Sensor Selection for Event Detection in Wireless Sensor Networks

Dragana Bajović, Bruno Sinopoli and João Xavier

Abstract

We consider the problem of sensor selection for event detection in wireless sensor networks (WSNs). We want to choose a subset of p out of n sensors that yields the best detection performance. As the sensor selection optimality criteria, we propose the Kullback-Leibler and Chernoff distances between the distributions of the selected measurements under the two hypothesis. We formulate the maxmin robust sensor selection problem to cope with the uncertainties in distribution means. We prove that the sensor selection problem is NP hard, for both Kullback-Leibler and Chernoff criteria. To (sub)optimally solve the sensor selection problem, we propose an algorithm of affordable complexity. Extensive numerical simulations on moderate size problem instances (when the optimum by exhaustive search is feasible to compute) demonstrate the algorithm's near optimality in a very large portion of problem instances. For larger problems, extensive simulations demonstrate that our algorithm outperforms random searches, once an upper bound on computational time is set. We corroborate numerically the validity of the Kullback-Leibler and Chernoff sensor selection criteria, by showing that they lead to sensor selections nearly optimal both in the Neyman-Pearson and Bayes sense.

Keywords: sensor selection, event detection, robust optimization, Chernoff distance, Kullback-Leibler distance

D. Bajović is with the Institute for Systems and Robotics (ISR), Instituto Superior Técnico (IST), Technical University of Lisbon, Portugal, and with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, USA `dragana@isr.ist.utl.pt, dbajovic@andrew.cmu.edu`

B. Sinopoli is with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, USA `brunos@ece.cmu.edu`

J. Xavier is with the Institute for Systems and Robotics (ISR), Instituto Superior Técnico (IST), Technical University of Lisbon, Portugal `jxavier@isr.ist.utl.pt`

I. INTRODUCTION

Wireless sensor networks (WSNs) operate with limited power and communication resources. When observing phenomena with WSNs, a major challenge is to balance the tradeoff between the quality and the cost of operation. A fundamental problem of this kind, sensor selection problem, is how to optimally select a limited subset of sensors (hence limiting the operation cost) that gives the most valuable information about the observed phenomena.

Problem statement. This paper studies the sensor selection problem for event detection in WSNs. Nature can be in one of two states: H_1 (event occurring, e.g., target present) and H_0 (event not occurring, e.g., target absent). A WSN, composed of n sensors, instruments the nature. The distribution of the n-dimensional measurement vector is assumed Gaussian under the two hypothesis, with different means m_i and covariances S_i, $i = 0, 1$. We assume that, due to inherent WSN constraints, such as power, only p (out of n) sensors can sense and transmit their readings to a fusion node; based on the received p readings, the fusion node performs detection, i.e., decides which of the two hypothesis is true. We ask the following question: *Which p sensors should be chosen to achieve the best detection performance?*

Each possible p-sensor selection induces, under hypothesis H_i, a p dimensional Gaussian distribution π_i (of selected sensors), $i = 0, 1$. Intuitively, a p-sensor selection that yields more distant distributions π_1 and π_0 leads to better detection. Hence, we propose, as sensor selection optimality criteria: 1) the Kullback-Leibler (KL) distance; and 2) the Chernoff (C) distance between π_1 and π_0. In practice, the distribution parameters (m_i, S_i) are estimated from training data, and may not be known exactly, but only within an uncertainty region. We thus formulate the robust maxmin sensor selection problem of maximizing the KL (or C) distance between the selected distributions π_1 and π_0, for the worst case of parameter drifts. In this paper, we address the case when only the means of the two distributions are uncertain, relegating the general case for future work.

Contributions. The problem of evaluating the best p-sensor subset is combinatorial; checking over all $\binom{n}{p}$ possible combinations becomes infeasible when n and p are sufficiently large. We indeed prove that the KL and C sensor selection problems are NP hard; hence, it is unlikely to find an efficient algorithm that solves large instances of these problems. To (suboptimally in general) solve the sensor selection problems, we propose a computationally affordable algorithm. For example, to select 10 out of 100 sensors, our algorithm takes only few seconds on a current generation personal computer. There is no theoretical guarantee that our algorithm produces an optimal or near optimal solution; however, extensive numerical experiments demonstrate that our algorithm produces an optimal (or near optimal) solution in most cases.

The KL and C distances are only heuristic optimality measures for the detection performance; the exact (yet in our case intractable) criterion is the probability of detection error (either in Neyman-Pearson or Bayes sense). However, interestingly enough, we show by simulations that optimizing the KL and C distances yields to sensor selections that are very close to the optimal probability of error (either in Bayes or Neyman-Pearson sense). Moreover, they are often indeed optimal.

Selecting p out of n sensors is equivalent to finding the $n \times p$ selection matrix with 0/1 entries that maps the n-dimensional vector x of all measurements to the p-dimensional vector y of selected measurements. Our methodology for solving the combinatorial sensor selection problems relies on enlarging the search space from the the set of 0/1 selection matrices to the set of Stiefel matrices (the matrices with orthonormal columns). Then, after solving for the Stiefel matrix, we project it back to the set of selection matrices. The relaxed Stiefel problem corresponds to finding the linear map $\mathbb{R}^n \to \mathbb{R}^p$ which maximizes the (worst-case) KL (C) distance between the projected distributions in the lower p-dimensional space. To our best knowledge, existing work on this topic either does not solve the problem in full generality (i.e., unequal means and covariance matrices) or does not guarantee global optimality of their solutions (see [1], [2] for problems involving Chernoff distance and the closely related J-divergence). A major contribution of this paper is that we solve this nonconvex problem globally for the case $p = 1$, and in full generality, by reducing it to a scalar (1D) problem over a compact interval. We tackle the generic case $p > 1$ via an incremental, greedy approach, based on the 1D case, which provides near optimal result with small computational cost.

This paper is related to our prior work [3], [4], which also considers sensor selection for event detection, but only based on the KL distance. With respect to KL distance, this paper provides a new heuristic with reduced complexity; more importantly, this paper studies the problem with respect to the Chernoff distance, which we did not consider in [3], [4]. With respect to [3], [4], this paper also contributes by validating the KL and C distances as good optimality criteria by showing their near optimality in the error probability sense, and by establishing NP hardness of the corresponding sensor selection problems.

Finally, we would like to note that the KL-based and C-based sensor selection problems could be, in principle, globally solved by, e.g., branch and bound methods [5], [6]. However, the computational time of such methods is often very long, even for modest values of n and p. We discuss in more detail the challenges to solve the sensor selection problems that we address at the end of subsection II-B.

Review of existing work on sensor selection. Sensor selection problems have been extensively studied in different contexts, including target localization [7], target tracking [8] and sensor querying in ad hoc sensor networks [9]. Generally, sensor selection problems appear in the form of optimizing a certain

measure of performance of the system (e.g., utility [10], information theoretic measure [8], estimation error [11]) subject to energy constraints (e.g., limited number of sensors to be selected [11]). Our work belongs to the class of sensor selection problems for inference tasks, i.e., sensor selection for estimation and detection. Reference [11] considers the problem of sensor selection for parameter estimation in WSNs, proposing to select the subset of p (out of n) sensors that minimize the determinant of the estimator covariance matrix. Reference [12] proposes distributed algorithms to (suboptimally) solve the sensor selection problem formulated in [11]. Reference [13] addresses the problem of selecting the maximal number of reliable sensors for estimation. Reference [14] shows, through the optimal experiment design framework [15] and using convex analysis, that optimal estimation is achievable by using only a relatively small number of sensors.

Paper organization. Section II introduces the model and formulates the sensor selection optimization problems. Section III details the algorithms for solving the robust sensor selection problems, in the presence of uncertainties in the means of the distributions. Section IV considers the special, yet important case, when there are no uncertainties in the distribution parameters. Section V demonstrates numerically that the KL and the C distances are appropriate metrics for sensor selection. Section VI shows near optimal performance of the proposed algorithms. Finally, Section VII concludes the paper.

II. PROBLEM FORMULATION

A. Problem model

We assume that nature can be in one of two states: H_1–event occurring, and H_0–event not occurring. Let $x \in \mathbb{R}^n$ denote the vector that collects all sensor measurements (one measurement per sensor). We assume that x is Gaussian under the hypothesis H_1 (respectively, H_0), with generic means and covariances (m_1, S_1) (respectively, (m_0, S_0)), i.e.,

$$\text{under } H_i \quad : \quad x \sim \mathcal{N}(m_i, S_i), \, i = 0, 1 \ ,$$

where $\mathcal{N}(\mu, \Sigma)$ denotes Gaussian distribution with mean vector μ and covariance matrix Σ. The Gaussian assumption on x is standard and can be, in many applications, justified, e.g, by central limit theorem arguments [16], [17]. Noise correlation (i.e., non diagonal covariance matrices) is important to take into account in dense deployments of WSNs. We note that our formulation allows for different covariances under the two hypothesis ($S_1 \neq S_0$); in many applications, e.g., power-based detection of primary users for cognitive radios [17], accounting for different covariances is essential.

Sensor selection. Sensors transmit their measurements to a fusion node, which conducts the hypothesis test. Due to power constraints, only p sensors, $p < n$, perform their measurements and transmit them

to the fusion node. We address the problem of selecting the p sensors that guarantee the best detection performance. Mathematically, selecting p out of n sensors can be represented by a linear map $\mathbb{R}^n \to \mathbb{R}^p$, $x \mapsto y = E^\top x$, where $E \in \mathbb{R}^{n \times p}$ is a rank-p matrix that has exactly one unit entry per column, corresponding to a chosen sensor, and the other entries in columns being zero. The columns of E are orthonormal, i.e., $E^\top E = I_p$, where I_p denotes the $p \times p$ identity matrix. We refer to matrix E as the sensor selection matrix.

Hypothesis test induced by E. Conditioned on H_i, $i = 0, 1$, y is a linear transformation of a Gaussian vector x. Thus, y, under H_i, has the following distribution:

$$\text{under } H_i \quad : \quad y \sim \mathcal{N}(E^\top m_i, E^\top S_i E), \ i = 0, 1. \tag{1}$$

The fusion node performs the following log-likelihood ratio (LLR) decision test:

$$\log \frac{f_1(y; E)}{f_0(y; E)} \underset{H_0}{\overset{H_1}{\gtrless}} \gamma, \tag{2}$$

where $f_i(\cdot; E)$, $i = 0, 1$, is the density of $\mathcal{N}\left(E^\top m_i, E^\top S_i E\right)$ and $\gamma \in \mathbb{R}$ is the test threshold [2].

B. Formulation of the sensor selection optimization problem

Sensor selection optimality criteria. Detection performance is, as noted above, generally quantified by the error probabilities, P_{FA}, P_{D} and P_{e}. However, in the problem that we consider, none of the probabilities above admit closed form expression, and their minimizations with respect to sensor selection are very hard problems. As sensor selection optimality criteria, we propose the Kullback-Leibler (KL) and the Chernoff (C) distance between the tested distributions. Given two distributions with densities f_1 and f_0, KL and C distances measure dissimilarity between f_1 and f_0 and they are defined as follows:

$$D_{\text{KL}}\left(f_1 \,\|\, f_0\right) := \int \log \frac{f_1(x)}{f_0(x)} f_1(x) dx, \quad D_{\text{C}}\left(f_1, \, f_0\right) := \max_{s \in [0,1]} -\log \int f_1^s(x) f_0^{1-s}(x) dx.$$

Thus, our goal is to find a sensor selection matrix E that yields the maximal KL or C distance between the projected Gaussian distributions, $\mathcal{N}(E^\top m_1, E^\top S_1 E)$ and $\mathcal{N}(E^\top m_0, E^\top S_0 E)$ (see (1)).

We are motivated to choose KL and C distances, as sensor selection optimality criteria, by two fundamental results from detection theory: Chernoff-Stein lemma and Chernoff lemma [18]. Chernoff-Stein lemma (resp. Chernoff lemma) states that, when the number of independent identically distributed (i.i.d.) observations grows, the rate of exponential decay of probability of false alarm, P_{FA}, (resp. probability of error, P_{e},) of the Neyman-Pearson optimal (resp. Bayes optimal) test equals the KL (resp. C) between the two distributions. Thus, for large number of samples, more distant distributions (in either KL or C sense) lead to better detection performance. Probabilistic distance measures have been often

used in the literature as heuristics for detection problems(e.g., see [1], [19] for applications in linear dimensionality reduction) and have shown excellent results, even when the number of samples is very small or even equal to one [19]. In section V, we demonstrate by numerical tests that the KL and C distances are indeed excellent criteria for sensor selection, exhibiting near optimal performance in the probability of error sense. Section V shows that, generally, C distance has an advantage in the regimes of high probability of detection (P_D) (upper part of the receiver operating characteristics (ROC) curve), while KL has an edge in the regimes of low P_FA (lower left part of the ROC curve).

Robustness against uncertainty in distribution means. We consider the case when the true distribution parameters (m_i, S_i), $i = 0, 1$ are not exactly known at the fusion node detector (2). Fusion node has their estimates, $\left(\widehat{m_i}, \widehat{S_i}\right)$, which can be obtained, e.g., in the network training phase. Thus, there is a mismatch between the distribution used by the fusion node detector and the distribution that generates the observations. Our goal is to design a sensor selection that yields detection (2) robust to these mismatches. In this paper, we restrict our attention to the case where only the mean values are uncertain (the true covariance matrices are known), and we allow the mean values to drift in the following ellipsoidal uncertainty regions:

$$m_i \in \mathcal{E}\left(\widehat{m_i}, k_i \, S_i^{-1}\right) \quad i = 0, 1. \tag{3}$$

Here $\widehat{m_i}$ denotes the estimated mean vector, $i = 0, 1$, $\mathcal{E}(a, A)$ (A is a positive definite matrix) denotes the ellipsoid $\mathcal{E}(a, A) = \{x \in \mathbb{R}^n : (x - a)^\top A(x - a) \leq 1\}$, and the parameter $k_i \in (0, +\infty]$ is a free parameter which controls the "size" of the uncertainty region; e.g., if $k_i = +\infty$, there is no uncertainty: $\widehat{m_i} = m_i$. The orientations of the uncertainty ellipsoids in (3) are induced by the covariance matrices S_0 and S_1. This choice of the form of uncertainty regions is motivated by the following fact: if the means are estimated via the sample mean estimator based on N i.i.d. observations (the minimum variance unbiased estimate for Gaussian distributions), then the covariance of the estimate $\widehat{m_i}$ equals $\frac{1}{N} S_i$, for $i = 0, 1$. A standard measure of uncertainty, a confidence region [2], for m_i is given by (3). The scaling constants k_0 and k_1 are, in this context, proportional to N; also, if N is known, k_0 and k_1 can be used to design the uncertainty regions of desired confidence levels. We address the uncertainties in the mean vectors adopting the worst case approach. That is, we search for the sensor selection that gives the maximal distance (KL and C) for the worst case of the mean parameter drift.

Optimization problems. We introduce the following two functions, f_KL and f_C, that capture the dependence of the KL and the C distance on the selection matrix E and the mean vectors m_0 and m_1:

$$f_\mathrm{KL}(E, m_0, m_1) \;\; := \;\; D_\mathrm{KL}\left(\mathcal{N}(E^\top m_1, E^\top S_1 E) \| \mathcal{N}(E^\top m_0, E^\top S_0 E)\right)$$

$$\max_{s \in [0,1]} f_{\mathrm{C}}(s, E, m_0, m_1) \quad := \quad D_{\mathrm{C}} \left(\mathcal{N}(E^\top m_1, E^\top S_1 E), \mathcal{N}(E^\top m_0, E^\top S_0 E) \right).$$

The robust sensor selection optimization problems are then given as follows:

$$
\begin{aligned}
\text{maximize} \quad & \min_{m_0 \in \mathcal{E}(\widehat{m}_0, k_0 S_0^{-1}), m_1 \in \mathcal{E}(\widehat{m}_1, k_1 S_1^{-1})} f_{\mathrm{KL}}(E, m_0, m_1) \\
\text{subject to} \quad & E_{ij} \in \{0, 1\} \\
& E^\top E = I_p
\end{aligned}
\tag{4}
$$

$$
\begin{aligned}
\text{maximize} \quad & \min_{m_0 \in \mathcal{E}(\widehat{m}_0, k_0 S_0^{-1}), m_1 \in \mathcal{E}(\widehat{m}_1, k_1 S_1^{-1})} \max_{s \in [0,1]} f_{\mathrm{C}}(s, E, m_0, m_1) \\
\text{subject to} \quad & E_{ij} \in \{0, 1\} \\
& E^\top E = I_p
\end{aligned}
\tag{5}
$$

It can be shown that:

$$
f_{\mathrm{KL}}(E, m_0, m_1) = \frac{1}{2} \left\{ (m_1 - m_0)^\top E \left(E^\top S_0 E \right)^{-1} E^\top (m_1 - m_0) + \mathrm{tr} \left(\left(E^\top S_0 E \right)^{-1} E^\top S_1 E \right) \right.
$$
$$
\left. - \log \frac{|E^\top S_1 E|}{|E^\top S_0 E|} - p \right\} \tag{6}
$$

$$
f_{\mathrm{C}}(s, E, m_0, m_1) = \frac{1}{2} \left\{ s(1-s)(m_1 - m_0)^\top E \left(s E^\top S_0 E + (1-s) E^\top S_1 E \right)^{-1} E^\top (m_1 - m_0) \right.
$$
$$
\left. - \log \frac{|E^\top S_0 E|^s \, |E^\top S_1 E|^{1-s}}{|s E^\top S_0 E + (1-s) E^\top S_1 E|} \right\}, \tag{7}
$$

where $\mathrm{tr}(\cdot)$ and $|\cdot|$ denote trace and determinant, respectively.

Optimization problems (4) and (5) are combinatorial. When n and p are small, a simple method for solving them is exhaustive search that checks all $\binom{n}{p}$ sensor subsets (i.e. selection matrices). For large n and p, however, this method becomes computationally infeasible. Indeed, we have the following result, which we prove in Appendix A:

Theorem 1 Optimization problems (4) and (5) are NP hard (even when $k_0 = k_1 = +\infty$ and $S_1 = S_0$).

In principle, besides exhaustive search, problems (4) and (5) can be solved by branch and bound methods [6]. However, complexity of these methods relies strongly on the choice of bounds on the cost function; finding tight bounds is a hard problem itself. An interesting method for solving a problem somewhat similar to ours is proposed in [20]. In [20], the authors address the problem of finding the most informative locations, for future sensor placements, in a discretized Gaussian field; the measure of informativeness that the authors propose is the mutual information between sensed and un-sensed locations. The authors show that the mutual information is a submodular function, which assures that the simple strategy of choosing sensor by sensor greedily gives a solution within $1 - \frac{1}{e}$ relative of the optimum. In our problems, however, such a bound on the greedy strategy (of selecting sensors one by

one) does not hold, as our cost functions (KL and C distance) are not submodular (see the proof in Appendix B). In fact, in our problems, greedy performs poorly in many cases. The reason for this lies in the fact that the correlations – an information that greedy discards can play an important role in discerning between the two hypothesis.

III. SENSOR SELECTION ALGORITHMS

In this section, we present our algorithms, R–KL (robust KL based selection) and R–C (robust C-based selection), that, respectively, solve problems (4) and (5). First, we explain geometrical intuition behind our methodology and the resulting structure of the algorithms, and then in subsections III-A and III-B, respectively, we provide the details for the R–KL and R–C algorithms.

Algorithms methodology and structure. Geometrically, one combination of p sensors defines one p-dimensional subspace of $\mathbb{R}^n$. This subspace is the range of the corresponding selection matrix E and, since columns of E are canonical basis vectors of $\mathbb{R}^n$, we call this subspace a canonical subspace. Simple analysis reveals that cost functions in (4) and (5) depend on selection matrix E only through its range. Therefore, problems (4) and (5) in this sense search for the best, among $\binom{n}{p}$ canonical subspaces, on which the original distributions should be projected. The way we relax these combinatorial problems is by allowing for projections to arbitrary p-dimensional subspaces. Mathematically, this translates into replacing the set of 0/1 selection matrices with the set of $n \times p$ Stiefel matrices (that represent all p-dimensional subspaces). Then, we use a solution of the relaxed problem and "round" it to the closest canonical subspace. We call the first phase of our algorithm, that solves the relaxed Stiefel problem, the *Relaxation phase*; the second phase, in which we find the closest canonical subspace, is the *Projection phase*. Finally, the last, third step in our algorithms, the *Refinement phase*, refines the solution by performing local optimization.

A. Algorithm for the robust Kullback-Leibler based selection: the R–KL algorithm

1) Relaxation phase: We solve the following Stiefel relaxation of problem (4):

$$
\begin{aligned}
\text{maximize} \quad & \min_{m_0 \in \mathcal{E}(\widehat{m_0}, k_0\, S_0^{-1}), m_1 \in \mathcal{E}(\widehat{m_1}, k_1\, S_1^{-1})} f_{\text{KL}}(E, m_0, m_1) \\
\text{subject to} \quad & E^\top E = I_p.
\end{aligned}
\tag{8}
$$

The relaxed problem (8) is nonconvex and still difficult to solve; we globally solve it for $p = 1$. The general, $p > 1$ case, is addressed by a greedy approach, using our approach for $p = 1$. We first detail the algorithm that solves the $p = 1$ case. For $p = 1$, the constraint set of Stiefel matrices reduces to a

sphere in $\mathbb{R}^n$ and the problem (8) takes a simplified form:

$$\text{maximize} \quad \min_{\substack{m_0 \in \mathcal{E}(\widehat{m}_0, k_0 \, S_0^{-1}), \\ m_1 \in \mathcal{E}(\widehat{m}_1, k_1 \, S_1^{-1})}} \frac{1}{2} \left\{ \frac{(e^\top (m_1 - m_0))^2}{e^\top S_0 e} + \frac{e^\top S_1 e}{e^\top S_0 e} - \log \frac{e^\top S_1 e}{e^\top S_0 e} - 1 \right\}$$

$$\text{subject to} \quad e^\top e = 1 \tag{9}$$

Our major contribution is showing that problem (9) reduces to a search over a compact (one-dimensional) interval. We achieve this by a series of judicious problem reformulations, and by invoking convexity of certain quadratic mappings [21], [22]. It is important to note that the original problem (9) has in general very high dimensionality (equal to the total number of sensor in the network n); also, due to its nonconvexity, it is very difficult to solve globally. By doing reformulations, we manage to map it to a tractable, scalar problem. Lemma 2 states the first step towards this goal. It shows that a solution of (9) can be reconstructed after solving the two-dimensional problem (10).

Lemma 2 Suppose $(x^\star, y^\star)$ solves

$$\text{maximize} \quad \psi_{\text{KL}}(x, y)$$

$$\text{subject to} \quad (x, y) \in \mathcal{R} \tag{10}$$

where

$$\psi_{\text{KL}}(x, y) = x - \log x + \left\{ \left(\sqrt{y} - \frac{1}{\sqrt{k_1}} \sqrt{x} - \frac{1}{\sqrt{k_0}} \right)^+ \right\}^2, \tag{11}$$

$$\mathcal{R} = \left\{ (x, y) \in \mathbb{R}^2 : \; x = v^\top S v, \; y = v^\top M v, \; \text{for some } v \in \mathbb{R}^n, \; v^\top v = 1 \right\}, \tag{12}$$

$x^+ = \max(0, x)$, $S = S_0^{-1/2} S_1 S_0^{-1/2}$, $m = S_0^{-1/2}(\widehat{m}_1 - \widehat{m}_0)$, $M = m m^\top$. Let $v^\star \in \mathbb{R}^n$ be an unit-norm vector that generates $x^\star$ and $y^\star$, i.e. $x^\star = v^{\star\top} S v^\star$ and $y^\star = v^{\star\top} M v^\star$. Then, $e^\star := S_0^{-1/2} v^\star / \|S_0^{-1/2} v^\star\|$ solves (9).

The detailed proof of Lemma 2 is in [4]; for completeness, we briefly explain here why Lemma 2 holds. First, we note that the (inner) objective function in (9) depends on e only through its direction. This allows for replacing the constraint $e^\top e = 1$ with $e^\top S_0 e = 1$ and introducing the new variable $v = S_0^{1/2} e$; the outer (maximization) constraint becomes $v^\top v = 1$. It can be shown that the inner minimization, with respect to (w.r.t.) m_0 and m_1, now has a closed form solution, equal to: $v^\top S v - \log v^\top S v + \left\{ \left(\sqrt{v^\top M v} - \frac{1}{\sqrt{k_1}} \sqrt{v^\top S v} - \frac{1}{\sqrt{k_0}} \right)^+ \right\}^2$. Finally, noting that the latter function depends on v only through the two quadratic forms, $v^\top S v$ and $v^\top M v$, allows for replacing the n-dimensional optimization variable v with only two scalars, $x = v^\top S v$ and $y = v^\top M v$, which yields the claim of Lemma 2.

Lemma 2 says that, in order to solve (9), it suffices to search over the set $\mathcal{R} \subset \mathbb{R}^2$. For $n \geq 3$, the set $\mathcal{R}$ is compact and convex, as $\mathcal{R}$ is the image set of a unit sphere under two quadratic mappings [21]. Note that, since $\psi_{\text{KL}}(\cdot)$ is continuous and $\mathcal{R}$ is compact, there is a global maximizer, by the Weierstrass

theorem. The next lemma 3 further simplifies the search by asserting that the boundary of $\mathcal{R}$ contains a global maximizer. For the proof of Lemma 3 see [4].

Lemma 3 The boundary $\partial\mathcal{R}$ of the set $\mathcal{R}$ contains a global maximizer of (10).

The boundary of $\mathcal{R}$ is a closed curve in $\mathbb{R}^2$. Our strategy consists in circulating along ∂R to spot a global maximizer. More precisely, we will sample ∂R with a finite set of points and pick the best point. To implement this strategy, we borrow the following theorem from [22].

Theorem 4 ([22]) Let $n \geq 3$ and let A, B be $n \times n$ symmetric matrices. Let

$$\mathcal{R}(A, B) = \left\{ (x, y) \in \mathbb{R}^2 : \ x = v^\top Av, \ y = v^\top Bv, \ \text{for some } v \in \mathbb{R}^n, \ v^\top v = 1 \right\}.$$

For $t \in [0, 2\pi]$, let $C(t) = A \cos t + B \sin t$ and let $\lambda_{\min}(t)$ be the minimal eigenvalue of the matrix $C(t)$ and $u_{\min}(t)$ an associated unit-norm eigenvector. Suppose that $\lambda_{\min}(t)$ is a simple eigenvalue of $C(t)$ for all $t \in [0, 2\pi]$. Then, the set $\mathcal{R}(A, B)$ is strictly convex and its boundary is given by

$$\partial\mathcal{R}(A, B) = \left\{ (x(t), y(t)) : \ t \in [0, 2\pi], \ x(t) = u_{\min}(t)^\top Au_{\min}(t), \ y(t) = u_{\min}(t)^\top Bu_{\min}(t) \right\}. \quad (13)$$

The quantity t in Theorem 4 parameterizes all supporting hyperplanes of the set $\mathcal{R}(A, B)$; namely, a hyperplane H_t, $t \in [0, 2\pi]$, is given by: $H_t = \{(x, y) \in \mathbb{R}^2 : x \cos t + y \sin t = x(t) \cos t + y(t) \sin t\}$, where $(x(t), y(t))$ is a point at which H_t intersects with (touches) $\mathcal{R}(A, B)$. Theorem 4 shows that, when $\lambda_{\min}(t)$ is simple for all t ($\mathcal{R}(A, B)$ is strictly convex), the boundary $\partial\mathcal{R}(A, B)$ can be parameterized by moving a single parameter t over the compact interval $[0, 2\pi]$. A parametrization of $\partial\mathcal{R}(A, B)$ (in our context, $\partial\mathcal{R}(S, M)$) is readily available, even when $\mathcal{R}(S, M)$ is not strictly convex. Namely, if, for some t, $\lambda_{\min}(t)$ is not unique, then the hyperplane H_t touches $\mathcal{R}(S, M)$ on a line segment, and not at a unique point. The sample of $\partial\mathcal{R}(S, M)$ can then be numerically obtained as follows: 1) slide t in a fine grid over the interval $[0, 2\pi]$ and acquire one point $(x(t), y(t))$ (per each t) as in (13); and 2) to identify the line segments, connect the consecutive points $(x(t), y(t))$ with a straight line, if these consecutive points are far apart. More precisely, the procedure is given as follows:

1) generate the points

$$(x_k, y_k) = (u_k^\top Su_k, u_k^\top Mu_k), \quad k = 1, 2, \ldots, K,$$

where u_k denotes an unit-norm eigenvector corresponding to the minimal eigenvalue of

$$C_k = S \cos\left((k - 1)2\pi/K\right) + M \sin\left((k - 1)2\pi/K\right).$$

Here, K is the user-defined grid size and $\{(x_k, y_k) : k = 1, \ldots, K\}$ is an initial sample of $\partial\mathcal{R}$;

2) if the distance between two consecutive points (x_k, y_k) and (x_{k+1}, y_{k+1}) is greater than a prescribed threshold, interpolate the line segment which connects them, i.e., consider

$$\left(x_k^{(j)}, y_k^{(j)}\right) = (1 - j/J)(x_k, y_k) + j/J(x_{k+1}, y_{k+1}), \quad j = 0, 1, \ldots, J.$$

In summary, our sampling of $\partial \mathcal{R}$ is $\widehat{\partial \mathcal{R}} = \{(x_k, y_k)\} \cup \left\{x_k^{(j)}, y_k^{(j)}\right\}$.

Solving (9). We explained how to solve (10) by parameterizing $\partial \mathcal{R}$. Now we explain how a solution of (10), $x^\star, y^\star$, is sufficient to reconstruct $v^\star$, a solution of (9). The vector $v^\star$ in Lemma 2 can be found as follows. Let

$$(x^\star, y^\star) \in \arg \max_{(x,y) \in \widehat{\partial \mathcal{R}}} \psi_{\mathrm{KL}}(x, y).$$

That is, $(x^\star, y^\star)$ denotes the best point in $\widehat{\partial \mathcal{R}}$. If $(x^\star, y^\star) \in \{(x_k, y_k)\}$, say $(x^\star, y^\star) = (x_{k^\star}, y_{k^\star})$, then we can take $v^\star$ as an unit-norm eigenvector associated with the minimal eigenvalue of $C_{k^\star}$. Otherwise, $(x^\star, y^\star) \in \left\{(x_k^{(j)}, y_k^{(j)})\right\}$, and we need to solve the system of 3 quadratic equations:

$$v^\top S v = x^\star, \quad v^\top M v = y^\star, \quad v^\top v = 1, \tag{14}$$

w.r.t. v. Any solution can be taken as $v^\star$. It can be shown that (14) can be efficiently solved by solving a convex problem.

Case $p > 1$: greedy algorithm Optimization problem (8) for the case $p > 1$ is very difficult to solve globally; we propose a greedy, suboptimal approach. We construct the columns of the matrix $E = [e_1 \, e_2 \, \ldots \, e_p]$ one by one (in the order $e_1, e_2, \ldots$). We construct the j-th column by solving (9), with the constraint that the column e_j must be orthogonal to the previously determined columns $e_1, e_2, \ldots e_{j-1}$, i.e., we solve:

$$\begin{aligned}
\text{maximize} \quad & \min_{m_0 \in \mathcal{E}(\widehat{m}_0, k_0 S_0^{-1}), \, m_1 \in \mathcal{E}(\widehat{m}_1, k_1 S_1^{-1})} f_{\mathrm{KL}}(e, m_0, m_1, S_0, S_1) \\
\text{subject to} \quad & e^\top e = 1 \\
& e^\top e_i = 0, \quad i = 1, \ldots, j - 1.
\end{aligned} \tag{15}$$

Let $U^{(j)} \in \mathbb{R}^{n \times (n-j+1)}$ be a matrix with orthonormal columns which spans the orthogonal complement of span $\{e_1, \ldots, e_{j-1}\}$. The restrictions in (15) mean that $e = U^{(j)} e^{(j)}$ for some unit-norm $e^{(j)} \in \mathbb{R}^{n-j+1}$. This means that (15) corresponds to

$$\begin{aligned}
\text{maximize} \quad & \min_{m_0 \in \mathcal{E}(\widehat{m}_0, k_0 S_0^{-1}), \, m_1 \in \mathcal{E}(\widehat{m}_1, k_1 S_1^{-1})} f_{\mathrm{KL}}(U^{(j)} e^{(j)}, m_0, m_1, S_0, S_1) \\
\text{subject to} \quad & {e^{(j)}}^\top e^{(j)} = 1.
\end{aligned} \tag{16}$$

The problem (16) is equivalent to (17) [4]:

$$\text{maximize} \quad \min_{m_0^{(j)} \in \mathcal{E}\left(\widehat{m_0^{(j)}}, k_0 \left(S_0^{(j)}\right)^{-1}\right), \, m_1^{(j)} \in \mathcal{E}\left(\widehat{m_1^{(j)}}, k_1 S_1^{(j)^{-1}}\right)} f_{\text{KL}}(e^{(j)}, m_0^{(j)}, m_1^{(j)}, S_0^{(j)}, S_1^{(j)})$$

$$\text{subject to} \quad {e^{(j)}}^\top e^{(j)} = 1 \tag{17}$$

where $\widehat{m_i^{(j)}} = {U^{(j)}}^\top \widehat{m_i}$ and $S_i^{(j)} = {U^{(j)}}^\top S_i U^{(j)}$, $i = 0, 1$. That is, (17) is simply an instance of (9) in the reduced dimensional space $\mathbb{R}^{n-j+1}$, for which we have developed a global solution. Algorithm 1 outlines the overall approach.

Algorithm 1: Greedy algorithm

1: **for** $j = 1$ to p **do**
2:    Compute $U^{(j)} \in \mathbb{R}^{n \times (n-j+1)}$ ($U^{(1)} := I_n$), an orthonormal basis for the orthogonal complement of the $j-1$ dimensional subspace span $\{e_1, e_2, \ldots, e_{j-1}\}$
3:    Compute the projected means and covariances $\widehat{m_i^{(j)}} = {U^{(j)}}^\top \widehat{m_i}$, $S_i^{(j)} = {U^{(j)}}^\top S_i U^{(j)}$ for $i = 0, 1$
4:    Compute $S^{(j)} = (S_0^{(j)})^{-1/2} S_1^{(j)} (S_0^{(j)})^{-1/2}$, $m^{(j)} = (S_0^{(j)})^{-1/2} (\widehat{m_1^{(j)}} - \widehat{m_0^{(j)}})$, $M^{(j)} = m^{(j)} m^{(j)^\top}$
5:    Solve (10) for $(M, S) := (M^{(j)}, S^{(j)})$; find $e^{(j)} \in \mathbb{R}^{n-j+1}$ as: $e^{(j)} := e^\star$, as in Lemma 2
6:    Compute the jth column of E as $e_j = U^{(j)} e^{(j)}$
7: **end for**

---

2) Projection phase: The relaxation phase(in subsection III-A1) produces a Stiefel matrix E. Now, we project the matrix E back to the set of $0/1$ selection matrices. We remark that the objective function $f_{\text{KL}}(\cdot)$ in eqn. (6) depends on the matrix E only through its range space; that is, $f_{\text{KL}}(EQ) = f_{\text{KL}}(E)$, for any Stiefel matrix E and for any orthogonal $p \times p$ matrix Q. Thus, we choose the canonical selection matrix $\widetilde{E}$ with the range space closest to the range space of matrix E. It can be shown [3] that $\widetilde{E}$ can be efficiently obtained as follows: if $(j_1, j_2, \ldots, j_p)$ denote the indices of the largest entries on the diagonal of $EE^\top$, then $\widetilde{E} = \begin{bmatrix} h_{j_1} h_{j_2} \ldots h_{j_p} \end{bmatrix}$ where h_j stands for the j-th column of the identity matrix I_n. Thus, the projection phase has very small computational cost.

3) Refinement phase: Once the projection to the set of $0/1$ selection matrices is done and the matrix $\widetilde{E}$ is obtained, we finalize our algorithm with a local maximization around $\widetilde{E}$ to get $E^\star$ (see [11], [15] for very similar local searches.) Namely, for a given selection matrix E in the neighborhood of $\widetilde{E}$, we find

$$f_{\text{KL, worst}}(E) := \min_{m_0 \in \mathcal{E}\left(\widehat{m_0}, k_0 \, S_0^{-1}\right), m_1 \in \mathcal{E}\left(\widehat{m_1}, k_1 \, S_1^{-1}\right)} f_{\text{KL}}(E, m_0, m_1).$$

The procedure has p steps. We start with the matrix $E := \widetilde{E}$. In the first step, all columns of the current selection matrix E are fixed except the first one, which is viewed as an optimization variable. The first column is swept through all canonical vectors h_j, $j = 1, \ldots, n$, different from the remaining $p - 1$

columns of E. After all possible choices for the first column are tested, the column is frozen to the choice that gives the maximal $f_{\text{KL, worst}}$. In the second step, this procedure is repeated for the second column, and so on, up to the p-th step; after the p-th step is done, we set $E^\star := E$.

We obtain the quantity $f_{\text{KL, worst}}(E)$ by first finding:

$$\min_{\substack{m_0 \in \mathcal{E}\left(\widehat{m_0}, k_0 \, S_0^{-1}\right), \\ m_1 \in \mathcal{E}\left(\widehat{m_1}, k_1 \, S_1^{-1}\right)}} (m_1 - m_0)^\top E (E^\top S_0 E)^{-1} E^\top (m_1 - m_0), \tag{18}$$

and then adding the remaining terms of $f_{\text{KL}}(\cdot)$ that do not depend on m_1 and m_0 (see eqn. (6)).

The minimum in eqn. (18) (n-dimensional problem) equals the following minimum (of corresponding p-dimensional problem):

$$\min_{m_0' \in \mathcal{E}\left(Q^\top (E^\top S_0 E)^{-1/2} E^\top \widehat{m_0}, k_0 \, I_p\right), m_1' \in \mathcal{E}\left(Q^\top (E^\top S_0 E)^{-1/2} E^\top \widehat{m_1}, k_1 \, \Lambda^{-1}\right)} \|m_1' - m_0'\|, \tag{19}$$

where Λ and Q are, respectively, the matrix of eigenvalues and the matrix of eigenvectors of $(E^\top S_0 E)^{-1/2} E^\top S_1 E (E^\top S_0 E)^{-1/2}$, and $\| \cdot \|$ denotes Euclidean norm. The first step in showing the equivalence between (18) and (19) is to introduce the change of variables $m_i' = Q^\top (E^\top S_0 E)^{-1/2} E^\top m_i$, $i = 0, 1$. Define $\widehat{m_i}' := Q^\top (E^\top S_0 E)^{-1/2} E^\top \widehat{m_i}$. Then, the cost function becomes as in (19) and, to show the equivalence, it remains to show that $m_0' \in \mathcal{E}\left(\widehat{m_0}', k_0 \, I_p\right), m_1' \in \mathcal{E}\left(\widehat{m_1}', k_1 \, \Lambda^{-1}\right)$ if and only if $m_i \in \mathcal{E}\left(\widehat{m_i}, k_i \, S_i^{-1}\right)$, $i = 0, 1$. It can be shown (the proof can be found in [4]) that $E^\top \mathcal{E}\left(0, S_i^{-1}\right) = \mathcal{E}\left(0, (E^\top S_i E)^{-1}\right)$ and, also, that

$$Q^\top (E^\top S_0 E)^{-1/2} \mathcal{E}\left(E^\top \widehat{m_i}, (E^\top S_i E)^{-1}\right) = \mathcal{E}\left(\widehat{m_i}', \left(Q^\top (E^\top S_0 E)^{-1/2} E^\top S_i E (E^\top S_0 E)^{-1/2} Q\right)^{-1}\right)$$

$$= \begin{cases} \mathcal{E}\left(\widehat{m_0}', I_p\right) & \text{if } i = 0 \\ \mathcal{E}\left(\widehat{m_1}', \Lambda^{-1}\right) & \text{if } i = 1 \end{cases},$$

which proves the equivalence. The problem (19) is a (convex) quadratically constrained quadratic problem (QCQP), and it can be solved with complexity $\mathcal{O}(p^3)$ [23].

B. Algorithm for the robust Chernoff-based selection: the R–C algorithm

In this subsection, we present the algorithm R–C, the Chernoff based sensor selection under the presence of uncertainties. As mentioned previously, we adopt the same methodology for solving both (5) and (4) and, consequently, the structure of R–C is the same as the one in R–KL. However, the problem (5) is more difficult than (4), due to the additional maximization over the parameter s. This will result in several specificities in R–C compared to R–KL. We present R–C by focusing on these specificities, phase by phase, whereas the overall structure remains the same as in R–KL.

The main difference between R–KL and R–C is in the Relaxation phase in the case $p = 1$. As with the KL case, we solve the resulting Chernoff problem globally; we next explain a solution.

1) Relaxation phase: case $p = 1$: global solution: The Stiefel relaxation of (5) for $p = 1$ is given by:

$$\text{maximize} \quad \min_{m_0 \in \mathcal{E}(\widehat{m}_0, k_0\, S_0^{-1}), m_1 \in \mathcal{E}(\widehat{m}_1, k_1\, S_1^{-1})} \max_{s \in [0,1]} f_{\mathrm{C}}(s, e, m_0, m_1) \tag{20}$$
$$\text{subject to} \quad e^\top e = 1,$$

where

$$f_{\mathrm{C}}(s, e, m_0, m_1) = \frac{s(1-s)}{2} \frac{\left(e^\top(m_1 - m_0)\right)^2}{s e^\top S_0 e + (1-s) e^\top S_1 e} - \frac{1}{2} \log \frac{(e^\top S_0 e)^s (e^\top S_1 e)^{1-s}}{s e^\top S_0 e + (1-s) e^\top S_1 e}.$$

The first reformulation of (20) that we make is the conversion of the inner minimax problem into a maximin problem, as Lemma 5 explains.

Lemma 5 Problem (20) is equivalent to:

$$\text{maximize} \quad \max_{e^\top e = 1} \min_{m_0 \in \mathcal{E}(\widehat{m}_0, k_0\, S_0^{-1}), m_1 \in \mathcal{E}(\widehat{m}_1, k_1\, S_1^{-1})} f_{\mathrm{C}}(s, e, m_0, m_1) \tag{21}$$
$$\text{subject to} \quad s \in [0, 1].$$

Proof: Function f_{C} is convex w.r.t. m_0 and m_1, and concave w.r.t. s, and the constraint sets are compact and convex. Thus, the equivalence follows by Sion's minimax theorem [24]. ∎

Next, we focus on the inner maximization in (21):

$$\text{maximize} \quad \min_{m_0 \in \mathcal{E}(\widehat{m}_0, k_0\, S_0^{-1}), m_1 \in \mathcal{E}(\widehat{m}_1, k_1\, S_1^{-1})} f_{\mathrm{C}}(s, e, m_0, m_1) \tag{22}$$
$$\text{subject to} \quad e^\top e = 1$$

where $s \in [0, 1]$ is fixed. The following lemma is the counterpart of Lemma 2.

Lemma 6 Suppose $(x^\star, y^\star)$ solves

$$\text{maximize} \quad \psi_{\mathrm{C}}(s, x, y) \tag{23}$$
$$\text{subject to} \quad (x, y) \in \mathcal{R}$$

where

$$\psi_{\mathrm{C}}(s, x, y) = \frac{s(1-s)}{2} \frac{\left\{ \left(\sqrt{y} - \frac{1}{\sqrt{k_1}} \sqrt{x} - \frac{1}{\sqrt{k_0}} \right)^+ \right\}^2}{s + (1-s)x} - \frac{1}{2}(1-s) \log x + \frac{1}{2} \log(s + (1-s)x),$$

and $s \in [0, 1]$. Let $v^\star \in \mathbb{R}^n$ be an unit norm vector that generates $x^\star$ and $y^\star$, i.e., $x^\star = v^{\star\top} S v^\star$ and $y^\star = v^{\star\top} m m^\top v^\star$. Then, $e^\star := S_0^{-1/2} v^\star / \|S_0^{-1/2} v^\star\|$ solves (22).

Proof: The proof is similar to the proof of Lemma 2 and is omitted. ∎

By similar analysis as in subsection III-A1, it can be shown that (23) can be solved by parametrization of the boundary of $\mathcal{R}$, given in eqn. (12). Thus, for fixed s, the algorithm that solves (22) is the same as the one that solves (9), except that $\psi_{\mathrm{KL}}(x, y)$ is replaced by $\psi_{\mathrm{C}}(s, x, y)$. Then, the function $\psi_{\mathrm{C}}(s, x^\star(s), y^\star(s))$ can be evaluated using this algorithm and problem (21) is solvable by, e.g., a grid search on $[0, 1]$.

The projection phase of R–C is the same as the projection phase of R–KL and the steps in the refinement phase of R–C are the same as the ones in the refinement phase of R–KL (with $f_{\mathrm{KL}}(E, m_0, m_1)$ replaced by $\max_{s \in [0,1]} f_{\mathrm{C}}(s, E, m_0, m_1)$). Similarly as in III-A3, in the refinement phase, for a given selection matrix E, we have to find $f_{\mathrm{C, worst}}(E) := \min_{m_0 \in \mathcal{E}(\widehat{m}_0, k_0\, S_0^{-1}), m_1 \in \mathcal{E}(\widehat{m}_1, k_1\, S_1^{-1})} \max_{s \in [0,1]} f_{\mathrm{C}}(s, E, m_0, m_1)$. Applying again the minimax theorem, we first exchange the order of min and max in $f_{\mathrm{C, worst}}$; then, for fixed s, by applying the analogous transformation as between (18) and (19), we find

$$\min_{\substack{m_0 \in \mathcal{E}(\widehat{m}_0, k_0\, S_0^{-1}), \\ m_1 \in \mathcal{E}(\widehat{m}_1, k_1\, S_1^{-1})}} (m_1 - m_0)^\top E \left(s E^\top S_0 E + (1 - s) E^\top S_1 E \right)^{-1} E^\top (m_1 - m_0), \tag{24}$$

by solving the equivalent p-dimensional QCQP (25) (with equal minimum):

$$\min_{\substack{m_0' \in \mathcal{E}\left(Q^\top (E^\top S_0 E)^{-1/2} E^\top \widehat{m}_0, k_0\, I_p\right), \\ m_1' \in \mathcal{E}\left(Q^\top (E^\top S_0 E)^{-1/2} E^\top \widehat{m}_1, k_1\, \Lambda^{-1}\right)}} (m_1' - m_0')^\top (s I_p + (1 - s)\Lambda)^{-1}(m_1' - m_0'). \tag{25}$$

Denote by $m_i'(s)$ the optimal value of m_i', $i = 0, 1$. The function $f_{\mathrm{C}}'(s, E, m_0'(s), m_1'(s))$,

$$\begin{aligned} f_{\mathrm{C}}'(s, E, m_0'(s), m_1'(s)) \;=\; & \frac{s(1 - s)}{2}(m_1'(s) - m_0'(s))^\top (s I_p + (1 - s)\Lambda)^{-1}(m_1'(s) - m_0'(s)) \\ & - \frac{1}{2} \log \frac{\left| E^\top S_0 E \right|^s \left| E^\top S_1 E \right|^{1-s}}{\left| s E^\top S_0 E + (1 - s) E^\top S_1 E \right|} \end{aligned}$$

is concave w.r.t. s (as a minimum of concave functions); we find its maximum over $s \in [0, 1]$, which is equal to $f_{\mathrm{C, worst}}(E)$, via a bisection method.

C. Complexities of R–KL and R–C

The complexity of both R–KL and R–C algorithm is $\mathcal{O}(n^3 p + n p^4)$, although the hidden constant in R–C is larger than the one in R–KL. The least computational effort is required for the projection phase 2, which for both R–C and R–KL is $\mathcal{O}(n^2)$ and is dominated by complexities of the other two phases. It can be shown that Phase 1 has complexity $\mathcal{O}(n^3 p)$ and Phase 3 has complexity $\mathcal{O}(n p^4)$.

IV. SENSOR SELECTION ALGORITHMS: NO UNCERTAINTIES CASE

In this section, we address a special, yet important case, when there are no uncertainties in the mean vectors and the problems (4) and (5) simplify by dropping the inner minimizations. We first remark that algorithms R–KL and R–C can readily solve the simplified versions of (4) and (5). However, we derive

in this section a more efficient algorithm. We exploit the structure of the problem and the knowledge of exact distribution parameters m_i, $i = 0, 1$ (more specifically, their difference $m_1 - m_0$), and we resort to a heuristic approach (as detailed below), to reduce the computational load of the relaxation phase of R–KL and R–C algorithm, while keeping the second and the third phase the same[1]. The key to reducing the complexity of the relaxation phase is a *simple, analytic solution* of the relaxed, Stiefel problem, in the special case of equal mean values. We refer to the overall simplified algorithms as MD–KL (mean-difference based KL algorithm), and MD–C (mean-difference based C algorithm).

A. Kullback-Leibler based selection without uncertainties: The mean-difference KL algorithm (MD–KL)

In this subsection, we only explain the relaxation phase of MD–KL, as the other two phases are the same as in R–KL. We first consider a special case of equal mean values of the problem (8) (without inner minimization), and we show that this problem has a simple analytic solution. Based on the solution for the equal means, we derive an algorithm that solves the general case.

1) Relaxation phase: the case $m_0 = m_1$: Consider the problem (8) when there is no uncertainty in the mean values (uncertainty ellipsoids shrink to a point, by letting k_0, $k_1 = \infty$, and inner minimization drops from the problem). First, remark that, when $k_i = \infty$, $m_1 = m_0$, (8) can be written as:

$$
\begin{aligned}
&\text{maximize} && \tfrac{1}{2}\left(\mathrm{tr}(P^\top S P) - \log\left|P^\top S P\right| - p\right) \\
&\text{subject to} && P^\top P = I_p,
\end{aligned}
\tag{26}
$$

where $S = S_0^{-1/2} S_1 S_0^{-1/2}$. This equivalence can be shown by noting that the constraint $E^\top E = I_p$ in (8) can be replaced by $E^\top S_0 E = I_p$, and by introducing the new variable $P = S_0^{1/2} E$.

The objective function in (26) can be further simplified to $\sum_{i=1}^{p} \frac{1}{2}\phi_{\mathrm{KL}}\left(\lambda_i\left(P^\top S P\right)\right)$, where $\phi_{\mathrm{KL}}(x) = x - \log x - 1$ and λ_i denotes the i-th largest eigenvalue. We show in Appendix C that a solution $P^\star$ of (26) is given by the a of p orthonormal eigenvectors that correspond to the p largest eigenvalues of the matrix $S - \log S$, where $\log S$ is the matrix logarithm, given by $\log S = U \mathrm{Diag}\left(\log \lambda_1(S), ..., \log \lambda_n(S)\right) U^\top$. Here, U denotes the matrix of eigenvectors of S, and $\mathrm{Diag}(z_1, ..., z_n)$ is the diagonal matrix with the diagonal entries equal to z_i.

We give the intuition behind the solution of (26). Recall that the matrix E is chosen such that the projection of the covariance matrix S_0 equals $E^\top S_0 E = I_p$. Then, the projection of S_1, $E^\top S_1 E$, equals $P^\top S P$. Thus, the further from point 1 are the eigenvalues of $P^\top S P$, the more separated are the projected distributions. The function $\phi_{\mathrm{KL}}(\cdot)$ measures the distance from 1 in this sense.

---

[1]We remark that the refinement phase of R–C and R–KL simplifies significantly in the no uncertainties case, as in this case computing $f_{\mathrm{KL, worst}}(E)$ boils down to computing $f_{\mathrm{KL}}(E)$, i.e. there is no need to solve intermediate minimization problems (see (18) and (24)).

2) Relaxation phase: general case: The main idea behind the relaxation phase of MD–KL is the following heuristic: set one column of the solution Stiefel matrix E in the direction of the vector $m_1 - m_0$, i.e., in the direction of the difference of the distribution means. The remaining $p - 1$ columns of E are then obtained in the following way: we project the distribution parameters m_i, S_i, $i = 0, 1$, onto the orthogonal complement of $m_1 - m_0$, and then solve an ($p-1$ dimensional) instance of (8) when $k_i = \infty$, $i = 0, 1$, with the projected distribution parameters. This $p - 1$-dimensional instance of (8) is in fact the special one, with equal means, and is hence very efficiently solved, as explained above. The relaxation phase of MD–KL is summarized in Algorithm 2.

We give the intuition behind the heuristic choice of $m_1 - m_0$ as the direction of the first column of E: the Euclidean distance between the means $E^\top m_1$ and $E^\top m_0$ of the projected distributions is the maximal possible (and equal to $\|m_1 - m_0\|$), when one column of E lies in the direction of $m_1 - m_0$. Another justification is that choosing a column of E in the direction $m_1 - m_0$ provides an optimal solution of problem (8) when $k_i = \infty$, $i = 0, 1$, and $S_0 = S_1 = I_n$.

Algorithm 2: MD–KL algorithm: Relaxation phase

1: Set $e_1 := (m_1 - m_0) / \|m_1 - m_0\|$
2: Compute $U \in \mathbb{R}^{n \times (n-1)}$, an orthonormal basis for the orthogonal complement of e_1
3: Compute $S' = \left(U^\top S_0 U\right)^{-1/2} U^\top S_1 U \left(U^\top S_0 U\right)^{-1/2}$
4: Set the columns of P to be equal to $p - 1$ maximal eigenvectors of $S' - \log S'$
5: $[e_2, e_3, ..., e_p] := U \left(U^\top S_0 U\right)^{-1/2} P$, $E := [e_1, e_2, ..., e_p]$

B. Chernoff based selection without uncertainties: The mean-difference C algorithm (MD–C)

1) Relaxation phase: the case $m_0 = m_1$: The counterpart of problem (26) for the C criterion is:

$$\begin{aligned} \text{maximize} \quad & \max_{s \in [0,1]} \; -\log \frac{|P^\top S P|^{1-s}}{|s I_p + (1-s) P^\top S P|}, \\ \text{subject to} \quad & P^\top P = I_p \end{aligned} \tag{27}$$

The objective in (27) can be written as $\sum_{i=1}^p \phi_C \left(s, \lambda_i \left(P^\top S P\right)\right)$, where $\phi_C(s, x) := \log(s + (1-s)x) - (1-s) \log x$. Similarly to the proof of (26) (Appendix C), it can be shown that, for a fixed s, a maximizer of the objective in (27) w.r.t. $P^\top P = I_p$ is a set of p maximal eigenvectors of $\log\left(s I_n + (1-s)S\right) - (1-s) \log S$. Further, it can be shown, using unimodularity of $\phi_C(s, \cdot)$, that the corresponding p maximal eigenvalues must be of the form: $\lambda_1(S), \ldots, \lambda_{j(s)}(S), \; \lambda_{n-p+j(s)+1}(S), \ldots, \lambda_n(S)$, for some index $j(s) \in \{0, 1, ..., p\}$ (that depends on s). Thus, the solution to (27) is given by:

$$\max_{s \in [0,1]} \sum_{i \le j(s)} \phi_C\left(s, \lambda_i(S)\right) + \sum_{i > j(s)} \phi_C\left(s, \lambda_{n-p+i}(S)\right) = \max_{s \in [0,1]} \max_{j=0,...,p} \sum_{i \le j} \phi_C\left(s, \lambda_i(S)\right) + \sum_{i > j} \phi_C\left(s, \lambda_{n-p+i}(S)\right).$$

$$\tag{28}$$

Finally, switching the order of maximization (w.r.t. s and j) in (28) yields the procedure to solve (27) given in Algorithm 3. The general case of the relaxation phase of MD–C algorithm is the same as the

---

Algorithm 3: Procedure for solving (27) when $m_1 = m_0$

1: Set $\phi_C^\star = 0$
2: **for** $j = 0$ to p **do**
3:   if $j = 0$ $x = (\lambda_{n-p+1}(S), \ldots, \lambda_n(S))$, else $x = (\lambda_1(S), \ldots, \lambda_j(S), \lambda_{n-p+j+1}(S), \ldots, \lambda_n(S))$
4:   Find $s(x) \in \arg \max_{s \in [0,1]} \sum_{i=1}^p \phi_C(s, x_i)$ by Newton method
5:   Compute $\phi = \sum_{i=1}^p \phi_C(s(x), x_i)$
6:   if $\phi > \phi_C^\star$, then $j^\star = j$, $s^\star = s(x)$, $\phi_C^\star = \phi$
7: **end for**
8: $x^\star = (\lambda_1(S), \ldots, \lambda_{j^\star}(S), \lambda_{n-p+j^\star+1}(S), \ldots, \lambda_n(S))$; $s^\star = s^\star$;
   $P^\star$ is the set of orthonormal eigenvectors of S corresponding to the eigenvalues of S given in $x^\star$.

---

one in MD–KL given in 2, except that, in step 4), it calls the procedure given by Algorithm 3.

C. Complexities of MD–KL and MD–C

We briefly comment on the complexity of MD–KL and MD–C. The complexity of both MD–KL and MD–C is $\mathcal{O}(n^3 + np^3)$, although the hidden constant in MD–C is larger than the one in MD–C. The main computational burden in the relaxation phase of MD–KL is computation of the orthogonal complement U of $m_1 - m_0$, and the subsequent eigenvalue decomposition of the auxiliary matrix S (see Algorithm 2): the total is of order $\mathcal{O}(n^3)$. The relaxation phase of MD–C requires more computational effort; besides finding U and the eigenvalue decomposition of S, each of the p steps in Algorithm 3 requires finding the optimal s by, e.g., Newton method. However, the number of operations in p Newton runs is still dominated by the number of operations to find U. Therefore, the complexity of the relaxation phase of MD–C is $\mathcal{O}(n^3)$. Finally, it can be shown that, for both MD–KL and MD–C, the refinement phase is of complexity $\mathcal{O}(np^3)$.

V. NUMERICAL STUDIES: TESTING THE OPTIMIZATION CRITERIA

This subsection tests how good are the Kullback-Leibler and the Chernoff distance as optimality criteria for sensor selection. To this end, we want to compare the sensor selections that optimize the KL and C distances with: 1) the sensor selection that minimizes the Bayes probability of error (Bayes optimality); 2) the sensor selection that minimizes the probability of miss subject to a given probability of false alarm (Neyman-Pearson optimality). We find numerically the Bayes optimal and the Neyman-Pearson optimal sensor selections by Monte Carlo simulations. With respect to Bayes optimality, for each possible (out of $\binom{n}{p}$) sensor selection, we estimate the Bayes probability of error P_e by Monte Carlo simulations with $100,000$ instantiations of the maximum-likelihood detector tests (with zero treshold), with equal prior

probabilities. With respect to Neyman-Pearson optimality, for the fixed probability of false alarm P_{FA}, we find the sensor selection that maximizes the probability of detection P_{D}. This is achieved by estimating the receiver operating characteristic (ROC) curve (by Monte Carlo simulations with $20,000$ instantiations of the likelihood ratio tests) in the neighborhood of P_{FA} in a fine grid, and then interpolating the ROC curve (i.e., P_{D}) at the desired point P_{FA}. This is done for each possible selection and the selection with maximal obtained P_{D} is set as Neyman-Pearson optimal.

Table I shows the Bayes probability of error for: 1) the sensor selection that maximizes KL distance (KL); 2) the sensor selection that maximizes C distance (C); 3) the best sensor selection (that minimizes P_{e}); 4) the worst sensor selection (that maximizes P_{e}); and 5) the average P_{e} over all selections. Table I (left) is for $n = 12$ and $p = 2, 3, 4, 5$ and Table I (right) is for $n = 15$ and $p = 2, 3, 4, 6$. We can see that the C selection matches the best selection. KL selection is in many cases very close or equal to C and best selections in P_{e}. As could be predicted by theory (Chernoff lemma), the C selection is better than the KL selection in terms of P_{e}.

Comparison of: 1) C, KL and best selections; with 2) worst and average selections justifies the sensor selection problem; namely, by finding the optimal selection, P_{e} can be for an order of magnitude smaller than for the average selection. (See Table 1 and Figure 1-right.)

Table II shows the probability of detection for: 1) the KL-optimal selection (KL); 2) the C-optimal selection (C); and 3) the Neyman-Pearson (NP) optimal selection, for $n = 15$, $p = 2, 3, 4, 5$. We can see that, for $p = 2$ and $p = 3$, both KL and C selections match the NP optimal selections, and for $p = 4$ and $p = 5$, P_{D} for KL and C selections is at most 3.5% from the optimum ($p = 4$, $P_{\text{FA}} = 0.005$).

Figure 1 (left) plots the ROC curves for all possible selections, for $n = 5$ and $p = 2$. We plot the ROC curves for: 1) KL-optimal selection; 2) C-optimal selection; 3) the pointwise envelope of all possible curves (Neyman-Pearson optimal). Remark that the Neyman-Pearson optimal curve (envelope) is not obtained for a single selection; in different regions, it corresponds to different selections. We can see that for lower values of P_{FA}, KL selection is optimal; for higher values of P_{D}, C-selection is optimal.

Figure 1 (right) plots the ROC curve for all possible $\binom{n}{p} = 1365$ selections for a larger example, with $n = 15$ and $p = 4$. Interestingly, we can see that the KL and C selections are very close to the optimum, in whole range of P_{FA}. In addition, we plot the average of the ROC curves (pointwise average of P_{D} for each fixed P_{FA}). This average curve thus represents what performance would be, on average, achieved, if we choose a subset of sensors uniformly at random. We can see that there is a a large gain of the C and KL selections over this average curve; thus, selecting the optimal, rather than a random subset of sensors, provides large performance gain.

TABLE I

P_{e} FOR KL, C AND OPTIMAL SELECTION; LEFT: $n = 12$, $p = 2, 3, 4, 5$; RIGHT: $n = 15$, $p = 2, 3, 4, 6$

P_{e}	$p = 2$	$p = 3$	$p = 4$	$p = 5$		P_{e}	$p = 2$	$p = 3$	$p = 4$	$p = 6$
KL	0.100	0.078	0.052	0.046		KL	0.086	0.061	0.037	0.022
C	0.100	0.078	0.052	0.043		C	0.085	0.051	0.029	0.012
Bayes-best	0.100	0.078	0.052	0.043		Bayes-best	0.085	0.051	0.029	0.012
worst	0.457	0.439	0.396	0.319		worst	0.480	0.440	0.396	0.311
average	0.275	0.216	0.170	0.134		average	0.240	0.180	0.136	0.077

TABLE II

P_{D} FOR $P_{\mathrm{FA}} = [0.005\,0.03\,0.1]$, $n = 15$, $p = 2, 3, 4, 5$ FOR KL, C AND OPTIMAL SELECTION

P_{D}	$P_{\mathrm{FA}} = 0.005$			$P_{\mathrm{FA}} = 0.03$			$P_{\mathrm{FA}} = 0.1$		
	KL	C	NP	KL	C	NP	KL	C	NP
$p = 2$	0.720	0.720	0.720	0.889	0.889	0.889	0.961	0.961	0.961
$p = 3$	0.812	0.812	0.812	0.941	0.941	0.941	0.985	0.985	0.985
$p = 4$	0.868	0.838	0.868	0.962	0.967	0.970	0.990	0.995	0.995
$p = 5$	0.891	0.898	0.903	0.968	0.981	0.983	0.991	0.994	0.996

Finally, we remark that, in extensive simulations, we observe similar behavior as in representative Tables I and II, and Figure 1 (left and right). That is, the KL and C selections are very close to optimal and even equal to optimal in a certain range of P_{FA}. We also report that C-selection is generally better than KL for large P_{D}'s (upper right part of the ROC curve,) while KL is generally better for low P_{FA} (lower left in the ROC). Finally, improvement of KL over C for low P_{FA} is smaller than the improvement of C over KL for large P_{D}.
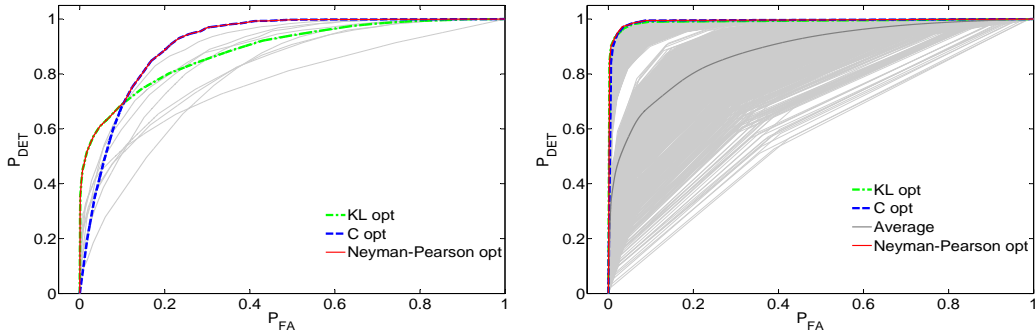

Fig. 1.   Left: $n = 5$, $p = 2$; ROC curves for all possible sensor selections (gray); the ROC curves for the KL-optimal and the C-optimal sensor selections, as well as the Neyman-Pearson optimal "envelope" are represented in different color. Right: $n = 15$, $p = 4$, Figure plots the same ROC curves as on the left; in addition, the Figure plots the average ROC among all sensor selections.

VI. NUMERICAL STUDIES: TESTING THE ALGORITHMS

This subsection tests our algorithms for solving the sensor selection problem. Subsection VI-A shows that the algorithms R–KL and R–C show near optimal performance in the presence of uncertainties (i.e., for solving the problems (4) and (5)), Subsection (VI-B) shows that R–KL and R–C have near optimal performance when also applied to the case of no uncertainties (i.e., when $k_0 = k_1 = \infty$ in (4) and (5)). This subsection also shows that MD–KL (resp. MD–C) has comparable performance to R–KL (resp. R–C), while reducing the computational time.

For smaller problem instances, i.e., for smaller values of n and p, we compare the solutions produced by our algorithms with the optimum obtained by exhaustive search. For larger examples, when the optimum is infeasible to compute, we generate randomly a number of sensor selections; then, we compare the selection obtained by our algorithms with the best among generated random selections.

For a fixed size of the problem (for fixed pair (n, p)), we generate randomly K_{exper} instances of the problem parameters $(m_i, S_i, \ i = 0, 1)$; for the case with uncertainties, $K_{\mathrm{exper}} = 50$, and for the case without uncertainties, $K_{\mathrm{exper}} = 200$. In the case with uncertainties, the parameters k_0 and k_1 are chosen such that the norm of the mean vector drift does not exceed $0.15\|m_1 - m_0\|$, i.e., we set them as $k_i = |S_i| / (0.15\|m_1 - m_0\|)^2, \ i = 0, 1$. Based on K_{exper} solved problems of fixed size (n, p), we create statistics on how our algorithms behave for a given size of the problem. The entries of the (nominal) mean vectors m_1, m_0 are drawn independently from the standard normal distribution. The covariance matrices S_1, S_0, are generated in two steps: 1) we generated a symmetric matrix F by drawing its entries independently from the standard normal distribution (we set $F_{ij} = F_{ji}, \ i \neq j$); 2) we generate the matrix S_i to have the same eigenvectors as F, and the eigenvalues of S_i are equal to the absolute values of the eigenvalues of F.

A. Testing the R–KL and R–C algorithms: The case with uncertainties

Testing the robustness against the uncertainty in distribution means. Table III (left) shows the performance of the R–KL and R–C algorithms, for solving (4) and (5), respectively. We evaluate the optimal solutions $f^{\mathrm{OPT-KL}}$ and $f^{\mathrm{OPT-C}}$ by exhaustive search. We then compute the ratio $r_{\mathrm{R-KL}} := \frac{f^{\mathrm{R-KL}}}{f^{\mathrm{OPT-KL}}}$ (resp. $r_{\mathrm{R-C}} := \frac{f^{\mathrm{R-C}}}{f^{\mathrm{OPT-C}}}$) that says how close is the solution value obtained by R–KL (resp. R–C) to the optimum. Table III (left) shows the results for $n = 10, 12, 15$ and $p = 3$. We present the maximum (max), the average (avg), the minimum (min), and the standard deviation (dev) of the ratio $r_{\mathrm{R-KL}}$ over $K_{\mathrm{exper}} = 50$ problem instances, for each pair (n, p). We can see that both R–KL and R–C show very good performance; the R–C shows better performance than R–KL. The maximal value of both $r_{\mathrm{R-KL}}$ and $r_{\mathrm{R-C}}$ in all the examples is equal to 1. With the R–KL algorithm, the average value of

TABLE III
LEFT: STATISTICS FOR $r_{\mathrm{R-KL}}$ AND $r_{\mathrm{R-C}}$, SMALLER EXAMPLES: $n = 10, 12, 15$, $p = 3$, $K_{\mathrm{exper}} = 50$; RIGHT: STATISTICS
FOR $\rho_{\mathrm{R-KL}}$ AND $\rho_{\mathrm{R-C}}$, $n = 50$, $p = 5$, $K_{\mathrm{exper}} = 50$

	max	avg	min	dev
$r_{\mathrm{R-KL}}$				
$n = 10$	1.000	0.964	0.606	0.090
$n = 12$	1.000	0.918	0.551	0.140
$n = 15$	1.000	0.939	0.512	0.118
$r_{\mathrm{R-C}}$				
$n = 10$	1.000	0.981	0.7862	0.051
$n = 12$	1.000	0.982	0.834	0.037
$n = 15$	1.000	0.961	0.595	0.095

	max	avg	min	dev
$\rho_{\mathrm{R-KL}}$	1.777	1.267	0.817	0.191
$\rho_{\mathrm{R-C}}$	1.868	1.277	1.005	0.166

$r_{\mathrm{R-KL}}$ is always above 91.8%, and the minimum value is always above 51%; with the R–C algorithm, the average of $r_{\mathrm{R-C}}$ is always above 96%, and the minimum is always above 59%. Also, we report that the refinement phase for both R–KL and R–C algorithm improves significantly the solution. Typically, the average value of $r_{\mathrm{R-KL}}$ before refinement is about 65%, while the one for R–C is slightly higher and is about 70%[2].

Table III (right) shows the performance of R–KL and R–C for a larger example $n = 50$, $p = 5$, for which it was infeasible to find the optimal solution by exhaustive search. We thus randomly generate 2500 sensor selections, and we evaluate the quantity $f^{\mathrm{BEST-RAND-KL}}$–the maximal objective function in (4) over all 2500 randomly generated selections. Define the ratios $\rho_{\mathrm{R-KL}} := \frac{f^{\mathrm{R-KL}}}{f^{\mathrm{BEST-RAND-KL}}}$. (Analogously define the corresponding ratios for the Chernoff-based sensor selection.) We report that, to obtain $f^{\mathrm{BEST-RAND-KL}}$ (for 2500 50×5 selection matrices) it takes about 10 times longer than for R–KL algorithm to find a solution. Also, to obtain $f^{\mathrm{BEST-RAND-C}}$ it takes about 6 times longer than for R–C algorithm to find a solution. From Table III (right) we can see that R–KL in many cases outperforms random strategy: with significant savings in time (90%) it gives a 26% larger objective function on average. More importantly, R–C always stays above the random strategy in terms of the objective function, with computational savings of 83%.

---

[2]Denote by $E^{\bullet}$ the selection matrix obtained prior to the refinement phase of R–KL, and by $f_{\mathrm{KL}}^{\star}$ the KL distance for the optimal selection. The average value of $f_{\mathrm{KL}}(E^{\bullet})/f_{\mathrm{KL}}^{\star}$, which is typically about 65%, is relatively high (although far from optimum). Namely, if we choose a selection matrix E_r at random, then the (simulated) average ratio $f_{\mathrm{KL}}(E_r)/f_{\mathrm{KL}}^{\star}$ (the average is over both randomly generated data m_i, S_i and E_r) is typically about 30%, which is more than twice smaller than $f_{\mathrm{KL}}(E^{\bullet})/f_{\mathrm{KL}}^{\star}$. Similar behavior is observed for the C distance case, also.

B. Testing the algorithms: The case without uncertainties

This subsection tests the algorithms R–KL, MD–KL, R–C, and MD–C in the no uncertainties case. Simulations show that R–KL and MD–KL show comparable (both very good) performance w.r.t. accuracy, while MD–KL has smaller computational cost. Thus, MD–KL is a better algorithm in the no uncertainties case, while MD–KL does not apply in the presence of uncertainties. Similar conclusions hold for R–C and MD–C algorithms.

Smaller examples: Comparison with the global optimum. Table IV shows the performance of the R–KL and MD–KL algorithms for solving (4) in the case without uncertainties ($k_0 = k_1 = \infty$). For fixed n and p, we generate $K_{\mathrm{exper}} = 200$ sets of problem parameters $m_i, S_i, i = 0, 1$; for each fixed $m_i, S_i, i = 0, 1$, we evaluate the ratios $r_{\mathrm{R-KL}}$, $r_{\mathrm{MD-KL}}$, $r_{\mathrm{R-C}}$, and $r_{\mathrm{MD-C}}$. Table IV shows the maximum (max), the average (avg) and the minimum (min) of the quantities $r_{\mathrm{R-KL}}$ and $r_{\mathrm{MD-KL}}$ over all $K_{\mathrm{exper}} = 200$ experiments. The standard deviation in these experiments varies from 0.03 to 0.06. We can see from the table that the maximum of both $r_{\mathrm{R-KL}}$ and $r_{\mathrm{MD-KL}}$ is always 1. Table IV demonstrates very good performance of both R–KL and MD–KL algorithms. On average, $r_{\mathrm{R-KL}}$ and $r_{\mathrm{MD-KL}}$ are always (for each pair n, p) above 97.7%; the minimum is always above 67.2%. We see that R–KL and MD–KL algorithms have comparable performance with respect to (near)optimality, while MD–KL has smaller computational cost (We note, however, that MD–KL does not apply for the case with uncertainties in distribution means.) We performed similar simulations for R–C and MD–C algorithms also. Results are omitted due to the lack of space, but we report that R–C and MD–C show mutually similar (and near optimal) performance. Also, performance of R–C and MD–C is very similar to performance of R–KL and MD–KL.

Larger examples: Comparison with the best randomly generated selection. We now consider larger n and p, when computing the optimum by the exhaustive search is infeasible. Similarly as in subsection VI-A (larger examples), we randomly generate 10^5 sensor selections and find the maximal objective function $f^{\mathrm{BEST-RAND-KL}}$ over these selections. For a fixed pair n and p, we generate $K_{\mathrm{exper}} = 100$ sets of the distribution parameters $m_i, S_i, i = 0, 1$; for each set of the distribution parameters, we evaluate $\rho_{\mathrm{R-KL}}$, $\rho_{\mathrm{MD-KL}}$, and $\rho_{\mathrm{R-C}}$, $\rho_{\mathrm{MD-C}}$. We are also interested in the ratio of the computational time of the R–KL (resp. MD–KL) algorithm over the computational time of checking 10^5 random selections (and the analogous quantities for the Chernoff-based selection).

Table VI shows the average (avg), the maximal (max), and the minimal (min) values of $\rho_{\mathrm{R-KL}}$ and $\rho_{\mathrm{MD-KL}}$ over 100 data sets $m_i, S_i, i = 0, 1$; Table VII (left) shows the computational time ratios. We can see that both R–KL and MD–KL outperform best random selection strategy, as they achieve larger

TABLE IV
STATISTICS FOR $r_{\mathrm{R-KL}} = f^{\mathrm{R-KL}}/f^{\mathrm{OPT-KL}}$ AND $r_{\mathrm{MD-KL}} = f^{\mathrm{MD-KL}}/f^{\mathrm{OPT-KL}}$

	$p = 3$			$p = 4$			$p = 5$		
	max	avg	min	max	avg	min	max	avg	min
$r_{\mathrm{R-KL}}$									
$n = 20$	1.000	0.990	0.789	1.000	0.985	0.688	1.000	0.977	0.672
$n = 30$	1.000	0.989	0.830	1.000	0.988	0.826	1.000	0.983	0.795
$n = 40$	1.000	0.989	0.729	1.000	0.985	0.842	1.000	0.983	0.817
$r_{\mathrm{MD-KL}}$									
$n = 20$	1.000	0.992	0.744	1.000	0.982	0.688	1.000	0.975	0.672
$n = 30$	1.000	0.989	0.809	1.000	0.987	0.832	1.000	0.981	0.742
$n = 40$	1.000	0.985	0.729	1.000	0.980	0.802	1.000	0.981	0.834

objective function while reducing computational time. For example, for $n = 80$, $p = 0.1 \times n = 8$, R–KL can give 54% larger objective function, while it has 3 times smaller computational time. MD–KL is better than R–KL, as it shows comparable performance in terms of the objective function; at the same time, it significantly reduces the computational time. Thus, for sensor selection without uncertainties, MD–KL is a very good tool that can handle large problem instances. Similar conclusions hold for the Chernoff based algorithms (see Table VII and Table VIII (right)). Chernoff based algorithms have larger computational time than Kullback-Leibler based algorithms, which is expected due to maximization over s (see (5)). Section V shows generally that the Chernoff criterion has some advantage over the Kullback-Leibler criterion, which trades off the computational requirements for these criteria.

VII. CONCLUSIONS

In this paper, we addressed the problem of finding the most informative subset of p out of n sensors, for the task of deciding between the two possible hypothesis on the monitored environment. We proposed two different information theoretic criteria for the best sensor selection: the Kullback-Leibler distance and the Chernoff distance between the distributions induced by the selected sensors. We tackled the case where the distributions are Gaussian, but the mean vectors are known only up to confidence regions. We formulated the corresponding maxmin optimization problems, and developed the R–KL and R–C algorithms, that efficiently solve the problems with complexity $\mathcal{O}(n^3 p + n p^4)$. We also addressed the case when the mean vectors are known and, for this case, we exploited the structure of the problems to develop more efficient algorithms, MD–KL and MD–C, of complexity $\mathcal{O}(n^3 + n p^3)$. We performed Monte-Carlo based experiments to test both the proposed sensor selection criteria and the sensor selection algorithms.

TABLE V
STATISTICS FOR $\rho_{\text{R}-\text{KL}} = f^{\text{R}-\text{KL}}/f^{\text{BEST}-\text{RAND}-\text{KL}}$ AND $\rho_{\text{MD}-\text{KL}} = f^{\text{MD}-\text{KL}}/f^{\text{BEST}-\text{RAND}-\text{KL}}$

	$p = 0.1 \times n$			$p = 0.2 \times n$			$p = 0.3 \times n$		
	avg	max	min	avg	max	min	avg	max	min
$\rho_{\text{R}-\text{KL}}$									
$n = 50$	1.069	1.246	0.855	1.225	1.577	1.072	1.274	1.733	1.071
$n = 80$	1.298	1.544	1.162	1.474	1.860	1.233	1.472	1.962	1.229
$n = 100$	1.427	1.705	1.169	1.556	1.739	1.337	1.598	1.894	1.378
$\rho_{\text{MD}-\text{KL}}$									
$n = 50$	1.072	1.246	0.919	1.223	1.464	1.002	1.265	1.684	1.077
$n = 80$	1.299	1.544	1.130	1.475	1.782	1.248	1.468	2.002	1.225
$n = 100$	1.429	1.705	1.169	1.563	1.864	1.384	1.617	1.973	1.374

TABLE VI
STATISTICS FOR $\rho_{\text{R}-\text{C}} = f^{\text{R}-\text{C}}/f^{\text{BEST}-\text{RAND}-\text{C}}$ AND $\rho_{\text{MD}-\text{C}} = f^{\text{MD}-\text{C}}/f^{\text{BEST}-\text{RAND}-\text{C}}$

	$p = 0.1 \times n$			$p = 0.2 \times n$			$p = 0.3 \times n$		
	avg	max	min	avg	max	min	avg	max	min
$\rho_{\text{R}-\text{C}}$									
$n = 50$	1.074	1.214	0.992	1.182	1.341	1.072	1.194	1.307	1.095
$n = 80$	1.262	1.454	1.132	1.357	1.621	1.199	1.338	1.464	1.206
$n = 100$	1.375	1.517	1.251	1.447	1.616	1.299	1.402	1.629	1.287
$\rho_{\text{MD}-\text{C}}$									
$n = 50$	1.074	1.214	0.991	1.182	1.341	1.038	1.195	1.307	1.100
$n = 80$	1.262	1.471	1.132	1.357	1.621	1.194	1.338	1.464	1.212
$n = 100$	1.375	1.517	1.254	1.445	1.616	1.296	1.403	1.632	1.296

Numerical studies of the criteria show that the Kullback-Leibler based and the Chernoff based selections have near optimal performance, both in the Neyman-Pearson and Bayes sense. The performance of our algorithms was compared with 1) the optimal sensor selections, when the exhaustive search to compute them is feasible (smaller n and p); and 2) with best random selections, when n and p are large. Comparison with the exhaustive search shows that the proposed algorithms in many cases find the optimal selection and, on average, are at most 5% below the optimal value; at the same time, computational savings are significant. For larger problems, simulation results demonstrate that our algorithm outperforms random

TABLE VII
AVERAGE TIME RATIOS, K_{exper}=100; LEFT: KL ALGORITHMS; RIGHT: C ALGORITHMS

	$p = 10\%$	$p = 20\%$	$p = 30\%$		$p = 10\%$	$p = 20\%$	$p = 30\%$
R–KL				R–C			
$n = 50$	0.070	0.104	0.107	$n = 50$	0.105	0.146	0.153
$n = 80$	0.276	0.295	0.291	$n = 80$	0.472	0.549	0.507
$n = 100$	0.417	0.472	0.349	$n = 100$	0.875	0.990	0.721
MD–KL				MD–C			
$n = 50$	0.003	0.005	0.006	$n = 50$	0.002	0.004	0.006
$n = 80$	0.007	0.012	0.014	$n = 80$	0.006	0.011	0.014
$n = 100$	0.011	0.017	0.021	$n = 100$	0.010	0.017	0.021

searches, once an upper bound on computational time is set.

APPENDIX

A. Proof of NP hardness of optimization problems (4) *and* (5)

We will prove that both problems (4) and (5) are NP hard by reducing them to the maximal clique problem (MQP), which is known to be NP hard [25]. We first define the MQP. Consider an undirected, simple (i.e., without self-loops) graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, where $\mathcal{N}$ is the set of vertices with cardinality $|\mathcal{N}| = n$, and $\mathcal{E}$ is the set of undirected edges $\{i, j\}$, $|\mathcal{E}| = m$. A clique of the graph $\mathcal{G}$, of size p, is a complete subgraph of $\mathcal{G}$ that has p vertices. The decision version of MQP is as follows:

MQP: The maximal clique problem: "For given graph $\mathcal{G}$ and a positive integer p, $1 \leq p \leq n$, determine whether $\mathcal{G}$ has a clique of size at least p."

We conduct a reduction to MQP by attaching to a graph $\mathcal{G}$ a $n \times n$ matrix $S(\mathcal{G})$ of a special structure. Namely, we define a positive definite matrix $S(\mathcal{G})$ as follows: $[S(\mathcal{G})]_{ij} = 2n$, if $i = j$; $[S(\mathcal{G})]_{ij} = -1$, if $i \neq j$, and $\{i, j\} \in \mathcal{E}$; and $[S(\mathcal{G})]_{ij} = 0$, otherwise. The matrix $S(\mathcal{G})$ is positive definite because it has positive diagonal elements and it is strictly diagonally dominant. Now, fix an integer p, $1 \leq p \leq n$, and consider a set of matrices $\mathcal{A}_p$ defined as $\mathcal{A}_p = \left\{ A \in \mathbb{R}^{p \times p} : A = A^\top, A_{ii} = 2n, \forall i, A_{ij} \in \{0, -1\}, i \neq j \right\}$. Clearly, all matrices in $\mathcal{A}_p$ are positive definite, as they are strictly diagonally dominant, with positive diagonal entries. Denote by $\mathbf{1}_p$ the column vector with all entries equal to 1 and define the function $g : \mathcal{A}_p \to \mathbb{R}$ as $g(A) = \mathbf{1}_p^\top A^{-1} \mathbf{1}_p$. We have the following result on the matrices in $\mathcal{A}_p$.

Lemma 7 For all matrices $A \in \mathcal{A}_p$, there holds:

1) Denote $B := A^{-1}$. Then, for all i, j $B_{ij} \geq 0$.

2) If $A_{ij} = 0$, then $g\left(A - h_i h_j^\top - h_j h_i^\top\right) \geq g(A)$. (Recall that h_i denotes ith canonical vector.)

3) $g(A) \leq \frac{p}{2n-p+1}$, where the equality holds if and only if $A = A^\star := 2nI - 11^\top + I$.

Proof: The claim 1) in Lemma 7 follows from the fact that the matrix A is an M matrix [26], and, consequently, the matrix $B = A^{-1}$ has all entries greater than or equal to zero [26]. We now show the claim 2) in Lemma 7. Remark first that the function $g(\cdot)$ is convex and differentiable on the set of positive definite matrices. Applying the first order Taylor expansion lower bound at A, we get:

$$g\left(A - h_i h_j^\top - h_j h_i^\top\right) \geq g(A) + \mathrm{tr}\left(\nabla g(A)\left(-h_i h_j^\top - h_j h_i^\top\right)\right), \tag{29}$$

where $\nabla g(A)$ stands for the (matrix form) gradient of $g(\cdot)$ at A and is equal to $-A^{-1}\mathbf{1}_p \mathbf{1}_p^\top A^{-1}$. Now, by claim 1), all entries of A^{-1} are nonnegative, and, thus, the second term on the right hand side of the inequality (29) is nonnegative as well. This completes the proof of claim 2.

We proceed and prove the claim 3) in Lemma 7. From claim 2) we know that the more -1's a matrix $A \in \mathcal{A}_p$ has on its off-diagonal entries, the higher the value $g(A)$ can be. Therefore, $A^\star$ is a maximizer of g over the set $\mathcal{A}_p$. Also, it is straightforward to check that $g(A^\star) = \frac{p}{2n-p+1}$. We will show next that $A^\star$ is in fact the only maximizer of g (over $\mathcal{A}_p$). To show this, it suffices to show that, for any choice of $1 \leq i, j \leq p, i \neq j$, the following strict inequality holds:

$$g\left(A^\star + h_i h_j^\top + h_j h_i^\top\right) < g(A^\star). \tag{30}$$

To this end represent the matrix $h_i h_j^\top + h_j h_i^\top$ as $h_i h_j^\top + h_j h_i^\top = HCH^\top$, where $H = [h_i\, h_j] \in \mathbb{R}^{n \times 2}$ and $C \in \mathbb{R}^{2 \times 2}$, $C_{12} = C_{12} = 1$, $C_{11} = C_{22} = 0$. Using the matrix inversion lemma, we get:

$$g\left(A^\star + h_i h_j^\top + h_j h_i^\top\right) = \mathbf{1}_p^\top B^\star \mathbf{1}_p - \mathbf{1}_p B^\star H \left(C^{-1} + H^\top B^\star H\right)^{-1} H^\top B^\star \mathbf{1}_p. \tag{31}$$

After straightforward algebra, we obtain:

$$\mathbf{1}_p B^\star H \left(C^{-1} + H^\top B^\star H\right) H^\top B^\star \mathbf{1}_p = \frac{2n+1}{n+1+\frac{1}{2n-p+1}}.$$

Since this term is greater than zero for all $1 \leq p \leq n$, the inequality (30) follows. This completes the proof of claim 3) and the proof of the Lemma ∎

We proceed with the proof of Theorem (1). The decision version of (4), for $k_1 = k_0 = +\infty$, is:

D–KL: Decision version of (4) "For given data: 1) vectors $m_0, m_1 \in \mathbb{R}^n$; 2) positive definite matrices S_0 and S_1; 3) positive integer p, $p \leq n$; and 4) a number $f_{\mathrm{KL}}^\bullet$, determine whether there is a $n \times p$ sensor selection matrix E, such that $f_{\mathrm{KL}}(E)$ defined in eqn. (6) is at least $f_{\mathrm{KL}}^\bullet$."

We now reduce the MQP to D-KL. Consider a simple, undirected graph $\mathcal{G}$ and consider MQP of

determining whether there is a clique in $\mathcal{G}$ of the size at least p. Define the matrix $S(\mathcal{G})$ by $[S(\mathcal{G})]_{ij} = 2n$, if $i = j$; $[S(\mathcal{G})]_{ij} = -1$, if $i \neq j$, and $\{i, j\} \in \mathcal{E}$; and $[S(\mathcal{G})]_{ij} = 0$, otherwise. Consider an instance of D–KL, for some fixed p, with the following data: 1) $m_1 = \mathbf{1}_n$, $m_0 = \mathbf{0}_n$; 2) $S_1 = S_0 = S(\mathcal{G})$; 3) p; and 4) $f_{\mathrm{KL}}^{\bullet} = \frac{1}{2}\frac{p}{2n-p+1}$. Now, the answer to D–KL is YES (resp. NO) if and only if $\mathcal{G}$ has (resp does not have) a clique of size at least p. Thus, MQP problem is reduced to D–KL.

MQP problem reduces to the decision version of problem (5) (denoted by D–C), for $k_1 = k_0 = +\infty$, in a very similar way as it reduces to D–KL, by considering the instance of D–C with the data m_1, m_0, S_1, S_0, p, same as we considered for D–KL, but $f_{\mathrm{KL}}^{\bullet} = \frac{1}{2}\frac{p}{2n-p+1}$ is replaced by $f_{\mathrm{C}}^{\bullet} = \frac{1}{8}\frac{p}{2n-p+1}$.

B. Proof that the Kullback-Leibler and Chernoff distances are not submodular functions

See [20] for the definition of a submodular function. Consider two Gaussian distributions $\mathcal{N}(m_i, S_i)$, $i = 0, 1$, with parameters $m_0 = m_1$, $S_0 = I_3$ and $S_1 = I_3 + \epsilon(h_2 h_3^{\top} + h_3 h_2^{\top})$ $(h_2, h_3 \in \mathbb{R}^3)$, where $0 < \epsilon < 1$. Let $E_1 = h_1 \in \mathbb{R}^3$, $E_{13} = [h_1\, h_3]$, $E_{12} = [h_1\, h_2]$ and $E_{123} = I_3$. Computing the Kullback Leibler distance for E_1, E_{12} and E_{13}, we get $f_{\mathrm{KL}}(E_1) = f_{\mathrm{KL}}(E_{12}) = f_{\mathrm{KL}}(E_{13}) = 0$, whereas $f_{\mathrm{KL}}(E_{123}) = -\frac{1}{2}\log(1 - \epsilon^2) > 0$. Thus, $f_{\mathrm{KL}}(E_{13}) - f_{\mathrm{KL}}(E_1) < f_{\mathrm{KL}}(E_{123}) - f_{\mathrm{KL}}(E_{12})$ which proves that function f_{KL} is not submodular. The proof for Chernoff distance can be done in a similar way.

C. Solution to problem (26)

We consider the Lagrangian form for the problem (26), which equals $L(P, \mathcal{M}) = \frac{1}{2}\left(\mathrm{tr}(P^{\top}SP) - \log\left|P^{\top}SP\right| - p\right) + \mathrm{tr}\left(\mathcal{M}(P^{\top}P - I)\right)$, where $\mathcal{M}$ is a symmetric matrix of the Lagrange multipliers associated with the constraint $P^{\top}P = I_p$. (We note that, since S is positive definite and P is full rank, the matrix $P^{\top}SP$ is also positive definite, and, thus, invertible.) A necessary condition for P to be a maximizer of (26) is that P is a stationary point of the Lagrangian form. Lemma 8 characterizes the set of candidates for stationary points of the Lagrangian form.

Lemma 8 Let Q_P and Λ_P denote, respectively, the (orthogonal) matrix of eigenvectors and the matrix of eigenvalues of $P^{\top}SP$, i.e., $P^{\top}SP = Q_P\Lambda_P Q_P^{\top}$. Then, stationary points P of the Lagrangian $L(P, \mathcal{M})$ have the following form: $P = [U_1 A]Q_P^{\top}$, where U_1 is a matrix of some $p_1 \leq p$ eigenvectors of S, and the matrix A (that depends on S) is such that $A^{\top}SU_1 = 0$ and $A^{\top}SA = I_{p-p_1}$.

We now find the maximizer of (26) by searching over the set of stationary points of $L(P, \mathcal{M})$. By Lemma 8, a stationary point is of the form $P = [U_1 A]Q_P^{\top}$. Note that $\Lambda_P = [U_1 A]^{\top} S [U_1 A] = \Lambda_1 \oplus I_{p-p_1}$, where $\oplus$ denotes the matrix direct sum operation, and Λ_1 denotes the matrix of eigenvalues of S that correspond

to the eigenvectors U_1. Then, the value of the objective function in (26) at P equals:

$$f(P) = \frac{1}{2}\left(\operatorname{tr}(\Lambda_P) - \log|\Lambda_P| - p\right) = \frac{1}{2}\left(\operatorname{tr}(\Lambda_1) - \log|\Lambda_1| - p_1\right). \tag{32}$$

The quantity $f(P)$ in (32) equals the sum of the functions $\phi_{\mathrm{KL}}(x) = x - \log x - 1$ evaluated at some p_1 $(p_1 \leq p)$ eigenvalues of S. Clearly, since $\phi_{\mathrm{KL}}(x) > 0$ for $x \neq 1$, (32) is maximized for P for which $p_1 = p$ (A is an empty matrix), and PQ_P equals the set of the p maximal eigenvectors of $S - \log S$. As $f(\cdot)$ is invariant with respect to rotations ($f(PG) = f(P)$ for any orthogonal $p \times p$ matrix G), (26) is also maximized at $P' = PQ_P$, which we needed to show. We proceed with the proof of Lemma 8.

Proof: Taking the derivative of $L(P, \mathcal{M})$ with respect to P, setting the derivative to zero, and multiplying the corresponding equality from the left by $(I - PP^\top)$ yields:

$$(I - PP^\top)SP(I - (P^\top SP)^{-1}) = 0 \tag{33}$$

Since $(I - PP^\top)$ is a projection matrix onto the orthogonal complement of P, condition (33) is satisfied if and only if all columns of $SP(I - (P^\top SP)^{-1})$ belong to $\operatorname{span}(P)$, i.e., when $SP(I - (P^\top SP)^{-1}) = PR$, for some $R \in \mathbb{R}^{p \times p}$; multiplying the latter condition by $P^\top$ from the left, and using the fact that $P^\top P = I$, we get $R = P^\top SP - I$. Therefore, P satisfies:

$$SP(I - (P^\top SP)^{-1}) = P(P^\top SP - I). \tag{34}$$

Then, plugging in (34) the eigenvalue decomposition of $P^\top SP$, multiplying both sides in equality (34) by Q_P from the right, and using the identity $\Lambda_P - I = \Lambda_P(I - \Lambda_P^{-1})$, (34) implies:

$$SPQ_P(I - \Lambda_P^{-1}) = PQ_P\Lambda_P(I - \Lambda_P^{-1}). \tag{35}$$

Without loss of generality, we can reorder the columns of PQ_P, such that $\Lambda_P = \Lambda_1 \oplus I_{p-p_1}$, where Λ_1 collects the eigenvalues of Λ_P not equal to one. Then, condition (35) does not "see" the last $p-p_1$ columns of the matrix PQ_P, whereas the condition on the first p_1 columns of PQ_P, denoted by $(PQ_P)_1$, becomes: $S(PQ_P)_1 = (PQ_P)_1\Lambda_1$. Therefore, $(PQ_P)_1$ must be equal to a matrix U_1 of some p_1 eigenvectors of S. Summing up, a stationary point takes the form $PQ_P = [U_1 A]$, where $A^\top SA = I_{p-p_1}$, $A^\top SU_1 = 0$. ∎

REFERENCES

[1] L. Rueda and M. Herrera, "Linear dimensionality reduction by maximizing the Chernoff distance in the transformed space," *Pattern Recognition*, vol. 41, no. 10, pp. 3138–3152, October 2008.

[2] L. L. Scharf, *Statistical signal processing: detection, estimation and time series analysis*. New York: Addison-Wesley, 1991.

[3] D. Bajović, B. Sinopoli, and J. Xavier, "Sensor selection for hypothesis testing in wireless sensor networks: a Kullback-Leibler based approach," in *48th IEEE Conf. on Decision and Control*, Shanghai, China, Dec. 2009, pp. 1659 – 1664.

[4] ——, "Robust linear dimensionality reduction for hypothesis testing with application to sensor selection," in *47th Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, September 2009, pp. 363 – 370.

[5] C.-W. Ko, J. Lee, and M. Queyranne, "An exact algorithm for maximum entropy sampling," *Operations Research*, vol. 43, no. 4, pp. 684–691, July–August 1995.

[6] E. L. Lawler and N. Draper, "Branch-and-bound methods: a survey," *Operations Research*, vol. 14, no. 4, pp. 699–719, July–August 1966.

[7] A. L. Buczak, H. H. Wang, H. Darabi, and M. A. Jafari, "Genetic algorithm convergence study for sensor network optimization," *Information Sciences*, vol. 133, pp. 267–282, 2001.

[8] F. Zhao, J. Shin, and J. Reich, "Information-driven dynamic sensor collaboration for tracking applications," *IEEE Signal Processing Magazine*, vol. 19, no. 2, pp. 61 – 72, March 2002.

[9] M. Chu, H. Haussecker, and F. Zhao, "Scalable information-driven sensor querying and routing for ad hoc heterogenous sensor networks," *Int. Journ. High Performance Computing Applications*, vol. 16, no. 3, pp. 293–313, Aug. 2002.

[10] F. Bian, D. Kempe, and R. Govindan, "Utility based sensor selection," in *IPSN '06, 5th International Conference on Information Processing in Sensor Networks*, New York, NY, USA, 2006, pp. 11–18.

[11] S. Joshi and S. Boyd, "Sensor selection via convex optimization," *IEEE Transactions on Signal Processing*, vol. 57, no. 2, pp. 451–462, February 2009.

[12] D. Bickson and D. Dolev, "Distributed sensor selection using a truncated Newton method," submitted for publication, available on arXiv:0907.0931v2 [cs.IT].

[13] V. Kekatos and G. B. Giannakis, "Selecting reliable sensors via convex optimization," in *SPAWC 2010, IEEE International Workshop on Signal Processing Advances in Wireless Communications*, Marrakech, Morocco, June 2010.

[14] Z. Song, C. R. Sastry, N. C. Tas, and Y. Q. Chen, "Feasibility analysis on optimal sensor selection in cyber-physical systems," in *ACC '09, American Control Conference*, St. Louis, MO, June 2009, pp. 5368–5373.

[15] V. Fedorov, *Theory of optimal experiments*. New York: Academic, 1972.

[16] Z. Quan, S. Cui, K. Ma, and A. H. Sayed, "Optimal linear fusion for distributed detection via semidefinite programming," *IEEE Transactions on Signal Processing*, vol. 58, no. 4, pp. 2431–2436, April 2010.

[17] Z. Quan, S. Cui, and A. H. Sayed, "Optimal linear cooperation for spectrum sensing in cognitive radio networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, no. 1, pp. 28–40, Feb. 2008.

[18] T. M. Cover and J. A. Thomas, *Elements of information theory*. Wiley - Interscience, 2006.

[19] M. K. Ali, L. Rueda, and M. Herrera, "On the performance of Chernoff-distance-based linear dimensionality reduction techniques," in *Canadian AI 2006, 19th Conference of the Canadian Society for Computational Studies of Intelligence*, vol. 4013, Quebec City, Canada, June 2006, pp. 3467–3478.

[20] C. Guestrin, A. Krause, and A. P. Singh, "Near-optimal sensor placements in Gaussian processes," in *22nd International Conference on Machine learning*, vol. 119, Bonn, Germany, August 2005, pp. 265 – 272.

[21] L. Brickman, "On the field of values of a matrix," *Proceedings of the American Mathematical Society*, vol. 12, no. 1, pp. 61–66, February 1961.

[22] B. T. Polyak, "Convexity of quadratic transformations and its use in control and optimization," *Journal of Optimization Theory and Applications*, vol. 99, no. 3, pp. 553 – 583, December 1998.

[23] Y. Nesterov and A. Nemirovskii, *Interior-point polynomial algorithms in convex programming.* Society for Industrial Mathematics, 1995.

[24] B. Ricceri and S. Simons, *Minimax theory and applications.* Kluwer Academic Publishers, 1998.

[25] M. R. Garey and D. S. Johnson, *Computers and intractability: A guide to the theory of NP–completeness.* San Francisco: W. H. Freeman, 1979.

[26] C.-Y. Zhang and Y.-T. Li, "On schur complement of block diagonally dominant matrices," *Linear Algebra and its Applications*, vol. 414, pp. 533–546, April 2006.

Dragana Bajović (S'11) received her Engineer Diploma from the School of Electrical Engineering, Department of Automatic Control, at University of Belgrade in 2007. She is currently working toward the Ph.D. degree in the Dual PhD program between Carnegie-Mellon University, Pittsburgh, PA, and Instituto de Sistemas e Robótica (ISR), Instituto Superior Técnico (IST), Lisbon, Portugal. Her research interests are in the area of optimization, dimensionality reduction, and information processing in sensor networks.

João Xavier (S'97–M'03) received the Ph.D. degree in Electrical and Computer Engineering from Instituto Superior Técnico (IST), Technical University of Lisbon, Portugal, in 2002. Currently, he is an Assistant Professor in the Department of Electrical and Computer Engineering, IST. He is also a Researcher at the Institute of Systems and Robotics (ISR), Lisbon, Portugal. His current research interests are in the area of optimization, sensor networks and signal processing on manifolds.

Bruno Sinopoli (M'03) Bruno Sinopoli received the Dr. Eng. degree from the University of Padova in 1998 and his M.S. and Ph.D. in Electrical Engineering from the University of California at Berkeley, in 2003 and 2005 respectively. After a postdoctoral position at Stanford University, Dr. Sinopoli joined the faculty at Carnegie Mellon University where he is an assistant professor in the Department of Electrical and Computer Engineering with courtesy appointments in Mechanical Engineering and in the Robotics Institute. Dr. Sinopoli was awarded the 2006 Eli Jury Award for outstanding research achievement in the areas of systems, communications, control and signal processing at U.C. Berkeley and the NSF Career award in 2010. His research interests include networked embedded control systems, distributed estimation and control over wireless sensor-actuator networks and cyber-physical systems security.