IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 29 May 2023, accepted 6 July 2023, date of publication 17 July 2023, date of current version 14 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3296098

RESEARCH ARTICLE

A Comparative Analysis of Early and Late Fusion for the Multimodal Two-Class Problem

LUIS MANUEL PEREIRA, ADDISSON SALAZAR, (Member, IEEE), AND LUIS VERGARA
Institute of Telecommunications and Multimedia Applications, Universitat Politècnica de València, 46022 Valencia, Spain

Corresponding author: Luis Vergara (lvergara@dcom.upv.es)

This work was supported in part by MCIN/AEI/10.13039/501100011033 under Grant PRE2018-085092, and in part by Universitat Politècnica de València.

ABSTRACT In this article we carry out a comparison between early (feature) and late (score) multimodal fusion, for the two-class problem. The comparison is made first from a general perspective, and then from a specific mathematical analysis. Thus, we deduce the error probability expressions for the uncorrelated and correlated multivariate Gaussian distribution, assuming perfect model knowledge (Bayes error rates). We also deduce the corresponding expressions when the model is to be learned from a finite training set, demonstrating its convergence to the Bayes error rates as the training set size goes to infinite. These expressions also demonstrates that early fusion is the best option with model knowledge, and that both early and late fusion degrade due to a finite training set. This degradation is showed to be greater for early fusion due to the dimensionality increase of the feature space, so, eventually, late fusion could be a better option in a practical setting. The mathematical analysis also suggests the convenience of using a, so called, convergence factor, to quantify if a training set size is appropriate for the error probability to be close enough to the Bayes error rate. Different simulated experiments have been made to verify the validity of the mathematical analysis, as well as its possible extension to non-Gaussian models.

INDEX TERMS Multimodal two-class classification, early fusion, late fusion, probability of error, training set size.

I. INTRODUCTION

Data fusion is a consolidated concept in the areas of pattern recognition, machine learning and artificial intelligence. It refers to methods which combine data from different channels, with the aim of improving the performance of the overall data processing system. There exists a number of comprehensive reviews which introduce a variety of taxonomies, categorizations and bibliographical analysis. Although some approaches are general [1], most of the reviews have different perspectives: modality fusion [2], [3], [4], sensor fusion [5], [6] or classifier combination [7], [8]. Federated analytics [9] is also a recent related paradigm where fusion from different parties is implemented without revealing the private raw data. Apart from specific issues, all the approaches share some common essential concepts. One especially relevant is the distinction between early and late fusion. Early fusion normally refers to directly combining signals from different source sensors or, most commonly, to combine features separately extracted from every source and then provided to a unique classifier. On the other hand, late fusion refers to fuse scores (soft late fusion) or decisions (hard late fusion) provided by different classifiers operating on the same target problem.

So far, there exists some research comparing early and late fusion [10], [11], [12], [13], [14], [15], [16], [17]. Unfortunately it is focused in specific application domains and are mostly experimental. There is a lack of contributions in establishing first principles from theoretical approaches in this area. In this article we intend to contribute to fill this gap by providing some understanding about the implications of selecting early or late fusion and even some guidelines for selecting the most appropriate option. The contributions have general interest (not application domain dependence), and are supported by mathematical analysis.

The associate editor coordinating the review of this manuscript and approving it for publication was Jad Nasreddine.

To be specific we will consider early fusion of features corresponding to different modalities. Any preprocessing related to modality synchronization, feature normalization or dimension reduction is out of the scope of this work. Thus early fusion will consider a feature vector formed by the direct concatenation of features from individual modalities. This vector will be the input to a unique classifier which provides a unique score to be compared with a threshold to decide between class 1 and class 2. However, in (soft) late fusion every modality feature vector is the input to a separate classifier, then scores from all the modalities are fused to give a unique score before comparison with a threshold. Fig. 1 describes the two options. We consider I modalities, the feature vector of the i-th modality is $\mathbf{x}_i$, so the early fusion concatenation of all the feature vectors is vector $\mathbf{x} = [\mathbf{x}_1 \ldots \mathbf{x}_I]^T$, this is the input to a unique classifier providing the score s and, after thresholding, a binary decision d of early fusion. Correspondingly, s_i is the score given by every modality classifier, s_s the (soft) fused score and d_s the binary decision of late fusion.

The paper is organized as follows. In the next section we provide a general comparison between early and late fusion, in terms of classifier error probability, both assuming known or estimated models. The comparison is valid for any class-conditional distributions of the data. In Section III we present a mathematical analysis for the multivariate uncorrelated Gaussian case. We deduce expressions for the probabilities of error assuming knowledge of the model parameters (Bayes error rates) as well as considering that the parameters are to be learned from training data. In Section IV the analysis is extended to the correlated case. Discussion and conclusions are respectively considered in Sections V and VI.

II. A GENERAL COMPARISON

Let us assume the scenario of two classes ($k = 1, 2$). Suppose we perform early fusion by selecting the class that maximizes the *a posteriori* probability given $\mathbf{x} = [\mathbf{x}_1 \ldots \mathbf{x}_I]^T$, i.e., the decision rule will be:

$$P\left(k = 1/\mathbf{x}\right) \underset{k=2}{\overset{k=1}{\gtrless}} P\left(k = 2/\mathbf{x}\right) \Leftrightarrow P\left(k = 1/\mathbf{x}\right) \underset{k=2}{\overset{k=1}{\gtrless}} 0.5. \tag{1}$$

where we have taken into account that $P\left(k = 1/\mathbf{x}\right) + P\left(k = 2/\mathbf{x}\right) = 1$. Note that rule (1) implies minimization of the probability of error if we assume that the costs of being wrong in the decisions are symmetric and normalized to 1. That is, with exact knowledge of the *a posteriori* probability $P\left(k = 1/\mathbf{x}\right)$, there is no other fusion rule that allows us to reduce the probability of error more than (1).

Let us now assume that we perform late fusion. For this we generate an *a posteriori* probability (score) separately for each modality $s_i = P\left(k = 1/\mathbf{x}_i\right)$ $i = 1 \ldots I$. Notice again that since $P\left(k = 1/\mathbf{x}_i\right) + P\left(k = 2/\mathbf{x}_i\right) = 1$ $i = 1 \ldots I$, it is sufficient to consider the defined scores s_i $i = 1 \ldots I$. Then we generate a new score s_s by means of a certain fusion function $s_s = f\left(s_1, \ldots, s_I\right)$ $0 \leq s \leq 1$, and the new decision rule will be

$$s \underset{k=2}{\overset{k=1}{\gtrless}} 0.5. \tag{2}$$

Notice that $f\left(s_1, \ldots, s_I\right) = f\left(P\left(k = 1/\mathbf{x}_1\right), \ldots, P\left(k = 1/\mathbf{x}_I\right)\right)$ is ultimately a function of the multivariate random variables $\mathbf{x}_i$ $i = 1 \ldots I$, which will in general be different from the function $P\left(k = 1/\mathbf{x}\right)$ in (1). Therefore (2) can never achieve a lower error probability than (1). However, exact knowledge of $P\left(k = 1/\mathbf{x}\right)$ is only strictly possible if we have an infinite training set and assume statistical consistency in the estimation of $P\left(k = 1/\mathbf{x}\right)$, i.e. $\hat{P}\left(k = 1/\mathbf{x}\right) \overset{N\to\infty}{\to} P\left(k = 1/\mathbf{x}\right)$. Where N is the size of the training set for each class (for simplicity we will assume the same for both classes throughout the paper). From a practical point of view accurate estimation implies having "sufficiently" large training set sizes for each class. Unfortunately it is not easy to determine what should be an appropriate size, as it depends on several factors like type of classifier, data distributions and class separability. Efforts have been given to this problem ([18], [19], [20], [21], [22], [23] are some representative examples), the research was basically experimental and the conclusions not easy to generalize. Recently some theoretical learning curves have been deduced in the framework of Bayesian classifiers with parametric models [24]. In any case, many real data problems in machine learning are constrained by a limited amount of available samples for training. Moreover, as far as we are concerned here, it is important to bear in mind that training sizes must increase with increasing feature vector dimension, and they do so in a generally nonlinear fashion. So, given a training set size N, the increase in the dimension of $\mathbf{x} = [\mathbf{x}_1 \ldots \mathbf{x}_I]^T$ with respect to the dimensions of $\mathbf{x}_i$ $i = 1 \ldots I$, means that the actual performance of early fusion could severely degrade and be under the performance of late fusion. Next we delve into this problem.

Let us focus on Bayesian generative methods. Suppose a feature vector $\mathbf{y}$. The estimation of the *a posteriori* probability is done by applying Bayes' rule

$$\hat{P}\left(k = 1/\mathbf{y}\right) = \frac{\hat{p}\left(\mathbf{y}/k = 1\right)\hat{P}_1}{\hat{P}\left(\mathbf{y}\right)} = \frac{\hat{p}\left(\mathbf{y}/k = 1\right)\hat{P}_1}{\hat{p}\left(\mathbf{y}/k = 1\right)\hat{P}_1 + \hat{p}\left(\mathbf{y}/k = 2\right)\hat{P}_2}. \tag{3}$$

So we must estimate the *a priori* probabilities P_1, P_2 and the probability densities conditional on each class $p\left(\mathbf{y}/k = 1\right), p\left(\mathbf{y}/k = 2\right)$. The first two are usually estimated by considering the proportion of instances (feature vectors) of each class in the training set or simply assumed to be equal ($\hat{P}_1 = \hat{P}_2 = 0.5$), if there is no collateral information suggesting other values. However, the probabilities conditional on each class are more complex to estimate. Nonparametric methods and parametric methods can be used.

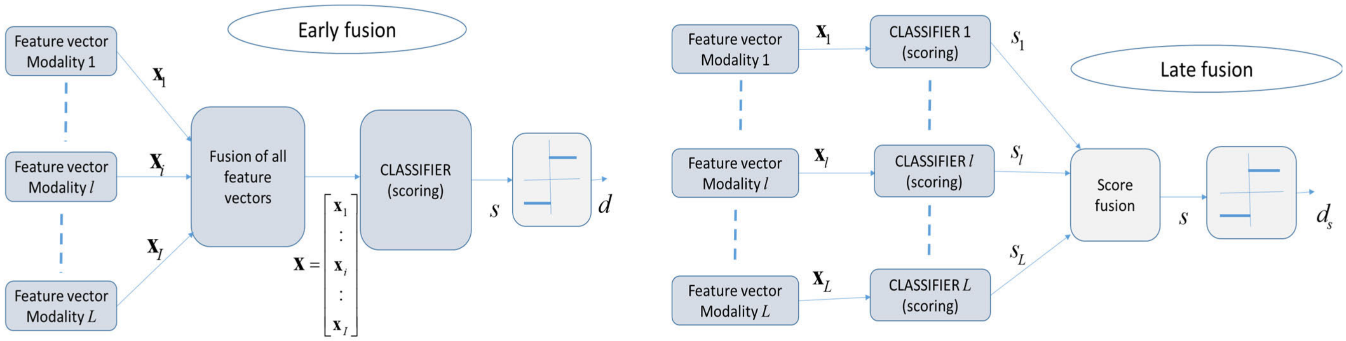
Early fusion
Feature vector Modality 1
Feature vector Modality l
Feature vector Modality L
$\mathbf{x}_1$
$\mathbf{x}_l$
$\mathbf{x}_l$
Fusion of all feature vectors
$\mathbf{x} = [\mathbf{x}_1 : \mathbf{x}_i : \mathbf{x}_l]$
CLASSIFIER (scoring)
s
d
Late fusion
Feature vector Modality 1
Feature vector Modality l
Feature vector Modality L
$\mathbf{x}_1$
$\mathbf{x}_l$
$\mathbf{x}_L$
CLASSIFIER 1 (scoring)
CLASSIFIER l (scoring)
CLASSIFIER L (scoring)
s_1
s_l
s_L
Score fusion
s
d_s

FIGURE 1. Early and late fusion schemes.

The simplest nonparametric method is to calculate a multidimensional histogram for each class from the training data. Essentially, one divides the hyperspace into small hypercubes and measures the proportion of instances of a class that are inside the hypercube with respect to the total number of instances of that class available in the training set. Let us call M the dimension of vector $\mathbf{y}$. Suppose initially that $M = 1$, so $\mathbf{y}$ is really a scalar y, and the hypercubes are simply intervals. Let us also assume that we have a number $N^{\langle 1\rangle}$ of training instances to compute the histogram and a number h of intervals. The $N^{\langle 1\rangle}$ instances will be unevenly distributed among the intervals, producing estimates of the conditional probability densities associated with each of the intervals. If we increase the dimension of the feature vector to $M = 2$, (i.e. in this case the hypercubes are square two-dimensional cells), and we wish to maintain the same accuracy in the histogram calculation, we must consider h intervals for each of the two components of the vector. This implies a number of h^2 two-dimensional cells. For an arbitrary M, the number of hypercubes will be h^M, therefore to keep the same quality of the estimation we should also increase the number of training instances to be distributed among the total number of hypercubes, i.e.

$$\frac{N^{\langle 1\rangle}}{h} = \frac{N^{\langle M\rangle}}{h^M} \Rightarrow N^{\langle M\rangle} = \frac{N^{\langle 1\rangle}}{h} h^M. \quad (4)$$

This implies an exponential growth of the training set with the dimension of the feature vector. This is part of what is commonly referred to as "the curse of dimensionality". Let us apply (4) to the case of fusing I modalities with feature vectors $\mathbf{x}_i$ $i = 1 \ldots I$ of equal dimension M. Early fusion will involve feature vectors $\mathbf{x} = [\mathbf{x}_1 \ldots \mathbf{x}_I]^T$ of dimension $M \times I$ and late fusion will involve I vectors of dimension M, hence considering (4)

$$\left.\begin{aligned} &\textit{early fusion} \quad N^{\langle M\times I\rangle} = \frac{N^{\langle 1\rangle}}{h} h^{MxI} \\ &\textit{late fusion} \quad N^{\langle M\rangle} \times I = \frac{N^{\langle 1\rangle}}{h} h^M \times I \end{aligned}\right\} \frac{N^{\langle M\times I\rangle}}{N^{\langle M\rangle} \times I}$$

$$= \frac{h^{MxI}}{h^M \times I} = \frac{1}{I} h^{M\times(I-1)}. \quad (5)$$

where we observe the large increase in the size of the training set required in early fusion with respect to late fusion. Even for low values of h and M, the increase required for early fusion to maintain accuracy, is exorbitant. For example for $h = 10, I = 2$ and $M = 3$ turns out to be $10^3/2 = 500$, we need to multiply by 500 the size of the training set for each class and modality.

Similar considerations can be made with respect to other nonparametric methods such as Parzen's window method [25], which can be understood as a generalization of the hypercubic kernel implicit in a histogram to other multidimensional (typically Gaussian) kernels. The result is basically a "smoothed" version of the histogram. Although usually considered a discriminative method, the popular "K-nearest neighbours" (K-nn) method [26] can also be interpreted as a nonparametric generative method. In this case the size of the hypercubes is variable as it must include the K instances of the total training set closest to the instance under test. Given a training set of size $N^{\langle 1\rangle}$ and an initial dimension $M = 1$, if we increase the dimension of the hyperspace by adding more components to the feature vector, the distance between the $N^{\langle 1\rangle}$ instances increases and consequently so does the size of the hypercubes that include the K-nearest neighbors, which are no longer really close to the instance under test. To avoid this problem we must increase the size of the training set to $N^{\langle M\rangle}$, which can be seen to grow exponentially with M [26].

The huge increases in the size of the training set of nonparametric generative methods can be partly alleviated by considering parametric methods. In these methods, certain models are assumed for $p\left(\mathbf{y}/k = 1\right), p\left(\mathbf{y}/k = 2\right)$ specified by a finite (and parsimonious) number of parameters to be estimated for each class from the training instances. Normally the number of parameters grows with M, so again it is necessary to increase N as we increase M. The required increase depends on the model, so for example if we assume a multivariate Gaussian model we need to estimate a $M \times M$ symmetric covariance matrix plus a mean vector of dimension M. This amounts to a total number of $\left(M^2 + 3M\right)/2$ parameters to be estimated. Let us assume that in a first approximation ("rule of thumb") a certain number of constant instances C

are required for each parameter to be estimated, then let us consider a comparison similar to (5) between early and late fusion

$$\left.\begin{array}{l} \textit{early fusion} \\ N^{\langle M\times I\rangle} = C\left(\frac{1}{2}\left((M\times I)^2 + 3\,(M\times I)\right)\right) \\ \textit{late fusion} \\ N^{\langle M\rangle}\times I = C\frac{1}{2}\left(M^2+3M\right)\times I \end{array}\right\} \frac{N^{\langle M\times I\rangle}}{N^{\langle M\rangle}\times I}$$

$$= \frac{\left(M^2\times I\right)+3M}{M^2+3M}. \qquad (6)$$

Notice that for M and/or I large $\frac{N^{\langle M\times I\rangle}}{N^{\langle M\rangle}\times I} \simeq I$, i.e., we need to multiply by I, the number of instances per class and modality. This is clearly a more conservative requirement than the one deduced from (5), but it may nevertheless imply a significant increase of the training set size. Furthermore the assumed model may not be sufficient to capture the probabilistic structure of the data, other models may require more parameters. For example the Gaussian mixture model will require to estimate a number of parameters $\left(M^2+3M\right)/2$ multiplied by the number of mixture components.

In the next two sections we present a mathematical analysis to get a more formal comparison of early and late fusion. A (parametric) multivariate Gaussian model will be assumed to make the analysis tractable, but conclusions will be consistent with this previous general discussions. Firstly the case of uncorrelated features will be considered so that only centroids are to be estimated, then the analysis will be extended to the correlated case where covariance matrices estimates are required.

III. A CASE OF ANALYSIS, UNCORRELATED DATA

A. KNOWN MODEL PARAMETERS

We will now consider a particular case that allows us to illustrate and verify the conclusions of the previous section. This is a simple parametric model scenario, where the number of parameters involved has been reduced as much as possible in order to make the analysis tractable and to focus on the essential ideas outlined above. Thus, we will consider two equiprobable classes with multivariate Gaussian generative models. Furthermore, all the feature vectors of every modality $\mathbf{x}_i$ $i = 1\ldots I$ will have the same dimension M.

Let us call $\mathbf{m}_i^{(k)}$ the mean vector (centroid) of modality i and class k, and $\mathbf{C}_i^{(k)}$ the corresponding covariance matrix. We will assume in principle that

$$\left.\begin{array}{l} \mathbf{m}_i^{(1)} = \mathbf{m}_i \quad \|\mathbf{m}_i\|^2 = \mathbf{m}_i^T\mathbf{m}_i \neq 0 \\ \mathbf{m}_i^{(2)} = \mathbf{0}_M = \left[\underbrace{0\ldots 0}_{M}\right]^T \end{array}\right\} i = 1\ldots I$$

$$\mathbf{C}_i^{(k)} = \mathbf{I}_{MxM}, \quad i = 1,\ldots I \quad k = 1, 2. \qquad (7)$$

where $\mathbf{I}_{MxM}$ is the identity matrix of dimension $M \times M$. We see that the centroids of class 2 are the origin of coordinates in all modalities, whereas the centroids of class 1 are $\mathbf{m}_i$ $i = 1\ldots I$. These later can take arbitrary values other than $\mathbf{0}_M$, since we impose the nonzero Euclidean norm condition. Moreover, as indicated in (7), we consider that the features of each class are uncorrelated and variance normalized. Thus, the model is defined only from $\mathbf{m}_i$ $i = 1\ldots I$. We will assume in this section that $\mathbf{m}_i$ $i = 1\ldots I$ are known. In Section III-B we will consider the effects that occur in a practical scenario in which they must be estimated from training sets (estimated model). On the other hand, in Section IV we will extend the case analysis to arbitrary covariance matrices, both with known and estimated model.

In the following, we will calculate the error probabilities corresponding to the classification problem defined above for early and late fusion. As we are assuming perfect model knowledge, the calculated error probabilities coincide with the so called Bayes error rates, so they are the minimum achievable probability of errors for the defined classification problem.

Let us start with the early fusion and take into account (7), hence vector $\mathbf{x} = [\mathbf{x}_1 \ldots \mathbf{x}_I]^T$ will be multivariate Gaussian characterized by

$$\begin{aligned} \mathbf{m}^{(1)} &= \mathbf{m} = [\mathbf{m}_1, \ldots, \mathbf{m}_I]^T, \quad \mathbf{m}^{(2)} = \mathbf{0}_{(M\times I)} \\ \mathbf{C}^{(k)} &= \mathbf{I}_{(M\times I)\times(M\times I)}, \quad k = 1, 2. \end{aligned} \qquad (8)$$

where we have made the additional assumption that features from different modalities are also uncorrelated. Let us consider the test (1) that minimizes the probability of error for the early fusion case. For convenience we will express it in the form

$$\frac{P\left(k=1/\mathbf{x}\right)}{P\left(k=2/\mathbf{x}\right)} = \frac{\frac{p(\mathbf{x}/k=1)P_1}{p(\mathbf{x})}}{\frac{p(\mathbf{x}/k=2)P_2}{p(\mathbf{x})}} = \frac{p\left(\mathbf{x}/k=1\right)}{p\left(\mathbf{x}/k=2\right)} \underset{k=2}{\overset{k=1}{\gtrless}} 1. \qquad (9)$$

where we have applied Bayes' theorem and considered that both classes are equiprobable, $P_1 = P_2 = 0.5$. On the other hand, according to the Gaussian model (8)

$$\begin{aligned} p\left(\mathbf{x}/k=1\right) &= \frac{1}{(2\pi)^M}\exp\left(-\frac{1}{2}\left(\mathbf{x}-\mathbf{m}\right)^T\left(\mathbf{x}-\mathbf{m}\right)\right) \\ p\left(\mathbf{x}/k=2\right) &= \frac{1}{(2\pi)^M}\exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{x}\right). \end{aligned} \qquad (10)$$

Therefore it is fulfilled

$$\begin{aligned} &\frac{p\left(\mathbf{x}/k=1\right)}{p\left(\mathbf{x}/k=2\right)} \\ &\quad = \exp\left(-\frac{1}{2}\left(\mathbf{x}-\mathbf{m}\right)^T\left(\mathbf{x}-\mathbf{m}\right)+\frac{1}{2}\mathbf{x}^T\mathbf{x}\right) \\ &\quad = \exp\left(\mathbf{m}^T\mathbf{x}-\frac{1}{2}\mathbf{m}^T\mathbf{m}\right) \\ &\quad \Rightarrow \frac{p\left(\mathbf{x}/k=1\right)}{p\left(\mathbf{x}/k=2\right)} \underset{k=2}{\overset{k=1}{\gtrless}} 1 \Leftrightarrow \ln\frac{p\left(\mathbf{x}/k=1\right)}{p\left(\mathbf{x}/k=2\right)} \underset{k=2}{\overset{k=1}{\gtrless}} 0 \\ &\quad \Leftrightarrow \frac{\mathbf{m}^T\mathbf{x}}{\mathbf{m}^T\mathbf{m}} \underset{k=2}{\overset{k=1}{\gtrless}} \frac{1}{2}. \end{aligned} \qquad (11)$$

Let us denote $z = \frac{\mathbf{m}^T\mathbf{x}}{\mathbf{m}^T\mathbf{m}}$. Being a linear combination of Gaussian random variables, z will also be a Gaussian random variable, let us calculate its mean and variance for each class.

$$\begin{aligned}
E^{(1)}(z) &= \frac{\mathbf{m}^T E^{(1)}(\mathbf{x})}{\mathbf{m}^T\mathbf{m}} = \frac{\mathbf{m}^T\mathbf{m}}{\mathbf{m}^T\mathbf{m}} = 1\\
\mathrm{var}^{(1)}(z) &= E^{(1)}\left(z^2\right) - E^{(1)^2}(z)\\
&= \frac{1}{\left(\mathbf{m}^T\mathbf{m}\right)^2} E^{(1)}\left(\mathbf{m}^T\mathbf{x}\mathbf{x}^T\mathbf{m}\right) - 1\\
&= \frac{1}{\left(\mathbf{m}^T\mathbf{m}\right)^2}\left(\mathbf{m}^T \underbrace{E^{(1)}\left(\mathbf{x}\mathbf{x}^T\right)}_{\mathbf{I}_{2Mx2M}+\mathbf{m}\mathbf{m}^T}\mathbf{m}\right) - 1 = \frac{1}{\mathbf{m}^T\mathbf{m}}\\
E^{(2)}(z) &= \frac{\mathbf{m}^T E^{(2)}(\mathbf{x})}{\mathbf{m}^T\mathbf{m}} = 0;\\
\mathrm{var}^{(2)}(z) &= E^{(2)}\left(z^2\right) - E^{(2)^2}(z) = \frac{1}{\left(\mathbf{m}^T\mathbf{m}\right)^2} E^{(2)}\left(\mathbf{m}^T\mathbf{x}\mathbf{x}^T\mathbf{m}\right)\\
&= \frac{1}{\left(\mathbf{m}^T\mathbf{m}\right)^2}\left(\mathbf{m}^T \underbrace{E^{(2)}\left(\mathbf{x}\mathbf{x}^T\right)}_{\mathbf{I}_{2Mx2M}}\mathbf{m}\right) = \frac{1}{\mathbf{m}^T\mathbf{m}}. \qquad (12)
\end{aligned}$$

For simplicity of notation, we will call $\nu = \frac{1}{\mathbf{m}^T\mathbf{m}}$, a parameter which is the inverse of the Euclidean distance between the centroids of the two classes in the fusion hyperspace. Considering (12) we can write the probability densities of z for each class

$$p^{(1)}(z) = \frac{1}{\sqrt{2\pi\nu}} e^{-\frac{(z-1)^2}{2\nu}} \quad p^{(2)}(z) = \frac{1}{\sqrt{2\pi\nu}} e^{-\frac{z^2}{2\nu}}. \quad (13)$$

We are now in a position to calculate the probability of error for the early fusion case

$$\begin{aligned}
P_e &= \frac{1}{2}\Pr\left(z > \frac{1}{2}\Big/ k=2\right) + \frac{1}{2}\Pr\left(z < \frac{1}{2}\Big/ k=1\right)\\
&= \frac{1}{2}\int_{\frac{1}{2}}^{\infty} \frac{1}{\sqrt{2\pi\nu}} e^{-\frac{z^2}{2\nu}} dz + \frac{1}{2}\int_{-\infty}^{\frac{1}{2}} \frac{1}{\sqrt{2\pi\nu}} e^{-\frac{(z-1)^2}{2\nu}} dz\\
&= \frac{1}{2}\int_{\frac{1}{2}}^{\infty} \frac{1}{\sqrt{2\pi\nu}} e^{-\frac{z^2}{2\nu}} dz + \frac{1}{2}\int_{-\infty}^{-\frac{1}{2}} \frac{1}{\sqrt{2\pi\nu}} e^{-\frac{z^2}{2\nu}} dz\\
&= \int_{\frac{1}{2}}^{\infty} \frac{1}{\sqrt{2\pi\nu}} e^{-\frac{z^2}{2\nu}} dz = \int_{\frac{1}{2\sqrt{2\nu}}}^{\infty} \frac{1}{\sqrt{\pi}} e^{-u^2} du\\
&= \frac{1}{2} erfc\left(\frac{1}{2\sqrt{2\nu}}\right). \qquad (14)
\end{aligned}$$

where $erfc(x) = 2\int_x^{\infty} \frac{1}{\sqrt{\pi}} e^{-w^2} dw$ is the so-called complementary error function. Note that $\underbrace{P_e}_{\nu\to\infty} \to 0.5$, $\underbrace{P_e}_{\nu\to 0} \to 0$, i.e. as the two classes approach each other, the classifier becomes fully random, and as they separate, the classifier converges to an error-free classifier. On the other hand we note that the key parameter of the error probability is the "separability" between classes $1/\nu$.

All of the above is applicable to each modality separately by considering the statistics $z_i = \frac{\mathbf{m}_i^T\mathbf{x}}{\mathbf{m}_i^T\mathbf{m}_i}$ $i = 1\ldots I$, which will have Gaussian distributions in each of the two classes characterized by

$$\begin{aligned}
&E^{(1)}(z_i) = 1, \quad E^{(2)}(z_i) = 0\\
&\mathrm{var}^{(1)}(z_i) = \mathrm{var}^{(2)}(z_i) = \frac{1}{\mathbf{m}_i^T\mathbf{m}_i} = \nu_i \quad i = 1\ldots I. (15)
\end{aligned}$$

We can therefore take advantage of (14), to deduce the error probability P_{ei} corresponding to the use of only modality i by simply replacing ν by ν_i

$$P_{ei} = \frac{1}{2} erfc\left(\frac{1}{2\sqrt{2\nu_i}}\right) \quad i = 1\ldots I. \qquad (16)$$

Note that $\nu = \frac{1}{[\mathbf{m}_1,\ldots,\mathbf{m}_I]^T[\mathbf{m}_1,\ldots,\mathbf{m}_I]} = \frac{1}{\sum_{i=1}^{I}\mathbf{m}_i^T\mathbf{m}_i} = \frac{1}{\sum_{i=1}^{I}\frac{1}{\nu_i}}$ is the harmonic mean of $\nu_1, \ldots, \nu_I$ divided by I. Also notice that $\frac{1}{\nu} = \sum_{i=1}^{I}\frac{1}{\nu_i} \Rightarrow \frac{1}{\nu} > \frac{1}{\nu_i}$ $i = 1\ldots I$, i.e., the class separability in the $M \times I$ hyperspace (early fusion) is greater than de separability in any of the M-dimensional hyperspace of every separate modality. Moreover as $erfc(x)$ is a monotonically decreasing function and $\nu < \nu_i$ $i = 1\ldots I$, the error probabilities will decrease as the variance parameter decreases and it will be true that $P_e < P_{ei}$ $i = 1\ldots I$. In other words, early fusion with knowledge of the model parameters always reduces the probability of error with respect to any individual modality.

Let us now consider late fusion. From the above it is clear that the optimal test applicable in each modality, can be expressed in an equivalent way

$$P\left(k = 1/\mathbf{x}_i\right) \mathop{\gtrless}\limits^{k=1}_{k=2} \frac{1}{2} \Leftrightarrow \frac{\mathbf{m}_i^T\mathbf{x}_i}{\mathbf{m}_i^T\mathbf{m}_i} \mathop{\gtrless}\limits^{k=1}_{k=2} \frac{1}{2} \quad i = 1\ldots I. \quad (17)$$

Therefore, to make the analysis tractable we will assume that the scores to be fused are $z_i = \frac{\mathbf{m}_i^T\mathbf{x}_i}{\mathbf{m}_i^T\mathbf{m}_i}$ $i = 1\ldots I$. Furthermore, for simplicity, we will focus on a fusion function that calculates the average of the I previous scores, which we will call z_s. Bearing in mind that the mean of a sum is the sum of the means, and that the variance of a sum (assuming independence) is the sum of the variances, we conclude that:

$$\begin{aligned}
z_s &= \frac{1}{I}\sum_{i=1}^{I} z_i\\
E^{(1)}(z_s) &= \frac{1}{I}\sum_{i=1}^{I} E^{(1)}(z_i) = 1;\\
E^{(2)}(z_s) &= \frac{1}{I}\sum_{i=1}^{I} E^{(2)}(z_i) = 0\\
\mathrm{var}^{(1)}(z_s) &= \frac{1}{I^2}\sum_{i=1}^{I}\mathrm{var}^{(1)}(z_i) = \frac{1}{I^2}\sum_{i=1}^{I}\nu_i = \nu_s = \mathrm{var}^{(2)}(z_s).
\end{aligned} \qquad (18)$$

TABLE 1. Error probabilities assuming knowledge of model parameters.

Method	Probability of error	Properties
Modality i	$P_{ei} = \frac{1}{2} erfc\left(\frac{1}{2\sqrt{2\nu_i}}\right)$ (16)	ν_i parameter inverse to separability between classes in modality i
Early fusion	$P_e = \frac{1}{2} erfc\left(\frac{1}{2\sqrt{2\nu}}\right); \quad \nu = \frac{1}{\sum_{i=1}^{I} \frac{1}{\nu_i}}$ (14) (14)	$P_e < P_{ei} \quad i = 1 \ldots I$ $P_e \le P_{es}$
Late fusion	$P_{es} = \frac{1}{2} erfc\left(\frac{1}{2\sqrt{2\nu_s}}\right); \quad \nu_s = \frac{1}{I^2}\sum_{i=1}^{I} \nu_i$ (19)	$P_{es} \le P_{ei}$ if $\nu_s \le \nu_i$

Again, we can take advantage of (14) to derive the error probability corresponding to late fusion by simply replacing ν with ν_s, where ν_s is the arithmetic mean of $\nu_1, \ldots, \nu_I$ divided by I.

$$P_{es} = \frac{1}{2} erfc\left(\frac{1}{2\sqrt{2\nu_s}}\right). \tag{19}$$

It is well known that the harmonic mean is always less than or equal to the arithmetic mean, therefore $\nu \le \nu_s \Rightarrow P_e \le P_{es}$, the early fusion gives a probability of error less than or equal to the late fusion. Equality $P_e = P_{es}$ holds if $\nu_1 = \nu_2 = \ldots = \nu_I$. That is, if the separability between classes is the same in all modalities, the mean of the scores provides the same error probability as early fusion.

With respect to considering each modality separately, late fusion will provide a lower probability of error than either modality separately if it is satisfied that $\nu_s = \frac{1}{I^2}\sum_{i=1}^{I} \nu_i \le \nu_j \; j = 1 \ldots I$. For example, if $I = 2$ it should be $\frac{1}{4}(\nu_1 + \nu_2) \le \nu_1$ and $\frac{1}{4}(\nu_1 + \nu_2) \le \nu_2$, i.e. $\nu_2 \le 3\nu_1$ and $\nu_1 \le 3\nu_2$.Therefore, late fusion outperforms both separate modalities if certain conditions are met. This is illustrated by Fig. 2 for ν_1 and ν_2 varying between 0 and 1. The region between the two lines is the geometric locus of the pairs ν_1, ν_2 where late fusion produces improvement over both separate modalities. Intuitively, what this indicates is that if the separability between classes in one of the modalities is clearly higher than the separability between classes in the other modality, it could be better to work only with the modality with the higher separability, and not incorporate the other modality where the separability is small. For example, if $\nu_2 > 3\nu_1 \Leftrightarrow \|\mathbf{m}_2\|^2 < (1/3)\|\mathbf{m}_1\|^2$, the separability of modality 2 is less than 1/3 of the separability of modality 1, and it is better to work only with modality 1 than to late fusing with modality 2. This is not the case with early fusion, where there will always be an improvement of the error probability with respect to both modalities working separately.

Table 1 summarizes the results obtained in the analysis of the case under consideration. These results are consistent with what was stated in previous section: knowing the parameters of the models, early fusion outperforms late fusion. It is also worth noting that in all cases, the corresponding error probability is determined by the inverse measures of separabilities $\nu_1, \ldots, \nu_I$.

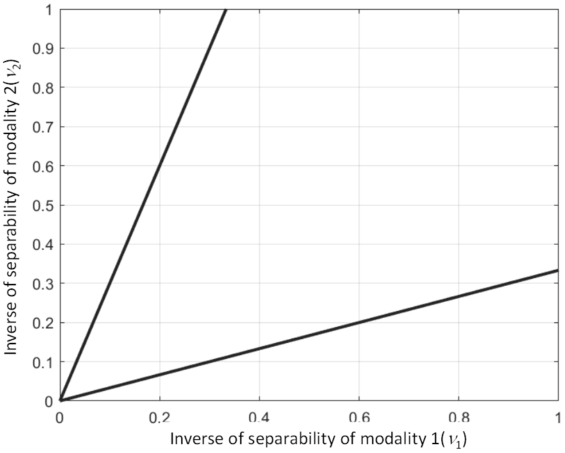

FIGURE 2. Late fusion improves both modalities for separabilities inside the two lines.

We illustrate the above results in Fig. 3. It represents the different probabilities of error for the case $I = 2$, assuming that $\nu_2 = 0.2$ is fixed and ν_1 varies between 0 and 1. Thus, the early fusion always outperforms late fusion as well as the modalities working individually. Only for the case $\nu_2 = \nu_1 = 0.2$ does late fusion provides the same error probability than early fusion (harmonic mean coincidences with arithmetic mean). Therefore, for equal separability of both modalities, the mean is the optimum late fuser. On the other hand, late fusion always outperforms the individual modalities if the aforementioned conditions $\nu_2 \le 3\nu_1, \nu_1 \le 3\nu_2$ are met, i.e. in the case of Fig. 3, $0.2 \le 3\nu_1 \Rightarrow \nu_1 \ge \frac{0.2}{3} = 0.0\widehat{6}$.

B. ESTIMATED MODEL PARAMETERS

In a real situation, models of probabilistic distributions must be estimated from training data. In our example, we have assumed Gaussian models defined by the parameters $\mathbf{m}_i$

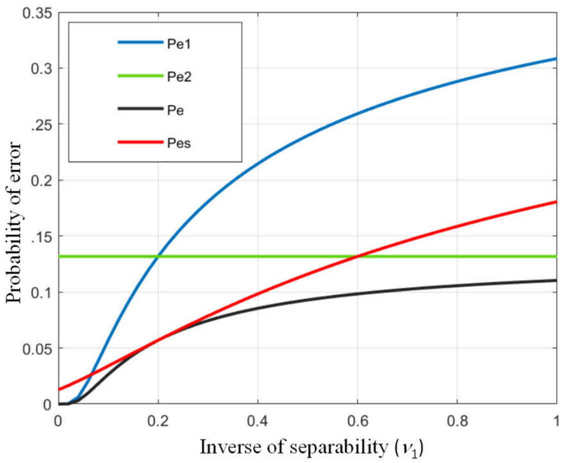

FIGURE 3. Error probability for $I = 2$, $\nu_2 = 0.2$ and ν_1 varying between 0 and 1: Modality 1, Modality 2, Early fusion, Late fusion.

$i = 1 \ldots I$. These are the centroids of class 1 respectively corresponding to each modality since, for simplicity, we have assumed class 2 to be centered at the origin and the covariance equal to the identity matrix in both classes. Let us therefore assume that we have N training vectors of class 1 in every modality $\mathbf{x}_{im}^{(1)}$ $m = 1 \ldots N$ $i = 1 \ldots I$. We assume equal training set sizes for all modalities in order to reduce the number of variables involved as much as possible and without impeding the illustration of the general considerations of Section I. We will use the sample means to estimate the centroids

$$\hat{\mathbf{m}}_i = \frac{1}{N}\sum_{m=1}^{N}\mathbf{x}_{im}^{(1)} \quad i = 1 \ldots I. \tag{20}$$

It is well known that, under Gaussian models, (20) is the maximum likelihood estimator, which turns out to be unbiased and reduces the original variance of the data by a factor N, thus

$$E\left(\hat{\mathbf{m}}_i\right) = \mathbf{m}_i; \quad \mathbf{C}_{\hat{\mathbf{m}}_i\hat{\mathbf{m}}_i} = E\left(\hat{\mathbf{m}}_i\hat{\mathbf{m}}_i^T\right) = \frac{1}{N}\mathbf{I}_{MxM} \quad i = 1 \ldots I. \tag{21}$$

Our aim is to deduce how the fact that we work with centroid estimates affects the expressions of the error probabilities. Let us start, as we did in the previous section, with early fusion. To do so, we begin from the equivalent form of the optimal test (17), but considering the estimated value of $\mathbf{m}$

$$\hat{z} = \frac{\hat{\mathbf{m}}^T\mathbf{x}}{\hat{\mathbf{m}}^T\hat{\mathbf{m}}} \underset{k=2}{\overset{k=1}{\gtrless}} \frac{1}{2}. \tag{22}$$

where $\hat{\mathbf{m}} = \left[\hat{\mathbf{m}}_1, \ldots, \hat{\mathbf{m}}_I\right]^T$, therefore, taking into account (21), the following will be satisfied

$$E\left(\hat{\mathbf{m}}\right) = \mathbf{m}; \quad \mathbf{C}_{\hat{\mathbf{m}}\hat{\mathbf{m}}} = E\left(\hat{\mathbf{m}}\hat{\mathbf{m}}^T\right) = \frac{1}{N}\mathbf{I}_{2Mx2M}. \tag{23}$$

In Appendix VI-B we deduce that the probability of error $P_{e(N)}$ corresponding to early fusion with a training set of size N, is given by the integral

$$P_{e(N)} = \frac{1}{4}\int_0^{\infty}\left(erfc\left(\frac{1}{2\sqrt{2\frac{1}{\hat{\eta}}}}\right) + erfc\left(\frac{2\frac{1}{\nu\hat{\eta}} - 1}{2\sqrt{2\frac{1}{\hat{\eta}}}}\right)\right) f\left(\hat{\eta}\right) d\hat{\eta}$$
$$E\left(\hat{\eta}\right) = \frac{M \times I}{N} + \frac{1}{\nu}; \quad \operatorname{var}\left(\hat{\eta}\right) = 2\left(\frac{M \times I}{N^2} + \frac{2}{N\nu}\right). \tag{24}$$

where $f\left(\hat{\eta}\right)$ is the probability density function of the random variable $\hat{\eta} = \hat{\mathbf{m}}^T\hat{\mathbf{m}} = \frac{1}{\hat{\nu}} = \frac{1}{N}\chi$ where χ is a non-central chi-squared random variable with $M \times I$ degrees of freedom and non-centrality parameter $N\mathbf{m}^T\mathbf{m} = \frac{N}{\nu}$ [27]. We indicate in (24) the mean and variance of $\hat{\eta}$, it is easy to check that $\lim_{N\to\infty} E\left(\hat{\eta}\right) = \frac{1}{\nu}$; $\lim_{N\to\infty} \operatorname{var}\left(\hat{\eta}\right) = 0$, so that $\lim_{N\to\infty} f\left(\hat{\eta}\right) = \delta\left(\hat{\eta} - \frac{1}{\nu}\right) \Rightarrow \lim_{N\to\infty} P_{e(N)} = P_e$.

The extension of the above to each individual modality is straightforward, we only have to consider in (24) $\frac{1}{\nu_i}$ instead of $\frac{1}{\nu}$ and M instead of $M \times I$, i.e.,

$$P_{ei(N)} = \frac{1}{4}\int_0^{\infty}\left(erfc\left(\frac{1}{2\sqrt{2\frac{1}{\hat{\eta}_i}}}\right) + erfc\left(\frac{2\frac{1}{\nu_i\hat{\eta}_i} - 1}{2\sqrt{2\frac{1}{\hat{\eta}_i}}}\right)\right) f\left(\hat{\eta}_i\right) d\hat{\eta}_i$$
$$E\left(\hat{\eta}_i\right) = \frac{M}{N} + \frac{1}{\nu_i}; \quad \operatorname{var}\left(\hat{\eta}_i\right) = 2\left(\frac{M}{N^2} + \frac{2}{N\nu_i}\right) \quad i = 1 \ldots I. \tag{25}$$

Obviously, training consistency is still fulfilled in (25) since again it is $\lim_{N\to\infty} E\left(\hat{\eta}_i\right) = \frac{1}{\nu_i}$; $\lim_{N\to\infty} \operatorname{var}\left(\hat{\eta}_i\right) = 0$ so that $\lim_{N\to\infty} f\left(\hat{\eta}_i\right) = \delta\left(\hat{\eta}_i - \frac{1}{\nu_i}\right) \Rightarrow \lim_{N\to\infty} P_{ei(N)} = P_{ei}$

For the extension to late fusion we can start from (22) since all the previous development would be valid other than considering the statistic $\hat{z}_s = \frac{1}{I}\sum_{i=1}^{I}\hat{z}_i = \frac{1}{I}\sum_{i=1}^{I}\frac{\hat{\mathbf{m}}_i^T\mathbf{x}_i}{\hat{\mathbf{m}}_i^T\hat{\mathbf{m}}_i}$ instead of $\hat{z} = \frac{\hat{\mathbf{m}}^T\mathbf{x}}{\hat{\mathbf{m}}^T\hat{\mathbf{m}}}$. From this consideration, in Appendix VI-B we derive the error probability $P_{es(N)}$ corresponding to late fusion with a training set of size N for each modality, which is given by the multiple integral

$$P_{es(N)} = \frac{1}{4}\int_0^{\infty}\ldots\int_0^{\infty}\left(erfc\left(\frac{1}{2\sqrt{\frac{2}{I^2}\sum_{i=1}^{I}\frac{1}{\hat{\eta}_i}}}\right) + erfc\left(\frac{2\frac{1}{I}\sum_{i=1}^{I}\frac{1}{\nu_i\hat{\eta}_i} - 1}{2\sqrt{\frac{2}{I^2}\sum_{i=1}^{I}\frac{1}{\hat{\eta}_i}}}\right)\right) \times f\left(\hat{\eta}_1\right)\ldots f\left(\hat{\eta}_I\right) d\hat{\eta}_1 \ldots d\hat{\eta}_I$$

TABLE 2. Error probabilities with model parameters estimated from training sets of size *N*.

Method	Probability of error
Modality i	$P_{ei(N)} = \int_0^\infty \left(erfc\left(\frac{1}{2\sqrt{2\frac{1}{\hat{\eta}_i}}} \right) + erfc\left(\frac{2\frac{1}{\nu_i \hat{\eta}_i} - 1}{2\sqrt{2\frac{1}{\hat{\eta}_i}}} \right) \right) f\left(\hat{\eta}_i\right) d\hat{\eta}_i$, see (25) $E\left(\hat{\eta}_i\right) = \frac{M}{N} + \frac{1}{\nu_i}$; $\mathrm{var}\left(\hat{\eta}_i\right) = 2\left(\frac{M}{N^2} + \frac{2}{N\nu_i} \right)$
Early fusion	$P_{e(N)} = \frac{1}{4}\int_0^\infty \left(erfc\left(\frac{1}{2\sqrt{2\frac{1}{\hat{\eta}}}} \right) + erfc\left(\frac{2\frac{1}{\nu\hat{\eta}} - 1}{2\sqrt{2\frac{1}{\hat{\eta}}}} \right) \right) f\left(\hat{\eta}\right) d\hat{\eta}$, see (24) $E\left(\hat{\eta}\right) = \frac{M \times I}{N} + \frac{1}{\nu}$; $\mathrm{var}\left(\hat{\eta}\right) = 2\left(\frac{M \times I}{N^2} + \frac{2}{N\nu} \right)$
Late fusion	$P_{es(N)} = \frac{1}{4}\int_0^\infty \ldots \int_0^\infty \left(erfc\left(\frac{1}{2\sqrt{\frac{2}{I^2}\sum_{i=1}^{I}\frac{1}{\hat{\eta}_i}}} \right) + erfc\left(\frac{2\frac{1}{I}\sum_{i=1}^{I}\frac{1}{\nu_i\hat{\eta}_i} - 1}{2\sqrt{\frac{2}{I^2}\sum_{i=1}^{I}\frac{1}{\hat{\eta}_i}}} \right) \right) f\left(\hat{\eta}_1\right) \ldots f\left(\hat{\eta}_I\right) d\hat{\eta}_1 \ldots d\hat{\eta}_I$,see (26) $E\left(\hat{\eta}_i\right) = \frac{M}{N} + \frac{1}{\nu_i}$ $\mathrm{var}\left(\hat{\eta}_i\right) = 2\left(\frac{M}{N^2} + \frac{2}{N\nu_i} \right)$ $i = 1 \ldots I$

$$E\left(\hat{\eta}_i\right) = \frac{M}{N} + \frac{1}{\nu_i} \quad \mathrm{var}\left(\hat{\eta}_i\right) = 2\left(\frac{M}{N^2} + \frac{2}{N\nu_i}\right) \quad i = 1 \ldots I. \tag{26}$$

Notice that $\lim_{N\to\infty} f\left(\hat{\eta}_1\right) \ldots f\left(\hat{\eta}_2\right) = \delta\left(\hat{\eta}_1 - \frac{1}{\nu_1}\right) \ldots \delta\left(\hat{\eta}_I - \frac{1}{\nu_I}\right)$, hence $\lim_{N\to\infty} P_{es(N)} = P_{es}$.

Table 2 summarizes the error probabilities with estimated model parameter. The equations in Table 2 allow us to assess the effects on the error probabilities of finite training set sizes, with respect to the ideal situation of exact model knowledge, embodied in the equations of Table 1.

In Fig. 4 we present some experiments that illustrate these results. As in Fig. 3, we show the curves of probability of error for the case of two modalities $I = 2$, assuming that $\nu_2 = 0.2$ is fixed and ν_1 varies between 0 and 1. They have been obtained by numerical integration of the equations in Table 2 and are compared with the curves corresponding to knowledge of the model in Fig. 3. In Fig. 4 we have considered four pairs of values M and N.

It can be seen that there is always an increase in the error probability when the model parameters have to be estimated (dotted lines) with respect to the same case with known model parameters (solid lines). This increase is smaller the larger N is for a given M(the closer we are to convergence to the ideal values). It can also be observed that the results of Fig. 4a ($M = 10, N = 20$) are similar to that of Fig. 4c ($M = 25, N = 50$), and the same with respect to Fig. 4b ($M = 10, N = 50$) and Fig. 4d ($M = 25, N = 125$). This is consistent with the equations of Table 2, notice that in all three cases, convergence of $f(\cdot)$ towards a delta function approximately depends on the quotient M/N. It is also significant that despite the fact that in early fusion (equation (24)) M is multiplied by the number of modalities, the degradation of early fusion for a given N is quite similar to the degradation of the other cases.

This can be explained by noting that the separability $\frac{1}{\nu} = \sum_{i=1}^{I} \frac{1}{\nu_i}$ is the arithmetic mean of the individual separabilities also multiplied by I. Let us assume for simplicity that the individual separabilities are equal, so that $\frac{1}{\nu} = I\frac{1}{\nu_i}$. Furthermore the convergences $P_{e(N)} \underset{N\to\infty}{\to} P_e$ $P_{ei(N)} \underset{N\to\infty}{\to} P_{ei}$ $P_{es(N)} \underset{N\to\infty}{\to} P_{es}$ are equivalent to the convergences of the random variables $\hat{\eta} \underset{N\to\infty}{\to} \frac{1}{\nu}$ $\hat{\eta}_i \underset{N\to\infty}{\to} \frac{1}{\nu_i}$, so for a fair comparison of these convergences we should consider normalized means and variances in (24), (25) and (26)

$$\frac{E\left(\hat{\eta}\right)}{1/\nu} = \frac{\frac{M\times I}{N} + \frac{1}{\nu}}{1/\nu} = \frac{\frac{M}{N}}{1/\nu_i} + 1 = \frac{E\left(\hat{\eta}_i\right)}{1/\nu_i}$$
$$\frac{\mathrm{var}\left(\hat{\eta}\right)}{1/\nu} = \frac{2\left(\frac{M\times I}{N^2} + \frac{2}{N\nu}\right)}{1/\nu} = \frac{2\frac{M}{N^2}}{1/\nu_i} + \frac{2}{N} = \frac{\mathrm{var}\left(\hat{\eta}_i\right)}{1/\nu_i}. \tag{27}$$

We see that normalized means and variances are the same in early fusion, late fusion and for every separate modality. This can be explained from a different, perspective. On the one hand, in the analyzed case, early and late fusion require the same number of estimated parameters ($M \times I$). On the other hand, no additional parameters are required in the early

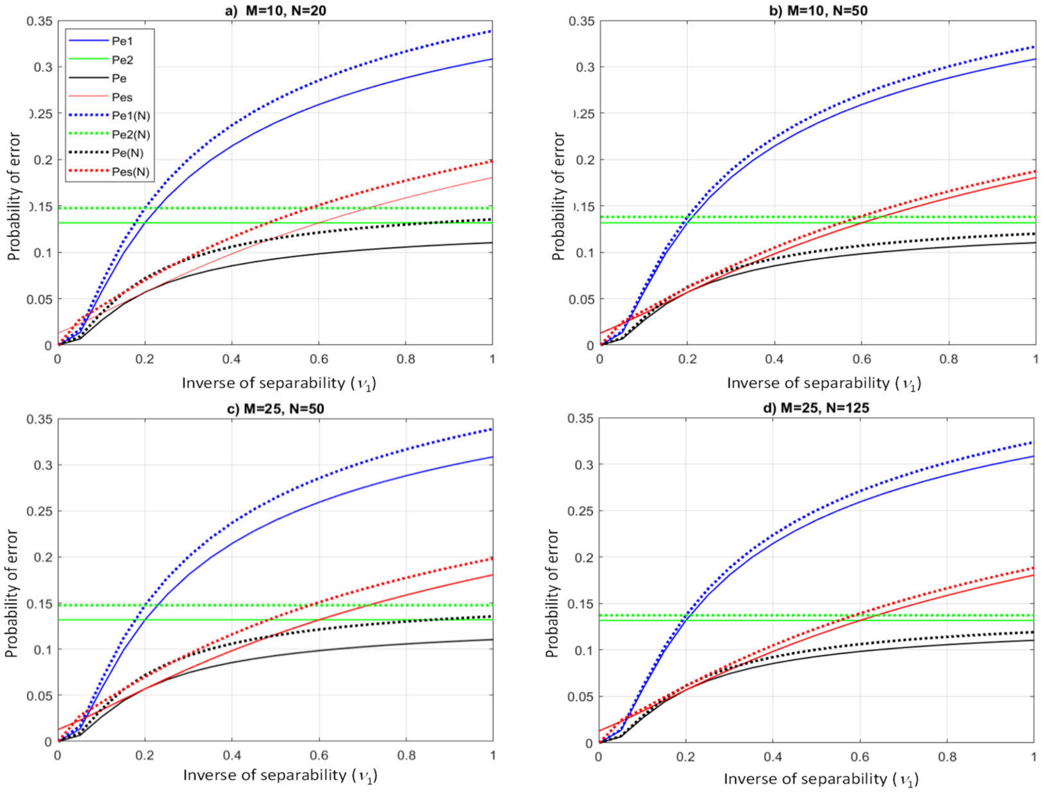

FIGURE 4. Error probability with known model (solid) and estimated (dotted) for different M and N, $I = 2\nu_2 = 0.2$ and ν_1 varying between 0 and 1: Modality 1, Modality 2, Early fusion, Late fusion.

fusion model apart from the union of all of each modality. Thus the finite sample size effects of the training set for a given dimension M affects the same to all the options and the comparative analysis of Section III-A is still valid. In the next section we are going to consider the correlated case where covariance matrices are to be estimated. Then different amounts of parameters will be required in early and late fusion, and the parameters of every separate modality will not define the complete set of parameters estimates required for early fusion (estimates of the correlation among modalities will be necessary).

Before going to Section IV, let us make some verification of the approximations made in the derivation of (24) and (26) in Appendices VI-B and VI-B. We have resorted to Monte Carlo simulation to validate these approximations. For this purpose, we have generated sets of N independent training vectors with multivariate Gaussian distributions for each modality ($I = 2$) and each class. According to model (8), in class 2, the vectors have mean $\mathbf{0}_M$ in both modalities, and in class 1 they have mean $\mathbf{m}_1$ for modality 1, and $\mathbf{m}_2$ for modality 2. The centroids $\mathbf{m}_1$ and $\mathbf{m}_2$ are randomly generated from multivariate Gaussian distributions of mean $\mathbf{0}_M$ and covariance $\mathbf{C}_{\mathbf{m}_i\mathbf{m}_i} = \mathbf{I}_{MxM}, i = 1, 2$ and are adjusted in their Euclidean norm to achieve the desired separability values $\frac{1}{\nu_1}$ and $\frac{1}{\nu_2}$. The covariance matrices generated in both modalities are $\mathbf{C}_i^{(k)} = \mathbf{I}_{MxM}, i = 1, 2\ k = 1, 2$. Additionally, 50 test vectors are generated for each modality and each class under the same conditions as those defined for the training vectors. We apply the different methods trained with the training sets ((20)) and estimate the probability of error as the percentage of misclassification on the 50 test vectors. Finally, we repeat the above process 500 times and average the error probability estimates obtained in each round to achieve stable results. Fig. 5a and Fig. 5b show respectively the error probability curves of Fig. 4a and Fig. 4b obtained from the equations of Table 2 (dotted lines), superimposed with those obtained by simulation (solid lines). It can be seen that they are quite close.

IV. EXTENDING THE ANALYSIS TO DATA WITH ARBITRARY COVARIANCE MATRICES

A. KNOWN MODEL PARAMETERS

Let us now complicate the Gaussian model so far considered, assuming that the elements of the feature vectors present arbitrary correlations, although equal in each class. Thus, in the model (8) corresponding to the early fusion we will assume that $\mathbf{C}^{(1)} = \mathbf{C}^{(2)} = \mathbf{C}$ and therefore

$$p\left(\mathbf{x}/k=1\right) = \frac{1}{(2\pi)^M}\exp\left(-\frac{1}{2}\left(\mathbf{x}-\mathbf{m}\right)^T\mathbf{C}^{-1}\left(\mathbf{x}-\mathbf{m}\right)\right)$$
$$p\left(\mathbf{x}/k=2\right) = \frac{1}{(2\pi)^M}\exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{C}^{-1}\mathbf{x}\right). \quad (28)$$

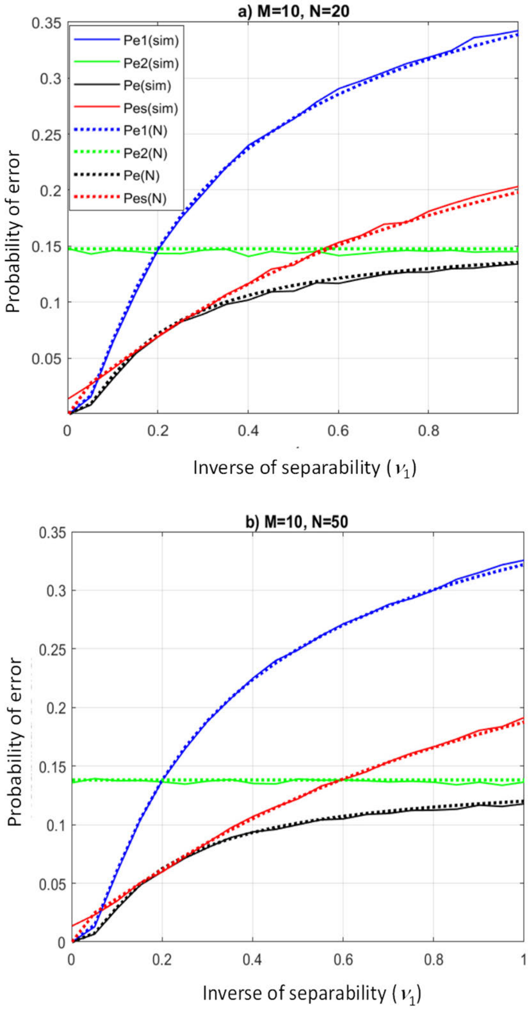

FIGURE 5. Error probability with estimated model from the equations of Table 2 (dotted) and by simulation (solid) for different M and N, $I = 2\nu_2 = 0.2$ and ν_1 varying between 0 and 1. Modality 1, Modality 2, Early fusion, Late fusion.

And the optimal test in this case will be

$$z_C = \frac{\mathbf{m}^T\mathbf{C}^{-1}\mathbf{x}}{\mathbf{m}^T\mathbf{C}^{-1}\mathbf{m}} \overset{k=1}{\underset{k=2}{\gtrless}} \frac{1}{2}. \tag{29}$$

And for each modality we will assume $\mathbf{C}_i^{(1)} = \mathbf{C}_i^{(2)} == \mathbf{C}_i$ $i = 1 \ldots I$, so the optimal test for each modality separately will be

$$z_{iC} = \frac{\mathbf{m}_i^T\mathbf{C}_i^{-1}\mathbf{x}_i}{\mathbf{m}_i^T\mathbf{C}_i^{-1}\mathbf{m}_i^T} \overset{k=1}{\underset{k=2}{\gtrless}} \frac{1}{2} \quad i = 1, 2. \tag{30}$$

Note that $\mathbf{C} = \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} & \ldots & \mathbf{C}_{1I} \\ \mathbf{C}_{21} & \mathbf{C}_{22} & \ldots & \mathbf{C}_{1I} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{C}_{I1} & \mathbf{C}_{I2} & \ldots & \mathbf{C}_{II} \end{bmatrix}$, being $\mathbf{C}_{ii} = \mathbf{C}_i$ and $\mathbf{C}_{ij}$ $i \neq j$ the cross-covariance between the feature vectors of each pair of different modalities. Let us first consider that the model is known. We can directly exploit the results of the uncorrelated case by using the prewhitened vectors $\mathbf{x}_{iC} = \mathbf{C}_i^{-\frac{1}{2}}\mathbf{x}_i$, $\mathbf{m}_{iC} = \mathbf{C}_i^{-\frac{1}{2}}\mathbf{m}_i$, $\mathbf{x}_C = \mathbf{C}^{-\frac{1}{2}}\mathbf{x}$, $\mathbf{m}_C = \mathbf{C}^{-\frac{1}{2}}\mathbf{m}$, so that we can write (29), (30) in the form

$$z_C = \frac{\mathbf{m}_C^T\mathbf{x}_C}{\mathbf{m}_C^T\mathbf{m}_C} \overset{k=1}{\underset{k=2}{\gtrless}} \frac{1}{2}; \quad z_{Ci} = \frac{\mathbf{m}_{Ci}^T\mathbf{x}_{Ci}}{\mathbf{m}_{Ci}^T\mathbf{m}_{Ci}} \overset{k=1}{\underset{k=2}{\gtrless}} \frac{1}{2} \quad i = 1 \ldots I. \tag{31}$$

Therefore, we can apply the same analysis from (12) to (19), simply by properly defining the separabilities from the prewhitened centroids.

$$\begin{aligned} \nu_C &= \frac{1}{\mathbf{m}_C^T\mathbf{m}_C} = \frac{1}{\mathbf{m}^T\mathbf{C}^{-1}\mathbf{m}} \\ \nu_{iC} &= \frac{1}{\mathbf{m}_{iC}^T\mathbf{m}_{iC}} = \frac{1}{\mathbf{m}_i^T\mathbf{C}_i^{-1}\mathbf{m}_i} \quad i = 1 \ldots I \\ \nu_{sC} &= \frac{1}{I^2}\sum_{i=1}^{I}\nu_{iC}. \end{aligned} \tag{32}$$

And the equations in Table 1 would be applicable with the modified parameters, i.e.,

$$\begin{aligned} P_{eC} &= \frac{1}{2}erfc\left(\frac{1}{2\sqrt{2\nu_C}}\right); \\ P_{eiC} &= \frac{1}{2}erfc\left(\frac{1}{2\sqrt{2\nu_{iC}}}\right) \quad i = 1 \ldots I \\ P_{esC} &= \frac{1}{2}erfc\left(\frac{1}{2\sqrt{2\nu_{sC}}}\right). \end{aligned} \tag{33}$$

Notice that if $\mathbf{C}_{ij} = \mathbf{0}_{M\times M}$ $i \neq j$ then $\mathbf{C}^{-1} = \begin{bmatrix} \mathbf{C}_1^{-1} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_2^{-1} & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{C}_I^{-1} \end{bmatrix}$ and $\nu_C = \frac{1}{\mathbf{m}^T\mathbf{C}^{-1}\mathbf{m}} = \frac{1}{\sum_{i=1}^{I}\mathbf{m}_i^T\mathbf{C}_i^{-1}\mathbf{m}_i} = \frac{1}{\sum_{i=1}^{I}\frac{1}{\nu_{iC}}}$ is still the harmonic mean of $\nu_1, \ldots, \nu_I$ divided by I. Therefore all the properties indicated in the third column of Table 1 are still valid. However this will not be true in general because the particular covariances between every pair of modalities would affect the comparative performance of early fusion with respect to late fusion or to every separate modality.

B. ESTIMATED MODEL PARAMETERS

To maintain homogeneity with the previous approach, we will continue to assume that we have N training vectors of class 1 in every modality $\mathbf{x}_{im}^{(1)}$ $m = 1 \ldots N$ $i = 1 \ldots I$. From these vectors we perform maximum likelihood estimates of both the mean vectors and the covariance matrices that we will substitute in (29) and (30)

$$\hat{\mathbf{m}}_i = \frac{1}{N}\sum_{m=1}^{N}\mathbf{x}_{im}^{(1)}$$

$$\begin{aligned}
\hat{\mathbf{C}}_i &= \frac{1}{N}\sum_{m=1}^{N}\left(\mathbf{x}_{im}^{(1)} - \hat{\mathbf{m}}_i\right)\left(\mathbf{x}_{im}^{(1)} - \hat{\mathbf{m}}_i\right)^T \; i = 1\ldots I \\
\hat{\mathbf{m}} &= \left[\hat{\mathbf{m}}_1, \hat{\mathbf{m}}_2\right]^T \\
\mathbf{x}_m^{(1)} &= \left[\mathbf{x}_{1m}^{(1)}, \mathbf{x}_{2m}^{(1)}\right]^T \quad \hat{\mathbf{C}} = \frac{1}{N}\sum_{m=1}^{N}\left(\mathbf{x}_m^{(1)} - \hat{\mathbf{m}}\right)\left(\mathbf{x}_m^{(1)} - \hat{\mathbf{m}}\right)^T.
\end{aligned} \tag{34}$$

We will follow a similar procedure as in Section III-B. Starting with early fusion, we have deduced in Appendix VI-B that the error probability with estimated covariance matrices is given by

$$\begin{aligned}
P_{eC(N)} &= \frac{1}{4}\int_0^{\infty}\left(erfc\left(\frac{1}{2\sqrt{2\frac{1}{\hat{\eta}_C}}}\right)\right. \\
&\qquad \left. + erfc\left(\frac{2\frac{1}{\nu_C\hat{\eta}_C} - 1}{2\sqrt{2\frac{1}{\hat{\eta}_C}}}\right)\right) f\left(\hat{\eta}_C\right) d\hat{\eta}_C \\
E\left(\hat{\eta}_C\right) &= \frac{1}{c_{M\times I,N}}\left(\frac{M\times I}{N} + \frac{1}{\nu_C}\right) \quad N > M\times I + 2 \\
\mathrm{var}\left(\hat{\eta}_C\right) &= \frac{1}{c_{M\times I,N}^2} 2\left(\frac{M\times I}{N^2} + \frac{2}{N\nu_C}\right) \\
&\quad \times\left(1 + \frac{2}{N - M\times I - 4}\right) \\
&\quad + E^2\left(\hat{\eta}_C\right)\frac{2}{N - M\times I - 4} \quad N > M\times I + 4.
\end{aligned} \tag{35}$$

where $\hat{\nu}_C = \frac{1}{\hat{\mathbf{m}}_C^T\hat{\mathbf{m}}_C} = \frac{1}{\hat{\mathbf{m}}^T\hat{\mathbf{C}}^{-1}\hat{\mathbf{m}}} = \frac{1}{\hat{\eta}_C}$, $c_{M\times I,N} = \frac{N-M\times I-2}{N-1}$ and again $f\left(\cdot\right)$ denotes probability density. Unfortunately, in this case, unlike in (24), it is not possible to know $f\left(\hat{\eta}_C\right)$ due to the randomness of the covariance matrix in the definition $\hat{\eta}_C = \hat{\mathbf{m}}^T\hat{\mathbf{C}}^{-1}\hat{\mathbf{m}}$. However, we have been able to calculate, as expressed in (35), the mean and variance of $\hat{\eta}_C$. Firstly it is observed in (35) that $\lim_{N\to\infty} E\left(\hat{\eta}_C\right) = \frac{1}{\nu_C}$; $\lim_{N\to\infty}\mathrm{var}\left(\hat{\eta}_C\right) = 0 \Rightarrow \lim_{N\to\infty} f\left(\hat{\eta}_C\right) = \delta\left(\hat{\eta}_C - \frac{1}{\nu_C}\right)$ so that $\lim_{N\to\infty} P_{eC(N)} = P_{eC}$. On the other hand (35) allows us to understand the effects of estimating the covariance matrix with respect to the case of only estimating the centroids. Thus, comparing $E\left(\hat{\eta}_C\right)$ in (35) with $E\left(\hat{\eta}_C\right)$ in (24) we can observe the presence of the factor $\frac{1}{c_{M\times I,N}}$ being $0 \leq c_{M\times I,N} \leq 1$; $\lim_{N\to\infty} c_{M\times I,N} = 1$. The inverse of this factor increases $\left(\frac{M\times I}{N} + \frac{1}{\nu_C}\right)$ delaying the convergence of $E\left(\hat{\eta}_C\right)$ towards $\frac{1}{\nu_C}$ as we increase the size of the training set. Moreover, comparing $\mathrm{var}\left(\hat{\eta}_C\right)$ in (35) with $\mathrm{var}\left(\hat{\eta}\right)$ in (24), we also observe an increasing effect of $2\left(\frac{M\times I}{N^2} + \frac{2}{N\nu_C}\right)$ due in part to the factor $\frac{1}{c_{M\times I,N}^2}$ as well as the term $E^2\left(\hat{\eta}_C\right)\frac{2}{N-M\times I-4}$(whose convergence is in turn affected by $\frac{1}{c_{M\times I,N}^2}$). So, we will call $c_{M\times I,N}$ the *convergence factor*. All this suggests that we can use the *convergence factor* as a criterion for determining the appropriate size of the training set. By solving for N in the definition of the convergence factor we can write

$$N = \frac{M\times I + 2 - c_{M\times I,N}}{1 - c_{M\times I,N}}. \tag{36}$$

Thus if we wish to approximate the performance of early fusion with known model we must choose a value $c_{M\times I,N}$ close to 1. For example, for $I = 2$, $M = 10$ and $c_{2M,N} = 0.9$, the required size of the training set is $N = 211$.

We can take advantage of (35) to directly derive the rest of the error probabilities. Thus $P_{eiC(N)}$ $i = 1\ldots I$ can be obtained from (35), replacing ν_C by ν_{iC}, $\hat{\eta}_C$ by $\hat{\eta}_{iC} = \hat{\mathbf{m}}_i^T\hat{\mathbf{C}}_i^{-1}\hat{\mathbf{m}}_i$ and $M\times I$ by M. Likewise, considering these changes and starting from (26), it is immediate to obtain the expression of the error probability $P_{esC(N)}$ for the late fusion with estimation of the covariance matrix.

Note that both in the case of individual modality and in the case of the late fusion the *convergence factor* to be considered is $c_{M,N} = \frac{N-M-2}{N-1}$, so that solving for N as in (36)

$$N = \frac{M + 2 - c_{M,N}}{1 - c_{M,N}} \quad i = 1, 2. \tag{37}$$

Hence if we want the same value $c_{M,N} = 0.9$ as above, being $I = 2; M = 10$, the required size of the training set for each modality is 111. However, we have seen that for the late fusion, 211 feature vectors are needed, which are obtained from 211 vectors for each modality. This confirm the result of (6) in Section II, where we have deduced than in the multivariate Gaussian model, the number of parameters to estimate in early fusion was slightly less than twice the required for late fusion. This simple example illustrates the problem of the increased dimensionality of early fusion when the model has to be estimated.

For further verification of the above, we present some illustrative results in Fig. 6. The curves have been obtained by simulation in the way described at the end of Section III to obtain Fig. 5, but considering covariance matrices whose generic element is $\mathbf{C}_i\left(m, n\right) = r^{m-n}$ $i = 1, 2$. Additionally, we have assumed uncorrelation between the feature vectors of the two modalities, i.e. $\mathbf{C}_{ij} = \mathbf{0}_{M\times M}$ $i \neq j$, although we will not consider this knowledge available in the training, thus estimating the whole covariance matrix $\mathbf{C}$, as well as the centroids from (34). We observe in Fig. 6a, that early fusion gets substantially worse with respect to previous cases, being practically always worse than late fusion, and being the worst option in the interval from $\nu_1 = 0.05$ to $\nu_1 = 0.35$. This is due to the fact that the *convergence factor* $c_{2\times 10,24} = 0.08$ is very close to zero, indicating that the value $N = 24$ is far from adequate to achieve the probability of error with known model in early fusion. The corresponding value for the rest of the methods is $c_{10,24} = 0.52$ which, although clearly improvable, is substantially higher than 0.08. In Fig. 6b we increase N

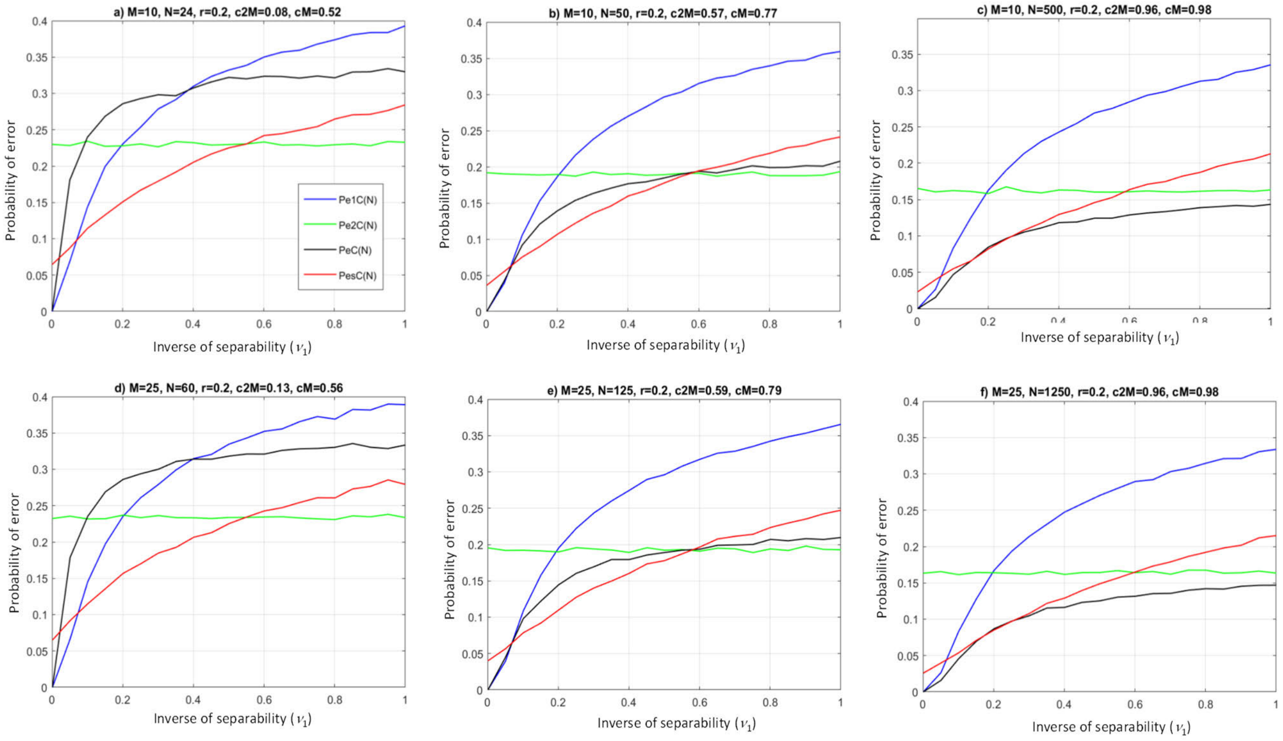

FIGURE 6. **Error probability with estimated model assuming correlation obtained by simulation for different values of M and N. Where $r = 0.2$, $\nu_2 = 0.2$ and ν_1 varying between 0 and 1. Modality 1, Modality 2, Early fusion, Late fusion.**

to 50, there is a notable improvement of the early fusion which is consistent with the increase of the *convergence factor* $c_{2\times10,50} = 0.57$. Despite this, it is still a worse choice than late fusion in the interval from $\nu_1 = 0.05$ to $\nu_1 = 0.55$, as the factor $c_{10,50} = 0.77$ has also improved. In Fig. 6c we increase N to 500, which results in $c_{2\times10,500} = 0.96$ and $c_{2\times10,500} = 0.98$, indicating that all methods have practically reached their maximum performance (known model) and early fusion is the best choice in all cases. In Fig. 6d, 6e and 6f we have increased M to 25. We have also increased N proportionally according to the values respectively considered in Fig. 6a, 6b and 6c. We note that the result obtained in each upper figure is comparable with that of the corresponding lower figure, which is consistent with the similar values of $c_{2M,N}$ and $c_{M,N}$ in both figures. This is due to the fact that for large N both $c_{2M,N} \simeq 1 - 2\left(M/N\right)$ and $c_{M,N} \simeq 1 - \left(M/N\right)$ are practically dependent on the ratio M/N.

V. DISCUSSION

We claimed in Section II that assuming perfect knowledge of the data model, early fusion will always be better or, in the worst case, equal to late fusion. This was made in terms of minimizing the classifier error probability by comparing the posterior probabilities of each class conditional to the extracted features. This claim is valid for any class-conditional distributions of the data. However, in a practical setting where the models are to be learned from training data, late fusion could eventually be a better option than early fusion. This is due to the dimensionality increase of the feature vectors in early fusion, which directly affects the training set size. Actually we have showed that an exponential increase in the training set size is required for early fusion in nonparametric approaches ((5)). In parametric models, the required increase depends mainly on the number of parameters to be estimated as we have illustrated in (6) for the multivariate Gaussian model.

In a general setting, a mathematical analysis comparing early and late fusion is unapproachable. On the data side, we can find different distribution models. On the side of the classifiers, there are a myriad of options. Both concepts, data models and classifiers, could even vary from one modality to another. Thus, in general, it is not a simple matter to obtain the Bayes error rates nor the corresponding probabilities of error for models estimated from a finite training set. That is why we have focused in Sections III and IV on the multivariate Gaussian model. Despite its specificity, the analysis confirms the general considerations of Section II. It should also serve the purpose of having some quantitative criterion to determine if the size of the training set is adequate for a given dimension. Even for non-Gaussian models or different type of classifiers, it may be a tentative criterion at least.

Just to illustrate this general interest we have performed a similar experiment to that of Fig. 6, but considering a non-Gaussian data model. Thus, it is assumed a mixture of two equiprobable multivariate Gaussian probability densities in

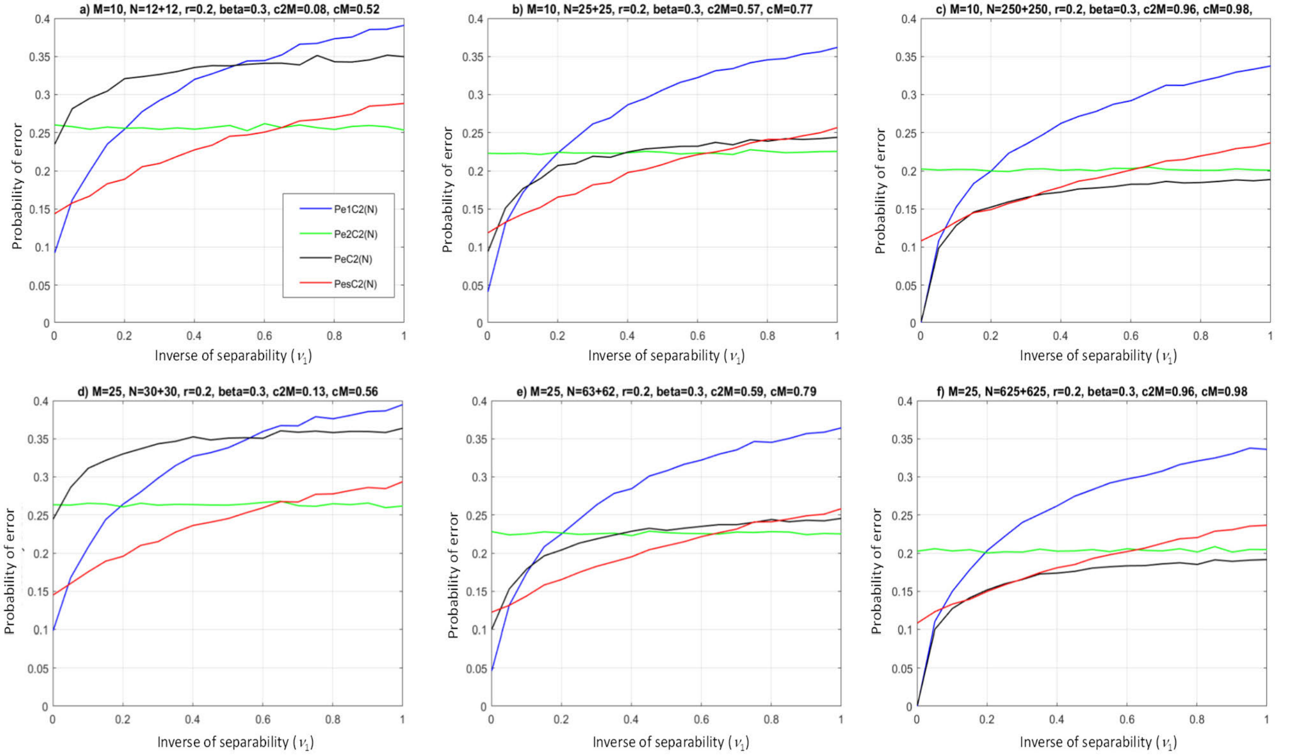

FIGURE 7. Error probability with estimated model assuming a mixture of two Gaussians for every modality and class, obtained by simulation for different values of M and N. Where $r = 0.2$, $\beta = 0.3$, $\nu_2 = 0.2$ and ν_1 varying between 0 and 1. Modality 1, Modality 2, Early fusion, Late fusion.

every class and modality.

$$\left.\begin{aligned} \mathbf{x}_i^{(k)} &\sim \frac{1}{2}N\left(\mathbf{m}_i^{(k)}\left(1+\beta_i^{(k)}\right), \mathbf{C}_{i1}^{(k)}\right) \\ &\quad + \frac{1}{2}N\left(\mathbf{m}_i^{(k)}\left(1-\beta_i^{(k)}\right), \mathbf{C}_{i2}^{(k)}\right) \\ \mathbf{m}_i^{(1)} &= \mathbf{m}_i \;\; \|\mathbf{m}_i\|^2 = \mathbf{m}_i^T\mathbf{m}_i \neq 0, \\ \mathbf{m}_i^{(2)} &= \mathbf{0}_M = \left[\underbrace{0\ldots0}_{M}\right]^T \end{aligned}\right\} \begin{aligned} i &= 1\ldots I \\ k &= 1,2 \end{aligned}. \tag{38}$$

where $\beta_i^{(k)}$ indicates a symmetric shift of the class centroid to define the respective location of the mean of every Gaussian mixture component. Notice that the corresponding model for the early fused vectors $\mathbf{x}^{(k)} = \left[\mathbf{x}_1^{(k)} \ldots \mathbf{x}_I^{(k)}\right]$ cannot be clearly defined from (38).

In Fig. 7 we present some illustrative results which can be directly compared with those of Fig. 6. The case $I = 2$ has been considered. Again, the curves have been obtained by simulation in the way described at the end of Section III. The training instances have been most equally divided between both components of the Gaussian mixtures so that the total training sizes of Fig. 6 are maintained. For simplicity we have considered

$$\left.\begin{aligned} \beta_i^{(k)} &= \beta \\ \mathbf{C}_{i1}^{(k)}(m,n) &= \mathbf{C}_{i2}^{(k)}(m,n) = r^{m-n} \end{aligned}\right\} \quad i = 1,2\; k = 1,2. \tag{39}$$

The detectors (29) and (30), with the estimates (34) have been implemented. As these detectors are optimum for the multivariate Gaussian case, a general increase of the error probability is observed when comparing each subfigure in Fig. 7 with the corresponding subfigure in Fig. 6. However the relative comparison among methods leads to similar conclusions to those from Fig. 6.

VI. CONCLUSION

The comparative mathematical analysis carried out, together with the experimental verifications, allow us to reach the conclusions indicated below:

A. WITH KNOWLEDGE OF THE MODEL PARAMETERS

Early fusion is always the best option. If early fusion is not possible due to lack of access to feature vectors, late fusion is an option that can improve the performance of all modalities acting separately. For this later to be true, certain conditions must be met for the separabilities of each of them. In the case studied, when using the mean as the fusing function, the separabilities must be comparable. This makes sense for a late fusion that gives more weight to the modalities with higher separability. Therefore, we can consider learning optimal fusers from training data (α-integration is an example [31]).

B. WITH ESTIMATED MODEL PARAMETERS

In practice, the model parameters have to be estimated, which leads to performance losses for all the methods analysed.

We have found that avoiding such losses involves large (ideally infinite) training sets. If the training set sizes are limited but the number of parameters to be estimated is the same in both early and late fusion, the superiority of early fusion still holds. That is, both options suffer a comparable degradation for a given training set size. But if the training set sizes are limited and the number of parameters to be estimated is greater in early fusion, its degradation could be higher. In other words, we have verified the problem of increased dimensionality of early fusion.

We have proposed a measure to quantify whether a training set size is sufficiently large, which we have called the *convergence factor*. This factor varies between 0 and 1, with a value close to 1 indicating that the performance will be close to that corresponding to knowledge of the model parameters (Bayes error rate). If we have to choose between early or late fusion and we have fixed the size N of the training set, we can calculate the *convergence factors* for early (feature vectors of dimension $M \times I$) and late fusion (feature vectors of dimension M). If the later is significantly higher than the former, we should choose this option.

APPENDIX A
DERIVATION OF (24)

To analyse the effect of working with parameter estimates, we will start by defining a binary random variable $\hat{d}$ associated to the early fusion rule (22) (notation congruent with that used in Section I), which will be 1 if $\hat{z} > 0.5$ and 0 if $\hat{z} < 0.5$. The expected value of this variable conditional on each class shall be

$$\begin{aligned} E\left[\hat{d}\Big/k=1\right] &= \Pr\left(\hat{d}=1\Big/k=1\right) \\ &= \Pr\left(\hat{z}>0.5/k=1\right) \\ &= 1-\Pr\left(\hat{z}<0.5/k=1\right) \\ E\left[\hat{d}\Big/k=2\right] &= \Pr\left(\hat{d}=1\Big/k=2\right) \\ &= \Pr\left(\hat{z}>0.5/k=2\right). \end{aligned} \tag{A1}$$

We will call $P_{e(N)}$ the error probability corresponding to the test (22), as it will depend on N according to (20). But

$$P_{e(N)} = \frac{1}{2}\Pr\left(\hat{z}>\frac{1}{2}\Big/k=2\right)+\frac{1}{2}\Pr\left(\hat{z}<\frac{1}{2}\Big/k=1\right). \tag{A2}$$

Considering (A1) we can write

$$P_{e(N)} = \frac{1}{2}E\left[\hat{d}\Big/k=2\right]+\frac{1}{2}\left(1-E\left[\hat{d}\Big/k=1\right]\right). \tag{A3}$$

But the random variable $\hat{d}$ depends on the random variable $\hat{z} = \frac{\hat{\mathbf{m}}^T\mathbf{x}}{\hat{\mathbf{m}}^T\hat{\mathbf{m}}}$ which in turn depends on two multivariate random variables $\hat{\mathbf{m}}$ and $\mathbf{x}$. We can therefore apply the law of total expectation in (A3), by first averaging over $\mathbf{x}$ conditional on $\hat{\mathbf{m}}$, and then averaging over $\hat{\mathbf{m}}$, i.e.,

$$\begin{aligned} P_{e(N)} = E_{\hat{\mathbf{m}}}\Bigg(&\frac{1}{2}E_{\mathbf{x}}\left(\hat{d}\Big/k=2,\hat{\mathbf{m}}\right) \\ &+\frac{1}{2}\left(1-E_{\mathbf{x}}\left(\hat{d}\Big/k=1,\hat{\mathbf{m}}\right)\right)\Bigg). \end{aligned} \tag{A4}$$

But

$$\begin{aligned} E_{\mathbf{x}}\left(\hat{d}\Big/k=2,\hat{\mathbf{m}}\right) &= \Pr\left(\hat{d}=1\Big/k=2,\hat{\mathbf{m}}\right) \\ &= \Pr\left(\hat{z}>\frac{1}{2}\Big/k=2,\hat{\mathbf{m}}\right) \\ E_{\mathbf{x}}\left(\hat{d}\Big/k=1,\hat{\mathbf{m}}\right) &= \Pr\left(\hat{d}=1\Big/k=1,\hat{\mathbf{m}}\right) \\ &= \Pr\left(\hat{z}>\frac{1}{2}\Big/k=1,\hat{\mathbf{m}}\right). \end{aligned} \tag{A5}$$

Being

$$\begin{aligned} &E^{(1)}\left(\hat{z}/\hat{\mathbf{m}}\right) \\ &\quad= \frac{\hat{\mathbf{m}}^T E^{(1)}\left(\mathbf{x}\right)}{\hat{\mathbf{m}}^T\hat{\mathbf{m}}} = \frac{\hat{\mathbf{m}}^T\mathbf{m}}{\hat{\mathbf{m}}^T\hat{\mathbf{m}}} \\ &\mathrm{var}^{(1)}\left(\hat{z}/\hat{\mathbf{m}}\right) \\ &\quad= E^{(1)}\left(\hat{z}^2\Big/\hat{\mathbf{m}}\right)-E^{(1)^2}\left(\hat{z}/\hat{\mathbf{m}}\right) \\ &\quad= \frac{1}{\left(\hat{\mathbf{m}}^T\hat{\mathbf{m}}\right)^2}E^{(1)}\left(\hat{\mathbf{m}}^T\mathbf{x}\mathbf{x}^T\hat{\mathbf{m}}\right)-\left(\frac{\hat{\mathbf{m}}^T\mathbf{m}}{\hat{\mathbf{m}}^T\hat{\mathbf{m}}}\right)^2 \\ &\quad= \frac{1}{\left(\hat{\mathbf{m}}^T\hat{\mathbf{m}}\right)^2}\left(\hat{\mathbf{m}}^T\underbrace{E^{(1)}\left(\mathbf{x}\mathbf{x}^T\right)}_{\mathbf{I}_{2Mx2M}+\mathbf{m}\mathbf{m}^T}\hat{\mathbf{m}}\right)-\left(\frac{\hat{\mathbf{m}}^T\mathbf{m}}{\hat{\mathbf{m}}^T\hat{\mathbf{m}}}\right)^2 \\ &\quad= \frac{1}{\hat{\mathbf{m}}^T\hat{\mathbf{m}}} \\ &E^{(2)}\left(\hat{z}/\hat{\mathbf{m}}\right) \\ &\quad= \frac{\hat{\mathbf{m}}^T E^{(2)}\left(\mathbf{x}\right)}{\hat{\mathbf{m}}^T\hat{\mathbf{m}}} = 0; \\ &\mathrm{var}^{(2)}\left(\hat{z}/\hat{\mathbf{m}}\right) \\ &\quad= E^{(2)}\left(\hat{z}^2\Big/\hat{\mathbf{m}}\right)-E^{(2)^2}\left(\hat{z}/\hat{\mathbf{m}}\right) \\ &\quad= \frac{1}{\left(\hat{\mathbf{m}}^T\hat{\mathbf{m}}\right)^2}E^{(2)}\left(\hat{\mathbf{m}}^T\mathbf{x}\mathbf{x}^T\hat{\mathbf{m}}\right) \\ &\quad= \frac{1}{\left(\hat{\mathbf{m}}^T\hat{\mathbf{m}}\right)^2}\left(\hat{\mathbf{m}}^T\underbrace{E^{(2)}\left(\mathbf{x}\mathbf{x}^T\right)}_{\mathbf{I}_{2Mx2M}}\hat{\mathbf{m}}\right) = \frac{1}{\hat{\mathbf{m}}^T\hat{\mathbf{m}}}. \end{aligned} \tag{A6}$$

Taking into account (23), we will consider that $\hat{\mathbf{m}}^T\mathbf{m} \simeq \mathbf{m}^T\mathbf{m}$, this will make the analysis tractable by following a parallel path to that of (13)-(14). Thus, defining $\hat{\nu} = \frac{1}{\hat{\mathbf{m}}^T\hat{\mathbf{m}}}$, it will be $E^{(1)}\left(\hat{z}/\hat{\mathbf{m}}\right) \simeq \frac{\mathbf{m}^T\mathbf{m}}{\hat{\mathbf{m}}^T\hat{\mathbf{m}}} = \frac{\hat{\nu}}{\nu}$; $E^{(2)}\left(\hat{z}/\hat{\mathbf{m}}\right) = 0$; $\mathrm{var}^{(1)}\left(\hat{z}/\hat{\mathbf{m}}\right) = \mathrm{var}^{(2)}\left(\hat{z}/\hat{\mathbf{m}}\right) = \hat{\nu}$ so that the probability densities of $\hat{z}$ in each class will be

$$p^{(1)}\left(\hat{z}/\hat{\mathbf{m}}\right)=\frac{1}{\sqrt{2\pi\hat{\nu}}}e^{-\frac{\left(\hat{z}-\frac{\hat{\nu}}{\nu}\right)^{2}}{2\hat{\nu}}}\quad p^{(2)}\left(\hat{z}/\hat{\mathbf{m}}\right)=\frac{1}{\sqrt{2\pi\hat{\nu}}}e^{-\frac{\hat{z}^{2}}{2\hat{\nu}}}. \tag{A7}$$

And therefore

$$\begin{aligned}
&\frac{1}{2}E_{\mathbf{x}}\left(\hat{d}\Big/k=2,\ \hat{\mathbf{m}}\right)+\frac{1}{2}\left(1-E_{\mathbf{x}}\left(\hat{d}\Big/k=1,\ \hat{\mathbf{m}}\right)\right)\\
&=\frac{1}{2}\Pr\left(\hat{z}>\frac{1}{2}\Big/k=2,\hat{\mathbf{m}}\right)\\
&\quad+\frac{1}{2}\left(1-\Pr\left(\hat{z}>\frac{1}{2}\Big/k=1,\hat{\mathbf{m}}\right)\right)\\
&=\frac{1}{2}\int_{\frac{1}{2}}^{\infty}\frac{1}{\sqrt{2\pi\hat{\nu}}}e^{-\frac{\hat{z}^{2}}{2\hat{\nu}}}d\hat{z}+\frac{1}{2}\left(1-\int_{\frac{1}{2}}^{\infty}\frac{1}{\sqrt{2\pi\hat{\nu}}}e^{-\frac{\left(\hat{z}-\frac{\hat{\nu}}{\nu}\right)^{2}}{2\hat{\nu}}}d\hat{z}\right)\\
&=\frac{1}{2}\int_{\frac{1}{2\sqrt{2\hat{\nu}}}}^{\infty}\frac{1}{\sqrt{\pi}}e^{-u^{2}}du+\frac{1}{2}\left(1-\int_{\frac{1-2\frac{\hat{\nu}}{\nu}}{2\sqrt{2\hat{\nu}}}}^{\infty}\frac{1}{\sqrt{\pi}}e^{-u^{2}}du\right)\\
&=\frac{1}{4}erfc\left(\frac{1}{2\sqrt{2\hat{\nu}}}\right)+\frac{1}{2}\left(1-\frac{1}{2}erfc\left(\frac{1-2\frac{\hat{\nu}}{\nu}}{2\sqrt{2\hat{\nu}}}\right)\right)\\
&=\frac{1}{4}erfc\left(\frac{1}{2\sqrt{2\hat{\nu}}}\right)+\frac{1}{4}erfc\left(\frac{2\frac{\hat{\nu}}{\nu}-1}{2\sqrt{2\hat{\nu}}}\right).
\end{aligned} \tag{A8}$$

where we have taken into account that $erfc(x)=2-erfc(-x)$. We then incorporate the result of (A8) into (A4), taking into account that the dependence on $\hat{\mathbf{m}}$ manifests itself through the dependence on $\hat{\nu}$

$$P_{e(N)}=E_{\hat{\nu}}\left[\frac{1}{4}erfc\left(\frac{1}{2\sqrt{2\hat{\nu}}}\right)+\frac{1}{4}erfc\left(\frac{2\frac{\hat{\nu}}{\nu}-1}{2\sqrt{2\hat{\nu}}}\right)\right]. \tag{A9}$$

To calculate (A9) we can take into account that $N\hat{\mathbf{m}}^{T}\hat{\mathbf{m}}=\frac{N}{\hat{\nu}}=\hat{\chi}$ is a non-central chi-squared random variable, with $M\times I$ degrees of freedom and non-centrality parameter $N\mathbf{m}^{T}\mathbf{m}=\frac{N}{\nu}$. Its mean value is $M\times I+\frac{N}{\nu}$ and its variance $2\left(M\times I+2\frac{N}{\nu}\right)$. Calling $p\left(\hat{\chi}\right)$ the corresponding probability density we have that

$$\begin{aligned}
&P_{e(N)}\\
&=\frac{1}{4}\int_{0}^{\infty}\left(erfc\left(\frac{1}{2\sqrt{2\frac{N}{\hat{\chi}}}}\right)+erfc\left(\frac{2\frac{N}{\nu\hat{\chi}}-1}{2\sqrt{2\frac{N}{\hat{\chi}}}}\right)\right)p\left(\hat{\chi}\right)d\hat{\chi}\\
&\underbrace{=\frac{1}{4}\int_{0}^{\infty}\left(erfc\left(\frac{1}{2\sqrt{2\frac{1}{\hat{\eta}}}}\right)+erfc\left(\frac{2\frac{1}{\nu\hat{\eta}}-1}{2\sqrt{2\frac{1}{\hat{\eta}}}}\right)\right)Np\left(N\hat{\eta}\right)d\hat{\eta}\,.}_{\hat{\chi}=N\hat{\eta};d\hat{\chi}=Nd\hat{\eta};}
\end{aligned} \tag{A10}$$

where $Np\left(N\hat{\eta}\right)=f\left(\hat{\eta}\right)$ is the probability density of the random variable $\hat{\eta}$. In short

$$\begin{aligned}
P_{e(N)}=\frac{1}{4}\int_{0}^{\infty}&\left(erfc\left(\frac{1}{2\sqrt{2\frac{1}{\hat{\eta}}}}\right)\right.\\
&\left.+erfc\left(\frac{2\frac{1}{\nu\hat{\eta}}-1}{2\sqrt{2\frac{1}{\hat{\eta}}}}\right)\right)f\left(\hat{\eta}\right)d\hat{\eta}
\end{aligned}$$

$$\begin{aligned}
E\left(\hat{\eta}\right)&=\frac{1}{N}E\left(\hat{\chi}\right)=\frac{M\times I+\frac{N}{\nu}}{N}=\frac{M\times I}{N}+\frac{1}{\nu}\\
\operatorname{var}\left(\hat{\eta}\right)&=\frac{1}{N^{2}}\operatorname{var}\left(\hat{\chi}\right)=\frac{2\left(M\times I+2\frac{N}{\nu}\right)}{N^{2}}\\
&=2\left(\frac{M\times I}{N^{2}}+\frac{2}{N\nu}\right).
\end{aligned} \tag{A11}$$

APPENDIX B
DERIVATION OF (26)

We can follow a similar derivation as in Appendix A. We start from (A4), (A5), the equivalent equations in this case are

$$\begin{aligned}
&P_{es(N)}\\
&=E_{\hat{\mathbf{m}}_{1}\ldots\hat{\mathbf{m}}_{2}}\left(\frac{1}{2}E_{\mathbf{x}}\left(\hat{d}\Big/k=2,\ \hat{\mathbf{m}}_{1}\ldots\hat{\mathbf{m}}_{2}\right)\right)\\
&\quad+E_{\hat{\mathbf{m}}_{1}\ldots\hat{\mathbf{m}}_{2}}\left(\frac{1}{2}\left(1-E_{\mathbf{x}}\left(\hat{d}\Big/k=1,\ \hat{\mathbf{m}}_{1}\ldots\hat{\mathbf{m}}_{2}\right)\right)\right)\\
&=E_{\hat{\mathbf{m}}_{1}\ldots\hat{\mathbf{m}}_{2}}\left(\frac{1}{2}\Pr\left(\hat{z}_{s}>\frac{1}{2}\Big/k=2,\hat{\mathbf{m}}_{1}\ldots\hat{\mathbf{m}}_{2}\right)\right)\\
&\quad+E_{\hat{\mathbf{m}}_{1}\ldots\hat{\mathbf{m}}_{2}}\left(\frac{1}{2}\left(1-\Pr\left(\hat{z}_{s}>\frac{1}{2}\Big/k=1,\hat{\mathbf{m}}_{1}\ldots\hat{\mathbf{m}}_{2}\right)\right)\right).
\end{aligned} \tag{B1}$$

where $\hat{z}_{s}$ is a Gaussian random variable whose mean and variance in each class are

$$\begin{aligned}
E^{(1)}\left(\hat{z}_{s}/\hat{\mathbf{m}}_{1}\ldots\hat{\mathbf{m}}_{I}\right)&=\frac{1}{I}\sum_{i=1}^{I}E^{(1)}\left(\hat{z}_{i}/\hat{\mathbf{m}}_{i}\right)\\
&\simeq\frac{1}{I}\sum_{i=1}^{I}\frac{\mathbf{m}_{i}^{T}\mathbf{m}_{i}}{\hat{\mathbf{m}}_{i}^{T}\hat{\mathbf{m}}_{i}}=\frac{1}{I}\sum_{i=1}^{I}\frac{\hat{\nu}_{i}}{\nu_{i}}\\
\operatorname{var}^{(1)}\left(\hat{z}_{s}/\hat{\mathbf{m}}_{1}\ldots\hat{\mathbf{m}}_{I}\right)&=\frac{1}{I^{2}}\sum_{i=1}^{I}\operatorname{var}^{(1)}\left(\hat{z}_{i}/\hat{\mathbf{m}}_{i}\right)\\
&=\frac{1}{I^{2}}\sum_{i=1}^{I}\frac{1}{\hat{\mathbf{m}}_{i}^{T}\hat{\mathbf{m}}_{i}}=\frac{1}{I^{2}}\sum_{i=1}^{I}\hat{\nu}_{i}\\
E^{(2)}\left(\hat{z}_{s}/\hat{\mathbf{m}}_{1}\ldots\hat{\mathbf{m}}_{I}\right)&=\frac{1}{I}\sum_{i=1}^{I}E^{(2)}\left(\hat{z}_{i}/\hat{\mathbf{m}}_{i}\right)=0;\\
\operatorname{var}^{(2)}\left(\hat{z}_{s}/\hat{\mathbf{m}}_{1}\ldots\hat{\mathbf{m}}_{I}\right)&=\frac{1}{I^{2}}\sum_{i=1}^{I}\operatorname{var}^{(2)}\left(\hat{z}_{i}/\hat{\mathbf{m}}_{i}\right)\\
&=\frac{1}{I^{2}}\sum_{i=1}^{I}\frac{1}{\hat{\mathbf{m}}_{i}^{T}\hat{\mathbf{m}}_{i}}=\frac{1}{I^{2}}\sum_{i=1}^{I}\hat{\nu}_{i}.
\end{aligned} \tag{B2}$$

Therefore

$$p^{(1)}\left(\hat{z}_{s}/\hat{\mathbf{m}}_{1},\hat{\mathbf{m}}_{2}\right)=\frac{1}{\sqrt{2\pi\frac{1}{I^{2}}\sum_{i=1}^{I}\hat{\nu}_{i}}}e^{-\frac{\left(\hat{z}_{s}-\frac{1}{I}\sum_{i=1}^{I}\frac{\hat{\nu}_{i}}{\nu_{i}}\right)^{2}}{2\frac{1}{I^{2}}\sum_{i=1}^{I}\hat{\nu}_{i}}}$$

$$p^{(2)}\left(\hat{z}_s/\hat{\mathbf{m}}_1,\hat{\mathbf{m}}_2\right)=\frac{1}{\sqrt{2\pi\frac{1}{I^2}\sum_{i=1}^{I}\hat{\nu}_i}}e^{-\frac{\hat{z}_s^2}{2\frac{1}{I^2}\sum_{i=1}^{I}\hat{\nu}_i}}. \tag{B3}$$

Thus, using a development similar to that of (A8) we arrive at

$$P_{es(N)}=E_{\hat{\nu}_1,\hat{\nu}_2}\left[\frac{1}{4}erfc\left(\frac{1}{2\sqrt{\frac{2}{I^2}\sum_{i=1}^{I}\hat{\nu}_i}}\right)+\frac{1}{4}erfc\left(\frac{2\frac{1}{I}\sum_{i=1}^{I}\frac{\hat{\nu}_i}{\nu_i}-1}{2\sqrt{\frac{2}{I^2}\sum_{i=1}^{I}\hat{\nu}_i}}\right)\right]. \tag{B4}$$

Then, taking into account that $\hat{\nu}_1,\ldots,\hat{\nu}_I$ and, therefore $\hat{\eta}_1,\ldots,\hat{\eta}_I$, are independent, we arrive at

$$P_{es(N)}=\frac{1}{4}\int_0^\infty\ldots\int_0^\infty\left(erfc\left(\frac{1}{2\sqrt{\frac{2}{I^2}\sum_{i=1}^{I}\frac{1}{\hat{\eta}_i}}}\right)+erfc\left(\frac{2\frac{1}{I}\sum_{i=1}^{I}\frac{1}{\nu_i\hat{\eta}_i}-1}{2\sqrt{\frac{2}{I^2}\sum_{i=1}^{I}\frac{1}{\hat{\eta}_i}}}\right)\right)\times f\left(\hat{\eta}_1\right)\ldots f\left(\hat{\eta}_I\right)d\hat{\eta}_1\ldots d\hat{\eta}_I$$
$$E\left(\hat{\eta}_i\right)=\frac{M}{N}+\frac{1}{\nu_i}\quad \mathrm{var}\left(\hat{\eta}_i\right)=2\left(\frac{M}{N^2}+\frac{2}{N\nu_i}\right)\; i=1\ldots I. \tag{B5}$$

APPENDIX C
DERIVATION OF (35)

The test in this case is $\hat{z}_C=\frac{\hat{\mathbf{m}}^T\hat{\mathbf{C}}^{-1}\mathbf{x}}{\hat{\mathbf{m}}^T\hat{\mathbf{C}}^{-1}\hat{\mathbf{m}}}\underset{k=2}{\overset{k=1}{\gtrless}}\frac{1}{2}$, so $\hat{z}_C$, conditional on $\hat{\mathbf{m}}$ and $\hat{\mathbf{C}}^{-1}$, is a Gaussian random variable whose mean and variance in each class are

$$\begin{aligned}
&E^{(1)}\left(\hat{z}_C/\hat{\mathbf{m}},\hat{\mathbf{C}}^{-1}\right)\\
&\quad=\frac{\hat{\mathbf{m}}^T\hat{\mathbf{C}}^{-1}E^{(1)}(\mathbf{x})}{\hat{\mathbf{m}}^T\hat{\mathbf{C}}^{-1}\hat{\mathbf{m}}}=\frac{\hat{\mathbf{m}}^T\hat{\mathbf{C}}^{-1}\mathbf{m}}{\hat{\mathbf{m}}^T\hat{\mathbf{C}}^{-1}\hat{\mathbf{m}}}\\
&\mathrm{var}^{(1)}\left(\hat{z}/\hat{\mathbf{m}},\hat{\mathbf{C}}^{-1}\right)\\
&\quad=E^{(1)}\left(\hat{z}^2/\hat{\mathbf{m}},\hat{\mathbf{C}}^{-1}\right)-E^{(1)^2}\left(\hat{z}/\hat{\mathbf{m}},\hat{\mathbf{C}}^{-1}\right)\\
&\quad=\frac{1}{\left(\hat{\mathbf{m}}^T\hat{\mathbf{C}}^{-1}\hat{\mathbf{m}}\right)^2}E^{(1)}\left(\hat{\mathbf{m}}^T\hat{\mathbf{C}}^{-1}\mathbf{x}\mathbf{x}^T\hat{\mathbf{C}}^{-1}\hat{\mathbf{m}}\right)\\
&\qquad-\left(\frac{\hat{\mathbf{m}}^T\hat{\mathbf{C}}^{-1}\mathbf{m}}{\hat{\mathbf{m}}^T\hat{\mathbf{C}}^{-1}\hat{\mathbf{m}}}\right)^2\\
&\quad=\frac{1}{\left(\hat{\mathbf{m}}^T\hat{\mathbf{C}}^{-1}\hat{\mathbf{m}}\right)^2}\left(\hat{\mathbf{m}}^T\hat{\mathbf{C}}^{-1}\underbrace{E^{(1)}\left(\mathbf{x}\mathbf{x}^T\right)}_{\mathbf{C}+\mathbf{m}\mathbf{m}^T}\hat{\mathbf{C}}^{-1}\hat{\mathbf{m}}\right)\\
&\qquad-\left(\frac{\hat{\mathbf{m}}^T\hat{\mathbf{C}}^{-1}\mathbf{m}}{\hat{\mathbf{m}}^T\hat{\mathbf{C}}^{-1}\hat{\mathbf{m}}}\right)^2\\
&\quad=\frac{1}{\left(\hat{\mathbf{m}}^T\hat{\mathbf{C}}^{-1}\hat{\mathbf{m}}\right)^2}\left(\hat{\mathbf{m}}^T\hat{\mathbf{C}}^{-1}\mathbf{C}\hat{\mathbf{C}}^{-1}\hat{\mathbf{m}}\right)\\
&E^{(2)}\left(\hat{z}/\hat{\mathbf{m}},\hat{\mathbf{C}}^{-1}\right)\\
&\quad=\frac{\hat{\mathbf{m}}^T\hat{\mathbf{C}}^{-1}E^{(2)}(\mathbf{x})}{\hat{\mathbf{m}}^T\hat{\mathbf{C}}^{-1}\hat{\mathbf{m}}}=0;\\
&\mathrm{var}^{(2)}\left(\hat{z}/\hat{\mathbf{m}},\hat{\mathbf{C}}^{-1}\right)\\
&\quad=E^{(2)}\left(\hat{z}^2/\hat{\mathbf{m}}\right)-E^{(2)^2}\left(\hat{z}/\hat{\mathbf{m}}\right)\\
&\quad=\frac{1}{\left(\hat{\mathbf{m}}^T\hat{\mathbf{C}}^{-1}\hat{\mathbf{m}}\right)^2}\left(\hat{\mathbf{m}}^T\hat{\mathbf{C}}^{-1}\mathbf{C}\hat{\mathbf{C}}^{-1}\hat{\mathbf{m}}\right).
\end{aligned} \tag{C1}$$

Let us consider again a couple of reasonable approximations to naturally extend the results of the uncorrelated case. We will assume that $\hat{\mathbf{m}}^T\hat{\mathbf{C}}^{-1}\mathbf{m}\simeq\mathbf{m}^T\hat{\mathbf{C}}^{-1}\mathbf{m}$ and that $\hat{\mathbf{C}}^{-1}\mathbf{C}\simeq\mathbf{I}_{2M\times 2M}$. Thus, defining $\hat{\nu}_C=\frac{1}{\hat{\mathbf{m}}^T\hat{\mathbf{C}}^{-1}\hat{\mathbf{m}}}$, it will be $E^{(1)}\left(\hat{z}_C/\hat{\mathbf{m}},\hat{\mathbf{C}}^{-1}\right)=\frac{\hat{\mathbf{m}}^T\hat{\mathbf{C}}^{-1}\mathbf{m}}{\hat{\mathbf{m}}^T\hat{\mathbf{C}}^{-1}\hat{\mathbf{m}}}=\frac{\hat{\nu}_C}{\nu_C}$; $E^{(2)}\left(\hat{z}_C/\hat{\mathbf{m}},\hat{\mathbf{C}}^{-1}\right)=0$; $\mathrm{var}^{(1)}\left(\hat{z}/\hat{\mathbf{m}},\hat{\mathbf{C}}^{-1}\right)=\mathrm{var}^{(2)}\left(\hat{z}/\hat{\mathbf{m}},\hat{\mathbf{C}}^{-1}\right)\simeq\hat{\nu}_C$.

From this point on, we can directly apply (A7) to (A9) by simply substituting $\hat{\nu}$ by $\hat{\nu}_C$ and ν by ν_C, arriving at

$$P_{eC(N)}=\frac{1}{4}E_{\hat{\nu}_C}\left[erfc\left(\frac{1}{2\sqrt{2\hat{\nu}_C}}\right)+erfc\left(\frac{2\frac{\hat{\nu}_C}{\nu}-1}{2\sqrt{2\hat{\nu}_C}}\right)\right]. \tag{C2}$$

From (C2) we can arrive at an integral equivalent to (A11) by defining $\hat{\nu}_C=\frac{1}{\hat{\eta}_C}$

$$P_{eC(N)}=\frac{1}{4}\int_0^\infty\left(erfc\left(\frac{1}{2\sqrt{2\frac{1}{\hat{\eta}_C}}}\right)+erfc\left(\frac{2\frac{1}{\nu\hat{\eta}_C}-1}{2\sqrt{2\frac{1}{\hat{\eta}_C}}}\right)\right)f\left(\hat{\eta}_C\right)d\hat{\eta}_C. \tag{C3}$$

where in this case it is not possible to determine the probability density $f\left(\hat{\eta}_C\right)$ due to the presence of the esimated covariance, i.e., $\hat{\eta}_C=\hat{\mathbf{m}}^T\hat{\mathbf{C}}^{-1}\hat{\mathbf{m}}$. However we can calculate the mean and variance of $\hat{\eta}_C$, as we see below. First, we can take into account that matrix $\frac{1}{N-1}\hat{\mathbf{C}}^{-1}$ follows an

Wishart distribution $\frac{1}{N-1}\hat{\mathbf{C}}^{-1} \sim W_M^{-1}\left(\mathbf{C}^{-1}, N-1\right)$ [28], with $E_{\hat{\mathbf{C}}^{-1}}\left[\hat{\mathbf{C}}^{-1}\right] = \frac{N-1}{N-M\times I-2}\mathbf{C}^{-1}$ being satisfied for $N > M \times I + 2$.

On the other hand, based on the well-known result for the calculation of the mean of quadratic forms of random vectors [29] and applying the law of the total expectation, we arrive at

$$\begin{aligned} E\left(\hat{\eta}_C\right) &= E_{\hat{\mathbf{m}}}\left(E_{\hat{\mathbf{C}}^{-1}}\left(\hat{\eta}_C/\hat{\mathbf{m}}\right)\right) = E_{\hat{\mathbf{m}}}\left(\hat{\mathbf{m}}^T E_{\hat{\mathbf{C}}^{-1}}\left(\hat{\mathbf{C}}^{-1}\right)\hat{\mathbf{m}}\right) \\ &= trace\left(\frac{N-1}{N-M\times I-2}\mathbf{C}^{-1}\mathbf{C}_{\hat{\mathbf{m}}\hat{\mathbf{m}}}\right) \\ &\quad + \mathbf{m}^T\frac{N-1}{N-M\times I-2}\mathbf{C}^{-1}\mathbf{m} \\ &= trace\left(\frac{N-1}{N-M\times I-2}\mathbf{C}^{-1}\frac{1}{N}\mathbf{C}\right) + \frac{1}{\nu_C} \\ &= \frac{N-1}{N-M\times I-2}\frac{M\times I}{N} + \frac{1}{\nu_C} \\ &= \frac{1}{c_{M\times I,N}}\left(\frac{M\times I}{N} + \frac{1}{\nu_C}\right) \quad N > M\times I + 2. \end{aligned} \tag{C4}$$

where we have taken into account that, being $\hat{\mathbf{m}} = \frac{1}{N}\sum_{m=1}^{N}\mathbf{x}_m^{(1)}$, the covariance matrix of $\hat{\mathbf{m}}$ will be $\mathbf{C}_{\hat{\mathbf{m}}\hat{\mathbf{m}}} = \frac{1}{N}\mathbf{C}$. In (C4) we have defined the factor $c_{M\times I,N} = \frac{N-M\times I-2}{N-1}$, which varies between 0 $(N = M\times I+2)$, to 1 $(N=\infty)$. It can be interpreted as a *convergence factor* that slowdown the error probability convergence for increasing training size, with respect to the case of uncorrelated data. This slowdown is due to the need to estimate the covariance matrix.

On the other hand, applying the law of total variance

$$\mathrm{var}\left(\hat{\eta}_C\right) = \mathrm{var}_{\hat{\mathbf{m}}}\left(E_{\hat{\mathbf{C}}^{-1}}\left(\hat{\eta}_C/\hat{\mathbf{m}}\right)\right) + E_{\hat{\mathbf{m}}}\left(\mathrm{var}_{\hat{\mathbf{C}}^{-1}}\left(\hat{\eta}_C/\hat{\mathbf{m}}\right)\right). \tag{C5}$$

For the first term in (C5) we can apply the well-known result for calculating the variance of quadratic forms of random vectors [29]:

$$\begin{aligned} &\mathrm{var}_{\hat{\mathbf{m}}}\left(E_{\hat{\mathbf{C}}^{-1}}\left(\hat{\eta}_C/\hat{\mathbf{m}}\right)\right) \\ &= \mathrm{var}_{\hat{\mathbf{m}}}\left(\hat{\mathbf{m}}^T E_{\hat{\mathbf{C}}^{-1}}\left(\hat{\mathbf{C}}^{-1}\right)\hat{\mathbf{m}}\right) \\ &= 2trace\left(\frac{N-1}{N-M\times I-2}\mathbf{C}^{-1}\mathbf{C}_{\hat{\mathbf{m}}\hat{\mathbf{m}}}\frac{N-1}{N-M\times I-2}\right. \\ &\qquad \left.\times\,\mathbf{C}^{-1}\mathbf{C}_{\hat{\mathbf{m}}\hat{\mathbf{m}}}\right) \\ &\quad + 4\mathbf{m}^T\frac{N-1}{N-M\times I-2}\mathbf{C}^{-1}\mathbf{C}_{\hat{\mathbf{m}}\hat{\mathbf{m}}}\frac{N-1}{N-M\times I-2}\mathbf{C}^{-1}\mathbf{m} \\ &= 2\left(\frac{N-1}{N-M\times I-2}\right)^2\frac{M\times I}{N^2} \\ &\quad + 4\left(\frac{N-1}{N-M\times I-2}\right)^2\mathbf{m}^T\frac{1}{N}\mathbf{C}^{-1}\mathbf{m} \\ &= 2\left(\frac{N-1}{N-M\times I-2}\right)^2\left(\frac{M\times I}{N^2} + \frac{2}{N\nu_C}\right) \\ &\qquad N > M\times I+2. \end{aligned} \tag{C6}$$

As for the second term in (C4) let us take into account that a quadratic form of an inverse Wishart matrix, properly normalised, follows an inverse chi-squared distribution [30], which implies in our case that

$$\left(\frac{\hat{\mathbf{m}}^T\hat{\mathbf{C}}^{-1}\hat{\mathbf{m}}}{N-1}\Big/\hat{\mathbf{m}}\right)\Big/\left(\hat{\mathbf{m}}^T\mathbf{C}^{-1}\hat{\mathbf{m}}\big/\hat{\mathbf{m}}\right) \sim inv\chi^2_{N-2M}.$$

Considering further that $x \sim inv\chi^2_G \Rightarrow \mathrm{var}(x) = 2/(G-2)^2(G-4)$ $G > 4$, we can write

$$\begin{aligned} &E_{\hat{\mathbf{m}}}\left(\mathrm{var}_{\hat{\mathbf{C}}^{-1}}\left(\hat{\eta}_C/\hat{\mathbf{m}}\right)\right) \\ &= E_{\hat{\mathbf{m}}}\left(\mathrm{var}_{\hat{\mathbf{C}}^{-1}}\left(\hat{\mathbf{m}}^T\hat{\mathbf{C}}^{-1}\hat{\mathbf{m}}\right)\right) \\ &= E_{\hat{\mathbf{m}}}\left(\left(\hat{\mathbf{m}}^T\mathbf{C}^{-1}\hat{\mathbf{m}}\right)^2\mathrm{var}_{\hat{\mathbf{C}}^{-1}}\left(\left(\hat{\mathbf{m}}^T\hat{\mathbf{C}}^{-1}\hat{\mathbf{m}}\right)\Big/\left(\hat{\mathbf{m}}^T\mathbf{C}^{-1}\hat{\mathbf{m}}\right)\right)\right) \\ &= E_{\hat{\mathbf{m}}}\left(\left(\hat{\mathbf{m}}^T\mathbf{C}^{-1}\hat{\mathbf{m}}\right)^2\right)\frac{2\left(N-1\right)^2}{\left(N-M\times I-2\right)^2\left(N-M\times I-4\right)} \\ &\qquad N > M\times I+4. \end{aligned} \tag{C7}$$

But

$$\begin{aligned} &E_{\hat{\mathbf{m}}}\left(\left(\hat{\mathbf{m}}^T\mathbf{C}^{-1}\hat{\mathbf{m}}\right)^2\right) \\ &= \mathrm{var}_{\hat{\mathbf{m}}}\left(\hat{\mathbf{m}}^T\mathbf{C}^{-1}\hat{\mathbf{m}}\right) + E^2_{\hat{\mathbf{m}}}\left(\hat{\mathbf{m}}^T\mathbf{C}^{-1}\hat{\mathbf{m}}\right) \\ &= 2trace\left(\mathbf{C}^{-1}\mathbf{C}_{\hat{\mathbf{m}}\hat{\mathbf{m}}}\mathbf{C}^{-1}\mathbf{C}_{\hat{\mathbf{m}}\hat{\mathbf{m}}}\right) + 4\mathbf{m}^T\mathbf{C}^{-1}\mathbf{C}_{\hat{\mathbf{m}}\hat{\mathbf{m}}}\mathbf{C}^{-1}\mathbf{m} \\ &\quad + \left(trace\left(\mathbf{C}^{-1}\mathbf{C}_{\hat{\mathbf{m}}\hat{\mathbf{m}}}\right) + \mathbf{m}^T\mathbf{C}^{-1}\mathbf{m}\right)^2 \\ &= 2\frac{M\times I}{N^2} + 4\frac{1}{N\nu_C} + \left(\frac{M\times I}{N} + \frac{1}{\nu_C}\right)^2. \end{aligned} \tag{C8}$$

So finally, incorporating (C6), (C7) and (C8) in (C5), we arrive at

$$\begin{aligned} &\mathrm{var}\left(\hat{\eta}_C\right) \\ &= 2\left(\frac{N-1}{N-2M-2}\right)^2\left(\frac{M\times I}{N^2} + \frac{2}{N\nu_C}\right) \\ &\quad + \frac{2\left(N-1\right)^2}{\left(N-M\times I-2\right)^2\left(N-2M-4\right)} \\ &\quad \times\left(2\frac{M\times I}{N^2} + 4\frac{1}{N\nu_C} + \left(\frac{M\times I}{N} + \frac{1}{\nu_C}\right)^2\right) \\ &= \frac{1}{c^2_{M\times I,N}}2\left(\frac{M\times I}{N^2} + \frac{2}{N\nu_C}\right)\left(1 + \frac{2}{N-M\times I-4}\right) \\ &\quad + E^2\left(\hat{\eta}_C\right)\frac{2}{N-M\times I-4} \qquad N > M\times I+4. \end{aligned} \tag{C9}$$

REFERENCES

[1] F. Castanedo, "A review of data fusion techniques," *Scientific World J.*, vol. 2013, Oct. 2013, Art. no. 704504.

[2] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: An overview of methods, challenges, and prospects," *Proc. IEEE*, vol. 103, no. 9, pp. 1449–1477, Sep. 2015.

[3] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[4] A. Barua, M. U. Ahmed, and S. Begum, "A systematic literature review on multimodal machine learning: Applications, challenges, gaps and future directions," *IEEE Access*, vol. 11, pp. 14804–14831, 2023.

[5] B. Khaleghi, A. Khamis, F. O. Karray, and S. N. Razavi, "Multisensor data fusion: A review of the state-of-the-art," *Inf. Fusion*, vol. 14, no. 1, pp. 28–44, Jan. 2013.

[6] M. L. Fung, M. Z. Q. Chen, and Y. H. Chen, "Sensor fusion: A review of methods and applications," in *Proc. 29th Chin. Control Decis. Conf. (CCDC)*, Chongqing, China, May 2017, pp. 3853–3860.

[7] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, pp. 226–239, Mar. 1998.

[8] M. Mohandes, M. Deriche, and S. O. Aliyu, "Classifiers combination techniques: A comprehensive review," *IEEE Access*, vol. 6, pp. 19626–19639, 2018.

[9] A. R. Elkordy, Y. H. Ezzeldin, S. Han, S. Sharma, C. He, S. Mehrotra, and S. Avestimehr, "Federated analytics: A survey," *APSIPA Trans. Signal Inf. Process.*, vol. 12, no. 1, pp. 1–33, 2023.

[10] K. Gadzicki, R. Khamsehashari, and C. Zetzsche, "Early vs late fusion in multimodal convolutional neural networks," in *Proc. IEEE 23rd Int. Conf. Inf. Fusion (FUSION)*, Rustenburg, South Africa, Jul. 2020, pp. 1–6.

[11] J. H. Mervitz, J. P. de Villiers, J. P. Jacobs, and M. H. O. Kloppers, "Comparison of early and late fusion techniques for movie trailer genre labelling," in *Proc. IEEE 23rd Int. Conf. Inf. Fusion (FUSION)*, Rustenburg, South Africa, Jul. 2020, pp. 1–8.

[12] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders, "Early versus late fusion in semantic video analysis," in *Proc. 13th Annu. ACM Int. Conf. Multimedia*, Singapore, Nov. 2005, pp. 399–402.

[13] W. Yao, A. Moumtzidou, C. O. Dumitru, S. Andreadis, I. Gialampoukidis, S. Vrochidis, M. Datcu, and I. Kompatsiaris, "Early and late fusion of multiple modalities in Sentinel imagery and social media retrieval," in *Proc. Int. Workshops Challenges Pattern Recognit. (ICPR)*, vol. 12667, 2021, pp. 521–606.

[14] Y. Dong, S. Gao, K. Tao, J. Liu, and H. Wang, "Performance evaluation of early and late fusion methods for generic semantics indexing," *Pattern Anal. Appl.*, vol. 17, no. 1, pp. 37–50, Feb. 2014.

[15] G. Barnum, S. Talukder, and Y. Yue, "On the benefits of early fusion in multimodal representation learning," in *Proc. 2nd Workshop Shared Visual Represent. Human Machine Intell. (SVRHM)*, 2020, pp. 1–14.

[16] M. Pawłowski, A. Wróblewska, and S. Sysko-Romanczuk, "Effective techniques for multimodal data fusion: A comparative analysis," *Sensors*, vol. 23, no. 5, p. 2381, Feb. 2023.

[17] B. Li and A. Sano, "Early versus late modality fusion of deep wearable sensor features for personalized prediction of tomorrow's mood, health, and stress," in *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Montreal, QC, Canada, Jul. 2020, pp. 5896–5899.

[18] S. J. Raudys and A. K. Jain, "Small sample size effects in statistical pattern recognition: Recommendations for practitioners," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 2, pp. 252–264, Mar. 1991.

[19] M. Sordo and Q. Zeng, "On sample size and classification accuracy: A performance comparison," in *Biological and Medical Data Analysis*, J. L. Oliveira, V. Maojo, F. Martín-Sínchez, and A. Sousa, Eds. Berlin, Germany: Springer, 2005, pp. 193–201.

[20] C. Beleites, U. Neugebauer, T. Bocklitz, C. Krafft, and J. Popp, "Sample size planning for classification models," *Analytica Chim. Acta*, vol. 760, pp. 25–33, Jan. 2013.

[21] A. Alwosheel, S. van Cranenburgh, and C. G. Chorus, "Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis," *J. Choice Model.*, vol. 28, pp. 167–182, Sep. 2018.

[22] Y. Wahba, E. ElSalamouny, and G. ElTaweel, "Estimating the sample size for training intrusion detection systems," *Int. J. Comput. Netw. Inf. Secur.*, vol. 9, no. 12, pp. 1–10, Dec. 2017.

[23] R. L. Figueroa, Q. Zeng-Treitler, S. Kandula, and L. H. Ngo, "Predicting sample size required for classification performance," *BMC Med. Inform. Decis. Making*, vol. 12, pp. 1–10, Feb. 2012.

[24] A. Salazar, L. Vergara, and E. Vidal, "A proxy learning curve for the Bayes classifier," *Pattern Recognit.*, vol. 136, Apr. 2023, Art. no. 109240.

[25] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. New York, NY, USA: Wiley, 2001, pp. 2–18.

[26] V. Pestov, "Is the k-NN classifier in high dimensions affected by the curse of dimensionality?" *Comput. Math. Appl.*, vol. 65, no. 10, pp. 1427–1437, May 2013.

[27] S. M. Kay, *Fundamentals of Statistical Signal Processing: Detection Theory*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1998.

[28] V. Kanti, J. T. Mardia, and B. J. M. Kent, *Multivariate Analysis*. London, U.K.: Academic Press, 1979.

[29] A. C. Rencher and G. B. Schaalje, *Linear Models in Statistics*. Hoboken, NJ, USA: Wiley, 2008.

[30] D. A. Harville, *Linear Models and the Relevant Distributions and Matrix Algebra*. London, U.K.: Chapman & Hall, 2018.

[31] A. Soriano, L. Vergara, B. Ahmed, and A. Salazar, "Fusion of scores in a detection context based on alpha integration," *Neural Comput.*, vol. 27, no. 9, pp. 1983–2010, Sep. 2015.

LUIS MANUEL PEREIRA received the degree from Hermanos Saíz Montes de Oca University, Pinar del Río, Cuba, in 2015. He is currently pursuing the Ph.D. degree with the Institute of Telecommunications and Multimedia Applications (iTEAM), Universitat Politècnica de València. He is a telecommunications and electronics engineer. His current research interest includes the fusion of data associated with different biomedical modalities.

ADDISSON SALAZAR (Member, IEEE) received the Ph.D. degree in electrical engineering from Universitat Politècnica de València (UPV), in 2011. He is currently a Senior Researcher with the Institute of Telecommunications and Multimedia Applications, UPV. He has more than 100 articles in statistical signal processing, machine learning, decision fusion, and pattern recognition.

LUIS VERGARA received the Ph.D. degree in electrical engineering from Universidad Politécnica de Madrid, in 1983. He is currently a Full Professor of telecommunications, signal and data processing with Universitat Politècnica de València. He has more than 250 publications in theoretical and applied problems of signal and data processing and has led many important projects in these fields.