Mismatch in the Classification of Linear Subspaces: Sufficient Conditions for Reliable Classification

Jure Sokolić, *Student Member, IEEE*, Francesco Renna, *Member, IEEE*, Robert Calderbank, *Fellow, IEEE*, and Miguel R. D. Rodrigues, *Senior Member, IEEE*

***Abstract*—This paper considers the classification of linear subspaces with mismatched classifiers. In particular, we assume a model where one observes signals in the presence of isotropic Gaussian noise and the distribution of the signals conditioned on a given class is Gaussian with a zero mean and a low-rank covariance matrix. We also assume that the classifier knows only a mismatched version of the parameters of input distribution *in lieu* of the true parameters. By constructing an asymptotic low-noise expansion of an upper bound to the error probability of such a mismatched classifier, we provide sufficient conditions for reliable classification in the low-noise regime that are able to sharply predict the absence of a classification error floor. Such conditions are a function of the geometry of the true signal distribution, the geometry of the mismatched signal distributions as well as the interplay between such geometries, namely, the principal angles and the overlap between the true and the mismatched signal subspaces. Numerical results demonstrate that our conditions for reliable classification can sharply predict the behavior of a mismatched classifier both with synthetic data and in a motion segmentation and a hand-written digit classification applications.**

***Index Terms*—Classification, mismatch, linear subspace, maximum-a-posteriori classifier, error floor.**

I. Introduction

SIGNAL classification is a fundamental task in various fields, including statistics, machine learning and computer vision. One often approaches this problem by leveraging the Bayesian inference paradigm, where one infers the signal class from signal samples or measurements based on a model of the joint distribution of the signal and signal classes ([1], Chapter 2).

Such joint distribution is typically inferred by relying on pre-labeled data sets. However, in practical applications, the methods used to estimate the distributions from training data inevitably lead to signal models that are not perfectly matched to the underlying one. This can be due to an insufficient number of labeled data, the noise in the pre-labeled data [2]–[4], or due to the non-stationary statistical behavior [5].

It is therefore relevant to ask the question:

What is the impact that a mismatched classifier, i.e., a classifier that infers the signal classes based on an inaccurate model of the data distribution in lieu of the true underlying data distribution, has on classification performance?

We answer this question for the scenario where the data classes are constrained to lie approximately on a low-dimensional linear subspace embedded in the high-dimensional ambient space. Indeed, there are various problems in signal processing, image processing and computer vision that conform to such a model, some of which are:

- *Face Recognition*: It can be shown that, provided that the Lambertian reflectance assumption is verified, the set of images taken from the same subject under different lighting conditions can be well approximated by a low-dimensional linear subspace embedded in the high-dimensional space [6]. This is leveraged in several face recognition applications [7]–[9].
- *Motion Segmentation*: It can also be shown—under the assumption of the affine projection camera model—that the coordinates of feature points associated with rigidly moving objects through different video frames lie in a 4 dimensional linear space [10]–[12]. This is leveraged in [10] to design subspace clustering algorithms that can perform motion segmentation.
- In general, (affine) subspaces or unions of (affine) subspaces can also be used to model other data such as images of handwritten digits [13].

Our contributions include:

- We derive an upper bound to the error probability associated with the mismatched classifier for the case where the distribution of the signal in a given class is Gaussian with zero-mean and low-rank covariance matrix.
- We then derive sufficient conditions for reliable classification in the asymptotic low-noise regime. Such conditions are expressed in terms of the geometry of the true signal model, the geometry of the mismatched signal model and the interaction of these geometries (via the principal angles associated with the subspaces of the true and mismatched signal models as well as the dimension of the intersection of such subspaces).
- We finally provide a number of results, both with synthetic and real data, that show that our sufficient conditions for reliable classification are sharp. In particular, we also use our theoretical framework to determine the number of training

Manuscript received August 07, 2015; revised January 12, 2016; accepted February 08, 2016. Date of publication March 02, 2016; date of current version April 20, 2016. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Frederic Pascal. The work of J. Sokolić, F. Renna and M. R. D. Rodrigues was supported in part by EPSRC under grant EP/K033166/1. The work of R. Calderbank was supported in part by AFOSR under grant FA 9550-13-1-0076 and by NGA under grant HM017713-1-0006. This paper was presented in part at the IEEE International Symposium on Information Theory 2015.

J. Sokolić, F. Renna, and M. R. D. Rodrigues are with the Department of Electronic and Electrical Engineering, University College London, WC1E 7JE London, U.K. (e-mail: jure.sokolic.13@ucl.ac.uk; f.renna@ucl.ac.uk; m.rodrigues@ucl.ac.uk).

R. Calderbank is with the Department of Electrical and Computer Engineering, Duke University, NC, USA (e-mail: robert.calderbank@duke.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2016.2537272

samples needed to achieve reliable classification in a motion segmentation and a hand-written digit classification applications.

A. Related Work

The concept of model mismatch has been widely explored by the information theory and communication theory communities. For example, in lossless source coding problems, mismatch between the distribution used to encode the source and the true distribution is shown to lead to a compression rate penalty which is determined by the Kullback-Leibler (KL) distance between the mismatched and the true distributions [14, Theorem 5.4.3].

In channel coding problems, mismatch has an impact on the reliable information transmission rate that has been characterized via inner and outer bounds to the achievable rate and error exponents of different channel models [15]–[19]. The problem of mismatched quantization is considered in [20].

The concept of mismatch has also been explored in the machine learning literature [5]. In particular, [5] studies the impact on classification performance of training sets consisting of biased samples of the true distribution, expressing classification error bounds as a function of the sample bias severity and type. The effect of label noise in the training sets is also considered in classification algorithms such as Support Vector Machines [3] and Logistic Regression classifiers [4]. See also [2] for an overview of the literature on classification in presence of label noise.

Signal classification and estimation using mismatched models is also considered in [21]–[24]. For example, [23] expresses bounds to the error probability in the presence of mismatch via the f-Divergence between the true and mismatched source distributions, and [24] expresses the mean-squared error penalty in presence of mismatch in terms of the derivative of the KL distance between the true and the mismatched distributions with respect to the decoder signal to noise ratio (SNR). In particular, the work in [23] is closely related to our work in the sense that it also establishes bounds to the error probability in the presence of mismatch. The bounds presented in [23] are more general since they do not assume a particular form of probability density functions. Our work, on the other hand, leverages the assumption that signals are contained in linear subspaces in order to derive an upper bound that sharply predicts the presence or absence of an error floor. The bounds in [23] fail to capture the presence or absence of an error floor when specialized to the proposed signal model.

B. Organization

The remainder of this paper is organized as follows: Section II introduces the observation and signal models, the Mismatched Maximum-a-Posteriori (MMAP) classifier and the geometrical quantities associated with the signal and the mismatched model that are essential for the description of the MMAP classifier performance. The upper bound to the error probability associated with the MMAP classifier and the asymptotic expansion, which provide sufficient conditions for reliable classification in the low-noise regime, are given in Section III. In Section IV the theoretical results are validated via numerical experiments. Applications of the proposed bound in a motion segmentation task and in a hand-written digit classification task are given in Section V. The paper is concluded in Section VI. The proofs of the results are given in the Appendix.

C. Notation

We use the following notation in the sequel: matrices, column vectors and scalars are denoted by boldface upper-case letters ($\mathbf{X}$), boldface lower-case letters ($\mathbf{x}$) and italic letters (x), respectively. $\mathbf{I}_N \in \mathbb{R}^{N\times N}$ denotes the identity matrix and $\mathbf{0}_{M\times N} \in \mathbb{R}^{M\times N}$ denotes the zero matrix. The subscripts are omitted when the dimensions are clear from the context. $\mathbf{e}_k$ denotes the k-th basis vector in $\mathbb{R}^N$. The transpose, rank and determinant operators are denoted as $(\cdot)^T, \mathrm{rank}(\cdot)$ and $|\cdot|$, respectively. $\|\mathbf{x}\|$ denotes Euclidean norm of the vector $\mathbf{x}$ and $\|\mathbf{X}\|_2$ denotes the spectral matrix norm of the matrix $\mathbf{X}$. The image of a matrix is denoted by $\mathrm{im}(\cdot)$ and the kernel of a matrix is denoted by $\ker(\cdot)$. The sum of subspaces $\mathcal{A}$ and $\mathcal{B}$ is denoted as $\mathcal{A}+\mathcal{B}$ and the orthogonal complement of $\mathcal{A}$ is denoted as $\mathcal{A}^\perp$. $\log(\cdot)$ denotes the natural logarithm, and the multi-variate Gaussian distribution with the mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ is denoted as $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$. We also use the following asymptotic notation: $f(x)=\mathcal{O}(g(x))$ if $\lim_{x\to\infty}\frac{f(x)}{g(x)}=c$, where $c>0$, and $f(x)=o(g(x))$ if $\lim_{x\to\infty}\frac{f(x)}{g(x)}=0$.

II. Problem Statement

We consider a standard observation model:

$$\mathbf{y}=\mathbf{x}+\mathbf{n} \tag{1}$$

where $\mathbf{y}\in\mathbb{R}^N$ represents the observation vector, $\mathbf{x}\in\mathbb{R}^N$ represents the signal vector and $\mathbf{n}\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I})\in\mathbb{R}^N$ represents observation noise, where σ^2 denotes the noise variance per dimension.[1] We also assume that the signal $\mathbf{x}\in\mathbb{R}^N$ is drawn from a class $c\in\{1,\dots,C\}$ with prior probability $P(c=i)=p_i$, and that the distribution of the signal $\mathbf{x}$ conditioned on a given class $c=i$ is Gaussian with mean zero and (possibly) low-rank covariance matrix $\boldsymbol{\Sigma}_i\in\mathbb{R}^{N\times N}$, i.e.,

$$\mathbf{x}|c=i\sim\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma}_i)\,, \tag{2}$$

with $\mathrm{rank}(\boldsymbol{\Sigma}_i)=r_i\le N$. Therefore, conditioned on a given class $c=i$, the signal lies on the linear subspace spanned by the eigenvectors associated with the positive eigenvalues of the covariance matrix $\boldsymbol{\Sigma}_i$.

The classification problem involves inferring the correct class label c associated with the signal $\mathbf{x}$ from the signal observation $\mathbf{y}$. It is well known that the optimal classification rule, which minimizes the error probability, is given by the Maximum-A-Posteriori (MAP) classifier [1, Ch. 2.3]:

$$\hat{c}=\operatorname*{arg\,max}_{i\in\{1,\dots,C\}} p(c=i|\mathbf{y})=\operatorname*{arg\,max}_{i\in\{1,\dots,C\}} p(\mathbf{y}|c=i)p_i\,, \tag{3}$$

where $p(c=i|\mathbf{y})$ represents the *a posteriori* probability of class label $c=i$ given the observation $\mathbf{y}$ and

$$p(\mathbf{y}|c=i)=\frac{1}{\sqrt{(2\pi)^N|\boldsymbol{\Sigma}_i+\sigma^2\mathbf{I}|}}e^{-\frac{1}{2}\mathbf{y}^T(\boldsymbol{\Sigma}_i+\sigma^2\mathbf{I})^{-1}\mathbf{y}} \tag{4}$$

[1]This noise vector can also model the fact that data does not always lie exactly on a low-dimensional subspace but rather approximately on a low-dimensional subspace [13].

TABLE I
MAIN QUANTITIES USED IN THE ANALYSIS

Subspace	Dimension	Description
$\mathrm{im}(\mathbf{U}_i)$	r_i	signal space of class i
$\mathrm{im}(\tilde{\mathbf{U}}_i)$	$\tilde{r}_i$	mismatched signal space of class i
$\mathrm{im}(\tilde{\mathbf{U}}_{ij}^{\cap})$	$\tilde{r}_{ij}^{\cap}$	intersection of the mismatched signal spaces of classes i and j
$\mathrm{im}(\tilde{\mathbf{U}}'_{ij})$	$\tilde{r}'_{ij}$	subspace of mismatched signal space of class i that is not associated with the mismatched signal space of class j
$\mathrm{im}(\tilde{\mathbf{U}}'_{ji})$	$\tilde{r}'_{ji}$	subspace of mismatched signal space of class j that is not associated with the mismatched signal space of class i
$\mathrm{im}(\mathbf{W}_{ij})$	s_{ij}^{W}	subspace of signal space of class i that is orthogonal to $\mathrm{im}(\tilde{\mathbf{U}}'_{ij})$
$\mathrm{im}(\mathbf{V}_{ij})$	s_{ij}^{V}	subspace of signal space of class i that is not orthogonal to $\mathrm{im}(\tilde{\mathbf{U}}'_{ij})$, i.e. it complements $\mathrm{im}(\mathbf{W}_{ij})$ in $\mathrm{im}(\mathbf{U}_i)$

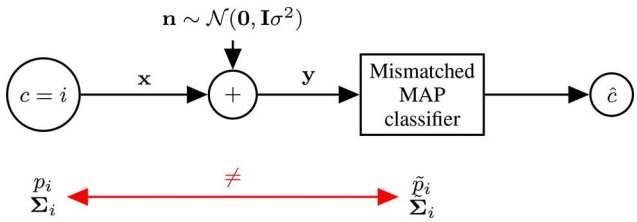

Fig. 1. System Model.

represents the probability density function of the observation $\mathbf{y}$ given the class label $c = i$.

However, we assume that the classifier does not have access to the true signal parameters $p_i, i = 1, \ldots, C$ and $\mathbf{\Sigma}_i, i = 1, \ldots, C$ but rather to a set of mismatched parameters $\tilde{p}_i, i = 1, \ldots, C$ and $\tilde{\mathbf{\Sigma}}_i, i = 1, \ldots, C$, where $\tilde{p}_i$ is the mismatched *a priori* probability of the i-th class and $\tilde{\mathbf{\Sigma}}_i$ is the mismatched covariance matrix associated with the class i with $\mathrm{rank}(\tilde{\mathbf{\Sigma}}_i) = \tilde{r}_i \leq N$.[2] (See Fig. 1.)

Such a Mismatched-MAP (MMAP) classifier delivers the class estimate

$$\tilde{c} = \underset{i \in \{1,\ldots,C\}}{\arg\max}\ \tilde{p}(c = i|\mathbf{y}) = \underset{i \in \{1,\ldots,C\}}{\arg\max}\ \tilde{p}(\mathbf{y}|c = i)\tilde{p}_i\,, \tag{5}$$

where $\tilde{p}(c = i|\mathbf{y})$ denotes the mismatched *a posteriori* probability of class label $c = i$ given observation $\mathbf{y}$ and

$$\tilde{p}(\mathbf{y}|c = i) = \frac{1}{\sqrt{(2\pi)^N |\tilde{\mathbf{\Sigma}}_i + \sigma^2 \mathbf{I}|}} e^{-\frac{1}{2}\mathbf{y}^T (\tilde{\mathbf{\Sigma}}_i + \sigma^2 \mathbf{I})^{-1} \mathbf{y}} \tag{6}$$

denotes the mismatched probability density function of the observation $\mathbf{y}$ given the class label $c = i$.

The probability of error associated with a MMAP classifier is given by:

$$P(e) = \sum_{i=1}^{C} p_i \cdot P(e|c = i) \tag{7}$$

where

$$P(e|c = i) = \int_{-\infty}^{\infty} p(\mathbf{y}|c = i) \cdot u\left(\max_{\substack{j \\ j \neq i}} \log \left(\frac{\tilde{p}_j \tilde{p}(\mathbf{y}|c = j)}{\tilde{p}_i \tilde{p}(\mathbf{y}|c = i)} \right) \right) \mathrm{d}\mathbf{y} \tag{8}$$

and $u(\cdot)$ is the unit-step function. This error probability cannot be calculated in closed form, but it can be easily bounded.

[2]We assume that C and σ^2 are known. Since we study the scenario where $\sigma^2 \to 0$, the assumption that σ^2 is known exactly is immaterial.

TABLE II
RELATIONSHIPS BETWEEN THE QUANTITIES USED IN THE ANALYSIS

1) $\mathrm{im}(\tilde{\mathbf{U}}_i) + \mathrm{im}(\tilde{\mathbf{U}}_j) = \mathrm{im}(\tilde{\mathbf{U}}'_{ij}) + \mathrm{im}(\tilde{\mathbf{U}}_{ij}^{\cap}) + \mathrm{im}(\tilde{\mathbf{U}}'_{ji})$
2) $\mathrm{im}(\tilde{\mathbf{U}}_{ij}^{\cap}) = \mathrm{im}(\tilde{\mathbf{U}}_{ji}^{\cap}) = \mathrm{im}(\tilde{\mathbf{U}}_i) \cap \mathrm{im}(\tilde{\mathbf{U}}_j)$
3) $\mathrm{im}(\tilde{\mathbf{U}}'_{ij}) = \mathrm{im}(\tilde{\mathbf{U}}_i) \cap \mathrm{im}(\tilde{\mathbf{U}}_{ij}^{\cap})^{\perp}$
4) $\mathrm{im}(\tilde{\mathbf{U}}'_{ji}) = \mathrm{im}(\tilde{\mathbf{U}}_j) \cap \mathrm{im}(\tilde{\mathbf{U}}_{ij}^{\cap})^{\perp}$
5) $\mathrm{im}(\tilde{\mathbf{U}}'_{ij})^{\perp} = \mathrm{im}(\tilde{\mathbf{U}}_i)^{\perp} + \mathrm{im}(\tilde{\mathbf{U}}_{ij}^{\cap})$
6) $\mathrm{im}(\mathbf{U}_i) = \mathrm{im}(\mathbf{W}_{ij}) + \mathrm{im}(\mathbf{V}_{ij})$
7) $\mathrm{im}(\mathbf{W}_{ij}) = \mathrm{im}(\mathbf{U}_i) \cap \mathrm{im}(\tilde{\mathbf{U}}'_{ij})^{\perp}$
8) $\mathrm{im}(\mathbf{V}_{ij}) = \mathrm{im}(\mathbf{U}_i) \cap \mathrm{im}(\mathbf{W}_{ij})^{\perp}$

Our goal is to study the performance of the MMAP classifier by establishing conditions, which are a function of the geometry of the true and mismatched signal models as well as the interaction of such geometries, for reliable classification in the low-noise regime i.e., such that $\lim_{\sigma^2 \to 0} P(e) = 0$.

A. Geometrical Description of the Signals

Our characterization of the performance of the MMAP classifier will be expressed via various quantities that embody the geometry of the true signal model, the geometry of the mismatched signal model, and their interplay. The quantities central to the analysis are given in Table I and the relationships between the presented quantities are summarized in Table II.

1) Quantities Associated With the Geometry of the True Signal Model or the Mismatched Signal Model: The signal space corresponding to class i and the mismatched signal space corresponding to class i, which are subspaces of $\mathbb{R}^N$, are denoted as $\mathrm{im}(\mathbf{\Sigma}_i)$ and $\mathrm{im}(\tilde{\mathbf{\Sigma}}_i)$, respectively. An orthonormal basis for $\mathrm{im}(\mathbf{\Sigma}_i)$ is denoted as $\mathbf{U}_i \in \mathbb{R}^{N \times r_i}$ and an orthonormal basis for $\mathrm{im}(\tilde{\mathbf{\Sigma}}_i)$ is denoted as $\tilde{\mathbf{U}}_i \in \mathbb{R}^{N \times \tilde{r}_i}$; these quantities follow directly from the truncated eigenvalue decompositions $\mathbf{\Sigma}_i = \mathbf{U}_i \mathbf{\Lambda}_i \mathbf{U}_i^T$ and $\tilde{\mathbf{\Sigma}}_i = \tilde{\mathbf{U}}_i \tilde{\mathbf{\Lambda}}_i \tilde{\mathbf{U}}_i^T$ where $\mathbf{\Lambda}_i = \mathrm{diag}\left(\lambda_1^i, \lambda_2^i, \ldots, \lambda_{r_i}^i\right) \in \mathbb{R}^{r_i \times r_i}$ and $\tilde{\mathbf{\Lambda}}_i = \mathrm{diag}\left(\tilde{\lambda}_1^i, \tilde{\lambda}_2^i, \ldots, \tilde{\lambda}_{\tilde{r}_i}^i\right) \in \mathbb{R}^{\tilde{r}_i \times \tilde{r}_i}$ are diagonal matrices containing the positive eigenvalues of $\mathbf{\Sigma}_i$ and $\tilde{\mathbf{\Sigma}}_i$, respectively. Note that $\mathrm{im}(\mathbf{\Sigma}_i) = \mathrm{im}(\mathbf{U}_i)$ and $\mathrm{im}(\tilde{\mathbf{\Sigma}}_i) = \mathrm{im}(\tilde{\mathbf{U}}_i)$.

2) Quantities Associated With the Interplay Between the Geometry of the Mismatched Signal Models: We consider quantities that reveal the relationship between the mismatched signal spaces of classes i and j. In particular, such quantities follow from the decomposition of the subspace $\mathrm{im}(\tilde{\mathbf{\Sigma}}_i + \tilde{\mathbf{\Sigma}}_j) = \mathrm{im}(\tilde{\mathbf{U}}_i) + \mathrm{im}(\tilde{\mathbf{U}}_j)$, which spans the mismatched signal subspaces of classes i and j, given by:

$$\mathrm{im}(\tilde{\mathbf{U}}_i) + \mathrm{im}(\tilde{\mathbf{U}}_j) = \underbrace{\mathrm{im}(\tilde{\mathbf{U}}'_{ij}) + \overbrace{\mathrm{im}(\tilde{\mathbf{U}}_{ij}^{\cap})}^{\mathrm{im}(\tilde{\mathbf{U}}_j) = \mathrm{im}(\tilde{\Sigma}_j)} + \mathrm{im}(\tilde{\mathbf{U}}'_{ji})}_{\mathrm{im}(\tilde{\mathbf{U}}_i) = \mathrm{im}(\tilde{\Sigma}_i)}$$

where

- $\tilde{\mathbf{U}}_{ij}^{\cap} \in \mathbb{R}^{N \times \tilde{r}_{ij}^{\cap}}$ represents an orthonormal basis for the intersection $\mathrm{im}(\tilde{\mathbf{\Sigma}}_i) \cap \mathrm{im}(\tilde{\mathbf{\Sigma}}_j)$ and $\tilde{r}_{ij}^{\cap}$ is the dimension of $\mathrm{im}(\tilde{\mathbf{\Sigma}}_i) \cap \mathrm{im}(\tilde{\mathbf{\Sigma}}_j)$. This intersection is associated with class i as well as class j;
- $\tilde{\mathbf{U}}'_{ij} \in \mathbb{R}^{N \times \tilde{r}'_{ij}}$ represents an orthonormal basis for the orthogonal complement of $\mathrm{im}(\tilde{\mathbf{\Sigma}}_i) \cap \mathrm{im}(\tilde{\mathbf{\Sigma}}_j)$ in $\mathrm{im}(\tilde{\mathbf{\Sigma}}_i)$ and $\tilde{r}'_{ij}$ is the codimension of $\mathrm{im}(\tilde{\mathbf{\Sigma}}_i) \cap \mathrm{im}(\tilde{\mathbf{\Sigma}}_j)$ in $\mathrm{im}(\tilde{\mathbf{\Sigma}}_i)$. $\mathrm{im}(\tilde{\mathbf{U}}'_{ij})$ can be interpreted as the subspace of the mismatched signal space corresponding to class i that is only associated with class i and not with class j;
- $\tilde{\mathbf{U}}'_{ji} \in \mathbb{R}^{N \times \tilde{r}'_{ji}}$ represents an orthonormal basis for the orthogonal complement of $\mathrm{im}(\tilde{\mathbf{\Sigma}}_i) \cap \mathrm{im}(\tilde{\mathbf{\Sigma}}_j)$ in $\mathrm{im}(\tilde{\mathbf{\Sigma}}_j)$ and $\tilde{r}'_{ji}$ is the codimension of $\mathrm{im}(\tilde{\mathbf{\Sigma}}_i) \cap \mathrm{im}(\tilde{\mathbf{\Sigma}}_j)$ in $\mathrm{im}(\tilde{\mathbf{\Sigma}}_j)$. $\mathrm{im}(\tilde{\mathbf{U}}'_{ij})$ can be interpreted as the subspace of the mismatched signal space corresponding to class j that is only associated with class j and not with class i.

Note that $\tilde{\mathbf{U}}_{ij}^{\cap}$ together with $\tilde{\mathbf{U}}'_{ij}$ and $\tilde{\mathbf{U}}'_{ji}$ complete the basis for $\mathrm{im}(\tilde{\mathbf{U}}_i)$ and $\mathrm{im}(\tilde{\mathbf{U}}_j)$, respectively, i.e., $\mathrm{im}(\tilde{\mathbf{U}}_i) = \mathrm{im}([\tilde{\mathbf{U}}_{ij}^{\cap}\ \tilde{\mathbf{U}}'_{ij}])$ and $\mathrm{im}(\tilde{\mathbf{U}}_j) = \mathrm{im}([\tilde{\mathbf{U}}_{ij}^{\cap}\ \tilde{\mathbf{U}}'_{ji}])$.

3) Quantities Associated With the Interplay Between the Geometry of the True Signal Model and the Mismatched Signal Model: We also consider quantities that capture the interaction between the signal space corresponding to class i and the mismatched signal spaces of classes i and j. Such quantities are given by the decomposition of $\mathrm{im}(\mathbf{\Sigma}_i) = \mathrm{im}(\mathbf{U}_i)$ given by:

$$\mathrm{im}(\mathbf{U}_i) = \mathrm{im}(\mathbf{W}_{ij}) + \mathrm{im}(\mathbf{V}_{ij})$$

where

- $\mathrm{im}(\mathbf{W}_{ij}) = \mathrm{im}(\mathbf{U}_i) \cap \mathrm{im}(\tilde{\mathbf{U}}'_{ij})^{\perp}$ and $\mathbf{W}_{ij} \in \mathbb{R}^{N \times s_{ij}^W}$ represents an orthonormal basis for $\mathrm{im}(\mathbf{U}_i) \cap \mathrm{im}\left(\tilde{\mathbf{U}}'_{ij}\right)^{\perp}$ where $s_{ij}^W = \dim(\mathrm{im}(\mathbf{U}_i) \cap \mathrm{im}(\tilde{\mathbf{U}}'_{ij})^{\perp})$. $\mathrm{im}(\mathbf{W}_{ij})$ can be interpreted as the subspace of signal space corresponding to class i that is orthogonal to $\mathrm{im}(\tilde{\mathbf{U}}'_{ij})$;
- $\mathrm{im}(\mathbf{V}_{ij}) = \mathrm{im}(\mathbf{U}_i) \cap (\mathrm{im}(\mathbf{U}_i) \cap \mathrm{im}(\tilde{\mathbf{U}}'_{ij})^{\perp})^{\perp}$ and $\mathbf{V}_{ij} \in \mathbb{R}^{N \times s_{ij}^V}$ represents an orthonormal basis for the orthogonal complement of $\mathrm{im}(\mathbf{U}_i) \cap \mathrm{im}(\tilde{\mathbf{U}}'_{ij})^{\perp}$ in $\mathrm{im}(\mathbf{U}_i)$ where $s_{ij}^V = r_i - \dim(\mathrm{im}(\mathbf{U}_i) \cap \mathrm{im}(\tilde{\mathbf{U}}'_{ij})^{\perp})$; then, $s_{ij}^V = \dim(\mathrm{im}(\mathbf{V}_{ij})) = \mathrm{rank}(\mathbf{V}_{ij})$ is the codimension of $\mathrm{im}(\mathbf{W}_{ij})$ in $\mathrm{im}(\mathbf{U}_i)$. $\mathrm{im}(\mathbf{V}_{ij})$ can be interpreted as the subspace of signal space of class i that is not orthogonal to $\mathrm{im}(\tilde{\mathbf{U}}'_{ij})$, i.e., it complements $\mathrm{im}(\mathbf{W}_{ij})$ in $\mathrm{im}(\mathbf{U}_i)$.

Note that $\mathrm{im}([\mathbf{V}_{ij}\ \mathbf{W}_{ij}]) = \mathrm{im}(\mathbf{U}_i)$.

4) Principal Angles and Distance Between Subspaces: Finally, our results will also be expressed via the principal angles between certain subspaces. In particular, consider a subspace $\mathcal{Y}$ with an orthonormal basis $\mathbf{Y} \in \mathbb{R}^{N \times y}$, where $y = \dim(\mathcal{Y})$, and a subspace $\mathcal{Z}$ with an orthonormal basis $\mathbf{Z} \in \mathbb{R}^{N \times z}$, where $z = \dim(\mathcal{Z})$, and define $k = \min(y, z)$. Then the principal angles $0 \le \theta_1 \le \cdots \le \theta_k \le \frac{\pi}{2}$ between $\mathcal{Y}$ and $\mathcal{Z}$ are given by the singular value decomposition (SVD):

$$\mathbf{Y}^T\mathbf{Z} = \mathbf{H}\mathbf{D}\mathbf{J}^T \tag{9}$$

where $\mathbf{H} \in \mathbb{R}^{y \times y}$ and $\mathbf{J} \in \mathbb{R}^{z \times z}$ are orthonormal matrices and $\mathbf{D} \in \mathbb{R}^{y \times z}$ is a rectangular diagonal matrix containing the singular values: $1 \ge d_1 \ge \cdots \ge d_k \ge 0$. Each singular value d_l corresponds to the cosine of the principal angle θ_l between $\mathcal{Y}$ and $\mathcal{Z}$, i.e., $d_l = \cos(\theta_l)$ [25, Ch. 8.7].

The principal angles are used to define various distances on a Grassmann manifold [26]. We will be predominantly using the max correlation distance between two subspaces

$$d_{\max}(\mathcal{Y}, \mathcal{Z}) = d_{\max}(\mathbf{Y}, \mathbf{Z}) = \sin\theta_1 \tag{10}$$

which is a function of the smallest principal angle θ_1, and the min correlation distance between two subspaces

$$d_{\min}(\mathcal{Y}, \mathcal{Z}) = d_{\min}(\mathbf{Y}, \mathbf{Z}) = \sin\theta_k \tag{11}$$

which is a function of the largest principal angle θ_k between the two subspaces. Note that we slightly abuse the notation in the second term of (10) and (11), as $\mathbf{Y}$ and $\mathbf{Z}$ are bases for the subspaces, not subspaces.

5) Interpretation: It is instructive to cast some insight on the role of these various quantities in the characterization of the performance of the MMAP classifier.

Consider a two-class classification problem that involves distinguishing class 1 from class 2 in the low-noise regime (so $\mathbf{y} \approx \mathbf{x}$). It is clear that the MMAP classifier will associate an observation $\mathbf{y} \in \mathrm{im}(\tilde{\mathbf{U}}'_{12})$ with class 1 and an observation $\mathbf{y} \in \mathrm{im}(\tilde{\mathbf{U}}'_{21})$ with class 2; in turn, the MMAP classifier may associate an observation $\mathbf{y} \in \mathrm{im}(\tilde{\mathbf{U}}_{12}^{\cap})$ either with class 1 or 2. In general, the observation associated with class 1 is such that $\mathbf{y} \in \mathrm{im}(\mathbf{U}_1) = \mathrm{im}(\mathbf{V}_{12}) + \mathrm{im}(\mathbf{W}_{12})$.

The following example demonstrates the classification of $\mathbf{y}|c = 1$ by the MMAP classifier where the covariance matrices are assumed to be diagonal.

Example 1: We take the covariance matrices to be

$$\mathbf{\Sigma}_1 = \mathrm{diag}(1,1,1,0)\,, \quad \mathbf{\Sigma}_2 = \mathrm{diag}(0,1,1,1)$$
$$\tilde{\mathbf{\Sigma}}_1 = \mathrm{diag}(1,1,0,0)\,, \quad \tilde{\mathbf{\Sigma}}_2 = \mathrm{diag}(0,1,1,0)\,.$$

The relevant quantities (see Table I) are given as:

$$\mathbf{U}_1 = [\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3]\,, \quad \mathbf{U}_2 = [\mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4]$$
$$\tilde{\mathbf{U}}_1 = [\mathbf{e}_1, \mathbf{e}_2]\,, \quad \tilde{\mathbf{U}}_2 = [\mathbf{e}_2, \mathbf{e}_3]\,,$$

and

$$\tilde{\mathbf{U}}'_{12} = \mathbf{e}_1\,, \quad \tilde{\mathbf{U}}_{12}^{\cap} = \mathbf{e}_2\,, \quad \tilde{\mathbf{U}}'_{21} = \mathbf{e}_3\,.$$

We also determine $\mathrm{im}(\mathbf{W}_{12})$ and $\mathrm{im}(\mathbf{V}_{12})$:

$$\begin{aligned}\mathrm{im}(\mathbf{W}_{12}) &= \mathrm{im}(\mathbf{U}_1) \cap \mathrm{im}(\tilde{\mathbf{U}}'_{12})^{\perp} \\ &= \mathrm{im}([\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3]) \cap \mathrm{im}([\mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4]) = \mathrm{im}([\mathbf{e}_2, \mathbf{e}_3]) \\ \mathrm{im}(\mathbf{V}_{12}) &= \mathrm{im}(\mathbf{U}_1) \cap \mathrm{im}(\mathbf{W}_{12})^{\perp} = \mathbf{e}_1\,.\end{aligned}$$

Assume now that $\mathbf{y} \in \mathrm{im}(\mathbf{V}_{12})$ and note that $\mathrm{im}(\mathbf{V}_{12}) = \mathrm{im}(\tilde{\mathbf{U}}'_{12})$. Therefore, $\mathbf{y} \in \mathrm{im}(\mathbf{V}_{12})$ will be classified as class 1 by the MMAP classifier. In contrast, assume now that $\mathbf{y} \in \mathrm{im}(\mathbf{W}_{12})$ and note that $\mathrm{im}(\mathbf{W}_{12})$ contains $\mathrm{im}(\tilde{\mathbf{U}}'_{21})$. Therefore $\mathbf{y} \in \mathrm{im}(\mathbf{W}_{12})$ may be classified as class 2.

Next, we modify the mismatched model of class 2 as

$$\tilde{\mathbf{\Sigma}}_2 = \mathrm{diag}(0,1,0,1)\,,$$

which leads to $\tilde{\mathbf{U}}'_{21} = \mathbf{e}_4$. Note now that $\mathrm{im}(\mathbf{W}_{12})$ does not contain $\mathrm{im}(\tilde{\mathbf{U}}'_{21})$ and $\mathbf{y} \in \mathrm{im}(\mathbf{W}_{12})$ will not be associated uniquely with class 2 by the MMAP classifier.

It is now clear that the relationship between subspaces $\mathrm{im}(\mathbf{W}_{12})$ and $\mathrm{im}(\tilde{\mathbf{U}}'_{21})$ will play a role in the characterization of conditions for perfect classification in the low-noise regime.

The next example demonstrates the role of principal angles in the conditions for perfect classification in the low-noise regime.

Example 2: We take the signal space bases as:

$$\mathbf{U}_1 = [0,1]^T, \quad \mathbf{U}_2 = \left[\cos\left(\frac{\pi}{4}\right), \sin\left(\frac{\pi}{4}\right)\right]^T$$
$$\tilde{\mathbf{U}}_1 = \left[\cos\left(\frac{5\pi}{6}\right), \sin\left(\frac{5\pi}{6}\right)\right]^T, \quad \tilde{\mathbf{U}}_2 = \mathbf{U}_2\,.$$

The relevant quantities (see Table I) are given as:

$$\tilde{\mathbf{U}}'_{12} = \tilde{\mathbf{U}}_1, \quad \tilde{\mathbf{U}}^{\cap}_{12} = \{0\}, \quad \tilde{\mathbf{U}}'_{21} = \tilde{\mathbf{U}}_2$$

and $\mathbf{W}_{12} = \{0\}, \mathbf{V}_{12} = \mathbf{U}_1$. The geometry of the signals and decision regions is presented in Fig. 2(a). Note now that $\mathbf{y}|c = 1 \in \mathrm{im}(\mathbf{U}_1)$ can potentially be associated to the correct class 1 depending on the distance (computed according to an appropriate metric) between $\mathrm{im}(\mathbf{V}_{12})$ and $\mathrm{im}(\tilde{\mathbf{U}}'_{12})$ and the distance between $\mathrm{im}(\mathbf{V}_{12})$ and $\mathrm{im}(\tilde{\mathbf{U}}'_{21})$. In particular, the angle between $\mathrm{im}(\mathbf{V}_{12})$ and $\mathrm{im}(\tilde{\mathbf{U}}'_{12})$ is greater than the angle between $\mathrm{im}(\mathbf{V}_{12})$ and $\mathrm{im}(\tilde{\mathbf{U}}'_{21})$, which leads to misclassification of signals from class 1. On the other hand, if we take

$$\tilde{\mathbf{U}}_1 = \left[\cos\left(\frac{4\pi}{6}\right), \sin\left(\frac{4\pi}{6}\right)\right]^T$$

the angle between $\mathrm{im}(\mathbf{V}_{12})$ and $\mathrm{im}(\tilde{\mathbf{U}}'_{12})$ is smaller than the angle between $\mathrm{im}(\mathbf{V}_{12})$ and $\mathrm{im}(\tilde{\mathbf{U}}'_{21})$, which leads to perfect classification of signals from class 1 in the low-noise regime. This case is presented in Fig. 2(b).

The ensuing analysis shows how these various quantities—which are readily computed from the underlying geometry of the true subspaces and the mismatched ones—can be used as a proxy to define sufficient conditions for perfect classification in the low-noise regime. In particular, these quantities bypass the need to compute the decision regions associated with the MMAP classifier in order to quantify the performance.

III. Conditions for Reliable Classification

We now consider (sufficient) conditions for reliable classification in the low-noise regime. We derive these conditions directly from a low-noise expansion of an upper bound to the error probability associated with the MMAP classifier.

The following upper bound to the probability of error associated with a MMAP classifier will play a key role in the analysis.

Theorem 1: Set $\alpha_{ij} > 0\,\forall(i,j)\,, i \neq j$. Set

$$\mathbf{\Sigma}_{ij} = (\mathbf{\Sigma}_i + \sigma^2\mathbf{I})^{-1} + \alpha_{ij}(\tilde{\mathbf{\Sigma}}_j + \sigma^2\mathbf{I})^{-1} - \alpha_{ij}(\tilde{\mathbf{\Sigma}}_i + \sigma^2\mathbf{I})^{-1}\,. \quad (12)$$

Then the error probability associated with the MMAP classifier in (7) can be bounded as follows:

- If $\mathbf{\Sigma}_{ij} \succ \mathbf{0}\ \forall(i,j)$ with $i \neq j$, then

$$P(e) \leq \bar{P}(e) = \sum_{i=1}^{C} p_i \cdot \left(\sum_{j=1, j\neq i}^{C} \bar{P}(e_{ij})\right) \quad (13)$$

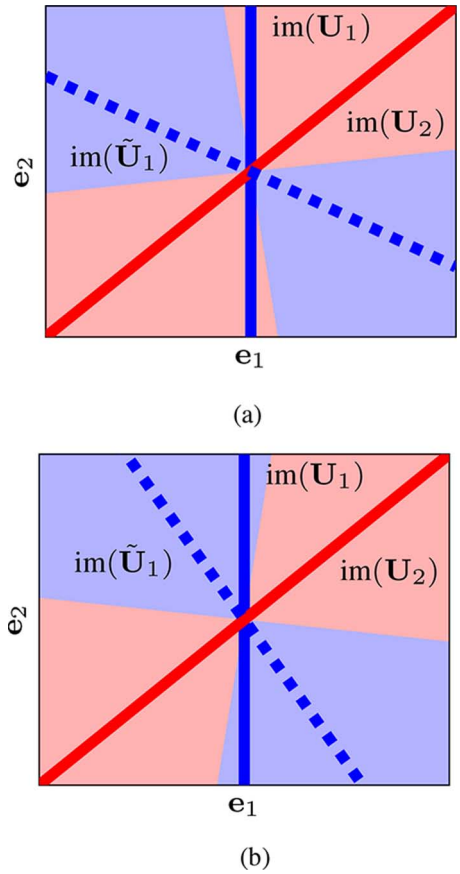

Fig. 2. The two plots illustrate the decision regions associated with the 2-class MMAP classifier for different values of $\mathbf{U}_1, \mathbf{U}_2, \tilde{\mathbf{U}}_1$ and $\tilde{\mathbf{U}}_2$ in the limit $\sigma^2 \to 0$. Transparent blue and red regions indicate the decision region where MMAP outputs class labels 1 and 2, respectively. Blue line represent the signal subspace $\mathrm{im}(\mathbf{U}_1)$ and red line represent the signal subspace $\mathrm{im}(\mathbf{U}_2)$. Dashed blue line represents the mismatched signal subspace $\mathrm{im}(\tilde{\mathbf{U}}_1)$. The subspace bases are given in Example 2. (a) Example of wrong classification with the MMAP classifier, (b) Example of correct classification with the MMAP classifier.

where

$$\bar{P}(e_{ij}) = \left(\frac{\tilde{p}_j}{\tilde{p}_i}\sqrt{\frac{|\tilde{\mathbf{\Sigma}}_i + \sigma^2\mathbf{I}|}{|\tilde{\mathbf{\Sigma}}_j + \sigma^2\mathbf{I}|}}\right)^{\alpha_{ij}} \cdot \left(|\mathbf{\Sigma}_i + \sigma^2\mathbf{I}||\mathbf{\Sigma}_{ij}|\right)^{-\frac{1}{2}}. \quad (14)$$

- If $\exists(i,j)$ with $i \neq j : \mathbf{\Sigma}_{ij} \not\succ \mathbf{0}$ then $P(e) \leq \bar{P}(e) = 1$.

Proof: The proof appears in Appendix. ■

This upper bound to the error probability of the MMAP classifier can capture the fact that the error probability may tend to zero as the noise power approaches zero, depending on the relation between the true signal parameters and the mismatched ones. In particular, the upper bound to the misclassification probability of class i is expressed as a function of the covariance matrix of class i, the mismatched covariance matrix of class i and the mismatched covariance matrices of classes $j \neq i$. In contrast, the bound proposed in [23] expresses the upper bound to the error probability as a function of the sum of f-divergences between the true and the mismatched distributions of class i, for all classes i. Therefore, it does not capture the interplay between mismatched models of different classes. In addition, when specialized to the proposed signal model, the bound in [23] always predicts the presence of an error floor (see Section IV).

The following Theorem presents a low-noise expansion of the upper bound to the error probability of the MMAP classifier.

Theorem 2: The upper bound to the error probability of the MMAP classifier in (13) can be expanded as follows:

- Assume that $\forall (i,j)\,, i \neq j$, the following conditions hold:

$$\mathrm{im}(\mathbf{W}_{ij}) \subseteq \mathrm{im}\left(\tilde{\mathbf{U}}'_{ji}\right)^{\perp}\,, \tag{15}$$

$$\mathbf{V}_{ij}^T\left(\tilde{\mathbf{U}}'_{ij}\left(\tilde{\mathbf{U}}'_{ij}\right)^T - \tilde{\mathbf{U}}'_{ji}\left(\tilde{\mathbf{U}}'_{ji}\right)^T\right)\mathbf{V}_{ij} \succ \mathbf{0} \text{ or } s_{ij}^V = 0\,, \tag{16}$$

and take $d = \min_{(i\neq j)} d_{ij}$, where

$$d_{ij} = \frac{1}{2}\left(s_{ij}^V + \alpha_{ij}(\tilde{r}_j - \tilde{r}_i)\right)\,, \tag{17}$$

and $\alpha_{ij} \in \left(0, \alpha_{ij}^0\right)$ where the value of $\alpha_{ij}^0 > 0$ is given in the Appendix. Then

— If $d \leq 0$:

$$\bar{P}(e) = \mathcal{O}(1), \quad \sigma^2 \to 0\,. \tag{18}$$

— If $d > 0$:

$$\bar{P}(e) = A \cdot (\sigma^2)^d + o\left((\sigma^2)^d\right)\,, \quad \sigma^2 \to 0\,, \tag{19}$$

where $A > 0$.

- Assume $\exists (i,j), i \neq j$, such that conditions (15) or (16) do not hold. Then

$$\bar{P}(e) = \mathcal{O}(1), \quad \sigma^2 \to 0\,. \tag{20}$$

Proof: The proof appears in Appendix. ∎

The expansion of the upper bound to the error probability embodied in Theorem 2 provides a set of conditions, which are a function of the geometry of the true signal model, the geometry of the mismatched signal model, and the interaction of the geometries, that enable us to understand whether or not the upper bound to the error probability may exhibit an error floor. In particular, in view of the fact that we use the union bound in order to bound the error probability of a multi-class problem in terms of the error probabilities of two-class problems, these conditions have to hold for every pair of class labels $(i,j), i \neq j$. We can note that:

- The upper bound to the probability of error exhibits an error floor if either (15) or (16) are not satisfied for some pair $(i,j), i \neq j$. The interpretation of condition (15) is straightforward by noting that the subspace $\mathrm{im}(\mathbf{W}_{ij})$ contains vectors of class i that are orthogonal to the subspace $\mathrm{im}\left(\mathbf{U}'_{ij}\right)$, which is the subspace uniquely associated with class i. Then, condition (15) states that such vectors must also be orthogonal to the mismatched subspace uniquely associated with class j, i.e., $\mathrm{im}\left(\mathbf{U}'_{ji}\right)$. The interpretation of condition (16) is obtained by reformulating the expression as:

$$\begin{aligned}
&\mathbf{V}_{ij}^T\left(\tilde{\mathbf{U}}'_{ij}\left(\tilde{\mathbf{U}}'_{ij}\right)^T - \tilde{\mathbf{U}}'_{ji}\left(\tilde{\mathbf{U}}'_{ji}\right)^T\right)\mathbf{V}_{ij} \succ \mathbf{0}\\
&\qquad\Longleftrightarrow\\
&\mathbf{x}^T\tilde{\mathbf{U}}'_{ij}\left(\tilde{\mathbf{U}}'_{ij}\right)^T\mathbf{x} > \mathbf{x}^T\tilde{\mathbf{U}}'_{ji}\left(\tilde{\mathbf{U}}'_{ji}\right)^T\mathbf{x} \quad \forall \mathbf{x} \in \mathrm{im}(\mathbf{V}_{ij})\\
&\qquad\Longleftrightarrow\\
&\left\|\left(\tilde{\mathbf{U}}'_{ij}\right)^T\mathbf{x}\right\|_2 > \left\|\left(\tilde{\mathbf{U}}'_{ji}\right)^T\mathbf{x}\right\|_2 \quad \forall \mathbf{x} \in \mathrm{im}(\mathbf{V}_{ij})\,.
\end{aligned} \tag{21}$$

Note that $\|(\tilde{\mathbf{U}}'_{ij})^T\mathbf{x}\|_2 = \|\tilde{\mathbf{U}}_{ij}(\tilde{\mathbf{U}}'_{ij})^T\mathbf{x}\|_2$ is the norm of the projection of $\mathbf{x}$ onto $\mathrm{im}\left(\tilde{\mathbf{U}}'_{ij}\right)$. Therefore, (16) requires that the norm of vectors in $\mathrm{im}(\mathbf{V}_{ij})$, which are associated with class i, projected onto $\mathrm{im}\left(\tilde{\mathbf{U}}'_{ij}\right)$, which is also associated with class i, is greater than the norm of vectors in $\mathrm{im}(\mathbf{V}_{ij})$ projected onto $\mathrm{im}\left(\tilde{\mathbf{U}}'_{ji}\right)$, which is associated with class j.

Equation (21) is also implied by

$$d_{\min}(\mathbf{V}_{ij}, \tilde{\mathbf{U}}'_{ij}) < d_{\max}(\mathbf{V}_{ij}, \tilde{\mathbf{U}}'_{ji}) \tag{22}$$

which requires that the largest principal angle between $\mathrm{im}(\mathbf{V}_{ij})$ and $\mathrm{im}\left(\tilde{\mathbf{U}}'_{ij}\right)$ is smaller than the smallest principal angle between $\mathrm{im}(\mathbf{V}_{ij})$ and $\mathrm{im}\left(\tilde{\mathbf{U}}'_{ji}\right)$.[3] Demonstration of this condition is provided by Example 2 in Section II.A.

- On the other hand, the upper bound to the probability of error does not exhibit an error floor if conditions (15) and (16) are satisfied for all pairs $(i,j), i \neq j$ and $d > 0$. In particular, necessary and sufficient conditions for $d > 0$ depend on the dimension of the various subspaces and their relation, i.e., $s_{ij}^V > 0$ for all pairs (i,j) such that $\tilde{r}_j - \tilde{r}_i \leq 0$ is necessary and sufficient for $d > 0$. For example, if the rank of all covariance matrices associated to the mismatched model is the same, i.e., if $\tilde{r}_i = \tilde{r}$, for $i = 1, \ldots, C$, then $s_{ij}^V > 0, \forall (i,j), i \neq j$ is necessary and sufficient for $d > 0$. Note that a positive value for s_{ij}^V indicates that there is at least one vector in $\mathrm{im}(\mathbf{U}_i)$ that is not contained in $\mathrm{im}(\tilde{\mathbf{U}}'_{ij})^{\perp}$, or equivalently, there exists at least one vector in $\mathrm{im}(\mathbf{U}_i)$ that has a non-zero projection onto $\mathrm{im}(\tilde{\mathbf{U}}'_{ij})$, therefore leading to reliable classification of signals from class i.
- Note that parameters α_{ij} do not play a role in the characterization of the necessary and sufficient conditions for $d > 0$. In fact, the conditions for $d_{ij} > 0$ do not depend on a particular value of α_{ij}, provided that $\alpha_{ij} \in \left(0, \frac{1}{|\tilde{r}_j - \tilde{r}_i|}\right)$.
- Note also that the value of d represents a measure of robustness against noise in the low-noise regime, as it determines the speed at which the upper bound of the error probability decays with $1/\sigma^2$. In particular, higher values of d will represent higher robustness against noise, in the low-noise regime. For example, on assuming $\tilde{r}_i = \tilde{r}$ for $i = 1, \ldots, C$, we observe that larger values of s_{ij}^V correspond to larger values of d. Therefore, as expected, higher levels of robustness are obtained when the overlap between $\mathrm{im}(\mathbf{U}_i)$ and $\mathrm{im}(\tilde{\mathbf{U}}'_{ij})^{\perp}$, i.e., dimension of $\mathrm{im}(\mathbf{W}_{ij})$, is reduced.

We also discuss how the value of d_{ij} in (17) relates to the value of d_{ij} for the non-mismatched case.[4] In particular, we assume that $r_i = r_j = \tilde{r}_i = \tilde{r}_j$ and that the true and the mismatched covariance matrices are diagonal. Then for the non-mismatched case

$$d_{ij} = \frac{1}{2}(r_i - \dim(\mathrm{im}(\mathbf{U}_i) \cap \mathrm{im}(\mathbf{U}_j)))$$

[3]The detailed derivation of this statement is reported in Appendix.

[4]Note that our comparison involves upper bounds on the error probabilities rather than the actual error probabilities.

and for the mismatched case

$$d_{ij} = \frac{1}{2}(r_i - \dim(\mathrm{im}(\mathbf{U}_i) \cap \mathrm{im}(\tilde{\mathbf{U}}_j)) - \dim(\mathrm{im}(\mathbf{U}_i) \cap \ker(\tilde{\mathbf{U}}_i) \cap \ker(\tilde{\mathbf{U}}_j)))\,.$$

Therefore, in the non-mismatched case d_{ij} is at most r_i and it decreases as the dimension of the intersection of the signal spaces of classes i and j increases. In the mismatched case d_{ij} is also at most r_i, but it decreases as the dimension of the intersection of the signal space of class i and the mismatched signal space of class j increases, and as the dimension of the intersection of the signal space of class i and the noise subspace of the mismatched classifier, i.e., $\ker(\tilde{\mathbf{U}}_i) \cap \ker(\tilde{\mathbf{U}}_j)$, increases. It can also be easily verified that the value of d for a non-mismatched 2 class problem obtained in [27] matches the value of d derived via the proposed bound. Note that the bound analyzed in [27] is different than the bound proposed in this paper and it is only valid for non-mismatched models.

- The constant A in (19) distinguishes the upper bounds for different mismatched models with a constant d, in the low-noise regime, and is determined as the ratio of volumes of subspaces associated with true and mismatched signal subspaces and their interaction. See Appendix for the detailed expression.

Theorem 2 therefore leads immediately to sufficient conditions for reliable classification in the low-noise regime.

Corollary 1: If

$$\mathrm{im}(\mathbf{W}_{ij}) \subseteq \mathrm{im}\left(\tilde{\mathbf{U}}'_{ji}\right)^{\perp} \ \forall (i,j), i \neq j\,, \tag{23}$$

$$\mathbf{V}_{ij}^T \left(\tilde{\mathbf{U}}'_{ij}\left(\tilde{\mathbf{U}}'_{ij}\right)^T - \tilde{\mathbf{U}}'_{ji}\left(\tilde{\mathbf{U}}'_{ji}\right)^T\right)\mathbf{V}_{ij} \succ \mathbf{0} \ \forall (i,j), i \neq j \tag{24}$$

and $s_{ij}^V > 0 \ \forall (i,j)$ such that $\tilde{r}_j - \tilde{r}_i \leq 0$, then $\lim_{\sigma^2 \to 0} P(e) = 0$.

Proof: This follows directly from Theorem 1, since $\lim_{\sigma^2 \to 0} \bar{P}(e) = 0 \Longrightarrow \lim_{\sigma^2 \to 0} P(e) = 0$. ∎

Corollary 2: If

$$d_{\min}(\mathbf{U}_i, \tilde{\mathbf{U}}_i) < d_{\max}(\mathbf{U}_i, \tilde{\mathbf{U}}_j) \text{ and } s_{ij}^V > 0 \,\forall (i,j), i \neq j \tag{25}$$

then $\lim_{\sigma^2 \to 0} P(e) = 0$.

Proof: The proof appears in Appendix. ∎

Note that the conditions in Corollary 2 are implied by (hence are weaker) the conditions in Corollary 1.

The conditions for reliable classification are particularly simple for the scenario where true and mismatched covariance matrices are diagonal.

Corollary 3: Assume $\mathbf{\Sigma}_i, \tilde{\mathbf{\Sigma}}_i, i = 1, \ldots, C$ are diagonal. If

$$\mathrm{im}(\mathbf{W}_{ij}) \subseteq \mathrm{im}\left(\tilde{\mathbf{U}}'_{ji}\right)^{\perp} \ \forall (i,j), i \neq j \tag{26}$$

and $s_{ij}^V > 0 \,\forall (i,j)$ such that $\tilde{r}_j - \tilde{r}_i \leq 0$, then $\lim_{\sigma^2 \to 0} P(e) = 0$. ∎

Proof: The proof appears in Appendix.

Note that in diagonal case the sufficient conditions for perfect classification simplify only to inclusion of subspaces. Recall the Example 1 where we demonstrate that the signals in $\mathrm{im}(\mathbf{W}_{ij})$ may be associated with class i or with class j. Condition (26) formalizes the intuition that the signals in $\mathrm{im}(\mathbf{W}_{ij})$ must be orthogonal to $\mathrm{im}\left(\tilde{\mathbf{U}}'_{ji}\right)$, which is uniquely associated with class j.

We finally illustrate how our conditions cast insight onto the impact of mismatch for a two-class case where the mismatched subspaces are a rotated version of the true signal subspaces.

Example 3: Consider a two-class classification problem where $\mathbf{x}|c=1 \sim \mathcal{N}\left(\mathbf{0}, \mathbf{U}_1\mathbf{U}_1^T\right)$ and $\mathbf{x}|c=2 \sim \mathcal{N}\left(\mathbf{0}, \mathbf{U}_2\mathbf{U}_2^T\right)$ and

$$\tilde{\mathbf{U}}_1 = \mathbf{Q}_1\mathbf{U}_1\,,\ \tilde{\mathbf{U}}_2 = \mathbf{Q}_2\mathbf{U}_2\,, \tag{27}$$

where $\mathbf{Q}_1 \in \mathbb{R}^{N\times N}$ and $\mathbf{Q}_2 \in \mathbb{R}^{N\times N}$ are orthogonal matrices, and $s_{12}, s_{21} > 0$.[5] By defining

$$\epsilon_1 = \|\mathbf{I} - \mathbf{Q}_1\|_2\,,\ \epsilon_2 = \|\mathbf{I} - \mathbf{Q}_2\|_2 \tag{28}$$

$$\delta_{12} = \max_l \cos\theta_l^{12} = \sqrt{1 - d_{\min}^2(\mathbf{U}_1, \mathbf{U}_2)}, \tag{29}$$

it follows that

$$1 - \delta_{12} > \epsilon_1 + \epsilon_2 \Longrightarrow \lim_{\sigma^2 \to 0} P(e) = 0\,. \tag{30}$$

The proof is in the Appendix.

This example provides sufficient conditions for reliable classification in the low-noise regime by relating the degree of mismatch—measured in terms of the spectral norm of the matrix $\mathbf{I} - \mathbf{Q}_i, i = 1, 2$—to the minimum principal angle between subspaces. It states that the larger the minimum principal angle between the spaces spanned by signals of class 1 and class 2, i.e., the larger $1 - \delta_{12}$, the more robust is the classifier against mismatch, where the level of mismatch is measured by $\epsilon_1 + \epsilon_2$. The maximum robustness against mismatch is obtained when $\delta_{12} = 0$, which means that signals from class 1 and class 2 are orthogonal.

This example also provides a rationale for state-of-the-art feature extraction mechanisms where the signal classes are transformed via a linear operator $\mathbf{\Phi}$ prior to classification. In particular, assume that $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ correspond to the covariances of signals in class 1 and 2 after the transformation $\mathbf{\Phi}$: the example suggests that the operator $\mathbf{\Phi}$ should transform the signal covariances so that δ_{12} is small (i.e., so that the signals from class 1 and 2 are close to orthogonal) in order to create robustness against mismatch. Such an approach is considered, for example, in [28], where signals are transformed by a matrix, which promotes large principal angles between the subspaces. Note that the work in [28] is not motivated on the basis of robustness against mismatch, but rather on intuitive insight about classification of signals that lie on subspaces.

IV. Numerical Results

We now show that our conditions for reliable classification in the low-noise regime are sharp, by revisiting the Examples

[5]This condition insures that the mismatched subspaces are not completely orthogonal to the signal subspaces.

TABLE III
MISMATCH EXAMPLES GIVEN IN SECTION II.A

	Model	Theory $\lim_{\sigma^2\to 0}\bar{P}(e)$	Simulation $\lim_{\sigma^2\to 0}P(e)$
(a)	$\mathbf{U}_1=[\mathbf{e}_1,\mathbf{e}_2,\mathbf{e}_3]\,,\mathbf{U}_2=[\mathbf{e}_2,\mathbf{e}_3,\mathbf{e}_4]\,,\tilde{\mathbf{U}}_1=[\mathbf{e}_1,\mathbf{e}_2]\,,\tilde{\mathbf{U}}_2=[\mathbf{e}_2,\mathbf{e}_3]$	>0	>0
(b)	$\mathbf{U}_1=[\mathbf{e}_1,\mathbf{e}_2,\mathbf{e}_3]\,,\mathbf{U}_2=[\mathbf{e}_2,\mathbf{e}_3,\mathbf{e}_4]\,,\tilde{\mathbf{U}}_1=[\mathbf{e}_1,\mathbf{e}_2]\,,\tilde{\mathbf{U}}_2=[\mathbf{e}_2,\mathbf{e}_4]$	$=0$	$=0$
(c)	$\mathbf{U}_1=[0,1]^T\,,\mathbf{U}_2=[\cos\left(\frac{\pi}{4}\right),\sin\left(\frac{\pi}{4}\right)]\,,\tilde{\mathbf{U}}_1=[\cos\left(\frac{5\pi}{6}\right),\sin\left(\frac{5\pi}{6}\right)]\,,\tilde{\mathbf{U}}_2=\mathbf{U}_2$	>0	>0
(d)	$\mathbf{U}_1=[0,1]^T\,,\mathbf{U}_2=[\cos\left(\frac{\pi}{4}\right),\sin\left(\frac{\pi}{4}\right)]\,,\tilde{\mathbf{U}}_1=[\cos\left(\frac{4\pi}{6}\right),\sin\left(\frac{4\pi}{6}\right)]\,,\tilde{\mathbf{U}}_2=\mathbf{U}_2$	$=0$	$=0$

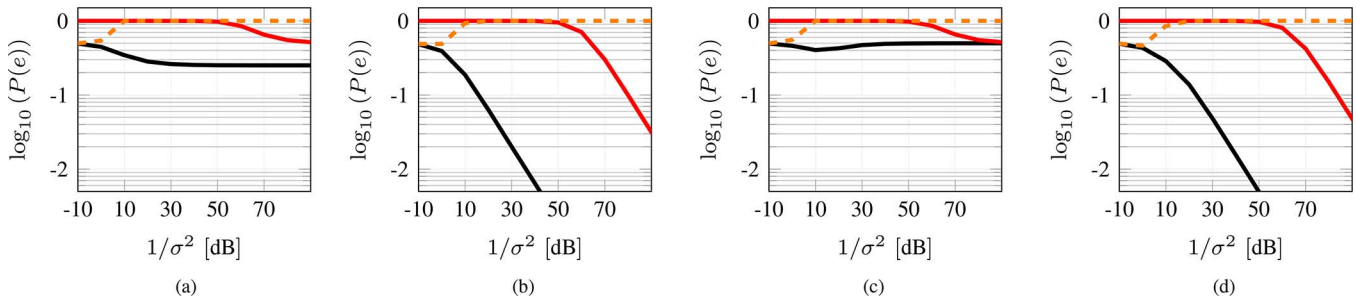

Fig. 3. Simulation results for the examples in Table I. In all plots, the black line corresponds to the true error probability $P(e)$ obtained via simulation, the red line corresponds to the proposed upper bound to error probability $\bar{P}(e)$ given in Theorem 1 and the dashed orange line corresponds to the upper bound in [23] (with KL-divergence). (a) $\lim_{\sigma^2\to 0}P(e)>0$. (b) $\lim_{\sigma^2\to 0}P(e)=0$. (c) $\lim_{\sigma^2\to 0}P(e)>0$. (d) $\lim_{\sigma^2\to 0}P(e)=0$.

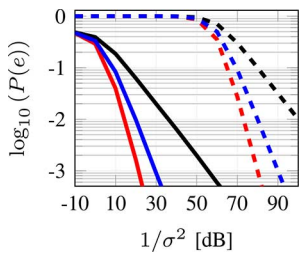

Fig. 4. Black, blue and red lines correspond to the simulated error probabilities for examples given by (32), (33) and (34), respectively. Dashed black, blue and red lines correspond to the upper bound given in Theorem 1 for examples given by (32), (33) and (34), respectively.

1 and 2 presented in Section II.A. The model parameters and results are summarized in Table III.

Fig. 3 shows the estimated true error probability, which is obtained from simulation[6], the upper bound to the error probability given in Theorem 1 and the bound proposed in [23] (using the KL-divergence) as a function of σ^2. Note that the proposed upper bound to the error probability and the derived sufficient conditions give a sharp predictions of an error floor, and also that the bound proposed in [23] always exhibits an error floor.

In case (a), condition (15) in Theorem 2 is not satisfied for $(i,j)=(1,2)$, i.e., $\mathrm{im}(\mathbf{W}_{12})=\mathrm{im}([\mathbf{e}_2,\mathbf{e}_3])\not\subseteq\mathrm{im}\left(\tilde{\mathbf{U}}'_{21}\right)^{\perp}=\mathrm{im}(\mathbf{e}_3)^{\perp}$, therefore, via Theorem 2 we conclude that the upper bound exhibits an error floor. The results in Fig. 3 show that in this case the true error probability also exhibits an error floor. In case (b), conditions (15) and (16) are satisfied and $d>0$. Therefore, via Theorem 2, the upper bound to the error probability approaches zero, which also implies that the true error probability approaches zero, in the low-noise regime.

For cases (c) and (d) the intuition is provided by the Corollary 2, where in the case of the one-dimensional subspaces the concept of principal angles simply reduces to the notion of angle between two lines. In particular, in case (c) the condition (25) in Corollary 2 is not satisfied for $(i,j)=(1,2)$, and we observe an error floor in the true error probability. On the contrary, in case (d) the conditions (25) in Corollary 2 are satisfied which immediately implies perfect classification in the low-noise regime.

We now explore how different mismatched models affect the value of d. Consider the following 2-class example in $\mathbb{R}^6$ with orthonormal basis vectors $\mathbf{e}_i, i=1,\ldots,6$, where the signal spaces are:

$$\mathbf{U}_1=[\mathbf{e}_1,\mathbf{e}_2,\mathbf{e}_3]\,,\quad \mathbf{U}_2=[\mathbf{e}_4,\mathbf{e}_5,\mathbf{e}_6] \tag{31}$$

and various mismatched signal spaces are:

$$\tilde{\mathbf{U}}_1=[\mathbf{e}_1]\,,\quad \tilde{\mathbf{U}}_2=[\mathbf{e}_4] \tag{32}$$
$$\tilde{\mathbf{U}}_1=[\mathbf{e}_1,\mathbf{e}_2]\,,\quad \tilde{\mathbf{U}}_2=[\mathbf{e}_4,\mathbf{e}_5] \tag{33}$$
$$\tilde{\mathbf{U}}_1=\mathbf{U}_1\,,\quad \tilde{\mathbf{U}}_2=\mathbf{U}_2\,. \tag{34}$$

It is straightforward to verify that the sufficient conditions for perfect classification given by Theorem 2 hold for all three pairs of mismatch models (32), (33) and (34). Furthermore, one can also determine the values of d as 0.5, 1 and 1.5, where values of d do not depend on α_{ij}, for the mismatched models given by (32), (33) and (34), respectively. As observed in Section III, a higher value of d implies a higher robustness to noise. Simulation results of the true error probability and the values of the upper bounds as given in Theorem 1 are plotted in Fig. 4. One can observe that increasing values of d (associated with the upper bound to the error probability) correspond to steeper decrease of the true error probability as $\sigma^2\to 0$. Moreover, the values of d obtained via the upper bound match the values of d obtained from the simulation of the true error probability for all the examples (32)–(34).

V. APPLICATIONS

We finally show how theory can also capture the impact of mismatch on classification performance in applications involving real world data. We consider a motion segmentation

[6]In our simulations, signals are drawn independently from the true distribution and are classified by the MMAP classifier.

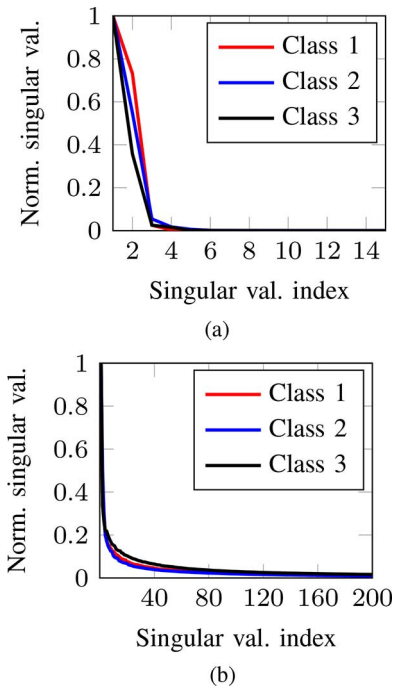

Fig. 5. Normalized singular values of data matrices corresponding to: (a) motions in the Hopkins dataset and (b) digits in the MNIST dataset. For Hopkins dataset only the first 15 out of 58 singular values are shown. For MNIST dataset only the first 200 out of $784 (= 28 \times 28)$ singular values are shown for the first 3 classes. (a) Hopkins dataset, (b) MNIST dataset.

application, where the goal is to segment a video in multiple rigidly moving objects, and a hand-written digit classification application. In both tasks we concentrate on a supervised learning approach, in which we are given a number of labeled samples, which are used to estimate the model (training set) and a number of unlabeled samples that we want to classify (testing set). Our aim is to determine the minimum size of the training set needed to guarantee reliable classification of the testing set.

A. Datasets

For the motions segmentation task we use the Hopkins 155 dataset [29], which consists of video sequences with 2 or 3 motions in each video. The motion segmentation problem is usually solved by extracting feature points from the video and tracking their position over different frames. In more details, in this application, observation vectors $\mathbf{y}$ are obtained by stacking the coordinate values associated to a given feature point corresponding to different frames, and the objective of motion segmentation is that of classifying each feature point as belonging to one of the moving objects in the video [10].

Theoretical results show that the features points trajectories belonging to a given motion lie on approximately 3 dimensional affine space or 4 dimensional linear space [10]–[12]. We validate that empirically by observing the decay of singular values of the data matrix associated with a given motion, which is shown in Fig. 5(a). Note that singular values are close to zero for singular value indices that are greater than 4.

For the experiment we consider a video with 3 motions[7], where number of samples of class 1, class 2 and class 3 is 236, 142 and 114, respectively. The rule adopted to pick the video

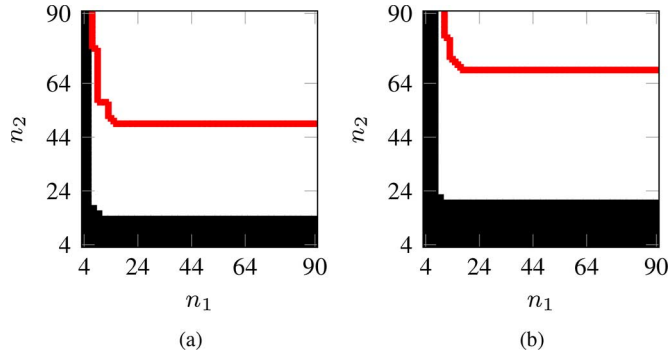

Fig. 6. Phase transition of true error rate and phase transition given by the upper bound to the error probability as a function of number of training samples n_1, n_2. Black corresponds to an error floor of the true error rate, white corresponds to reliable classification, and red line denotes the phase transition predicted via Theorem 2 for a given probability p_p. (a) $p_p = 0.8, n_3 = 70$, (b) $p_p = 0.9, n_3 = 70$.

was the maximal possible feature points—samples—for each motion. The ranks of the true and the mismatched covariances is always set to 4. We also split the dataset samples randomly into a training set and a testing set, where the training set contains $n_{\max} = 90$ samples per class.

For the hand-written digit classification task we use the MNIST dataset [30], which consists of 28×28 grey scale images of hand-written digits between 0 and 9. We obtain observation vectors $\mathbf{y}$ by vectorizing the images.

The decay of singular values associated with the data matrix of MNIST digits is shown in Fig. 5(b). Note that the singular values do not approach zero as fast as in the case of the Hopkins dataset. We can argue that the classes in the MNIST dataset are only "approximately low-rank", i.e., the covariance matrix associated with the class i can be expressed as $\mathbf{\Sigma}_i = \bar{\mathbf{\Sigma}}_i + \delta \mathbf{I}$, where $\bar{\mathbf{\Sigma}}_i$ is low-rank and $\delta > 0$ accounts for the deviation from the perfectly low-rank model. In view of the presented signal model this can be interpreted as a classification of signals with low-rank covariance matrix $\bar{\mathbf{\Sigma}}_i$ at finite $\sigma^2 = \delta$. The sufficient conditions for perfect classification in the case of "approximately low-rank" model will now predict what number of training samples is required to achieve the best possible error rate for the given classification problem.

The ranks of the true and the mismatched covariances is always set to 20 in the experiments. Such rank leads to capturing approximately 90% of the energy of the signals. The split into training and testing set is provided by the MNIST dataset, where the training set contains approximately $n_{\max} = 5000$ samples per class.

B. Methodology

We obtain the class-conditioned covariance matrices by retaining only the first r principal components of the estimated covariances obtained via the maximum likelihood (ML) estimator[8] for each class. The covariance matrix associated with the "true model" of class i is obtained by estimating the covariance matrix on all available data samples of class i, and the covariance matrices associated with the "mismatched model" of class i are obtained by estimating the covariance matrix on n_i data samples of class i.

[7]Denoted as "1RT2RCR" in the dataset.

[8]Note that this is equivalent to computing the empirical covariance matrix.

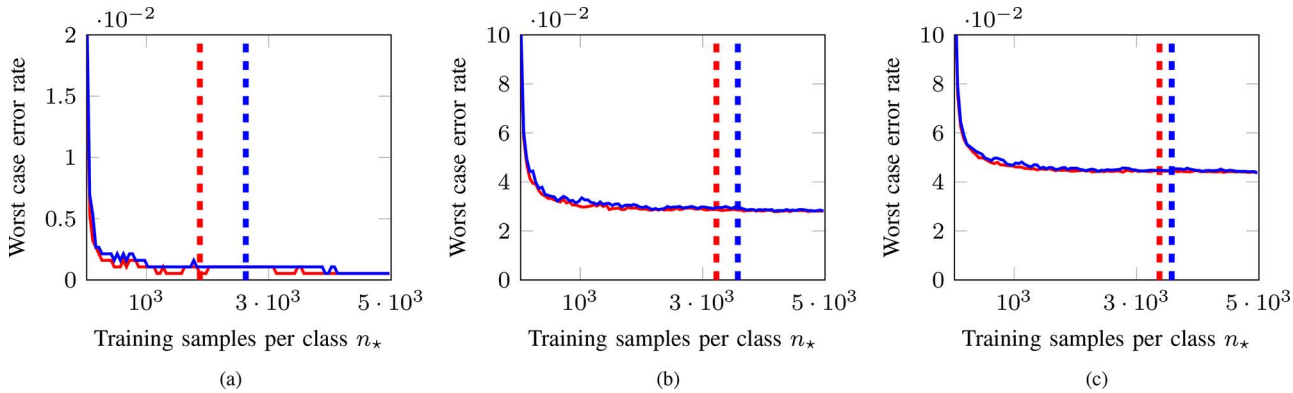

Fig. 7. The worst case error rate and phase transition predicted via Theorem 2 for a given probability p_p are plotted for classification of MNIST digits. Solid red and blue lines correspond to worst case error rates for $p_p = 0.9$ and $p_p = 1$, respectively. Dashed vertical lines denote the phase transition predicted via Theorem 2 for $p_p = 0.9$ (red) and $p_p = 1$ (blue). (a) Classification of 2 digits. (b) Classification of 5 digits. (c) Classification of 10 digits.

Results are produced as follows: in each run n_i samples are drawn at random from the training set for various values of $n_i, i = 1, \ldots, C$, and the signal covariances are estimated. The error rate of the MMAP classifier is then evaluated on the testing set. At the same time, we also determine if sufficient conditions for perfect classification as in Theorem 2 hold. We run 1000 experimental runs with the Hopkins dataset, where in each run dataset is split at random into training and testing sets. We run 20 experimental runs with the MNIST dataset, where in each run the draw of the n_i samples from the training set is random for $i = 1, \ldots, C$.

The particular choice of samples in the training set can lead to high variability in the mismatched models, especially for small number of training samples. Therefore, in the following, we have chosen to report the results as follows:

- we state that analysis predicts reliable classification if the sufficient conditions in Theorem 2 hold with probability p_p over the different experiment runs;
- we also state that simulation predicts reliable classification if the true error probability is 0 with probability p_p over different experiment runs;
- if the simulated error rate exhibits an error floor we report the worst case error rate with probability p_p: the error rate that is achieved at least with probability p_p over all experimental runs.

C. Results

The results for the Hopkins dataset are reported in Fig. 6.

We observe that the phase transition predicted by analysis approximates reasonably well the phase transition obeyed by simulation. In particular, we can use our theory to gauge the number of training samples required for perfect classification in the low-noise regime. As expected, we also observe that the larger value of p_p gives more conservative estimates of the required training samples. This holds for both simulation and analysis.

We also observe that identical trends hold for other values of n_3. In particular, for $n_3 < 30$ simulation does not show a phase transition and likewise analysis does not show a phase transition either (these experiments are not reported in view of space limitations). In contrast, for $n_3 \geq 30$ both simulation and analysis predict a phase transition in the error probability.

The results for the MNIST dataset are reported in Fig. 7. Note that the number of training samples per class is the same for all classes, i.e., $n_i = n_\star, i = 1, \ldots, C$.

In contrast to the results with the Hopkins dataset, the error rate obtained on the MNIST dataset exhibits an error floor. However, we observe that the worst case error rate reduces with a higher number of training samples and reaches an error floor at sufficiently large number of training samples. We also observe that the phase transition obtained via Theorem 2 predicts reasonably well the number of training samples needed to reach the error floor.

Finally, note that real data are not drawn from Gaussian distributions or perfect linear subspaces (the two main ingredients underlying our analysis). Nevertheless, we have shown that the derived bound has practical value even when the two assumptions do not hold strictly.

VI. Conclusion

This paper studies the classification of linear subspaces with mismatched classifiers, i.e., classifiers that operate on a mismatched version of the signal parameters *in lieu* of the true signal parameters. In particular, we have developed a low-noise expansion of an upper bound to the error probability of such a mismatched classifier that equips one with a set of sufficient conditions—which are a function of the geometry of the true signal distributions, the geometry of the mismatched signal distributions, and their interplay—in order to understand whether it is possible to classify reliably in the presence of mismatch in the low-noise regime.

Such sufficient conditions are shown to be sharp in the sense that they can predict the presence (and the absence) of a classification error floor both in experiments involving synthetic data as well as experiments involving real data. These conditions have also been shown to gauge well the number of training samples required for reliable classification in a motion segmentation application using the Hopkins 155 dataset and a hand-written digit classification application using the MNIST dataset.

Overall, we argue that our conditions can also be used as a proxy to develop linear feature extraction methods that are robust to mismatch. In particular, our study suggests that such methods ought to orthogonalize the different classes as much as possible in order to tolerate model mismatch. This intuition

has been pursued in recent state-of-the-art linear feature extraction methods.

APPENDIX

A. Preliminaries

We introduce additional quantities and Lemmas that are useful for the proofs.

a) Quantities: We define the projection operators:

$$\mathbf{P}_i = \mathbf{U}_i\mathbf{U}_i^T\,,\ \tilde{\mathbf{P}}_i = \tilde{\mathbf{U}}_i\tilde{\mathbf{U}}_i^T \tag{35}$$

$$\tilde{\mathbf{P}}'_{ij} = \tilde{\mathbf{U}}'_{ij}\left(\tilde{\mathbf{U}}'_{ij}\right)^T\,, \tag{36}$$

where $\mathbf{U}_i, \tilde{\mathbf{U}}_i$ and $\tilde{\mathbf{U}}'_{ij}$ are given as in Section II.A. In addition to the bases $\mathbf{U}_i$ and $\tilde{\mathbf{U}}_i$ for the $\mathrm{im}(\mathbf{\Sigma}_i)$ and $\mathrm{im}(\tilde{\mathbf{\Sigma}}_i)$, respectively, we also introduce the bases for the $\ker(\mathbf{\Sigma}_i)$ and $\ker(\tilde{\mathbf{\Sigma}}_i)$ as $\mathbf{U}_i^\perp \in \mathbb{R}^{N\times N-r_i}$ and $\tilde{\mathbf{U}}_i^\perp \in \mathbb{R}^{N\times N-\tilde{r}_i}$, respectively. We define the projection operators onto this subspaces:

$$\mathbf{K}_i = \mathbf{U}_i^\perp\left(\mathbf{U}_i^\perp\right)^T\,,\ \tilde{\mathbf{K}}_i = \tilde{\mathbf{U}}_i^\perp\left(\tilde{\mathbf{U}}_i^\perp\right)^T\,. \tag{37}$$

We also define

$$\mathbf{L}_i = \mathbf{U}_i\left(\mathrm{diag}\left(\lambda_1^i,\ldots,\lambda_{r_i}^i\right)+\sigma^2\mathbf{I}\right)^{-1}(\mathbf{U}_i)^T \tag{38}$$

$$\tilde{\mathbf{L}}_i = \tilde{\mathbf{U}}_i\left(\mathrm{diag}(\tilde{\lambda}_1^i,\ldots,\tilde{\lambda}_{\tilde{r}_i}^i\right)+\sigma^2\mathbf{I})^{-1}(\tilde{\mathbf{U}}_i)^T\,, \tag{39}$$

and write

$$\mathbf{\Sigma}_{ij} = \mathbf{L}_{ij} + \frac{1}{\sigma^2}\mathbf{K}_{ij}\,, \tag{40}$$

where $\mathbf{L}_{ij} = \mathbf{L}_i + \alpha_{ij}\tilde{\mathbf{L}}_j - \alpha_{ij}\tilde{\mathbf{L}}_i$ and $\mathbf{K}_{ij} = \mathbf{K}_i + \alpha_{ij}\tilde{\mathbf{K}}_j - \alpha_{ij}\tilde{\mathbf{K}}_i$. Note that

$$\mathbf{K}_{ij} = \mathbf{K}_i + \alpha_{ij}\tilde{\mathbf{P}}_i - \alpha_{ij}\tilde{\mathbf{P}}_j \tag{41}$$

$$= \mathbf{K}_i + \alpha_{ij}\tilde{\mathbf{P}}'_{ij} - \alpha_{ij}\tilde{\mathbf{P}}'_{ji} \tag{42}$$

in view of the fact that $\mathbf{P}_i + \mathbf{K}_i = \mathbf{I}$ and $\tilde{\mathbf{P}}_i + \tilde{\mathbf{K}}_i = \mathbf{I}$ and $\tilde{\mathbf{P}}_i - \tilde{\mathbf{P}}_j = \tilde{\mathbf{P}}'_{ij} - \tilde{\mathbf{P}}'_{ji}$. The last equality simply follows from the definition of $\tilde{\mathbf{P}}'_{ij}$ and $\tilde{\mathbf{P}}'_{ji}$, and the definitions of $\tilde{\mathbf{U}}'_{ij}, \tilde{\mathbf{U}}'_{ji}$ and $\tilde{\mathbf{U}}^\cap_{ij}$ given in Section II.A:

$$\begin{aligned}\tilde{\mathbf{P}}_i - \tilde{\mathbf{P}}_j &= \tilde{\mathbf{U}}'_{ij}\left(\tilde{\mathbf{U}}'_{ij}\right)^T + \tilde{\mathbf{U}}^\cap_{ij}\left(\tilde{\mathbf{U}}^\cap_{ij}\right)^T \\ &\quad - \left(\tilde{\mathbf{U}}'_{ji}\left(\tilde{\mathbf{U}}'_{ji}\right)^T + \tilde{\mathbf{U}}^\cap_{ij}\left(\tilde{\mathbf{U}}^\cap_{ij}\right)^T\right) \qquad (43)\\ &= \tilde{\mathbf{P}}'_{ij} - \tilde{\mathbf{P}}'_{ji}\,. \qquad (44)\end{aligned}$$

Finally, we present a decomposition of $\mathbf{x}\in\mathbb{R}^N$. We write

$$\mathbf{x} = \mathbf{x}_\parallel + \mathbf{x}_\perp = \mathbf{x}_\mathbf{V} + \mathbf{x}_\mathbf{W} + \mathbf{x}_\perp, \tag{45}$$

where

$$\mathbf{x}_\parallel = \mathbf{U}_i\mathbf{z}_\parallel \tag{46}$$

$$\mathbf{x}_\perp = \mathbf{U}_i^\perp\mathbf{z}_\perp \tag{47}$$

$$\mathbf{x}_\mathbf{V} = \mathbf{V}_{ij}\mathbf{z}_\mathbf{V} \tag{48}$$

$$\mathbf{x}_\mathbf{W} = \mathbf{W}_{ij}\mathbf{z}_\mathbf{W}, \tag{49}$$

for some vectors $\mathbf{z}_\parallel \in \mathbb{R}^{r_i}, \mathbf{z}_\perp \in \mathbb{R}^{N-r_i}, \mathbf{z}_\mathbf{V} \in \mathbb{R}^{s_{ij}^V}$ and $\mathbf{z}_\mathbf{W} \in \mathbb{R}^{s_{ij}^W}$. Note also that $\|\mathbf{x}_\mathbf{V}\| = \|\mathbf{z}_\mathbf{V}\|, \|\mathbf{x}_\mathbf{W}\| = \|\mathbf{z}_\mathbf{W}\|$ and $\|\mathbf{x}_\perp\| = \|\mathbf{z}_\perp\|$.

b) Lemmas:

Lemma 1: The following equality holds:

$$\mathrm{im}\left(\tilde{\mathbf{U}}'_{ij}\right)^\perp = \ker\left(\tilde{\mathbf{P}}'_{ij}\right) = \ker(\tilde{\mathbf{U}}_i) + (\mathrm{im}(\tilde{\mathbf{U}}_i)\cap\mathrm{im}(\tilde{\mathbf{U}}_j))\,.$$

Proof: By leveraging the definition of $\tilde{\mathbf{P}}'_{ij}$ in (36) we have

$$\begin{aligned}\ker\left(\tilde{\mathbf{P}}'_{ij}\right) &= \left(\mathrm{im}\left(\tilde{\mathbf{P}}'_{ij}\right)\right)^\perp \\ &= \left(\mathrm{im}\left(\tilde{\mathbf{U}}'_{ij}\right)\right)^\perp = \mathrm{im}\left(\left[\tilde{\mathbf{U}}^\cap_{ij}, \tilde{\mathbf{U}}_i^\perp\right]\right) \\ &= \mathrm{im}(\tilde{\mathbf{U}}_i)^\perp + \left(\mathrm{im}(\tilde{\mathbf{U}}_i)\cap\mathrm{im}(\tilde{\mathbf{U}}_j)\right).\end{aligned}$$

■

Lemma 2: The following statement holds:

$$d_{\min}\left(\mathbf{V}_{ij}, \tilde{\mathbf{U}}'_{ij}\right) < d_{\max}\left(\mathbf{V}_{ij}, \tilde{\mathbf{U}}'_{ji}\right) \tag{50}$$

$$\Longrightarrow$$

$$\mathbf{V}_{ij}^T\left(\tilde{\mathbf{U}}'_{ij}\left(\tilde{\mathbf{U}}'_{ij}\right)^T - \tilde{\mathbf{U}}'_{ji}\left(\tilde{\mathbf{U}}'_{ji}\right)^T\right)\mathbf{V}_{ij} \succ \mathbf{0}\,. \tag{51}$$

Proof: First, note that

$$\begin{aligned}&\mathbf{V}_{ij}^T\left(\tilde{\mathbf{U}}'_{ij}\left(\tilde{\mathbf{U}}'_{ij}\right)^T - \tilde{\mathbf{U}}'_{ji}\left(\tilde{\mathbf{U}}'_{ji}\right)^T\right)\mathbf{V}_{ij} \\ &\qquad = (\mathbf{V}_{ij})^T\left(\tilde{\mathbf{P}}'_{ij} - \tilde{\mathbf{P}}'_{ji}\right)\mathbf{V}_{ij}\,.\end{aligned}$$

Then we write the following

$$(\mathbf{V}_{ij})^T\tilde{\mathbf{P}}'_{ij}\mathbf{V}_{ij} = (\mathbf{V}_{ij})^T\tilde{\mathbf{U}}'_{ij}\left(\tilde{\mathbf{U}}'_{ij}\right)^T\mathbf{V}_{ij} \tag{52}$$

$$(\mathbf{V}_{ij})^T\tilde{\mathbf{P}}'_{ji}\mathbf{V}_{ij} = (\mathbf{V}_{ij})^T\tilde{\mathbf{U}}'_{ji}\left(\tilde{\mathbf{U}}'_{ji}\right)^T\mathbf{V}_{ij}. \tag{53}$$

Note that the singular values of $(\mathbf{V}_{ij})^T\tilde{\mathbf{U}}'_{ij}$ and $(\mathbf{V}_{ij})^T\tilde{\mathbf{U}}'_{ji}$ correspond to the cosines of the principal angles between and $\mathrm{im}(\mathbf{V}_{ij})$ and $\mathrm{im}\left(\tilde{\mathbf{U}}'_{ij}\right)$, and $\mathrm{im}(\mathbf{V}_{ij})$ and $\mathrm{im}\left(\tilde{\mathbf{U}}'_{ji}\right)$, respectively. We then consider the SVDs

$$(\mathbf{V}_{ij})^T\tilde{\mathbf{U}}'_{ij} = \mathbf{H}_{ij}\mathbf{D}_{ij}\mathbf{J}_{ij}^T \tag{54}$$

$$(\mathbf{V}_{ij})^T\tilde{\mathbf{U}}'_{ji} = \mathbf{H}_{ji}\mathbf{D}_{ji}\mathbf{J}_{ji}^T \tag{55}$$

where the dimensions of matrices $\mathbf{H}_{ij}, \mathbf{H}_{ji}, \mathbf{D}_{ij}, \mathbf{D}_{ji}, \mathbf{J}_{ij}$ and $\mathbf{J}_{ji}$ follow from the dimension of the $\mathbf{V}_{ij}, \tilde{\mathbf{U}}'_{ij}$ and $\tilde{\mathbf{U}}'_{ji}$ as shown in (9). We can now express (51) as

$$\mathbf{H}_{ij}\mathbf{D}_{ij}\mathbf{D}_{ij}^T\mathbf{H}_{ij}^T \succ \mathbf{H}_{ji}\mathbf{D}_{ji}\mathbf{D}_{ji}^T\mathbf{H}_{ji}^T. \tag{56}$$

It is straightforward to see that (50) implies (51). ■

Lemma 3: The following equalities and inequalities hold:

$$\mathbf{x}^T\mathbf{L}_i\mathbf{x} \geq \frac{1}{\lambda_1^i+1}\|\mathbf{x}_\parallel\|^2 \tag{57}$$

$$\mathbf{x}^T(\tilde{\mathbf{L}}_j - \tilde{\mathbf{L}}_i)\mathbf{x} \geq -\frac{1}{\tilde{\lambda}_{\tilde{r}_i}^i}\|\mathbf{x}\|^2 \tag{58}$$

$$\mathbf{x}^T\mathbf{K}_i\mathbf{x} = \|\mathbf{x}_\perp\|^2\,. \tag{59}$$

Proof: The inequality in (57) is due to the fact that $\mathbf{x}_\parallel \in \mathrm{im}(\mathbf{L}_i) = \mathrm{im}(\mathbf{U}_i)$ and $\frac{1}{\lambda_1^i+1}$ is a lower bound to the minimum positive eigenvalue of $\mathbf{L}_i$. The inequality in (58) is due to the fact that $\tilde{\mathbf{L}}_j$ is positive semidefinite and that $\frac{1}{\tilde{\lambda}_{\tilde{r}_i}^i}$ is an upper

bound for the largest eigenvalue of $\tilde{\mathbf{L}}_i$. The equality in (59) follows from the definition of the projector $\mathbf{K}_i$. ∎

Lemma 4: Assume that

$$\mathrm{im}(\mathbf{W}_{ij}) \subseteq \mathrm{im}\left(\tilde{\mathbf{U}}'_{ji}\right)^{\perp} \text{ and} \tag{60}$$

$$\mathbf{V}_{ij}^T\left(\tilde{\mathbf{U}}'_{ij}\left(\tilde{\mathbf{U}}'_{ij}\right)^T - \tilde{\mathbf{U}}'_{ji}\left(\tilde{\mathbf{U}}'_{ji}\right)^T\right)\mathbf{V}_{ij} \succ \mathbf{0}\,. \tag{61}$$

Denote by c_0 the smallest eigenvalue of

$$\mathbf{V}_{ij}^T\left(\tilde{\mathbf{U}}'_{ij}\left(\tilde{\mathbf{U}}'_{ij}\right)^T - \tilde{\mathbf{U}}'_{ji}\left(\tilde{\mathbf{U}}'_{ji}\right)^T\right)\mathbf{V}_{ij} = (\mathbf{V}_{ij})^T\left(\tilde{\mathbf{P}}'_i - \tilde{\mathbf{P}}'_j\right)\mathbf{V}_{ij}\,.$$

Then

$$\mathbf{x}^T(\tilde{\mathbf{K}}_j - \tilde{\mathbf{K}}_i)\mathbf{x} \geq c_0\|\mathbf{x}_{\mathbf{V}}\|^2 - 2\|\mathbf{x}_{\mathbf{V}}\|\|\mathbf{x}_\perp\| - \|\mathbf{x}_\perp\|^2\,. \tag{62}$$

Proof: Note that (49) implies $\mathbf{x}_{\mathbf{W}} \in \mathrm{im}(\mathbf{W}_{ij}) = \mathrm{im}(\mathbf{\Sigma}_i) \cap \ker\left(\tilde{\mathbf{P}}'_{ij}\right)$, and the condition (61) also implies $\mathbf{x}_{\mathbf{W}} \in \ker\left(\tilde{\mathbf{P}}'_{ji}\right) = \mathrm{im}\left(\tilde{\mathbf{U}}'_{ji}\right)^{\perp}$. Then, we can write

$$\mathbf{x}^T(\tilde{\mathbf{K}}_j - \tilde{\mathbf{K}}_i)\mathbf{x} = \mathbf{x}^T\left(\tilde{\mathbf{P}}'_{ij} - \tilde{\mathbf{P}}'_{ji}\right)\mathbf{x} \tag{63}$$

$$= \mathbf{x}_{\mathbf{V}}^T\left(\tilde{\mathbf{P}}'_{ij} - \tilde{\mathbf{P}}'_{ji}\right)\mathbf{x}_{\mathbf{V}} + 2\mathbf{x}_{\mathbf{V}}^T\left(\tilde{\mathbf{P}}'_{ij} - \tilde{\mathbf{P}}'_{ji}\right)\mathbf{x}_\perp + \mathbf{x}_\perp^T\left(\tilde{\mathbf{P}}'_{ij} - \tilde{\mathbf{P}}'_{ji}\right)\mathbf{x}_\perp \tag{64}$$

and we note that condition (61) implies the lower bound $\mathbf{x}_{\mathbf{V}}^T\left(\tilde{\mathbf{P}}'_{ij} - \tilde{\mathbf{P}}'_{ji}\right)\mathbf{x}_{\mathbf{V}} \geq c_0\|\mathbf{x}_{\mathbf{V}}\|^2$. Moreover, all the eigenvalues of $\tilde{\mathbf{P}}'_{ij} - \tilde{\mathbf{P}}'_{ji}$ are contained in the interval $[-1, 1]$ [31, Theorem 26], so that $\mathbf{x}_\perp^T\left(\tilde{\mathbf{P}}'_{ij} - \tilde{\mathbf{P}}'_{ji}\right)\mathbf{x}_\perp \geq -\|\mathbf{x}_\perp\|^2$, and, on leveraging Cauchy-Schwarz inequality, we also have $\mathbf{x}_{\mathbf{V}}^T\left(\tilde{\mathbf{P}}'_{ij} - \tilde{\mathbf{P}}'_{ji}\right)\mathbf{x}_\perp \geq -2\|\mathbf{x}_{\mathbf{V}}\|\|\mathbf{x}_\perp\|$. ∎

B. Proof of Theorem 1

We prove Theorem 1 by using the fact that $u(x) \leq \exp(\alpha x), \forall x, \alpha > 0$ and by leveraging the union bound.

Recall from (7) that the error probability associated with the MMAP classifier can be expressed as

$$P(e) = \sum_{i=1}^{C} p_i \cdot P(e|c=i) \tag{65}$$

where $P(e|c=i) = P(\hat{c} \neq i|c=i)$ is the error probability for signals in class i. Via the union bound, we can state that

$$P(e|c=i) = P(\hat{c} \neq i|c=i) \leq \sum_{j=1, j\neq i}^{C} P(\hat{c}=j|c=i) \tag{66}$$

where

$$P(\hat{c}=j|c=i) = \int_{-\infty}^{\infty} p(\mathbf{y}|c=i) \tag{67}$$

$$\cdot\, u\left(\log\left(\frac{\tilde{p}_j\tilde{p}(\mathbf{y}|c=j)}{\tilde{p}_i\tilde{p}(\mathbf{y}|c=i)}\right)\right)\mathrm{d}\mathbf{y}\,. \tag{68}$$

We will denote $P(\hat{c}=j|c=i) = P(e_{ij})$. Now, by letting $\alpha_{ij} > 0\,\forall i \neq j$ we can upper bound the step function to obtain

$$P(e_{ij}) \leq \int_{-\infty}^{\infty} p(\mathbf{y}|c=1) \cdot \exp\left(\alpha_{ij}\log\left(\frac{\tilde{p}_j\tilde{p}(\mathbf{y}|c=j)}{\tilde{p}_i\tilde{p}(\mathbf{y}|c=i)}\right)\right)\mathrm{d}\mathbf{y}$$

$$= \left(\frac{\tilde{p}_j}{\tilde{p}_i}\right)^{\alpha_{ij}}\left(\frac{\left|\tilde{\mathbf{\Sigma}}_i + \sigma^2\mathbf{I}\right|}{\left|\tilde{\mathbf{\Sigma}}_j + \sigma^2\mathbf{I}\right|}\right)^{\frac{\alpha_{ij}}{2}} \cdot \left((2\pi)^N\left|\mathbf{\Sigma}_i + \sigma^2\mathbf{I}\right|\right)^{-\frac{1}{2}} \cdot \int_{-\infty}^{\infty}\exp\left(-\frac{1}{2}\mathbf{y}^T\mathbf{\Sigma}_{ij}\mathbf{y}\right)\mathrm{d}\mathbf{y} = \bar{P}(e_{ij}), \tag{69}$$

where we recall

$$\mathbf{\Sigma}_{ij} = (\mathbf{\Sigma}_i + \sigma^2\mathbf{I})^{-1} + \alpha_{ij}(\tilde{\mathbf{\Sigma}}_j + \sigma^2\mathbf{I})^{-1} - \alpha_{ij}(\tilde{\mathbf{\Sigma}}_i + \sigma^2\mathbf{I})^{-1}\,.$$

If $\mathbf{\Sigma}_{ij} \succ \mathbf{0}\ \forall i \neq j$, then the integral in (69) converges $\forall i \neq j$. Therefore, we can bound the error probability as follows:

$$P(e) \leq \bar{P}(e) = \sum_{i=1}^{C} p_i \cdot \left(\sum_{j=1, j\neq i}^{C}\bar{P}(e_{ij})\right) \tag{70}$$

where

$$\bar{P}(e_{ij}) = \left(\frac{\tilde{p}_j}{\tilde{p}_i}\sqrt{\frac{\left|\tilde{\mathbf{\Sigma}}_i + \sigma^2\mathbf{I}\right|}{\left|\tilde{\mathbf{\Sigma}}_j + \sigma^2\mathbf{I}\right|}}\right)^{\alpha_{ij}} \cdot \left(\left|\mathbf{\Sigma}_i + \sigma^2\mathbf{I}\right|\left|\mathbf{\Sigma}_{ij}\right|\right)^{-\frac{1}{2}}\,. \tag{71}$$

If $\exists i \neq j : \mathbf{\Sigma}_{ij} \not\succ \mathbf{0}$ then the integral in (69) does not converge. Therefore, we trivially bound the error probability as $P(e) \leq \bar{P}(e) \leq 1$.

C. Proof of Theorem 2

The proof is presented in two parts. First, we establish sufficient conditions for $\mathbf{\Sigma}_{ij} \succ \mathbf{0}$; second, we establish conditions for the upper bound to the probability of misclassification to approach zero as the noise approaches zero.

1) Positive Definiteness of $\mathbf{\Sigma}_{ij}$: The following two Lemmas gives sufficient conditions for $\mathbf{\Sigma}_{ij} \succ \mathbf{0}$.

Lemma 5: Assume that $s_{ij}^V > 0$,

$$\mathrm{im}(\mathbf{W}_{ij}) \subseteq \mathrm{im}\left(\tilde{\mathbf{U}}'_{ji}\right)^{\perp}, \tag{72}$$

$$\mathbf{V}_{ij}^T\left(\tilde{\mathbf{U}}'_{ij}\left(\tilde{\mathbf{U}}'_{ij}\right)^T - \tilde{\mathbf{U}}'_{ji}\left(\tilde{\mathbf{U}}'_{ji}\right)^T\right)\mathbf{V}_{ij} \succ \mathbf{0}\,, \tag{73}$$

$$\alpha_{ij} < \min\left(\frac{\tilde{\lambda}^i_{\tilde{r}_i}}{\lambda^i_1 + 1}, \frac{c_0}{1 + c_0\left(1 + \frac{1}{\lambda^i_{\tilde{r}_i}}\right)}, 1\right), \tag{74}$$

where c_0 is the smallest eigenvalue of

$$\mathbf{V}_{ij}^T\left(\tilde{\mathbf{U}}'_{ij}\left(\tilde{\mathbf{U}}'_{ij}\right)^T - \tilde{\mathbf{U}}'_{ji}\left(\tilde{\mathbf{U}}'_{ji}\right)^T\right)\mathbf{V}_{ij} = (\mathbf{V}_{ij})^T\left(\tilde{\mathbf{P}}'_i - \tilde{\mathbf{P}}'_j\right)\mathbf{V}_{ij}\,.$$

Then

$$\mathbf{\Sigma}_{ij} \succ \mathbf{0},\, \forall \sigma^2 \in \left(0, \min\left(1, \frac{1-\alpha_{ij}}{\alpha_{ij}} \tilde{\lambda}_i^{\tilde{r}_i}\right)\right). \tag{75}$$

Proof: To show this we first produce a lower bound:

$$\begin{aligned}\mathbf{x}^T \mathbf{\Sigma}_{ij} \mathbf{x} &= \mathbf{x}^T \mathbf{L}_{ij} \mathbf{x} + \frac{1}{\sigma^2} \mathbf{x}^T \mathbf{K}_{ij} \mathbf{x} && (76)\\ &= \mathbf{x}^T \mathbf{L}_i \mathbf{x} + \alpha_{ij} \mathbf{x}^T (\tilde{\mathbf{L}}_j - \tilde{\mathbf{L}}_i) \mathbf{x} \\ &\quad + \frac{1}{\sigma^2}\left(\mathbf{x}^T \mathbf{K}_i \mathbf{x} + \alpha_{ij} \mathbf{x}^T (\tilde{\mathbf{K}}_j - \tilde{\mathbf{K}}_i)\mathbf{x}\right) && (77)\\ &\geq \mathbf{c}^T \mathbf{C} \mathbf{c}, && (78)\end{aligned}$$

where $\mathbf{c} = [\|\mathbf{x}_{\mathbf{W}}\|, \|\mathbf{x}_{\mathbf{V}}\|, \|\mathbf{x}_{\perp}\|]^T$ and

$$\mathbf{C} = \begin{bmatrix} \frac{1}{\lambda_1^i+1} - \frac{\alpha_{ij}}{\lambda_{\tilde{r}_i}^i} & 0 & 0 \\ 0 & \frac{1}{\lambda_1^i+1} - \frac{\alpha_{ij}}{\lambda_{\tilde{r}_i}^i} + \frac{c_0\alpha_{ij}}{\sigma^2} & -\frac{\alpha_{ij}}{\sigma^2} \\ 0 & -\frac{\alpha_{ij}}{\sigma^2} & \frac{1-\alpha_{ij}}{\sigma^2} - \frac{\alpha_{ij}}{\lambda_{\tilde{r}_i}^i} \end{bmatrix}, \tag{79}$$

by using the equalities and inequalities (57)–(59) and (62).

Now, by using standard algebraic manipulations, it is possible to show that the choice (74) leads to $\mathbf{C} \succ \mathbf{0}$, hence (75) holds. ■

Lemma 6: Assume that $s_{ij}^V = 0$,

$$\mathrm{im}(\mathbf{W}_{ij}) \subseteq \mathrm{im}\left(\tilde{\mathbf{U}}'_{ji}\right)^{\perp}, \tag{80}$$

$$\alpha_{ij} < \min\left(\frac{\tilde{\lambda}_{\tilde{r}_i}^i}{\lambda_1^i + 1}, \frac{\tilde{\lambda}_{\tilde{r}_i}^i}{\tilde{\lambda}_{\tilde{r}_i}^i + 1}, 1\right). \tag{81}$$

Then

$$\mathbf{\Sigma}_{ij} \succ \mathbf{0},\, \forall \sigma^2 \in \left(0, \min\left(1, \frac{1-\alpha_{ij}}{\alpha_{ij}} \tilde{\lambda}_i^{\tilde{r}_i}\right)\right). \tag{82}$$

Proof: We prove the Lemma by constructing the lower bound

$$\begin{aligned}\mathbf{x}^T \mathbf{\Sigma}_{ij} \mathbf{x} &\geq \frac{1}{\lambda_1^i + 1}\|\mathbf{x}_{\|}\|^2 - \frac{\alpha_{ij}}{\tilde{\lambda}_{\tilde{r}_i}^i}\|\mathbf{x}\|^2 \\ &\quad + \frac{1}{\sigma^2}\left(\|\mathbf{x}_{\perp}\|^2 + \alpha_{ij}\mathbf{x}^T\left(\tilde{\mathbf{P}}'_{ij} - \tilde{\mathbf{P}}'_{ji}\right)\mathbf{x}\right) && (83)\\ &\geq \left(\frac{1}{\lambda_1^i+1} - \frac{\alpha_{ij}}{\tilde{\lambda}_{\tilde{r}_i}^i}\right)\|\mathbf{x}_{\|}\|^2 + \left(\frac{1-\alpha_{ij}}{\sigma^2} - \frac{\alpha_{ij}}{\tilde{\lambda}_{\tilde{r}_i}^i}\right)\|\mathbf{x}_{\perp}\|^2, && (84)\end{aligned}$$

by using the inequalities equalities and inequalities (57)–(59) and (62), and by noting that $\mathbf{x}_{\mathbf{V}} = \mathbf{0}$. The choice (81) then leads to (82). ■

2) Part 2: Low-Noise Expansion: To obtain the low-noise expansion of the upper bound to the error probability we first present two supporting Lemmas.

Lemma 7: Assume that condition (72) given in Lemma 5 holds. Assume also that $s_{ij}^V > 0$ and (73) and (74) given in Lemma 5 hold, or that $s_{ij}^V = 0$ and (81) given in Lemma 6 holds. Then $\mathbf{K}_{ij} \succeq \mathbf{0}$ and $\mathrm{rank}(\mathbf{K}_{ij}) = N + s_{ij}^V - r_i$.

Proof: Assume that (72), $s_{ij}^V > 0$ (73) and (74) are satisfied. By definition, $\mathrm{im}(\mathbf{W}_{ij}) = \mathrm{im}(\mathbf{\Sigma}_i) \cap \ker\left(\tilde{\mathbf{P}}'_{ij}\right)$ and, as a consequence of (72), it also holds $\mathrm{im}(\mathbf{W}_{ij}) \subseteq \ker\left(\tilde{\mathbf{P}}'_{ji}\right)$, which leads to $\mathrm{im}(\mathbf{W}_{ij}) \subseteq \ker(\mathbf{K}_{ij})$. Moreover, it is straightforward to note that $\mathrm{im}\left(\left[\mathbf{V}_{ij}, \mathbf{U}_i^{\perp}\right]\right) = (\mathrm{im}(\mathbf{W}_{ij}))^{\perp}$. Then, in order to prove that $\mathbf{K}_{ij} \succeq \mathbf{0}$, we show that $\mathbf{x}^T\mathbf{K}_{ij}\mathbf{x} = \mathbf{x}^T\left(\mathbf{K}_i + \alpha_{ij}\left(\tilde{\mathbf{P}}'_{ij} - \tilde{\mathbf{P}}'_{ji}\right)\right)\mathbf{x} > 0, \forall \mathbf{x} \in \mathrm{im}\left(\left[\mathbf{V}_{ij}, \mathbf{U}_i^{\perp}\right]\right)$. Namely, by leveraging the equality in (59) and inequality in (62), we can write

$$\begin{aligned}\mathbf{x}^T\left(\mathbf{K}_i + \alpha_{ij}\left(\tilde{\mathbf{P}}'_{ij} - \tilde{\mathbf{P}}'_{ji}\right)\right)\mathbf{x} &\geq (1-\alpha_{ij})\|\mathbf{x}_{\perp}\|^2 \\ &\quad -2\alpha_{ij}\|\mathbf{x}_{\perp}\|\|\mathbf{x}_{\mathbf{V}}\| + \alpha_{ij}c_0\|\mathbf{x}_{\mathbf{V}}\|^2, \end{aligned} \tag{85}$$

where $\mathbf{x}_{\perp}, \mathbf{x}_{\mathbf{V}}$ have been defined in (47) and (48). If $\alpha_{ij} < \frac{c_0}{c_0+1}$ then the right hand side of (85) is always strictly positive, unless $\mathbf{x} = \mathbf{0}$. Then, since the condition in (74) implies $\alpha_{ij} < \frac{c_0}{c_0+1}$, we can conclude that $\mathbf{K}_{ij} \succeq \mathbf{0}$ and $\mathrm{im}(\mathbf{W}_{ij}) = \ker(\mathbf{K}_{ij})$ and $\mathrm{im}\left(\left[\mathbf{V}_{ij}, \mathbf{U}_i^{\perp}\right]\right) = \mathrm{im}(\mathbf{K}_{ij})$. Therefore, $\mathrm{rank}(\mathbf{K}_{ij}) = \mathrm{rank}\left(\left[\mathbf{V}_{ij}, \mathbf{U}_i^{\perp}\right]\right) = s_{ij}^V + (N - r_i)$.

Assume now that (72), $s_{ij}^V = 0$ and (81) are satisfied. In this case $\mathbf{x}^T\mathbf{K}_{ij}\mathbf{x} = \mathbf{x}^T\left(\mathbf{K}_i + \alpha_{ij}\left(\tilde{\mathbf{P}}'_{ij} - \tilde{\mathbf{P}}'_{ji}\right)\right)\mathbf{x} = \mathbf{x}_{\perp}^T\left(\mathbf{K}_i + \alpha_{ij}\left(\tilde{\mathbf{P}}'_{ij} - \tilde{\mathbf{P}}'_{ji}\right)\right)\mathbf{x}_{\perp} \geq \|\mathbf{x}_{\perp}\|^2(1-\alpha_{ij})$, where we have used the fact that eigenvalues of $\tilde{\mathbf{P}}'_{ji} - \tilde{\mathbf{P}}'_{ji}$ contained in the interval $[-1, 1]$. Since (81) implies $\alpha_{ij} < 1$ we conclude, via an argument similar to that in previous paragraph, that $\mathbf{K}_{ij} \succeq \mathbf{0}$ and $\mathrm{rank}(\mathbf{K}_{ij}) = \mathrm{rank}\left(\mathbf{U}_i^{\perp}\right) = s_{ij}^V + (N - r_i)$. ■

Lemma 8: Assume that condition (72) given in Lemma 5 holds. Assume also that $s_{ij}^V > 0$ and (73) and (74) given in Lemma 5 hold, or that $s_{ij}^V = 0$ and (81) given in Lemma 6 holds. Then, as $\sigma^2 \to 0$, we can write

$$\left|\mathbf{L}_{ij} + \frac{1}{\sigma^2}\mathbf{K}_{ij}\right| = v_{ij} \cdot \left(\frac{1}{\sigma^2}\right)^{r_{\mathbf{K}_{ij}}} + \mathcal{O}\left(\left(\frac{1}{\sigma^2}\right)^{r_{\mathbf{K}_{ij}}-1}\right), \tag{86}$$

where $r_{\mathbf{K}_{ij}} = \mathrm{rank}(\mathbf{K}_{ij})$, and v_{ij} is given as

$$v_{ij} = \begin{cases} \mathrm{pdet}(\mathbf{K}_{ij}) \left|\left(\mathbf{U}_{\mathbf{K}_{ij}}^{\perp}\right)^T \mathbf{L}_{ij}^0 \mathbf{U}_{\mathbf{K}_{ij}}^{\perp}\right| & \text{if } r_{\mathbf{K}_{ij}} < N \\ |\mathbf{K}_{ij}| & \text{if } r_{\mathbf{K}_{ij}} = N \end{cases}, \tag{87}$$

$\mathbf{L}_{ij}^0 = \lim_{\sigma^2 \to 0} \mathbf{L}_{ij} = \mathbf{L}_i^0 + \alpha_{ij}\tilde{\mathbf{L}}_j^0 - \alpha_{ij}\tilde{\mathbf{L}}_i^0$ and

$$\mathbf{L}_i^0 = \mathbf{U}_i \left(\mathrm{diag}\left(\lambda_1^i, \ldots, \lambda_{r_i}^i\right)\right)^{-1} (\mathbf{U}_i)^T \tag{88}$$

$$\tilde{\mathbf{L}}_i^0 = \tilde{\mathbf{U}}_i \left(\mathrm{diag}\left(\tilde{\lambda}_1^i, \ldots, \tilde{\lambda}_{\tilde{r}_i}^i\right)\right)^{-1} (\tilde{\mathbf{U}}_i)^T. \tag{89}$$

Proof: Note first that the sufficient conditions imply $\mathbf{K}_{ij} \succeq \mathbf{0}$ via Lemma 7. We can write the eigenvalue decomposition of $\mathbf{K}_{ij}$:

$$\mathbf{K}_{ij} = \mathbf{U}_{\mathbf{K}_{ij}} \begin{bmatrix} \mathbf{\Lambda}_{\mathbf{K}_{ij}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{U}_{\mathbf{K}_{ij}}^T, \tag{90}$$

where $\mathbf{U}_{\mathbf{K}_{ij}} \in \mathbb{R}^{N \times N}$ is orthogonal and $\mathbf{\Lambda}_{\mathbf{K}_{ij}} = \mathrm{diag}\left(\lambda_1^{\mathbf{K}_{ij}}, \ldots, \lambda_{r_{\mathbf{K}_{ij}}}^{\mathbf{K}_{ij}}\right)$ contains the positive eigenvalues of $\mathbf{K}_{ij}$, with $r_{\mathbf{K}_{ij}} = \mathrm{rank}(\mathbf{K}_{ij})$.

Now, we can write

$$\left|\mathbf{L}_{ij} + \frac{1}{\sigma^2}\mathbf{K}_{ij}\right| = \left|\begin{bmatrix} \frac{1}{\sigma^2}\mathbf{\Lambda}_{\mathbf{K}_{ij}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \mathbf{E}\right|, \tag{91}$$

where $\mathbf{E} = \mathbf{U}_{\mathbf{K}_{ij}}^T \mathbf{L}_{ij} \mathbf{U}_{\mathbf{K}_{ij}}$. We also denote by $\mathbf{E}_{i_1 \dots i_m}$ the principal submatrix of order $N-m$ obtained by deleting the rows and the columns $i_1, \dots, i_m$ of the matrix $\mathbf{E}$. Note that $\mathbf{E}_{i_1 \dots i_m} = \mathbf{P}_{i_1 \dots i_m}^T \mathbf{E} \mathbf{P}_{i_1 \dots i_m}$, where the matrix $\mathbf{P}_{i_1 \dots i_m} \in \mathbb{R}^{N \times N-m}$ is obtained by picking all the columns from the identity matrix with the column indices different from $i_1, \dots, i_m$. Then, the Poincaré separation theorem [32, Corollary 4.3.37] guarantees that the eigenvalues $\mathbf{E}_{i_1 \dots i_m}$ are bounded by the minimum and the maximum eigenvalues of $\mathbf{E}$, which correspond to the minimum and maximum eigenvalues of $\mathbf{L}_{ij}$. Moreover, as $\sigma^2 \to 0$, while the diagonal elements of $\frac{1}{\sigma^2}\mathbf{\Lambda}_{\mathbf{K}_{ij}}$ grow unbounded, the eigenvalues of $\mathbf{L}_{ij}$, and therefore, also the determinant of $\mathbf{E}_{i_1 \dots i_m}$, are bounded.

Then, we can use the determinant decomposition in [33, Theorem 2.3] to express $\left|\mathbf{L}_{ij} + \frac{1}{\sigma^2}\mathbf{K}_{ij}\right|$ as follows. If $r_{\mathbf{K}_{ij}} = N$:

$$\left|\mathbf{L}_{ij} + \frac{1}{\sigma^2}\mathbf{K}_{ij}\right| = \left|\frac{1}{\sigma^2}\mathbf{K}_{ij}\right| + |\mathbf{L}_{ij}| + S_1 + \cdots + S_{N-1}\,, \quad (92)$$

where

$$S_m = \sum\nolimits_{1 \le i_1 < \cdots < i_m \le N} \left(\frac{1}{\sigma^2}\lambda_{i_1}^{\mathbf{K}_{ij}}\right) \cdots \left(\frac{1}{\sigma^2}\lambda_{i_m}^{\mathbf{K}_{ij}}\right) |\mathbf{E}_{i_1 \dots i_m}| \quad 1 \le m \le N-1 \quad (93)$$

and the summation is over all possible ordered subsets of m indices from the set $\{1, \dots, r_{\mathbf{K}_{ij}}\}$. Otherwise, if $r_{\mathbf{K}_{ij}} < N$:

$$\left|\mathbf{L}_{ij} + \frac{1}{\sigma^2}\mathbf{K}_{ij}\right| = |\mathbf{L}_{ij}| + S_1 + \ldots + S_{r_{\mathbf{K}_{ij}}}\,, \quad (94)$$

where

$$S_m = \sum\nolimits_{1 \le i_1 < \cdots < i_m \le r_{\mathbf{K}_{ij}}} \left(\frac{1}{\sigma^2}\lambda_{i_1}^{\mathbf{K}_{ij}}\right) \cdots \left(\frac{1}{\sigma^2}\lambda_{i_m}^{\mathbf{K}_{ij}}\right) |\mathbf{E}_{i_1 \dots i_m}| \quad 1 \le m \le r_{\mathbf{K}_{ij}}\,. \quad (95)$$

Now we show that (87) holds. We first assume $\mathrm{rank}(\mathbf{K}_{ij}) = N$ and take the right hand side of (92) and multiply it by $\left(\frac{1}{\sigma^2}\right)^{r_{\mathbf{K}_{ij}}} \left(\sigma^2\right)^{r_{\mathbf{K}_{ij}}}$ to get

$$\left(\frac{1}{\sigma^2}\right)^{r_{\mathbf{K}_{ij}}} \left(|\mathbf{K}_{ij}| + \left(\sigma^2\right)^{r_{\mathbf{K}_{ij}}} \left(|\mathbf{L}_{ij}| + S_1 + \cdots + S_{N-1}\right)\right)\,. \quad (96)$$

Note now that for all $S_m, m = 1, \dots, N-1$, $\lim_{\sigma^2 \to 0} (\sigma^2)^{r_{\mathbf{K}_{ij}}} S_m = 0$ and $\lim_{\sigma^2 \to 0} (\sigma^2)^{r_{\mathbf{K}_{ij}}} |\mathbf{L}_{ij}| = 0$. Therefore, (87) holds for the case $\mathrm{rank}(\mathbf{K}_{ij}) = N$. To show the derivation of v_{ij} for the case $\mathrm{rank}(\mathbf{K}_{ij}) < N$ we use the same technique where we multiply by $\left(\frac{1}{\sigma^2}\right)^{r_{\mathbf{K}_{ij}}} (\sigma^2)^{r_{\mathbf{K}_{ij}}}$ the right hand side of (94) to get

$$\left(\frac{1}{\sigma^2}\right)^{r_{\mathbf{K}_{ij}}} \left((\sigma^2)^{r_{\mathbf{K}_{ij}}} S_{r_{\mathbf{K}_{ij}}} + (\sigma^2)^{r_{\mathbf{K}_{ij}}} \left(|\mathbf{L}_{ij}| + S_1 + \cdots + S_{r_{\mathbf{K}_{ij}}-1}\right)\right)\,. \quad (97)$$

As $\sigma^2 \to 0$ we can write $(\sigma^2)^{r_{\mathbf{K}_{ij}}} S_{r_{\mathbf{K}_{ij}}} = \mathrm{pdet}(\mathbf{K}_{ij}) \left|\left(\mathbf{U}_{\mathbf{K}_{ij}}^{\perp}\right)^T \mathbf{L}_{ij}^0 \mathbf{U}_{\mathbf{K}_{ij}}^{\perp}\right|$. This concludes the derivation of (87). Note also that $v_{ij} > 0$, since the pseudo-determinant and the determinants in (87) are greater than zero. ∎

We now provide the low-noise expansion of the upper bound to the probability of misclassification.

Assume that sufficient conditions for positive definiteness of $\mathbf{\Sigma}_{ij}, \forall i \neq j$ do not hold. Then, the upper bound to the probability of error is chosen to be $\bar{P}(e) = 1$, so that in general it does not tend to zero as σ^2 tends to zero.

Assume now that the sufficient conditions for $\mathbf{\Sigma}_{ij} \succ \mathbf{0}$ as given in the first part of the proof hold $\forall i \neq j$. Then, the upper bound to the probability of misclassification can be written as follows:[9]

$$\bar{P}(e) = \sum_i \sum_{j \neq i} p_i \left(\frac{\tilde{p}_j}{\tilde{p}_i} \sqrt{\frac{\left|\tilde{\mathbf{\Sigma}}_i + \sigma^2\mathbf{I}\right|}{\left|\tilde{\mathbf{\Sigma}}_j + \sigma^2\mathbf{I}\right|}}\right)^{\alpha_{ij}} \cdot \left(\left|\mathbf{\Sigma}_i + \sigma^2\mathbf{I}\right| |\mathbf{\Sigma}_{ij}|\right)^{-\frac{1}{2}}\,. \quad (98)$$

We will now produce a low-noise expansion of (98) in order to understand whether or not $\lim_{\sigma^2 \to 0} \bar{P}(e) = 0$. The following low-noise expansions are trivial:

$$\left|\mathbf{\Sigma}_i + \sigma^2\mathbf{I}\right| = \left(\prod_{k=1}^{r_i} \left(\lambda_k^i + \sigma^2\right)\right) (\sigma^2)^{N-r_i} = \mathcal{O}\left((\sigma^2)^{N-r_i}\right), \quad \sigma^2 \to 0 \quad (99)$$

$$\left|\tilde{\mathbf{\Sigma}}_i + \sigma^2\mathbf{I}\right| = \left(\prod_{k=1}^{\tilde{r}_i} \left(\tilde{\lambda}_k^i + \sigma^2\right)\right) (\sigma^2)^{N-\tilde{r}_i} = \mathcal{O}\left((\sigma^2)^{N-\tilde{r}_i}\right), \quad \sigma^2 \to 0\,. \quad (100)$$

The low-noise expansion of $|\mathbf{\Sigma}_{ij}|$ is more involved and it is provided in Lemma 8.

Then, it follows immediately that the low-noise expansion of each term in the upper bound to the probability of error in (98) is given by

$$A_{ij}(\sigma^2)^{d_{ij}} + o\left((\sigma^2)^{d_{ij}}\right), \quad (101)$$

where

$$\begin{aligned} d_{ij} &= -\frac{\alpha_{ij}}{2}\left((N - \tilde{r}_i) - (N - \tilde{r}_j)\right) - \frac{1}{2}(N - r_i) - \frac{1}{2}(-\mathrm{rank}(\mathbf{K}_{ij})) \\ &= \frac{1}{2}(\alpha_{ij}(\tilde{r}_j - \tilde{r}_i) + s_{ij}^V)\,, \end{aligned} \quad (102)$$

$$A_{ij} = \left(\frac{\tilde{p}_j}{\tilde{p}_i}\right)^{\alpha_{ij}} \left(\frac{\tilde{v}_i}{\tilde{v}_j}\right)^{\frac{\alpha_{ij}}{2}} (v_i v_{ij})^{-\frac{1}{2}} > 0 \quad (103)$$

and

$$v_i = \mathrm{pdet}\left(\mathbf{\Sigma}_i\right)\,,\ \tilde{v}_i = \mathrm{pdet}\left(\tilde{\mathbf{\Sigma}}_i\right)\,. \quad (104)$$

It follows immediately that the low-noise expansion of the upper bound to the probability of error in (98) is given by

$$\bar{P}(e) = A(\sigma^2)^d + o\left((\sigma^2)^d\right)\,, \quad (105)$$

where $d = \min_{(i \neq j)} d_{ij}$ and $A = \sum_{(i,j) \in \mathcal{S}_d} A_{ij}$ where $\mathcal{S}_d = \{(i,j) : d_{ij} = d\}$.

D. Proof of Corollary 2

Assume $s_{ij}^V > 0\, \forall (i,j), i \neq j$ and

$$d_{\min}(\mathbf{U}_i, \tilde{\mathbf{U}}_i) < d_{\max}(\mathbf{U}_i, \tilde{\mathbf{U}}_j)\, \forall (i,j), i \neq j\,. \quad (106)$$

[9]Note that a value for which α_{ij} satisfies the conditions for $\mathbf{\Sigma}_{ij} \succ \mathbf{0}$ always exists and therefore does not affect the derivation of the low-noise expansion.

Note that $d_{\min}(\mathbf{U}_i, \tilde{\mathbf{U}}_i) < d_{\max}(\mathbf{U}_i, \tilde{\mathbf{U}}_j)$ implies

$$\mathbf{U}_i^T \left(\tilde{\mathbf{U}}_i \tilde{\mathbf{U}}_i^T - \tilde{\mathbf{U}}_j \tilde{\mathbf{U}}_j^T \right) \mathbf{U}_i \succ \mathbf{0} \tag{107}$$

$$\Longleftrightarrow$$

$$\mathbf{U}_i^T \left(\tilde{\mathbf{U}}'_{ij} \left(\tilde{\mathbf{U}}'_{ij} \right)^T - \tilde{\mathbf{U}}'_{ji} \left(\tilde{\mathbf{U}}'_{ji} \right)^T \right) \mathbf{U}_i \succ \mathbf{0}\,, \tag{108}$$

where we have used result in Lemma 2 in the Appendix A.

By taking $\mathbf{x} \in \mathbf{W}_{ij}$ or $\mathbf{x} \in \mathbf{V}_{ij}$ it is straightforward to show that (108) implies (15) and (16), thus obtaining conditions identical to those in Corollary 1.

E. Proof of Corollary 3

We prove the corollary by showing that in diagonal case (16) always holds. Note first that

$$\begin{aligned} \operatorname{im}(\mathbf{V}_{ij}) &= \operatorname{im}(\mathbf{U}_i) \cap \left(\operatorname{im}(\mathbf{U}_i) \cap \operatorname{im}\left(\tilde{\mathbf{U}}'_{ij} \right)^{\perp} \right)^{\perp} \\ &= \operatorname{im}(\mathbf{U}_i) \cap \operatorname{im}\left(\tilde{\mathbf{U}}'_{ij} \right) \subseteq \operatorname{im}\left(\tilde{\mathbf{U}}'_{ij} \right). \end{aligned}$$

It is also straightforward to establish that (16) holds if and only if $\operatorname{im}(\mathbf{V}_{ij}) \subseteq \operatorname{im}\left(\tilde{\mathbf{U}}'_{ji} \right)^{\perp}$, and this always holds since $\operatorname{im}(\mathbf{V}_{ij}) \subseteq \operatorname{im}(\tilde{\mathbf{U}}'_{ij}$ and $\operatorname{im}(\tilde{\mathbf{U}}'_{ij} \subseteq \operatorname{im}(\tilde{\mathbf{U}}'_{ji})^{\perp}$.

F. Derivation of Example 3

We prove statement (30), by showing that

$$1 - \delta_{12} > N(\epsilon_1 + \epsilon_2) \tag{109}$$

together with $s_{12}, s_{21} > 0$ implies the sufficient conditions for perfect classification in Corollary 2.

Assume $\mathbf{U}_i$ and $\mathbf{U}_j$ are given and the singular values of $(\mathbf{U}_i)^T \mathbf{U}_j$ are known. We also know that $\tilde{\mathbf{U}}_j = \mathbf{Q}_j \mathbf{U}_j$. We can write

$$(\mathbf{U}_i)^T \tilde{\mathbf{U}}_j = (\mathbf{U}_i)^T \mathbf{U}_j + (\mathbf{U}_i)^T (\mathbf{Q}_j - \mathbf{I}) \mathbf{U}_j \tag{110}$$

On leveraging [34, Theorem 1], we can state that the i-th singular value d_i associated with $(\mathbf{U}_i)^T \tilde{\mathbf{U}}_j$ lies in the interval $\left[d_i - \left\| (\mathbf{U}_i)^T (\mathbf{Q}_j - \mathbf{I}) \mathbf{U}_j \right\|_2, d_i + \left\| (\mathbf{U}_i)^T (\mathbf{Q}_j - \mathbf{I}) \mathbf{U}_j \right\|_2 \right]$, where d_i is the i-th singular value of $(\mathbf{U}_i)^T \mathbf{U}_j$. Then, we can write the upper bound

$$\left\| (\mathbf{U}_i)^T (\mathbf{Q}_j - \mathbf{I}) \mathbf{U}_j \right\|_2 \leq \|\mathbf{Q}_j - \mathbf{I}\|_2 = \epsilon_j \tag{111}$$

where the first inequality follows from the SVD separation theorem [35, Theorem 2.2]. Note also that

$$(\mathbf{U}_i)^T \tilde{\mathbf{U}}_i = \mathbf{I} + (\mathbf{U}_i)^T (\mathbf{Q}_i - \mathbf{I}) \mathbf{U}_i \tag{112}$$

where the singular values of $(\mathbf{U}_i)^T \tilde{\mathbf{U}}_i$ are bounded by $1 \pm \left\| (\mathbf{U}_i)^T (\mathbf{Q}_i - \mathbf{I}) \mathbf{U}_i \right\|_2$. By leveraging (111) we can further bound the singular values as $1 \pm \epsilon_i$.

Note now that $1 - \delta_{12} > (\epsilon_1 + \epsilon_2)$ if and only if $1 - \epsilon_1 > \delta_{12} + \epsilon_2$, which implies

$$d_{\min}(\mathbf{U}_1, \tilde{\mathbf{U}}_1) < d_{\max}(\mathbf{U}_1, \tilde{\mathbf{U}}_2)\,, \tag{113}$$

and is also equivalent to

$$\max_k \cos_k((\mathbf{U}_1)^T \tilde{\mathbf{U}}_2) < \min_l \cos_l \left((\mathbf{U}_1)^T \tilde{\mathbf{U}}_1 \right), \tag{114}$$

where $\max_k \cos_k((\mathbf{U}_1)^T \tilde{\mathbf{U}}_2)$ denotes the cosine of the smallest principal angle between $\operatorname{im}(\mathbf{U}_1)$ and $\operatorname{im}(\tilde{\mathbf{U}}_2)$, $\max_k \cos_k((\mathbf{U}_1)^T \tilde{\mathbf{U}}_2)$ denotes the cosine of the largest principal angle between $\operatorname{im}(\mathbf{U}_1)$ and $\operatorname{im}(\tilde{\mathbf{U}}_1)$. The equivalence between (113) and (114) follows straight from the definition of min and max correlation distances. It is now easy to verify that $1 - \epsilon_1 > \delta_{12} + \epsilon_2$ implies (114), since $1 - \epsilon_1$ is a lower bound for the cosine of the largest principal angles between $\mathbf{U}_1$ and $\tilde{\mathbf{U}}_1$, and $\delta_{12} + \epsilon_2$ is an upper bound to the cosine of the smallest principal angles between $\mathbf{U}_1$ and $\tilde{\mathbf{U}}_2$.

Finally, the same arguments can be used to show that $d_{\min}(\mathbf{U}_2, \tilde{\mathbf{U}}_2) < d_{\max}(\mathbf{U}_2, \tilde{\mathbf{U}}_1)$. This concludes the proof.

REFERENCES

[1] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, 2nd ed. New York, NY, USA: Wiley-Interscience, 2000.

[2] B. Frénay and M. Verleysen, "Classification in the presence of label noise: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 5, pp. 845–869, May 2014.

[3] N. Natarajan, I. S. Dhillon, P. K. Ravikumar, and A. Tewari, "Learning with noisy labels," *Adv. Neural Inf. Process. Syst.*, vol. 26, pp. 1196–1204, 2013.

[4] J. Bootkrajang and A. Kabán, "Label-noise robust logistic regression and its applications," in *Mach. Learn. Knowl. Discov. Databases*. New York, NY, USA: Springer, 2012, vol. 7523, pp. 143–158.

[5] J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, *Dataset Shift in Machine Learning*. Cambridge, U.K.: MIT Press, 2009.

[6] R. Basri and D. W. Jacobs, "Lambertian reflectance and linear subspaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 2, pp. 218–233, Feb. 2003.

[7] K.-C. Lee, J. Ho, and D. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 684–698, May 2005.

[8] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[9] Q. Zhang and B. Li, "Discriminative K-SVD for dictionary learning in face recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2691–2698, IEEE.

[10] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2765–2781, Nov. 2013.

[11] C. Tomasi and T. Kanade, "Shape and motion from image streams under orthography: A factorization method," *Int. J. Comput. Vis.*, vol. 9, no. 2, pp. 137–154, Nov. 1992.

[12] T. E. Boult and L. G. Brown, "Factorization-based segmentation of motions," in *Proc. IEEE Workshop Vis. Motion*, Oct. 1991, pp. 179–186.

[13] M. Chen, J. Silva, J. Paisley, C. Wang, D. Dunson, and L. Carin, "Compressive sensing on manifolds using a nonparametric mixture of factor analyzers: Algorithm and performance bounds," *IEEE Trans. Signal Process.*, vol. 58, no. 12, pp. 6140–6155, Dec. 2010.

[14] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley, 2012.

[15] N. Merhav, G. Kaplan, A. Lapidoth, and S. Shamai (Shitz), "On information rates for mismatched decoders," *IEEE Trans. Inf. Theory*, vol. 40, no. 6, pp. 1953–1967, Nov. 1994.

[16] A. Lapidoth and P. Narayan, "Reliable communication under channel uncertainty," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2148–2177, Oct. 1998.

[17] A. Ganti, A. Lapidoth, and I. E. Telatar, "Mismatched decoding revisited: General alphabets, channels with memory, and the wide-band limit," *IEEE Trans. Inf. Theory*, vol. 46, no. 7, pp. 2315–2328, Nov. 2000.

[18] J. Scarlett, A. Martinez, and A. Guillen i Fabregas, "Mismatched decoding: Error exponents, second-order rates and saddlepoint approximations," *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 2647–2666, May 2014.

[19] A. Somekh-Baruch, "On achievable rates and error exponents for channels with mismatched decoding," *IEEE Trans. Inf. Theory*, vol. 61, no. 2, pp. 727–740, Feb. 2015.

[20] R. M. Gray and T. Linder, "Mismatch in high-rate entropy-constrained vector quantization," *IEEE Trans. Inf. Theory*, vol. 49, no. 5, pp. 1204–1217, May 2003.

[21] D. Kazakos, "Signal detection under mismatch (corresp.)," *IEEE Trans. Inf. Theory*, vol. 28, no. 4, pp. 681–684, Jul. 1982.

[22] R. Schluter and H. Ney, "Model-based MCE bound to the true Bayes' error," *IEEE Signal Process. Lett.*, vol. 8, no. 5, pp. 131–133, May 2001.
[23] R. Schluter, M. Nussbaum-Thom, E. Beck, T. Alkhouli, and H. Ney, "Novel tight classification error bounds under mismatch conditions based on f-divergence," in *Proc. IEEE Inf. Theory Work. (ITW)*, Sep. 2013, pp. 432–436.
[24] S. Verdú, "Mismatched estimation and relative entropy," *IEEE Trans. Inf. Theory*, vol. 56, no. 8, pp. 3712–3720, Aug. 2010.
[25] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed. Baltimore, MD, USA: Johns Hopkins Univ. Press, 1996.
[26] J. Hamm and D. D. Lee, "Grassmann discriminant analysis: A unifying view on subspace-based learning," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 376–383, ACM.
[27] H. Reboredo, F. Renna, R. Calderbank, and M. R. D. Rodrigues, "Compressive classification of a mixture of Gaussians: Analysis, designs and geometrical interpretation," 2014, arXiv Preprint arXiv1401.6962.
[28] Q. Qiu and G. Sapiro, "Learning transformations for clustering and classification," *J. Mach. Learn. Res.*, vol. 16, pp. 187–225, Feb. 2015.
[29] R. Tron and R. Vidal, "A benchmark for the comparison of 3-D motion segmentation algorithms," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, June 2007, pp. 1–8.
[30] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[31] A. Galántai, "Subspaces, angles and pairs of orthogonal projections," *Linear Multilinear Algebr.*, vol. 56, no. 3, pp. 227–260, May 2008.
[32] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2012.
[33] I. Ipsen and R. Rehman, "Perturbation bounds for determinants and characteristic polynomials," *SIAM J. Matrix Anal. Appl.*, vol. 30, no. 2, pp. 762–776, July 2008.
[34] G. W. Stewart, "Perturbation theory for the singular value decomposition," Comput. Sci. Dept., Univ. of Maryland, College Park, MD, USA, Tech. Rep. UMIACS-TR-90-124, 1990.
[35] C. R. Rao, "Separation theorems for singular values of matrices and their applications in multivariate analysis," *J. Multivar. Anal.*, vol. 9, no. 3, pp. 362–377, Sep. 1979.

Jure Sokolić (S'14) received his Diploma in electrical engineering from University of Ljubljana in 2013. Currently, he is working towards his Ph.D. in the Department of Electrical and Electronic Engineering at University College London. His research interest focus on high-dimensional data processing and machine learning.

Francesco Renna (S'09–M'11) received his Laurea Specialistica Degree in telecommunication engineering and Ph.D. degree in information engineering, both from University of Padova, in 2006 and 2011, respectively. Between 2007 and 2015 he held visiting researcher and postdoctoral appointments at Infineon Technology AG, Princeton University, Georgia Institute of Technology (Lorraine Campus), Supélec, University of Porto, Duke University and University College London. Since 2016 he is a Marie Curie fellow at University of Cambridge. His research interests focus on high-dimensional information processing but also include physical layer security for multicarrier and multiantenna systems.

Robert Calderbank (M'89–SM'97–F'98) received the B.Sc. degree in 1975 from Warwick University, England, the M.Sc. degree in 1976 from Oxford University, England, and the Ph.D. degree in 1980 from the California Institute of Technology, all in mathematics.

Dr. Calderbank is Professor of Electrical Engineering at Duke University where he now directs the Information Initiative at Duke (iiD) after serving as Dean of Natural Sciences (2010–2013). Dr. Calderbank was previously Professor of Electrical Engineering and Mathematics at Princeton University where he directed the Program in Applied and Computational Mathematics. Prior to joining Princeton in 2004, he was Vice President for Research at AT&T, responsible for directing the first industrial research lab in the world where the primary focus is data at scale. At the start of his career at Bell Labs, innovations by Dr. Calderbank were incorporated in a progression of voiceband modem standards that moved communications practice close to the Shannon limit. Together with Peter Shor and colleagues at AT&T Labs he showed that good quantum error correcting codes exist and developed the group theoretic framework for quantum error correction. He is a co-inventor of space-time codes for wireless communication, where correlation of signals across different transmit antennas is the key to reliable transmission.

Dr. Calderbank served as Editor in Chief of the IEEE TRANSACTIONS ON INFORMATION THEORY from 1995 to 1998, and as Associate Editor for Coding Techniques from 1986 to 1989. He was a member of the Board of Governors of the IEEE Information Theory Society from 1991 to 1996 and from 2006 to 2008. Dr. Calderbank was honored by the IEEE Information Theory Prize Paper Award in 1995 for his work on the Z4 linearity of Kerdock and Preparata Codes (joint with A.R. Hammons Jr., P.V. Kumar, N.J.A. Sloane, and P. Sole), and again in 1999 for the invention of space-time codes (joint with V. Tarokh and N. Seshadri). He has received the 2006 IEEE Donald G. Fink Prize Paper Award, the IEEE Millennium Medal, the 2013 IEEE Richard W. Hamming Medal, and he was elected to the U.S. National Academy of Engineering in 2005.

Miguel R. D. Rodrigues (S'98–M'02–SM'15) received the Licenciatura degree in electrical engineering from the Faculty of Engineering of the University of Porto, Portugal in 1998 and the Ph.D. degree in electronic and electrical engineering from University College London, U.K., in 2002.

He is currently a Reader with the Department of Electronic and Electrical Engineering, University College London, U.K. He was previously with the Department of Computer Science, University of Porto, Portugal, where he also led the Information Theory and Communications Research Group at Instituto de Telecomunicações Porto. He has also held visiting research appointments at Princeton University, USA, Duke University, USA, Cambridge University, UK and University College London, UK in the period 2007 to 2013. His research interests are in the general areas of information theory, communications theory and signal processing.

Dr. Rodrigues was the recipient of the IEEE Communications and Information Theory Societies Joint Paper Award in 2011 for the work on Wireless Information-Theoretic Security (with M. Bloch, J. Barros and S. W. McLaughlin). He was also the recipient of the Prize Engenheiro António de Almeida, the Prize Engenheiro Cristiano Spratley, and the Merit Scholarship from the University of Porto, and the best student poster prize at the 2nd IMA Conference on Mathematics in Communications. He was also awarded doctoral and postdoctoral research fellowships from the Portuguese Foundation for Science and Technology, and research fellowships from Foundation Calouste Gulbenkian.

Dr. Rodrigues is an associate editor to IEEE COMMUNICATIONS LETTERS.