Linear Discriminant Analysis under f-divergence Measures

Anmol Dwivedi, Sihui Wang, and Ali Tajer
Electrical, Computer, and Systems Engineering Department
Rensselaer Polytechnic Institute
Troy, NY 12180

***Abstract*—In statistical inference, the information-theoretic performance limits can be often expressed in terms of a notion of divergence between the underlying statistical models (e.g., in binary hypothesis testing, the total error probability is related to the total variation between the 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 (divergence reduces due to the data processing inequality for divergence). This paper considers linear dimensionality reduction such that the divergence between the models is *maximally* preserved. Specifically, this paper focuses on the Gaussian models and characterizes an optimal projection of the data onto a lower dimensional subspace with respect to four f-divergence measures (Kullback-Leibler, χ^2, Hellinger, and total variation). There are two key observations. First, projections are not necessarily along the largest modes of the covariance matrix of the data, and even in some situations can 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, independently of the inference problem of interest.**

2021 IEEE International Symposium on Information Theory (ISIT) | 978-1-5386-8209-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ISIT45174.2021.9518004

I. Introduction

A. 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:

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

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 between $\mathbb{P}$ and $\mathbb{Q}$. A key practical obstacle to solving 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 computation power [1] and [2]. Such a gap grows as the dimension n grows. Besides such a performance gap, the stability of the statistics is also compromised in high dimensions.

Dimensionality reduction techniques have become integral parts of statistical analysis in high dimensions [3]–[6]. Linear dimensionality reduction methods linearly map the high-dimensional-data to lower dimensions while ensuring that desired features of the data are preserved. There exist two general approaches to linear dimensionality reduction in one dataset X that we review next.

This research was supported in part by the U. S. National Science Foundation under Grants ECCS-1933107 and CAREER Award ECCS-1554482.

B. Related Literature

*1) **Feature extraction**:* In one set of approaches, the objective is to select and extract relevant features in the data X. These approaches are generally unsupervised. Two widely-used techniques include principal component analysis (PCA), and its variations [7]–[9] and multidimensional scaling (MDS) [10]–[13]. 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. There exist extensive variations to both approaches, and we refer the reader to [6] for more discussions.

*2) **Class separation**:* In another set of approaches, the objective is to maximize between-class variability of the lower dimensional data. These approaches are supervised. One approach pertinent to this paper's scope is linear discriminant analysis (LDA) that leverages the distinction between given models and designs a linear projection such that its lower-dimensional output exhibits maximum separation across different models [14]–[18]. In general, LDA approaches generate two scatter matrices: within-class and between-class scatter matrices. The within-class scatter matrix shows the scatter of the samples around their respective class model. 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 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. The LDA approaches focus on reducing the dimension to one and maximizing separability between two classes. There exist, however, studies that consider reducing to dimensions higher than one and separation across more than two classes.

C. Contribution

The contribution of this paper has two main distinctions from the existing literature on LDA. First, LDA 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 select

f-divergence as a measure of separation between 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; (iii) it enables characterizing the information-theoretic performance limits after linear mapping. Our analyses are focused on Gaussian models.

The remainder of the paper is organized as follows. In Section II we provide the linear dimensionality reduction model and provide an overview of the f-divergence measures considered in this paper. Section III provides a motivating operational interpretation for each measure and then characterizes an optimal design of the linear mapping under each. Although we observe that the design of the linear mapping has differences under different measures, we have two main observations: (i) the optimal design of the linear mapping is not necessarily along the most dominant eigenvalues of the covariance matrix; (ii) in certain regimes, the linear mapping design is identical under different f-divergence measures making the design independent of the inference problem at hand. Section IV concludes the paper.

II. Preliminaries

Consider two n-dimensional zero-mean Gaussian models with different structures:

$$\mathbb{P}: \ \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{\mathbb{P}}) \ , \quad \text{and} \quad \mathbb{Q}: \ \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{\mathbb{Q}}) \ , \tag{2}$$

where $\mathbf{\Sigma}_{\mathbb{P}}$ and $\mathbf{\Sigma}_{\mathbb{Q}}$ are two distinct covariance matrices, and $\mathbb{P}$ and $\mathbb{Q}$ denote their associated Gaussian 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

$$Y \ \triangleq \mathbf{A} \cdot X \ . \tag{3}$$

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

$$\begin{array}{ll} \mathbb{P}_{\mathbf{A}}: & \mathcal{N}(0, \mathbf{A} \cdot \mathbf{\Sigma}_{\mathbb{P}} \cdot \mathbf{A}^\top) \\ \mathbb{Q}_{\mathbf{A}}: & \mathcal{N}(0, \mathbf{A} \cdot \mathbf{\Sigma}_{\mathbb{Q}} \cdot \mathbf{A}^\top) \end{array} \ . \tag{4}$$

Motivated by the inference problems that we discuss in Section III, our objective is to design the linear mapping matrix $\mathbf{A}$ that ensures 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., $\mathbf{\Sigma}_{\mathbb{P}}$ and $\mathbf{\Sigma}_{\mathbb{Q}}$) such that relevant notions of distance between $\mathbb{P}_{\mathbf{A}}$ and $\mathbb{Q}_{\mathbf{A}}$ are maximized. We use a number of f-divergence measures for capturing the distance 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[1]

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

[1]We use the shorthand $D_f(\mathbf{A})$ for the canonical notation $D_f(\mathbb{Q}_{\mathbf{A}} \,\|\, \mathbb{P}_{\mathbf{A}})$ for emphasizing the dependence on $\mathbf{A}$ and for the simplicity in notations.

where $\mathbb{E}_{\mathbb{Q}_{\mathbf{A}}}$ denotes expectation with respect to $\mathbb{Q}_{\mathbf{A}}$ and $f : (0, +\infty) \to \mathbb{R}$ is a convex function that is strictly convex at 1 and $f(1) = 0$. Given the 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 $f(t) = t \log t$:

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

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

$$\chi^2(\mathbf{A}) \ \triangleq \int_Y \frac{(\mathrm{d}\mathbb{Q}_{\mathbf{A}} - \mathrm{d}\mathbb{P}_{\mathbf{A}})^2}{\mathrm{d}\mathbb{P}_{\mathbf{A}}} \ . \tag{8}$$

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

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

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

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

III. Main Results

In this section, we provide an optimal design of $\mathbf{A}$ under the choices of f-divergence measures specified in Section II. One key observation is that the optimal choices of $\mathbf{A}$ under different measures have strong similarities. We first note that by defining $\bar{\mathbf{A}} = \mathbf{A}\mathbf{\Sigma}_{\mathbb{P}}^{1/2}$ and $\mathbf{\Sigma} = \triangleq \mathbf{\Sigma}_{\mathbb{P}}^{-1/2}\mathbf{\Sigma}_{\mathbb{Q}}\mathbf{\Sigma}_{\mathbb{P}}^{-1/2}$, designing $\mathbf{A}$ for maximally distinguishing

$$\mathcal{N}(0, \mathbf{A} \cdot \mathbf{\Sigma}_{\mathbb{P}} \cdot \mathbf{A}^\top) \quad \text{versus} \quad \mathcal{N}(0, \mathbf{A} \cdot \mathbf{\Sigma}_{\mathbb{Q}} \cdot \mathbf{A}^\top) \ , \tag{11}$$

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

$$\mathcal{N}(0, \bar{\mathbf{A}} \cdot \bar{\mathbf{A}}^\top) \quad \text{versus} \quad \mathcal{N}(0, \bar{\mathbf{A}} \cdot \mathbf{\Sigma} \cdot \bar{\mathbf{A}}^\top) \ . \tag{12}$$

Hence, without loss of generality, we focus on the setting $\mathbf{\Sigma}_{\mathbb{P}} = \mathbf{I}_r$ 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: Corresponding to any matrix $\mathbf{A}$ there exists a semi-orthogonal matrix $\bar{\mathbf{A}}$ such that $D_f(\bar{\mathbf{A}}) = D_f(\mathbf{A})$.

This observation indicates that, without loss of generality, we can reduce the unconstrained problem in (6) to the following constrained problem:

$$\mathcal{Q}: \quad \max_{\mathbf{A} \in \mathbb{R}^{r \times n}} \quad D_f(\mathbf{A}) \quad \text{s.t.} \quad \mathbf{A} \cdot \mathbf{A}^\top = \mathbf{I}_r \ . \tag{13}$$

Design of $\mathbf{A}$ under all measures directly relates to analyzing the eigenspace of matrix $\mathbf{\Sigma}$. For this purpose, we denote the non-negative eigenvalues of $\mathbf{\Sigma}$ ordered in the descending order by $\{\lambda_i : i \in [n]\}$. For an arbitrary permutation function $\pi : [n] \to [n]$, we also 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 $\mathbf{A}\mathbf{\Sigma}\mathbf{A}^\top$ ordered in the descending order by $\{\gamma_i : i \in [r]\}$, where for an integer m we have defined

$[m] = \{1, \ldots, m\}$. Throughout the analysis we frequently use Poincaré separation theorem [19] for finding the row space of matrix $\mathbf{A}$ with respect to the eigenvalues of $\mathbf{\Sigma}$.

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

$$\lambda_{n-(r-i)} \leq \gamma_i \leq \lambda_i \ , \quad \forall i \in [r] \ . \tag{14}$$

Finally, we define function $h : \mathbb{R}^{r \times n} \to \mathbb{R}^{r \times r}$ as

$$h(\mathbf{A}) \triangleq \mathbf{A} \cdot \mathbf{\Sigma} \cdot \mathbf{A}^\top \ . \tag{15}$$

In the remainder of this section, we analyze the optimal design of $\mathbf{A}$ under different f-divergence measures. In each case, we provide an operational interpretation of the measure in the dichotomous mode in (4).

A. Kullback-Leibler Divergence

1) Motivation: The KL divergence, being the expected value of the log-likelihood ratio, captures the performance of a wide range of inference problems. One specific problem whose performance is completely captured by the KL divergence measure 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.,

$$\begin{array}{ll} X_t \sim \mathbb{P} & t < \kappa \\ X_t \sim \mathbb{Q} & t \geq \kappa \end{array} \ . \tag{16}$$

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). Irrespectively of their discrepancies in settings and the nature of performance guarantees, the ADD in the (asymptotically) optimal algorithms are in the form [20]

$$\mathsf{ADD} \sim \frac{c}{D_{\mathsf{KL}}(\mathbb{Q} \parallel \mathbb{P})} \ , \tag{17}$$

which after the mapping induced by matrix $\mathbf{A}$ changes to

$$\mathsf{ADD} \sim \frac{c}{D_{\mathsf{KL}}(\mathbf{A})} \ , \tag{18}$$

where c is a constant specified by the false alarm constraints. Clearly, the 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 $\mathbb{Q}_{\mathbf{A}}$.

A similar KL divergence maximization appears also in variational inference, in which the objective involves maximizing an evidence lower bound (ELBO), also known as the variational lower bound, between an intractable posterior distribution $\mathbb{P}$ and a sought distribution $\mathbb{Q}$. As shown in [21], maximizing ELBO is equivalent to minimizing the KL divergence between $\mathbb{P}$ and $\mathbb{Q}$. Optimizing the ELBO has formed a basis for numerous approximate inference algorithms (e.g., mean field approximation) in various probabilistic graphical models [22].

2) Results and Observations: By noting that $\mathbf{A}$ is a semi-orthogonal matrix and recalling that the eigenvalues of $h(\mathbf{A})$ are denoted by $\{\gamma_i : i \in [r]\}$, simple algebraic manipulations simplify the KL divergence defined in (7) and is given by

$$D_{\mathsf{KL}}(\mathbf{A}) = \frac{1}{2}\left[\log \frac{1}{|h(\mathbf{A})|} - r + \mathsf{tr}\left[h(\mathbf{A})\right] \right] \tag{19}$$

$$= \sum_{i=1}^{r} g_{\mathsf{KL}}(\gamma_i) \ , \tag{20}$$

where we have defined

$$g_{\mathsf{KL}}(x) \triangleq \frac{1}{2}(x - \log x - 1) \ . \tag{21}$$

Hence, by leveraging Theorem 2 the optimal design of interest $\mathcal{Q}$ formalized in (22) can be restated as

$$\mathcal{Q} : \quad \left\{ \begin{array}{ll} \max\limits_{\{\gamma_i : i \in [r]\}} & \sum\limits_{i=1}^{r} g_{\mathsf{KL}}(\gamma_i) \\ \text{s.t.} & \lambda_{n-(r-i)} \leq \gamma_i \leq \lambda_i \ \ \forall i \in [r] \end{array} \right. . \tag{22}$$

Based on this, an optimal design of $\mathbf{A}$ is constructed by choosing r eigenvectors of $\mathbf{\Sigma}$ as the rows of $\mathbf{A}$. The results and observations are formalized in the next theorem and corollaries.

Theorem 3: Define the permutation $\pi^* : [n] \to [n]$ as a solution to:

$$\pi^* = \arg\max_{\pi} \sum_{i=1}^{r} g_{\mathsf{KL}}(\lambda_{\pi(i)}) \ . \tag{23}$$

Then, for maximizing $D_{\mathsf{KL}}(\mathbf{A})$:

1) The eigenvalues of $\mathbf{A}\mathbf{\Sigma}\mathbf{A}^\top$ are given by $\gamma_i = \lambda_{\pi^*(i)}$.
2) Row i of matrix $\mathbf{A}$ is the eigenvector of $\mathbf{\Sigma}$ associated with the eigenvalue $\gamma_i = \lambda_{\pi^*(i)}$.

By noting that g_{KL} is strictly convex taking its global minima at $x = 1$, we have the following additional observations.

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

Corollary 2: For maximizing $D_{\mathsf{KL}}(\mathbf{A})$, 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 the eigenvectors of $\mathbf{\Sigma}$ associated with its r smallest eigenvalues, i.e., $\{\lambda_i : i \in \{n - r + 1, \ldots, n\}\}$.

Remark 1: We note that when maximizing $D_{\mathsf{KL}}(\mathbf{A})$ for cases when $\lambda_n \leq 1 \leq \lambda_1$, finding the best permutations of eigenvectors involves sorting n eigenvalues and subsequently

performing r comparisons. This amounts to $\mathcal{O}(n \cdot \log(n) + r)$ time complexity instead of the $\mathcal{O}(n \cdot log(n))$ time complexity involved in the case of Corollaries 1 and 2.

Remark 2: It is noteworthy that the optimal design of $\mathbf{A}$ often does not involve being aligned with the largest eigenvalues of the covariance matrix $\mathbf{\Sigma}$. This is in contrast to some of the key approaches to linear dimensionality reduction which generally perform linear mapping along the eigenvectors associated with the largest eigenvalues of the covariance matrix for the purpose of preserving as much data variation as possible in the lower dimension. 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.

B. χ^2 Divergence

1) Motivation: χ^2 divergence appears in a wide range of statistical estimation problems for the purpose of finding 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 $p(\theta)$ and $q(\theta)$. Define $\mathbb{P}$ and $\mathbb{Q}$ as the probability distributions of $p(\theta)$ and $q(\theta)$, respectively. According to the Hammersly-Chapman-Robbins (HCR) bound on the estimation quadratic loss function for any estimator $\hat{\theta}$ we have

$$\mathsf{var}_\theta(\hat{\theta}) \geq \sup_{p \neq q} \frac{[\mathbb{E}_{\mathbb{Q}}[q(X)] - \mathbb{E}_{\mathbb{P}}[p(X)]]^2}{\chi^2(\mathbb{Q} \parallel \mathbb{P})} \ , \tag{24}$$

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

$$\mathsf{var}_\theta(\hat{\theta}) \geq \sup_{p \neq q} \frac{(q-p)^2}{\chi^2(\mathbb{Q} \parallel \mathbb{P})} \ , \tag{25}$$

which depends 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 the 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, and it is easier compute compared to the KL divergence and the 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.

2) Results and Observations: For a given matrix $\mathbf{A}$, from (8) we have the following closed-form expression

$$\chi^2(\mathbf{A}) = \frac{1}{|h(\mathbf{A})|\sqrt{|2(h(\mathbf{A}))^{-1} - \mathbf{I}_r|}} - 1 \tag{26}$$

$$= \prod_{i=1}^{r} g_\chi(\gamma_i) - 1 \ , \tag{27}$$

where we have defined

$$g_\chi(x) \triangleq \frac{1}{\sqrt{x(2-x)}} \ . \tag{28}$$

It indicates that for having a finite χ^2 divergence, all the eigenvalues are bounded away from 2 (a singularity point) and the number of eigenvalues larger than 2 must be even. To place the emphasis on the key observations, we initially focus on the case in which for all $i \in [n]$ we have $\lambda_i \in (0, \varepsilon]$, where $\varepsilon < 2$ is a given constant. Based on this and by following a similar line of argument as in the case of KL divergence, designing an optimal $\mathbf{A}$ reduces to identifying a subset of the eigenvalues of $\mathbf{\Sigma}$ and assigning their associated eigenvalues as the rows of matrix $\mathbf{A}$. These observations are formalized next.

Theorem 4: Define the permutation $\pi^* : [n] \to [n]$ as a solution to:

$$\pi^* \triangleq \arg\max_{\pi} \prod_{i=1}^{r} g_\chi(\lambda_{\pi(i)}) \ . \tag{29}$$

Then for maximizing $\chi^2(\mathbf{A})$:

1) The eigenvalues of $\mathbf{A}\mathbf{\Sigma}\mathbf{A}^\top$ are given by $\gamma_i = \lambda_{\pi^*(i)}$.
2) Row i of matrix $\mathbf{A}$ is the eigenvector of $\mathbf{\Sigma}$ associated with the eigenvalue $\lambda_{\pi^*(i)}$.

By noting that g_χ is strictly convex over $(0, 2)$ and takes its global minima at $x = 1$, we have the following additional observations.

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

Corollary 4: For maximizing $\chi^2(\mathbf{A})$, 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 the eigenvectors of $\mathbf{\Sigma}$ associated with its r smallest eigenvalues, i.e., $\{\lambda_i : i \in \{n-r+1, \ldots, n\}\}$.

C. Squared Hellinger Distance

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 $[0, 2]$.

2) Results and Observations: For a given matrix $\mathbf{A}$ we have the following closed-form expression:

$$\mathsf{H}^2(\mathbf{A}) = 2 - 2\frac{|4h(\mathbf{A})|^{\frac{1}{4}}}{|h(\mathbf{A}) + \mathbf{I}_r|^{\frac{1}{2}}} \tag{30}$$

$$= 2 - 2\prod_{i=1}^{r} \sqrt[4]{\frac{4}{g_{\mathsf{H}}(\gamma_i)}} \ , \tag{31}$$

where we have defined

$$g_{\mathsf{H}}(x) \triangleq \frac{(x+1)^2}{x} \ . \tag{32}$$

The structure of the results and the key observations are consistent with the case of KL divergence, as formalized next.

Theorem 5: Define the permutation $\pi^* : [n] \to [n]$ as a solution to:

$$\pi^* = \arg\max_{\pi} \prod_{i=1}^{r} g_{\mathsf{H}}(\lambda_{\pi(i)}) \ . \tag{33}$$

Then, for maximizing $\mathsf{H}^2(\mathbf{A})$:

1) The eigenvalues of $\mathbf{A}\mathbf{\Sigma}\mathbf{A}^\top$ are given by $\gamma_i = \lambda_{\pi^*(i)}$.
2) Row i of matrix $\mathbf{A}$ is the eigenvector of $\mathbf{\Sigma}$ associated with the eigenvalue $\lambda_{\pi^*(i)}$.

By noting that g_{H} is strictly convex over taking its global minima at $x = 1$, we have the following additional observations.

Corollary 5: If $\lambda_n \geq 1$, then $\gamma_i = \lambda_i$ for all $i \in [r]$, and the rows of $\mathbf{A}$ are the eigenvectors of $\mathbf{\Sigma}$ associated with its r largest eigenvalues, i.e., $\{\lambda_i : i \in [r]\}$.

Corollary 6: If $\lambda_1 \leq 1$, then $\gamma_i = \lambda_{n-r+i}$ for all $i \in [r]$, and the rows of $\mathbf{A}$ are the eigenvectors of $\mathbf{\Sigma}$ associated with its r smallest eigenvalues, i.e., $\{\lambda_i : i \in \{n-r+1, \ldots, n\}\}$.

Remark 3: An observation can be made that for the case of squared Hellinger distance, λ_i's $(\lambda_i > 1)$ are equally favored as that of $\frac{1}{\lambda_i}$. This is different from the KL divergence favoring larger eigenvalues $(\lambda_i > 1)$ over the smaller ones $(\frac{1}{\lambda_i} < 1)$.

D. Total Variation

1) Motivation: Total variation 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, in a simple binary hypothesis test

$$\mathsf{H}_0 \ : X \sim \mathbb{P} \quad \text{versus} \quad \mathsf{H}_1 \ : X \sim \mathbb{Q} \ , \tag{34}$$

the minimum total probability of error (sum of type I and type II error probabilities) is related to the total variation $d_{\mathsf{TV}}(\mathbb{P} \parallel \mathbb{Q})$. If we define $d : X \to \{\mathsf{H}_0, \mathsf{H}_1\}$ as a decision rule, then

$$\inf_d [\mathbb{P}(d = \mathsf{H}_1) + \mathbb{Q}(d = \mathsf{H}_0)] = 1 - d_{\mathsf{TV}}(\mathbb{P} \parallel \mathbb{Q}) \ . \tag{35}$$

The total variation distance 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 have 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 by 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 [23].

2) Matching Bounds up to a Constant: As shown in [23], the total variation of interest is given by

$$\frac{1}{100} \leq \frac{d_{\mathsf{TV}}(\mathbf{A})}{\min\{1, \sqrt{\sum_{i=1}^r g_{\mathsf{TV}}(\gamma_i)}\}} \leq \frac{3}{2} \ , \tag{36}$$

where we have defined

$$g_{\mathsf{TV}}(x) \triangleq \left(\frac{1}{x} - 1\right)^2 \ . \tag{37}$$

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 $\mathbf{A}$. Next we show that optimizing the bounds lead to a design for $\mathbf{A}$ for which we have observations consistent with those of the KL divergence and the squared Hellinger distance.

Theorem 6: Define the permutation $\pi^* : [n] \to [n]$ as a solution to:

$$\pi^* = \arg\max_{\pi} \sum_{i=1}^r g_{\mathsf{TV}}(\lambda_{\pi(i)}) \ . \tag{38}$$

Then for maximizing the bounds on $g_{\mathsf{TV}}(x)$ in (36):

1) The eigenvalues of $\mathbf{A}\mathbf{\Sigma}\mathbf{A}^\top$ are given by $\gamma_i = \lambda_{\pi^*(i)}$.
2) Row i of matrix $\mathbf{A}$ is the eigenvector of $\mathbf{\Sigma}$ associated with the eigenvalue $\lambda_{\pi^*(i)}$.

By noting that g_{TV} is strictly convex taking its global minima at $x = 1$, we have the following additional observations.

Corollary 7: For maximizing the bounds on $g_{\mathsf{TV}}(x)$ in (36), when $\lambda_n \geq 1$ we have $\gamma_i = \lambda_i$ for all $i \in [r]$, and the rows of $\mathbf{A}$ are the eigenvectors of $\mathbf{\Sigma}$ associated with its r largest eigenvalues, i.e., $\{\lambda_i : i \in [r]\}$.

Corollary 8: For maximizing the bounds on $g_{\mathsf{TV}}(x)$ in (36), 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 the eigenvectors of $\mathbf{\Sigma}$ associated with its r smallest eigenvalues, i.e., $\{\lambda_i : i \in \{n-r+1, \ldots, n\}\}$.

Remark 4: We notice that the evaluation function g_{TV} favors smaller eigenvalues $(\lambda_i < 1)$ over the larger ones $(\frac{1}{\lambda_i} > 1)$. This is in contrast to other forms of f-divergence measures we have analyzed in the previous subsections.

3) Bounding by Hellinger Distance: Total variation can be also bounded by the Hellinger distance as follows.

$$\frac{1}{2}\mathsf{H}^2(\mathbf{A}) \leq d_{\mathsf{TV}}(\mathbf{A}) \leq \mathsf{H}(\mathbf{A})\sqrt{1 - \frac{\mathsf{H}^2(\mathbf{A})}{4}} \ . \tag{39}$$

It can be readily verified that these bounds are monotonously increasing with $\mathsf{H}^2(\mathbf{A})$ in the interval $[0, 2]$. Hence, they are maximized simultaneously by maximizing the squared Hellinger distance discussed in Section III-C.

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 eigenvectors of $\mathbf{\Sigma}$ to construct $\mathbf{A}$ according to the functions

$$g_{\mathsf{H}}(x) = \frac{(x+1)^2}{x} \quad \text{versus} \quad g_{\mathsf{TV}}(x) = \left(\frac{1}{x} - 1\right)^2 \ . \tag{40}$$

IV. Conclusion

In this paper, we have considered the problem of linear discriminant analysis (LDA) such that separation is maximized under f-divergence measures. This approach is motivated by dimensionality reduction for inference problems, where we have investigated LDA under Kullback-Leibler, χ^2, Hellinger, and total variation measures. We have characterized an optimal design for the linear transformation of the data onto a lower-dimensional subspace in each case for Gaussian models. We have shown that 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.

REFERENCES

[1] D. Kunisky, A. S. Wein, and A. S. Bandeira, "Notes on computational hardness of hypothesis testing: Predictions using the low-degree likelihood ratio," 2019.

[2] D. Gamarnik, A. Jagannath, and A. S. Wein, "Low-degree hardness of random optimization problems," 2020.

[3] D. DeMers and G. W. Cottrell, "Non-linear dimensionality reduction," in *Advances in neural information processing systems*, November 1993, pp. 580–587.

[4] J. A. Lee and M. Verleysen, *Nonlinear dimensionality reduction*. Springer Science & Business Media, 2007.

[5] L. Van Der Maaten, E. Postma, and J. Van den Herik, "Dimensionality reduction: a comparative," *J Mach Learn Res*, vol. 10, no. 66-71, p. 13, 2009.

[6] J. P. Cunningham and Z. Ghahramani, "Linear dimensionality reduction: Survey, insights, and generalizations," *The Journal of Machine Learning Research*, vol. 16, no. 1, pp. 2859–2900, December 2015.

[7] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philosophical Magazine*, vol. 2, no. 11, pp. 559–572, 1901.

[8] C. Eckart and G. Young, "The approximation of one matrix by another of lower rank," *Psychometrika*, vol. 1, no. 3, pp. 211–218, 1936.

[9] I. Jolliffe and Springer-Verlag, *Principal Component Analysis*, ser. Springer Series in Statistics. Springer, 2002.

[10] W. S. Torgerson, "Multidimensional scaling: I. theory and method," *Psychometrika*, vol. 17, no. 4, pp. 401–419, 1952.

[11] T. F. Cox and M. A. Cox, "Multidimensional scaling," in *Handbook of data visualization*. Springer, 2008.

[12] I. Borg and P. J. Groenen, *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005.

[13] A. J. Izenman, "Linear discriminant analysis," in *Modern multivariate statistical techniques*. Springer, 2013, pp. 237–280.

[14] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[15] C. R. Rao, "The utilization of multiple measurements in problems of biological classification," *Journal of the Royal Statistical Society. Series B*, vol. 10, no. 2, pp. 159–203, 1948.

[16] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. Elsevier, 2013.

[17] S. Balakrishnama and A. Ganapathiraju, "Linear discriminant analysis-a brief tutorial," *Institute for Signal and information Processing*, vol. 18, no. 1998, pp. 1–8, 1998.

[18] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[19] C. R. Rao and M. Statistiker, *Linear statistical inference and its applications*. Wiley New York, 1973, vol. 2.

[20] H. V. Poor and O. Hadjiliadis, *Quickest detection*. Cambridge, UK: Cambridge University Press, 2008.

[21] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Machine learning*, vol. 37, no. 2, pp. 183–233, November 1999.

[22] M. J. Wainwright and M. I. Jordan, *Graphical models, exponential families, and variational inference*. Now Publishers Inc, 2008.

[23] L. Devroye, A. Mehrabian, and T. Reddad, "The total variation distance between high-dimensional gaussians," 2020.