PERGAMON

Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE @DIRECT®

Pattern Recognition 37 (2004) 857–874

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

On the relation between discriminant analysis and mutual information for supervised linear feature extraction

Sergios Petridis, Stavros J. Perantonis*

Computational Intelligence Laboratory, Institute of Informatics and Telecommunications, National Center for Scientific Research "Demokritos", 15310 Aghia Paraskevi, Athens,Greece

Received 27 November 2002; received in revised form 27 October 2003

Abstract

This paper provides a unifying view of three discriminant linear feature extraction methods: linear discriminant analysis, heteroscedastic discriminant analysis and maximization of mutual information. We propose a model-independent reformulation of the criteria related to these three methods that stresses their similarities and elucidates their differences. Based on assumptions for the probability distribution of the classification data, we obtain sufficient conditions under which two or more of the above criteria coincide. It is shown that these conditions also suffice for Bayes optimality of the criteria. Our approach results in an information-theoretic derivation of linear discriminant analysis and heteroscedastic discriminant analysis. Finally, regarding linear discriminant analysis, we discuss its relation to multidimensional independent component analysis and derive suboptimality bounds based on information theory.

Keywords: Linear feature extraction; Linear discriminant analysis; Heteroscedastic discriminant analysis; Maximization of mutual information; Bayes error; Negentropy

1. Introduction

Discriminant linear feature extraction (DLFE) is the task of reducing the dimension of the observation space by finding a suitable linear subspace in which the class separability is optimally maintained. DLFE is mainly used for two purposes: for data visualization, in which case the target subspace is usually of very small dimension (2 or 3), or as a preprocessing step in a pattern recognition system [1], since a reduced feature space dimension may lead to better classifier training with improved generalization ability. The importance and benefits of DLFE in a pattern recognition system have been emphasized even when combined with very competent classifier models, such as support vector machines [2].

In this paper we focus on three different methods for DLFE: the statistical *linear discriminant analysis* criterion (LDA), the *heteroscedastic discriminant analysis* criterion (HDA) and the information-theoretic *maximization of mutual information* criterion (MMI). LDA has a long tradition in statistics and pattern recognition [1] having been used in many application fields, such as face recognition [3,4], or document classification [5]. Several extensions and variations of the basic LDA algorithm have been developed concerning either implementation and robustness issues [6–8] or deviations from the model assumptions [9,10]. HDA [11] is a more recent approach derived by applying the maximum likelihood principle in an heteroscedastic model and has been successfully applied to speech recognition [11].

* Corresponding author. Tel.: +30-210-6503174; fax: +30-210-6532175.

E-mail addresses: petridis@iit.demokritos.gr (S. Petridis), sper@iit.demokritos.gr (S.J. Perantonis).

[1] Its popularity accounts also for its numerous viewpoints and names, such as multiple discriminant analysis, generalized Fisher criterion or canonical variate analysis.

0031-3203/$30.00 © 2003 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2003.12.002

The MMI principle has also been known for long as a natural criterion for evaluating the separability quality of features [12,13]. Due to its computational complexity [14] it has mostly been used in an approximate way for individual feature selection [15,16]. However, its main advantage of making no assumptions about the underlying probability model of data has motivated the development of successful linear feature extraction algorithms [17,18].

Each of the above three DLFE criteria has been originally derived by a different rationale. Moreover, their mathematical expressions differ in form, so that a comparison of these criteria is generally difficult. It must also be stressed that, although conditions that ensure the Bayes optimality of the LDA criterion are known, the same is not true for the other two criteria.

The purpose of this paper is to provide a unified view of these DLFE criteria by

- emphasizing their similarity through expressing them in a common framework and with mathematical forms that resemble each other
- investigating conditions under which these criteria recover subspaces that are Bayes optimal, in the sense that the minimum possible Bayes error is obtained, and hence optimal classification accuracy is ensured
- proposing conditions on the underlying probabilities of the observation data model under which two or more of the above criteria coincide, in the sense that they recover identical subspaces of the original observation space.

The first objective is met by proposing new model-independent mathematical forms for the three criteria that stress their similarities and elucidate their differences. To meet the second and third objectives, we state the DLFE problem as a source recovering problem. We introduce the following hierarchy of models on the observation space variables: the homoscedastic gaussian model (HOG) [19, p. 59], Kumar and Andreou's heteroscedastic model (KAH) [11] and a more general model which we call zero information loss model (ZIL). It is shown that each of these models is a special case of the next one in the hierarchy. It is also demonstrated that under the ZIL model the MMI criterion is Bayes optimal. Moreover, our analysis shows that under the KAH model the MMI criterion coincides with the HDA criterion and both criteria reach Bayes optimal solutions. Also, under the more restricted HOG model all three criteria coincide and are Bayes optimal. These results allow for an information-theoretic derivation of the LDA criterion. Finally, our discussion allows for an alternative interpretation of LDA as a special case of multidimensional independent component analysis [20] and for the derivation of suboptimality bounds for the LDA criterion based on information theory.

The rest of this paper is organized as follows: In Section 2, the LDA, HDA and MMI criteria are described and their derivations are briefly explained. In Section 3, the ZIL model is proposed and Bayes optimality of the MMI criterion under the ZIL model is proved. In Section 4, a unified model-independent view of the three DLFE criteria is presented by expressing them in forms that enable a straight comparison. In Section 5, a deeper investigation of the relation among the three criteria is carried out and equivalence as well as Bayes optimality in the framework of the HOG and KAH models is established. An information theoretic derivation of HDA and LDA is also proposed. In Section 6, a number of two-dimensional classification problems are given that illustrate the main results of the paper. Finally, in Section 7, the connection with multidimensional independent component analysis is established, suboptimality bounds of LDA are discussed and future work is outlined.

2. Foundations of LDA, HDA and MMI

2.1. The DLFE problem

Consider the classification problem which is concerned with finding an optimal rule for the assignment of a given observation, assumed to be an n-dimensional vector $\boldsymbol{x} \in \mathscr{X} \subseteq \mathbb{R}^n$, to one of K known classes ω_k among the set $\mathscr{C} = \{\omega_k,\ k = 1\dots K\}$.

From a probabilistic point of view, the above classification problem is solved in an optimal way by using the *Bayes classification rule* to determine the optimal class choice for a given observation. Let the observation vector and the class be the jointly distributed random variables X and Ω taking values from the sets $\mathscr{X}$ and $\mathscr{C}$ and having a-priori distribution functions $p(\boldsymbol{x})$ and $p(\omega_k)$ respectively.[2] The conditional probability of the class given the observation vector $\boldsymbol{x}$ will be denoted, as usual, by $p(\omega_k|\boldsymbol{x})$. The optimal class choice for a given observation, called *Bayes classification rule*, is made by selecting the class with maximum probability given the observation. The *Bayes error*, defined as the expected classification error resulting from assigning classes to the observations according to the Bayes classification rule, is given by

$$P_e(\Omega|\mathsf{X}) = 1 - \mathscr{E}_{\mathsf{X}}\left[\max_{\omega_k} p(\omega_k|\boldsymbol{x})\right], \tag{1}$$

where $\mathscr{E}_{\mathsf{X}}$ denotes expectation value with respect to X.

Consider now the DLFE problem. The task is to solve the classification problem using only an m-dimensional subspace of $\mathbb{R}^n$, where m is a given positive integer with $m < n$. Hence, it is sought to find a linear transformation of the observation vector, effected by a full rank matrix A of dimension $n \times m$, such that optimal classification accuracy is obtained using only the reduced dimensionality vector $\mathrm{A}^\top \boldsymbol{x}$.

To each choice of A there corresponds a certain Bayes error $P_e(\Omega|\mathrm{A}^\top\mathsf{X})$. The DLFE problem is now naturally

[2] By convention, capital letters will denote random variables and lowercase letters will denote particular values that random variables can take on.

defined as finding the set $\hat{\mathrm{A}}^{Bayes}$ of matrices that yield the minimum Bayes error $P_e(\Omega|\mathrm{A}^\top \mathsf{X})$:

$$\hat{\mathrm{A}}^{Bayes} = \operatorname*{argmin}_{\mathrm{A}} P_e(\Omega|\mathrm{A}^\top \mathsf{X}). \tag{2}$$

Any matrix that yields minimum Bayes error in Eq. (2) will be called *Bayes optimal matrix*.

It should be stressed that Bayes optimality is related to the subspace of $\mathbb{R}^n$ induced by matrix A rather than to the exact matrix [21, p. 441]. This means that any transformed vector of the form $T(\mathrm{A}^\top \mathsf{X})$, where T is a non-singular $m \times m$ matrix, gives the same Bayes error.

In practice, finding the Bayes optimal matrix by direct optimization of Eq. (2) can be very hard. As a common alternative, one may choose to find a matrix that maximizes other feature extraction criteria that have been proposed in the literature whose evaluation may be less demanding. However, the success of these criteria is still measured by their ability to extract subspaces with classification error close to the Bayes error. In the remainder of this section we review the LDA, HDA and MMI criteria, each one derived using a different path.

2.2. The LDA criterion

The LDA criterion has been proposed as a class separatory measure. The basic idea behind its derivation is to extract a subspace in which the classes means are far from each other, whereas the within class covariance matrices (i.e. class conditional covariance matrices) are small. To formulate the criterion, consider the overall mean $\boldsymbol{\mu}$ and class conditional means $\boldsymbol{\mu}_k$ of X defined by

$$\boldsymbol{\mu}_k = \mathscr{E}_{\mathsf{X}|\omega_k}[\boldsymbol{x}], \tag{3}$$

$$\boldsymbol{\mu} = \mathscr{E}_{\mathsf{X}}[\boldsymbol{x}] \tag{4}$$

and second order statistics, namely the class conditional covariance matrices Σ_k and their average $\bar{\Sigma}$ as well as the overall class covariance matrix Σ, defined by

$$\Sigma_k = \mathscr{E}_{\mathsf{X}|\omega_k}[(\boldsymbol{x}-\boldsymbol{\mu}_k)(\boldsymbol{x}-\boldsymbol{\mu}_k)^\top], \tag{5}$$

$$\bar{\Sigma} = \mathscr{E}_{\Omega}[\Sigma_k], \tag{6}$$

$$\Sigma = \mathscr{E}_{\mathsf{X}}[(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top]. \tag{7}$$

Given a positive integer $m < n$, the LDA criterion requires to find among all possible $n \times m$ full rank matrices the set $\hat{\mathrm{A}}^{LDA}$ of matrices defined by [21, p. 446]: [3]

$$\hat{\mathrm{A}}^{LDA} = \operatorname*{argmax}_{\mathrm{A}} \log \frac{|\mathrm{A}^\top \Sigma \mathrm{A}|}{|\mathrm{A}^\top \bar{\Sigma} \mathrm{A}|}. \tag{8}$$

Any matrix that yields the maximum in Eq. (8) will be called *LDA optimal matrix*. [4]

The above optimum can be found by solving a generalized eigenvalue problem. Namely, the matrix formed by the m eigenvectors of $\bar{\Sigma}^{-1}\Sigma$ with greater eigenvalues is LDA optimal (see Ref. [21, p. 449]).

In general, the LDA optimal matrix is not Bayes optimal. However, we can guarantee the Bayes optimality of LDA under some conditions. First, let us define the *homoscedastic gaussian model*.

Definition 1 (HOG model). Let X and Ω be random variables taking values at $\mathscr{X} \subseteq \mathbb{R}^n$ and $\mathscr{C} = \{\omega_k\}_1^K$ respectively. We say that X and Ω assume the homoscedastic gaussian model (HOG) if

$$\mathsf{X}|\Omega = \omega_k \sim \mathscr{N}(\boldsymbol{\mu}_k, \bar{\Sigma}) \quad \forall k \in \{1,\ldots,K\}$$

i.e. the probability distribution function (PDF) of X given ω_k follows the gaussian distribution, with distinct means but the same covariance matrix for all classes.

Proposition 1. *When X and Ω assume the HOG model, and $m \geqslant K-1$, any LDA optimal matrix is also Bayes optimal.*

For the proof of this proposition we refer the reader to Ref. [19, pp. 87–90]. The proof is derived by considering explicitly the optimal decision boundaries, which are hyperplanes. In fact, not only is it shown that an LDA optimal matrix is Bayes optimal, but it is also shown that the Bayes error in the projected subspace is equal to the Bayes error in the original space.

2.3. The HDA criterion

The homoscedasticity assumption needed for LDA Bayes optimality is quite strict and may not be applicable to real data. In Ref. [11] Kumar and Andreou considered a model in which the classes PDFs are still gaussian, yet they are allowed to have different covariance matrices, under the condition that both means and covariance matrices coincide in a subspace of the observation space. They then followed a maximum likelihood approach and derived a criterion for DLFE to find the matrix that maximizes the probability of this model given the available sample.

Before proceeding, we need to introduce the following notational convention, which will be used throughout the paper: Given any full rank projection matrix T, we will denote by T^{c} a full rank matrix whose columns span the orthogonal complement of the space spanned by the columns of T. We will also denote by $\tilde{T}$ the matrix $\tilde{T} = [TT^{\mathrm{c}}]$, whose

[3] The criterion involves actually the *sample estimates* of means, covariance matrices and class probabilities. Here, we do not consider sample estimation issues, and consider instead the asymptotic limit of big sample size, assuming that the sample estimates of means and covariance matrices converge to the true means and covariance matrices. This assumption will be made also in Section 2.3, when we examine the HDA criterion.

[4] The LDA criterion can be expressed in several other equivalent forms. For more information, see Ref. [21, p. 446].

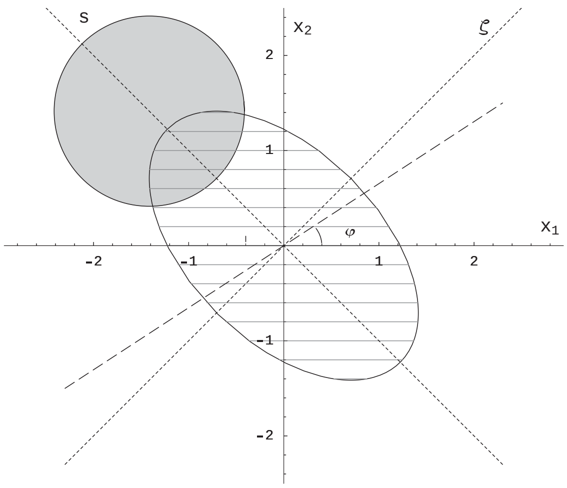

Fig. 1. A two-dimensional, two-class KAH model example. The shaded areas correspond to $(\boldsymbol{x}-\boldsymbol{\mu}_k)^\top \Sigma_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k) \leqslant 1$.

columns span the whole space. Using this convention, we now formally define Kumar and Andreou's heteroscedastic model:

Definition 2 (KAH model). Let X and Ω be random variables taking values at $\mathscr{X} \subseteq \mathbb{R}^n$ and $\mathscr{C} = \{\omega_k\}_1^K$ respectively. We say that X and Ω assume Kumar and Andreou's heteroscedastic model (KAH) if

(1) $\mathsf{X}|\Omega=\omega_k \sim \mathscr{N}(\boldsymbol{\mu}_k, \Sigma_k)$, $\forall k \in \{1,\dots,K\}$, i.e. the PDF of each class follows the gaussian distribution, with distinct mean and covariance matrix (heteroscedasticity assumption) and

(2) There exists an integer d, $d < n$ and a full rank matrix F of dimension $n \times d$, such that, for all the class conditional means $\tilde{\mathrm{F}}^\top \boldsymbol{\mu}_k$ and class conditional covariance matrices $\tilde{\mathrm{F}}^\top \Sigma_k \tilde{\mathrm{F}}$ of $\tilde{\mathrm{F}}^\top \boldsymbol{x}$, it holds

$$\tilde{\mathrm{F}}^\top \boldsymbol{\mu}_k = \begin{bmatrix} \mathrm{F}^\top \boldsymbol{\mu}_k \\ \mathrm{F}^{\mathrm{c}\top} \boldsymbol{\mu} \end{bmatrix} \quad \text{and}$$

$$\tilde{\mathrm{F}}^\top \Sigma_k \tilde{\mathrm{F}} = \begin{bmatrix} \mathrm{F}^\top \Sigma_k \mathrm{F} & 0 \\ 0 & \mathrm{F}^{\mathrm{c}\top} \Sigma \mathrm{F}^{\mathrm{c}} \end{bmatrix}. \tag{9}$$

In words, the assumptions made are that the classes may follow different gaussian distributions but their differences lie solely in a subspace of $\mathbb{R}^n$ of d dimensions, whereas in the complementary subspace of $n-d$ dimensions they are identical, i.e. this latter subspace is useless for distinguishing the classes.

Fig. 1 shows a two-dimensional two-class example where the KAH model holds. The two class conditional PDFs assume different gaussian distributions:

$$\mathscr{N}\left(\begin{bmatrix} -\sqrt{2} \\ \sqrt{2} \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) \quad \text{and}$$

$$\mathscr{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}\right).$$

Notice that the means and covariances along the $\boldsymbol{\zeta}$-axis ($\varphi = \pi/4$) are the same for both classes and therefore the $\boldsymbol{\zeta}$-axis is useless for classification. Thus the KAH model is satisfied with $d = 1$ and

$$B = -\frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix},$$

the $3\pi/4$ rotation matrix.

Kumar and Andreou assume that d is somehow known, so that the DLFE task reduces to extracting $m=d$ features. They construct a criterion to determine, in a maximum likelihood sense, among all full rank matrices A of dimension $n \times d$ and their corresponding A^{c}, the matrices which recover the "useful" and "useless" space for distinguishing the classes. Thus, the HDA criterion requires to find the set $\hat{\mathrm{A}}^{HDA}$ of matrices defined by

$$\hat{\mathrm{A}}^{HDA} = \operatorname*{argmax}_{\mathrm{A}} \left[-\log |\mathrm{A}^{\mathrm{c}\top} \Sigma \mathrm{A}^{\mathrm{c}}| - \sum_{k=1}^{K} p(\omega_k) \log |\mathrm{A}^\top \Sigma_k \mathrm{A}| + 2 \log |\tilde{\mathrm{A}}| \right], \tag{10}$$

where $\tilde{\mathrm{A}} = [\mathrm{A}\mathrm{A}^{\mathrm{c}}]$. Any matrix that yields the maximum in Eq. (10) will be called HDA *optimal matrix*.[5,6]

In Ref. [23], it is shown that, for equal class-covariance matrices, the HDA criterion finds the same subspace as the LDA criterion. However, there has not been any direct proof that the HDA solution is Bayes optimal within the KAH model.

2.4. The MMI criterion

In contrast with LDA and HDA, the MMI criterion for feature extraction has been derived from a different path, using concepts from information theory and their relation with the Bayes error.

Consider first the *entropy* $\mathscr{H}(\Omega)$ as a measure of uncertainty about the class value. Entropy is a functional of the PDF of the class variable, averaged over its possible values, such that complete knowledge of class value (positive probability for only one value) corresponds to zero entropy

[5] In Eq. (10), it is assumed that $|\tilde{\mathrm{A}}|$ is positive. As noticed by Kumar, this is not a real constraint, since we can always consider an equivalent A by multiplying any columns of A with -1 (see Ref. [11]).

[6] In Ref. [22], Schukat-Talamazini et al. derive a special case of criterion (10) in a hidden Markov model context with an additional orthonormality constraint on $\tilde{\mathrm{A}}$.

whereas equal probability for all classes values corresponds to maximum entropy. The most popular form of entropy is the Shannon entropy

$$\mathscr{H}(\Omega) = -\sum_{k=1\ldots K} p(\omega_k)\log[p(\omega_k)]. \tag{11}$$

The knowledge of a feature vector X related somehow to the class reduces our uncertainty about the class. Thus, we define the uncertainty of Ω when X is known, i.e. the feature vector-conditional class entropy, as *equivocation*

$$\begin{aligned}\mathscr{H}(\Omega|\mathsf{X}) &= \mathscr{E}_{\mathsf{X}}[\mathscr{H}(\Omega|\mathsf{X}=\boldsymbol{x})]\\ &= -\mathscr{E}_{\mathsf{X}}\left[\sum_k p(\omega_k|\boldsymbol{x})\log[p(\omega_k|\boldsymbol{x})]\right].\end{aligned} \tag{12}$$

In addition, we define the *gain* in information on Ω by knowledge of X as

$$\mathscr{I}(\Omega;\mathsf{X}) = \mathscr{H}(\Omega) - \mathscr{H}(\Omega|\mathsf{X}). \tag{13}$$

For Shannon entropies, the above quantity has the important property of being symmetrical in its two arguments [24], i.e. it holds that

$$\mathscr{I}(\Omega;\mathsf{X}) = \mathscr{I}(\mathsf{X};\Omega) = \mathscr{H}(\mathsf{X}) - \mathscr{H}(\mathsf{X}|\Omega) \tag{14}$$

and $\mathscr{I}(\Omega;\mathsf{X})$ is commonly known as the *mutual information* between the class and the observation. For a detailed discussion of entropy and mutual information the reader is referred to Ref. [25]. In this paper, we shall keep referring the reader to Ref. [25] to help establish some of the more technical points in the proofs of various propositions.

Since equivocation and mutual information relate the knowledge of X to that of Ω, they can be used to formulate criteria for discriminative feature selection: given a set of features, one should choose a subset that minimizes equivocation or, equivalently, maximizes mutual information. The argument holds also for DLFE. Thus, given a positive integer $m < n$, the MMI criterion requires to find among all possible full rank $n \times m$ matrices the set $\hat{\mathrm{A}}^{HDA}$ of matrices defined by

$$\hat{\mathrm{A}}^{MMI} = \underset{\mathrm{A}}{\operatorname{argmax}}\, \mathscr{I}(\Omega;\mathrm{A}^\top\mathsf{X}). \tag{15}$$

Any matrix $\hat{\mathrm{A}}^{MMI}$ that yields the maximum in Eq. (15) will be called *MMI optimal matrix*.

It has been shown that mutual information between the class and the observation is related to the Bayes error of the class given the observation, via lower and upper bounds (see Ref. [13]). For the DLFE case, these bounds become[7]

$$\frac{\mathscr{H}(\Omega) - \mathscr{I}(\Omega;\mathrm{A}^\top\mathsf{X}) - 1}{\log(K-1)} \leqslant P_e(\Omega|\mathrm{A}^\top\mathsf{X}) \leqslant \frac{\mathscr{H}(\Omega) - \mathscr{I}(\Omega;\mathrm{A}^\top\mathsf{X})}{2}. \tag{16}$$

To conclude this section, let us summarize the properties of the criteria: The LDA criterion has been defined as a separability measure and is Bayes optimal under the HOG model. The HDA criterion is derived by applying the maximum likelihood principle to the KAH model. For equal covariance matrices it reduces to the LDA criterion and is therefore also Bayes optimal under the HOG model, but there is no direct proof of its Bayes optimality within any more general model. The MMI criterion originates from an information theoretic point of view and is connected to the Bayes error only via lower and upper bounds. The remainder of this article is concerned with creating a bridge between the criteria, by giving answers to the following questions:

- Are there any conditions under which Bayes optimality can be guaranteed for HDA and MMI?
- Are there any conditions under which the DLFE subspaces extracted by two or more of the above criteria coincide?

3. The ZIL model and conditions for Bayes optimality of the MMI criterion

In this section we examine sufficient conditions under which the MMI optimal matrix is also Bayes optimal. To this end, we propose to view the DLFE problem as a source recovering problem, and impose constraints on the probability densities, so that Bayes optimality of the MMI optimal matrix can be guaranteed. The results of this section are used later in Section 5, where the relationship among DLFE criteria is investigated.

The DLFE task, as defined in Section 2.1 is viewed as finding a suitable matrix which transforms the original observation vector $\boldsymbol{x}$ to an observation vector of reduced dimensionality. However, it is possible to view DLFE from a reverse point of view. Let us assume that the observation vector $\boldsymbol{x}$ is in fact the result of linearly mixing up a "source" vector and a "noise" vector by a non-singular $n \times n$ matrix B, as

$$\boldsymbol{x} = \mathrm{B}^\top \begin{bmatrix} \boldsymbol{s} \\ \boldsymbol{\zeta} \end{bmatrix}.$$

The source vector $\boldsymbol{s} \in \mathbb{R}^d$ contains the information useful for identifying classes. On the other hand, the noise vector, $\boldsymbol{\zeta} \in \mathbb{R}^{(n-d)}$ contains information that is *redundant* for the classification, when the source vector is known. In probabilistic terms, with $\boldsymbol{s}$, $\boldsymbol{\zeta}$, ω_k considered as instances of random variables S, Z and Ω, we can formulate this assumption as follows:

$$p(\omega_k|\boldsymbol{s},\boldsymbol{\zeta}) = p(\omega_k|\boldsymbol{s})$$

for all $\boldsymbol{s}$, $\boldsymbol{\zeta}$ and ω_k, i.e. the conditional probability of each class given both the source and the noise vector equals its conditional probability given solely the source vector. This requirement makes sense, since the common implicit assumption in DLFE is that there is part of the observation

[7] For a tighter lower bound, known as Fano's bound, as well as for a general discussion of bounds for the probability of error, see Ref. [13].

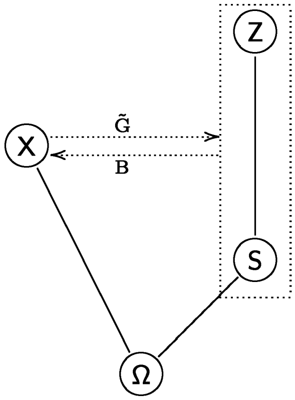

Fig. 2. The dependence graph between the variables. Straight lines denote probabilistic dependence, whereas the dashed line functional dependence. Notice that Z is related to Ω only through S.

vector lying in a *noise* subspace, the knowledge of which does not change the class probability when the other part, which lies in the *source* subspace, is known.

When B is known, we can evaluate its inverse $\tilde{\mathrm{G}} = \mathrm{B}^{-1}$ and recover complete knowledge of S and Z from X. However, the mixing matrix is not known, and the DLFE task can be viewed as estimating the un-mixing matrix $\tilde{\mathrm{G}}$ using solely the probability assumptions about the source and noise vectors.

Based on the above rationale, we now formally define the proposed model, under the name of Zero Information Loss model. The name stems from the fact that, as it will be proved shortly, mutual information with the class is not lost if only the source vector is retained.

Definition 3 (ZIL model). Let X and Ω be random variables taking values at $\mathscr{X} \subseteq \mathbb{R}^n$ and $\mathscr{C} = \{\omega_k\}_1^K$ respectively. We say that X and Ω assume the zero information loss (ZIL) model, if there exist

(1) an integer d with $d < n$ and
(2) a full rank matrix G of dimension $n \times d$,

such that for the vectors $\boldsymbol{s} = \mathrm{G}^\top \boldsymbol{x}$ (source vector) and $\boldsymbol{\zeta} = \mathrm{G}^{\mathrm{c}\top} \boldsymbol{x}$ (noise vector), considered as instances of the random variables S and Z respectively, it holds that

$$p(\omega_k|\boldsymbol{s},\boldsymbol{\zeta}) = p(\omega_k|\boldsymbol{s}), \qquad \forall \mathrm{X} \in \mathscr{X},\ \omega_k \in \mathscr{C}. \tag{17}$$

The subspaces of $\mathbb{R}^n$ spanned by the columns of G and G^{c} will be termed source subspace and noise subspace respectively.

Fig. 2 shows a graph representing the relations that hold between the random variables involved in the ZIL model. First, the observation, X, is connected in probability with the class, which reflects the fact that knowing the observation changes the probabilities for the class. The same holds for the source, S. However, the noise, Z, is only indirectly connected with the class, through the source, i.e. although knowing the noise may be pertinent for the class, this knowledge is already contained in the source. Moreover, the direct connection between the source and the noise reflects the fact that these variables are not necessarily independent, i.e. they may be jointly distributed. Finally, the dotted line represents the functional relation that exists between the observation, on the one hand, and the source and the noise on the other, through the mixing/unmixing matrices B and $\tilde{\mathrm{G}}$.

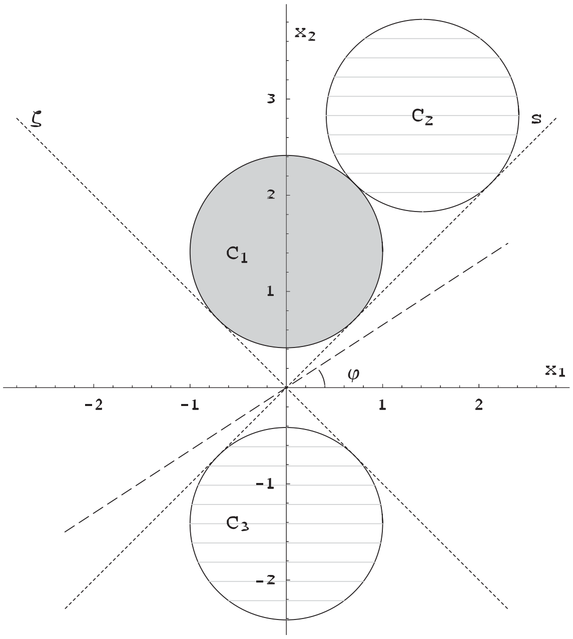

Fig. 3. A two-dimensional, two-class ZIL model example.

To illustrate the ZIL model, a two-dimensional, two-class setting is shown in Fig. 3. The observation vector assumes uniform distribution inside three unit-radius circles, centered at

$$\begin{bmatrix} 0 \\ \sqrt{2} \end{bmatrix}, \begin{bmatrix} \sqrt{2} \\ 2\sqrt{2} \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 0 \\ -\sqrt{2} \end{bmatrix},$$

denoted by C_1, C_2 and C_3 respectively. The probability of the observation outside the circles is null and thus we may take $\mathscr{X} = C_1 \cup C_2 \cup C_3$ and $p(\boldsymbol{x}) = 1/3\pi$, $\forall \boldsymbol{x} \in \mathscr{X}$. Moreover, the probability of the first class given the observation vector is

$$p(\omega_1|\boldsymbol{x}) = \begin{cases} 1 & \forall \boldsymbol{x} \in C_1, \\ 0 & \forall \boldsymbol{x} \in C_2 \cup C_3 \end{cases}$$

and reversely for the second class. Now, to see that the ZIL model holds, consider the $\boldsymbol{s}$ and $\boldsymbol{\zeta}$ axes drawn at angles $\varphi = \pi/4$ and $\varphi = 3\pi/4$, respectively, and the projections of the circles along these axes, $C_1^s = (0,2)$, $C_1^\zeta = (0,2)$, $C_2^s = (2,4)$, $C_2^\zeta = (0,2)$, $C_3^s = (-2,0)$, and $C_3^\zeta = (-2,0)$. Note that, in this example, $\boldsymbol{s}$ and $\boldsymbol{\zeta}$ are jointly distributed. It can easily be seen that, along the $\boldsymbol{s}$-axis,

$$p(\omega_1|\boldsymbol{s}) = \begin{cases} 1 & \forall \boldsymbol{s} \in C_1^s, \\ 0 & \forall \boldsymbol{s} \in C_2^s \cup C_3^s \end{cases}$$

and reversely for the second class, which implies that $p(\omega_k|\boldsymbol{x}) = p(\omega_k|\boldsymbol{s})$, $\forall \boldsymbol{x} \in \mathscr{X}$, $k = 1, 2$, and thus the ZIL model holds with $d = 1$ and

$$B = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix},$$

i.e. the $\pi/4$ rotation matrix. In addition, notice that the same does not hold for $\boldsymbol{\zeta}$, since the projections of clusters C_1 and C_2 completely overlap, giving

$$p(\omega_1|\boldsymbol{s}) = \begin{cases} \frac{1}{2} & \forall \boldsymbol{\zeta} \in C_1^\zeta \cup C_2^\zeta, \\ 0 & \forall \boldsymbol{\zeta} \in C_3^\zeta \end{cases}$$

i.e. $p(\omega_1|\boldsymbol{x}) \neq p(\omega_1|\boldsymbol{\zeta})$.

Another ZIL model example will be given in Section 6.

We now prove the following fundamental properties of the ZIL model, which will help us examine the issue of Bayes optimality of the MMI and other DLFE criteria within this model:

Lemma 1. *Let* X *and* Ω *conform to the ZIL model for source space dimension equal to d. Then* X *and* Ω *conform to the ZIL model for any* d', $d < d' \leqslant n$.

Lemma 2. *Under the ZIL model*,

(1) $P_e(\Omega|\mathsf{X}) = P_e(\Omega|\mathsf{S})$
(2) $\mathscr{I}(\Omega;\mathsf{X}) = \mathscr{I}(\Omega;\mathsf{S})$

Lemma 3. *Consider a classification problem that complies to the assumptions of the ZIL model. Given an integer* $m \geqslant d$, *let* A *be an* $n \times m$ *full rank matrix. Then*, A *is MMI optimal if and only if*

$$p(\omega_k|\mathrm{A}^\top \boldsymbol{x}, \mathrm{A}^{\mathrm{c}\top} \boldsymbol{x}) = p(\omega_k|\mathrm{A}^\top \boldsymbol{x}) \quad \forall \boldsymbol{x} \in \mathscr{X},\ \omega_k \in \mathscr{C}. \tag{18}$$

The proofs are given in Appendices A.1, A.2 and A.3.

Note that Lemma 2 means that performing classification keeping solely the source vector does not affect either the Bayes error or mutual information. This assertion justifies the name given to the model (zero information loss).

As it has been said earlier in Section 2.4, the MMI criterion is related to the Bayes error only by means of lower and upper bounds. This implies that, in general, an optimal MMI extracted subspace is not guaranteed to be Bayes optimal. However, we are now in a position to show that under the ZIL model, any MMI optimal matrix yields minimum Bayes error:

Proposition 2. *Let* X *and* Ω *assume the ZIL model with source subspace dimension d. Then any MMI optimal matrix* A *of dimension* $n \times m$, $m \geqslant d$ *is also Bayes optimal.*

Proof. By Lemma 2, the Bayes error using any vector that complies to the source vector assumption of the ZIL model equals the Bayes error using the whole observation vector, and hence it is minimum. Moreover, by Lemma 3, the MMI optimal matrices recover only such vectors. It follows that any MMI optimal matrix is also Bayes optimal. □

4. Formal similarity of DLFE criteria

From the discourse in Section 2 it is evident that the LDA, HDA and MMI criteria are derived using different rationales and result in different mathematical forms. The aim of this section is to put these criteria in a mathematical form that facilitates their comparison and elucidates the points in which they differ.

The results of this section which are summarized in Proposition 3, together with the results of Section 3 will be used in Section 5 to investigate a deeper similarity of the criteria. Before proceeding to the main proposition of this section, however, we need to introduce the concepts of gaussian entropy and negentropy.

To begin, consider a continuous n-dimensional random variable following the gaussian distribution $\mathscr{N}(\boldsymbol{\mu}, \Sigma)$. It can easily be seen that its entropy is given by $\frac{1}{2}\log(2\pi\mathrm{e})^n|\Sigma|$. The reader is referred to Ref. [25, p. 230] for a proof. Now, consider an n-dimensional continuous random variable X, with covariance Σ.

- The *gaussian entropy* of $\mathsf{X}, \mathscr{H}_g(\mathsf{X})$, is the entropy of a gaussian variable with the same covariance matrix as X:
$$\mathscr{H}_g(\mathsf{X}) = \frac{1}{2}\log(2\pi\mathrm{e})^n|\Sigma|. \tag{19}$$
- The *negentropy* of X, $\mathscr{J}(\mathsf{X})$, is the difference of the entropy of X and its gaussian entropy:
$$\mathscr{J}(\mathsf{X}) = \mathscr{H}_g(\mathsf{X}) - \mathscr{H}(\mathsf{X}). \tag{20}$$

It can be shown [26] that negentropy is always a positive quantity and vanishes when the random variable is gaussian. Negentropy may be viewed as a measure of non-gaussianity and has thus been extensively used in the ICA literature (see Ref. [26]). Here we emphasize the fact that these definitions allow for a split of the entropy in two terms: the gaussian term, which depends on the covariance matrix, and the negentropy term which may be attributed to higher order statistics.

The concepts of gaussian entropy and negentropy can be naturally extended to account for the class conditional case, as follows:

- The *conditional gaussian entropy* of $\boldsymbol{x}$ given $\Omega = \omega_k$ and the *average conditional gaussian entropy* of $\boldsymbol{x}$ given Ω are
$$\mathscr{H}_g(\mathsf{X}|\omega_k) = \frac{1}{2}\log(2\pi\mathrm{e})^n|\Sigma_k| \quad \text{and} \tag{21}$$
$$\mathscr{H}_g(\mathsf{X}|\Omega) = \sum_k p(\omega_k)\mathscr{H}_g(\mathsf{X}|\omega_k). \tag{22}$$

- The *conditional negentropy* of $\boldsymbol{x}$ given $\Omega = \omega_k$ and the *average conditional negentropy* of $\boldsymbol{x}$ given Ω are

$$\mathscr{J}(\mathsf{X}|\omega_k) = \mathscr{H}_g(\mathsf{X}|\omega_k) - \mathscr{H}(\mathsf{X}|\omega_k), \quad \text{and} \tag{23}$$

$$\mathscr{J}(\mathsf{X}|\Omega) = \sum_k p(\omega_k)\mathscr{J}(\mathsf{X}|\omega_k). \tag{24}$$

It is straightforward to show that for the conditional case, as in Eq. (20), it holds that

$$\mathscr{J}(\mathsf{X}|\Omega) = \mathscr{H}_g(\mathsf{X}|\Omega) - \mathscr{H}(\mathsf{X}|\Omega). \tag{25}$$

We now proceed to the following proposition which reformulates the three DLFE criteria using similar mathematical formulas:

Proposition 3. *Let* X *be a continuous random variable taking values from* $\mathscr{X} \subseteq \mathbb{R}^n$ *and* Ω *be a discrete random variable taking values from* $\mathscr{C} = \{\omega_k\}_1^K$. *Let* Σ *be the covariance matrix of* X *and* Σ_k *be the covariance matrix of* X *given the class* ω_k. *Given a positive integer m with* $m < n$, *let* A *be a full rank matrix of dimension* $n \times m$. *Then, the sets of LDA, HDA and MMI optimal matrices can be found as*

$$\hat{\mathrm{A}}^{LDA} = \underset{\mathrm{A}}{\operatorname{argmax}} \left[\log \frac{|\mathrm{A}^\top \Sigma \mathrm{A}|}{\sum_{k=1}^K p(\omega_k)|\mathrm{A}^\top \Sigma_k \mathrm{A}|} \right], \tag{26}$$

$$\hat{\mathrm{A}}^{HDA} = \underset{\mathrm{A}}{\operatorname{argmax}} \left[\log \frac{|\mathrm{A}^\top \Sigma \mathrm{A}|}{\prod_{k=1}^K |\mathrm{A}^\top \Sigma_k \mathrm{A}|^{p(\omega_k)}} \right], \tag{27}$$

$$\hat{\mathrm{A}}^{MMI} = \underset{\mathrm{A}}{\operatorname{argmax}} \left[\log \frac{|\mathrm{A}^\top \Sigma \mathrm{A}|}{\prod_{k=1}^K |\mathrm{A}^\top \Sigma_k \mathrm{A}|^{p(\omega_k)}} - 2(\mathscr{J}(\mathrm{A}^\top \mathsf{X}) - \mathscr{J}(\mathrm{A}^\top \mathsf{X}|\Omega)) \right]. \tag{28}$$

The proof is given in Appendix A.4. We stress the fact that the formulas in Proposition 3 are quite general and hold independently of the model assumed for the underlying data. The formal similarity is evident: HDA differs from LDA only in the aggregation operator (weighted product vs weighted sum), whereas MMI differs from HDA only in the additional negentropy terms. [8]

5. Relation of models and equivalence of DLFE criteria

In this section we explore a deeper relation that exists between the LDA, HDA and MMI criteria. The connection between the criteria is established by relating the models under which sufficient conditions for Bayes optimality hold.This allows us to re-derive the LDA criterion based on the MMI criterion and to prove the Bayes optimality of the HDA criterion under the KAH model.

5.1. Relations between the ZIL, KAH and HOG models

The following proposition asserts that the KAH model is a special case of the ZIL model and that the HOG model is a special case of the KAH model. This creates an “hierarchy” of models, summarized schematically in Fig. 4.

Proposition 4. *Consider the HOG, KAH and ZIL models, as defined in* Definitions 1, 2, *and* 3 *respectively.*

- *If the classification problem conforms to the assumptions of the KAH model, it also conforms to the assumptions of the ZIL model. The unmixing matrix is obtained by* $\mathrm{G} = \tilde{\mathrm{F}}$ *while the source and noise vectors are obtained by* $\boldsymbol{s} = \mathrm{F}^\top \boldsymbol{x}$ *and* $\boldsymbol{\zeta} = \mathrm{F}^{c\top} \boldsymbol{x}$ *respectively.*
- *If the classification problem conforms to the assumptions of the HOG model, it also conforms to the assumptions of the KAH model with the source space dimension d equal to* $\min(K-1, n)$.

The proof is given in Appendix A.5.

Summarizing the proof, we can say that the KAH model is a ZIL model with two additional constraints: (a) both the source vector and the noise vector are gaussian and (b) the noise vector is also completely independent (as well as source-conditionally independent) of the class. Also, as it is evident from Ref. [19], the HOG model may be viewed as a KAH model with two additional constraints: (a) the class conditional covariance matrices are equal, and (b) the number of classes is at most $d+1$. These identifications allow us to speak of “source” and “noise” vectors and the corresponding “source” and “noise” subspaces in connection with the KAH and HOG models.

5.2. Equivalence and Bayes optimality of feature extraction criteria

Given the similarity between the formal expressions of the three DLFE criteria established by Proposition 3, a question that naturally arises is the following: Are there any conditions under which these criteria do actually coincide? In this section, we show that the criteria do coincide under certain assumptions about the models obeyed by the classification data. Moreover, the issue of equivalence of two or more criteria is closely related with the issue of Bayes optimality.

First, we examine the MMI criterion under the KAH model. However, before presenting the basic result, we prove the following lemma:

[8] A one-dimensional form of Eq. (27) is reported in Ref. [27]. Moreover, in Ref. [28], a criterion of similar form, though not the same, has been proposed, based on heuristic grounds, where instead of the overall covariance matrix, the between class covariance matrix is used.

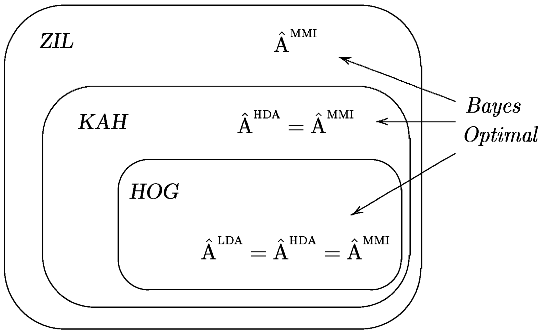
ZIL
$\hat{A}^{MMI}$
KAH
$\hat{A}^{HDA} = \hat{A}^{MMI}$
HOG
$\hat{A}^{LDA} = \hat{A}^{HDA} = \hat{A}^{MMI}$
Bayes Optimal

Fig. 4. This figure summarises most of the results of this paper. It shows the hierarchy of models (for a specific value of the source subspace dimensionality, the HOG model is a special case of the KAH model which in turn is a special case of the ZIL model). It also depicts the equivalence and Bayes optimality of DLFE criteria under the various models, when the number of the extracted features is greater or equal to the source subspace dimension.

Lemma 4. *Consider a classification problem conforming to the KAH model and the corresponding source vector* $\mathsf{S}=\mathrm{F}^\top\mathsf{X}$. *Then, for any matrix* A *of dimension* $n\times m$, $m \leqslant n$, *it holds*

$$\mathscr{J}(\mathsf{S}) \geqslant \mathscr{J}(\mathrm{A}^\top\mathsf{X}) \tag{29}$$

i.e. the source vector has maximum negentropy among all projected vectors.

The proof is given in Appendix A.6. Using this lemma, we are now ready to show the following

Proposition 5. *Under the KAH model, when* $m \geqslant d$, *the MMI criterion reduces to the HDA criterion. Moreover, the MMI and HDA optimal matrices are also Bayes optimal.*

Proof. First, notice that when the class-conditional PDFs are gaussian, the conditional negentropies of the observation given the class vanish, so that the MMI criterion becomes

$$\hat{\mathrm{A}}^{MMI} = \underset{\mathrm{A}}{\operatorname{argmax}}\, f_{\mathrm{MMI}}(\mathrm{A}^\top\mathsf{X}), \tag{30}$$

where

$$f_{\mathrm{MMI}}(\mathrm{A}^\top\mathsf{X}) = f_{\mathrm{HDA}}(\mathrm{A}^\top\mathsf{X}) - 2\mathscr{J}(\mathrm{A}^\top\mathsf{X}) \tag{31}$$

and

$$f_{\mathrm{HDA}}(\mathrm{A}^\top\mathsf{X}) = \log\frac{|\mathrm{A}^\top\Sigma\mathrm{A}|}{\prod_{k=1}^{K}|\mathrm{A}^\top\Sigma_k\mathrm{A}|^{p(\omega_k)}}. \tag{32}$$

With a minor rearrangement, we obtain:

$$f_{\mathrm{HDA}}(\mathrm{A}^\top\mathsf{X}) = f_{\mathrm{MMI}}(\mathrm{A}^\top\mathsf{X}) + 2\mathscr{J}(\mathrm{A}^\top\mathsf{X}) \tag{33}$$

from where it can be seen that $f_{\mathrm{HDA}}(\mathrm{A}^\top\mathsf{X})$ is maximized if and only if $f_{\mathrm{MMI}}(\mathrm{A}^\top\mathsf{X}) + 2\mathscr{J}(\mathrm{A}^\top\mathsf{X})$ is maximized.

Now, by Lemma 3, it is known that, under the ZIL model, which includes the KAH model, and when $m \geqslant d$, $f_{\mathrm{MMI}}(\mathrm{A}^\top\mathsf{X})$ is maximized if and only if A recovers vectors that comply to the source vector assumption. On the other hand, Lemma 4 asserts that, for $m = d$, such vectors also maximise the negentropy. Moreover, by Lemma 1, this will also hold for any $m \geqslant d$.

This implies that $f_{\mathrm{MMI}}(\mathrm{A}^\top\mathsf{X}) + 2\mathscr{J}(\mathrm{A}^\top\mathsf{X})$, and thus $f_{\mathrm{HDA}}(\mathrm{A}^\top\mathsf{X})$, is also maximized if and only if A recovers vectors that comply to the source vector assumption. Hence, under the KAH model, to find the matrices that maximize the MMI criterion, one can ignore $\mathscr{J}(\mathrm{A}^\top\mathsf{X})$ and find the matrices that maximize solely $f_{\mathrm{HDA}}(\mathrm{A}^\top\mathsf{X})$. Therefore, under the KAH model, the MMI and HDA criteria are equivalent.

Furthermore, we know that the MMI criterion yields a Bayes optimal matrix under the ZIL model (Proposition 2). Since the KAH model is a special case of the ZIL model (Proposition 4), MMI also yields a Bayes-optimal matrix under the KAH model. Since MMI reduces to the HDA criterion, as concluded above, HDA also yields a Bayes-optimal matrix under the KAH model. □

We now examine the relation between HDA and LDA under the HOG model.

Proposition 6. *Under the HOG model, when* $m \geqslant K-1$, *the MMI and HDA criteria reduce to the LDA criterion. Moreover, the MMI, HDA and LDA optimal matrices are also Bayes optimal.*

Proof. First we show that the HDA criterion reduces to the LDA criterion under the HOG model. In fact, this has already been proved by Kumar [11]. Still, we present here a direct proof, based on the formal similarity of the criteria. Under the HOG model, all class-conditional covariance matrices are equal, i.e. $\Sigma_k = \bar{\Sigma}\ \forall k \in \{1\ldots K\}$, and hence HDA (as given in Eq. (27)) directly simplifies to

$$\underset{\mathrm{A}}{\operatorname{argmax}}\,\log\frac{|\mathrm{A}^\top\Sigma\mathrm{A}|}{|\mathrm{A}^\top\bar{\Sigma}\mathrm{A}|}. \tag{34}$$

However, when all class-conditional covariance matrices are equal, LDA, as given in Eq. (26), also simplifies to Eq. (34). Hence, HDA and LDA are equivalent. Moreover, by Proposition 5, HDA recovers Bayes-optimal matrices under the KAH model and hence (by Proposition 4) also under the HOG model.

To conclude the proof, notice that MMI is equivalent to HDA under the KAH model, and thus also under the HOG model. As a result, in the HOG model, MMI is equivalent to LDA and recovers Bayes optimal matrices. □

Fig. 4 summarizes the results of Propositions 1, 2, 4, 5 and 6 by showing the hierarchy of the models as well as the equivalence and Bayes optimality of the criteria.

We conclude this subsection by stressing that Propositions 5 and 6 have a very important interpretation: applying the MMI criterion to the KAH model we end up with the HDA criterion, which simplifies to the LDA criterion under the

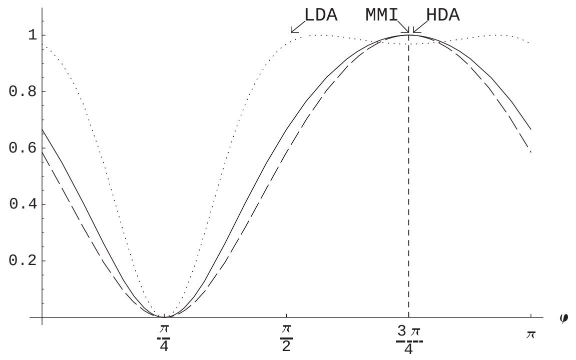

Fig. 5. Graph for the example of Fig. 1. The curves show the values of the three criteria as functions of the projection angle. Each curve has been rescaled between 0 and 1 to facilitate visualization. The solid line corresponds to the MMI criterion, the dotted line to the HDA criterion and the dashed line to the LDA criterion. The angles selected by the criteria correspond to the points where the curves exhibit their maxima.

homoscedasticity assumption. Hence *HDA and LDA can be derived directly from an information theoretic path.*

6. Examples

To illustrate the similarities and differences of LDA, HDA and MMI we apply them in a series of two-dimensional two-class examples. In all cases, the objective of the criteria is to find a one-dimensional projection, such that the two classes, given the projection, are best separated. For easiness, we will refer to the projection by the angle of the projection axis relative to the x_1 axis, φ, $\varphi \in [0, 2\pi]$. The evaluation of the criteria has been carried out numerically using the distributions and the class conditional distributions of the observations.

Example 1. Consider the classification problem illustrated in Fig. 1. As discussed in Section 2.3, this problem conforms to the KAH model with $d = 1$ and source projection at $\varphi = 3\pi/4$. In this problem, the Bayes optimal angle is also $\varphi = 3\pi/4$, because the Bayes error for this projection is equal to the Bayes error for the whole plane. This can be easily verified either by direct evaluation of the marginal probabilities, or by applying Lemma 2.

According to Proposition 5, the HDA and MMI criteria are expected to find the same Bayes optimal angle. On the other hand, LDA is not guaranteed to find the optimal projection, since the covariance matrices of the classes are different, and thus the HOG model is not satisfied. Indeed, as is shown in the graph of Fig. 5, both HDA and MMI succeed in finding the Bayes optimal angle, $\varphi = 3\pi/4$, whereas LDA fails, giving maxima at $\varphi \sim 0.54\pi$ and $\varphi \sim 0.96\pi$.

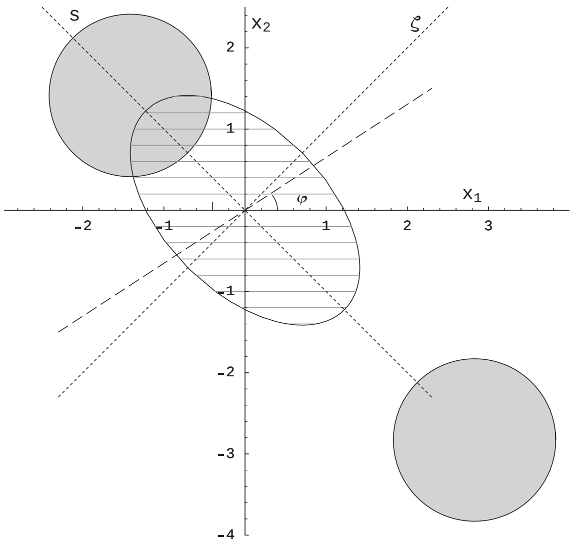

Fig. 6. A two-dimensional, two-class ZIL model example.

Example 2. As a second example, consider the classification problem shown in Fig. 6. This is similar to the problem of Fig. 1, but now the second class is split into two equiprobable clusters, the extra cluster following the gaussian distribution

$$\mathcal{N}\left(\begin{bmatrix} 2\sqrt{2} \\ -2\sqrt{2} \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right).$$

Since not all the class conditional observations assume gaussian distributions, the KAH model does not hold any more. However, the ZIL model still holds. This can be verified as follows: First, along the source and noise axes, which are the same as before, $p(\boldsymbol{s}, \boldsymbol{\zeta}) = p(\boldsymbol{s})p(\boldsymbol{\zeta})$, i.e. the source and the noise are independent. The independence statement holds also for each class separately, and, since the distribution of the observation along the $\boldsymbol{\zeta}$ projection is the same for all classes, it follows that $p(\boldsymbol{s}, \boldsymbol{\zeta}|\omega) = p(\boldsymbol{s}|\omega)p(\boldsymbol{\zeta})$. This leads directly to the probability assumption of the ZIL model (cf. also the path followed in the proof of Lemma 4, Eqs. (A.35)–(A.37)).

According to Proposition 2 the MMI criterion is expected to find the Bayes optimal angle, which remains $\varphi = 3\pi/4$. On the other hand, LDA and HDA are not guaranteed to find the optimal projection, since the distributions are not gaussian, and thus the KAH and HOG model are not satisfied. As seen from the graph in Fig. 7, while the MMI criterion indeed succeeds in finding the Bayes optimal angle, the HDA and LDA criteria fail.

Example 3. Consider now the classification problem of Fig. 3. As discussed in Section 3, the ZIL model holds with $d = 1$ and source subspace at $\varphi = \pi/4$. From the figure, one can easily see that the classes are perfectly separated

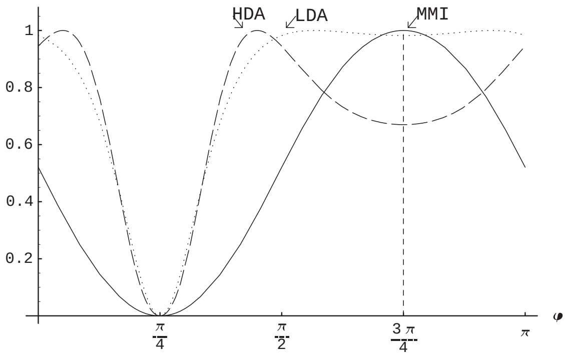

Fig. 7. Values of the criteria for the example of Fig. 6.

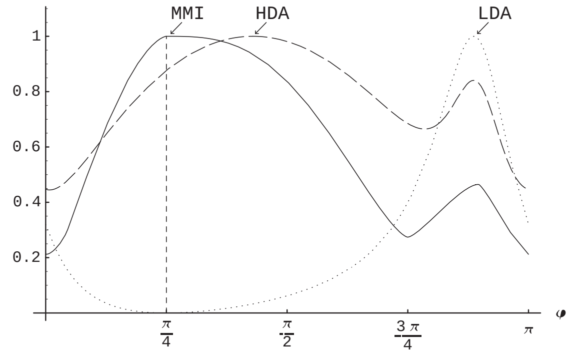

Fig. 8. Graph for the example of Fig. 3. Notice that only MMI is maximized at angle $\varphi = \pi/4$, where Bayes optimality holds.

along the $\boldsymbol{s}$-axis, and therefore the Bayes optimal angle is $\varphi = \pi/4$. Notice also that neither the KAH nor the HOG model conditions are satisfied, since the class conditional distributions are not gaussians.

According to Proposition 2, only the MMI criterion is guaranteed to be optimal. Indeed, as is shown in Fig. 8 the MMI criterion succeeds in finding the Bayes optimal projection, $\varphi = \pi/4$, whereas HDA and LDA fail, giving maxima at $\varphi \sim 0.43\pi$ and $\varphi \sim 0.88\pi$ respectively.

Example 4. Propositions 2, 5 and 6 give sufficient conditions for Bayes optimality of the three DLFE criteria. However, these conditions are not *necessary*. Hence, it should not be induced that LDA, respectively HDA, will fail whenever the HOG, respectively KAH, model is not satisfied, nor that MMI is guaranteed to be superior to HDA and LDA outside the ZIL model.

To illustrate this fact, consider, as a last case, the following variation of the classification setting of Fig. 1: the second class assumes the gaussian distribution

$$\mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \frac{1}{2}\begin{bmatrix}3 & -1\\-1 & 3\end{bmatrix}\right),$$

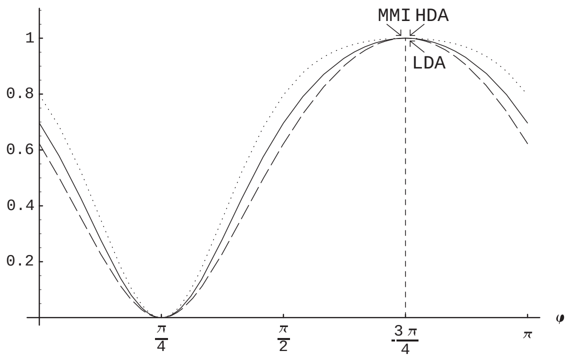

Fig. 9. Graph for the variation of Fig. 1. Notice that all criteria exhibit maxima at the Bayes optimal angle ($\varphi = 3\pi/4$).

i.e. its variance is still stressed along the $\boldsymbol{s}$-axis but less than before. It can be verified that the KAH, and hence the ZIL, model hold, whereas the HOG does not, since the covariance matrices of the two classes are still not the same. However, as shown in Fig. 9, this time, all three methods succeed in finding the optimum, which shows that the LDA criterion may be Bayes optimal outside the HOG model.

7. Discussion and prospects

7.1. LDA as a special case of ICA

Independent component analysis (ICA) is an unsupervised method for source separation relying on higher order statistics [26]. ICA results also in a linear transform, requiring that the discovered sources be maximally independent. It has been shown that this amounts to finding sources with maximum negentropy (see, for instance, Ref. [29]). The basic assumption of ICA is that all sources are non-gaussian, except possibly one. Moreover, an extension of ICA, *multidimensional independent component analysis* [20], allows for sources to be multidimensional.

Looking back at Lemma 4, notice that to recover the source space, one may as well seek the subspace that maximizes negentropy. This is actually equivalent to performing multidimensional ICA distinguishing between the source subspace and the gaussian subspace. This result may be also viewed as a transition point between supervised and unsupervised feature extraction: *should the classes assume gaussian distributions*, *we do not need supervised training to recover the source subspace*. However, one should be aware that this only holds if classes PDFs are also gaussian in the noise subspace, i.e. the noise is gaussian.

7.2. Bounds for suboptimal LDA

The propositions on the Bayes optimality of the LDA, HDA and MMI criteria are relying on the assumption of a *totally noisy* subspace, i.e. a subspace that does not further reduce the probability of misclassification, given the complementary *source* space. When such a space does not exist, or its dimensionality is smaller than the one we seek (i.e. the number of linear features we wish to extract is less than the number of features that contain all the discriminatory information), these criteria are not guaranteed to be Bayes optimal. A question that arises, then, is how suboptimal are the extracted features with respect to the optimal ones.

In particular, under the HOG model, LDA is known to be suboptimal when extracting less than $K-1$ features [19, p. 92]. By the same arguments that led to the formal similarity of the criteria (see Section 4), we may derive bounds for LDA, using the bounds that hold for the MMI criterion, as discussed in Section 2.4. Under the HOG model,

$$\mathscr{H}(\mathrm{A}^\top \mathsf{X}|\Omega) = \mathscr{H}_g(\mathrm{A}^\top \mathsf{X}|\Omega) = \frac{1}{2}\log(2\pi\mathrm{e})^m|\mathrm{A}^\top\bar{\Sigma}\mathrm{A}|,$$

$$\mathscr{H}(\mathrm{A}^\top \mathsf{X}) = \mathscr{H}_g(\mathrm{A}^\top \mathsf{X}) - \mathscr{J}(\mathrm{A}^\top \mathsf{X}) = \frac{1}{2}\log(2\pi\mathrm{e})^m|\mathrm{A}^\top\Sigma\mathrm{A}|$$

and hence

$$\mathscr{I}(\Omega,\mathsf{X}) = \frac{1}{2}\log(2\pi\mathrm{e})^m\frac{|\mathrm{A}^\top\bar{\Sigma}\mathrm{A}|}{|\mathrm{A}^\top\Sigma\mathrm{A}|} - \mathscr{J}(\mathrm{A}^\top \mathsf{X}). \tag{35}$$

However,

$$\frac{|\mathrm{A}^\top\bar{\Sigma}\mathrm{A}|}{|\mathrm{A}^\top\Sigma\mathrm{A}|} = \prod_{i=1}^{m}\lambda_i,$$

where λ_i are the m larger generalized eigenvalues of $\bar{\Sigma}^{-1}\Sigma$ (see Ref. [21, p. 449]) and thus, bounds (16) become

$$\frac{\mathscr{H}(\Omega) - 1/2\sum_{i=1}^{m}[\log(2\pi\mathrm{e})^m\lambda_i] + \mathscr{J}(\mathrm{A}^\top \mathsf{X}) - 1}{\log(K-1)} \leqslant P_e(\Omega|\mathrm{A}^\top \mathsf{X}) \leqslant \frac{\mathscr{H}(\Omega) - 1/2\sum_{i=1}^{m}[\log(2\pi\mathrm{e})^m\lambda_i] + \mathscr{J}(\mathrm{A}^\top \mathsf{X})}{2}. \tag{36}$$

These bounds provide an estimate of the Bayes error in the extracted subspace. Concerning the negentropy term, notice that there is no analytical expression in general to evaluate $\mathscr{J}(\mathrm{A}^\top \mathsf{X})$. However, $\mathrm{A}^\top \mathsf{X}$ follows a mixture of gaussian distributions and, therefore, we conjecture that it can be approximated based on second order statistics of the class conditional distributions.

7.3. Prospects

Beyond the equivalence of the MMI criterion with HDA and LDA under the KAH and HOG models respectively, a "bottom line" question is which criterion to choose for linear feature extraction, when we have no knowledge about the distributions of the classes. According to Proposition 5, any method based on MMI can never be worse than LDA or HDA, when the ZIL model holds. However, one should be aware of two caveats. (Table 1).

First, in a practical setting, the search for the maxima of the criteria has to be carried out using some optimization technique *and* the functions have to be estimated from a limited size sample set. Consequently, robustness or complexity issues have to be considered before choosing the criterion to apply and it may turn out in the end that LDA is a preferable criterion because of its simplicity.

Second, the results of this article do not guarantee that the theoretical superiority of MMI against HDA and the superiority of HDA against LDA hold outside the ZIL and KAH model respectively: our work has resulted in sufficient, but not necessary, conditions for criteria equivalence and optimality. It could be claimed that MMI will continue to have a theoretical advantage, since it exploits much more than first and second order statistics. This claim has been supported by experimental evidence (see Refs. [11,18]), where both HDA and MMI do compare favorably to LDA. Still, work similar to ours that would yield necessary conditions for Bayes optimality and equivalence of DLFE criteria would be a very significant step for gaining important insight into the foundations of DLFE.

Moreover, we believe that this paper can be used as a starting point for a more thorough theoretical investigation of the relation of these criteria under departure from the ZIL assumptions. Finally, as mentioned before, although MMI is generally considered as a very efficient DLFE criterion, its evaluation is quite difficult and usually requires approximations to make it tractable. Our result that LDA can be derived as a special case of MMI through a purely information theoretic approach could initiate a search for other simplifications of MMI along the same line that could yield new DLFE criteria combining the efficiency of MMI with improved computational tractability.

Summary

Discriminant linear feature extraction (DLFE) is the task of reducing the dimension of the pattern observation space by finding a suitable linear subspace in which the class separability is optimally maintained. The importance and benefits of DLFE in a pattern recognition system have been emphasized even when combined with very competent classifier models.

In this paper we focus on three methods for DLFE, stemming from different points of view: the statistical "*linear discriminant analysis criterion* (LDA), the *heteroscedastic discriminant analysis criterion* (HDA) and the information-theoretic *maximization of mutual information criterion* (MMI). LDA, has a long tradition in statistics and pattern recognition with many variations and applications. HDA is a more recent approach derived by applying the maximum likelihood principle in a heteroscedastic model.

Table 1
Nomenclature

DLFE	Discriminant linear feature extraction
LDA	Linear discriminant analysis
HDA	Heteroscedastic discriminant analysis
MMI	Maximization of mutual information
ICA	Independent component analysis
HOG	Homoscedastic Gaussian (model)
KAH	Kumar and Andreou's heteroscedastic (model)
ZIL	Zero information loss (model)
P_e	Bayes classification error
$\boldsymbol{x} \in \mathscr{X} \subseteq \mathbb{R}^n$	n-dimensional observation vector
X	n-dimensional observation random variable
$\omega_k \in \mathscr{C} = \{\omega_k\}_1^K$	kth class in classification problem
Ω	Random variable taking values from $\mathscr{C}$
$\boldsymbol{\mu}, \Sigma$	Overall mean and covariance matrices of X
μ_k, Σ_k	Class conditional mean and covariance matrix corresponding to kth class
$\bar{\Sigma}$	Average class conditional covariance matrix of X
m	Number of extracted features
$\mathrm{A}, \mathrm{A}^{\mathrm{c}}, \tilde{\mathrm{A}}$	Matrices extracting DLFE subspace, its orthogonal complementary subspace and $\tilde{\mathrm{A}} = [\mathrm{A}\mathrm{A}^{\mathrm{c}}]$
$\hat{\mathrm{A}}^{\mathrm{LDA}}$, $\hat{\mathrm{A}}^{\mathrm{HDA}}$, $\hat{\mathrm{A}}^{\mathrm{MMI}}$, $\hat{\mathrm{A}}^{\mathrm{Bayes}}$	Set of optimal matrices (with respect to the LDA, HDA, MMI and Bayes error criteria)
$\tilde{\mathrm{F}}, \mathrm{F}, \mathrm{F}^{\mathrm{c}}$	Unmixing matrices for the KAH model
B, $\tilde{\mathrm{G}}$, G, G^{c}	Mixing and unmixing matrices for the ZIL model
$\boldsymbol{s}$, S	Source vector and random variable
$\boldsymbol{\zeta}$, Z	Noise vector and random variable
Q	$n \times n$ Sphering matrix
$\Lambda, \Lambda^c, \tilde{\Lambda}$	$n \times m$-dimensional sphered matrices
$\mathscr{H}(\Omega)$	Entropy of the class
$\mathscr{H}(\Omega\|\mathsf{X})$	Equivocation of the class given X
$\mathscr{I}(\Omega;\mathsf{X})$	Mutual information between Ω and X
$\mathscr{J}(\mathsf{X})$	Negentropy of X

The MMI principle has also been known for long as a natural criterion for evaluating the separability quality of features and has recently motivated the development of successful linear feature extraction algorithms.

The purpose of this paper is to provide a unified view of these DLFE criteria by

- emphasizing their similarity through expressing them in a common framework and with mathematical forms that resemble each other
- investigating conditions under which these criteria recover subspaces that are Bayes optimal, in the sense that the minimum possible Bayes error is obtained, and hence optimal classification accuracy is ensured
- proposing conditions on the underlying probabilities of the observation data model under which two or more of the above criteria coincide, in the sense that they recover identical subspaces of the original observation space.

The first objective is met by proposing new model-independent mathematical forms for the three criteria that stress their similarities and elucidate their differences. To meet the second and third objectives, we state the DLFE problem as a source recovering problem. We introduce the following hierarchy of models on the observation space variables: the homoscedastic gaussian model (HOG), Kumar and Andreou's heteroscedastic class-conditional gaussian model (KAH) and a more general model which we call zero information loss model (ZIL). It is shown that each of these models is a special case of the next one in the hierarchy. It is also demonstrated that under the ZIL model the MMI criterion is Bayes optimal. Moreover, our analysis shows that under the KAH model the MMI criterion coincides with the HDA criterion and both criteria reach Bayes optimal solutions. Also, under the more restricted HOG model all three criteria coincide and are Bayes optimal. These results allow for an information-theoretic derivation of the LDA criterion. Finally, our discussion allows for an alternative interpretation of LDA as a special case of multidimensional independent component analysis and for the derivation of suboptimality bounds for the LDA criterion based on information theory.

Appendix A. Proofs

A.1. Proof of Lemma 1

Since the classification problem conforms to the ZIL model with source space dimension d, there will exist a noise vector $\boldsymbol{\zeta}$ of dimension $(n-d)$. Now consider that $\boldsymbol{\zeta}$ is split in two parts $\boldsymbol{\zeta}=[\zeta_1,\zeta_2]$, and define $\boldsymbol{s}'=[\boldsymbol{s},\zeta_1]$ and $\boldsymbol{\zeta}'=\zeta_2$, such that the dimension of $\boldsymbol{s}'$ becomes d'.

Then

$$p(\omega_k|\boldsymbol{s}') = p(\omega_k|\boldsymbol{s},\zeta_1) = \mathscr{E}_{\mathsf{Z}_2}[p(\omega|\boldsymbol{s},\zeta_1,\zeta_2)]$$
$$= \mathscr{E}_{\mathsf{Z}_2}[p(\omega_k|\boldsymbol{s},\boldsymbol{\zeta})] = \mathscr{E}_{\mathsf{Z}_2}[p(\omega_k|\boldsymbol{s})]$$
$$= p(\omega_k|\boldsymbol{s}) \tag{A.1}$$

and

$$p(\omega_k|\boldsymbol{s}',\boldsymbol{\zeta}') = p(\omega_k|\boldsymbol{s},\zeta_1,\zeta_2) = p(\omega_k|\boldsymbol{s},\boldsymbol{\zeta})$$
$$= p(\omega_k|\boldsymbol{s}). \tag{A.2}$$

Thus,

$$p(\omega_k|\boldsymbol{s}',\boldsymbol{\zeta}') = p(\omega_k|\boldsymbol{s}') \tag{A.3}$$

i.e. the probability assumption of the ZIL model holds for $\mathscr{S}'$.

A.2. Proof of Lemma 2

Since linearly transforming the space by a non-singular matrix does not change the Bayes error, it holds that

$$P_e(\Omega|\mathsf{X}) = P_e(\Omega|[\mathrm{G}\ \mathrm{G}^{\mathrm{c}}]^\top\mathsf{X}) = P_e(\Omega|\mathrm{G}^\top\mathsf{X},\ \mathrm{G}^{\mathrm{c}\top}\mathsf{X})$$
$$= P_e(\Omega|\mathsf{S},\mathsf{Z}). \tag{A.4}$$

By using the Bayes error definition and the ZIL model assumption (17), the above equation can be rewritten as

$$P_e(\Omega|\mathsf{X}) = P_e(\Omega|\mathsf{S},\mathsf{Z}) = 1-\mathscr{E}_{\mathsf{S},\mathsf{Z}}\left[\max_{\omega_k} p(\omega_k|\boldsymbol{s},\boldsymbol{\zeta})\right]$$
$$= 1-\mathscr{E}_{\mathsf{S}}\left[\max_{\omega_k} p(\omega_k|\boldsymbol{s})\right] = P_e(\Omega|\mathsf{S}). \tag{A.5}$$

To prove the second part of the lemma, first note that mutual information is invariant under a non-singular linear transform. To see that, consider the non-singular $n\times n$ matrix $\tilde{\mathrm{A}}$, $\tilde{\mathrm{A}}=[\mathrm{A}\mathrm{A}^{\mathrm{c}}]$ such that it spans the whole n-dimensional space. By expressing mutual information as a difference of entropies

$$\mathscr{I}(\Omega;\tilde{\mathrm{A}}^\top\mathsf{X}) = \mathscr{H}(\tilde{\mathrm{A}}^\top\mathsf{X}) - \mathscr{H}(\tilde{\mathrm{A}}^\top\mathsf{X}|\Omega) \tag{A.6}$$

and noticing that the entropy of a linearly transformed vector equals the entropy of the vector plus the logarithm of the absolute value of the determinant of the transformation matrix (see Ref. [25, p. 234]), we can write

$$\mathscr{I}(\Omega;\tilde{\mathrm{A}}^\top\mathsf{X}) = \mathscr{H}(\mathsf{X}) + \log|\tilde{\mathrm{A}}| - (\mathscr{H}(\mathsf{X};\Omega) + \log|\tilde{\mathrm{A}}|)$$
$$= \mathscr{H}(\mathsf{X}) - \mathscr{H}(\mathsf{X};\Omega) = \mathscr{I}(\Omega;\mathsf{X}). \tag{A.7}$$

Now, by the probability assumption of the ZIL model, and the definition of Shannon entropy, it follows that

$$\mathscr{H}(\Omega|\mathsf{S},\mathsf{Z}) = \mathscr{H}(\Omega|\mathsf{S}). \tag{A.8}$$

Therefore, using (A.7) and (A.8), we can write

$$\mathscr{I}(\Omega;\mathsf{X}) = \mathscr{I}(\Omega;\tilde{\mathrm{G}}^\top\mathsf{X}) = \mathscr{I}(\Omega;\mathrm{G}^\top\mathsf{X},\mathrm{G}^{\mathrm{c}\top}\mathsf{X})$$
$$= \mathscr{H}(\Omega) - \mathscr{H}(\Omega|\mathrm{G}^\top\mathsf{X},\mathrm{G}^{\mathrm{c}\top}\mathsf{X})$$
$$= \mathscr{H}(\Omega) - \mathscr{H}(\Omega|\mathsf{S},\mathsf{Z})$$
$$= \mathscr{H}(\Omega) - \mathscr{H}(\Omega|\mathsf{S}) = \mathscr{I}(\Omega;\mathsf{S}). \tag{A.9}$$

A.3. Proof of Lemma 3

To prove the lemma, we need two properties concerning equivocation and mutual information. First for any two random variables $\mathsf{R}_1,\mathsf{R}_2$ it holds that (see Ref. [25, p. 232])

$$\mathscr{H}(\mathsf{R}_1|\mathsf{R}_2) \leqslant \mathscr{H}(\mathsf{R}_1) \tag{A.10}$$

with the equality holding if and only if

$$p(\boldsymbol{r}_1|\boldsymbol{r}_2) = p(\boldsymbol{r}_1)\forall\boldsymbol{r}_1,\boldsymbol{r}_2 \tag{A.11}$$

i.e., conditioning reduces entropy except if the variables are independent.

Second, for a $n\times m$, $m\leqslant n$ matrix A,

$$\mathscr{I}(\Omega;\mathsf{X}) \geqslant \mathscr{I}(\Omega;\mathrm{A}^\top\mathsf{X}) \tag{A.12}$$

with equality holding if $m=n$. The equality part has been shown in Lemma 2, Eq. (A.7). To see that the inequality holds, notice that $\mathscr{I}(\Omega;\mathsf{X})$ can be rewritten as

$$\mathscr{I}(\Omega;\mathsf{X}) = \mathscr{I}(\Omega;\tilde{\mathrm{A}}^\top\mathsf{X}) = \mathscr{I}(\Omega;\mathrm{A}^\top\mathsf{X},\mathrm{A}^{\mathrm{c}\top}\mathsf{X}). \tag{A.13}$$

However, by the chain rule for mutual information (see Ref. [25, p. 22])

$$\mathscr{I}(\Omega;\mathrm{A}^\top\mathsf{X},\mathrm{A}^{\mathrm{c}\top}\mathsf{X})$$
$$= \mathscr{I}(\Omega;\mathrm{A}^\top\mathsf{X}) + \mathscr{I}(\Omega;\mathrm{A}^{\mathrm{c}\top}\mathsf{X}|\mathrm{A}^\top\mathsf{X}) \tag{A.14}$$

which implies

$$\mathscr{I}(\Omega;\mathrm{A}^\top\mathsf{X},\mathrm{A}^{\mathrm{c}\top}\mathsf{X}) \geqslant \mathscr{I}(\Omega;\mathrm{A}^\top\mathsf{X}) \tag{A.15}$$

and therefore, by Eq. (A.13), property (A.12) holds.

Now let A be an $n\times m$ MMI-optimal matrix. By Lemma 1, since $m\geqslant d$, we can always find a vector $\boldsymbol{s}'$ of dimension m that qualifies as source vector of the ZIL model. Moreover, since A is MMI optimal, $\mathscr{I}(\Omega;\mathrm{A}^\top\mathsf{X})$ cannot be less than $\mathscr{I}(\Omega;\mathsf{S}')$. Also, as shown in Lemma 2, $\mathscr{I}(\Omega;\mathsf{S}')=\mathscr{I}(\Omega;\mathsf{X})$ and therefore $\mathscr{I}(\Omega;\mathrm{A}^\top\mathsf{X})$ cannot be less than $\mathscr{I}(\Omega;\mathsf{X})$. However, by Eq. (A.12), $\mathscr{I}(\Omega;\mathrm{A}^\top\mathsf{X})$ also cannot exceed $\mathscr{I}(\Omega;\mathsf{X})$. Hence,

$$\mathscr{I}(\Omega;\mathrm{A}^\top\mathsf{X}) = \mathscr{I}(\Omega;\mathsf{X}). \tag{A.16}$$

Furthermore, rewriting mutual information as a difference of entropy and equivocation, and using Eq. (A.13), we have

$$\mathscr{H}(\Omega) - \mathscr{H}(\Omega|\mathrm{A}^\top\mathsf{X})$$
$$= \mathscr{H}(\Omega) - \mathscr{H}(\Omega|\mathrm{A}^\top\mathsf{X},\mathrm{A}^{\mathrm{c}\top}\mathsf{X}) \tag{A.17}$$

or

$$\mathscr{H}(\Omega|\mathrm{A}^\top\mathsf{X}) = \mathscr{H}(\Omega|\mathrm{A}^\top\mathsf{X}, \mathrm{A}^{\mathrm{c}\top}\mathsf{X}) \tag{A.18}$$

which leads, by applying Eqs. (A.10)–(A.11), to

$$p(\omega_k|\mathrm{A}^\top\boldsymbol{x}, \mathrm{A}^{\mathrm{c}\top}\boldsymbol{x}) = p(\omega_k|\mathrm{A}^\top\boldsymbol{x}) \quad \forall \boldsymbol{x}, \omega_k. \tag{A.19}$$

Reversely, let A and A^{c} be matrices for which Eq. (A.19) holds. Then, by Eqs. (A.10)–(A.11), Eq. (A.18) also holds and, by following the reverse path, so does Eq. (A.16). Therefore, by Eq. (A.12), $\mathrm{A}^\top\mathsf{X}$ attains the maximum possible mutual information with the class and A is an MMI optimal matrix.

A.4. Proof of Proposition 3

The LDA expression is easily derived from Eq. (8). By replacing the average class conditional covariance matrix $\bar{\Sigma}$ by its definition (6), the denominator of Eq. (8) is rewritten as

$$\begin{aligned} |\mathrm{A}^\top\bar{\Sigma}\mathrm{A}| &= \left|\mathrm{A}^\top \left[\sum_{k=1}^{K} p(\omega_k)\Sigma_k\right] \mathrm{A}\right| \\ &= \sum_{k=1}^{K} p(\omega_k)|\mathrm{A}^\top\Sigma_k\mathrm{A}| \end{aligned} \tag{A.20}$$

from which Eq. (27) follows.

The derivation of the HDA form is not so straightforward. First, consider an $n \times n$ matrix Q which transforms the original space, such that the overall covariance matrix of the observation in the transformed space becomes the unity matrix, i.e.

$$\mathrm{Q}^\top\Sigma\mathrm{Q} = \mathrm{I}_n. \tag{A.21}$$

Such a matrix is called a *sphering matrix* or *whitening matrix* and can always be evaluated as the inverse of the product of a matrix containing the eigenvectors of the covariance matrix and a diagonal matrix containing the square roots of the corresponding eigenvalues (see Ref. [30]).

Using sphering, the transformation of matrix $\tilde{\mathrm{A}}$, as well as of A and A^{c} can be decomposed as $\tilde{\mathrm{A}} = \mathrm{Q}\tilde{\Lambda}$, respectively $\mathrm{A} = \mathrm{Q}\Lambda$ and $\mathrm{A}^{\mathrm{c}} = \mathrm{Q}\Lambda^c$, i.e. a sphering step using Q and a unmixing step using $\tilde{\Lambda} = [\Lambda\Lambda^c]$.

Now, by looking at Eq. (10), $\log|\mathrm{A}^{\mathrm{c}\top}\Sigma\mathrm{A}^{\mathrm{c}}|$ can be rewritten as

$$\begin{aligned} \log|\mathrm{A}^{\mathrm{c}\top}\Sigma\mathrm{A}^{\mathrm{c}}| &= \log|(\mathrm{Q}\Lambda^c)^\top\Sigma(\mathrm{Q}\Lambda^c)| \\ &= \log|\Lambda^{c\top}(\mathrm{Q}^\top\Sigma\mathrm{Q})\Lambda^c| \\ &= \log|\Lambda^{c\top}\Lambda^c|, \end{aligned} \tag{A.22}$$

where we have made use of Eq. (A.21). Moreover, notice that

$$\begin{aligned} |\tilde{\Lambda}^\top\tilde{\Lambda}| = |[\Lambda\Lambda^c]^\top[\Lambda\Lambda^c]| &= \begin{vmatrix} \Lambda^\top\Lambda & 0 \\ 0 & \Lambda^{c\top}\Lambda^c \end{vmatrix} \\ &= |\Lambda^\top\Lambda||\Lambda^{c\top}\Lambda^c|, \end{aligned} \tag{A.23}$$

where the zero submatrices originate from the fact that Λ and Λ^c are formed by vectors orthogonal to each other, and hence they have zero product. Hence Eq. (A.22) becomes

$$\log|\mathrm{A}^{\mathrm{c}\top}\Sigma\mathrm{A}^{\mathrm{c}}| = \log|\tilde{\Lambda}^\top\tilde{\Lambda}| - \log|\Lambda^\top\Lambda|. \tag{A.24}$$

Since $\Lambda = \mathrm{Q}^{-1}\mathrm{A}$, the second term on the right-hand side of Eq. (A.24) becomes

$$\begin{aligned} \log|\Lambda^\top\Lambda| &= \log|(\mathrm{Q}^{-1}\mathrm{A})^\top(\mathrm{Q}^{-1}\mathrm{A})| \\ &= \log|\mathrm{A}^\top(\mathrm{Q}^{-1\top}\mathrm{Q}^{-1})\mathrm{A}| \\ &= \log|\mathrm{A}^\top\Sigma\mathrm{A}|, \end{aligned} \tag{A.25}$$

where we have made use of Eq. (A.21). Furthermore, since $\tilde{\Lambda} = \mathrm{Q}^{-1}\tilde{\mathrm{A}}$, the first term on the right-hand side of Eq. (A.24) becomes

$$\begin{aligned} \log|\tilde{\Lambda}^\top\tilde{\Lambda}| &= \log|(\mathrm{Q}^{-1}\tilde{\mathrm{A}})^\top(\mathrm{Q}^{-1}\tilde{\mathrm{A}})| \\ &= 2\log|\mathrm{Q}^{-1}| + 2\log|\tilde{\mathrm{A}}| \\ &= -2\log|\mathrm{Q}| + 2\log|\tilde{\mathrm{A}}|, \end{aligned} \tag{A.26}$$

where we have used the fact that Q and $\tilde{\mathrm{A}}$ are square matrices and hence the determinant of their product equals the product of their determinants. Thus, by Eqs. (A.25) and (A.26), Eq. (A.24) becomes

$$\begin{aligned} \log|\mathrm{A}^{\mathrm{c}\top}\Sigma\mathrm{A}^{\mathrm{c}}| &= -2\log|\mathrm{Q}| + 2\log|\tilde{\mathrm{A}}| \\ &\quad + \log|\mathrm{A}^\top\Sigma\mathrm{A}|. \end{aligned} \tag{A.27}$$

Introducing Eq. (A.27) into Eq. (10), the terms containing the matrix $\tilde{\mathrm{A}}$ cancel each other, and the criterion simplifies to

$$\begin{aligned} \hat{\mathrm{A}}^{\mathrm{HDA}} = \operatorname*{argmax}_{\mathrm{A}} \Bigg[&\log|\mathrm{A}^\top\Sigma\mathrm{A}| \\ &- \sum_{k=1}^{K} p(\omega_k)\log|\mathrm{A}^\top\Sigma_k\mathrm{A}| + 2\log|\mathrm{Q}|\Bigg]. \end{aligned} \tag{A.28}$$

Now, notice that the determinant of the sphering matrix Q is a constant term that does not affect the optimization and thus it can be ignored. Hence, we end up with

$$\begin{aligned} \hat{\mathrm{A}}^{\mathrm{HDA}} = \operatorname*{argmax}_{\mathrm{A}} \Bigg[&\log|\mathrm{A}^\top\Sigma\mathrm{A}| \\ &- \sum_{k=1}^{K} p(\omega_k)\log|\mathrm{A}^\top\Sigma_k\mathrm{A}|\Bigg] \end{aligned} \tag{A.29}$$

or, in a more compact form,

$$\hat{\mathrm{A}}^{\mathrm{HDA}} = \operatorname*{argmax}_{\mathrm{A}} \log \frac{|\mathrm{A}^\top\Sigma\mathrm{A}|}{\prod_{k=1}^{K}|\mathrm{A}^\top\Sigma_k\mathrm{A}|^{p(\omega_k)}}. \tag{A.30}$$

The derivation for the MMI criterion is done using the expansion of entropies and conditional entropies in their gaussian and negentropy parts. First, notice that the mutual information of the extracted vector with the class can be written as the difference between its unconditional and the average class conditional entropy (14):

$$\mathscr{I}(\Omega;\mathrm{A}^\top\mathsf{X}) = \mathscr{H}(\mathrm{A}^\top\mathsf{X}) - \mathscr{H}(\mathrm{A}^\top\mathsf{X}|\Omega). \tag{A.31}$$

Now, based on Eqs. (20) and (23), the above equation is rewritten as

$$\mathscr{I}(\Omega;\mathrm{A}^\top \mathsf{X}) = (\mathscr{H}_g(\mathrm{A}^\top \mathsf{X}) - \mathscr{H}_g(\mathrm{A}^\top \mathsf{X}|\Omega)) \\ - (\mathscr{J}(\mathrm{A}^\top \mathsf{X}) - \mathscr{J}(\mathrm{A}^\top \mathsf{X}|\Omega)). \tag{A.32}$$

Finally the difference of gaussian entropies is further expressed as

$$\begin{aligned}\mathscr{H}_g(\mathrm{A}^\top \mathsf{X}) - \mathscr{H}_g(\mathrm{A}^\top \mathsf{X}|\Omega) &= \frac{1}{2}\log[(2\pi \mathrm{e})^m|\mathrm{A}^\top \Sigma \mathrm{A}|] \\ &\quad - \sum_k p(\omega_k)\frac{1}{2}\log[(2\pi \mathrm{e})^m|\mathrm{A}^\top \Sigma_k \mathrm{A}|] \\ &= \frac{1}{2}\log\frac{|\mathrm{A}^\top \Sigma \mathrm{A}|}{\prod_{k=1}^K |\mathrm{A}^\top \Sigma_k \mathrm{A}|^{p(\omega_k)}}\end{aligned} \tag{A.33}$$

and by insertion of Eq. (A.33) into Eq. (A.32), and multiplication by 2 the criterion becomes

$$\bar{\mathrm{A}}^{\mathrm{MNI}} = \underset{\mathrm{A}}{\operatorname{argmax}}\left[\log\frac{|\mathrm{A}^\top \Sigma \mathrm{A}|}{\prod_{k=1}^K |\mathrm{A}^\top \Sigma_k \mathrm{A}|^{p(\omega_k)}} \right. \\ \left. -\ 2(\mathscr{J}(\mathrm{A}^\top \mathsf{X}) - \mathscr{J}(\mathrm{A}^\top \mathsf{X}|\Omega))\right].$$

A.5. Proof of Proposition 4

Here we prove the first part of the proposition. For the proof of the second part, the reader is referred to Ref. [11].

Consider the matrix $\tilde{\mathrm{F}} = [\mathrm{F}\mathrm{F}^{\mathrm{c}}]$ as defined by the KAH model. Then, by defining $\boldsymbol{s} = \mathrm{F}^\top \boldsymbol{x}$ and $\boldsymbol{\zeta} = \mathrm{F}^{\mathrm{c}\top}\boldsymbol{x}$, the probability assumption of the ZIL model is derived as follows: First, since both source vector and noise vector are gaussian, given the class, their joint distribution given the class will be gaussian, and, under the KAH model, their covariance matrix and its determinant will be

$$\begin{bmatrix} \mathrm{F}^\top \Sigma_k \mathrm{F} & 0 \\ 0 & \mathrm{F}^{\mathrm{c}\top}\Sigma \mathrm{F}^{\mathrm{c}} \end{bmatrix},$$

$$\begin{vmatrix} \mathrm{F}^\top \Sigma_k \mathrm{F} & 0 \\ 0 & \mathrm{F}^{\mathrm{c}\top}\Sigma \mathrm{F}^{\mathrm{c}} \end{vmatrix} = |\mathrm{F}^\top \Sigma_k \mathrm{F}||\mathrm{F}^{\mathrm{c}\top}\Sigma \mathrm{F}^{\mathrm{c}}|. \tag{A.34}$$

It follows that

$$\begin{aligned} p(\boldsymbol{s},\boldsymbol{\zeta}|\omega_k) &= \frac{\mathrm{e}^{-\frac{1}{2}\begin{bmatrix}\boldsymbol{s}-\mathrm{F}^\top\boldsymbol{\mu}_k \\ \boldsymbol{\zeta}-\mathrm{F}^{\mathrm{c}\top}\boldsymbol{\mu}\end{bmatrix}^\top \begin{bmatrix}(\mathrm{F}^\top \Sigma_k \mathrm{F})^{-1} & 0 \\ 0 & (\mathrm{F}^{\mathrm{c}\top}\Sigma \mathrm{F}^{\mathrm{c}})^{-1}\end{bmatrix}\begin{bmatrix}\boldsymbol{s}-\mathrm{F}^\top\boldsymbol{\mu}_k \\ \boldsymbol{\zeta}-\mathrm{F}^{\mathrm{c}\top}\boldsymbol{\mu}\end{bmatrix}}}{(2\pi)^{\frac{n}{2}}\begin{vmatrix} \mathrm{F}^\top \Sigma_k \mathrm{F} & 0 \\ 0 & \mathrm{F}^{\mathrm{c}\top}\Sigma \mathrm{F}^{\mathrm{c}} \end{vmatrix}^{\frac{1}{2}}} \\ &= \frac{\mathrm{e}^{-\frac{1}{2}(\boldsymbol{s}-\mathrm{F}^\top\boldsymbol{\mu}_k)^\top(\mathrm{F}^\top\Sigma_k\mathrm{F})^{-1}(\boldsymbol{s}-\mathrm{F}^\top\boldsymbol{\mu}_k)}\mathrm{e}^{-\frac{1}{2}(\boldsymbol{\zeta}-\mathrm{F}^{\mathrm{c}\top}\boldsymbol{\mu})^\top(\mathrm{F}^{\mathrm{c}\top}\Sigma\mathrm{F}^{\mathrm{c}})^{-1}(\boldsymbol{\zeta}-\mathrm{F}^{\mathrm{c}\top}\boldsymbol{\mu})}}{(2\pi)^{\frac{d}{2}}|\mathrm{F}^\top\Sigma_k\mathrm{F}|^{\frac{1}{2}}(2\pi)^{\frac{(n-d)}{2}}|\mathrm{F^c}\top\Sigma\mathrm{F^c}|^{\frac{1}{2}}} \\ &= p(\boldsymbol{s}|\omega_k)p(\boldsymbol{\zeta}). \end{aligned} \tag{A.35}$$

This implies that

$$\begin{aligned} p(\boldsymbol{s},\boldsymbol{\zeta}) &= \sum_k p(\boldsymbol{s},\boldsymbol{\zeta}|\omega_k)p(\omega_k) \\ &= \sum_k p(\boldsymbol{s}|\omega_k)p(\omega_k)p(\boldsymbol{\zeta}) \\ &= p(\boldsymbol{s})p(\boldsymbol{\zeta}) \end{aligned} \tag{A.36}$$

i.e. the source and noise vectors are independent, and, finally,

$$\begin{aligned} p(\omega_k|\boldsymbol{s},\boldsymbol{\zeta}) &= \frac{p(\boldsymbol{s},\boldsymbol{\zeta}|\omega_k)p(\omega_k)}{p(\boldsymbol{s},\boldsymbol{\zeta})} = \frac{p(\boldsymbol{s}|\omega_k)p(\boldsymbol{\zeta})p(\omega_k)}{p(\boldsymbol{s})p(\boldsymbol{\zeta})} \\ &= \frac{p(\boldsymbol{s}|\omega_k)p(\omega_k)}{p(\boldsymbol{s})} = p(\omega_k|\boldsymbol{s}). \end{aligned} \tag{A.37}$$

A.6. Proof of Lemma 4

First, we show that for any $n\times m, m\leqslant n$ matrix A, it holds

$$\mathscr{J}(\mathsf{X}) \geqslant \mathscr{J}(\mathrm{A}^\top \mathsf{X}), \tag{A.38}$$

i.e. there is no projection of the observation vector that has larger negentropy than the observation vector itself.

The negentropy is invariant under a non singular linear transformation [26] and thus

$$\mathscr{J}(\mathsf{X}) = \mathscr{J}(\tilde{\mathrm{A}}^\top \mathsf{X}) \tag{A.39}$$

and inequality (A.38) becomes

$$\mathscr{J}(\tilde{\mathrm{A}}^\top \mathsf{X}) \geqslant \mathscr{J}(\mathrm{A}^\top \mathsf{X}). \tag{A.40}$$

Next, by rewriting both negentropies in terms of entropies and gaussian entropies, we obtain

$$\mathscr{H}_g(\tilde{\mathrm{A}}^\top \mathsf{X}) - \mathscr{H}(\tilde{\mathrm{A}}^\top \mathsf{X}) \geqslant \mathscr{H}_g(\mathrm{A}^\top \mathsf{X}) - \mathscr{H}(\mathrm{A}^\top \mathsf{X}) \tag{A.41}$$

or, by re-arranging the terms,

$$\mathscr{H}_g(\tilde{\mathrm{A}}^\top \mathsf{X}) - \mathscr{H}_g(\mathrm{A}^\top \mathsf{X}) \geqslant \mathscr{H}(\tilde{\mathrm{A}}^\top \mathsf{X}) - \mathscr{H}(\mathrm{A}^\top \mathsf{X}). \tag{A.42}$$

Thus, to prove Eq. (A.38) it suffices to prove Eq. (A.42).

Now, both the left- and the right-hand side of the inequality can be rewritten in simple forms. Beginning with the right-hand side and using the conditional entropy definition [25, p. 230],

$$\begin{aligned}\mathscr{H}(\tilde{\mathrm{A}}^\top \mathsf{X}) - \mathscr{H}(\mathrm{A}^\top \mathsf{X}) &= \mathscr{H}(\mathrm{A}^\top \mathsf{X}, \mathrm{A}^{\mathrm{c}\top}\mathsf{X}) - \mathscr{H}(\mathrm{A}^\top \mathsf{X}) \\ &= \mathscr{H}(\mathrm{A}^{\mathrm{c}\top}\mathsf{X}|\mathrm{A}^\top \mathsf{X})\end{aligned} \tag{A.43}$$

i.e. the difference of entropies equals the conditional entropy of the complementary vector given the vector.

The same arguments hold for gaussian random variables, and thus

$$\mathscr{H}_g(\tilde{\mathrm{A}}^\top \mathsf{X}) - \mathscr{H}_g(\mathrm{A}^\top \mathsf{X}) = \mathscr{H}_g(\mathrm{A}^{\mathrm{c}\top}\mathsf{X}|\mathrm{A}^\top \mathsf{X}). \tag{A.44}$$

Introducing Eqs. (A.43) and (A.44) in inequality (A.42), yields finally

$$\mathscr{H}_g(\mathrm{A}^{\mathrm{c}\top}\mathsf{X}|\mathrm{A}^\top \mathsf{X}) \geqslant \mathscr{H}(\mathrm{A}^{\mathrm{c}\top}\mathsf{X}|\mathrm{A}^\top \mathsf{X}) \tag{A.45}$$

or

$$\mathscr{J}(\mathrm{A}^{\mathrm{c}\top}\mathsf{X}|\mathrm{A}^{\top}\mathsf{X}) \geqslant 0. \tag{A.46}$$

Since negentropy is always non negative, Eq. (A.46) holds always, and thus Eq. (A.38) holds.

We will show now that Eq. (A.38), and thus Eq. (A.42) holds with equality when $\mathrm{A} = \mathrm{F}$, i.e. when $\mathrm{A}^{\top}\mathsf{X} = \mathsf{S}$.

To facilitate the proof, single linearly transforming a variable does not change its negentropy [26], we may assume without loss of generality that $\tilde{\mathrm{F}}^{\top}\mathsf{X}$, and thus $\mathrm{F}^{\top}\mathsf{X}$ and $\mathrm{F}^{\mathrm{c}\top}\mathsf{X}$, are sphered, i.e. their covariance matrix equals the unity matrix. The sphering process has been discussed in the proof of Proposition 3 (Appendix A.4). In this case, as in Eq. (A.23),

$$|\tilde{\mathrm{F}}^{\top}\Sigma\tilde{\mathrm{F}}| = |\tilde{\mathrm{F}}^{\top}\tilde{\mathrm{F}}| = |\mathrm{F}^{\top}\mathrm{F}|\,|\mathrm{F}^{\mathrm{c}\top}\mathrm{F}^{\mathrm{c}}| \tag{A.47}$$

and, the left-hand side of Eq. (A.42) becomes

$$\begin{aligned}
&\mathscr{H}_g(\tilde{\mathrm{F}}^{\top}\mathsf{X}) - \mathscr{H}_g(\mathrm{F}^{\top}\mathsf{X})\\
&\quad = \frac{1}{2}\log[(2\pi e)^n|\tilde{\mathrm{F}}^{\top}\tilde{\mathrm{F}}|] - \frac{1}{2}\log[(2\pi e)^m|\mathrm{F}^{\top}\mathrm{F}|]\\
&\quad = \frac{1}{2}\log[(2\pi e)^{n-m}|\mathrm{F}^{\mathrm{c}\top}\mathrm{F}^{\mathrm{c}}|]\\
&\quad = \mathscr{H}_g(\mathrm{F}^{\mathrm{c}\top}\mathsf{X}) = \mathscr{H}(\mathrm{F}^{\mathrm{c}\top}\mathsf{X}) = \mathscr{H}(\mathsf{Z}),
\end{aligned} \tag{A.48}$$

where the last step follows by the KAH assumption that the distribution in the noise subspace is gaussian, and thus the gaussian entropy equals the entropy. Moreover, by Eq. (A.36), the noise subspace on the KAH model is independent of the source subspace, which leads to [25, p. 230]

$$\mathscr{H}(\mathrm{F}^{\mathrm{c}\top}\mathsf{X}|\mathrm{F}^{\top}\mathsf{X}) = \mathscr{H}(\mathrm{F}^{\mathrm{c}\top}\mathsf{X}) = \mathscr{H}(\mathsf{Z}). \tag{A.49}$$

Thus, the right-hand side of Eq. (A.42) becomes

$$\mathscr{H}(\tilde{\mathrm{F}}^{\top}\mathsf{X}) - \mathscr{H}(\mathrm{F}^{\top}\mathsf{X}) = \mathscr{H}(\mathsf{Z}). \tag{A.50}$$

Introducing Eqs. (A.48) and (A.50) in Eq. (A.42), we see that the relation holds with equality, and thus

$$\mathscr{J}(\mathsf{X}) = \mathscr{J}(\mathrm{F}^{\top}\mathsf{X}) = \mathscr{J}(\mathsf{S}) \tag{A.51}$$

which, together with Eq. (A.38) proves that the negentropy is maximum in the source subspace.

References

[1] A.K. Jain, R.P. Duin, J. Mao, Statistical pattern recognition: a review, IEEE Trans. Pattern Anal. Mach. Intell. 22 (1) (2000) 4–37.

[2] J. Kittler, A framework for classifier fusion: is it still needed?, in: Proceedings of the Joint SPR& SPRR Workshop, Alicante, Spain, 2000, pp. 45–56.

[3] H. Yu, J. Yang, A direct LDA algorithm for high-dimensional data—with application to face recognition, Pattern Recognition 34 (10) (2001) 2067–2070.

[4] Y.-D. Guo, T.-T. Shu, J.-Y. Yang, S.-J. Li, Feature extraction method based on the generalised Fisher discriminant criterion and facial recognition, Pattern Anal. Appl. 4 (2001) 61–66.

[5] K. Torkkola, Linear discriminant analysis in document classification, in: IEEE ICDM, San Jose, CA, USA, 2001.

[6] D.M. Hawkins, G.J. McLachlan, High breakdown C/LDA, J. Am. Statist. Assoc. 92 (437) (1997) 136–143.

[7] J. Yang, J.-Y. Yang, Why can LDA be performed in PCA transformed space? Pattern Recognition 36 (2003) 563.

[8] D. Xu, J.C. Principe, H.-C. Wu, Generalized eigen-decomposition with an on-line local algorithm, IEEE Signal Process. Lett. 5 (1999) 298–301.

[9] T. Hastie, R. Tibshirani, Discriminant analysis by gaussian mixtures, J. R. Statist. Soc. Ser. B-Methodological 58 (1) (1996) 155–176.

[10] R. Lotlikar, R. Kothari, Fractional step dimensionality reduction, IEEE Trans. Pattern Anal. Mach. Intell. 22 (6) (2000) 623–627.

[11] N. Kumar, A.G. Andreou, Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition, Speech Commun. 26 (4) (1998) 283–297.

[12] P. Lewis II, The characteristic selection problem in recognition systems, IRE Trans. Inform. Theory IT-8 (1962) 171–178.

[13] M. Ben-Bassat, Use of distance measures, information measures and error bounds in feature evaluation, in: P. Krishnaiah, L. Kanal (Eds.), Handbook of Statistics, North-Holland, Amsterdam, 1982, pp. 773–791.

[14] G.A. Darbellay, An estimator of the mutual information based on a criterion for independence, Comput. Statist. Data Anal. 32 (1999) 1–17.

[15] R. Battiti, Using mutual information for selecting features in supervised neural net learning, IEEE Trans. Neural Networks 5 (4) (1994) 537–550.

[16] N. Kwak, Chong-Ho, Input feature selection for classification problems, IEEE Trans. Neural Networks 13 (2002) 143–159.

[17] J.C. Principe, D. Xu, Q. Zhao, J.W. Fisher III, Learning from examples with information theoretic criteria, in: NNSP'99 Special Issue, Kluwer Publishing Co, Dordrecht, 2000.

[18] K. Torkkola, Learning discriminative feature transforms to low dimensions in low dimensions, in: Advances in Neural Information Processing Systems, MIT Press, Vancouver, BC, Canada, 2001.

[19] G.J. McLachlan, Discriminant Analysis and Statistical Pattern Recognition, Wiley, New York, 1992.

[20] J.-F. Cardoso, Multidimensional independent component analysis, in: Proceedings of ICASSP, Seattle, WA, USA, 1998, pp. 1941–1944.

[21] K. Fukunaga, Introduction to Statistical Pattern Recognition, Academic Press, Boston, MA, 1990.

[22] E. Schukat-Talamazini, J. Hornegger, H. Nieman, Optimal linear feature transformations for semi-continuous hidden Markov models, in: Proceedings of ICASSP, Detroit, Vol. I, 1995, pp. 369–372.

[23] N. Kumar, A.G. Andreou, A generalization of linear discriminant analysis in maximum likelihood framework, in: Proceedings of Joint Meeting of American Statistical Association, Chicago, IL, USA, 1996.

[24] R.B. Ash, Information Theory, Dover Publications, New York, 1990.

[25] T.M. Cover, J.A. Thomas, Elements of Information Theory, Wiley, New York, 1991.

[26] P. Comon, Independent component analysis, a new concept? Signal Process. 36 (1994) 287–314.

[27] W.A. Gardner, A unifying view of second-order measures of quality for signal classification, IEEE Trans. Comm. 28 (1980) 800–816.

[28] G. Saon, M. Padmanabhan, R. Gopinath, S. Chen, Maximum likelihood discriminant feature spaces, in: Proceedings of ICASSP, Vol. II, Seattle, 2000, pp. 1129–1132.

[29] A. Hyvarinen, E. Oja, Independent component analysis: algorithms and applications, Neural Networks 13 (4–5) (2000) 411–430.

[30] A.K. Nandi, Blind Estimation Using Higher-Order Statistics, Kluwer Academic Publishers, Boston, MA, 1999.

About the Author—SERGIOS PETRIDIS was born in Athens in 1973. He received the Diploma Degree in Electrical and Computer Engineering from the National Technical University of Athens, Athens in 1996 and the D.E.A in Artificial Intelligence, Pattern Recognition and Applications from University "Pierre et Marie Curie", Paris in 1997. Currently he is a Ph.D. candidate at the Department of Informatics and Telecommunications of the University of Athens and member of the Computational Intelligence Laboratory at the Institute of Informatics and Telecommunications of the NCSR "Demokritos". His interests include statistical pattern recognition, speech recognition, natural language processing and neural networks.

About the Author—STAVROS J. PERANTONIS holds a B.Sc. degree in Physics from the Department of Physics, University of Athens, an M.Sc. degree in Computer Science from the Department of Computer Science, University of Liverpool and a D. Phil. Degree in Computational Physics from the Department of Physics, University of Oxford. Since 1992 he has been with the Institute of Informatics and Telecommunications, National Centre for Scientific Research "Demokritos", where he currently holds the position of Senior Researcher and heads the Computational Intelligence Program of the Institute. His current research activities are in the areas of computational intelligence, neural networks, pattern recognition and image processing. Dr. Perantonis is the author of more than 90 papers in journals and refereed conference proceedings and has been involved in numerous European or national research and development projects concerning industrial or web-based intelligent computing applications.