Differentially Private Outlier Detection in Multivariate Gaussian Signals

Kwassi H. Degue, Karthik Gopalakrishnan, Max Z. Li, Hamsa Balakrishnan, Jerome Le Ny

***Abstract*— The detection of outliers in data, while preserving the privacy of individual agents who contributed to the data set, is an increasingly important task when monitoring and controlling large-scale systems. In this paper, we use an algorithm based on the sparse vector technique to perform differentially private outlier detection in multivariate Gaussian signals. Specifically, we derive analytical expressions to quantify the trade-off between detection accuracy and privacy. We validate our analytical results through numerical simulations.**

***Index Terms*— Differential privacy; Outlier detection; Data storage systems; Pattern recognition and classification.**

I. Introduction

The identification of outliers in a data set plays a major role in the monitoring and control of intelligent systems such as transportation networks, power grids, and other urban infrastructures. The underlying data, however, often consist of privacy-sensitive information such as the real-time locations [1] or identity-revealing characteristics of individuals. It is therefore necessary to implement privacy-preserving mechanisms when sensitive data is shared for the purposes of inference and control. A privacy guarantee incentivizes agents to participate truthfully, allowing for accurate outlier detection and subsequent control actions. Other examples that motivate the need for privacy in control systems can be found in [2]–[5].

Many privacy-preserving data analysis techniques adopt the notion of *differential privacy* [6]–[8], which provides a much stronger privacy guarantee than anonymization techniques such as k-anonymity [9]. Differential privacy guarantees that the participation or absence of an individual agent does not significantly alter the output of any query (e.g., "is $\sum_i x_i$ an outlier?"). Differential privacy can be achieved through *input* (i.e., random noise is added to $x_i, \forall i$) or *output perturbation* (i.e., random noise is added to $\sum_i x_i$) [7]. Higher noise provides more privacy at the cost of query accuracy. A more sophisticated method is the sparse vector technique (SVT) [6], [10], which provides, for certain types of queries, higher accuracy for the same level of privacy.

Several differentially private algorithms have been proposed for classical hypothesis testing. For example, [11] and [12] assume categorical data that follow a multinomial distribution in order to prove the privacy properties of their proposed algorithms. Differential privacy has also been considered in the context of anomaly detection, using Monte Carlo (MC) [11], [12] and machine learning-based [13] techniques. [14] and [15] propose statistical tests for normally distributed data under differential privacy constraints to decide whether or not the mean of a sequence of scalar, independent, and identically distributed (i.i.d.) Gaussian random variables attains a given value. Furthermore, [16] proposes a differentially private mechanism to detect distributional changes at an unknown change-point in a sequence of scalar i.i.d. random variables. However, these prior works assume that the privacy-sensitive data provided by each individual are i.i.d., which may not be the case in data generated from networked systems where individuals' data are correlated [5], [8], [17].

Our contribution in this paper is the design and analysis of SVT for detecting outliers in multivariate Gaussian signals. This setting considers agents whose signals may be correlated. Our approach differs from prior work in two ways: Unlike the Monte Carlo approaches presented in [11], [12] and the machine learning-based approach of [13], we derive analytic expressions for the accuracy of a differentially private mechanism; unlike [14]–[16] that only consider i.i.d. scalar quantities, we explore multivariate correlated data.

The remainder of the paper is organized as follows: We

K. H. Degue and J. Le Ny were supported in part by NSERC under Grant RGPIN-5287-2018 and RGPAS-2018-522686, by the Mitacs Globalink Research Award and by a doctoral scholarship of the FRQNT. M. Z. Li was supported in part by a NSF fellowship. K. Gopalakrishnan, M. Z. Li, and H. Balakrishnan were supported in part by NSF CPS Award No. 1739505. The NASA University Leadership Initiative (grant #80NSSC20M0163) provided funds to assist the MIT authors and K. H. Degue with their research, but this article solely reflects the opinions and conclusions of its authors and not any NASA entity.

K. H. Degue and J. Le Ny are with the Department of Electrical Engineering, Polytechnique Montreal and GERAD, QC H3T-1J4, Montreal, Canada. K. Gopalakrishnan, M. Z. Li, and H. Balakrishnan are with the Department of Aeronautics and Astronautics at the Massachusetts Institute of Technology, Cambridge, MA. This work was done while the first author was visiting MIT. {kwassi-holali.degue,jerome.le-ny}@polymtl.ca, {karthikg, maxli, hamsa}@mit.edu

set up the problem in Sec. II, present our main results in Sec. III, validate our approach with numerical simulations in Sec. IV, and provide concluding remarks in Sec. V.

II. Problem Statement

A. Notation

A generic probability triple is denoted $(\Omega, \mathcal{F}, \mathbb{P})$, where $\mathcal{F}$ stands for a σ-algebra on the sample space Ω, and $\mathbb{P}$ is a probability measure defined on $\mathcal{F}$. The ℓ_p-norm of a vector $\mathbf{x} \in \mathbb{R}^n$ is denoted by $\|\mathbf{x}\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$, for $p \in [1, \infty]$; we use $|\cdot|$ for absolute values. We denote an n-dimensional Gaussian vector with mean $\boldsymbol{\mu} = [\mu_i] \in \mathbb{R}^{n\times 1}$ and covariance $\Sigma \in \mathbb{S}_{\succeq 0}^{n\times n}$ as $\mathbf{X} = [X_i] \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{\mu}, \Sigma)$. We denote by $\mathrm{Lap}(b)$ a zero-mean Laplace distribution with variance $2b^2$ and probability density function $f(x; b) = \frac{1}{2b}\exp\left(-\frac{|x|}{b}\right)$. Let $\mu = \mathbb{E}\left[\sum_{i=1}^n X_i\right] = \mathbf{1}^\intercal \boldsymbol{\mu}$, $\sigma^2 = \mathrm{Var}\left[\sum_{i=1}^n X_i\right] = \mathbf{1}^\intercal \Sigma \mathbf{1}$, and $\sigma = \sqrt{\mathbf{1}^\intercal \Sigma \mathbf{1}}$. When μ, σ, and σ^2 have no indices, they refer respectively to the mean, standard deviation, and variance of the sum of the elements in $\mathbf{X}$, i.e., $\sum_{i=1}^n X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$. We denote the error function by $\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2} dt$, and the complementary error function by $\mathrm{erfc}(x) = 1 - \mathrm{erf}(x)$.

B. Outlier Detection Problem

Consider a sequence of m observations $\mathcal{O}_m \coloneqq \left\{\mathbf{x}^{(k)}\right\}_{k=1}^{k=m}$, with $\mathbf{x}^{(k)} = \left[x_i^{(k)}\right] \in \mathbb{R}^{n\times 1}$. For example, each $\mathbf{x}^{(k)}$ can represent a vector at each time k, and $x_i^{(k)}$ can represent the value corresponding to an individual in the vector at the time k. The data $\mathcal{O}_m$ is observed sequentially. Assume that the signal vectors $\mathbf{x}^{(k)}$ are realizations of $\mathbf{X} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, meaning that the samples across k are i.i.d. However, individual values across each vector $\mathbf{x}^{(k)}$ are correlated. We design a statistical test to check whether or not there is an outlier in terms of the magnitude of the signal. We first formally define the term *outlier*, as used throughout this article.

Definition 1. *An observation $\mathbf{x}^{(k)}$ is labeled as an outlier of level κ if $\sum_{i=1}^n x_i^{(k)} \notin [\mu - \kappa\sigma,\ \mu + \kappa\sigma]$.*

In other words, we label an observation as an outlier of level κ if the sum of its elements is at least κ standard deviations away from its expected value. This notion of outliers captures the impact of the signal magnitude. Accordingly, we map observations in the data set $\mathcal{O}_m$ to one of the two following hypotheses:

$$\begin{cases} \mathrm{H}_0: & \left|\sum_{i=1}^n x_i^{(k)} - \mu\right| < h : \mathbf{x}^{(k)} \text{ is not an outlier} \\ \mathrm{H}_1: & \left|\sum_{i=1}^n x_i^{(k)} - \mu\right| \geq h : \mathbf{x}^{(k)} \text{ is an outlier} \end{cases}$$

where the threshold is $h = \kappa\sigma$. Consequently, we can compute the following decision rule:

$$d\left(\mathbf{x}^{(k)}\right) = \begin{cases} 0 & \text{if } q\left(\mathbf{x}^{(k)}\right) < h : \mathrm{H}_0 \text{ is chosen} \\ 1 & \text{if } q\left(\mathbf{x}^{(k)}\right) \geq h : \mathrm{H}_1 \text{ is chosen,} \end{cases} \tag{1}$$

where

$$q\left(\mathbf{x}^{(k)}\right) = \left|\sum_{i=1}^n x_i^{(k)} - \mu\right|. \tag{2}$$

We note that rule (1) determines whether or not an observation is an outlier, and it depends on the data set $\mathcal{O}_m$. In this article, we consider cases in which the data set $\mathcal{O}_m$ is privacy-sensitive. As we discuss in the next subsection, our goal is to publish the results of outlier detection under a differential privacy constraint. In order to satisfy the differential privacy requirement, we will modify the decision rule (1) in Sec. III. In the rest of this section, we briefly review the concepts of differential privacy and differentially private mechanisms [18].

C. Differential Privacy

Consider a space $\mathcal{H}$ of data sets. Throughout this article, we have that $\mathcal{H} \equiv \mathbb{R}^{n\times m}$ denotes the space containing the observation sequence $\mathcal{O}_m$. A *mechanism* M is defined as a random map from $\mathcal{H}$ to some measurable output space. The goal of a differentially private mechanism is to produce outputs with similar distributions for inputs that we wish to make indistinguishable [6].

We define a symmetric binary relation Adj on $\mathcal{H}$, called adjacency, to describe which inputs are considered "close" in some sense. For example, two inputs are termed *adjacent* if all the entries are the same for all individuals, except for at most one entry corresponding to one individual, that has a bounded difference. More formally, two sequences of observations $\mathcal{O}_m \coloneqq \left\{\mathbf{x}^{(k)}\right\}_{k=1}^{k=m}$ and $\tilde{\mathcal{O}}_m \coloneqq \left\{\tilde{\mathbf{x}}^{(k)}\right\}_{k=1}^{k=m}$ are adjacent if, and only if:

$$\begin{aligned} &\left|x_i^{(k)} - \tilde{x}_i^{(k)}\right| \leq \rho^{(k)}, \text{for some } 1 \leq k \leq m \text{ and } 1 \leq i \leq n, \\ &\text{and } x_j^{(\ell)} = \tilde{x}_j^{(\ell)}, \text{ for all } \ell \neq k \text{ and } j \neq i, \end{aligned} \tag{3}$$

with $\left\{\rho^{(k)}\right\}_{k=1}^m \in \mathbb{R}_{>0}^m$ a given set of positive numbers. If $\mathcal{O}_m$ and $\tilde{\mathcal{O}}_m$ are adjacent, we say $\mathrm{Adj}(\mathcal{O}_m, \tilde{\mathcal{O}}_m)$. In other words, two observed sequences are adjacent if and only if they differ only by the value of a single element $x_i^{(k)}$ of a single vector $\mathbf{x}^{(k)}$, with bounded deviations in the value of that element. In what follows, we denote $\rho = \max_{1\leq k\leq m} \rho^{(k)}$.

Next, we provide the formal definition of differential privacy as presented in [18], [19].

Definition 2. *Consider $\mathcal{H}$, a space provided with a symmetric binary relation denoted* Adj*, and let $(\mathcal{P}, \mathcal{M})$ be a*

measurable space, where $\mathcal{M}$ is a given σ-algebra over $\mathcal{P}$. Let $\epsilon \geq 0$. A randomized mechanism M from $\mathcal{H}$ to $\mathcal{P}$ is ϵ-differentially private (for Adj*) if the following property holds for all $\mathcal{O}_m, \tilde{\mathcal{O}}_m \in \mathcal{H}$ such that* $\mathrm{Adj}(\mathcal{O}_m, \tilde{\mathcal{O}}_m)$*, for all sets S in $\mathcal{M}$:*

$$\mathbb{P}\left(M\left(\mathcal{O}_m\right) \in S\right) \leq e^{\epsilon}\, \mathbb{P}\left(M\left(\tilde{\mathcal{O}}_m\right) \in S\right). \qquad (4)$$

Note that (4) implies that the distributions of the random variables $M\left(\mathcal{O}_m\right)$ and $M\left(\tilde{\mathcal{O}}_m\right)$ are close when $\mathcal{O}_m$ and $\tilde{\mathcal{O}}_m$ are adjacent. We now define a quantity that plays a key role in the design of differentially private mechanisms.

Definition 3. *Consider a space of data sets $\mathcal{H}$ with an adjacency relation* Adj*, and let $\mathcal{P}$ be a vector space with norm $\|\cdot\|_{\mathcal{P}}$. The sensitivity of a query $q : \mathcal{H} \mapsto \mathcal{P}$ is the quantity $\Delta_{\mathcal{P}q} := \sup_{\{\mathcal{O}_m, \tilde{\mathcal{O}}_m \,:\, \mathrm{Adj}(\mathcal{O}_m, \tilde{\mathcal{O}}_m)\}} \left\| q\left(\mathcal{O}_m\right) - q\left(\tilde{\mathcal{O}}_m\right)\right\|_{\mathcal{P}}$. In particular, when $\mathcal{P} = \mathbb{R}^n$ (with $n = +\infty$ being a possibility), and given the p-norm for $p \in [1, \infty]$, this definition of $\Delta_{\mathcal{P}q}$ is called the ℓ_p-sensitivity. For notational brevity, we simply write Δ instead of $\Delta_{\mathcal{P}q}$ when the context is clear.*

In this article, we design a differentially private outlier detection algorithm for multivariate Gaussian signals, namely, an outlier detection algorithm which publishes a decision that is differentially private with respect to the adjacency relation (3) for queries on $\mathcal{O}_m$.

III. Differentially Private Detection of Outliers in Magnitude

Following the SVT as presented in [20], we design the following differentially private outlier detection algorithm for multivariate signals:

Algorithm 1: OutlierDetect$(\mathcal{O}_m, q(\cdot), h, \rho, \epsilon)$

Set $\Delta = \rho$
Compute noisy threshold $\widetilde{h} = h + \mathrm{Lap}\left(\frac{2\Delta}{\epsilon}\right)$
$d\left(\mathbf{x}^{(k)}\right) \leftarrow 0, \quad \forall k$
for *each query* k **do**
  Compute $\zeta^{(k)} \sim \mathrm{Lap}\left(\frac{4\Delta}{\epsilon}\right)$
  if $q\left(\mathbf{x}^{(k)}\right) + \zeta^{(k)} \geq \widetilde{h}$ **then**
    $d\left(\mathbf{x}^{(k)}\right) = 1$
$\mathcal{K} \leftarrow \sum_{k=1}^{m} d\left(\mathbf{x}^{(k)}\right)$
Return $\mathcal{K}$, $\left\{d\left(\mathbf{x}^{(k)}\right)\right\}_{k=1}^{k=m}$

Theorem 1. OutlierDetect$(\mathcal{O}_m, q(\cdot), h, \rho, \epsilon)$ *is $\frac{\mathcal{K}+1}{2}\epsilon$-differentially private for the sequence of queries $\left\{q\left(\mathbf{x}^{(k)}\right)\right\}_{k=1}^{k=m}$ as defined in* (2) *and adjacency relation* (3).

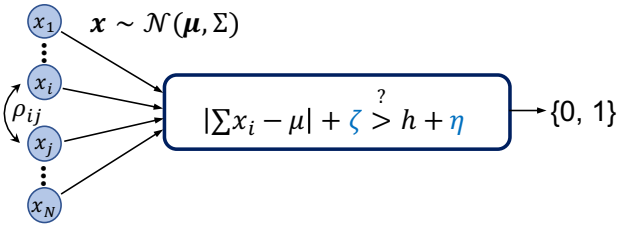

Fig. 1. Differentially private outlier detection algorithm for each observation vector $\mathbf{x}^{(k)} = \left[x_i^{(k)}\right] \in \mathbb{R}^{n\times 1}$. We omit k for brevity.

Proof. For two observation sequences $\mathcal{O}_m$ and $\tilde{\mathcal{O}}_m$ with an adjacency relation defined in (3), the sensitivity can be bounded as follows:

$$\begin{aligned}\Delta &= \sup_{\substack{1\leq k\leq m \\ \mathcal{O}_m, \tilde{\mathcal{O}}_m \,:\, \mathrm{Adj}\left(\mathcal{O}_m, \tilde{\mathcal{O}}_m\right)}} \left| q\left(\mathbf{x}^{(k)}\right) - q\left(\tilde{\mathbf{x}}^{(k)}\right)\right| \\ &= \sup_{\substack{1\leq k\leq m \\ \mathcal{O}_m, \tilde{\mathcal{O}}_m \,:\, \mathrm{Adj}\left(\mathcal{O}_m, \tilde{\mathcal{O}}_m\right)}} \left|\left|\sum_{i=1}^{n} x_i^{(k)} - \mu\right| - \left|\sum_{i=1}^{n} \tilde{x}_i^{(k)} - \mu\right|\right|.\end{aligned}$$

By using the reverse triangle inequality, it follows that

$$\begin{aligned}\Delta &\leq \sup_{\substack{1\leq k\leq m \\ \mathcal{O}_m, \tilde{\mathcal{O}}_m \,:\, \mathrm{Adj}\left(\mathcal{O}_m, \tilde{\mathcal{O}}_m\right)}} \left|\sum_{i=1}^{n} x_i^{(k)} - \sum_{i=1}^{n} \tilde{x}_i^{(k)}\right| \\ &= \sup_{\substack{1\leq k\leq m \\ \mathcal{O}_m, \tilde{\mathcal{O}}_m \,:\, \mathrm{Adj}\left(\mathcal{O}_m, \tilde{\mathcal{O}}_m\right)}} \left|x_i^{(k)} - \tilde{x}_i^{(k)}\right| \leq \rho. \qquad (5)\end{aligned}$$

We deduce the result from the proof argument of [20, Theorem 1], and find that

$$\begin{aligned}\frac{\mathbb{P}(\text{OutlierDetect}(\mathcal{O}_m, q(\cdot), h, \rho, \epsilon) = d)}{\mathbb{P}(\text{OutlierDetect}(\tilde{\mathcal{O}}_m, q(\cdot), h, \rho, \epsilon) = d)} &\leq e^{\frac{\epsilon}{2}} \left(e^{\frac{\epsilon}{2}}\right)^{\mathcal{K}}, \\ &\leq e^{\frac{(\mathcal{K}+1)\epsilon}{2}}.\end{aligned}$$

□

Fig. 1 summarizes the OutlierDetect algorithm.

A. Performance Analysis

In this section, we characterize the privacy-utility trade-off of our privacy-preserving algorithm, OutlierDetect. Our analysis relies on the following calculation.

Proposition 1. Consider two independent random variables $Z_1 \sim \mathrm{Lap}\left(\frac{4\Delta}{\epsilon}\right)$ and $Z_2 \sim \mathrm{Lap}\left(\frac{2\Delta}{\epsilon}\right)$. The probability density function (pdf) of the difference $Z = Z_1 - Z_2$ can be computed as follows:

$$f_Z(z) = \frac{\epsilon}{12\Delta} e^{-\frac{|z|\epsilon}{4\Delta}} \left(2 - e^{-\frac{|z|\epsilon}{4\Delta}}\right). \qquad (6)$$

Proof. Since the Laplace pdf is symmetric about 0, the pdf of Z_2 is the same as pdf of $-Z_2$. Thus, the pdf of $Z = Z_1 + (-Z_2)$ is given by

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{\infty} f_{Z_1}(z-\tau) f_{Z_2}(\tau) d\tau \\ &= \frac{\epsilon^2}{32\Delta^2} \int_{-\infty}^{\infty} \exp\left(\frac{-\epsilon|z-\tau|}{4\Delta} - \frac{\epsilon|\tau|}{2\Delta}\right) d\tau. \end{aligned} \tag{7}$$

We evaluate (7) for two cases: $z \geq 0$ and $z < 0$. When $z \geq 0$, we split the integration limits in (7) into $(-\infty, 0] \cup [0, z] \cup [z, \infty)$ in order to remove the absolute values on $z - \tau$ and τ. Performing a similar decomposition of the integration domain for $z < 0$ and combining the two cases give the desired result in (6). □

Next, we give formal definitions of the classification errors that will be used to characterize the performance of the OUTLIERDETECT algorithm.

Error definitions

Two types of errors are important for any classification or hypothesis testing problem: Type I (or false positives) and Type II (or false negatives). In our case, the Type I error rate (P_I) is the probability that a nominal data point is classified incorrectly as an outlier by the OUTLIERDETECT algorithm, and the Type II error rate (P_{II}) is the probability that an outlier is classified incorrectly as nominal by OUTLIERDETECT. Complementary to P_I is the true negative rate given by $1 - P_I$; similarly, the true positive rate is given by $1 - P_{II}$. Figure 2 shows a geometric perspective of these four probabilities with respect to the threshold h, the query $q\left(\mathbf{x}^{(k)}\right)$, and noise Z drawn according to the density (6) from Proposition 1. Next, we use these error definitions to discuss the performance of the OUTLIERDETECT algorithm.

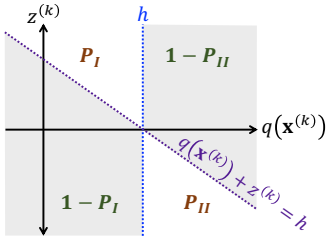

Fig. 2. Geometry of the detection probabilities for Algorithm OUTLIERDETECT $(\mathcal{O}_m, q(\cdot), h, \epsilon)$.

First, we need to derive the pdf of the queries $q\left(\mathbf{x}^{(k)}\right)$. For conciseness, let $f_Q(q)$ denote the pdf of $q\left(\mathbf{x}^{(k)}\right)$. The following proposition gives us an expression for $f_Q(q)$:

Proposition 2. Each query $q\left(\mathbf{x}^{(k)}\right)$ defined in (2) is a realization of a random variable Q whose pdf is

$$f_Q(q) = \frac{2}{\sqrt{2\pi\sigma^2}} e^{-\frac{q^2}{2\sigma^2}}, \text{ for } q \geq 0. \tag{8}$$

Proof. Note that $Q = |\sum_{i=1}^n X_i - \mu|$, where $\sum_{i=1}^n X_i \sim \mathcal{N}\left(\mu, \sigma^2\right)$. Equivalently, $Q = |Y|$, where $Y = \mathcal{N}\left(0, \sigma^2\right)$. Hence, $f_Q(q) = f_Y(q) + f_Y(-q)$, and the result follows. □

Next, in Theorem 2, we derive an analytical formula for the true positive rate $1 - P_{II}$ of the OUTLIERDETECT algorithm.

Theorem 2. *For each data point indexed by $k = 1, \ldots, m$, the algorithm* OUTLIERDETECT $(\mathcal{O}_m, q(\cdot), h, \rho, \epsilon)$ *achieves the following true positive probability:*

$$\begin{aligned} 1 - P_{II} = 1 &+ \frac{c}{6} e^{2a_1\epsilon + 4a_2\epsilon^2} \operatorname{erfc}\left(\frac{a_1}{2\sqrt{a_2}} + 2\epsilon\sqrt{a_2}\right) \\ &- \frac{2c}{3} e^{a_1\epsilon + a_2\epsilon^2} \operatorname{erfc}\left(\frac{a_1}{2\sqrt{a_2}} + \epsilon\sqrt{a_2}\right), \end{aligned}$$

with $c = 1/\operatorname{erfc}\left(h/\left(\sigma\sqrt{2}\right)\right)$, $a_1 = h/(4\rho)$, *and* $a_2 = \sigma^2/\left(32\rho^2\right)$.

Proof. The outlier detection algorithm gives a true positive result when $q\left(\mathbf{x}^{(k)}\right) + \zeta^{(k)} \geq h + \eta$ and $q\left(\mathbf{x}^{(k)}\right) \geq h$, with $\zeta^{(k)} \sim \text{Lap}\left(\frac{4\rho}{\epsilon}\right)$ and $\eta \sim \text{Lap}\left(\frac{2\rho}{\epsilon}\right)$.

Combining the noise terms, we can define $z^{(k)} = \zeta^{(k)} - \eta$, where $z^{(k)}$ has density (6) from Proposition 1. We then have the true positive rate, defined in terms of a conditional probability, as:

$$\begin{aligned} 1 - P_{II} &= \mathbb{P}\left(q\left(\mathbf{x}^{(k)}\right) + z^{(k)} \geq h \,\middle|\, q\left(\mathbf{x}^{(k)}\right) \geq h\right) \\ &= \mathbb{P}\left(Q + Z \geq h \,\middle|\, Q \geq h\right). \end{aligned}$$

Denoting the true positive region in Figure 2 (i.e., the $1 - P_{II}$ region) by $\mathcal{R}$, and using the fact that Q and Z are independent, we have:

$$1 - P_{II} = \frac{\iint_{\mathcal{R}} f_Q(q) f_Z(z)\, dz\, dq}{\int_h^\infty f_Q(q)\, dq}.$$

Expanding using Propositions 1 and 2, we get:

$$\begin{aligned} 1 - P_{II} = &\frac{\epsilon \int_h^\infty \int_0^\infty \left(2e^{\frac{z\epsilon}{4\rho}} - 1\right) e^{-\frac{q^2}{2\sigma^2} - \frac{z\epsilon}{2\rho}}\, dz\, dq}{6\rho\sigma\sqrt{2\pi} \int_h^\infty \frac{2}{\sigma\sqrt{2\pi}} e^{-\frac{q^2}{2\sigma^2}}\, dq} \\ &- \frac{\epsilon \int_h^\infty \int_{h-q}^0 \left(e^{\frac{z\epsilon}{4\rho}} - 2\right) e^{\frac{1}{4}\left(\frac{z\epsilon}{\rho} - \frac{2q^2}{\sigma^2}\right)}\, dz\, dq}{6\rho\sigma\sqrt{2\pi} \int_h^\infty \frac{2}{\sigma\sqrt{2\pi}} e^{-\frac{q^2}{2\sigma^2}}\, dq}. \end{aligned} \tag{9}$$

The rest of the proof involves algebraic simplification of Equation (9) in terms of erfc, and defining appropriate constants c, a_1 and a_2. □

Corollary 1. The true positive probability as derived in Theorem 2 is $1/2$ for $\epsilon \to 0$, and 1 for $\epsilon \to \infty$.

Proof. $\lim_{\epsilon\to 0}(1 - P_{II}) = 1/2$ follows from direct substitution. $\lim_{\epsilon\to\infty}(1 - P_{II}) = 1$ is obtained using L'Hôpital's rule and the fact that $\frac{d}{dx}\operatorname{erfc}(x) = -\frac{2}{\sqrt{\pi}}e^{-x^2}$. □

Theorem 3. *For each data point indexed by $k = 1, \ldots, m$, the algorithm* OUTLIERDETECT $(\mathcal{O}_m, q(\cdot), h, \rho, \epsilon)$ *incurs the following false positive probability:*

$$P_I = \frac{ce^{-2a_1\epsilon+a_2\epsilon^2}}{6(c-1)} \times \left[4e^{a_1\epsilon}\left(\operatorname{erf}\left(\epsilon\sqrt{a_2}\right) - \operatorname{erf}\left(-\frac{a_1}{2\sqrt{a_2}} + \epsilon\sqrt{a_2}\right)\right) - e^{3a_2\epsilon^2}\left(\operatorname{erf}\left(2\epsilon\sqrt{a_2}\right) - \operatorname{erf}\left(-\frac{a_1}{2\sqrt{a_2}} + 2\epsilon\sqrt{a_2}\right)\right)\right],$$

with constants c, a_1, and a_2 as defined in Theorem 2.

Proof. The proof is similar to that of Theorem 2. □

Corollary 2. The probability P_I of incurring a Type I error (as derived in Theorem 3) is $1/2$ for $\epsilon \to 0$, and is 0 for $\epsilon \to \infty$.

Proof. $\lim_{\epsilon\to 0} P_I = 1/2$ follows from direct substitution. The other limit, $\lim_{\epsilon\to\infty} P_I = 0$ is derived using L'Hôpital's rule and $\frac{d}{dx}\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}}e^{-x^2}$. □

B. Discussion

In the proposed OUTLIERDETECT algorithm, the differential privacy parameter ϵ directly affects the scale of the noise terms for both the query and the threshold. Corollaries 1 and 2 confirm our intuitions of the behavior of the true positive and false positive probabilities of OUTLIERDETECT at the two extreme regimes: infinite noise ($\epsilon \to 0$) and no noise ($\epsilon \to \infty$). In the former case, the ϵ-differentially private OUTLIERDETECT algorithm classifies no better than a random guess, with both $1 - P_{II} = P_{II} = 1/2$ and $P_I = 1 - P_I = 1/2$. In the latter case, the ϵ-differentially private variant of OUTLIERDETECT behaves exactly like its non-differentially private counterpart, so no Type I nor Type II errors are expected. In other words, for the case with no noise, we have that $P_I = P_{II} = 0$, and the true positive and true negative probabilities are both equal to 1.

IV. NUMERICAL SIMULATIONS

We generate a data set of residuals $\mathcal{O}_m$ with $m = 1,000$, $\mathbf{x}^{(k)} \in \mathbb{R}^{30\times 1}$ with $\mathbf{x}^{(k)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, and $\mathbf{1}^\intercal\boldsymbol{\mu} = 1.73 \times 10^4, \mathbf{1}^\intercal\Sigma\mathbf{1} = 3.01 \times 10^7$. We set a sensitivity of $\Delta = \rho = 500$. We first compare the analytical and empirical performance of Theorems 2 and 3, using $\mathcal{O}_m$. We set $h \approx$

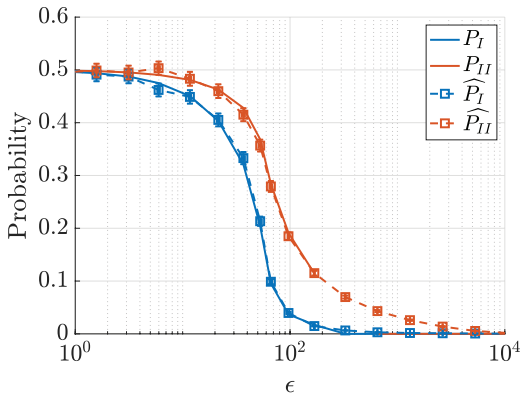

Fig. 3. P_I and P_{II} are direct evaluations of the true positive (Theorem 2) and false positive (Theorem 3) probabilities, whereas $\hat{P}_I$ and $\hat{P}_{II}$ are probabilities from numerical simulations.

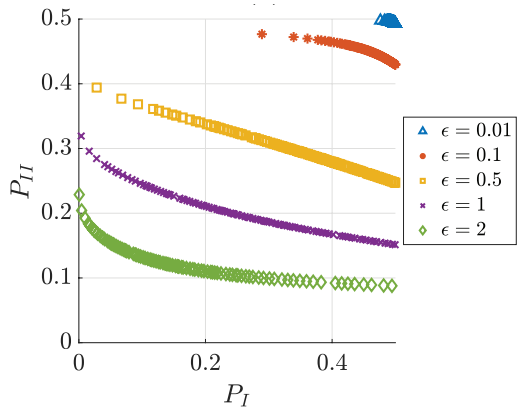

Fig. 4. P_I and P_{II} evaluated using Theorems 2 and 3 for different levels of privacy (i.e., different values of ϵ): each curve is constructed by varying the value of the threshold h.

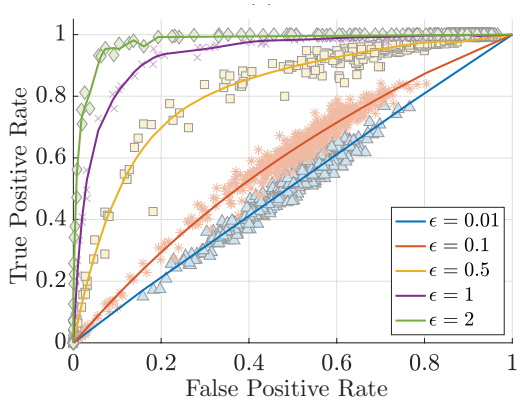

Fig. 5. ROC curves for OUTLIERDETECT, for the example described in Section IV.

9.13×10^3, resulting in 10% of observations $\mathbf{x}^{(k)}$ being classified as outliers. This comparison is plotted in Figure 3. Note the validation of our analytical expressions in the limiting regimes of $\epsilon \to 0$ and $\epsilon \to \infty$. Additionally, the classification performance degrades ("i.e., less information is revealed") with higher levels of privacy (i.e., ϵ decreases).

In Figure 4, we examine the performance of Theorems 2 and 3 with respect to $\mathcal{O}_m$, parameterized by threshold h. Each ϵ-level curve in Figure 4 sweeps out the P_I versus P_{II} probabilities from left to right corresponding to decreasing h. We see that for high privacy requirements (i.e., $\epsilon = 0.01$), the classifier conceals information regardless of the threshold h. As the privacy requirements relax, the incidence of false positive classifications decrease with higher h, at the cost of incurring more false negatives.

Finally, in Figure 5 we show via a receiver operating characteristic (ROC) curve the trade-off between detection accuracy (i.e., the true positive rate $1 - P_{II}$ versus false positive rate P_I) and privacy requirements when we use OUTLIERDETECT to analyze $\mathcal{O}_m$. Once again, we set $h \approx 9.13 \times 10^3$. As expected, the performance of the algorithm is worse in the high privacy regime (i.e., as ϵ becomes smaller).

V. CONCLUSION

In this article, we considered the problem of conducting norm-based outlier detection in multivariate Gaussian signals under a differentially private constraint. We designed a differentially private outlier detection algorithm, and derived closed-form expressions for its classification probabilities. Using a numerical example, we quantified the trade-off between classification accuracy and privacy, and empirically validated our theoretical results. Our ongoing work investigates the applications of the proposed framework to mode detection in hybrid systems, and to the classification of graph signals.

REFERENCES

[1] G. Piacentini, P. Goatin, and A. Ferrara, "Traffic control via platoons of intelligent vehicles for saving fuel consumption in freeway systems," *IEEE Control Systems Letters*, vol. 5, no. 2, pp. 593–598, 2021.

[2] J. Cortés, G. E. Dullerud, S. Han, J. Le Ny, S. Mitra, and G. J. Pappas, "Differential privacy in control and network systems," in *2016 IEEE 55th Conference on Decision and Control (CDC)*. IEEE, 2016, pp. 4252–4272.

[3] J. He and L. Cai, "Differential private noise adding mechanism: Basic conditions and its application," in *2017 American Control Conference (ACC)*. IEEE, 2017, pp. 1673–1678.

[4] E. Nozari, P. Tallapragada, and J. Cortés, "Differentially private average consensus: Obstructions, trade-offs, and optimal algorithm design," *Automatica*, vol. 81, pp. 221–231, 2017.

[5] K. H. Degue and J. Le Ny, "Differentially private interval observer design with bounded input perturbation," in *2020 American Control Conference (ACC)*, 2020, pp. 1465–1470.

[6] C. Dwork and A. Roth, *The Algorithmic Foundations of Differential Privacy*. Foundations and Trends in Theoretical Computer Science, 2014, vol. 9, no. 3-4.

[7] J. Le Ny and G. Pappas, "Differential private filtering," *IEEE Transactions on Automatic Control*, vol. 59, no. 2, pp. 341–354, February 2014.

[8] Y. Wang, Z. Huang, S. Mitra, and G. E. Dullerud, "Differential privacy in linear distributed control systems: Entropy minimizing mechanisms and performance tradeoffs," *IEEE Transactions on Control of Network Systems*, vol. 4, no. 1, pp. 118–130, Mar. 2017.

[9] L. Sweeney, "*k*-anonymity: a model for protecting privacy," *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, vol. 10, pp. 557–570, 2002.

[10] C. Dwork, M. Naor, O. Reingold, G. N. Rothblum, and S. Vadhan, "On the complexity of differentially private data release: efficient algorithms and hardness results," in *Proceedings of the forty-first annual ACM symposium on Theory of computing*, 2009, pp. 381–390.

[11] M. Gaboardi, H. W. Lim, R. Rogers, and S. P. Vadhan, "Differentially Private Chi-Squared Hypothesis Testing: Goodness of Fit and Independence Testing," in *Proceedings of the 33rd International Conference on Machine Learning*, New York City, NY, USA, Jun. 2016, pp. 2111–2120.

[12] Y. Wang, J. Lee, and D. Kifer, "Revisiting differentially private hypothesis tests for categorical data," *ArXiv e-prints*, Mar. 2017.

[13] M. Ghassemi, A. D. Sarwate, and R. Wright, "Differentially private online active learning with applications to anomaly detection," in *Proceedings of the 9th ACM Workshop on Artificial Intelligence and Security*, October 2016.

[14] X. Tong, B. Xi, M. Kantarcioglu, and A. Inan, "Gaussian mixture models for classification and hypothesis tests under differential privacy," in *31st Annual IFIP WG 11.3 Conference on Data and Applications Security and Privacy (DBSec'17)*, Philadelphia, PA, USA, Jul. 2017.

[15] K. H. Degue and J. Le Ny, "On differentially private Gaussian hypothesis testing," in *Proceedings of the 56th Annual Allerton Conference on Communication, Control, and Computing*, Allerton Park and Retreat Center, Monticello, Illinois, USA, Oct. 2018.

[16] R. Cummings, S. Krehbiel, Y. Mei, R. Tuo, and W. Zhang, "Differentially private change-point detection," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS 2018)*, Montreal, Canada, Dec. 2018.

[17] G. Liao, X. Chen, and J. Huang, "Social-aware privacy-preserving mechanism for correlated data," *IEEE/ACM Transactions on Networking*, vol. 28, no. 4, pp. 1671–1683, 2020.

[18] C. Dwork, "Differential privacy," in *Proceedings of the 33rd International Colloquium on Automata, Languages and Programming (ICALP)*, ser. Lecture Notes in Computer Science, vol. 4052. Springer-Verlag, 2006.

[19] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proceedings of the Third Theory of Cryptography Conference*, 2006, pp. 265–284.

[20] M. Lyu, D. Su, and N. Li, "Understanding the sparse vector technique for differential privacy," *Proceedings of the VLDB Endowment*, vol. 10, no. 6, pp. 637–648, Feb. 2017.