\mu_0 = 0$ for the multiplicative model) and $\sigma_0^2 = \mathrm{VAR}[y_i|H_0]$ (with $\sigma_0 = \sigma$ for the additive model and $\sigma_0 = \alpha$ for the multiplicative model). As to H_1 and H_0, they denote the anomalous and regular operating conditions, respectively. The value of β in the multiplicative model and of ν in the additive model are then varied so as to obtain different ranges of the ARR. More specifically, low values of ARR are representative of challenging scenarios where the fault is hardly distinguishable from a normal operating condition; on the other hand, increasing values of ARR will make the fault condition progressively more evident.

With the aim of evaluating the performance of the considered algorithms, the tests are conducted by means of $M = 1000$ Monte Carlo trials and the performance are measured using the mean squared error (MSE) on the estimation of θ, corroborated by the accuracy, sensitivity, and specificity of sensor fault detection. Without loss of generality, in the following we assume a single measurement coming from each individual sensor, which means that the total number of processed data amounts to N.

2) RESULTS AND DISCUSSION

We start the evaluation by investigating the performance of the algorithms designed for the multiplicative error model.

Multiplicative Model - Analysis in Absence of Anomalies: in Fig. 2, we report the results obtained by considering a network operating with 0 faulty sensors. It is worth noting that this represents a scenario of practical interest, being it the most common operating condition of a properly deployed sensing network. As it can be observed, in this setup the MSEs of all the considered algorithms are almost constant in the ARR range, with the two Bayesian algorithms and the DBSCAN-based approach providing the best (comparable) performance, characterized by estimation errors always below 0.06. All algorithms also generally provide excellent sensors classification accuracy. In particular, the DBSCAN-based and the SEC approaches guarantee a 100% of classification accuracy, while the two Bayesian algorithms experience practically negligible deviations from 1 (as highlighted by the insets of the figure) only when the probability of sensor fault p_d, used as a fixed design parameter, is set to $p_d = 10^{-1}$. This behavior reveals an intrinsic robustness of the two Bayesian algorithms to erroneous values of p: in fact, they are still able to achieve an accuracy close to 100% despite being fed with a value of p_d significantly different from the actual one (i.e., $p = 0$). On the other hand, when p is inferred from the data (and can thus change across different trials), the Bayesian algorithms correctly classify the sensors state with practically 100% accuracy. Overall, it is interesting to notice that the reduced-complexity EC algorithm provides almost the same performance (in terms of both estimation and anomaly detection) of the joint ML-MAP.

Multiplicative Model - Analysis in Presence of Anomalies: Fig. 3 illustrates the results obtained when four faulty sensors (i.e., 20% of the network size) are present. It is worth highlighting that the analysis is performed by keeping the number of faulty nodes fixed over all the Monte Carlo realizations. Therefore, the operating conditions are always mismatched with respect to the statistical model assumed to derive the Bayesian algorithms at the design stage. In this case, the MSE of the DBSCAN-based approach worsens as the ARR (namely, the value of β) increases. This counterintuitive trend can be explained by noting that although DBSCAN is able to correctly identify sensors operating under normal conditions (specificity in Fig. 3 is always 1), it is not as effective in revealing all the faulty ones, as confirmed by the low values of the sensitivity achieving at most around 0.5. Such wrong decisions negatively impact on the subsequent estimation step performed using (29), with almost 50% of faulty sensors that are erroneously weighted as if they were normal ones. On the other hand, the SEC approach exhibits a rather good estimation performance, due to the fact that the estimator of θ based on the median value of the measurements is robust to the presence of a few outliers. Remarkably, the two Bayesian algorithms guarantee the best estimation performance, with their MSEs that tend to decrease as the ARR increases, attaining almost the same small errors provided in absence of faulty sensors. Such algorithms are also effective in terms of sensors classification, with an accuracy that is always above 90% even for hardly distinguishable faults (ARR lower than 5 dB), and rapidly increases as soon as the ARR increases.

Some interesting considerations can be drawn by comparing the curves of two Bayesian algorithms for different settings of the design parameter p_d. First, it can be noticed that the overall performance in terms of both estimation and sensors classification tends to improve as p_d gets close to the actual value of $p = 2 \cdot 10^{-1}$ (labeled as "matched" in the legend). Second, the algorithms are still very robust to erroneous settings of p_d, with a difference in the achieved performance that tends to vanish as the ARR increases. Third, the variants of the Bayesian algorithms that additionally estimate the hyperparameter p from the data exhibit performance almost equal to that obtained in the case of perfect knowledge of p. This excellent performance comes at the price of an increased complexity of the resulting estimators, which involve a 2D grid search instead of a 1D search, as discussed in Sec. III-B. Not least, the Bayesian two-step EC algorithm keeps providing almost the same performance of the optimal joint ML-MAP, which is a remarkable fact.

Additive Model - Analysis in Absence of Anomalies: We now consider the algorithms designed for the additive error model. Fig. 4 shows the results obtained by considering a network with 0 faulty sensors. As apparent from the MSEs, in this case the SEC approach and the two Bayesian algorithms generally provide the best performance, whereas the DBSCAN-based approach starts to attain low values of the MSE only for higher values of the ARR. Almost the same behavior can be observed in terms of sensors classification,

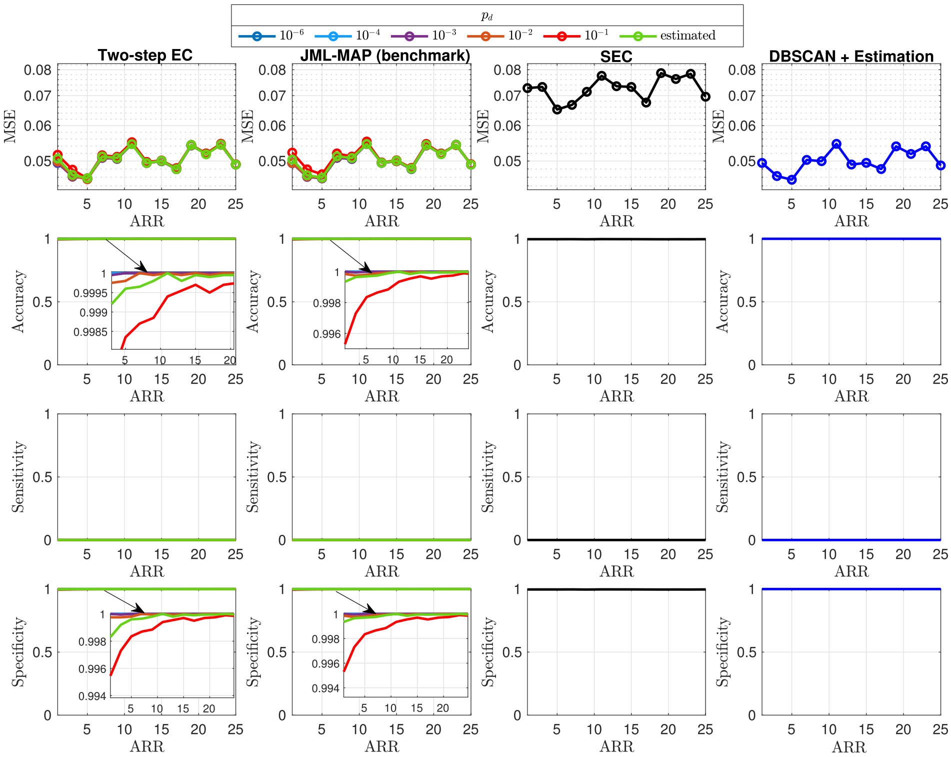


FIGURE 2. Performance of the algorithms for the multiplicative error model, in case of 0 faulty sensors in the network.

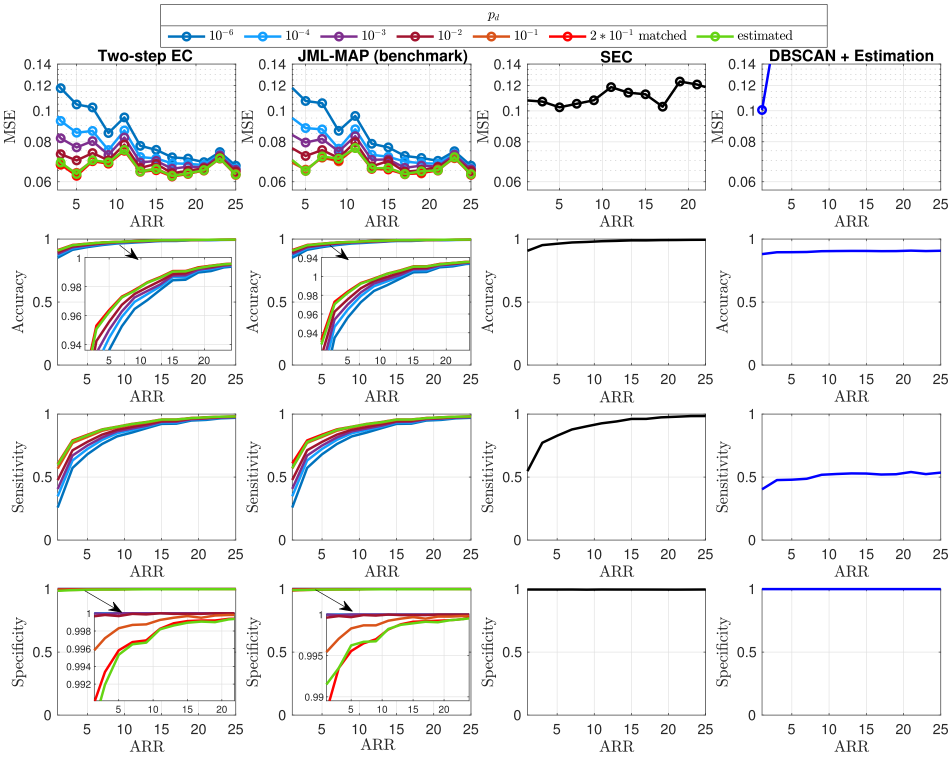


FIGURE 3. Performance of the algorithms for the multiplicative error model, in case of 4 faulty sensors in the network.

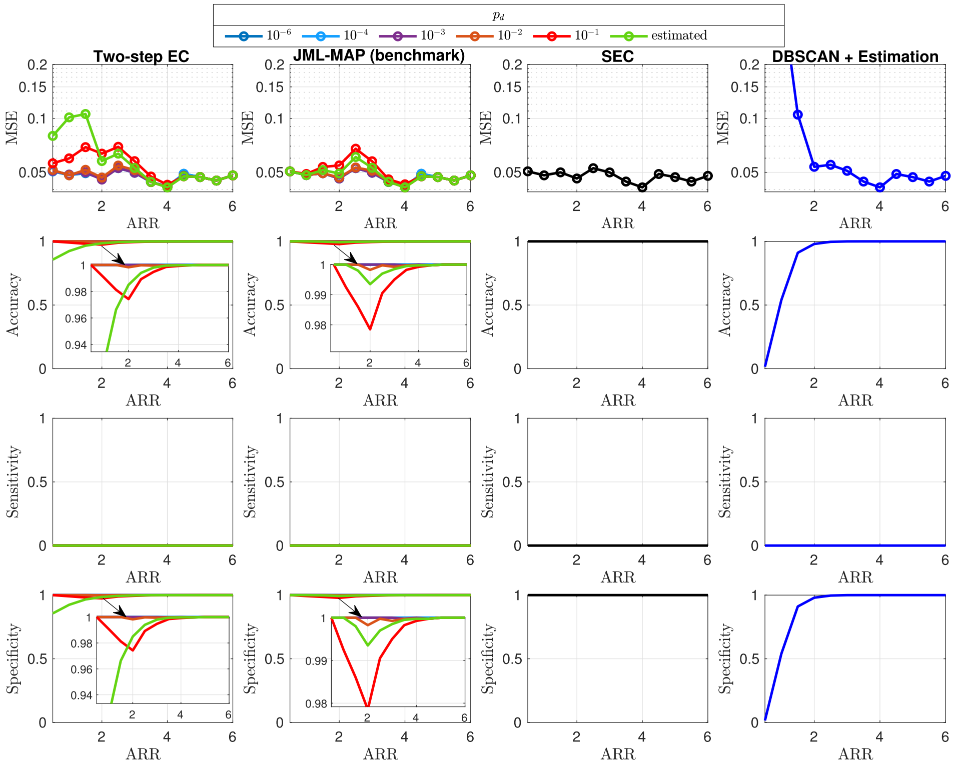


FIGURE 4. Performance of the algorithms for the additive error model, in case of 0 faulty sensors in the network.

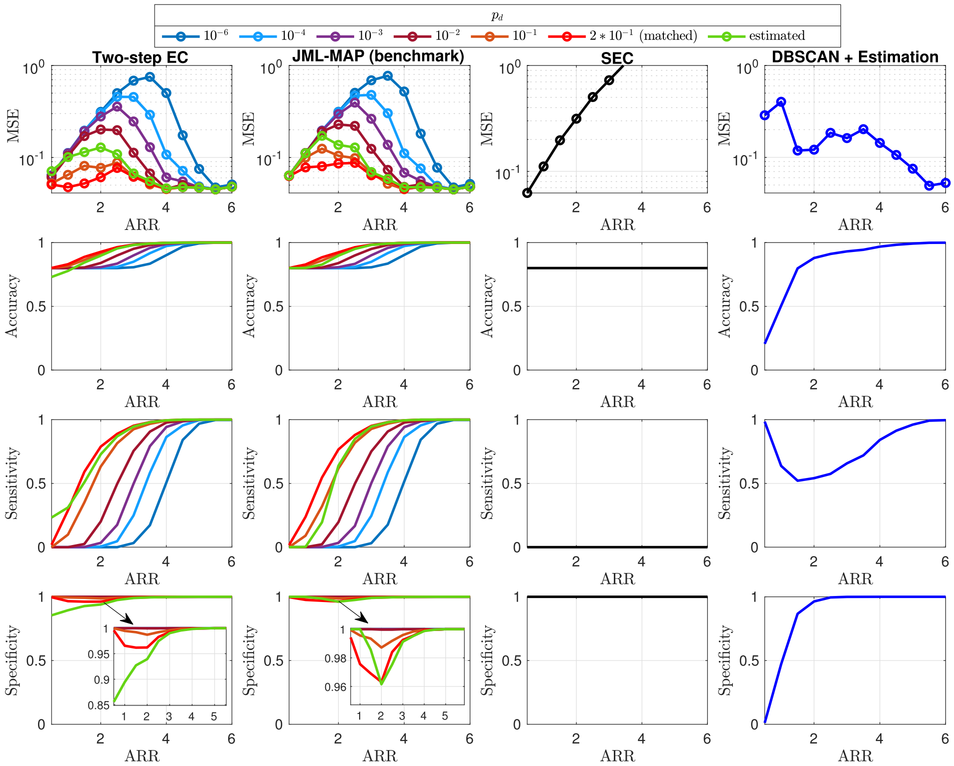


FIGURE 5. Performance of the algorithms for the additive error model, in case of 4 faulty sensors in the network.

with the SEC algorithm and the two Bayesian approaches exhibiting an accuracy close to 100%. The impact of the p_d parameter in the Bayesian approaches is similar to that observed in Fig. 2, with the overall performance that tends to improve as its setting gets close to the actual value of $p = 0$. Compared to the multiplicative error case, some more marked differences can be appreciated between the two-step and the joint ML-MAP algorithms, with the latter exhibiting better performance, especially for lower values of the ARR. The same relative trend is confirmed when the hyperparameter p is inferred from the data.

Additive Model - Analysis in Presence of Anomalies: In Fig. 5, we report the results for the case of 4 faulty nodes in the network. It can be seen that the SEC algorithm, which was effective in absence of faulty sensors, has a MSE that dramatically increases as the ARR increases. On the other hand, the DBSCAN-based approach has a higher value of MSE in the low ARR range, and tends to reduce its estimation error only for ARR greater than 5 dB. The two Bayesian algorithms instead tend to behave essentially the same and, remarkably, achieve superior estimation and sensors classification performance (compared to the other algorithms) either when the value of p_d approaches the actual one or when p is inferred from the data (though at an increased computational cost), with MSEs as low as 0.045 and accuracy close to 100% already for ARR $>$ 3 dB.

Sensitivity Analysis: The novel estimation and anomaly detection approaches we propose rely on two different error models based on Gaussian mixtures, summarized as follows: i) the *multiplicative* model accounting for anomalies affecting the variance of the measurement error; ii) the *additive* model taking into account anomalies that introduce deterministic biases in the mean of the measurement error. To inspect the performance when the errors do not follow the exact distribution (with its related parameters) assumed at the design stage, we perform a sensitivity analysis aimed at investigating the algorithms robustness when they are fed with a value of the distribution parameters encoding the entity of the anomaly (namely β for the multiplicative model and ν for the additive model) that differs from the actual one used to generate the data. This is tantamount to assuming a misknowledge of the actual entity of the anomaly affecting the sensors. In Fig. 6 we report the results obtained for the multiplicative error model, considering the more challenging case of 4 faulty sensors in the network, for ARR = 5 dB. Remarkably, the two Bayesian algorithms are very robust to mismatches, as revealed by the corresponding curves exhibiting a rather constant trend despite the increased mismatch between assumed and actual value of β. Furthermore, they generally exhibit the best performance both in terms of estimation and sensors classification. On the other hand, the DBSCAN-based approach significantly suffers for the presence of a misknowledge on β, with its MSE that significantly increases and the accuracy that progressively decreases as the mismatch degree (encoded through $\Delta\beta$) increases. As to the DBSCAN-based approach, only the first MSE value corresponding to the lowest value on the x-axis is actually visible in Fig. 6 for the chosen limits of the y-axis, being all the remaining MSE values much larger and therefore falling outside the considered range. This is due to the fact that the DBSCAN-based approach significantly suffers from mismatches on the assumed parameter encoding the entity of the anomaly.

Fig. 7 reports the results for the additive error model, assuming the same scenario with 4 faulty sensors and an ARR = 2 dB. It is evident that the two Bayesian algorithms and the DBSCAN-based approach are more sensitive to mismatches on the assumed ν, with MSEs and accuracy that get worse as the mismatch degree $\Delta\nu$ increases. Nevertheless, the Bayesian algorithms still outperform all the competitors, with evident gaps especially when p_d is chosen sufficiently close to the actual p or when it is inferred from the data, and the mismatch level is not too severe. Overall, these analyses demonstrate the effectiveness of the unified Bayesian estimation and classification framework, which correctly deals with the different operating conditions experienced by the sensor network and with partial or inaccurate a priori knowledge of the parameters related to the anomaly.

Scalability Analysis: to further corroborate the above results, we now investigate the scalability of the considered approaches with respect to the size of the sensor network N. The analysis is performed by varying N between 10 and 1000 and by keeping the percentage of anomalous sensors fixed to 20% of the network size. In doing so, the number of anomalous nodes increases proportionally to the network size. The obtained results are reported in Fig. 8 and Fig. 9 for the multiplicative and additive error model, respectively. It is evident that the two-step EC and joint ML-MAP algorithms keep superior performance in terms of both estimation and anomaly detection compared to the DBSCAN-based and SEC algorithms. In particular, their MSEs tend to decrease as the network size increases, achieving values below 10^{-2} already for $N \geq 120$. Moreover, they guarantee very high levels of accuracy over almost all the span of considered N, especially when p_d is not too far from the actual one or when p is inferred from data, achieving in those cases around 100% accuracy. These results are quite interesting since they demonstrate that having an increased number of sensors is anyway beneficial for the two Bayesian algorithms, despite for $N = 1000$ the network accounts for the presence of 200 anomalous sensors.

Analysis for Varying Number of Anomalous Nodes: we now conduct an additional analysis aimed at studying how the number of faulty nodes impacts on the algorithms estimation and detection performance. Considering the results for different network sizes discussed above, for this analysis we opted to keep $N = 20$ and varied the number of faulty nodes from a lower percentage of 5% (compared to the 20% already considered in the previous figures) up to a more challenging percentage of 50%, i.e., when half of the sensor nodes in the network is corrupted. In this case, thus, the true value of p will change as the percentage of anomalous nodes increases. We depict the obtained performance in Fig. 10 and Fig. 11 for the multiplicative and additive error model, respectively.

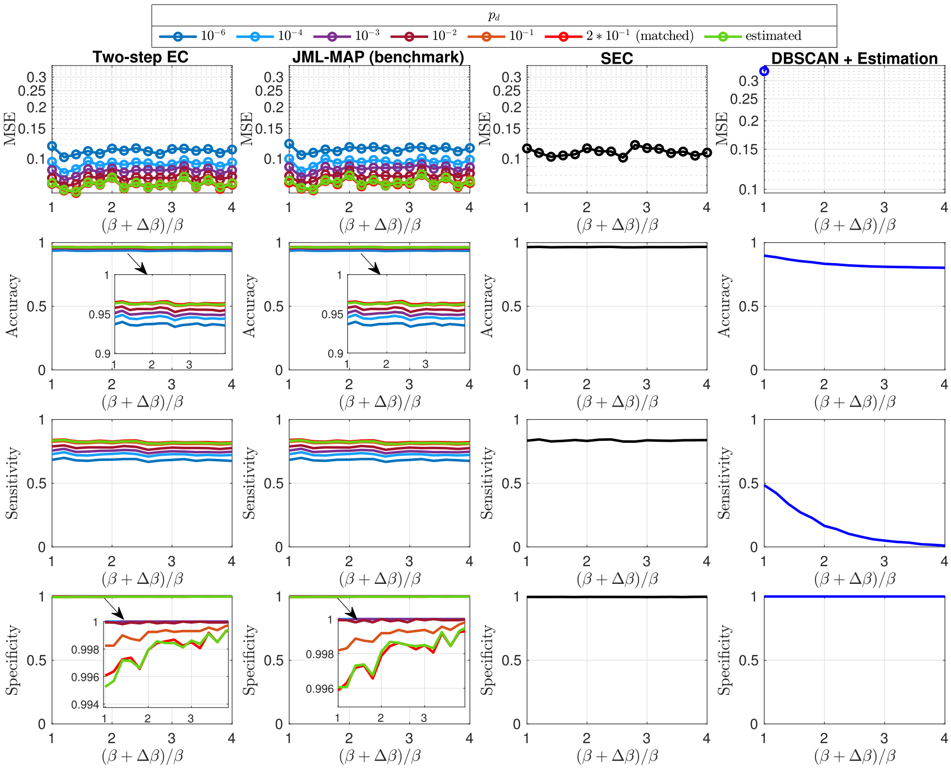


FIGURE 6. Sensitivity analysis of the algorithms for the multiplicative error model, in case of 4 faulty sensors in the network and for ARR = 5 dB.

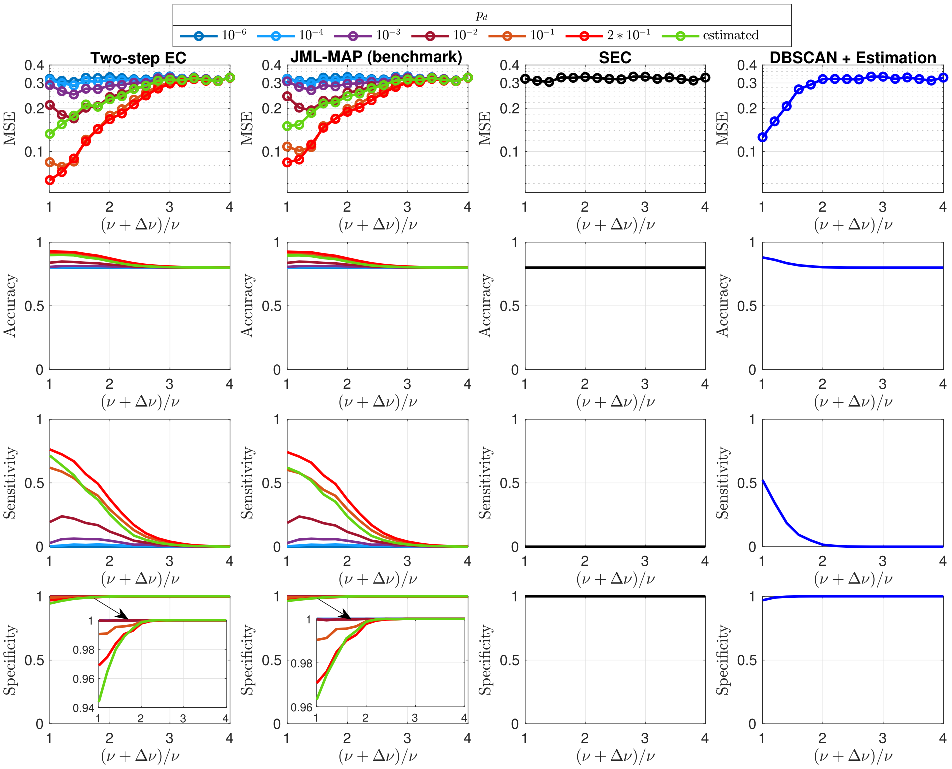


FIGURE 7. Sensitivity analysis of the algorithms for the additive error model, in case of 4 faulty sensors in the network and for ARR = 2 dB.

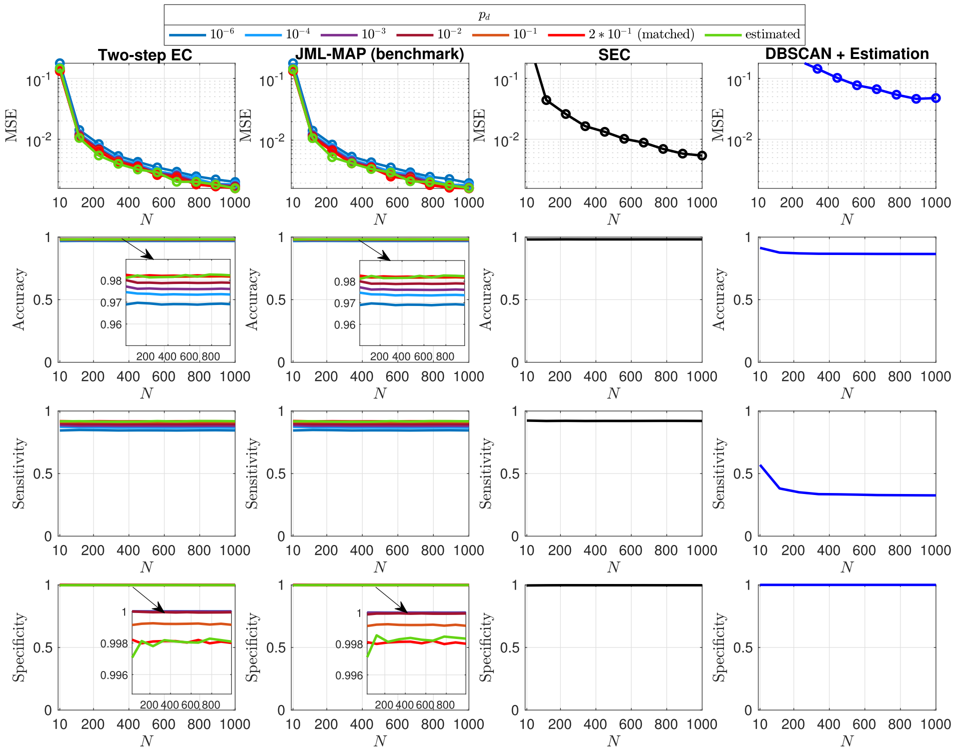


FIGURE 8. Performance of the algorithms for the multiplicative error model as a function of the network size N.

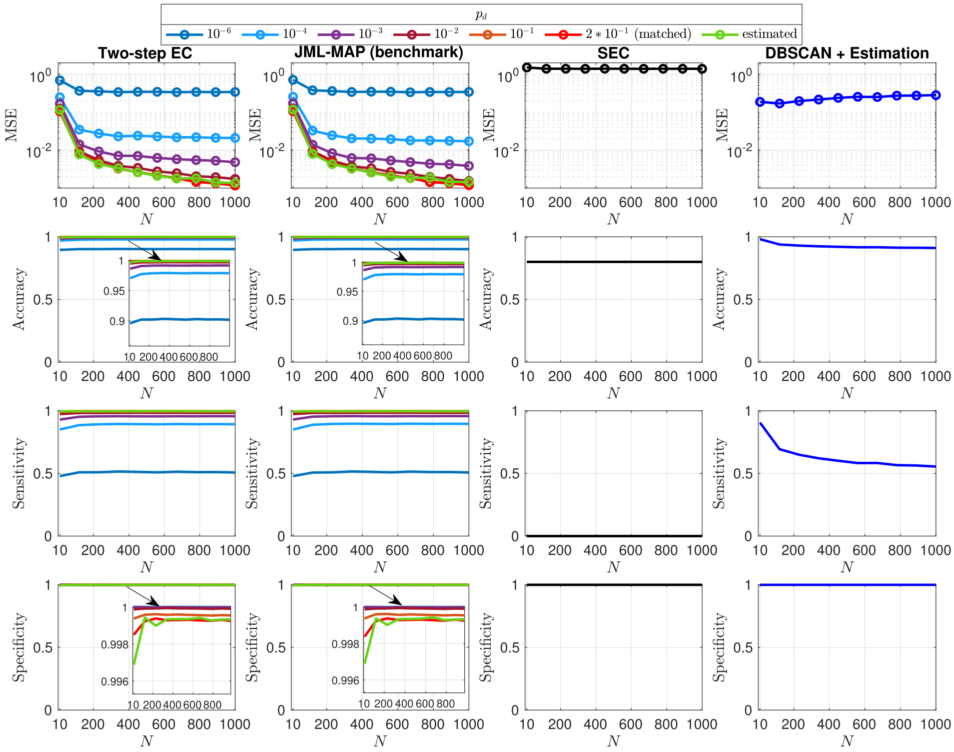


FIGURE 9. Performance of the algorithms for the additive error model as a function of the network size N.

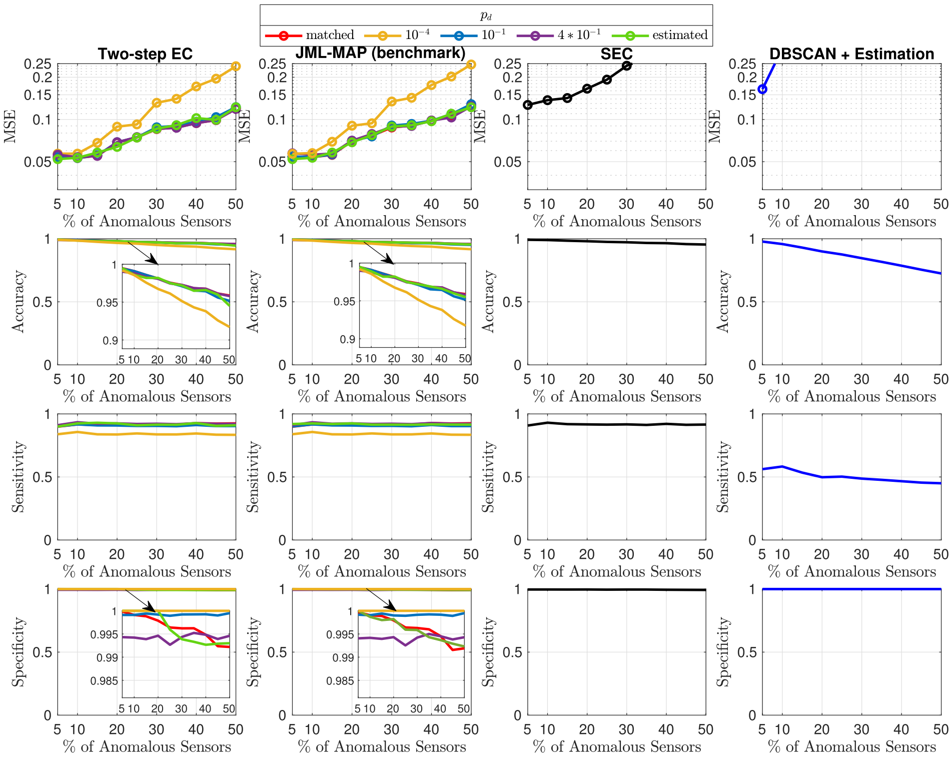


FIGURE 10. Performance of the algorithms for the multiplicative error model as a function of the percentage of anomalous sensors in the network.

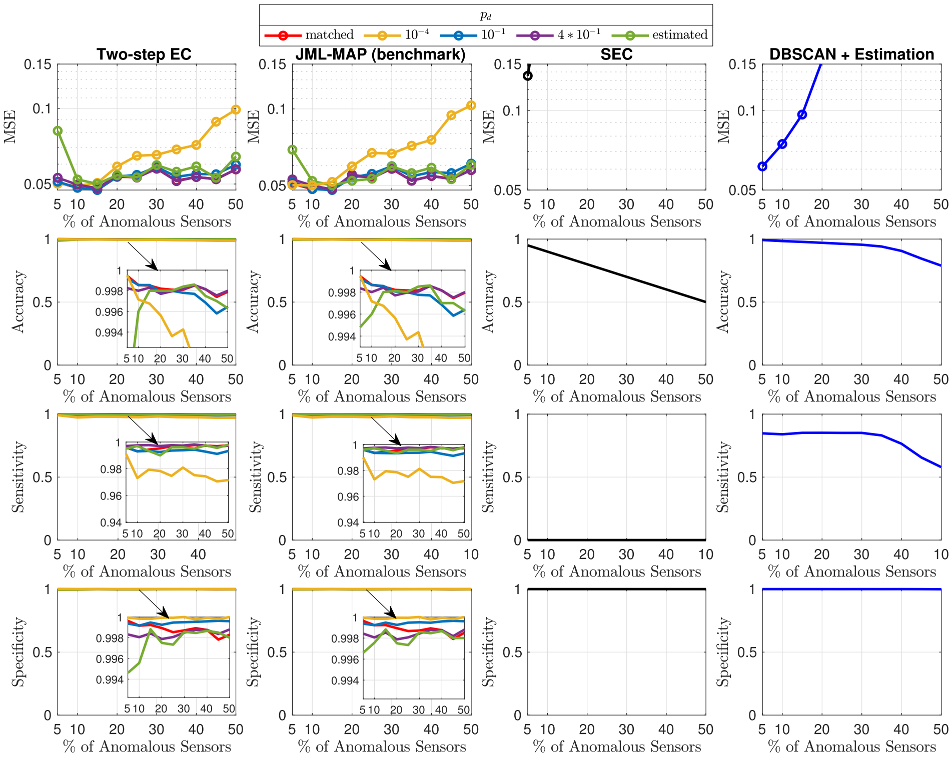


FIGURE 11. Performance of the algorithms for the additive error model as a function of the percentage of anomalous sensors in the network.

Remarkably, the two-step EC and joint ML-MAP outperform the SEC and DBSCAN-based approaches for all the considered percentages of faulty nodes, with MSEs that remain contained in the order of about 0.25 even in the worst case of 50% of corrupted nodes in the network. The two-step EC and joint ML-MAP algorithms also admit a quite intuitive interpretation of their behavior with respect to the design parameter p_d: indeed, for small values of p_d both the MSEs and accuracy tend to be better in the lower region of the x-axis, that is, when the p_d is close to the actual p found in the data. The opposite behavior can be observed instead for higher values of p_d, with better performance obtained in the higher region of the x-axis. Interestingly, when p is estimated via empirical Bayes from data, the two-step EC and joint ML-MAP keep providing excellent performance. From this analysis we can conclude that the two Bayesian algorithms can correctly cope also with the presence of an increased number of faulty nodes in the network.

Complexity Analysis: to conclude the numerical assessment, we perform a complexity analysis to quantify the required computational cost of each considered algorithm. Specifically, we record the runtime of the algorithms when executed on a standard laptop for 200 different trials and compare the corresponding average runtimes, normalized by the average runtime of the most costly algorithm, namely the joint ML-MAP approach that also estimates p from data. The obtained results are reported in Fig. 12. As it would be expected, the SEC approach is the least complex, followed by the DBSCAN-based approach that involves some additional processing in the first classification step. Remarkably, the Bayesian two-step EC algorithm using p_d as a design tunable parameter has a complexity comparable to that of the DBSCAN-based approach (with a normalized average runtime below $5 \cdot 10^{-4}$), while the joint ML-MAP approach (with tunable p_d) has a complexity about an order of magnitude greater. This outcome demonstrates the goodness of the Bayesian two-step EC approach, which is able to attain almost the same performance of the joint ML-MAP approach in almost all operating conditions, but at a fraction of its complexity. The advantages in terms of cost saving become even more evident when the Bayesian algorithms infer the value of p from data (which as shown generally leads to performance as good as in case of perfect knowledge of p): in this case, the two-step EC approach provides a complexity reduction of about 80% compared to the joint ML-MAP approach. From this analysis, it also emerges that having an accurate prior knowledge of the hyperparameter p (which avoids the need to estimate it from data) brings a significant complexity reduction.

C. EVALUATION ON TEMPERATURE DATA FROM A REAL SENSOR NETWORK

In this section, we test the algorithms effectiveness when they are applied on timeseries of real data acquired by a sensor network. As a practical case study, we consider a network of environmental monitoring stations deployed in the province of Lecce, Apulia region, in the south of Italy. The stations provide a real-time monitoring of several parameters including temperature, humidity, pressure, and PMx concentrations, on a daily basis. In the month of May 2022, the temperature sensor installed on a station near Torre dell'Orso experienced an unexpected fault, resulting in a stream of inaccurate measurements, as shown in the snapshot reported in Fig. 13a.

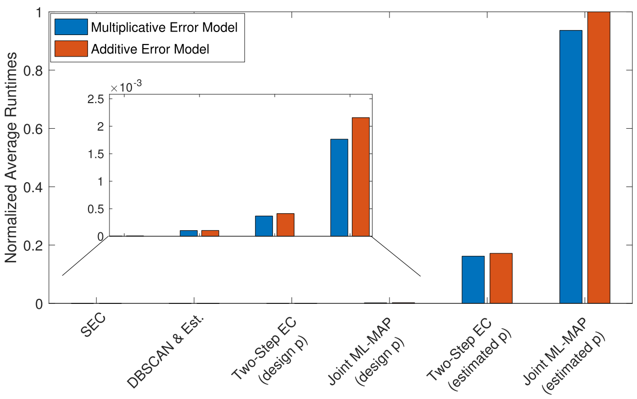
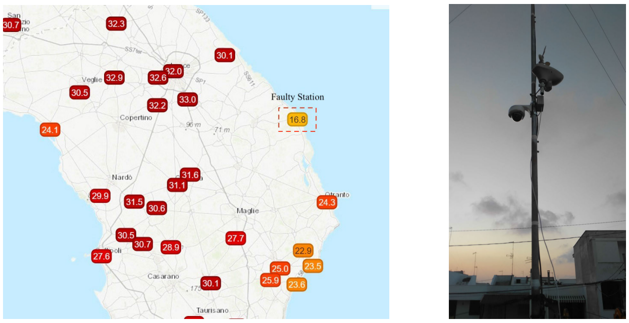


FIGURE 12. Normalized average runtimes of the considered algorithms under both multiplicative and additive error models.

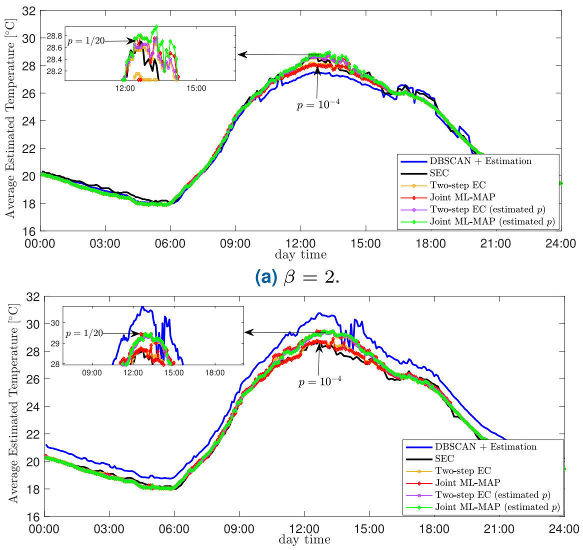


(a) snapshot of May 2022

(b) IPUGLIAT20 station

FIGURE 13. Anomalous temperature monitoring station near Torre dell'Orso, south of Italy, in the month of May 2022 (source: www.meteonetwork.eu).

For the sake of the analysis, we selected a subset of $N = 20$ monitoring stations (including the faulty one reported in Fig. 13b, whose public identifier is IPUGLIAT20) and retrieved the associated temperature measurements from the *WeatherUnderground* database, which is publicly available at https://www.wunderground.com. The monitoring stations provide streams of data starting from the midnight everyday, at regular intervals of 5 minutes. We processed all the measurements collected on May 23, consisting in a time series of 288 average temperatures from each station. The algorithms are applied to the whole time series and return an estimate of the average temperature over time, with associated classification of each monitoring station as either regular or anomalous. Unless otherwise specified, the setting of the algorithms parameters are the same as those in Sec. VI-B. It is worth remarking that the processed data are thus real measurements carried by the temperature sensors installed on the fixed monitoring stations, hence they do not follow



(a) $\beta = 2$.

(b) $\beta = 4$.

FIGURE 14. Average estimated temperature as a function of the daytime using the algorithms designed for the multiplicative error model, for (a) $\beta = 2$ and (b) $\beta = 4$.

any stochastic generative model (i.e., they do not match the Gaussian mixture models assumed ad the design stage).

1) RESULTS WITH ALGORITHMS FOR MULTIPLICATIVE ERROR MODEL

a: TEMPERATURE ESTIMATION PERFORMANCE

We start the analysis by testing the algorithms designed for the multiplicative error model. In Fig. 14 we report the average estimated temperature at different daytime for two values of the standard deviation parameter β encoding the entity of the anomaly in the multiplicative model. More specifically, Fig. 14a shows the resulting estimates when a value of $\beta = 2$ is used. As it can be noticed, all the algorithms generally experience a similar trend, except for the window between 10:00 and 17:00, which coincides with the hottest hours of the day. In that time window, measurements coming from the stations can be rather different among each other owing to fluctuations caused by different local environmental conditions (e.g., mitigation effects due to the vicinity of the sea). Since the anomalous station returned temperatures even 10/15 degrees lower than those measured by the other stations, a value of $\beta = 2$ may be not sufficient to correctly capture the actual entity of the fault. The consequent effects are visible on the resulting curves: the DBSCAN-based approach, for instance, tends to underestimate the average temperature, with a curve falling below that of the SEC algorithm. The two Bayesian algorithms are instead able to capture variations in the temperature in a more accurate manner when p_d is chosen close to the actual value of $p = 1/20$ (or when p is estimated from data), though their curves are only slightly better than that provided by the SEC algorithm.

In Fig. 14b, we report the average estimated temperatures when a value of $\beta = 4$ is used. Since this choice can more accurately capture the deviations present in the anomalous measurements, the two Bayesian algorithms begin to unveil their potentials: as it can be seen, their curves become visibly better than those of the SEC and DBSCAN-based approaches, especially for p_d close to the actual p or when p is inferred from data. The DBSCAN-based approach tends to overestimate the actual values of the temperature, providing in turn a non-smooth curve with notable spiky fluctuations.

TABLE 1. Accuracy of temperature sensors classification for the algorithms based on the multiplicative error model.

Algorithms		Accuracy (%)	
		$\beta = 2$	$\beta = 4$
SEC		92.15	
DBSCAN & Estim.		81.9	92.3
Two-step EC	$p_d = 10^{-4}$	92.7	93.4
	$p_d = 1/20$	95.5	99.3
	estimated p	94.3	98.6
Joint ML-MAP	$p_d = 10^{-4}$	93.2	94.1
	$p_d = 1/20$	95.7	99.6
	estimated p	94.8	99.1

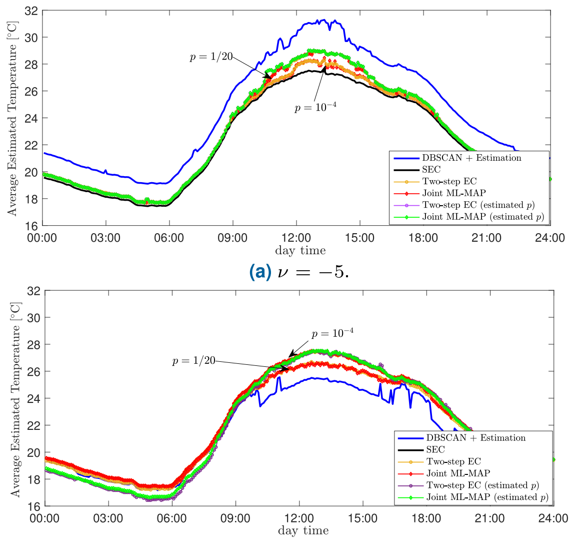
b: MONITORING STATIONS CLASSIFICATION ACCURACY

In Table 1 we report the accuracy (expressed in percentage) achieved by each algorithm in terms of correct classification of temperature sensors. The SEC algorithm provides an accuracy of 92.15%, while the DBSCAN-based approach achieves at most 92.3% when a value of $\beta = 4$ is used. Interestingly, the joint ML-MAP algorithms provides the best accuracy for all the possible cases, with values varying between 93.2% in the less convenient setting of $\beta = 2$ and $p_d = 10^{-4}$, up to about 99.6% when $p_d = 1/20$ and $\beta = 4$. As for the two-step EC algorithm, it also outperforms both DBSCAN-based and SEC algorithms, with an accuracy that is about 95.5% when a $p_d = 1/20$ and a $\beta = 2$ are assumed, and increases up to 99.3% when a value of $\beta = 4$ that better capture the entity of the anomaly is considered. Remarkably, when p is inferred from data, its accuracy ranges from 94.3% (for $\beta = 2$) up to 98.6% (for $\beta = 4$). This analysis confirms the same findings of Sec. VI-B, with the two step EC algorithm offering overall performance very close to that of the optimal joint ML-MAP approach.

2) RESULTS WITH ALGORITHMS FOR ADDITIVE ERROR MODEL

a: TEMPERATURE ESTIMATION PERFORMANCE

Fig. 15 shows the average estimated temperature as a function of the daytime when the algorithms designed for the additive error model are applied to the real measurements dataset, for two different values of the mean parameter ν (which encodes the entity of the anomaly in the additive model). In Fig. 15a, we report the average estimated temperature for a value of $\nu = -5$. It should be noticed that, differently from the multiplicative model, the algorithms designed for the additive model are also affected by the sign of the anomaly, being the latter encoded through the mean of the additive



(a) $\nu = -5$.

(b) $\nu = 2$.

FIGURE 15. Average estimated temperature as a function of the daytime using the algorithms designed for the additive error model, for (a) $\nu = -5$ and (b) $\nu = 2$.

TABLE 2. Accuracy of temperature sensors classification for the algorithms based on the additive error model.

Algorithms		**Accuracy (%)**	
		$\nu = -5$	$\nu = 2$
SEC		95	
DBSCAN & Estim.		87	81.9
Two-step EC	$p_d = 10^{-4}$	98	96
	$p_d = 1/20$	96.4	95.2
	estimated p	97.1	95.6
Joint ML-MAP	$p_d = 10^{-4}$	98.8	96.3
	$p_d = 1/20$	96.7	95.8
	estimated p	98.1	96.2

error term in (2). For this choice, the Bayesian algorithms are "informed" with a correct sign for the anomaly (being the measurements from the faulty station underestimates of the actual temperature), but with a magnitude that does not completely capture the whole errors (deviations can even achieve 10/15 degrees, as previously shown in Fig. 13). Remarkably, the two Bayesian algorithms are able to provide very good performance for values of p_d close to the actual $p = 1/20$ or when p is estimated from data, with curves that correctly estimate the average temperatures even in the hottest and most dynamic hours of the day. On the other hand, the SEC algorithm performs a too severe smoothing, resulting in a curve that does not follows all the temperature variations over time. As to the DBSCAN-based approach, it again tends to overestimate the average temperature and exhibits a curve with evident abrupt fluctuations.

Fig. 15b shows the average estimated temperature when a rather wrong value of $\nu = 2$ is fed to the algorithms. It can be seen that the DBSCAN-based approach has performance similar to the case of Fig. 14a, being the value of $|\nu|/2$ (analogously to $\beta/2$) used to set its search radius insensitive to the sign of the anomaly. On the other hand, the two Bayesian algorithms suffer from a more severe mismatch on the assumed parameter and tend to have better performance for a value of $p_d = 10^{-4}$. This behavior can be explained by observing that by setting $\nu = 2$, the Bayesian algorithms follow a model whose corresponding parameters do not match the actual situation found in the data; therefore, they tend to be blind with respect to an anomaly characterized by a mean value significantly different (taking into account also the sign) from $\nu = 2$, even if the probability of having a faulty sensor p_d is set close to the actual one.

b: MONITORING STATIONS CLASSIFICATION ACCURACY

To conclude the analysis, in Table 2 we report the sensors classification accuracy of the considered algorithms, under the different parameters settings. In this case, the DBSCAN-based approach provides accuracy levels that do not exceed 87%, whereas the SEC algorithm guarantees 95% accuracy. It should be however remarked that the latter approach always decides for the normal operational behavior of all sensors, which for the specific case at hand holds true 95% of the time (there is only a single faulty sensor in the network with dimension $N = 20$). Clearly, its accuracy would dramatically decrease when a more significant percentage of anomalous sensors appear in the network. Remarkably, the two Bayesian algorithms outperform the competitors for all the considered configurations. More specifically, the two-step EC approach has an accuracy ranging from a minimum value of 95.2%, obtained despite the rather erroneous value for the anomaly parameter $\nu = 2$ (and $p_d = 1/20$), up to about 98% for $\nu = -5$ and $p_d = 10^{-4}$. Also in this case, the joint ML-MAP approach provides the best performance, with accuracy greater than 96% in most of the cases. Notably, when p is inferred from data, the two Bayesian algorithms guarantee an accuracy of 97.1% (two-step EC) and 98.1% (joint ML-MAP), respectively.

VII. CONCLUSION

This paper addressed the problem of joint estimation and anomaly detection in environmental sensor networks, starting from possibly unreliable measurements of a common physical quantity of interest. The problem has been formulated within a novel unified Bayesian framework, accounting for two general error models that capture different types of anomalies in the measurement process. The optimal joint ML-MAP estimators have been illustrated for both models, and novel reduced-complexity two-step EC algorithms have been presented. The novel approaches employ a MAP classifier to make decisions about the sensor states and identify the presence of anomalies, while the estimation task is based on the ML criterion and may vary in complexity depending on the availability of some prior information about the probability of fault occurrence. Remarkably, the novel two-step EC approaches attain almost the same performance of the optimal joint ML-MAP estimator (which represents the benchmark for the problem at hand), but at a fraction of its computational complexity, as demonstrated by the theoretical cost analysis. The obtained results revealed that the proposed algorithms provide very low estimation errors, high accuracy, excellent scalability, and satisfactory robustness to potential model mismatches and to increasing percentages of faulty nodes in the network, on both synthetic and real-world data.

REFERENCES

[1] N. Q. Pham, V. P. Rachim, and W.-Y. Chung, "EMI-free bidirectional real-time indoor environment monitoring system," *IEEE Access*, vol. 7, pp. 5714–5722, 2018.

[2] A. Kumar and N. P. Pathak, "Wireless monitoring of volatile organic compounds/water vapor/gas pressure/temperature using RF transceiver," *IEEE Trans. Instrum. Meas.*, vol. 67, no. 9, pp. 2223–2234, Sep. 2018.

[3] F. Adamo, F. Attivissimo, C. G. C. Carducci, and A. M. L. Lanzolla, "A smart sensor network for sea water quality monitoring," *IEEE Sensors J.*, vol. 15, no. 5, pp. 2514–2522, May 2015.

[4] A. Boubrima, W. Bechkit, and H. Rivano, "Optimal WSN deployment models for air pollution monitoring," *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 2723–2735, May 2017.

[5] T. Li, H. Shen, C. Zeng, and Q. Yuan, "A validation approach considering the uneven distribution of ground stations for satellite-based $PM_{2.5}$ estimation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 1312–1321, 2020.

[6] F. Gasparin, E. Greiner, J.-M. Lellouche, O. Legalloudec, G. Garric, Y. Drillet, R. Bourdallé-Badie, P.-Y.-L. Traon, E. Rémy, and M. Drévillon, "A large-scale view of oceanic variability from 2007 to 2015 in the global high resolution monitoring and forecasting system at Mercator Océan," *J. Mar. Syst.*, vol. 187, pp. 260–276, Nov. 2018.

[7] S. L. Ullo and G. R. Sinha, "Advances in smart environment monitoring systems using IoT and sensors," *Sensors*, vol. 20, no. 11, p. 3113, May 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/11/3113

[8] G. Tmušić, S. Manfreda, H. Aasen, M. R. James, G. Gonçalves, E. Ben-Dor, A. Brook, M. Polinova, J. J. Arranz, J. Mészáros, R. Zhuang, K. Johansen, Y. Malbeteau, I. P. de Lima, C. Davids, S. Herban, and M. F. McCabe, "Current practices in UAS-based environmental monitoring," *Remote Sens.*, vol. 12, no. 6, p. 1001, Mar. 2020. [Online]. Available: https://www.mdpi.com/2072-4292/12/6/1001

[9] A. Capponi, C. Fiandrino, B. Kantarci, L. Foschini, D. Kliazovich, and P. Bouvry, "A survey on mobile crowdsensing systems: Challenges, solutions, and opportunities," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 2419–2465, 3rd Quart., 2019.

[10] A. Fascista, "Toward integrated large-scale environmental monitoring using WSN/UAV/crowdsensing: A review of applications, signal processing, and future perspectives," *Sensors*, vol. 22, no. 5, p. 1824, Feb. 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/5/1824

[11] A. Mahapatro and P. M. Khilar, "Fault diagnosis in wireless sensor networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 4, pp. 2000–2026, 4th Quart., 2013.

[12] L. Deng, X. Hao, Z. Mao, Y. Yan, J. Sun, and A. Zhang, "A subband radiometric calibration method for UAV-based multispectral remote sensing," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 8, pp. 2869–2880, Aug. 2018.

[13] W. Feng, Z. Yan, H. Zhang, K. Zeng, Y. Xiao, and Y. T. Hou, "A survey on security, privacy, and trust in mobile crowdsourcing," *IEEE Internet Things J.*, vol. 5, no. 4, pp. 2971–2992, Aug. 2018.

[14] S. Bhattacharjee, N. Ghosh, V. K. Shah, and S. K. Das, "Qn: Quality and quantity based unified approach for secure and trustworthy mobile crowdsensing," *IEEE Trans. Mobile Comput.*, vol. 19, no. 1, pp. 200–216, Jan. 2020.

[15] T. Strutz, *Data Fitting Uncertainty: A Practical Introduction to Weighted Least Squares Beyond*. Cham, Switzerland: Springer, 2011.

[16] P. J. Rousseeuw and A. M. Leroy, *Robust Regression Outlier Detection*. Hoboken, NJ, USA: Wiley, 2005.

[17] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. 2nd Int. Conf. Knowl. Discovery Data Mining*, 1996, pp. 226–231.

[18] L. Carlone, A. Censi, and F. Dellaert, "Selecting good measurements via ℓ_1 relaxation: A convex approach for robust estimation over graphs," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Sep. 2014, pp. 2667–2674.

[19] K. Huang and K. Yuen, "Hierarchical outlier detection approach for online distributed structural identification," *Struct. Control Health Monitor.*, vol. 27, no. 11, Nov. 2020, Art. no. e2623, doi: 10.1002/stc.2623.

[20] H.-Q. Mu and K.-V. Yuen, "Novel outlier-resistant extended Kalman filter for robust online structural identification," *J. Eng. Mech.*, vol. 141, no. 1, Jan. 2015, Art. no. 04014100.

[21] H.-Q. Mu, S.-C. Kuok, and K.-V. Yuen, "Stable robust extended Kalman filter," *J. Aerosp. Eng.*, vol. 30, no. 2, 2017, Art. no. B4016010, doi: 10.1061/(ASCE)AS.1943-5525.0000665.

[22] D. D. Palma and G. Indiveri, "Output outlier robust state estimation," *Int. J. Adapt. Control Signal Process.*, vol. 31, no. 4, pp. 581–607, Apr. 2017, doi: 10.1002/acs.2673.

[23] D. D. Palma and G. Indiveri, "Outlier robust state estimation through smoothing on a sliding window," *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 14636–14641, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405896320318851

[24] S. Li and X. Wang, "Optimal joint detection and estimation based on decision-dependent Bayesian cost," *IEEE Trans. Signal Process.*, vol. 64, no. 10, pp. 2573–2586, May 2016.

[25] B. Dulek, "A restricted Bayes approach to joint detection and estimation under prior uncertainty," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 54, no. 4, pp. 1767–1782, Aug. 2018.

[26] Y. Xu, Z. Wu, J. Chanussot, and Z. Wei, "Joint reconstruction and anomaly detection from compressive hyperspectral images using Mahalanobis distance-regularized tensor RPCA," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2919–2930, May 2018.

[27] A. Yeredor, "The joint MAP-ML criterion and its relation to ML and to extended least-squares," *IEEE Trans. Signal Process.*, vol. 48, no. 12, pp. 3484–3492, Dec. 2000.

[28] X. Rong Li, "Optimal Bayes joint decision and estimation," in *Proc. 10th Int. Conf. Inf. Fusion*, Jul. 2007, pp. 1–8.

[29] F. Fagnani, S. M. Fosson, and C. Ravazzi, "Consensus-like algorithms for estimation of Gaussian mixtures over large scale networks," *Math. Models Methods Appl. Sci.*, vol. 24, no. 2, pp. 381–404, Feb. 2014.

[30] A. Chiuso, F. Fagnani, L. Schenato, and S. Zampieri, "Gossip algorithms for simultaneous distributed estimation and classification in sensor networks," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 4, pp. 691–706, Aug. 2011.

[31] F. Fagnani, S. M. Fosson, and C. Ravazzi, "A distributed classification/estimation algorithm for sensor networks," *SIAM J. Control Optim.*, vol. 52, no. 1, pp. 189–218, Jan. 2014.

[32] C. Ravazzi, N. P. K. Chan, and P. Frasca, "Distributed estimation from relative measurements of heterogeneous and uncertain quality," *IEEE Trans. Signal Inf. Process. Over Netw.*, vol. 5, no. 2, pp. 203–217, Jun. 2019.

ALESSIO FASCISTA (Member, IEEE) received the Ph.D. degree in engineering of complex systems from the University of Salento, Lecce, Italy, in 2019. He has held a visiting position at the Department of Telecommunications and Systems Engineering, Universitat Autonoma de Barcelona (UAB), Spain, in 2018, and the Department of Electrical Engineering, Chalmers University of Technology, Gothenburg, Sweden, in 2022. He is currently an Assistant Professor of telecommunications with the Department of Innovation Engineering, University of Salento. His main research interests include telecommunications with focus on statistical signal processing for detection, estimation, and localization in terrestrial wireless systems. He serves as an Associate Editor for the IEEE Open Journal of the Communications Society.

ANGELO COLUCCIA (Senior Member, IEEE) received the Ph.D. degree in information engineering, in 2011. He has been a Research Fellow at Forschungszentrum Telekommunikation Wien, Vienna, Austria. He has held a visiting position with the Department of Electronics, Optronics, and Signals, Institut Supérieur de l'Aéronautique et de l'Espace (ISAE-Supaero), Toulouse, France. He is currently an Associate Professor of telecommunications with the Department of Engineering, University of Salento, Lecce, Italy. His research interests include multi-channel, multi-sensor, and multi-agent statistical signal processing for detection, estimation, localization, and learning problems. Relevant application fields are radar, wireless networks (including 5G and beyond), and emerging network contexts (including intelligent cyber-physical systems, smart devices, and social networks). He is a member of the Technical Area Committee in Signal Processing for Multisensor Systems of EURASIP.

CHIARA RAVAZZI (Member, IEEE) received the Ph.D. degree in mathematics for engineering sciences from the Politecnico di Torino, in 2011. She was a Visiting Member at the Massachusetts Institute of Technology (LIDS), in 2010, and a Postdoctoral Researcher at the Politecnico di Torino (DISMA, DET), from 2011 to 2016. Since 2017, she has been a Tenured Researcher of the National Research Council (CNR-IEIIT), Italy. Her current research interests include control and information theory, signal processing, optimization, and learning algorithms for network systems. She has been serving as an Associate Editor for the IEEE Transactions on Signal Processing, since 2019, and the IEEE Transactions on Control Systems Letters, since 2021.