entropy

MDPI

Article

Discriminant Analysis under f-Divergence Measures

Anmol Dwivedi, Sihui Wang and Ali Tajer *

Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180, USA; dwivea2@rpi.edu (A.D.); scottwon@bupt.edu.cn (S.W.)
* Correspondence: tajer@ecse.rpi.edu; Tel.: +1-518-276-8237

Abstract: In statistical inference, the information-theoretic performance limits can often be expressed in terms of a statistical divergence between the underlying statistical models (e.g., in binary hypothesis testing, the error probability is related to the total variation distance between the statistical models). As the data dimension grows, computing the statistics involved in decision-making and the attendant performance limits (divergence measures) face complexity and stability challenges. Dimensionality reduction addresses these challenges at the expense of compromising the performance (the divergence reduces by the data-processing inequality). This paper considers linear dimensionality reduction such that the divergence between the models is maximally preserved. Specifically, this paper focuses on Gaussian models where we investigate discriminant analysis under five f-divergence measures (Kullback–Leibler, symmetrized Kullback–Leibler, Hellinger, total variation, and χ^2). We characterize the optimal design of the linear transformation of the data onto a lower-dimensional subspace for zero-mean Gaussian models and employ numerical algorithms to find the design for general Gaussian models with non-zero means. There are two key observations for zero-mean Gaussian models. First, projections are not necessarily along the largest modes of the covariance matrix of the data, and, in some situations, they can even be along the smallest modes. Secondly, under specific regimes, the optimal design of subspace projection is identical under all the f-divergence measures considered, rendering a degree of universality to the design, independent of the inference problem of interest.

Keywords: dimensionality reduction; discriminant analysis; f-divergence; statistical inference

Citation: Dwivedi, A.; Wang, S.; Tajer, A. Discriminant Analysis under f-Divergence Measures. *Entropy* **2022**, *24*, 188. https://doi.org/10.3390/e24020188

Academic Editor: Igal Sason

Received: 18 November 2021
Accepted: 25 January 2022
Published: 27 January 2022

1. Introduction

1.1. Motivation

Consider a simple binary hypothesis testing problem in which we observe an n-dimensional sample X and aim to discern the underlying model according to:

$$\mathsf{H}_0: \ X \sim \mathbb{P} \qquad \text{vs.} \qquad \mathsf{H}_1: \ X \sim \mathbb{Q}\,. \tag{1}$$

The optimal decision rule (in the Neyman-Pearson sense) involves computing the likelihood ratio $\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(X)$ and the performance limit (sum of type I and type II errors) is related to the total variation distance between $\mathbb{P}$ and $\mathbb{Q}$. We emphasize that our focus is on the settings in which the n elements of X are not statistically independent, in which case the likelihood ratio $\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(X)$ cannot be decomposed into the product of the coordinate-level likelihood ratios. One of the key practical obstacles to solve such problems pertains to the computational cost of finding and performing the statistical tests. This renders a gap between the performance that is information-theoretically viable (unbounded complexity) versus a performance possible under bounded computational complexity [1,2].

Dimensionality reduction techniques have become an integral part of statistical analysis in high dimensions [3–6]. In particular, linear dimensionality reduction methods have been developed and used for over a century for various reasons, such as their low computational complexity and simple geometric interpretation, as well as for a multitude of applications, such as data compression, storage, and visualization, to name only a few.

These methods linearly map the high-dimensional data to lower dimensions while ensuring that the desired features of the data are preserved. There exist two broad sets of approaches to linear dimensionality reduction in one dataset X, which we review next.

1.2. Related Literature

(1) ***Feature extraction****:* In one set of approaches, the objective is to select and extract informative and non-redundant features in the dataset X. These approaches are generally unsupervised. These widely-used approaches are principal component analysis (PCA), and its variations [7–9], multidimensional scaling (MDS) [10–13], and sufficient dimensionality reduction (SDR) [14]. The objective of PCA is to retain as much variation in the data in a lower dimension by minimizing the reconstruction error. In contrast, MDS aims to maximize the scatter of the projection and maximizes an aggregate scatter metric. Finally, the objective of SDR is to design an orthogonal mapping of the data that makes the data X and the responses conditionally independent (given the projected data). There exist extensive variations to the three approaches, and we refer the reader to Reference [6] for more discussions.

(2) ***Class separation****:* In another set of approaches, the objective is to perform classification in the lower dimensional space. These approaches are supervised. Depending on the problem formulation and the underlying assumptions, the resulting decision boundaries between the models can be linear or non-linear. One approach pertinent to this paper's scope is discriminant analysis (DA), that leverages the distinction between given models and designs a mapping such that its lower-dimensional output exhibits maximum separation across different models [15–20]. In general, this approach generates two matrices: within-class and between-class scatter matrices. The within-class scatter matrix shows the scatter of the samples around their respective class means, whereas, in contrast, the between-class scatter matrix captures the scatter of the samples around the mixture mean of all the models. Subsequently, a univariate function of these matrices is formed such that it increases when the between-class scatter becomes larger, or when the within-class scatter becomes smaller. Examples of such a function of between-class and within-class matrices is a classification index that includes the ratio of their determinants, difference of their determinants, and ratio of their traces [17]. These approaches focus on reducing the dimension to one and maximize separability between the two classes. There exist, however, studies that consider reducing to dimensions higher than one and separation across more than two classes. Finally, depending on the structure of the class-conditional densities, the resulting shape of the decision boundaries give rise to linear and quadratic DA.

The f-divergences between a pair of probability measures quantifies the similarity between them. Shannon [21] introduced the mutual information as a divergence measure, which was later studied comprehensively by Kullback and Leibler [22] and Kolmogorov [23], establishing the importance of such measures in information theory, probability theory, and related disciplines. The family of f-divergences, independently introduced by Csiszár [24], Ali and Silvey [25], and Morimoto [26], generalize the Kullback–Leibler divergence which enable characterizing the information-theoretic performance limits of a wide range of inference, learning, source coding, and channel coding problems. For instance, References [27–30] consider their application to various statistical decision-making problems [31–34]. More recent developments on the properties of f-divergence measures can be found in Reference [31,35–37].

1.3. Contributions

The contribution of this paper has two main distinctions from the existing literature on DA. First, DA generally focuses on the classification problem for determining the underlying model of the data. Secondly, motivated by the complexities of finding the optimal decision rules for classification (e.g., density estimation), the existing criteria used for separation are selected heuristically. In this paper, we study this problem by referring to the family of f-divergences as measures of the distinction between a pair of

probability distributions. Such a choice has three main features: (i) it enables designing linear mappings for a wider range of inference problems (beyond classification); (ii) it provides the designs that are optimal for the inference problem at hand; and (iii) it enables characterizing the information-theoretic performance limits after linear mapping. Our analyses are focused on Gaussian models. Even though we observe that the design of the linear mapping has differences under different f-divergence measures, we have two main observations in the case of zero-mean Gaussian models: (i) the optimal design of the linear mapping is not necessarily along the most dominant components of the data matrix; and (ii) in certain regimes, irrespective of the choice of the f-divergence measure, the design of the linear map that retains the maximal divergence between the two models is robust. In such cases, this makes the optimal design of the linear map independent of the inference problem at hand rendering a degree of universality (in the considered space of the Gaussian probability measures).

The remainder of the paper is organized as follows. Section 2 provides the linear dimensionality reduction model, and it provides an overview of the f-divergence measures considered in this paper. Section 3 formulates the problem, and it helps to facilitate the mathematical analysis in subsequent sections. In Section 4, we provide a motivating operational interpretation for each f-divergence measure and then characterize an optimal design of the linear mapping for zero-mean Gaussian models. Section 5 considers numerical simulations for inference problems associated with the f-divergence measure of interest for zero-mean Gaussian models. Section 6 generalizes the theory to non-zero mean Gaussian models and discusses numerical algorithms that help characterize the design of the linear map, and Section 7 concludes the paper. A list of abbreviations used in this paper is provided on page 22.

2. Preliminaries

Consider a pair of n-dimensional Gaussian models:

$$\mathbb{P}: \ \mathcal{N}(\boldsymbol{\mu}_{\mathbb{P}}, \boldsymbol{\Sigma}_{\mathbb{P}}) \,, \quad \text{and} \quad \mathbb{Q}: \ \mathcal{N}(\boldsymbol{\mu}_{\mathbb{Q}}, \boldsymbol{\Sigma}_{\mathbb{Q}}) \,, \tag{2}$$

where $\boldsymbol{\mu}_{\mathbb{P}}, \boldsymbol{\mu}_{\mathbb{Q}}$ and $\boldsymbol{\Sigma}_{\mathbb{P}}, \boldsymbol{\Sigma}_{\mathbb{Q}}$ are two distinct mean vectors and covariance matrices, respectively, and $\mathbb{P}$ and $\mathbb{Q}$ denote their associated probability measures. The nature selects one model and generates a random variable $X \in \mathbb{R}^n$. We perform linear dimensionality reduction on X via matrix $\mathbf{A} \in \mathbb{R}^{r \times n}$, where $r < n$, rendering

$$Y \ \stackrel{\triangle}{=} \mathbf{A} \cdot X \,. \tag{3}$$

After linear mapping, the two possible distributions of Y induced by matrix $\mathbf{A}$ are denoted by $\mathbb{P}_{\mathbf{A}}$ and $\mathbb{Q}_{\mathbf{A}}$, where

$$\begin{array}{ll} \mathbb{P}_{\mathbf{A}}: & \mathcal{N}(\mathbf{A} \cdot \boldsymbol{\mu}_{\mathbb{P}}, \mathbf{A} \cdot \boldsymbol{\Sigma}_{\mathbb{P}} \cdot \mathbf{A}^{\top}) \\ \mathbb{Q}_{\mathbf{A}}: & \mathcal{N}(\mathbf{A} \cdot \boldsymbol{\mu}_{\mathbb{Q}}, \mathbf{A} \cdot \boldsymbol{\Sigma}_{\mathbb{Q}} \cdot \mathbf{A}^{\top}) \end{array} \,. \tag{4}$$

Motivated by inference problems that we discuss in Section 3, our objective is to design the linear mapping parameterized by matrix $\mathbf{A}$ that ensures that the two possible distributions of Y, i.e., $\mathbb{P}_{\mathbf{A}}$ and $\mathbb{Q}_{\mathbf{A}}$, are maximally distinguishable. That is, to design $\mathbf{A}$ as a function of the statistical models (i.e., $\boldsymbol{\mu}_{\mathbb{P}}$, $\boldsymbol{\mu}_{\mathbb{Q}}$, $\boldsymbol{\Sigma}_{\mathbb{P}}$ and $\boldsymbol{\Sigma}_{\mathbb{Q}}$) such that relevant notions of f-divergences between $\mathbb{P}_{\mathbf{A}}$ and $\mathbb{Q}_{\mathbf{A}}$ are maximized. We use a number of f-divergence measures for capturing the distinction between $\mathbb{P}_{\mathbf{A}}$ and $\mathbb{Q}_{\mathbf{A}}$, each with a distinct operational meaning under specific inference problems. For this purpose, we denote the f-divergence of $\mathbb{Q}_{\mathbf{A}}$ from $\mathbb{P}_{\mathbf{A}}$ by $D_f(\mathbf{A})$, where

$$D_f(\mathbf{A}) \ \stackrel{\triangle}{=} \mathbb{E}_{\mathbb{P}_{\mathbf{A}}}\left[f\left(\frac{\mathrm{d}\mathbb{Q}_{\mathbf{A}}}{\mathrm{d}\mathbb{P}_{\mathbf{A}}} \right) \right] . \tag{5}$$

We use the shorthand $D_f(\mathbf{A})$ for the canonical notation $D_f(\mathbb{Q}_\mathbf{A} \parallel \mathbb{P}_\mathbf{A})$ for emphasizing the dependence on $\mathbf{A}$ and for the simplicity in notations. $\mathbb{E}_{\mathbb{P}_\mathbf{A}}$ denotes the expectation with respect to $\mathbb{P}_\mathbf{A}$, and $f : (0, +\infty) \to \mathbb{R}$ is a convex function that is strictly convex at 1 and $f(1) = 0$. Strict convexity at 1 ensures that the f-divergence between a pair of probability measures is zero if and only if the probability measures are identical. Given the linear dimensionality reduction model in (3), the objective is to solve

$$\mathcal{P}: \quad \max_{\mathbf{A} \in \mathbb{R}^{r \times n}} \quad D_f(\mathbf{A}) \,, \tag{6}$$

for the following choices of the f-divergence measures.

1. *Kullback–Leibler (KL) divergence* for $f(t) = t \log t$:

$$D_{\mathsf{KL}}(\mathbf{A}) \triangleq \mathbb{E}_{\mathbb{Q}_\mathbf{A}}\left[\log \frac{\mathrm{d}\mathbb{Q}_\mathbf{A}}{\mathrm{d}\mathbb{P}_\mathbf{A}}\right] . \tag{7}$$

   We also denote the KL divergence from $\mathbb{P}_\mathbf{A}$ to $\mathbb{Q}_\mathbf{A}$ by $D_{\mathsf{KL}}(\mathbb{P}_\mathbf{A} \parallel \mathbb{Q}_\mathbf{A})$.

2. *Symmetric KL divergence* for $f(t) = (t-1) \log t$:

$$D_{\mathsf{SKL}}(\mathbf{A}) \triangleq D_{\mathsf{KL}}(\mathbb{Q}_\mathbf{A} \parallel \mathbb{P}_\mathbf{A}) + D_{\mathsf{KL}}(\mathbb{P}_\mathbf{A} \parallel \mathbb{Q}_\mathbf{A}) \,. \tag{8}$$

3. *Squared Hellinger distance* for $f(t) = (1 - \sqrt{t})^2$:

$$\mathsf{H}^2(\mathbf{A}) \triangleq \int_{\mathbb{R}^r} \left(\sqrt{\mathrm{d}\mathbb{Q}_\mathbf{A}} - \sqrt{\mathrm{d}\mathbb{P}_\mathbf{A}}\right)^2 . \tag{9}$$

4. *Total variation distance* for $f(t) = \frac{1}{2} \cdot |t - 1|$:

$$d_{\mathsf{TV}}(\mathbf{A}) \triangleq \frac{1}{2} \int_{\mathbb{R}^r} |\mathrm{d}\mathbb{Q}_\mathbf{A} - \mathrm{d}\mathbb{P}_\mathbf{A}| \,. \tag{10}$$

5. *χ^2-divergence* for $f(t) = (t-1)^2$:

$$\chi^2(\mathbf{A}) \triangleq \int_{\mathbb{R}^r} \frac{(\mathrm{d}\mathbb{Q}_\mathbf{A} - \mathrm{d}\mathbb{P}_\mathbf{A})^2}{\mathrm{d}\mathbb{P}_\mathbf{A}} \,. \tag{11}$$

   We also denote the χ^2-divergence from $\mathbb{P}_\mathbf{A}$ to $\mathbb{Q}_\mathbf{A}$ by $\chi^2(\mathbb{P}_\mathbf{A} \parallel \mathbb{Q}_\mathbf{A})$.

3. Problem Formulation

In this section, without loss of generality, we focus on the setting where one of the covariance matrices is the identity matrix, and the other one has a covariance matrix $\mathbf{\Sigma}$ in order to avoid complex representations. One key observation is that the design of $\mathbf{A}$ under different measures has strong similarities. We first note that, by defining $\bar{\mathbf{A}} \triangleq \mathbf{A} \cdot \mathbf{\Sigma}_\mathbb{P}^{1/2}$, $\boldsymbol{\mu} \triangleq \mathbf{\Sigma}_\mathbb{P}^{-1/2} \cdot (\boldsymbol{\mu}_\mathbb{Q} - \boldsymbol{\mu}_\mathbb{P})$, and $\mathbf{\Sigma} \triangleq \mathbf{\Sigma}_\mathbb{P}^{-1/2} \cdot \mathbf{\Sigma}_\mathbb{Q} \cdot \mathbf{\Sigma}_\mathbb{P}^{-1/2}$, designing $\mathbf{A}$ for maximally distinguishing

$$\mathcal{N}(\mathbf{A} \cdot \boldsymbol{\mu}_\mathbb{P}, \mathbf{A} \cdot \mathbf{\Sigma}_\mathbb{P} \cdot \mathbf{A}^\top) \quad \text{and} \quad \mathcal{N}(\mathbf{A} \cdot \boldsymbol{\mu}_\mathbb{Q}, \mathbf{A} \cdot \mathbf{\Sigma}_\mathbb{Q} \cdot \mathbf{A}^\top) \tag{12}$$

is equivalent to designing $\bar{\mathbf{A}}$ for maximally distinguishing

$$\mathcal{N}(\mathbf{0}, \bar{\mathbf{A}} \cdot \bar{\mathbf{A}}^\top) \quad \text{and} \quad \mathcal{N}(\bar{\mathbf{A}} \cdot \boldsymbol{\mu}, \bar{\mathbf{A}} \cdot \mathbf{\Sigma} \cdot \bar{\mathbf{A}}^\top) \,. \tag{13}$$

Hence, without loss of generality, we focus on the setting where $\boldsymbol{\mu}_\mathbb{P} = \mathbf{0}$, $\mathbf{\Sigma}_\mathbb{P} = \mathbf{I}_n$, and $\mathbf{\Sigma}_\mathbb{Q} = \mathbf{\Sigma}$. Next, we show that determining an optimal design for $\mathbf{A}$ can be confined to the class of semi-orthogonal matrices.

Theorem 1. *For every $\mathbf{A}$, there exists a semi-orthogonal matrix $\bar{\mathbf{A}}$ such that $D_f(\bar{\mathbf{A}}) = D_f(\mathbf{A})$.*

Proof. See Appendix A. □

This observation indicates that we can reduce the unconstrained problem in (6) to the following constrained problem:

$$\mathcal{Q}: \quad \max_{\mathbf{A}\in\mathbb{R}^{r\times n}} \quad D_f(\mathbf{A}) \quad \text{s.t.} \quad \mathbf{A}\cdot\mathbf{A}^\top = \mathbf{I}_r \,. \tag{14}$$

We show that the design of $\mathbf{A}$ in the case of $\boldsymbol{\mu} = \mathbf{0}$, under the considered f-divergence measures, directly relates to analyzing the eigenspace of matrix $\boldsymbol{\Sigma}$. For this purpose, we denote the non-negative eigenvalues of $\boldsymbol{\Sigma}$ ordered in the descending order by $\{\lambda_i : i \in [n]\}$, where for an integer m we have defined $[m] = \{1, \ldots, m\}$. For an arbitrary permutation function $\pi : [n] \to [n]$, we denote the permutation of $\{\lambda_i : i \in [n]\}$ with respect to π by $\{\lambda_{\pi(i)} : i \in [n]\}$. We also denote the eigenvalues of $\mathbf{A}\cdot\boldsymbol{\Sigma}\cdot\mathbf{A}^\top$ ordered in the descending order by $\{\gamma_i : i \in [r]\}$. Throughout the analysis, we frequently use Poincaré separation theorem [38] for finding the row space of matrix $\mathbf{A}$ with respect to the eigenvalues of $\boldsymbol{\Sigma}$.

Theorem 2 (Poincaré Separation Theorem)**.** *Let $\boldsymbol{\Sigma}$ be a real symmetric $n \times n$ matrix and $\mathbf{A}$ be a semi-orthogonal $r \times n$ matrix. The eigenvalues of $\boldsymbol{\Sigma}$ denoted by $\{\lambda_i : i \in [n]\}$ (sorted in the descending order) and the eigenvalues of $\mathbf{A}\cdot\boldsymbol{\Sigma}\cdot\mathbf{A}^\top$ denoted by $\{\gamma_i : i \in [r]\}$ (sorted in the descending order) satisfy*

$$\lambda_{n-(r-i)} \leq \gamma_i \leq \lambda_i \,, \quad \forall i \in [r] \,. \tag{15}$$

Finally, we define the following functions, which we will refer to frequently throughout the paper:

$$h_1(\mathbf{A}) \triangleq \mathbf{A}\cdot\boldsymbol{\Sigma}\cdot\mathbf{A}^\top \,, \tag{16}$$

$$h_2(\mathbf{A}) \triangleq \boldsymbol{\mu}^\top\cdot\mathbf{A}^\top\cdot\mathbf{A}\cdot\boldsymbol{\mu} \,, \tag{17}$$

$$h_3(\mathbf{A}) \triangleq \boldsymbol{\mu}^\top\cdot\mathbf{A}^\top\cdot[h_1(\mathbf{A})]^{-1}\cdot\mathbf{A}\cdot\boldsymbol{\mu} \,. \tag{18}$$

In the next sections, we analyze the design of $\mathbf{A}$ under different f-divergence measures. In particular, in Sections 4 and 5, we focus on zero-mean Gaussian models for $\mathbb{P}$ and $\mathbb{Q}$ where we provide an operational interpretation of the measure in the dichotomous mode in (4). Subsequently, we will discuss the generalization to non-zero mean Gaussian models in Section 6.

4. Main Results for Zero-Mean Gaussian Models

In this section, we analyze problem $\mathcal{Q}$ defined in (14) for each of the f-divergence measures separately. Specifically, for each case, we briefly provide an inference problem as a motivating example, in the context of which we relate the optimal performance limit of that inference problem to the f-divergence of interest. These analyses are provided in Sections 4.1–4.5. Subsequently, we provide the main results on the optimal design of the linear mapping matrix $\mathbf{A}$ in Section 4.6.

4.1. Kullback–Leibler Divergence

4.1.1. Motivation

The KL divergence, being the expected value of the log-likelihood ratio, captures, at least partially, the performance of a wide range of inference problems. One specific problem whose performance is completely captured by $D_{\mathsf{KL}}(\mathbf{A})$ is the quickest change-point detection. Consider an observation process (time-series) $\{X_t : t \in \mathbb{N}\}$ in which the observations $X_t \in \mathbb{R}^n$ are generated by a distribution with probability measure $\mathbb{P}$ specified in (2). This distribution changes to $\mathbb{Q}$ at an unknown (random or deterministic) time κ, i.e.,

$$X_t \sim \mathbb{P} \quad t < \kappa \,, \qquad \text{and} \qquad X_t \sim \mathbb{Q} \quad t \geq \kappa \quad . \tag{19}$$

Change-point detection algorithms sample the observation process sequentially and aim to detect the change point with the minimal delay after it occurs subject to a false alarm constraint. Hence, the two key figures of merit capturing the performance of a sequential change-point detection algorithm are the average detection delay (ADD) and the rate of false alarms. Whether the change-point κ is random or deterministic gives rise to two broad classes of quickest change-point detection problems, namely the Bayesian setting (κ is random) and minimax setting (κ is deterministic). Irrespective of their discrepancies in settings and the nature of performance guarantees, the ADD for the (asymptotically) optimal algorithms are in the form [39]:

$$\mathsf{ADD} \sim \frac{c_1}{D_{\mathsf{KL}}(\mathbb{Q} \parallel \mathbb{P})} \,. \tag{20}$$

Hence, after the linear mapping induced by matrix $\mathbf{A}$, for the ADD, we have

$$\mathsf{ADD} \sim \frac{c_2}{D_{\mathsf{KL}}(\mathbb{Q}_{\mathbf{A}} \parallel \mathbb{P}_{\mathbf{A}})} \,, \tag{21}$$

where c_1 and c_2 are constants specified by the false alarm constraints. Clearly, the design of $\mathbf{A}$ that minimizes the ADD will be maximizing the disparity between the pre- and post-change distributions $\mathbb{P}_{\mathbf{A}}$ and $\mathbb{Q}_{\mathbf{A}}$, respectively.

4.1.2. Connection between D_{KL} and $\mathbf{A}$

By noting that $\mathbf{A}$ is a semi-orthogonal matrix and recalling that the eigenvalues of $h_1(\mathbf{A})$ are denoted by $\{\gamma_i : i \in [r]\}$, simple algebraic manipulations simplify $D_{\mathsf{KL}}(\mathbb{Q}_{\mathbf{A}} \parallel \mathbb{P}_{\mathbf{A}})$ to:

$$D_{\mathsf{KL}}(\mathbb{Q}_{\mathbf{A}} \parallel \mathbb{P}_{\mathbf{A}}) = \frac{1}{2}\left[\log \frac{1}{|h_1(\mathbf{A})|} - r + \mathrm{Tr}[h_1(\mathbf{A})] + h_2(\mathbf{A})\right] . \tag{22}$$

By setting, and leveraging, Theorem 2, the problem of finding an optimal design for $\mathbf{A}$ that solves (14) can be found as the solution to:

$$\max_{\{\gamma_i \,:\, i \in [r]\}} \sum_{i=1}^{r} g_{\mathsf{KL}}(\gamma_i) \qquad \text{s.t.} \qquad \lambda_{n-(r-i)} \leq \gamma_i \leq \lambda_i \;\; \forall i \in [r] \,, \tag{23}$$

where we have defined

$$g_{\mathsf{KL}}(x) \triangleq \frac{1}{2}(x - \log x - 1) \,. \tag{24}$$

Likewise, finding the optimal design for $\mathbf{A}$ that optimizes $D_{\mathsf{KL}}(\mathbb{P}_{\mathbf{A}} \parallel \mathbb{Q}_{\mathbf{A}})$ when $\boldsymbol{\mu} = 0$ can be found by replacing $g_{\mathsf{KL}}(\gamma_i)$ by $g_{\mathsf{KL}}\left(\frac{1}{\gamma_i}\right)$ in (23). In either case, the optimal design of $\mathbf{A}$ is constructed by choosing r eigenvectors of $\boldsymbol{\Sigma}$ as the rows of $\mathbf{A}$. The results and observations are formalized in Section 4.6.

4.2. Symmetric KL Divergence

4.2.1. Motivation

The KL divergence discussed in Section 4.1 is an asymmetric measure of separation between two probability measures. It is symmetrized by adding two directed divergence measures in reverse directions. The symmetric KL divergence has applications in model selection problems in which the model selection criteria is based on a measure of disparity between the true model and the approximating models. As shown in Reference [40], using the symmetric KL divergence outperforms the individual directed KL divergences since it better reflects the risks associated with underfitting and overfitting of the models, respectively.

4.2.2. Connection between D_{SKL} and $\mathbf{A}$

For a given $\mathbf{A}$, the symmetric KL divergence of interest specified in (8) is given by

$$D_{\mathsf{SKL}}(\mathbf{A}) = \frac{1}{2} \cdot \left[\mathrm{Tr}\Big([h_1(\mathbf{A})]^{-1} + h_1(\mathbf{A}) \Big) + h_2(\mathbf{A}) + h_3(\mathbf{A}) \right] - r \, . \tag{25}$$

By setting $\boldsymbol{\mu} = 0$, and leveraging Theorem 2, the problem of finding an optimal design for $\mathbf{A}$ that solves (14) can be found as the solution to:

$$\max_{\{\gamma_i \, : \, i \in [r]\}} \sum_{i=1}^{r} g_{\mathsf{SKL}}(\gamma_i) \qquad \text{s.t.} \qquad \lambda_{n-(r-i)} \leq \gamma_i \leq \lambda_i \;\; \forall i \in [r] \, , \tag{26}$$

where we have defined

$$g_{\mathsf{SKL}}(x) \triangleq \frac{1}{2} \left(x + \frac{1}{x} - 2 \right) . \tag{27}$$

4.3. Squared Hellinger Distance

4.3.1. Motivation

Squared Hellinger distance facilitates analysis in high dimensions, especially when other measures fail to take closed-form expressions. We will discuss an important instance of this in the next subsection in the analysis of d_{TV}. Squared Hellinger distance is symmetric, and it is confined in the range $[0, 2]$.

4.3.2. Connection between H^2 and $\mathbf{A}$

For a given matrix $\mathbf{A}$, we have the following closed-form expression:

$$\mathsf{H}^2(\mathbf{A}) = \quad 2 - 2 \frac{|4 \cdot h_1(\mathbf{A})|^{\frac{1}{4}}}{|h_1(\mathbf{A}) + \mathbf{I}_r|^{\frac{1}{2}}} \cdot \exp\left(- \frac{\boldsymbol{\mu}^\top \cdot \mathbf{A}^\top \cdot [h_1(\mathbf{A}) + \mathbf{I}_r]^{-1} \cdot \mathbf{A} \cdot \boldsymbol{\mu}}{4} \right) . \tag{28}$$

By setting $\boldsymbol{\mu} = 0$, and leveraging Theorem 2, the problem of finding an optimal design for $\mathbf{A}$ that solves (14) can be found as the solution to:

$$\max_{\{\gamma_i \, : \, i \in [r]\}} \prod_{i=1}^{r} g_{\mathsf{H}}(\gamma_i) \qquad \text{s.t.} \qquad \lambda_{n-(r-i)} \leq \gamma_i \leq \lambda_i \;\; \forall i \in [r] \, , \tag{29}$$

where we have defined

$$g_{\mathsf{H}}(x) \triangleq \frac{(x+1)^2}{x} \, . \tag{30}$$

4.4. Total Variation Distance

4.4.1. Motivation

The total variation distance appears as the key performance metric in binary hypothesis testing and in high-dimensional inference, e.g., Le Cam's method for the binary quantization and testing of the individual dimensions (which is in essence binary hypothesis testing). In particular, for the simple binary hypothesis testing model in (65), the minimum total probability of error (sum of type-I and type-II error probabilities) is related to the total variation $d_{\mathsf{TV}}(\mathbf{A})$. Specifically, for a decision rule $d : X \to \{\mathsf{H}_0, \mathsf{H}_1\}$, the following holds:

$$\inf_{d} \left[\mathbb{P}_{\mathbf{A}}(d = \mathsf{H}_1) + \mathbb{Q}_{\mathbf{A}}(d = \mathsf{H}_0) \right] = 1 - d_{\mathsf{TV}}(\mathbf{A}) \, . \tag{31}$$

The total variation between two Gaussian distributions does not have a closed-form expression. Hence, unlike the other settings, an optimal solution to (6) in this context cannot be obtained analytically. Alternatively, in order to gain intuition into the structure of a

near optimal matrix $\mathbf{A}$, we design $\mathbf{A}$ such that it optimizes known bounds on $d_{\mathsf{TV}}(\mathbf{A})$. In particular, we use two sets of bounds on $d_{\mathsf{TV}}(\mathbf{A})$. One set is due to bounding it via the Hellinger distance, and another set is due to a recent study that established upper and lower bounds that are identical up to a constant factor [41].

4.4.2. Connection between d_{TV} and $\mathbf{A}$

(1) Bounding by Hellinger Distance: The total variation distance can be bounded by the Hellinger distance according to

$$\frac{1}{2}\mathsf{H}^2(\mathbf{A}) \le d_{\mathsf{TV}}(\mathbf{A}) \le \mathsf{H}(\mathbf{A})\sqrt{1-\frac{\mathsf{H}^2(\mathbf{A})}{4}}\,. \tag{32}$$

It can be readily verified that these bounds are monotonically increasing with $\mathsf{H}^2(\mathbf{A})$ in the interval $[0,2]$. Hence, they are maximized simultaneously by maximizing the squared Hellinger distance as discussed in Section 4.3. We refer to this bound as the Hellinger bound.

(2) Matching Bounds up to a Constant: The second set of bounds that we used are provided in Reference [41]. These bounds relate the total variation between two Gaussian models to the Frobenius norm (FB) of a matrix related to their covariance matrices. Specifically, these FB-based bounds on the total variation $d_{\mathsf{TV}}(\mathbf{A})$ are given by

$$\frac{1}{100} \le \frac{d_{\mathsf{TV}}(\mathbf{A})}{\min\{1, \sqrt{\sum_{i=1}^{r} g_{\mathsf{TV}}(\gamma_i)}\}} \le \frac{3}{2}\,, \tag{33}$$

where we have defined

$$g_{\mathsf{TV}}(x) \triangleq \left(\frac{1}{x}-1\right)^2. \tag{34}$$

Since the lower and upper bounds on $d_{\mathsf{TV}}(\mathbf{A})$ are identical up to a constant, they will be maximized by the same design of $\mathbf{A}$.

4.5. χ^2-Divergence

4.5.1. Motivation

χ^2-divergence appears in a wide range of statistical estimation problems for the purpose of finding a lower bound on the estimation noise variance. For instance, consider the canonical problem of estimating a latent variable θ from the observed data X, and denote two candidate estimates by $p(X)$ and $q(X)$. Define $\mathbb{P}$ and $\mathbb{Q}$ as the probability measures of $p(X)$ and $q(X)$, respectively. According to the Hammersly-Chapman-Robbins (HCR) bound on the quadratic loss function, for any estimator $\hat{\theta}$, we have

$$\mathsf{var}_\theta(\hat{\theta}) \ge \sup_{p\neq q} \frac{\left[\mathbb{E}_{\mathbb{Q}}[q(X)] - \mathbb{E}_{\mathbb{P}}[p(X)]\right]^2}{\chi^2(\mathbb{Q}\parallel\mathbb{P})}\,, \tag{35}$$

which, for unbiased estimators p and q, simplifies to the Cramér-Rao lower bound

$$\mathsf{var}_\theta(\hat{\theta}) \ge \sup_{p\neq q} \frac{(q-p)^2}{\chi^2(\mathbb{Q}\parallel\mathbb{P})}\,, \tag{36}$$

depending on $\mathbb{P}$ and $\mathbb{Q}$ through their χ^2-divergence. Besides the applications to estimation problems, χ^2 is easier to compute compared to some of other f-divergence measures (e.g., total variation). Specifically, for product distributions χ^2 tensorizes to be expressed in terms of the one-dimensional components that are easier to compute than the KL divergence and TV variation distance. Hence, a combination of bounding other measures with χ^2 and then analyzing χ^2 appears in a wide range of inference problems.

4.5.2. Connection between χ^2 and $\mathbf{A}$

By setting $\boldsymbol{\mu} = 0$, for a given matrix $\mathbf{A}$, from (11), we have the following closed-form expression:

$$\chi^2(\mathbf{A}) = \frac{1}{|h_1(\mathbf{A})|\sqrt{|2(h_1(\mathbf{A}))^{-1} - \mathbf{I}_r|}} - 1 \tag{37}$$

$$= \prod_{i=1}^{r} g_{\chi_1}(\gamma_i) - 1\,, \tag{38}$$

where we have defined

$$g_{\chi_1}(x) \triangleq \frac{1}{\sqrt{x(2-x)}}\,. \tag{39}$$

As we show in Appendix C, for $\chi^2(\mathbf{A})$ to exist (i.e., be finite), all the eigenvalues $\{\lambda_i : i \in [r]\}$ should fall in the interval $(0, 2)$. Subsequently, finding the optimal design for $\mathbf{A}$ that optimizes $\chi^2(\mathbb{P}_{\mathbf{A}} \parallel \mathbb{Q}_{\mathbf{A}})$ when $\boldsymbol{\mu} = 0$ can be done by replacing g_{χ_1} in (38) by g_{χ_2}, which is given by

$$g_{\chi_2}(x) \triangleq \sqrt{\frac{x^2}{2x-1}}\,. \tag{40}$$

Based on this, and by following a similar line of argument as in the case of the KL divergence, designing an optimal $\mathbf{A}$ reduces to identifying a subset of the eigenvalues of $\boldsymbol{\Sigma}$ and assigning their associated eigenvectors as the rows of matrix $\mathbf{A}$. These observations are formalized in Section 4.6.

4.6. Main Results

In this section, we provide analytical closed-form solutions to design optimal matrices $\mathbf{A}$ for the following f-divergence measures: D_{KL}, D_{SKL}, H^2, and χ^2. The total variation measure d_{TV} does not admit a closed-form for Gaussian models. In this case, we provide a design for $\mathbf{A}$ that optimizes the bound we have provided for d_{TV} in Section 4.4. Due to their structural similarities of the results, we group and treat D_{KL}, D_{SKL}, and d_{TV} in Theorem 3. Similarly, we group and treat H^2 and χ^2 in Theorem 4.

Theorem 3 (D_{KL}, D_{SKL}, d_{TV})**.** *For a given function $g : \mathbb{R} \to \mathbb{R}$, define the permutations:*

$$\pi^* \triangleq \arg\max_{\pi} \sum_{i=1}^{r} g(\lambda_{\pi(i)})\,. \tag{41}$$

Then, for $D_f(\mathbf{A}) \in \{D_{\mathsf{KL}}(\mathbf{A}), D_{\mathsf{SKL}}(\mathbf{A}), d_{\mathsf{TV}}(\mathbf{A})\}$ and functions $g_f \in \{g_{\mathsf{KL}}, g_{\mathsf{SKL}}, g_{\mathsf{TV}}\}$:

1. *For maximizing D_f, set $g = g_f$ and select the eigenvalues of $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top$ as*

$$\gamma_i = \lambda_{\pi^*(i)}\,, \qquad \text{for} \qquad i \in [r]\,. \tag{42}$$

2. *Row $i \in [r]$ of matrix $\mathbf{A}$ is the eigenvector of $\boldsymbol{\Sigma}$ associated with the eigenvalue γ_i.*

Proof. See Appendix B. □

By further leveraging the structures of functions $g_{\mathsf{KL}}, g_{\mathsf{SKL}}$, and g_{TV}, we can simplify approaches for designing the matrix $\mathbf{A}$. Specifically, note that the functions $g_{\mathsf{KL}}, g_{\mathsf{SKL}}, and g_{\mathsf{TV}}$ are all strictly convex functions taking their global minima at $x = 1$. Based on this, we have the following observations.

Corollary 1 (D_{KL}, D_{SKL}, d_{TV}). *For maximizing* $D_f(\mathbf{A}) \in \{D_{\mathsf{KL}}(\mathbf{A}), D_{\mathsf{SKL}}(\mathbf{A}), d_{\mathsf{TV}}(\mathbf{A})\}$, *when* $\lambda_n \geq 1$, *we have* $\gamma_i = \lambda_i$ *for all* $i \in [r]$, *and the rows of* $\mathbf{A}$ *are eigenvectors of* $\mathbf{\Sigma}$ *associated with its r largest eigenvalues, i.e.,* $\{\lambda_i : i \in [r]\}$.

Corollary 2 (D_{KL}, D_{SKL}, d_{TV}). *For maximizing* $D_f(\mathbf{A}) \in \{D_{\mathsf{KL}}(\mathbf{A}), D_{\mathsf{SKL}}(\mathbf{A}), d_{\mathsf{TV}}(\mathbf{A})\}$, *when* $\lambda_1 \leq 1$, *we have* $\gamma_i = \lambda_{n-r+i}$ *for all* $i \in [r]$, *and the rows of* $\mathbf{A}$ *are eigenvectors of* $\mathbf{\Sigma}$ *associated with its r smallest eigenvalues, i.e.,* $\{\lambda_i : i \in \{n-r+1, \ldots, n\}\}$.

Remark 1. *In order to maximize* $D_f(\mathbf{A}) \in \{D_{\mathsf{KL}}(\mathbf{A}), D_{\mathsf{SKL}}(\mathbf{A}), d_{\mathsf{TV}}(\mathbf{A})\}$ *when* $\lambda_n \leq 1 \leq \lambda_1$, *finding the best permutation of eigenvalues involves sorting all the n eigenvalues* λ_i*'s and subsequently performing r comparisons as illustrated in Algorithm 1. This amounts to* $\mathcal{O}(n \cdot \log(n))$ *time complexity instead of* $\mathcal{O}(n \cdot \log(r))$ *time complexity involved in determining the design for* $\mathbf{A}$ *in the case of Corollaries 1 and 2, which require finding the r extreme eigenvalues in determining the design for* π^*.

Remark 2. *The optimal design of* $\mathbf{A}$ *often does not involve being aligned with the largest eigenvalues of the covariance matrix* $\mathbf{\Sigma}$, *which is in contrast to some of the key approaches to linear dimensionality reduction that generally perform linear mapping along the eigenvectors associated with the largest eigenvalues of the covariance matrix. When the eigenvalues of* $\mathbf{\Sigma}$ *are all smaller than 1, in particular,* $\mathbf{A}$ *will be designed by choosing eigenvectors associated with the smallest eigenvalues of* $\mathbf{\Sigma}$ *in order to preserve largest separability.*

Next, we provide the counterpart results for the H^2 and χ^2-divergence measures. Their major distinction from the previous three measures is that, for these two, $D_f(\mathbf{A})$ can be decomposed into a product of individual functions of the eigenvalues $\{\gamma_i : i \in [r]\}$. Next, we provide the counterparts of Theorem 3 and Corollaries 1 and 2 for H^2 and χ^2.

Theorem 4 (H^2, χ^2). *For a given function* $g : \mathbb{R} \to \mathbb{R}$, *define the permutations:*

$$\pi^* \triangleq \arg\max_{\pi} \prod_{i=1}^{r} g(\lambda_{\pi(i)}) \,. \tag{43}$$

Then, for $D_f(\mathbf{A}) \in \{\mathsf{H}^2(\mathbf{A}), \chi^2(\mathbf{A}), \chi^2(\mathbb{P}_{\mathbf{A}} \parallel \mathbb{Q}_{\mathbf{A}})\}$ *and functions* $g_f \in \{g_{\mathsf{H}}, g_{\chi_1}, g_{\chi_2}\}$:

1. *For maximizing* D_f, *set* $g = g_f$ *and select the eigenvalues of* $\mathbf{A}\mathbf{\Sigma}\mathbf{A}^\top$ *as*

$$\gamma_i = \lambda_{\pi^*(i)} \,, \qquad \text{for} \qquad i \in [r] \,. \tag{44}$$

2. *Row* $i \in [r]$ *of matrix* $\mathbf{A}$ *is the eigenvector of* $\mathbf{\Sigma}$ *associated with the eigenvalue* γ_i.

Proof. See Appendix C. □

Next, note that g_{H} is a strictly convex function taking its global minimum at $x = 1$. Furthermore, g_{χ_i} for $i \in [2]$ are strictly convex over $(0, 2)$ and take their global minimum at $x = 1$.

Corollary 3 (H^2, χ^2). *For maximizing* $D_f(\mathbf{A}) \in \{\mathsf{H}^2(\mathbf{A}), \chi^2(\mathbf{A}), \chi^2(\mathbb{P}_{\mathbf{A}} \parallel \mathbb{Q}_{\mathbf{A}})\}$, *when* $\lambda_n \geq 1$, *we have* $\gamma_i = \lambda_i$ *for all* $i \in [r]$, *and the rows of* $\mathbf{A}$ *are eigenvectors of* $\mathbf{\Sigma}$ *associated with its r largest eigenvalues, i.e.,* $\{\lambda_i : i \in [r]\}$.

Corollary 4 (H^2, χ^2). *For maximizing* $D_f(\mathbf{A}) \in \{\mathsf{H}^2(\mathbf{A}), \chi^2(\mathbf{A}), \chi^2(\mathbb{P}_{\mathbf{A}} \parallel \mathbb{Q}_{\mathbf{A}})\}$, *when* $\lambda_1 \leq 1$, *we have* $\gamma_i = \lambda_{n-r+i}$ *for all* $i \in [r]$, *and the rows of* $\mathbf{A}$ *are eigenvectors of* $\mathbf{\Sigma}$ *associated with its r smallest eigenvalues, i.e.,* $\{\lambda_i : i \in \{n-r+1, \ldots, n\}\}$.

Algorithm 1 Optimal Permutation π^* When $\lambda_n \leq 1 \leq \lambda_1$

Initialize $i \leftarrow n$, $j \leftarrow 1$, $p_k \leftarrow \lambda_k \; \forall k \in \{i, j\}$, $\pi^* \leftarrow \emptyset$
Sort the eigenvalues of $\mathbf{\Sigma}$ in descending order $\{\lambda_k : k \in [n]\}$
while $|\pi^*| \neq r$ **do**
    if $g_f(p_i) > g_f(p_j)$ **then**
        $\pi^* \leftarrow \pi^* \cup \{p_i\}$
        $i \leftarrow i - 1$
    else
        $\pi^* \leftarrow \pi^* \cup \{p_j\}$
        $j \leftarrow j + 1$
    end if
end while
return π^*

Finally, we remark that, unlike the other measures, total variation does not admit a closed-form, and we used two sets of tractable bounds to analyze this case of total variations. By comparing the design of $\mathbf{A}$ based on different bounds, we have the following observation.

Remark 3. *We note that both sets of bounds lead to the same design of* $\mathbf{A}$ *when either* $\lambda_1 \leq 1$ *or* $\lambda_n \geq 1$*. Otherwise, each will be selecting a different set of the eigenvectors of* $\mathbf{\Sigma}$ *to construct* $\mathbf{A}$ *according to the functions*

$$g_{\mathsf{H}}(x) = \frac{(x+1)^2}{x} \quad \text{versus} \quad g_{\mathsf{TV}}(x) = \left(\frac{1}{x} - 1\right)^2. \tag{45}$$

5. Zero-Mean Gaussian Models–Simulations

5.1. KL Divergence

In this section, we show gains of the above analysis for the KL divergence measure $D_{\mathsf{KL}}(\mathbf{A})$ through simulations on a change-point detection problem. We focus on the minimax setting in which the change-point κ is deterministic. The objective is to detect a change in the stochastic process X_t with minimal delay after the change in the probability measure occurs at κ and define $\tau \in \mathbb{N}$ as the time that we can form a confident decision. A canonical model to quantify the decision delay is the conditional average detection delay (CADD) due to Pollak [42]

$$\mathsf{CADD}(\tau) \triangleq \sup_{\kappa \geq 1} \mathbb{E}_\kappa \left[\tau - \kappa \mid \tau \geq \kappa\right], \tag{46}$$

where $\mathbb{E}_\kappa$ is the expectation with respect to the probability distribution when the change happens at time κ. The objective of this formulation is to optimize the decision delay for the worst-case realization of the random change-point κ (that is, the change-point realization that leads to the maximum decision delay), while the constraints on the false alarm rate are satisfied. In this formulation, this worst-case realization is $\kappa = 1$, in which case all the data points are generated from the post-change distribution. In the minimax setting, a reasonable measure of false alarms is the mean-time to false alarm, or its reciprocal, which is the false alarm rate (FAR) defined as

$$\mathsf{FAR}(\tau) \triangleq \frac{1}{\mathbb{E}_\infty[\tau]}, \tag{47}$$

where $\mathbb{E}_\infty$ is the expectation with respect to the distribution when a change never occurs, i.e., $\kappa \triangleq \infty$. A standard approach to balance the trade-off between decision delay and false alarm rates involves solving [42]

$$\min_{\tau} \mathsf{CADD}(\tau) \qquad \text{s.t.} \qquad \mathsf{FAR}(\tau) \leq \alpha \,, \tag{48}$$

where $\alpha \in \mathbb{R}_+$ controls the rate of false alarms. For the quickest change-point detection formulation in (48), the popular cumulative sum (CuSum) test generates the optimal solutions, involving computing the following test statistic:

$$W[t] \triangleq \max_{1 \leq k \leq t+1} \sum_{i=k}^{t} \log\left(\frac{d\mathbb{Q}_{\mathbf{A}}(X_i)}{d\mathbb{P}_{\mathbf{A}}(X_i)}\right) . \tag{49}$$

Computing $W[t]$ follows a convenient recursion given by

$$W[t] \triangleq \left(W[t-1] + \log\left(\frac{d\mathbb{Q}_{\mathbf{A}}(X_t)}{d\mathbb{P}_{\mathbf{A}}(X_t)}\right)\right)^+ , \tag{50}$$

where $W[0] = 0$. The CuSum statistic declares a change at a stopping time τ given by

$$\tau \triangleq \inf\{t \geq 1 : W[t] > C\} \,, \tag{51}$$

where C is chosen such that the constraint on $\mathsf{FAR}(\tau)$ in (48) is satisfied.

In this setting, we consider two zero-mean Gaussian models with the following pre- and post-linear dimensionality reduction structures:

$$\begin{array}{lllll} \mathbb{P}: & \mathcal{N}(\mathbf{0}, \mathbf{I}_n) & \text{and} & \mathbb{Q}: & \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}) \\ \mathbb{P}_{\mathbf{A}}: & \mathcal{N}(\mathbf{0}, \mathbf{I}_r) & \text{and} & \mathbb{Q}_{\mathbf{A}}: & \mathcal{N}(\mathbf{0}, h_1(\mathbf{A})) \end{array} \,, \tag{52}$$

where the covariance matrix $\mathbf{\Sigma}$ is generated randomly, and its eigenvalues are sampled from a uniform distribution. In particular, for the original data dimension n, $\lceil 0.9n \rceil$ eigenvalues are sampled such that $\{\lambda_i \sim \mathcal{U}(0.064, 1)\}$, and the remaining eigenvalues are sampled such that $\{\lambda_i \sim \mathcal{U}(1, 4.24)\}$. We note that this is done since the objective function lies in the same range for the eigenvalues within the range $[0.0649, 1]$ and $[1, 4.24]$. In order to consider the worst case detection delay, we set $\kappa = 1$ and generate stochastic observations according to the model described in (52) that follows the change-point detection model in (19). For every random realization of covariance matrix $\mathbf{\Sigma}$, we run the CuSum statistic (50), where we generate $\mathbf{A}$ according to the following two schemes:

(1) Largest eigen modes: In this scheme, the linear map $\mathbf{A}$ is designed such that its rows are eigenvectors associated with the r largest eigenvalues of $\mathbf{\Sigma}$.

(2) Optimal design: In this scheme, the linear map $\mathbf{A}$ is designed such that its rows are eigenvectors associated with r eigenvalues of $\mathbf{\Sigma}$ that maximize $D_{\mathsf{KL}}(\mathbf{A})$ according to Theorem 3.

In order to evaluate and compare the performance of the two schemes, we compute the ADD obtained by running a Monte-Carlo simulation over 5000 random realizations of the stochastic process X_t following the change-point detection model in (19) for every random realization of $\mathbf{\Sigma}$ and for each reduced dimension $1 \leq r \leq 9$. The detection delays obtained are then averaged again over 100 random realizations of covariance matrices $\mathbf{\Sigma}$ for each reduced dimension r. Figure 1 shows the plot for ADD versus r for multiple initial data dimension n and for a fixed $\mathsf{FAR} = \frac{1}{5000}$. Owing to the dependence on $D_{\mathsf{KL}}(\mathbf{A})$ given in (21), the delay associated with the optimal linear mapping in Theorem 3 achieves better performance.

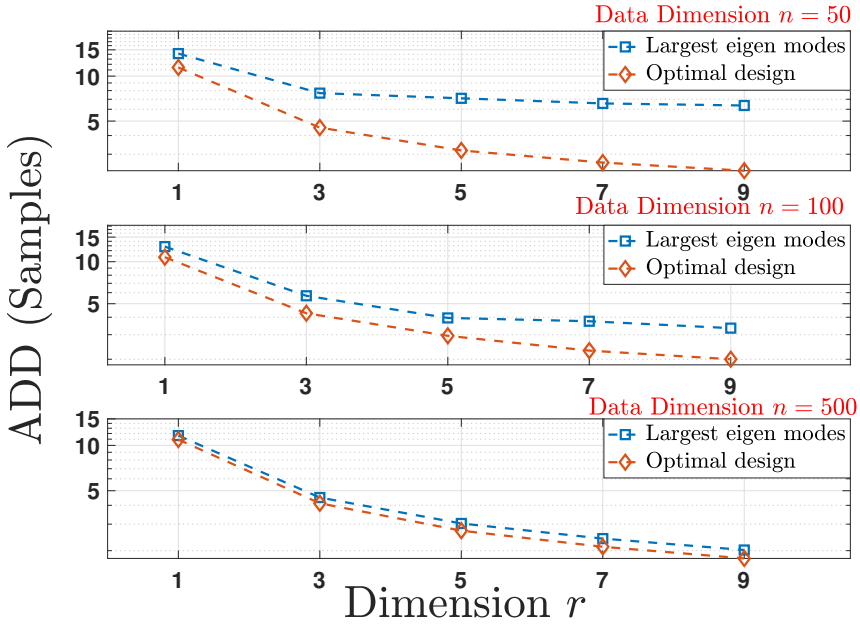

Figure 1. Comparison of the average detection delay (ADD) under the optimal design and largest eigen modes schemes for multiple reduced data dimensions r as a function of original data dimension n for a fixed false alarm rate (FAR) which is equal to $1/5000$.

5.2. Symmetric KL Divergence

In this section, we show the gains of the analysis by numerically computing $D_{\mathsf{SKL}}(\mathbf{A})$. We follow the pre- and post-linear dimensionality reduction structures given in (52), where the covariance matrix $\mathbf{\Sigma}$ is randomly generated following the setup used in Section 5.1. As plotted in Figure 2, by choosing the design scheme for $D_{\mathsf{SKL}}(\mathbf{A})$ according to Theorem 3, the optimal design outperforms other schemes.

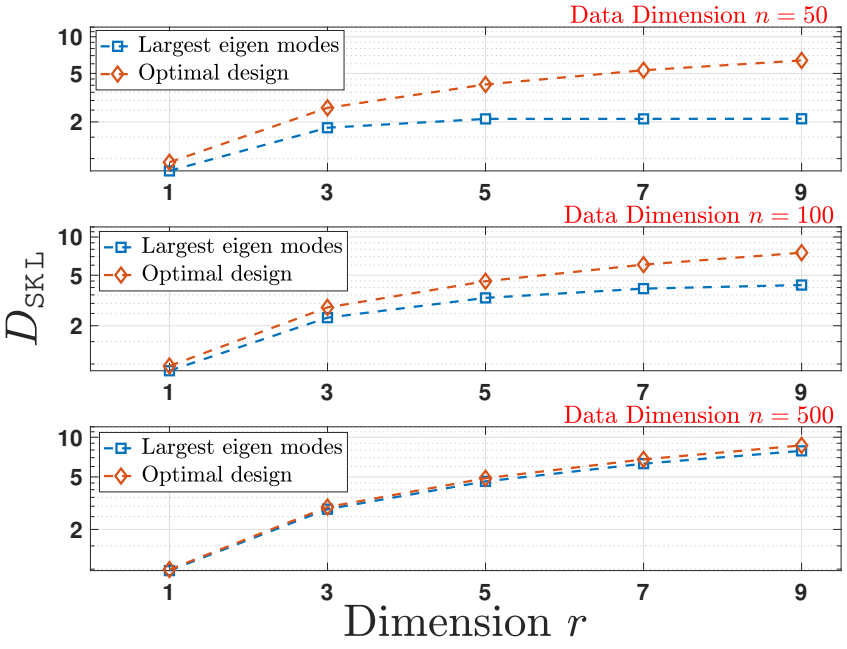

Figure 2. Comparison of the empirical average computed for the optimal design and largest eigen modes schemes for multiple reduced data dimensions r as a function of original data dimension n.

5.3. Squared Hellinger Distance

We consider a Bayesian hypothesis testing problem given class a priori parameters $p_{\mathbb{P}_{\mathbf{A}}}$, $p_{\mathbb{Q}_{\mathbf{A}}}$ and Gaussian class conditional densities for the linear dimensionality reduction model in (52). Without loss of generality, we assume a 0–1 loss function associated with misclassification for the hypothesis test. In order to quantify the performance of the Bayes decision rule, it is imperative to compute the associated probability of error, also known as the Bayes error, which we denote by P_e. Since, in general, computing P_e for the optimal decision rule for multivariate Gaussian conditional densities is intractable, numerous techniques have been devised to bound P_e. Owing to its simplicity, one of the most commonly employed metric is the Bhattacharyya coefficient given by

$$\mathsf{BC}(\mathbf{A}) \triangleq \int_{\mathbb{R}^r} \sqrt{\mathrm{d}\mathbb{P}_{\mathbf{A}} \cdot \mathrm{d}\mathbb{Q}_{\mathbf{A}}}\,. \tag{53}$$

The metric in (53) facilitates upper bounding the error probability as

$$P_e \leq \sqrt{p_{\mathbb{P}_{\mathbf{A}}}\, p_{\mathbb{Q}_{\mathbf{A}}}} \cdot \mathsf{BC}(\mathbf{A})\,, \tag{54}$$

which is widely referred to as the Bhattacharrya bound. Relevant to this study is that the squared Hellinger distance is related to the Bhattacharyya coefficient in (53) through

$$\mathsf{H}^2(\mathbf{A}) = 2 - \mathsf{BC}(\mathbf{A})\,. \tag{55}$$

Hence, maximizing the Hellinger distance $\mathsf{H}^2(\mathbf{A})$ results in a tighter bound on P_e from (54). To show the performance numerically, we compute the $\mathsf{BC}(\mathbf{A})$ via (55). For the pre- and post-linear dimensionality reduction structures as given in (52), the covariance matrix $\mathbf{\Sigma}$ is randomly generated following the setup used in Section 5.1. As plotted in Figure 3, by employing the design scheme according to Theorem 4, the optimal design results in a smaller $\mathsf{BC}(\mathbf{A})$ and, hence, a tighter upper bound on P_e in comparison to other schemes.

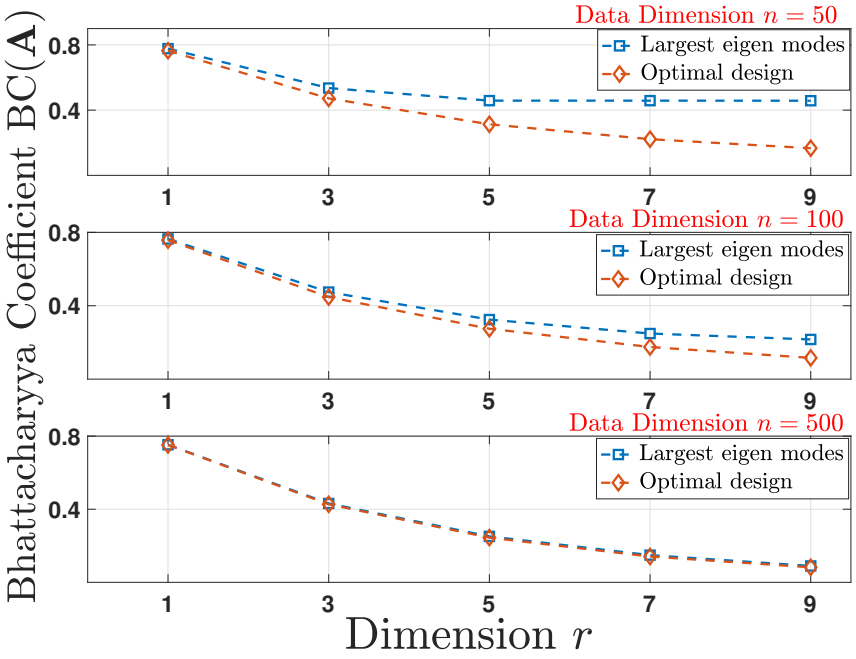

Figure 3. Comparison of the empirical average of the Bhattacharyya coefficient $\mathsf{BC}(\mathbf{A})$ under optimal design and largest eigen modes schemes for multiple reduced data dimensions r as a function of original data dimension n.

5.4. Total Variation Distance

Consider a binary hypothesis test with Gaussian class conditional densities following the model in (52) and equal class a priori probabilities, i.e., $p_{\mathbb{P}_{\mathbf{A}}} = p_{\mathbb{Q}_{\mathbf{A}}}$. We define c_{ij} as the cost associated with deciding in favor of H_i when the true hypothesis is H_j such that $0 \leq i, j \leq 1$, and denote the densities associated with measures $\mathbb{P}_{\mathbf{A}}$, $\mathbb{Q}_{\mathbf{A}}$ by $f_{\mathbb{P}_{\mathbf{A}}}$ and $f_{\mathbb{Q}_{\mathbf{A}}}$, respectively. Without loss of generality, we assume a 0–1 loss function such that $c_{ij} = 1 \; \forall \; i \neq j$ and $c_{ii} = 0 \; \forall \; i$. The optimal Bayes decision rule that minimizes the error probability is given by

$$\frac{f_{\mathbb{P}_{\mathbf{A}}}(x)}{f_{\mathbb{Q}_{\mathbf{A}}}(x)} \underset{d=\mathsf{H}_0}{\overset{d=\mathsf{H}_1}{\lessgtr}} 1 \,. \tag{56}$$

Since the total variation distance cannot be computed in closed-form, we numerically compute the error probability P_e under the two bounds (Hellinger-based and FB-based) introduced in Section 4.4.2 to quantify the performance of the design of matrix $\mathbf{A}$ for the underlying inference problem. The covariance matrix $\mathbf{\Sigma}$ is randomly generated following the setup used in Section 5.1. As plotted in Figure 4, by optimizing the Hellinger-based bound according to Theorem 4 and optimizing the FB-based bound according to Theorem 3, the two design schemes achieve a smaller P_e. We further observe that the bounds due to FB-based are loose in comparison to Hellinger-based bounds. Therefore, we choose not to plot the lower bound on P_e for the FB-based bounds in Figure 4.

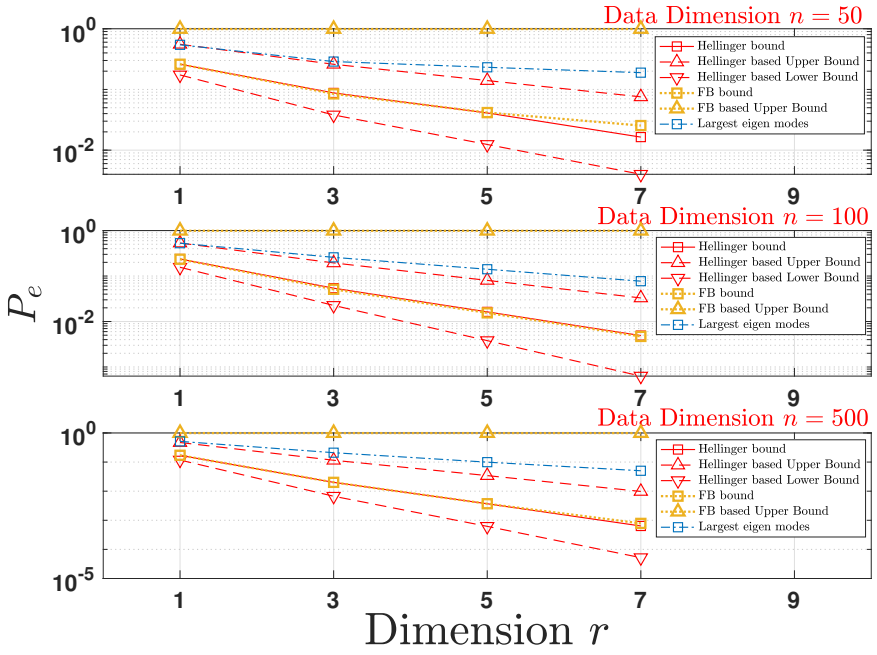

Figure 4. Comparing the logarithm of the empirical average value for P_e under the two bounds on $d_{\mathsf{TV}}(\mathbf{A})$ (Hellinger-based and Frobenius norm (FB)-based) with the largest eigen modes scheme for multiple projected data dimensions r as a function of initial data dimension n.

5.5. χ^2-Divergence

In this section, we show the gains of the proposed analysis through numerical evaluations by numerically computing $\chi^2(\mathbf{A})$, to find a lower bound on the noise variance $\mathsf{var}_\theta(\hat{\theta})$ up to a constant. Following the pre- and post-linear dimensionality reduction structures given in (52), the covariance matrix $\mathbf{\Sigma}$ is randomly generated following the setup used in Section 5.1. As shown in Figure 5, constructing the optimal design according to Theorem 4 achieves a tighter lower bound in comparison to the other scheme.

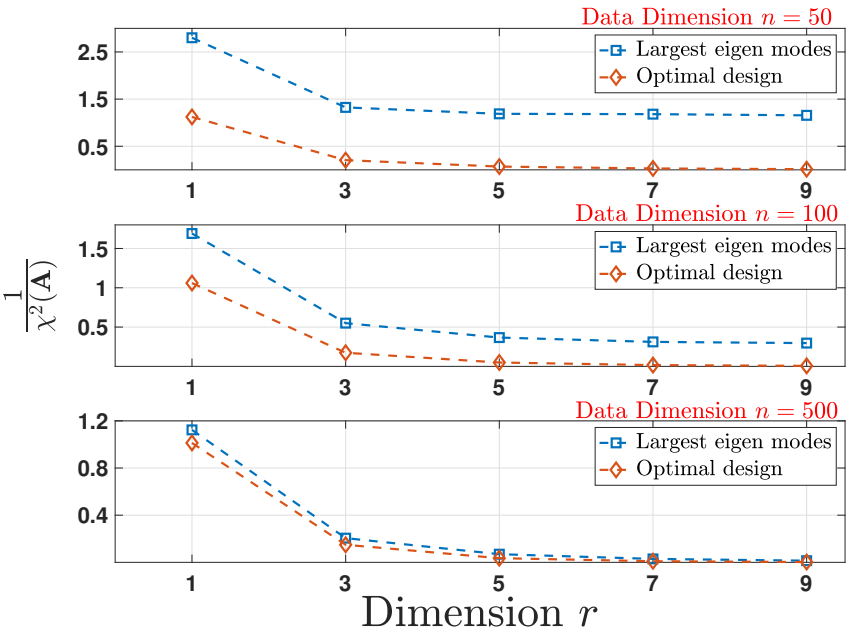

Figure 5. Comparison of the lower bound on noise variance given by $\frac{1}{\chi^2(\mathbf{A})}$ under the optimal and largest eigen modes schemes for multiple reduced data dimensions r as a function of original data dimension n.

6. General Gaussian Models

In the previous section, we focused on $\boldsymbol{\mu} = 0$. When $\boldsymbol{\mu} \neq 0$, optimizing each f-divergence measure under the semi-orthogonality constraint does not render closed-form expressions. Nevertheless, to provide some intuitions, we provide a numerical approach to the optimal design of $\mathbf{A}$, which might also enjoy some *local* optimality guarantees. To start, note that the feasible set of solutions given by $\mathcal{M}_n^r \triangleq \{\mathbf{A} \in \mathbb{R}^{r\times n} : \mathbf{A}\cdot\mathbf{A}^\top = \mathbf{I}_r\}$ owing to the orthogonality constraints in $\mathcal{Q}$ is often referred to as the Stiefel manifold. Therefore, solving $\mathcal{Q}$ requires designing algorithms that optimize the objective while preserving manifold constraints during iterations.

We employ the method of Lagrange multipliers to formulate the Lagrangian function. By denoting the matrix of Lagrangian multipliers by $\mathbf{L} \in \mathbb{R}^{r\times r}$, the Lagrangian function of problem (14) is given by

$$\mathcal{L}(\mathbf{A}, \mathbf{L}) = D_f(\mathbf{A}) + \langle \mathbf{L}, \mathbf{A}\cdot\mathbf{A}^\top - \mathbf{I}_r \rangle \,. \tag{57}$$

From the first order optimality condition, for any local maximizer $\mathbf{A}^*$ of (14), there exists a Lagrange multiplier $\mathbf{L}^*$ such that

$$\nabla_{\mathbf{A}} \mathcal{L}(\mathbf{A}, \mathbf{L})\Big|_{\mathbf{A}^*, \mathbf{L}^*} = 0 \,, \tag{58}$$

where we denote the partial derivative with respect to $\mathbf{A}$ by $\nabla_{\mathbf{A}}$. In what follows, we iterate the design mapping $\mathbf{A}$ using the gradient ascent algorithm in order to find a solution for $\mathbf{A}$. As discussed in the next subsection, this solution is guaranteed to be at least locally optimal.

6.1. Optimizing via Gradient Ascent

We use an iterative gradient ascent-based algorithm to find the local maximizer of $D_f(\mathbf{A})$ such that $\mathbf{A} \in \mathcal{M}_n^r$. The gradient ascent update at any given iteration $k \in \mathbb{N}$ is given by

$$\mathbf{A}^{k+1} = \mathbf{A}^k + \alpha \cdot \nabla_{\mathbf{A}} \mathcal{L}(\mathbf{A}, \mathbf{L})\Big|_{\mathbf{A}^k} \,. \tag{59}$$

Note that, following this update, since the new point $\mathbf{A}^{k+1}$ in (59) may not satisfy the semi-orthogonality, i.e., $\mathbf{A}^{k+1} \notin \mathcal{M}_n^r$, it is imperative to establish a relation between the multipliers $\mathbf{L}$ and $\mathbf{A}^k$ in every iteration k to ensure a constraint-preserving update scheme. In particular, to enforce the semi-orthogonality constraint on $\mathbf{A}^{k+1}$, a relationship between the multipliers and the gradients in every iteration k is derived. Following a similar line of analysis for gradient descent in Reference [43], the relationship between multipliers and the gradients is provided in Appendix E. More details on the analysis of the update scheme can be found in Reference [43], and a detailed discussion on the convergence guarantees of classical steepest descent update schemes adapted to semi-orthogonality constraints can be found in Reference [44].

In order to simplify $\nabla_{\mathbf{A}}\mathcal{L}(\mathbf{A}, \mathbf{L})$ and state the relationships, we define $\mathbf{\Lambda} \triangleq \mathbf{L} + \mathbf{L}^\top$ and subsequently find a relationship between $\mathbf{\Lambda}$ and $\mathbf{A}^k$ in every iteration k. This is obtained by right-multiplying (59) by $\mathbf{A}^{k+1}$ and solving for $\mathbf{\Lambda}$ that enforces the semi-orthogonality constraint on $\mathbf{A}^{k+1}$. To simplify the analysis, we take a finite Taylor series expansion of $\mathbf{\Lambda}$ around $\alpha = 0$ and choose α such that the error in forcing the constraint is a good approximation of the gradient of the objective subjected to $\mathbf{A} \cdot \mathbf{A}^\top = \mathbf{I}_r$. As derived in the Appendix E, by simple algebraic manipulations, it can be shown that the matrices $\mathbf{\Lambda}_0, \mathbf{\Lambda}_1$, and $\mathbf{\Lambda}_2$, for which the finite Taylor series expansion of $\mathbf{\Lambda} \approx \mathbf{\Lambda}_0 + \alpha \cdot \mathbf{\Lambda}_1 + \alpha^2 \cdot \mathbf{\Lambda}_2$ is a good approximation of the constraint, are given by

$$\mathbf{\Lambda}_0 \triangleq -\frac{1}{2}\left[\nabla_{\mathbf{A}} D_f(\mathbf{A}) \cdot (\mathbf{A})^\top + \mathbf{A} \cdot \nabla_{\mathbf{A}} D_f(\mathbf{A})^\top\right], \tag{60}$$

$$\mathbf{\Lambda}_1 \triangleq -\frac{1}{2}\left[\left(\nabla_{\mathbf{A}} D_f(\mathbf{A}) + \mathbf{\Lambda}_0\mathbf{A}\right) \cdot \left(\nabla_{\mathbf{A}} D_f(\mathbf{A}) + \mathbf{\Lambda}_0\mathbf{A}\right)^\top\right], \tag{61}$$

$$\mathbf{\Lambda}_2 \triangleq -\frac{1}{2}\left[\mathbf{\Lambda}_1 \cdot \mathbf{A} \cdot \nabla_{\mathbf{A}} D_f(\mathbf{A})^\top + \nabla_{\mathbf{A}} D_f(\mathbf{A}) \cdot (\mathbf{A})^\top \cdot \mathbf{\Lambda}_1 + \mathbf{\Lambda}_0 \cdot \mathbf{\Lambda}_1 + \mathbf{\Lambda}_1 \cdot \mathbf{\Lambda}_0\right]. \tag{62}$$

Additionally, we note that, since finding the global maximum is not guaranteed, it is imperative to initialize $\mathbf{A}^0$ close to the estimated maximum. In this regard, we leverage the structure of the objective function for each f-divergence measure as given in Appendix D. In particular, we observe that the objective of each f-divergence measure can be decomposed into two objectives: the first not involving $\boldsymbol{\mu}$ (making this objective a convex problem as shown in Section 4), and the second objective a function of $\boldsymbol{\mu}$. Hence, leveraging the structure of the solution from Section 4, we initialize $\mathbf{A}^0$ such that it maximizes the objective in the case of zero-mean Gaussian models. We further note that, while there are more sophisticated orthogonality constraint-preserving algorithms [45], we find that our method adopted from Reference [43] is sufficient for our purpose, as we show next through numerical simulations.

6.2. Results and Discussion

The design of $\mathbf{A}$ when $\boldsymbol{\mu} \neq 0$ is not characterized analytically. Therefore, we resort to numerical simulations to show the gains of optimizing f-divergence measures when $\boldsymbol{\mu} \neq 0$. In particular, we consider the linear discriminant analysis (LDA) problem where the goal is to design a mapping $\mathbf{A}$ and perform classification in the lower dimensional space (of dimension r). Without loss of generality, we assume $n = 10$ and consider Gaussian densities with the following pre- and post-linear dimensionality reduction structures:

$$\begin{array}{llll} \mathbb{P}: & \mathcal{N}(\mathbf{0}, \mathbf{I}_n) & \text{and} & \mathbb{Q}: \ \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma}) \\ \mathbb{P}_{\mathbf{A}}: & \mathcal{N}(\mathbf{0}, \mathbf{I}_r) & \text{and} & \mathbb{Q}_{\mathbf{A}}: \ \mathcal{N}(\mathbf{A} \cdot \boldsymbol{\mu}, h_1(\mathbf{A})) \end{array}, \tag{63}$$

where the covariance matrix $\mathbf{\Sigma}$ is generated randomly the eigenvalues of which are sampled from a uniform distribution $\{\lambda_i \sim \mathcal{U}(0,1)\}_{i=1}^{10}$. For the model in (63), we consider two kinds of performance metrics that have information-theoretic performance interpretations: (i) the total probability of error related to the $d_{\mathsf{TV}}(\mathbf{A})$, and (ii) the exponential decay of error probability related to $D_{\mathsf{KL}}(\mathbb{P}_{\mathbf{A}} \parallel \mathbb{Q}_{\mathbf{A}})$. In what follows, we demonstrate that optimizing

appropriate f-divergence measures between $\mathbb{P}_{\mathbf{A}}$ and $\mathbb{Q}_{\mathbf{A}}$ lead to better performance when compared to the performance of the popular Fisher's quadratic discriminant analysis (QDA) classifier [20]. In particular, the Fisher's approach sets $r = 1$ and designs $\mathbf{A}$ by solving

$$\underset{\mathbf{A}\in\mathbb{R}^{1\times n}}{\arg\max} \quad \frac{(\boldsymbol{\mu}\cdot\mathbf{A}^{\top})^2}{\mathbf{A}\cdot(\mathbf{I}_n+\boldsymbol{\Sigma})\cdot\mathbf{A}^{\top}}\,. \tag{64}$$

In contrast, we design $\mathbf{A}$ such that the information-theoretic objective functions associated with the total probability of error (captured by $d_{\mathsf{TV}}(\mathbf{A})$) and the exponential decay of error probability (captured by $D_{\mathsf{KL}}(\mathbb{P}_{\mathbf{A}} \parallel \mathbb{Q}_{\mathbf{A}})$) are minimized. The structure of the objective functions is discussed in Total probability of error and Type-II error subjected to type-I error constraints. Both methods and Fisher's method, after projecting the data into a lower dimension, deploy optimal detectors to discern the true model. It is noteworthy that, in both methods the data in the lower dimensions has a Gaussian model, and the conventional QDA [20] classifier is the optimal detector. Hence, we emphasize that our approach aims to have a design for $\mathbf{A}$ that maximizes the distance between the probability measures after reducing the dimensions, i.e., the distance between $\mathbb{P}_{\mathbf{A}}$ and $\mathbb{Q}_{\mathbf{A}}$. Since this distance captures the quality of the decisions, our design of $\mathbf{A}$ outperforms that of Fisher's. For each comparison, we consider various values for $\boldsymbol{\mu}$ and compare the appropriate performance metrics with that of Fisher's QDA for each. In all cases, the data is synthetically generated, i.e., sampled from a Gaussian distribution where we consider 2000 data points associated with each measure $\mathbb{P}$ and $\mathbb{Q}$.

6.2.1. Schemes for Linear Map

(1) Total Probability of Error: In this scheme, the linear map $\mathbf{A}$ is designed such that $d_{\mathsf{TV}}(\mathbf{A})$ is optimized via gradient ascent iterations until convergence. As discussed in Section 4.4.1, since the total probability of error is the key performance metric that arises while optimizing $d_{\mathsf{TV}}(\mathbf{A})$, it is expected that optimizing $d_{\mathsf{TV}}(\mathbf{A})$ will result in a smaller total error in comparison to other schemes that optimize other objective functions (e.g., Fisher's QDA). We note that, since there do not exist closed-form expressions for the total variation distance, we maximize bounds on $d_{\mathsf{TV}}(\mathbf{A})$ instead via the Hellinger bound in (33) as a proxy to minimize the total probability of error. The corresponding gradient expression to optimize $\mathsf{H}^2(\mathbf{A})$ (to perform iterative updates as in (59)) is derived in closed-form and is given in Appendix D.

(2) Type-II Error Subjected to Type-I Error Constraints: In this scheme, the linear map $\mathbf{A}$ is designed such that $D_{\mathsf{KL}}(\mathbb{P}_{\mathbf{A}} \parallel \mathbb{Q}_{\mathbf{A}})$ is optimized via gradient ascent iterations until convergence. In order to establish a relation, consider the following binary hypothesis test:

$$\mathsf{H}_0 \,: X \sim \mathbb{P}_{\mathbf{A}} \quad \text{versus} \quad \mathsf{H}_1 \,: X \sim \mathbb{Q}_{\mathbf{A}}\,. \tag{65}$$

When minimizing the probability of type-II error subjected to type-I error constraints, the optimal test guarantees that the probability of type-II error decays exponentially as

$$\lim_{s\to\infty} \frac{-\log(\mathbb{Q}_{\mathbf{A}}(d=\mathsf{H}_0))}{s} = D_{\mathsf{KL}}(\mathbb{P}_{\mathbf{A}} \parallel \mathbb{Q}_{\mathbf{A}})\,, \tag{66}$$

where we have define $d : X \to \{\mathsf{H}_0, \mathsf{H}_1\}$ as the decision rule for the hypothesis test, and s denotes the sample size. As a result, $D_{\mathsf{KL}}(\mathbb{P}_{\mathbf{A}} \parallel \mathbb{Q}_{\mathbf{A}})$ appears as the error exponent for hypothesis test in (65). Hence, it is expected that optimizing $D_{\mathsf{KL}}(\mathbb{P}_{\mathbf{A}} \parallel \mathbb{Q}_{\mathbf{A}})$ will result in a smaller type-II error for the same type-I error when comparing with a method that optimizes other objectives (e.g., Fisher's QDA). The corresponding gradient expression to optimize the $D_{\mathsf{KL}}(\mathbb{P}_{\mathbf{A}} \parallel \mathbb{Q}_{\mathbf{A}})$ is derived in closed-form and is given in Appendix D.

For the sake of comparison and reference, we also consider schemes in which $\mathbf{A}$ is designed to optimize the objectives $D_{\mathsf{KL}}(\mathbf{A})$, the largest eigen modes (LEM), and the smallest eigen modes (SEM), which carry no specific operational significance in the context

of the binary classification problem. In the case of LEM and SEM schemes, the linear map $\mathbf{A}$ is designed such that the rows of $\mathbf{A}$ are the eigenvector associated with the largest and smallest modes of the matrix $\mathbf{\Sigma}$, respectively. Furthermore, we define $\mathbb{1}$ as the vector of all those of appropriate dimension.

6.2.2. Performance Comparison

After learning the linear map $\mathbf{A}$ for each scheme described in Section 6.2.1, we perform classification in the lower dimensional space of dimension r to find the type-I, type-II, and total probability of error for each scheme. Tables 1–4 tabulate the results for various choices of the mean parameter $\boldsymbol{\mu}$. We have the following important observations: (i) we observe that optimizing $\mathsf{H}^2(\mathbf{A})$ results in a smaller total probability of error in comparison to the total error obtained by optimizing the Fisher's objective; it is important to note that the superior performance is observed despite maximizing bounds on $d_{\mathsf{TV}}(\mathbf{A})$ (that is sub-optimal) and not the distance itself; and (ii) we observe that except for the case of $\boldsymbol{\mu} = 0.8 \cdot \mathbb{1}$, optimizing $D_{\mathsf{KL}}(\mathbb{P}_{\mathbf{A}} \parallel \mathbb{Q}_{\mathbf{A}})$ results in a smaller type-II error in comparison to that obtained by optimizing the Fisher's objective indicating a gain in optimizing $D_{\mathsf{KL}}(\mathbb{P}_{\mathbf{A}} \parallel \mathbb{Q}_{\mathbf{A}})$ in comparison to the Fisher's objective in (64).

Table 1. $\boldsymbol{\mu} = 0.2 \cdot \mathbb{1}, r = 1$.

	Fisher's QDA	$D_{\mathsf{KL}}(\mathbb{P}_{\mathbf{A}} \parallel \mathbb{Q}_{\mathbf{A}})$	$\mathsf{H}^2(\mathbf{A})$	$D_{\mathsf{KL}}(\mathbf{A})$	**SEM**	**LEM**
$\mathbb{P}_{\mathbf{A}}(d = \mathsf{H}_1)$	331/2000	331/2000	331/2000	331/2000	337/2000	915/2000
$\mathbb{Q}_{\mathbf{A}}(d = \mathsf{H}_0)$	1226/2000	63/2000	63/2000	63/2000	64/2000	811/2000
Total Error	1557/4000	394/4000	394/4000	394/4000	401/4000	1726/4000

Table 2. $\boldsymbol{\mu} = 0.4 \cdot \mathbb{1}, r = 1$.

	Fisher's QDA	$D_{\mathsf{KL}}(\mathbb{P}_{\mathbf{A}} \parallel \mathbb{Q}_{\mathbf{A}})$	$\mathsf{H}^2(\mathbf{A})$	$D_{\mathsf{KL}}(\mathbf{A})$	**SEM**	**LEM**
$\mathbb{P}_{\mathbf{A}}(d = \mathsf{H}_1)$	344/2000	344/2000	344/2000	345/2000	347/2000	782/2000
$\mathbb{Q}_{\mathbf{A}}(d = \mathsf{H}_0)$	594/2000	63/2000	63/2000	63/2000	64/2000	739/2000
Total Error	938/4000	407/4000	407/4000	408/4000	411/4000	1521/4000

Table 3. $\boldsymbol{\mu} = 0.6 \cdot \mathbb{1}, r = 1$.

	Fisher's QDA	$D_{\mathsf{KL}}(\mathbb{P}_{\mathbf{A}} \parallel \mathbb{Q}_{\mathbf{A}})$	$\mathsf{H}^2(\mathbf{A})$	$D_{\mathsf{KL}}(\mathbf{A})$	**SEM**	**LEM**
$\mathbb{P}_{\mathbf{A}}(d = \mathsf{H}_1)$	326/2000	326/2000	335/2000	318/2000	335/2000	638/2000
$\mathbb{Q}_{\mathbf{A}}(d = \mathsf{H}_0)$	137/2000	55/2000	108/2000	57/2000	61/2000	669/2000
Total Error	463/4000	381/4000	443/4000	375/4000	396/4000	1307/4000

Table 4. $\boldsymbol{\mu} = 0.8 \cdot \mathbb{1}, r = 1$.

	Fisher's QDA	$D_{\mathsf{KL}}(\mathbb{P}_{\mathbf{A}} \parallel \mathbb{Q}_{\mathbf{A}})$	$\mathsf{H}^2(\mathbf{A})$	$D_{\mathsf{KL}}(\mathbf{A})$	**SEM**	**LEM**
$\mathbb{P}_{\mathbf{A}}(d = \mathsf{H}_1)$	264/2000	264/2000	159/2000	255/2000	307/2000	561/2000
$\mathbb{Q}_{\mathbf{A}}(d = \mathsf{H}_0)$	25/2000	53/2000	64/2000	55/2000	60/2000	580/2000
Total Error	289/4000	317/4000	214/4000	310/4000	367/4000	1141/4000

It is important to note that the convergence of the gradient ascent algorithm only guarantees a locally optimal solution. While we have restricted the results that consider a maximum separation of $\boldsymbol{\mu} = 0.8 \cdot \mathbb{1}$, we have performed additional simulations for larger separation between models (greater $\mu > 0.8$). We have the following observations: (i) solution for the linear map $\mathbf{A}$ obtained through gradient ascent becomes highly sensitive to the initialization $\mathbf{A}^0$; specifically, it was observed that optimizing the Fisher's objective outperforms optimizing $\mathsf{H}^2(\mathbf{A})$ for various initializations $\mathbf{A}^0$, and vice versa, for other random initializations; and (ii) the gradient ascent solver becomes more prone to getting stuck at the local maxima for larger separations between the models. We conjecture that the odd observation in the case of $\boldsymbol{\mu} = 0.8 \cdot \mathbb{1}$ when optimizing $D_{\mathsf{KL}}(\mathbb{P}_{\mathbf{A}} \parallel \mathbb{Q}_{\mathbf{A}})$ (where optimizing the Fisher's objective outperforms optimizing $D_{\mathsf{KL}}(\mathbb{P}_{\mathbf{A}} \parallel \mathbb{Q}_{\mathbf{A}})$) supports this

observation. Furthermore, we note that, since the problem is convex for $\boldsymbol{\mu} = \mathbf{0}$, a deviation from this assumption moves the problem further from being convex, making the solver prone to getting stuck at the locally optimal solutions for larger separation between the Gaussian models.

6.2.3. Subspace Representation

In order to gain more intuition towards the learned representations, we illustrate the 2-dimensional projections of the original 10-dimensional data obtained after optimizing the corresponding f-divergence measures. For brevity, we only show the plots for $D_{\mathsf{KL}}(\mathbb{P}_{\mathbf{A}} \parallel \mathbb{Q}_{\mathbf{A}})$ and $\mathsf{H}^2(\mathbf{A})$. Figures 6 and 7 plot the two-dimensional projections of the synthetic dataset that optimize $D_{\mathsf{KL}}(\mathbb{P}_{\mathbf{A}} \parallel \mathbb{Q}_{\mathbf{A}})$ and $\mathsf{H}^2(\mathbf{A})$, respectively. As expected, it is observed that the total probability of error is smaller when optimizing $\mathsf{H}^2(\mathbf{A})$. Figure 8 shows the variation in the objective function as a function of gradient ascent iterations. As the iterations grow, the objective functions eventually converges to a locally optimal solution.

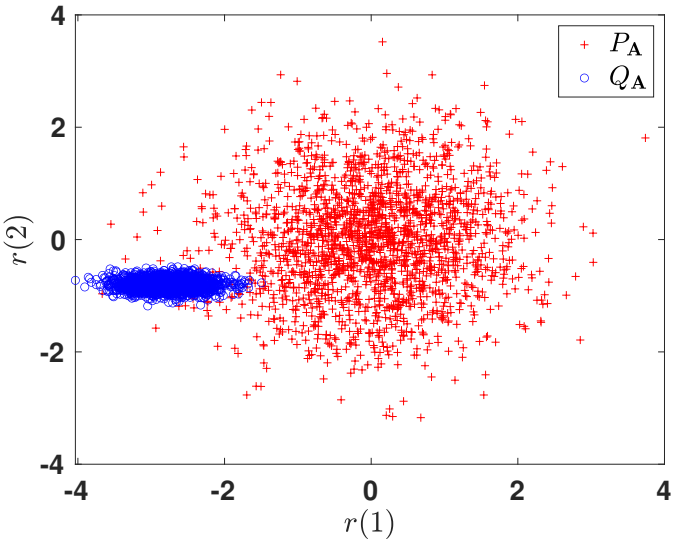

Figure 6. Two-dimensional projected data obtained by optimizing $D_{\mathsf{KL}}(\mathbb{P}_{\mathbf{A}} \parallel \mathbb{Q}_{\mathbf{A}})$.

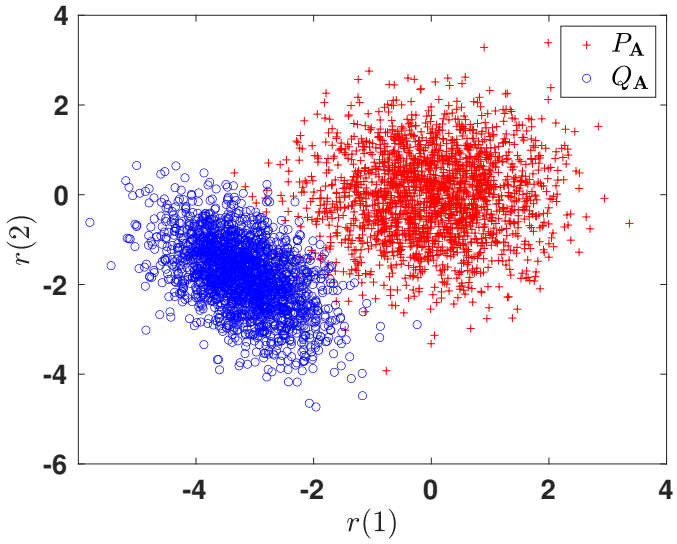

Figure 7. Two-dimensional projected data obtained by optimizing $\mathsf{H}^2(\mathbf{A})$.

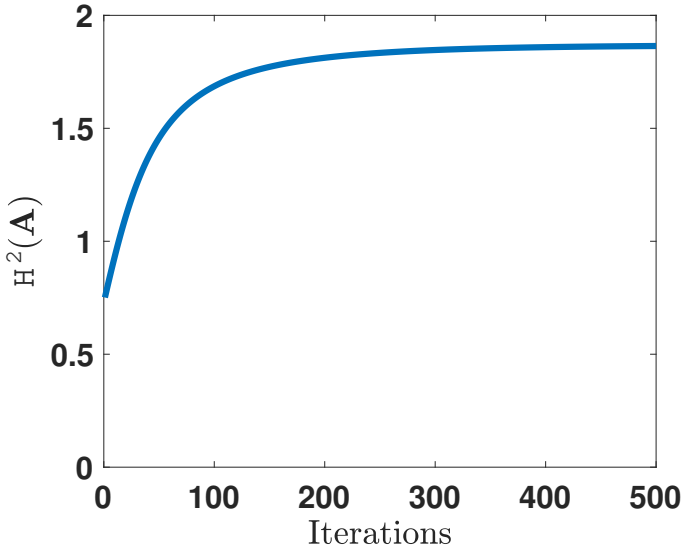

Figure 8. Convergence of the gradient ascent algorithm as a result of optimizing $\mathsf{H}^2(\mathbf{A})$.

7. Conclusions

In this paper, we have considered the problem of discriminant analysis such that separation between the classes is maximized under f-divergence measures. This approach is motivated by dimensionality reduction for inference problems, where we have investigated discriminant analysis under Kullback–Leibler, symmetrized Kullback–Leibler, Hellinger, χ^2, and total variation measures. We have characterized the optimal design for the linear transformation of the data onto a lower-dimensional subspace for each in the case of zero-mean Gaussian models and adopted numerical algorithms to find the design of the linear transformation in the case of general Gaussian models with non-zero means. We have shown that, in the case of zero-mean Gaussian models, the row space of the mapping matrix lies in the eigenspace of a matrix associated with the covariance matrix of the Gaussian models involved. While each f-divergence measure favors specific eigenvector components, we have shown that all the designs become identical in certain regimes, making the design of the linear mapping independent of the inference problem of interest.

Author Contributions: A.D., S.W. and A.T. contributed equally. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by the U. S. National Science Foundation under grants CAREER Award ECCS-1554482 and ECCS-1933107, and RPI-IBM Artificial Intelligence Research Collaboration (AIRC).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

PCA	Principal Component Analysis
MDS	Multidimensional Scaling
SDR	Sufficient Dimension Reduction
DA	Discriminant Analysis
KL	Kullback Leibler
TV	Total Variation
ADD	Average Detection Delay

FAR	False Alarm Rate
CuSum	Cumulative Sum
BC	Bhattacharyya Coefficient
LEM	Largest Eigen Modes
SEM	Smallest Eigen Modes
LDA	Linear Discriminant Analysis
QDA	Quadratic Discriminant Analysis

Appendix A. Proof of Theorem 1

Consider two pairs of probability measures $(\mathbb{P}_{\mathbf{A}}, \mathbb{Q}_{\mathbf{A}})$ and $(\mathbb{P}_{\bar{\mathbf{A}}}, \mathbb{Q}_{\bar{\mathbf{A}}})$ associated with the mapping $\mathbf{A}$ in space $\mathcal{X}$ and $\bar{\mathbf{A}}$ in space $\mathcal{Y}$, respectively. Let $g : \mathcal{X} \to \mathcal{Y}$ denote any invertible transformation. Under the invertible map, we have

$$\mathrm{d}\mathbb{Q}_{\bar{\mathbf{A}}} = \mathrm{d}\mathbb{Q}_{\mathbf{A}} \cdot |\mathcal{T}|^{-1}, \quad \text{and} \quad \mathrm{d}\mathbb{P}_{\bar{\mathbf{A}}} = \mathrm{d}\mathbb{P}_{\mathbf{A}} \cdot |\mathcal{T}|^{-1}, \tag{A1}$$

where $|\mathcal{T}|$ denotes the determinant of the Jacobian matrix associated with g. Leveraging (A1), the f-divergence measure $D_f(\bar{\mathbf{A}})$ simplifies as follows.

$$\begin{aligned}
D_f(\bar{\mathbf{A}}) &\stackrel{\triangle}{=} \mathbb{E}_{\mathbb{P}_{\bar{\mathbf{A}}}}\left[f\left(\frac{\mathrm{d}\mathbb{Q}_{\bar{\mathbf{A}}}}{\mathrm{d}\mathbb{P}_{\bar{\mathbf{A}}}}\right)\right] && \text{(A2)}\\
&= \int_{\mathcal{Y}} f\left(\frac{\mathrm{d}\mathbb{Q}_{\bar{\mathbf{A}}}}{\mathrm{d}\mathbb{P}_{\bar{\mathbf{A}}}}\right) \mathrm{d}\mathbb{P}_{\bar{\mathbf{A}}}(y) && \text{(A3)}\\
&= \int_{\mathcal{X}} |\mathcal{T}(x)|^{-1} \cdot f\left(\frac{\mathrm{d}\mathbb{Q}_{\mathbf{A}} \cdot |\mathcal{T}(x)|^{-1}}{\mathrm{d}\mathbb{P}_{\mathbf{A}} \cdot |\mathcal{T}(x)|^{-1}}\right) \cdot |\mathcal{T}(x)| \, \mathrm{d}\mathbb{P}_{\mathbf{A}}(x) && \text{(A4)}\\
&= \int_{\mathcal{X}} f\left(\frac{\mathrm{d}\mathbb{Q}_{\mathbf{A}}}{\mathrm{d}\mathbb{P}_{\mathbf{A}}}\right) \mathrm{d}\mathbb{P}_{\mathbf{A}}(x) && \text{(A5)}\\
&= D_f(\mathbf{A}) \, . && \text{(A6)}
\end{aligned}$$

Therefore, f-divergence measures are invariant under invertible transformations (both linear and non-linear) ensuring the existence of $\bar{\mathbf{A}}$ for every $\mathbf{A}$ as a special case for linear transformations.

Appendix B. Proof of Theorem 3

We observe that $D_{\mathsf{KL}}(\mathbf{A})$, $D_{\mathsf{SKL}}(\mathbf{A})$, and the objective to be optimized through the matching bound Section 4.4.2, Matching Bounds up to a Constant on $d_{\mathsf{TV}}(\mathbf{A})$ can be decomposed as the summation of strictly convex functions involving $g_{\mathsf{KL}}(x)$, $g_{\mathsf{SKL}}(x)$, and $g_{\mathsf{TV}}(x)$, respectively. Since the summation of strictly convex functions is strictly convex, we conclude that each objective $D_f \in \{D_{\mathsf{KL}}(\mathbf{A}), D_{\mathsf{SKL}}(\mathbf{A}), d_{\mathsf{TV}}(\mathbf{A})\}$ is strictly convex.

Next, the goal is to choose $\{\gamma_i\}_{i=1}^r$ such that $D_f \in \{D_{\mathsf{KL}}(\mathbf{A}), D_{\mathsf{SKL}}(\mathbf{A}), d_{\mathsf{TV}}(\mathbf{A})\}$ is maximized subjected to spectral constraints given by $\lambda_{n-(r-i)} \leq \gamma_i \leq \lambda_i$. In order to choose appropriate γ_i's, we first note that the global minimizer for functions $g_f \in \{g_{\mathsf{KL}}, g_{\mathsf{SKL}}, g_{\mathsf{TV}}\}$ appears at $x = 1$. By noting that each g_f is strictly convex, it can be readily verified that $g_f(x)$ is monotonically increasing for $x > 1$ and monotonically decreasing for $x < 1$. This will guide selecting $\{\gamma_i\}_{i=1}^r$, as explained next.

In the case of $\lambda_n \geq 1$, i.e., when all the eigenvalues are larger than or equal to 1, the objective of maximizing each $D_f \in \{D_{\mathsf{KL}}(\mathbf{A}), D_{\mathsf{SKL}}(\mathbf{A}), d_{\mathsf{TV}}(\mathbf{A})\}$ boils down to maximizing a monotonically increasing function (considering the domain). This is trivially done by choosing $\gamma_i = \lambda_i$ for $i \in [r]$, proving Corollary 1. On the other hand, when $\lambda_1 \leq 1$, i.e., when all the eigenvalues are smaller than or equal to 1, following the same line of argument, the objective boils down to maximizing each $D_f \in \{D_{\mathsf{KL}}(\mathbf{A}), D_{\mathsf{SKL}}(\mathbf{A}), d_{\mathsf{TV}}(\mathbf{A})\}$, where each D_f is a monotonically decreasing function (considering the domain). This is trivially done by choosing $\gamma_i = \lambda_{n-r+i}$ for $i \in [r]$.

When $\lambda_n \leq 1 \leq \lambda_1$, the selection process is not trivial. Rather, an iterative algorithm can be followed, where we start from the eigenvalues farthest away from 1 on both sides

and, subsequently, choose the one in every iteration that achieves a higher objective. This procedure can be repeated recursively until r eigenvalues are chosen. This procedure is also discussed in Algorithm 1 in Section 4.6.

Finally, constructing the optimal matrix $\mathbf{A}$, which maximizes D_f for any data matrix $\mathbf{\Sigma}$, becomes equivalent to choosing eigenvectors as the rows of $\mathbf{A}$ associated with the chosen permutation of eigenvalues for each of the aforementioned cases.

Appendix C. Proof for Theorem 4

We first find a closed-form expression for $\chi^2(\mathbf{A})$ and $\chi^2(\mathbb{P}_\mathbf{A} \parallel \mathbb{Q}_\mathbf{A})$. From the definition, we have

$$\chi^2(\mathbf{A}) \stackrel{\triangle}{=} \frac{|\mathbf{I}_r|^{\frac{1}{2}}}{(2\pi)^{\frac{r}{2}} \cdot |h_1(\mathbf{A})|} \cdot \int_{\mathbb{R}^r} \exp\left[-\frac{1}{2} \cdot \left(Y^\top \cdot \mathbf{K}_1 \cdot Y\right)\right] \mathrm{d}Y - 1\,, \tag{A7}$$

where we defined $\mathbf{K}_1 \stackrel{\triangle}{=} 2 \cdot h_1(\mathbf{A})^{-1} - \mathbf{I}_r$. We note that $\mathbf{K}_1$ is a real symmetric matrix since $h_1(\mathbf{A})$ is a real symmetric matrix. We denote the eigen decomposition of $\mathbf{K}_1$ as $\mathbf{K}_1 = \mathbf{U} \cdot \mathbf{\Theta} \cdot \mathbf{U}^\top$, where the matrix $\mathbf{\Theta}$ is a diagonal matrix with the eigenvalues $\{\theta_i\}_{i=1}^r$ as its elements. Based on this decomposition, we have

$$\chi^2(\mathbf{A}) = \frac{1}{(2\pi)^{\frac{r}{2}} \cdot |h_1(\mathbf{A})|} \cdot \int_{\mathbb{R}^r} \exp\left[-\frac{1}{2}\left(Y^\top \cdot \mathbf{U}\mathbf{\Theta}\mathbf{U}^\top \cdot Y\right)\right] \mathrm{d}Y - 1 \tag{A8}$$

$$= \frac{1}{(2\pi)^{\frac{r}{2}} \cdot |h_1(\mathbf{A})|} \cdot \int_{\mathbb{R}^r} \exp\left[-\frac{1}{2}\left(W^\top \cdot \mathbf{\Theta} \cdot W\right)\right] \mathrm{d}W - 1 \tag{A9}$$

$$= \frac{1}{(2\pi)^{\frac{r}{2}} \cdot |h_1(\mathbf{A})|} \cdot \prod_{i=1}^{r} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}\left(\theta_i \cdot w_i^2\right)\right] \mathrm{d}w_i - 1\,, \tag{A10}$$

where we have defined $W \stackrel{\triangle}{=} \mathbf{U}^\top \cdot Y$. We note that, in order for $\chi^2(\mathbf{A})$ to be finite, it is required that the eigenvalues $\{\theta_i\}_{i=1}^r$ be non-negative. Hence, based on the definition of $\mathbf{K}_1$, all the eigenvalues λ_i should fall in the interval $(0, 2)$. Hence, we obtain:

$$\chi^2(\mathbf{A}) = \frac{1}{(2\pi)^{\frac{r}{2}} \cdot |h_1(\mathbf{A})|} \cdot \prod_{i=1}^{r} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}\left(\theta_i \cdot w_i^2\right)\right] \mathrm{d}w_i - 1 \tag{A11}$$

$$= \frac{1}{(2\pi)^{\frac{r}{2}} \cdot |h_1(\mathbf{A})|} \cdot \prod_{i=1}^{r} \sqrt{\frac{2\pi}{\theta_i}} - 1 \tag{A12}$$

$$= \frac{1}{|h_1(\mathbf{A})|} \cdot \sqrt{\frac{1}{|\mathbf{K}_1|}} - 1\,. \tag{A13}$$

Recall that the eigenvalues of $h_1(\mathbf{A})$ are given by $\{\gamma_i\}_{i=1}^r$ in the descending order. Therefore, (A13) simplifies to:

$$\chi^2(\mathbf{A}) = \prod_{i=1}^{r} \sqrt{\frac{1}{\gamma_i \cdot (2 - \gamma_i)}} - 1 = \prod_{i=1}^{r} g_{\chi_1}(\gamma_i) - 1\,. \tag{A14}$$

Hence, from (A14), maximizing $\chi^2(\mathbf{A})$ is equivalent to choosing the eigenvalues $\{\gamma_i\}_{i=1}^r$ such that they maximize $g_{\chi_1}(x)$. Similarly, the closed-form expression for $\chi^2(\mathbb{P}_\mathbf{A} \parallel \mathbb{Q}_\mathbf{A})$ can be derived as follows:

$$\chi^2(\mathbb{P}_\mathbf{A} \parallel \mathbb{Q}_\mathbf{A}) = \frac{|h_1(\mathbf{A})|^{\frac{1}{2}}}{(2\pi)^{\frac{r}{2}} \cdot |\mathbf{I}_r|} \cdot \int_{\mathbb{R}^r} \exp\left[-\frac{1}{2} \cdot \left(Y^\top \cdot \mathbf{K}_2 \cdot Y\right)\right] \mathrm{d}Y - 1\,, \tag{A15}$$

where we defined $\mathbf{K}_2 \stackrel{\triangle}{=} 2 \cdot \mathbf{I}_r - h_1(\mathbf{A})^{-1}$. We note that $\mathbf{K}_2$ is a real symmetric matrix due to $h_1(\mathbf{A})$ being a real symmetric matrix. Hence, following a similar line of argument as in the case of $\chi^2(\mathbf{A})$, and as a consequence of Theorem 2, we conclude that all the

eigenvalues λ_i should fall in the interval $(0.5, \infty)$ to ensure a finite value for $\chi^2(\mathbb{P}_{\mathbf{A}} \parallel \mathbb{Q}_{\mathbf{A}})$. Following this requirement, since the integrands are bounded, we obtain the following closed-form expression:

$$\chi^2(\mathbb{P}_{\mathbf{A}} \parallel \mathbb{Q}_{\mathbf{A}}) = \frac{|h_1(\mathbf{A})|^{\frac{1}{2}}}{1} \cdot \sqrt{\frac{1}{|\mathbf{K}_2|}} - 1 \,. \tag{A16}$$

Recall that the eigenvalues of $h_1(\mathbf{A})$ are given by $\{\gamma_i\}_{i=1}^r$; then, (A16) simplifies to

$$\chi^2(\mathbb{P}_{\mathbf{A}} \parallel \mathbb{Q}_{\mathbf{A}}) = \prod_{i=1}^{r} \sqrt{\frac{\gamma_i^2}{(2\gamma_i - 1)}} - 1 = \prod_{i=1}^{r} g_{\chi_2}(\gamma_i) - 1 \,. \tag{A17}$$

Hence, from (A17), maximizing $\chi^2(\mathbb{P}_{\mathbf{A}} \parallel \mathbb{Q}_{\mathbf{A}})$ is equivalent to choosing the eigenvalues $\{\gamma_i\}_{i=1}^r$ such that they maximize $g_{\chi_2}(x)$.

We observe that $\mathsf{H}^2(\mathbf{A})$, $\chi^2(\mathbf{A})$, and $\chi^2(\mathbb{P}_{\mathbf{A}} \parallel \mathbb{Q}_{\mathbf{A}})$ can be decomposed as the product of r non-negative identical convex functions involving $g_{\mathsf{H}}(x)$, $g_{\chi_1}(x)$, and $g_{\chi_2}(x)$, respectively. Hence, the goal is to choose $\{\gamma_i\}_{i=1}^r$ such that $D_f \in \{\mathsf{H}^2(\mathbf{A}), \chi^2(\mathbf{A}), \chi^2(\mathbb{P}_{\mathbf{A}} \parallel \mathbb{Q}_{\mathbf{A}})\}$ is maximized subjected to spectral constraints given by $\lambda_{n-(r-i)} \leq \gamma_i \leq \lambda_i$. In order to choose appropriate γ_i's, we first note that the global minimizer for each $g_f \in \{g_{\mathsf{H}}, g_{\chi_1}, g_{\chi_2}\}$ is attained at $x = 1$. Leveraging this observation, along with the structure that each g_f is convex, it is easy to infer that each $g_f(x)$ is monotonically increasing for $x > 1$ and monotonically decreasing $x < 1$. From the exact same argument in Appendix B, we obtain Corollaries 3 and 4.

Therefore, similar to Appendix B, constructing the linear map $\mathbf{A}$ that maximizes $D_f \in \{\mathsf{H}^2(\mathbf{A}), \chi^2(\mathbf{A}), \chi^2(\mathbb{P}_{\mathbf{A}} \parallel \mathbb{Q}_{\mathbf{A}})\}$ for any data matrix $\mathbf{\Sigma}$ boils down to choosing eigenvectors as rows of $\mathbf{A}$ associated with the chosen permutation of eigenvalues for each of the aforementioned cases.

Appendix D. Gradient Expressions for f-Divergence Measures

For clarity in analysis, we define the following functions:

$$h_2(\mathbf{A}) \triangleq \boldsymbol{\mu}^\top \cdot \mathbf{A}^\top \cdot \mathbf{A} \cdot \boldsymbol{\mu} \,, \tag{A18}$$

$$h_3(\mathbf{A}) \triangleq \boldsymbol{\mu}^\top \cdot \mathbf{A}^\top \cdot [h_1(\mathbf{A})]^{-1} \cdot \mathbf{A} \cdot \boldsymbol{\mu} \,. \tag{A19}$$

Based on these definitions, we have the following representations for the divergence measures and their associated gradients:

$$D_{\mathsf{KL}}(\mathbf{A}) = \frac{1}{2}\left[\log \frac{1}{|h_1(\mathbf{A})|} - r + \mathrm{Tr}[h_1(\mathbf{A})] + h_2(\mathbf{A})\right] , \tag{A20}$$

$$\begin{aligned}\nabla_{\mathbf{A}} D_{\mathsf{KL}}(\mathbf{A}) = {} & [h_1(\mathbf{A})]^{-1} \cdot \left[\mathbf{I}_r - [h_1(\mathbf{A})]^{-1} - \mathbf{A} \cdot \boldsymbol{\mu} \cdot \boldsymbol{\mu}^\top \cdot \mathbf{A}^\top \cdot [h_1(\mathbf{A})]^{-1}\right] \cdot \mathbf{A} \cdot \mathbf{\Sigma} \\ & + [h_1(\mathbf{A})]^{-1} \cdot \mathbf{A} \cdot \boldsymbol{\mu} \cdot \boldsymbol{\mu}^\top \,.\end{aligned}$$

$$D_{\mathsf{KL}}(\mathbb{P}_{\mathbf{A}} \parallel \mathbb{Q}_{\mathbf{A}}) = \frac{1}{2}\left[\log |h_1(\mathbf{A})| - r + \mathrm{Tr}\left[h_1(\mathbf{A})^{-1}\right] + h_3(\mathbf{A})\right] , \tag{A21}$$

$$\nabla_{\mathbf{A}} D_{\mathsf{KL}}(\mathbb{P}_{\mathbf{A}} \parallel \mathbb{Q}_{\mathbf{A}}) = \left(\mathbf{I}_r - [h_1(\mathbf{A})]^{-1}\right) \cdot \mathbf{A} \cdot \mathbf{\Sigma} + \mathbf{A} \cdot \boldsymbol{\mu} \cdot \boldsymbol{\mu}^\top \,.$$

$$D_{\mathsf{SKL}}(\mathbf{A}) = \frac{1}{2} \cdot \left[\mathrm{Tr}\left([h_1(\mathbf{A})]^{-1} + h_1(\mathbf{A})\right) + h_2(\mathbf{A}) + h_3(\mathbf{A})\right] - r \,, \tag{A22}$$

$$\begin{aligned}\nabla_{\mathbf{A}} D_{\mathsf{SKL}}(\mathbf{A}) = {} & \left[\mathbf{I}_r - [h_1(\mathbf{A})]^{-2} - [h_1(\mathbf{A})]^{-1} \cdot \mathbf{A} \cdot \boldsymbol{\mu} \cdot \boldsymbol{\mu}^\top \cdot \mathbf{A}^\top \cdot [h_1(\mathbf{A})]^{-1}\right] \cdot \mathbf{A} \cdot \mathbf{\Sigma} \\ & + \left(\mathbf{I}_r + [h_1(\mathbf{A})]^{-1}\right) \cdot \mathbf{A} \cdot \boldsymbol{\mu} \cdot \boldsymbol{\mu}^\top \,.\end{aligned} \tag{A23}$$

$$\mathsf{H}^2(\mathbf{A}) = 2 - 2\frac{|4 \cdot h_1(\mathbf{A})|^{\frac{1}{4}}}{|h_1(\mathbf{A}) + \mathbf{I}_r|^{\frac{1}{2}}} \cdot \exp\left(-\frac{\boldsymbol{\mu}^\top \cdot \mathbf{A}^\top \cdot [h_1(\mathbf{A}) + \mathbf{I}_r]^{-1} \cdot \mathbf{A} \cdot \boldsymbol{\mu}}{4}\right), \tag{A24}$$

$$\begin{aligned}\frac{\nabla_{\mathbf{A}}\mathsf{H}^2(\mathbf{A})}{-[1-\mathsf{H}^2(\mathbf{A})]} &= \frac{1}{2} \cdot [h_1(\mathbf{A})]^{-1} \cdot \mathbf{A} \cdot \boldsymbol{\Sigma} + [h_1(\mathbf{A}) + \mathbf{I}_r]^{-1} \cdot \Big[-\mathbf{A} \cdot [\boldsymbol{\Sigma} + \mathbf{I}_n] - \frac{1}{2} \cdot \mathbf{A} \cdot \boldsymbol{\mu} \cdot \boldsymbol{\mu}^\top \\ &\quad + \frac{1}{2} \cdot \mathbf{A} \cdot \boldsymbol{\mu} \cdot \boldsymbol{\mu}^\top \cdot \mathbf{A}^\top \cdot [h_1(\mathbf{A}) + \mathbf{I}_r]^{-1} \cdot \mathbf{A} \cdot [\boldsymbol{\Sigma} + \mathbf{I}_n]\Big].\end{aligned}$$

Appendix E. Proof for Lagrange Multipliers

Denoting $\nabla_{\mathbf{A}}\mathcal{L}$ by $\tilde{\Delta}$ and $\nabla_{\mathbf{A}}D_f$ by Δ, and further post-multiplying (59) by $\mathbf{A}^{k+1}$, we have:

$$\mathbf{A}^{k+1} \cdot (\mathbf{A}^{k+1})^\top = \mathbf{A}^k \cdot (\mathbf{A}^{k+1})^\top + \alpha \cdot \tilde{\Delta} \cdot (\mathbf{A}^{k+1})^\top, \tag{A25}$$

$$\mathbf{I}_r = \mathbf{A}^k \cdot (\mathbf{A}^k + \alpha \cdot \tilde{\Delta})^\top + \alpha \cdot \tilde{\Delta} \cdot (\mathbf{A}^k + \alpha \cdot \tilde{\Delta})^\top, \tag{A26}$$

$$\mathbf{0} = \mathbf{A}^k \cdot \tilde{\Delta}^\top + \tilde{\Delta} \cdot (\mathbf{A}^k)^\top + \alpha \cdot \tilde{\Delta} \cdot \tilde{\Delta}^\top. \tag{A27}$$

Substituting $\tilde{\Delta} = \Delta + \boldsymbol{\Lambda} \cdot \mathbf{A}$ in (A27) and simplifying the expression, we obtain:

$$2 \cdot \boldsymbol{\Lambda} + \mathbf{A}^k \cdot \Delta^\top + \Delta \cdot (\mathbf{A}^k)^\top = -\alpha \cdot (\Delta \cdot \Delta^\top + \Delta \cdot (\mathbf{A}^k)^\top \boldsymbol{\Lambda} + \boldsymbol{\Lambda} \cdot \mathbf{A}^k \cdot \Delta^\top + \boldsymbol{\Lambda} \cdot \boldsymbol{\Lambda}^\top). \tag{A28}$$

By noting that $\boldsymbol{\Lambda}$ is symmetric, taking the Taylor series expansions of $\boldsymbol{\Lambda}$ around $\alpha = 0$ and equating the terms of α in both sides, we obtain the relationships in (60)–(62).

References

1. Kunisky, D.; Wein, A.S.; Bandeira, A.S. Notes on computational hardness of hypothesis testing: Predictions using the low-degree likelihood ratio. *arXiv* **2019**, arXiv:1907.11636.
2. Gamarnik, D.; Jagannath, A.; Wein, A.S. Low-degree hardness of random optimization problems. *arXiv* **2020**, arXiv:2004.12063.
3. van der Maaten, L.; Postma, E.; van den Herik, J. Dimensionality reduction: A comparative review. *J. Mach. Learn. Res.* **2009**, *10*, 66–71.
4. Lee, J.A.; Verleysen, M. *Nonlinear Dimensionality Reduction*; Springer Science: Berlin/Heidelberg, Germany, 2007.
5. DeMers, D.; Cottrell, G.W. Non-linear dimensionality reduction. In Proceedings of the Advances in Neural Information Processing Systems, Denver, CO, USA, 3–6 November 1993; pp. 580–587.
6. Cunningham, J.P.; Ghahramani, Z. Linear dimensionality reduction: Survey, insights, and generalizations. *J. Mach. Learn. Res.* **2015**, *16*, 2859–2900.
7. Pearson, K. On lines and planes of closest fit to systems of points in space. *Philos. Mag.* **1901**, *2*, 559–572. [CrossRef]
8. Eckart, C.; Young, G. The approximation of one matrix by another of lower rank. *Psychometrika* **1936**, *1*, 211–218. [CrossRef]
9. Jolliffe, I., *Principal Component Analysis*; Springer: Berlin/Heidelberg, Germany, 2002.
10. Torgerson, W.S. Multidimensional scaling: I. Theory and method. *Psychometrika* **1952**, *17*, 401–419. [CrossRef]
11. Cox, T.F.; Cox, M.A. Multidimensional scaling. In *Handbook of Data Visualization*; Springer: Berlin/Heidelberg, Germany, 2008.
12. Borg, I.; Groenen, P.J. *Modern Multidimensional Scaling: Theory and Applications*; Springer: Berlin/Heidelberg, Germany, 2005.
13. Izenman, A.J. Linear discriminant analysis. *Modern Multivariate Statistical Techniques*; Springer: New York, NY, USA, 2013; pp. 237–280.
14. Globerson, A.; Tishby, N. Sufficient dimensionality reduction. *J. Mach. Learn. Res.* **2003**, *3*, 1307–1331.
15. Fisher, R.A. The use of multiple measurements in taxonomic problems. *Ann. Eugen.* **1936**, *7*, 179–188. [CrossRef]
16. Rao, C.R. The utilization of multiple measurements in problems of biological classification. *J. R. Stat. Soc. Ser. B* **1948**, *10*, 159–203. [CrossRef]
17. Fukunaga, K. *Introduction to Statistical Pattern Recognition*; Elsevier: Amsterdam, The Netherlands, 2013.
18. Suresh, B.; Ganapathiraju, A. Linear discriminant analysis—A brief tutorial. *Inst. Signal Inf. Process.* **1998**, *18*, 1–8.
19. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2006.
20. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2009.
21. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]
22. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [CrossRef]
23. Gelfand, I.M.; Kolmogorov, A.N.; Yaglom, A.M. On the general definition of the amount of information. *Dokl. Akad. Nauk SSSR* **1956**, *11*, 745–748.

24. Csiszár, I. Eine Informationstheoretische Ungleichung und ihre Anwendung auf den Bewis der Ergodizität von Markhoffschen Ketten. *Magy. Tudományos Akad. Mat. Kut. Intézetének Közleményei* **1948**, *8*, 379–423.
25. Ali, S.M.; Silvey, S.D. General Class of Coefficients of Divergence of One Distribution from Another. *J. R. Stat. Soc.* **1966**, *28*, 131–142. [CrossRef]
26. Morimoto, T. Markov Processes and the H-Theorem. *J. Phys. Soc. Jpn.* **1963**, *18*, 328–331. [CrossRef]
27. Arimoto, S. Information-theoretical considerations on estimation problems. *Inf. Control* **1971**, *19*, 181–194. [CrossRef]
28. Barron, A.R.; Gyorfi, L.; Meulen, E.C. Distribution estimation consistent in total variation and in two types of information divergence. *IEEE Trans. Inf. Theory* **1992**, *38*, 1437–1454. [CrossRef]
29. Berlinet, A.; Vajda, I.; Meulen, E.C. About the asymptotic accuracy of Barron density estimates. *IEEE Trans. Inf. Theory* **1998**, *44*, 999–1009. [CrossRef]
30. Gyorfi, L.; Morvai, G.; Vajda, I. Information-theoretic methods in testing the goodness of fit. In Proceedings of the IEEE International Symposium on Information Theory, Sorrento, Italy, 25–30 June 2000.
31. Liese, F.; Vajda, I. On Divergences and Informations in Statistics and Information Theory. *IEEE Trans. Inf. Theory* **2006**, *52*, 4394–4412. [CrossRef]
32. Kailath, T. The Divergence and Bhattacharyya Distance Measures in Signal Selection. *IEEE Trans. Commun. Technol.* **1967**, *15*, 52–60. [CrossRef]
33. Poor, H. Robust decision design using a distance criterion. *IEEE Trans. Inf. Theory* **1980**, *26*, 575–587. [CrossRef]
34. Clarke, B.S.; Barron, A.R. Information-theoretic asymptotics of Bayes methods. *IEEE Trans. Inf. Theory* **1990**, *36*, 453–471. [CrossRef]
35. Harremoes, P.; Vajda, I. On Pairs of f-divergences and their joint range. *IEEE Trans. Inf. Theory* **2011**, *57*, 3230–3235. [CrossRef]
36. Sason, I.; Verdú, S. f-Divergence Inequalities. *IEEE Trans. Inf. Theory* **2016**, *62*, 5973–6006. [CrossRef]
37. Sason, I. On f-divergence: Integral representations, local behavior, and inequalities. *Entropy* **2018**, *20*, 383. [CrossRef]
38. Rao, C.R.; Statistiker, M. *Linear Statistical Inference and Its Applications*; Wiley: New York, NY, USA, 1973.
39. Poor, H.V.; Hadjiliadis, O. *Quickest Detection*; Cambridge University Press: Cambridge, UK, 2008.
40. Cavanaugh, J.E. Criteria for linear model selection based on Kullback's symmetric divergence. *Aust. N. Z. J. Stat.* **2004**, *46*, 257–274. [CrossRef]
41. Devroye, L.; Mehrabian, A.; Reddad, T. The total variation distance between high-dimensional Gaussians. *arXiv* **2020**, arXiv:1810.08693.
42. Pollak, M. Optimal detection of a change in distribution. *Ann. Stat.* **1985**, *13*, 206–227. [CrossRef]
43. Carter, K.M.; Raich, R.; Finn, W.G.; Hero, A.O. Information preserving component analysis: Data projections for flow cytometry analysis. *IEEE J. Sel. Top. Signal Process.* **2009**, *3*, 148–158. [CrossRef]
44. Wen, Z.; Yin, W. A feasible method for optimization with orthogonality constraints. *Math. Program.* **2013**, *142*, 397–434. [CrossRef]
45. Edelman, A.; Arias, T.; Smith, S. The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.* **1998**, *20*, 303–353. [CrossRef]