Robust Linear Dimensionality Reduction for Hypothesis Testing with Application to Sensor Selection

Dragana Bajović, Bruno Sinopoli and João Xavier

***Abstract*— This paper addresses robust linear dimensionality reduction (RLDR) for binary Gaussian hypothesis testing. The goal is to find a linear map from the high dimensional space where the data vector lives to a low dimensional space where the hypothesis test is carried out. The linear map is designed to maximize the detector performance. This translates into maximizing the Kullback-Leibler (KL) distance between the two projected distributions. In practice, the distribution parameters are estimated from training data, thus subject to uncertainty. This is modeled by allowing the distribution parameters to drift within some confidence regions. We address the case where only the mean values of the Gaussian distributions, m_0 and m_1, are uncertain with confidence ellipsoids defined by the corresponding covariance matrices, S_0 and S_1. Under this setup, we find the linear map that maximizes the KL distance for the worst case drift of the mean values. We solve the problem globally for the case of linear mapping to one dimension, reducing it to a grid search over a finite interval. Our solution shows superior performance compared to robust linear discriminant analysis techniques recently proposed in the literature. In addition, we use our RLDR solution as a building block to derive a sensor selection algorithm for robust event detection, in the context of sensor networks. Our sensor selection algorithm shows quasi-optimal performance: worst-case KL distance for suboptimal sensor selection is at most 15% smaller than worst-case KL distance for the optimal sensor selection obtained by exhaustive search.**

I. Introduction

In this paper, we study linear dimensionality reduction (LDR) for classification purposes. In the classification setup, one wants to project two given distributions to a lower dimensional space, so that the projected distributions are separated as much as possible, in a certain sense. One of the most often used LDR classification techniques is linear discriminant analysis (LDA) [1]. Its main advantage is small computational cost. However, it results in the Bayes optimal classifier only for the case of equal covariance matrices of the two Gaussian distributions. Moreover, its performance degrades when the mean values of the two distributions are close.

To overcome this problem, the authors in [12], [3] try to find the linear map that maximizes the Kullback-Leibler (KL) distance between the two projected distributions. This paper also uses the KL distance as optimality criterion. The motivation for choosing KL distance comes from Stein's lemma which states that the probability of false alarm in asymptotic regime (i.e., when the number of i.i.d. samples goes to infinity) goes exponentially fast to zero, with rate proportional to the KL distance between the two distributions. Alternative related choices of optimality criteria for LDR that have been studied are the Chernoff distance and the J-divergence [14], [15], [13]. For more details on Chernoff distance and J-divergence see [4]. An advantage of using KL and Chernoff distance or J-divergence over LDA is that they are able to distinguish between distributions with equal means (and different covariances). However, finding the linear projection that maximizes any of the mentioned distances is in general a hard nonconvex problem. In our previous work [3], we solve the problem of maximizing KL distance globally for the case of LDR to one dimension and in full generality.

In practice, the distribution parameters are estimated from training data. Thus, in general, there is a discrepancy between the parameters used by the classifier (obtained from training data) and the parameters of the distributions underlying the data samples to be classified. The goal is to design a linear map to a low dimensional space so that the resulting classifier works well for all possible distribution parameter mismatches within an uncertainty set. This leads to the formulation of a robust linear dimensionality reduction (RLDR) problem. To address this problem we follow a worst case analysis: we find the linear map that maximizes the KL distance between the projected distributions for the worst case of parameter drifts. We address the case when only means of the two distributions may drift. Related work to this paper is [8]. This reference also addresses LDR problem in the presence of parameter uncertainties, but the Fisher criterion, i.e., LDA is used as a cost function. As in this paper, a worst case approach in [8] is followed, but it is assumed that both means and covariances are uncertain, belonging to arbitrary convex sets. We refer to method proposed in [8] as robust linear discriminant analysis (RLDA).

Contributions. Maximizing worst case KL distance results in a max-min optimization problem that is in general nonconvex and difficult to solve. In this paper, we address the case where only the means of the distributions may drift.

Partially supported by grant SFRH/BD/33517/2008 (through the Carnegie Mellon/Portugal Program managed by ICTI) from Fundação para a Ciência e Tecnologia and also by ISR/IST plurianual funding (POSC program, FEDER).

D. Bajović is with the Institute for Systems and Robotics (ISR), Instituto Superior Técnico (IST), 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), Lisbon, Portugal jxavier@isr.ist.utl.pt

The confidence set for the mean m_i is an ellipsoid, with the orientation defined by the covariance matrix S_i, $i = 0, 1$. We solve this problem globally for the case when data are projected to a one-dimensional space ($p = 1$), reducing it to a grid search over a finite interval. We compare performance of our method with RLDA, proposed in [8]. Simulation examples (section IV-A) show that the classifier based on our proposed methodology outperforms the classifier based on RLDA for the case when the mean vectors of the two distributions are close. For the case of more distinct mean vectors, the performance of the two methods is comparable. For the problem of maximizing worst case KL distance when the projected space dimension is $p > 1$, an incremental greedy approach that capitalizes on the $p = 1$ method is proposed.

In addition, we consider an application of RLDR to sensor selection for event detection. A subset of p out of n sensors is to be chosen to report their measurements to a fusion node. After receiving the selected measurements, the fusion node conducts the hypothesis test. The goal is to choose the subset of p sensors that yields the best detector performance in the presence of allowed distribution parameter drifts. Mathematically, the sensor selection problem is a LDR problem where the linear map has a special structure: the matrix that defines the linear map has $0/1$ entries with a unique 1 entry per row. Again, as a selection criterion, we choose the KL distance between the distributions of the selected measurement. The optimal subset can be found by a combinatorial search, but this is infeasible for large n and p. Thus, we propose a sensor selection algorithm that gives a suboptimal solution. The algorithm is based on 1) relaxation of the zero-one selection matrix to the Stiefel space([16]); 2) solving the LDR problem to find the suboptimal Stiefel matrix; 3) projecting back the resulting Stiefel matrix to the zero-one selection matrices set. In section VI-C we tested the performance of our sensor selection algorithm against the optimal solution obtained by search over all possible sensor selections for cases when exhaustive search is feasible (up to $n = 30$, $p = 4$). Our solution shows near optimal performance with significant reductions in computational cost. For example, for case $n = 30, p = 3$, an exhaustive search takes more than 3 hours, while our algorithm produces solution only 2.3% below optimum in less than 15 seconds (or 0.1% of the time needed by the optimal one) on a 1.7 GHz personal computer.

Paper organization. The rest of the paper is organized as follows. In section II we introduce the dimensionality reduction problem when the distribution parameters are known. In section III we consider the case where they are uncertain. We formulate the robust linear dimensionality reduction as an optimization problem. In sections IV and V, we present our solution algorithm. In section VI we discuss one application to sensor selection and detail our sensor selection algorithm. In section VI-C simulation results are presented and section VII summarizes the work presented in this paper. An appendix contains the proofs of some lemmas.

II. Linear dimensionality reduction: known distributions

We consider a binary Gaussian hypothesis test. Let

$$x = (x_1, x_2, \ldots, x_n)^\top \in \mathbb{R}^n$$

be the vector of all features, where x_i denotes the ith feature. It is assumed that the data vector $x \in \mathbb{R}^n$ was generated by one of two known Gaussian distributions. Accordingly, we face the hypothesis test

$$\begin{array}{rcl} H_0 & : & x \sim \mathcal{N}(m_0, S_0) \\ H_1 & : & x \sim \mathcal{N}(m_1, S_1) \end{array},$$

where $\mathcal{N}(m_i, S_i)$ denotes the Gaussian distribution with mean m_i and covariance matrix S_i, $i = 0, 1$. Hypothesis H_i corresponds to the class i ($i = 0, 1$).

Linear dimensionality reduction. Since the hypothesis test in high dimensional space might be computationally too expensive, we linearly map the data $x \in \mathbb{R}^n$ to a lower dimensional space $\mathbb{R}^p$:

$$y = E^\top x.$$

This induces the hypothesis test in the lower dimension space $\mathbb{R}^p$:

$$\begin{array}{rcl} H_0 & : & y \sim \mathcal{N}(E^\top m_0, E^\top S_0 E) \\ H_1 & : & y \sim \mathcal{N}(E^\top m_1, E^\top S_1 E) \end{array}.$$

The detector operates in the lower dimension space $\mathbb{R}^p$. We are interested in finding the linear map $E : \mathbb{R}^n \mapsto \mathbb{R}^p$ which gives the detector the best possible performance. We choose the linear map E that maximizes the Kullback-Leibler distance between the two projected Gaussian distributions.

Stein's lemma. Motivation for the choice of Kullback-Leibler distance comes from the Stein's lemma, a fundamental result from detection theory. Stein's lemma [11] states that the probability of false alarm exponentially goes to zero in asymptotic regime with the exponential rate proportional to the Kullback-Leibler distance between the two tested distributions, if the Neyman-Pearson detector is used. Indeed, if $P_{\mathrm{FA}}(k)$ denotes the probability of false alarm when k i.i.d. samples are processed, then

$$\lim_{k\to\infty} \frac{\log P_{\mathrm{FA}}(k)}{k} = -D_{\mathrm{KL}}(p_1 \| p_0)$$

where

$$D_{\mathrm{KL}}(p_1 \| p_0) = \int p_1(x) \log\left(\frac{p_1(x)}{p_0(x)}\right) dx$$

denotes the Kullback-Leibler distance between the distributions p_0 and p_1. Thus, more dissimilar distributions lead to lower probabilities of false alarm in the asymptotic regime.

Inspired by Stein's lemma, we propose to search for the linear map E which yields the largest KL distance between the induced p-dimensional distributions. That is, we formulate the following optimization problem

$$\begin{array}{ll} \text{maximize} & f(E; m_0, m_1, S_0, S_1) \\ \text{subject to} & E^\top E = I_p \end{array} \qquad (1)$$

where I_p denotes the $p \times p$ identity matrix and

$$f(E;m_0,m_1,S_0,S_1) := \\ D_{\mathrm{KL}}\left(\mathscr{N}(E^\top m_1, E^\top S_1 E), \mathscr{N}(E^\top m_0, E^\top S_0 E)\right). \tag{2}$$

It can be shown that

$$f(E;m_0,m_1,S_0,S_1) = \frac{1}{2}\left\{\mathrm{tr}\left((E^\top S_0 E)^{-1}(E^\top S_1 E)\right) + (m_1-m_0)^\top E(E^\top S_0 E)^{-1}E^\top(m_1-m_0) - \log\frac{\det\left(E^\top S_1 E\right)}{\det\left(E^\top S_0 E\right)} - p\right\}. \tag{3}$$

Note that the constraint in (1) forces the matrix $E \in \mathbb{R}^{n\times p}$ to have orthonormal columns. That is, E is a Stiefel matrix. This entails no loss of generality. Indeed, it is straightforward to check from (3) that the invariance equation

$$f(EA;m_0,m_1,S_0,S_1) = f(E;m_0,m_1,S_0,S_1)$$

holds for all nonsingular $A \in \mathbb{R}^{p\times p}$. This means that f depends on E only through its range space, and we can restrict attention to Stiefel matrices.

This problem has already been addressed in [3]. This reference solved the problem globally for case $p = 1$ reducing it to a grid search over an interval and proposed suboptimal greedy approach for case $p > 1$.

III. Linear dimensionality reduction: uncertain distributions

In the previous section II, the parameters (m_i, S_i) of the Gaussian distributions were assumed known. In practice, these parameters are often estimated from training data and are affected by some uncertainty. That is, the distribution parameters may drift from nominal values within some prescribed regions of confidence. To obtain the optimal projection map $E : \mathbb{R}^n \mapsto \mathbb{R}^p$, we maximize the projected KL distance for the worst outcome of the allowed distribution parameters, i.e.,

$$\begin{array}{ll}\text{maximize} & \min_{(\tilde{m}_0,\tilde{m}_1,\tilde{S}_0,\tilde{S}_1)\in\mathscr{U}} \; f(E;\tilde{m}_0,\tilde{m}_1,\tilde{S}_0,\tilde{S}_1) \\ \text{subject to} & E^\top E = I_p\end{array} \tag{4}$$

where $\mathscr{U}$ denotes the uncertainty regions of the parameters of the distributions.

In this paper, we restrict attention to the case where only the mean values are uncertain. More precisely, we consider

$$\mathscr{U} = \left\{(\tilde{m}_0,\tilde{m}_1,\tilde{S}_0,\tilde{S}_1) : \tilde{m}_i \in \mathscr{E}\left(m_i, k_i S_i^{-1}\right) \text{ and } \tilde{S}_i = S_i \text{ for } i = 0,1\right\} \tag{5}$$

where the notation $\mathscr{E}(a,A)$ (A is positive definite) denotes the ellipsoid

$$\mathscr{E}(a,A) = \{x \in \mathbb{R}^n : \; (x-a)^\top A(x-a) \le 1\}.$$

In other words set (5) means that, for the ith Gaussian distribution $\mathscr{N}\left(\tilde{m}_i, \tilde{S}_i\right)$, we have $\tilde{S}_i = S_i$ (with S_i known) but the mean value $\tilde{m}_i$ is uncertain and belongs to the confidence ellipsoid $\mathscr{E}(m_i, k_i S_i)$, where m_i, k_i are known. Note that m_i can be interpreted as the nominal value of $\tilde{m}_i$ and $k_i > 0$ controls how "large" is the uncertainty region (if $k_i = +\infty$, there would be no uncertainty: $\tilde{m}_i = m_i$). Note that the orientations of the uncertainty ellipsoids are not generic, i.e., they are induced by the covariance matrices S_0 and S_1. This is motivated by the following fact: if the means are estimated as sample means (the minimum variance unbiased estimate for Gaussian distributions), then the covariance of the estimate of the ith mean is equal to $\frac{1}{N}S_i$, where N is the number of iid samples used. Therefore, it is natural to assume that the confidence region for m_i is given by (5), where the scaling constants k_0 and k_1 are proportional to the number of samples used.

Thus, the optimization problem that we address is the following:

$$\begin{array}{ll}\text{maximize} & \min_{\tilde{m}_0\in\mathscr{E}(m_0,k_0S_0^{-1}),\,\tilde{m}_1\in\mathscr{E}(m_1,k_1S_1^{-1})} f(\cdot) \\ \text{subject to} & E^\top E = I_p\end{array} \tag{6}$$

where $f(\cdot) = f(E;\tilde{m}_0,\tilde{m}_1,S_0,S_1)$. In the sequel, we refer to (6) as the robust linear dimensionality reduction (RLDR) problem.

IV. RLDR: solution for the case $p = 1$

For the case of dimensionality reduction to one dimension, i.e., $p = 1$, we provide an algorithm that gives a global solution to problem (6). This is achieved through series of problem reformulations. Note that, for the case $p = 1$, problem (6) is

$$\begin{array}{ll}\text{maximize} & \min_{\delta_0\in\mathscr{E}(0,k_0S_0^{-1}),\;\delta_1\in\mathscr{E}(0,k_1S_1^{-1})} h(\cdot) \\ \text{subject to} & e^\top e = 1\end{array} \tag{7}$$

where

$$h(\cdot) = h(e;\delta_0,\delta_1,m_0,m_1,S_0,S_1) = \\ \frac{1}{2}\left\{\frac{e^\top S_1 e}{e^\top S_0 e} + \frac{((m_1+\delta_1-m_0-\delta_0)^\top e)^2}{e^\top S_0 e} - \log\frac{e^\top S_1 e}{e^\top S_0 e} - 1\right\}. \tag{8}$$

We start with Lemma 1 which shows that problem (7) is equivalent to (9).

Lemma 1: Suppose $v^\star$ solves

$$\begin{array}{ll}\text{maximize} & v^\top S v - \log(v^\top S v) + \xi(v) \\ \text{subject to} & v^\top v = 1\end{array} \tag{9}$$

where $S := S_0^{-1/2} S_1 S_0^{-1/2}$,

$$\xi(v) = \left\{\left(\sqrt{v^\top m m^\top v} - \frac{1}{\sqrt{k_1}}\sqrt{v^\top S v} - \frac{1}{\sqrt{k_0}}\right)^+\right\}^2, \tag{10}$$

$m := S_0^{-1/2}(m_1 - m_0)$ and, for $x \in \mathbb{R}$, $x^+ := \max\{0,x\}$. Then, $e^\star := S_0^{-1/2}v^\star / \|S_0^{-1/2}v^\star\|$ solves (7).

Proof: See the appendix. ∎

We now focus on the optimization problem (9). Note that the objective function in (9) depends on the vector v only through the quadratic forms $v^\top m m^\top v$ and $v^\top S v$. Thus, we can reformulate (9) as

$$\begin{array}{ll}\text{maximize} & \psi(x,y) \\ \text{subject to} & (x,y) \in \mathscr{R}\end{array} \tag{11}$$

where

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

and

$$\mathscr{R} = \Big\{ (x,y) \in \mathbb{R}^2 : \; x = v^\top S v, \; y = v^\top m m^\top v, \\ \text{for some } v \in \mathbb{R}^n, \; v^\top v = 1 \Big\}. \tag{12}$$

For $n \geq 3$, the set $\mathscr{R} \subset \mathbb{R}^2$ is compact and convex, see [7]. Note that, since ψ is continuous and $\mathscr{R}$ is compact, there is a global maximizer. The next lemma asserts that the boundary of $\mathscr{R}$ contains a global maximizer.

Lemma 2: The boundary $\partial\mathscr{R}$ of the set $\mathscr{R}$ contains a global maximizer of (11).

Proof: See the appendix. ∎

The boundary of $\mathscr{R}$ is a closed curve in $\mathbb{R}^2$. Our strategy consists in circulating along $\partial\mathscr{R}$ to spot a global maximizer. More precisely, we will sample $\partial\mathscr{R}$ with a finite set of points and pick the best point. To implement this strategy, we borrow the following theorem from [6].

Theorem 3 ([6]): Let $n \geq 3$ and let A, B be $n \times n$ symmetric matrices. Let

$$\mathscr{R}(A,B) = \Big\{ (x,y) \in \mathbb{R}^2 : \; x = v^\top A v, \; y = v^\top B v, \\ \text{for some } v \in \mathbb{R}^n, \; v^\top v = 1 \Big\}.$$

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 $\mathscr{R}(A,B)$ is strictly convex and its boundary is given by

$$\partial\mathscr{R}(A,B) = \{ (x(t), y(t)) : \; t \in [0, 2\pi], \\ x(t) = u_{\min}(t)^\top A u_{\min}(t), \; y(t) = u_{\min}(t)^\top B u_{\min}(t) \Big\}. \tag{13}$$

Parametrization of $\partial\mathscr{R}$. Theorem 3 assumes that $\lambda_{\min}(t)$ is simple for all t. However, a parametrization of the boundary when this condition is not satisfied is readily available. Applied to our set $\mathscr{R}$ in (12), this leads to the following procedure:

1) generate the points

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

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

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

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

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

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

Finally, our sample of $\partial\mathscr{R}$ is $\widehat{\partial\mathscr{R}} = \{(x_k, y_k)\} \cup \left\{ x_k^{(j)}, y_k^{(j)} \right\}$.

Solution to problem (9)**.** A solution $v^\star$ of (9) can be found as follows. Let

$$(x^\star, y^\star) \in \arg \max_{(x,y) \in \widehat{\partial\mathscr{R}}} \psi(x,y).$$

That is, $(x^\star, y^\star)$ denotes the best point in $\widehat{\partial\mathscr{R}}$. 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 quadratics

$$\begin{cases} v^\top S v &= x^\star \\ v^\top m m^\top v &= y^\star \\ v^\top v &= 1 \end{cases} \tag{14}$$

with respect to v. Any solution can be taken as $v^\star$. It can be shown that (14) can be mapped into a convex problem.

A. Simulation example

We will present now simulation results on the performance of our RLDR method. The performance of our method is compared with RLDA in [8]. The classifier that is used is the Maximum Likelihood detector (MLD). The data samples to be classified are generated from two Gaussian distributions $\mathscr{N}(m_0, S_0)$ and $\mathscr{N}(m_1, S_1)$. However, the MLD uses $\tilde{m}_0$ and $\tilde{m}_1$. We want to assess the classifier performance when $\tilde{m}_0, \tilde{m}_1$ drift inside the confidence regions defined by (5). To this end, we generate 3000 pairs of points $(\tilde{m}_0, \tilde{m}_1)$ from the confidence region (5); for each of them we estimate probability of false alarm (P_{FA}). To estimate P_{FA} we generate 10000 data samples from $\mathscr{N}(m_0, S_0)$ and conduct 10000 single-sample hypothesis tests. We calculate P_{FA} as a ratio between the number of test instances when hypothesis 1 is chosen and total number of tests performed (10000). The data sample dimension is $n = 10$ and the parameters (m_0, S_0), (m_1, S_1) are generated randomly. Constants k_0 and k_1 that define the size of confidence regions (5) are chosen such that the highest uncertainty in $m_0(m_1)$ is equal to 20% of the norm of the vector $m_0 - m_1$. Figure 1 presents histograms for P_{FA} for RLDA and RLDR.

The pair of histograms on the right is obtained for the case when the norm $\|m_1 - m_0\|$ is large in comparison to the one of matrices S_0 and S_1. We see that for this case both RLDA and RLDR perform well (mean value is around 0.012 and variance is very small). This result was expected, since the more the two distributions are apart from each other in terms of mean vectors, the easier it is to discriminate between them. The pair of histograms at the top is obtained for the case when the norm $\|m_1 - m_0\|$ is small. In this case the projector obtained by RLDA results in very poor detector performance. This result is predicted by the theory: in the limit case when $m_0 = m_1$ LDA is not able to discern between the two distributions. On the other hand, Kullback -Leibler distance takes into account the orientation of the distributions. This explains why our KL-based RLDR projector results in a significantly better detector performance for the case when the norm $\|m_1 - m_0\|$ is small.

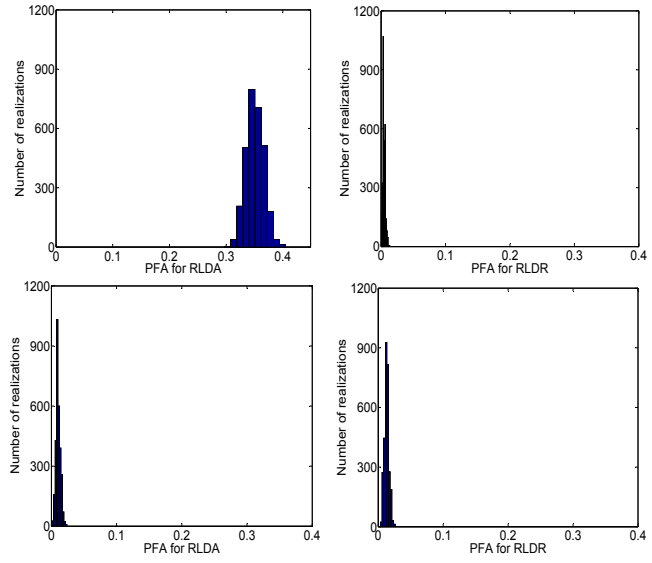

Fig. 1. Probability of false alarm

V. RLDR: SUBOPTIMAL SOLUTION FOR THE CASE $p > 1$

Optimization problem (6) for the case $p > 1$ seems to be very difficult to solve globally. Here, we propose a greedy, yet suboptimal, approach. Our method consists of constructing the columns of the matrix $E = \begin{bmatrix} e_1 & \cdots & e_p \end{bmatrix}$ one by one (in the order e_1, e_2, ...). We construct the jth column by solving (7) with the constraint that the column e_j must be orthogonal to the previously determined columns $e_1, ..., e_{j-1}$, i.e., we solve

$$\begin{array}{ll} \text{maximize} & \min_{\delta_0 \in \mathscr{E}(0,k_0 S_0^{-1}),\, \delta_1 \in \mathscr{E}(0,k_1 S_1^{-1})} h(\cdot) \\ \text{subject to} & e^\top e = 1 \\ & e^\top e_i = 0, \quad i = 1, \ldots, j-1. \end{array} \tag{15}$$

where $h(\cdot) = h(e; \delta_0, \delta_1, m_0, m_1, S_0, S_1)$. Let $U^{(j)} \in \mathbb{R}^{n \times (n-j+1)}$ be a matrix with orthonormal columns which spans the orthogonal complement of $\mathsf{span}\left\{e_1, \ldots, e_{j-1}\right\}$. The restrictions in (15) mean that $e = U^{(j)} e^{(j)}$ for some unit-norm $e^{(j)} \in \mathbb{R}^{n-j+1}$. Problem (15) therefore corresponds to

$$\begin{array}{ll} \text{maximize} & \min_{\delta_0 \in \mathscr{E}(0,k_0 S_0^{-1}), \delta_1 \in \mathscr{E}(0,k_1 S_1^{-1})} h(\cdot) \\ \text{subject to} & {e^{(j)}}^\top e^{(j)} = 1. \end{array} \tag{16}$$

where $h(\cdot) = h(U^{(j)} e^{(j)}; \delta_0, \delta_1, m_0, m_1, S_0, S_1)$. In the appendix, we show that (16) is equivalent to

$$\begin{array}{ll} \text{maximize} & \min_{\gamma_0 \in \mathscr{E}\left(0,k_0 \left(S_0^{(j)}\right)^{-1}\right),\, \gamma_1 \in \mathscr{E}\left(0,k_1 \left(S_1^{(j)}\right)^{-1}\right)} h(\cdot) \\ \text{subject to} & {e^{(j)}}^\top e^{(j)} = 1 \end{array} \tag{17}$$

where $h(\cdot) = h(e^{(j)}; \gamma_0, \gamma_1, m_0^{(j)}, m_1^{(j)}, S_0^{(j)}, S_1^{(j)})$, $m_i^{(j)} = {U^{(j)}}^\top m_i$ and $S_i^{(j)} = {U^{(j)}}^\top S_i U^{(j)}$, $i = 0, 1$. That is, (17) is simply an instance of (7) 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 $\mathsf{span}\left\{e_1, e_2, \ldots, e_{j-1}\right\}$
3: Compute the projected means and covariances $m_i^{(j)} = {U^{(j)}}^\top m_i$, $S_i^{(j)} = {U^{(j)}}^\top S_i U^{(j)}$ for $i = 0, 1$
4: Let $e^{(j)} \in \mathbb{R}^{n-j+1}$ denote a solution of (17)
5: Compute the jth column of E as $e_j = U^{(j)} e^{(j)}$
6: **end for**

VI. APPLICATION: ROBUST SENSOR SELECTION

A. Problem formulation

We now consider the sensor selection problem for event detection in wireless sensor networks (WSN). In WSNs, sensor nodes have a limited energy budget. Thus, a minimal number of sensors should be kept in active mode at any given time, in order to prolong the network lifetime. For a certain query, i.e., the task of determining whether an event occurred or not, we wish to minimize the number of reporting sensors without compromising performance. Based on the sensor readings the fusion node will decide whether the event has occurred or not. For a particular query, once we fix the number of reporting sensors, we want to choose the subset of them that will result in the best possible detection performance.

We wish to remark that WSNs are usually deployed to perform several detection and estimation tasks, i.e., a WSN will be required to answer several different queries. Since optimal sensor selection is query-dependent, during WSN operation time, occurrence of different events will be tested and thus different subsets of sensors will be activated over time. This provides motivation for selecting sensors rather than deploying only a subset of them in the first place. When one of the selected sensors for a given detection task cannot operate anymore (due to some failure, or the battery has ran out), one could re-select a subset of p out of the $n-1$ remaining ones by using the algorithm we propose below. Of course, other sensor replacement strategies can be envisaged. As this falls outside the focus of this paper, a more in-depth analysis is left to future work.

As in section III, we assume that the means of the two Gaussian distributions may drift within their regions of confidence, given by (5). Mathematically, the sensor selection problem can be formulated as follows:

$$\begin{array}{ll} \text{maximize} & \min_{\tilde{m}_0 \in \mathscr{E}(m_0,k_0 S_0^{-1}),\, \tilde{m}_1 \in \mathscr{E}(m_1,k_1 S_1^{-1})} f(\cdot) \\ \text{subject to} & E^\top E = 1 \\ & E_{ij} \in \{0,1\} \end{array} \tag{18}$$

where $f(\cdot) = f(E; \tilde{m}_0, \tilde{m}_1, S_0, S_1)$ is given by (3).

B. Sensor selection algorithm

The optimization problem (18) is combinatorial and solving it by searching over all $\binom{n}{p}$ combinations of sensors is

intractable for sufficiently large n and p. We propose a suboptimal approach. It consists of three phases, which we will detail below.

Phase 1: relaxation. We first find a Stiefel matrix using algorithm 1. Denote the result by $\widehat{E}$.

Phase 2: projection. We find a selection matrix $\widetilde{E}$ with range space closest to the range space of matrix $\widehat{E}$. This corresponds to solving:

$$\begin{array}{rll} \widetilde{E} = & \arg\min & \left\| EE^\top - \widehat{E}\widehat{E}^\top \right\| \\ & \text{subject to} & E_{ij} \in \{0,1\} \\ & & E^\top E = I_p \end{array} . \tag{19}$$

It can be shown that the solution of (19) is obtained as follows: if $(j_1, j_2, \ldots, j_p)$ denote the indices of the largest entries of the diagonal of the projector $\widehat{E}\widehat{E}^\top$, then $\widetilde{E} = \left[e_{j_1} e_{j_2} \cdots e_{j_p} \right]$ where e_j stands for the jth column of the identity matrix I_n.

Phase 3: local refinement. Finally, we polish the result by performing a local optimization in the neighborhood of the matrix $\widetilde{E}$. We start by computing mean vectors $m_0^\star$ and $m_1^\star$ that correspond to the closest distributions in the full space $\mathbb{R}^n$ by solving the convex optimization problem:

$$\begin{array}{ll} \text{minimize} & \frac{1}{2} \left\{ \operatorname{tr}\left(S_0^{-1} S_1\right) + (\tilde{m}_1 - \tilde{m}_0)^\top S_0^{-1} (\tilde{m}_1 - \tilde{m}_0) - \right. \\ & \qquad \left. \log \frac{\det(S_1)}{\det(S_0)} - p \right\} \\ \text{subject to} & (\tilde{m}_0, \tilde{m}_1) \in \mathscr{U}_m \end{array} . \tag{20}$$

Then, for the choice of parameters $m_0^\star$, $m_1^\star$, S_0, S_1, we evaluate the KL distance for the selection matrices in the neighborhood of $\widetilde{E}$. This local search consists of p steps. In the i-th step ($i = 1, 2, \ldots, p$) all columns of the current selection matrix are fixed except the i-th one, which is viewed as an optimization variable. The i-th column is swept through all possible choices, i.e., through the canonical vectors that are different from the remaining $p-1$ columns. After the i-th column is circulated through all possible choices, it is frozen to the choice that gives maximal KL distance between the projected distributions. The procedure is repeated for all p columns.

C. Simulation results

We provide simulation results on the performance of the proposed sensor selection algorithm. To test our algorithm, we compare the values of the worst-case KL distance for the subset of sensors obtained by our algorithm with the worst-case KL distance of the optimal subset of sensors (i.e., the one that solves (18)), obtained by exhaustive search. Due to the complexity of the optimal algorithm, we perform the comparison over a small number of sensors. It is important to clarify that our algorithm is capable of handling larger problems. We randomly generated parameters for two Gaussian distributions for cases n = 10, 11, 12, 13, 14, 20, 30 and for each of them we ran our algorithm for p = 1, 2, 3, 4 (with the exception of $n = 30$, $p = 4$). The results are shown in Table 1. It can be seen that the values of KL distance match in most cases. For p = 1,2,3 the suboptimal value differs from the optimal value in only one entry ($n = 30$, $p = 3$) for less than 3%; for $p = 4$ the suboptimal value differs from the optimal value less than 15%. Computational time of our algorithm is significantly shorter than the one of the exhaustive search, which requires a long time even for small n and p (as in Table 1). For example, in case $n = 30\,/\,p = 3$, exhaustive search ($\binom{30}{3} = 4060$ combinations) takes more than 3 hours, while our algorithm produces solution in less than 15 seconds on a 1.7 GHz personal computer, with savings of about 99.9% in computation. For $n = 50\,/\,p = 10$ the optimal solution is infeasible ($\binom{50}{10} \approx 10^{10}$ combinations) while our algorithm takes about 1.5 minutes to find a solution.

TABLE I

KL DISTANCES FOR OPTIMAL AND SUBOPTIMAL SELECTION

KL distance	$p=1$	$p=2$	$p=3$	$p=4$
SUBOPT				
n				
10	69.2481	86.7304	171.9278	243.5756
11	37.4097	69.2920	92.2846	110.8385
12	38.3409	89.8811	110.4657	129.9147
13	15.5787	31.6767	46.2754	60.1648
14	36.1556	76.4347	107.9220	124.5990
20	50.3118	91.6304	135.4058	168.5743
30	73.2230	122.1518	153.6667	
OPTIMAL				
n				
10	69.2481	86.7304	171.9278	284.6790
11	37.4097	69.2920	92.9244	115.4617
12	38.3409	89.8811	110.4657	129.9147
13	15.5787	31.6767	46.2754	61.5073
14	36.1556	76.4347	107.9220	127.3992
20	50.3118	91.6304	135.4058	168.5743
30	73.2230	122.1518	155.7737	

VII. Conclusion

In this paper, we considered robust linear dimensionality reduction for hypothesis testing. This translates in finding the linear map that maximizes the Kullback-Leibler distance between the projected Gaussian distributions for the worst case drift of the distribution parameters within the uncertainty set. The resulting optimization problem is difficult to solve with full generality. We solve the problem globally for the case $p = 1$ when only the distribution means are uncertain. We reduced the problem to a grid search over an finite interval. Simulations show that our $p = 1$ solution is superior to the robust LDA method for the case of small differences between the two distribution means. When applied to a sensor selection problem, our algorithm shows good performance: the suboptimal values of KL distance were always above 85% of the optimal (maximal) values.

Appendix

Proof of Lemma 1. To simplify notation within this proof we write $h(e; \delta_0, \delta_1)$ instead of $h(e; \delta_0, \delta_1, m_0, m_1, S_0, S_1)$.

Note that $h(e;\delta_0,\delta_1)$ depends on the vector e only through its direction. Therefore, we can change the sphere constraint $e^\top e = 1$ in (7) to the ellipsoid constraint $e^\top S_0 e = 1$. With this ellipsoid constraint in force, the function h simplifies significantly. We further change variables as $v := S_0^{1/2} e$, thus arriving at the equivalent problem

$$\begin{array}{ll} \text{maximize} & \min_{\delta_0 \in \mathscr{E}(0,k_0 S_0^{-1}),\, \delta_1 \in \mathscr{E}(0,k_1 S_1^{-1})} g(v;\delta_0,\delta_1) \\ \text{subject to} & v^\top v = 1 \end{array} \tag{21}$$

where

$$g(v;\delta_0,\delta_1) = \frac{1}{2}\left\{ v^\top S v + ((m_1+\delta_1-m_0-\delta_0)^\top S_0^{-1/2} v)^2 - \log v^\top S v - 1 \right\}$$

and $S := S_0^{-1/2} S_1 S_0^{-1/2}$. If $v^\star$ solves (21), then $e^\star = S_0^{-1/2} v^\star / \|S_0^{-1/2} v^\star\|$ solves (7).

In (21), the optimization variable is v and the objective function is

$$\phi(v) = \min_{\delta_0 \in \mathscr{E}(0,k_0 S_0^{-1}),\, \delta_1 \in \mathscr{E}(0,k_1 S_1^{-1})} g(v;\delta_0,\delta_1),$$

that is,

$$\phi(v) = \frac{1}{2}\left\{ v^\top S v - \log v^\top S v - 1 \right\} + \frac{1}{2}\xi(v) \tag{22}$$

where

$$\xi(v) := \min_{\delta_0 \in \mathscr{E}(0,k_0 S_0^{-1}),\, \delta_1 \in \mathscr{E}(0,k_1 S_1^{-1})} ((m_1+\delta_1-m_0-\delta_0)^\top S_0^{-1/2} v)^2. \tag{23}$$

It can be shown (details omitted) that previous problem has the closed-form solution

$$\xi(v) = \left\{ \left(\left| v^\top S_0^{-1/2}(m_1-m_0) \right| - \left\| \frac{1}{\sqrt{k_1}} S_1^{1/2} S_0^{-1/2} v \right\| - \left\| \frac{1}{\sqrt{k_0}} v \right\| \right)^+ \right\}^2 \tag{24}$$

where, for $x \in \mathbb{R}$, $x^+ = \max\{0,x\}$. We rewrite (24) as

$$\xi(v) = \left\{ \left(\sqrt{v^\top m m^\top v} - \frac{1}{\sqrt{k_1}} \sqrt{v^\top S v} - \frac{1}{\sqrt{k_0}} \|v\| \right)^+ \right\}^2. \tag{25}$$

where $m := S_0^{-1/2}(m_1 - m_0)$. Plugging (25) into (22), turns (21) into the problem (9). This completes the proof.

Proof of Lemma 2. Let $(x_0,y_0) \in \mathscr{R}$ be given and let $\mathscr{Y}_{x_0} := \{y : (x_0,y) \in \mathscr{R}\}$. Since $\mathscr{R}$ is compact and convex, $\mathscr{Y}_{x_0}$ is a finite interval, say $\mathscr{Y}_{x_0} = [y_1,y_2]$. Note that (x_0,y_2) belongs to $\partial\mathscr{R}$. Also, it is straightforward to check that the function $y \in \mathscr{Y}_{x_0} \mapsto \psi(x_0,y)$ is non-decreasing with respect to y (because $t \geq 0 \mapsto (t^+)^2$ is non-decreasing with respect to t). Thus, $\psi(x_0,y_0) \leq \psi(x_0,y_2)$. In sum, given any point in $\mathscr{R}$, we can produce another in $\partial\mathscr{R}$ with equal or greater objective ψ. This proves that $\partial\mathscr{R}$ contains a global maximizer.

Proof that (16) **is equivalent to** (17)**.** It is straightforward to check that (16) is equivalent to

$$\begin{array}{ll} \text{maximize} & \min_{\delta_0 \in \mathscr{E}(0,k_0 S_0^{-1}),\, \delta_1 \in \mathscr{E}(0,k_1 S_1^{-1})} h(\cdot) \\ \text{subject to} & {e^{(j)}}^\top e^{(j)} = 1. \end{array} \tag{26}$$

where $h(\cdot) = h\left(e^{(j)}; {U^{(j)}}^\top \delta_0, {U^{(j)}}^\top \delta_1, m_0^{(j)}, m_1^{(j)}, S_0^{(j)}, S_1^{(j)}\right)$, $m_i^{(j)} = {U^{(j)}}^\top m_i$ and $S_i^{(j)} = {U^{(j)}}^\top S_i U^{(j)}$. The ellipsoid constraint on δ_i in (26) means that

$$\delta_i \in \mathscr{U}_i = \left\{ \frac{1}{\sqrt{k_i}} S_i^{1/2} u_i : u_i^\top u_i \leq 1 \right\}.$$

Thus, if we introduce the variables $\gamma_i = {U^{(j)}}^\top \delta_i$, the optimization problem (26) is converted to

$$\begin{array}{ll} \text{maximize} & \min_{\gamma_0 \in \mathscr{V}_0,\, \gamma_1 \in \mathscr{V}_1} h\left(e^{(j)}; \gamma_0, \gamma_1, m_0^{(j)}, m_1^{(j)}, S_0^{(j)}, S_1^{(j)}\right) \\ \text{subject to} & {e^{(j)}}^\top e^{(j)} = 1 \end{array} \tag{27}$$

where

$$\mathscr{V}_i = \left\{ \frac{1}{\sqrt{k_i}} {U^{(j)}}^\top S_i^{1/2} u_i : u_i^\top u_i \leq 1 \right\}.$$

That is,

$$\mathscr{V}_i = \frac{1}{\sqrt{k_i}} {U^{(j)}}^\top S_i^{1/2} B_n(0,1) \tag{28}$$

where $B_n(0,1) = \{u \in \mathbb{R}^n : u^\top u \leq 1\}$ designates the unit-norm ball centered at the origin of $\mathbb{R}^n$. Also, for a matrix $A \in \mathbb{R}^{k \times l}$, the symbol $AB_l(0,1)$ denotes the set $\{Au : u \in B_l(0,1)\} \subset \mathbb{R}^k$.

To conclude the proof, we need to show that $\mathscr{V}_i = E\left(0, k_i \left(S_i^{(j)}\right)^{-1}\right)$. Let $S_i = Q_i \Lambda_i Q_i^\top$ be an eigenvalue decomposition of S_i where Q_i is an $n \times n$ orthogonal matrix ($Q_i^\top Q_i = I_n$) and Λ_i is an $n \times n$ diagonal matrix with positive diagonal entries. From (28), we have

$$\begin{aligned} \mathscr{V}_i &= \frac{1}{\sqrt{k_i}} {U^{(j)}}^\top Q_i \Lambda_i^{1/2} Q_i^\top B_n(0,1) \\ &= \frac{1}{\sqrt{k_i}} {U^{(j)}}^\top Q_i \Lambda_i^{1/2} B_n(0,1). \end{aligned} \tag{29}$$

Now, let

$$ {U^{(j)}}^\top Q_i \Lambda_i^{1/2} = W \Sigma V^\top \tag{30}$$

be a singular-value decomposition where W is an $(n-j+1) \times (n-j+1)$ orthogonal matrix,

$$\Sigma = \begin{bmatrix} \Sigma_1 & 0 \end{bmatrix}$$

is an $(n-j+1) \times n$ matrix with Σ_1 containing the singular values (Σ_1 is an $(n-j+1) \times (n-j+1)$ matrix) and V is an $n \times n$ orthogonal matrix. Plugging (30) into (29) gives

$$\begin{aligned} \mathscr{V}_i &= \frac{1}{\sqrt{k_i}} W \Sigma V^\top B_n(0,1) \\ &= \frac{1}{\sqrt{k_i}} W \Sigma B_n(0,1) \qquad (31) \\ &= \frac{1}{\sqrt{k_i}} W \Sigma_1 B_{n-j+1}(0,1). \qquad (32) \end{aligned}$$

On the other hand, we have

$$
\begin{aligned}
E\left(0, k_i\left(S_i^{(j)}\right)^{-1}\right) &= \frac{1}{\sqrt{k_i}}\left(S_i^{(j)}\right)^{1/2} B_{n-j+1}(0,1) \\
&= \frac{1}{\sqrt{k_i}}\left(U^{(j)\top} S_i U^{(j)}\right)^{1/2} B_{n-j+1}(0,1) \\
&= \frac{1}{\sqrt{k_i}}\left(W\Sigma_1^2 W^\top\right)^{1/2} B_{n-j+1}(0,1) \\
&= \frac{1}{\sqrt{k_i}} W\Sigma_1 W^\top B_{n-j+1}(0,1) \\
&= \frac{1}{\sqrt{k_i}} W\Sigma_1 B_{n-j+1}(0,1). \qquad (33)
\end{aligned}
$$

REFERENCES

[1] K. Fukunaga *Introduction to Statistical Pattern Recognition*, Academic Press, Boston, 1990

[2] S. Joshi, S.Boyd Sensor Selection via Convex Optimization, *IEEE Transactions on Signal Processing*, 57(2) pp 451-462, February 2009

[3] D. Bajovic, B. Sinopoli, J. Xavier Sensor Selection for Hypothesis Testing in Wireless Sensor Networks: a Kullback-Leibler based approach, *submitted to 48th IEEE Conference on Decision and Control*, 2009

[4] D. H. Johnson, S. Sinanovic Symmetrizing the Kullback-Leibler distance *IEEE Transactions on Information Theory*, 2000

[5] J. E. Weimer, B. Sinopoli, B. H. Krogh A Relaxation Approach to Dynamic Sensor Selection in Large-Scale Wireless Networks, *in Proceedings of the 2008 The 28th International Conference on Distributed Computing Systems Workshops* pp 501-506, 2008

[6] B. T. Polyak Convexity of Quadratic Transformations and Its Use in Control and Optimization, *Journal of Optimization theory and applications*, Vol. 99, N0. 3, pp. 553 - 583, Dec. 1998

[7] L. Brickman On the Field of Values of a Matrix, *Proceedings of the American Mathematical Society*, Vol. 12, pp. 61-66, 1961.

[8] S.-J. Kim, A. Magnani, and S. P. Boyd, Robust Fisher Linear Discriminant analysis *Advances in Neural Information Processing Systems* vol. 18, pp. 659–666, 2006

[9] F. Zhao, J. Shin, J. Reich Information-Driven Dynamic Sensor Collaboration for Tracking Applications, *IEEE Signal Processing Magazine* vol. 19(2), pp 61-72, March

[10] H. Rowaihy, S. Eswaran, M. Johnson, D. Verma, A. Bar-Noy, T. Brown, and T. L. Porta A survey of sensor selection schemes in wireless sensor networks, Technical report, Pennsylvania State University, March 2007

[11] T. M. Cover, J. A. Thomas *Elements of information theory*, Wiley, 1991

[12] F. De la Torre, T. Kanade Multimodal Oriented Discriminant Analysis *Proceedings on the 22nd International Conference on Machine Learning*, 2005

[13] M. Loog, R. P. W. Duin Linear dimensionality reduction via a heteroscedastic extension of LDA: the Chernoff criterion, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 26(6), June 2004

[14] L. Rueda, M. Herrera. Linear dimensionality reduction by maximizing the Chernoff distance in the transformed space, *Pattern Recognition*, vol. 41(10), pp 3138-3152, October 2008

[15] L. L. Scharf. *Statistical Signal Processing: Detection, Estimation and Time series Analysis*, Addison and Wesley, 1991

[16] A. Hatcher, *Algebraic Topology*, Cambridge University Press, 2002

[17] R. Ambrosino, B. Sinopoli, K. Poolla, S. S. Sastry, Optimal Sensor Density for Remote Estimation Over Wireless Sensor Networks, *Forty-Sixth Annual Allerton Conference*, September 2008