Pattern Recognition 45 (2012) 1853–1865

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

On signal representations within the Bayes decision framework

Jorge F. Silva [a],*, Shrikanth S. Narayanan [b]

[a] University of Chile, Department of Electrical Engineering, Av. Tupper 2007, Santiago 412-3, Chile
[b] University of Southern California, Department of Electrical Engineering, Los Angeles, CA 90089 2564, USA

ARTICLE INFO

Article history:
Received 22 September 2010
Received in revised form
24 September 2011
Accepted 22 November 2011
Available online 3 December 2011

Keywords:
Signal representation
Minimum risk decision
Bayes decision framework
Estimation-approximation error tradeoff
complexity regularization
Mutual information
Decision trees
Linear discriminant analysis

ABSTRACT

This work presents new results in the context of minimum probability of error signal representation (MPE-SR) within the Bayes decision framework. These results justify addressing the MPE-SR criterion as a complexity-regularized optimization problem, demonstrating the empirically well understood tradeoff between signal representation quality and learning complexity. Contributions are presented in three folds. First, the stipulation of conditions that guarantee a formal tradeoff between approximation and estimation errors under sequence of embedded transformations are provided. Second, the use of this tradeoff to formulate the MPE-SR as a complexity regularized optimization problem, and an approach to address this oracle criterion in practice is given. Finally, formal connections are provided between the MPE-SR criterion and two emblematic feature transformation techniques used in pattern recognition: the optimal quantization problem of classification trees (CART tree pruning algorithms), and some versions of Fisher linear discriminant analysis (LDA).

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

The optimal signal representation is a fundamental problem that the signal processing community has been addressing from different angles and under multiple research contexts. The formulation and solution to this problem has provided significant contributions in lossy compression, estimation and de-noising realms [1–3]. In the context of pattern recognition, signal representation issues are naturally associated with feature extraction (FE). In contrast to compression and de-noising scenarios, where the objective is to design bases that allow optimal representation of the observation source, for instance in the mean square error sense, in pattern recognition we seek representations that capture an unobserved finite alphabet phenomena, the class identity, from the observed signal. Consequently, a suitable optimality criterion is associated with minimizing the risk of taking the mentioned decision, as considered in Bayes decision framework [4–6]. In this context, the observation signal can be considered a combination of multiple sources not all of them related with the underlying target phenomenon. Hence, from a signal representation point of view, one objective is to characterize the observation subspace which is relevant for the decision problem, the well known concept of sufficient statistics [7–9]. An example of the use of sufficient statistics for feature representation is the basic detection problem formulated in communication theory [7]. In this scenario, the signal constellation and the statistics of the channel are known, and consequently there is an analytical solution for the observation subspace which captures the sufficient statistics, the well-known matching filter [7]. In pattern recognition, instead, we are dealing with a more challenging scenario, because we do not know the generative process in which different sources are combined to generate the observation phenomenon and we need to address the problem of FE in an unsupervised way.

Assuming that we know the observation-class distribution, the minimum risk decision can be obtained, the well-known Bayes rule [6,4]. However, in practice this distribution is unknown, which introduces the learning aspect of the problem. In this context, signal representation plays a key role, beyond the concept of sufficient statistics, as a techniques for doing dimensionally reduction (see a recent survey in [10]). The Bayes framework proposes to estimate this joint distribution based on a finite amount of training data [4,5]. It is well known that the accuracy of this estimation process is affected by the dimensionality of the observation space—the curse of dimensionality—which is proportional to the mismatch between the real and the estimated distributions, the estimation error effect of the learning problem. Then, an integral part of the feature extraction (FE) is to control the estimation error by finding suitable parsimonious signal

* Corresponding author. Tel.: +56 2 9784196; fax: +56 2 6953881.
E-mail addresses: josilva@ing.uchile.cl (J.F. Silva), shri@sipi.usc.edu (S.S. Narayanan).

0031-3203/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2011.11.015

transformations, particularly necessary in scenarios where the original raw observation measurements lie in a high-dimensional space and only a limited amount of training data are available (relative to the raw dimension of the problem), such as in most speech classification [11], image classification [12] and hyperspectral classification scenarios [13,14].

The FE problem in many cases considers a particular domain or task knowledge. This knowledge is used to characterize potentially salient features. For example, in the case of speech recognition, the short-term spectral envelope of the speech signal provides useful phonetic discrimination [15]. However, there are some problems in which it is not possible to characterize the set of relevant features in advance. A principle that can be used to select those salient features from a relatively large collection of potential representations hence is a central problem in FE. Many algorithms have been proposed along this direction for finding feature transformations that minimize some optimality criterion, directly or indirectly associated with the probability of error. Examples of these include information measures like the Kullback–Leibler divergence (KLD) [16] and mutual information [11], and empirical measures like the Mahalanobis distance and Fisher's class separability metric [4,12]. The proposed solutions for the FE problem variedly impose assumptions on the family of feature transformations, on the joint class-observation distributions—parametric or non-parametric, and on the optimality criterion, which allow to approximate or find closed-form solutions in a particular problem domain. Despite issues in finding feature representations of lower complexity which capture the most discriminant aspects of the full measurement-observation space, the problem is a well motivated one and good approximations have been presented under specific modeling assumptions [16–18,11]. Nevertheless, there has not been a concrete general formulation of the ultimate problem, which is to find the minimum probability error signal representation (MPE-SR) constrained on a given amount of training data or any additional operational cost that may constrain the decision task. Such a formulation would provide a better theoretical support and justification for the aforementioned FE problem and their existing algorithmic solutions.

Motivated by this need, new results in formalizing the MPE-SR problem have been presented in the seminal work by Vasconcelos [18]. Ref. [18] formalizes a tradeoff between the Bayes error and an information-theoretic indicator of the estimation error, and connects this result with the concept of optimal signal representation. The estimation and approximation error tradeoff was obtained with respect to a sequence of embedded representations (features) derived from coordinate projections, a special case of linear transformation. Silva et al. [19] provided a basic extension of these ideas for more general embedded feature collections. The present work extends the results in [19], and is motivated by, and is built upon the ideas in [18].

1.1. Specific contributions

The central result presented in this work is the stipulation of sufficient conditions that guarantee a formal tradeoff between Bayes error and estimation error across sequences of embedded feature transformations for continuous and finite alphabet feature spaces. These sufficient conditions not only take into consideration the embedded structure of the feature representation family, as the original results in [18], but also the consistent nature of the family of the empirical observation-class distributions estimated across the sequence of transformations—explicitly incorporating the role of the learning phase of the problem, and consequently generalizing the results presented in [18] for continuous feature representations. In addition this tradeoff is obtained for a rich collection of embedded features, significantly extending the scope of applicability of [18]. Furthermore, for the important scenario of finite alphabet representations (or quantization of the raw observation space) new results are presented where the notion of embedded representations; the estimation error quantity based on the KLD; and the tradeoff between estimation and approximation errors are developed in this work.

Following that, the Bayes-estimation error tradeoff is used to formulate the MPE-SR problem as a complexity-regularized optimization, with an objective function that considers a fidelity indicator, which represents the Bayes error, and a cost term—associated with the complexity of the representation—which reflects the estimation error. We show that the solution of this problem relies on a particular sequence of representations, which is the solution of a cost-fidelity problem. Interestingly restricting the problem and invoking some approximations, the well known CART pruning algorithm [5] and Fisher linear discriminant analysis [4], offer computationally efficient solutions for this cost-fidelity problem. Consequently, we are able to demonstrate that these well-known techniques are intrinsically addressing the MPE-SR problem.

1.2. Paper organization

Section 2 introduces the problem formulation, terminologies and key results that will be used in the rest of the exposition. Section 3 presents the Bayes-estimation tradeoff and Section 4 the MPE-SR problem and its cost-fidelity approximation. Sections 5.1 and 5.2 show how the MPE-SR can be addressed practically in two important scenarios: classification tree (CART pruning algorithms) and linear discriminant analysis. To conclude Sections 6 and 7 offer a discussion of the presented results and future work, respectively.

2. Preliminaries: Bayes decision approach

Let $X:(\Omega,\mathcal{F},\mathbb{P})\rightarrow(\mathcal{X},\mathcal{F}_{\mathcal{X}})$ be an observation random vector taking values in a finite dimensional Euclidean space $\mathcal{X}=\mathbb{R}^K$, and $Y:(\Omega,\mathcal{F},\mathbb{P})\rightarrow(\mathcal{Y},\mathcal{F}_{\mathcal{Y}})$ be a class label random variable with values in a finite alphabet space $\mathcal{Y}$.[1] $(\Omega,\mathcal{F},\mathbb{P})$ denotes the underlying probability space. Knowing the joint distribution $P_{X,Y}$ in $(\mathcal{X}\times\mathcal{Y},\sigma(\mathcal{F}_{\mathcal{X}}\times\mathcal{F}_{\mathcal{Y}}))$,[2] the problem is to find a decision function $g(\cdot)$ from $\mathcal{X}$ to $\mathcal{Y}$ such that for a given realization of X, infer its discrete counterpart Y with the minimum expected cost, or minimum risk given by $\mathbb{E}_{X,Y}[l(g(X),Y)]$, where $l(y_1,y_2)$ denotes the risk of labeling an observation with the value y_1, when its true label is y_2, $\forall\, y_1,y_2\in\mathcal{Y}$. The minimum risk decision is called *the Bayes rule*, where for the classical 0–1 risk function [4], $l(y_1,y_2)=\delta(y_1,y_2)$, the Bayes rule in (1) minimizes the probability error:

$$g_{P_{X,Y}}(\overline{x})\equiv\arg\max_{y\in\mathcal{Y}}P_{X,Y}(\overline{x},y),\quad\forall\,\overline{x}\in\mathcal{X}. \tag{1}$$

In this case the minimum probability error (MPE), or *Bayes error*, can be expressed by [6]:

$$\begin{aligned}L_{\mathcal{X}}&\equiv\mathbb{P}(\{u\in\Omega:g_{P_{X,Y}}(X(u))\neq Y(u)\})=P_{X,Y}(\{(x,y)\in\mathcal{X}\times\mathcal{Y}:g_{P_{X,Y}}(x)\neq y\})\\&=1-\mathbb{E}_X[\max_{i\in\mathcal{Y}}P_{Y|X}(i|X)].\end{aligned}\tag{2}$$

The subscript notation in $L_{\mathcal{X}}$ emphasizes that this is an indicator of the discrimination power of the observation space $\mathcal{X}$ and more precisely of the joint distribution $P_{X,Y}$. The following lemma states a version of the well-known result that a transformation of the

[1] Given that $\mathcal{X}=\mathbb{R}^K$, a natural choice for $\mathcal{F}_{\mathcal{X}}$ is the Borel sigma field [20], and for $\mathcal{F}_{\mathcal{Y}}$ the power set of $\mathcal{Y}$.

[2] $\sigma(\mathcal{F}_{\mathcal{X}}\times\mathcal{F}_{\mathcal{Y}})$ refers to the product sigma field [20].

observation space $\mathcal{X}$ can not provide discrimination gain or the *data-processing inequality*.

Lemma 1 (*Vasconcelos [18, Theorem 3], Wozencraft and Jacobs [7]*). *Consider* $\mathbf{f}:(\mathcal{X},\mathcal{F}_{\mathcal{X}})\to(\mathcal{X}',\mathcal{F}'_{\mathcal{X}})$ *to be a measurable mapping. If we define* $X'\equiv\mathbf{f}(X)$ *as a new observation random variable, with joint probability distribution* $P_{X',Y}$ *induced by* $\mathbf{f}(\cdot)$ *and* $P_{X,Y}$ [21], *we have that*

$$L_{X'} \geq L_X. \tag{3}$$

From the lemma, it is natural to say that the transformation $\mathbf{f}(\cdot)$ represents *sufficient statistics* for the inference problem if $L_{X'}=L_X$, see [7,4,8,22].

In practice we do not know the joint distribution $P_{X,Y}$. Instead we may have access to independent and identically distributed (i.i.d.) realizations of (X,Y), $D_N\equiv\{(x_i,y_i):i\in\{1,\ldots,N\}\}$, which in the Bayes approach are used to characterize an estimation of the joint observation-class distribution, the empirical distribution denoted by $\hat{P}_{X,Y}$. This estimated distribution $\hat{P}_{X,Y}$ is used to define *the plug-in empirical Bayes rule*, using (1), that we denote as $\hat{g}_{\hat{P}_{X,Y}}(\cdot)$. Note that the risk of the empirical Bayes rule in (4), differs from the Bayes error $L_{\mathcal{X}}$ as a consequence of what is called the estimation error effect in the learning process:

$$\mathbb{P}(\{u\in\Omega:\hat{g}_{\hat{P}_{X,Y}}(X(u))\neq Y(u)\}) \tag{4}$$

It is well understood that the magnitude of this estimation error is a function of some notion of complexity of the observation space [23,13,18]. This implies a strong relationship between the number of training examples and the complexity of the observation space, justifying the widely adopted dimensionality reduction during FE [10].

In this work, we focus on studying aspects of optimal feature representation for classification, assuming the Bayes decision approach, and that the learning framework satisfies certain conditions that will be detailed in the next section. Under these assumptions, we can formally consider two signal representation aspects that affect the performance of a Bayes decision framework. One relates to the signal representation quality, associated with the Bayes error, and the other to the signal space complexity, to quantify the effect of the estimation error in the problem. The formalization of this tradeoff and its implications are the main topics addressed in the following sections.

3. Signal representation results for the Bayes approach

Let us start with a result that provides an analytical expression to bound the performance deviation of the empirical Bayes rule with respect to the Bayes error.

Theorem 1 (*Vasconcelos [18, Theorem 4]*). *Let us consider the joint observation-class distribution* $P_{X,Y}$ *and its empirical counterpart* $\hat{P}_{X,Y}$, *assuming that they only differ in their class conditional probabilities (i.e.,* $\hat{P}_Y(\{y\})=P_Y(\{y\}),\forall\, y\in\mathcal{Y}$*). Then, the following inequality holds involving the performance of the empirical Bayes rule* $\hat{g}(\cdot)$, *and the Bayes error in* (2):

$$\mathbb{P}(\{u\in\Omega:\hat{g}_{\hat{P}_{X,Y}}(X(u))\neq Y(u)\})-L_{\mathcal{X}}\leq\Delta g_{MAP}(\hat{P}_{X,Y}), \tag{5}$$

where

$$\Delta g_{MAP}(\hat{P}_{X,Y})\equiv\sqrt{2\ln 2}\sum_{y\in\mathcal{Y}}P_Y(\{y\}) \times\sqrt{\min\{D(P_{X|Y}(\cdot|y)\|\hat{P}_{X|Y}(\cdot|y)),D(\hat{P}_{X|Y}(\cdot|y)\|P_{X|Y}(\cdot|y))\}} \tag{6}$$

and $D(\cdot\|\cdot)$ *is the Kullback–Leibler divergence (KLD)* [8] *between two probability distributions on* $(\mathcal{X},\mathcal{F}_{\mathcal{X}})$, *given by,*

$$D(P^1\|P^2)=\int_{\mathcal{X}}p^1(x)\cdot\log\frac{p^1(x)}{p^2(x)}\partial x,$$

where p^1 *and* p^2 *are the pdfs of* P^1 *and* P^2, *respectively.*[3]

Note that $\Delta g_{MAP}(\hat{P}_{X,Y})$ is the P_Y-average of a non-decreasing function of the KLD between the conditional class probabilities and their empirical counterparts. The KLD has a well known interpretation as a statistical discrimination measure between two probabilistic models [8,24,9], however in this case, it is an indicator of the performance deviation, relative to the fundamental performance bound, as a consequence of the statistical mismatch occurring in estimating the class-conditional probabilities. Vasconcelos has proved this result for the case when the classes are equally likely [18, Theorem 4]. The proof of Theorem 1 is a simple extension of that and not reported here for space considerations.

Remark 1. A necessary condition for $\Delta g_{MAP}(\hat{P}_{X,Y})$ to be well defined is that the empirical conditional class distributions are absolute continuous with respect to the other associated distributions [9,24], see (6). This assumption is not unreasonable because the empirical joint distribution is induced by i.i.d. realizations of the true distribution. As a result, it is assumed for the rest of the paper.

The next result shows an implication of Theorem 1 for the case when the observation random variable X takes values in a finite alphabet set (or a quantizations of $\mathcal{X}$), denoted by $\mathcal{A}_X$.

Corollary 1. *Let* (X,Y) *be a random vector taking values in the finite product space* $\mathcal{A}_X\times\mathcal{Y}$, *with* $P_{X,Y}$ *and* $\hat{P}_{X,Y}$ *being the probability and the empirical probability, respectively. Assuming that* $P_{X,Y}$ *and* $\hat{P}_{X,Y}$ *only differ in their class-conditional probabilities, then* (5) *and* (6) *hold, where* $D(P_{X|Y}(\cdot|y)\|\hat{P}_{X|Y}(\cdot|y))$ *in this context denotes the discrete version of the KLD* [8,24] *given by*:

$$D(P_{X|Y}(\cdot|y)\|\hat{P}_{X|Y}(\cdot|y))=\sum_{x\in\mathcal{A}_X}P_{X|Y}(x|y)\cdot\left(\log\frac{P_{X|Y}(x|y)}{\hat{P}_{X|Y}(x|y)}\right).$$

3.1. Tradeoff between Bayes and the estimation error

The following result introduces aspects of signal representation into the classification problem. Before that, we need to introduce the notion of an embedded space sequence, which provides a sort of order relationship among a family of feature observation spaces, and the notion of consistent probability measures associated with an embedded space sequence.

Definition 1. Let $\{\mathbb{F}_i(\cdot):i=1,\ldots,n\}$ be a family of measurable transformations from the same domain $(\mathcal{X},\mathcal{F}_{\mathcal{X}})$ and taking values in $\{\mathcal{X}_1,\ldots,\mathcal{X}_n\}$, where $\{\mathcal{X}_1,\ldots,\mathcal{X}_n\}$ has increasing finite dimensionality, i.e., $dim(X_i)<dim(X_{i+1}),\forall\, i\in\{1,\ldots,n-1\}$. We say that $\{\mathbb{F}_i(\cdot):i=1,\ldots,n\}$ is dimensionally embedded if, $\forall\, i\in\{1,\ldots,n-1\}$, $\exists\pi_{i+1,i}(\cdot)$, a measurable mapping from $(\mathcal{X}_{i+1},\mathcal{F}_{i+1})$ to $(\mathcal{X}_i,\mathcal{F}_i)$,[4] such that,

$$\mathbb{F}_i(x)=\pi_{i+1,i}(\mathbb{F}_{i+1}(x)),\quad\forall\, x\in\mathcal{X}.$$

[3] We assume that the distributions are absolutely continuous with respect to the Lebesgue measure for defining the KLD using their probability density functions [24].

[4] For all practical purposes $\mathcal{X}_i$ is a finite dimensional Euclidean space and $\mathcal{F}_i$ refers to the Borel sigma field.

In this context, we also say that $\{\mathcal{X}_1,\dots,\mathcal{X}_n\}$ is dimensionally embedded with respect to $\{\mathbb{F}_i(\cdot): i=1,\dots,n\}$ and $\{\pi_{i+1,i}(\cdot): i=1,\dots,n-1\}$.

Definition 2. Let $\{\mathcal{X}_i : i=1,\dots,n\}$ be a sequence of dimensionally embedded spaces, where $\{\pi_{i+1,i}:(\mathcal{X}_{i+1},\mathcal{F}_{i+1})\to(\mathcal{X}_i,\mathcal{F}_i): i=1,\dots,n-1\}$ is the set of measurable mapping stated in Definition 1. Associated with those spaces, let us consider a probability measure $\hat{P}_i$ defined on $(\mathcal{X}_i,\mathcal{F}_i),\forall\, i\in\{1,\dots n\}$. The family of probability measures $\{\hat{P}_i : i=1,\dots,n\}$ is consistent with respect to the embedded sequence if $\forall\, i,j\in\{1,\dots n\}$, $i<j,\forall\, B\in\mathcal{F}_i$

$$\hat{P}_i(B)=\hat{P}_j(\pi_{j,i}^{-1}(B)),$$

where $\pi_{j,i}(\cdot)\equiv\pi_{j,j-1}(\pi_{j-1,j-2}(\cdots\pi_{i+1,i}(\cdot)\cdots))$.

Definition 2 is equivalent to saying that if we induce a probability measure on $(\mathcal{X}_i,\mathcal{F}_i)$ by using the measurable mapping $\pi_{j,i}(\cdot)$ and the probability measure $\hat{P}_j$ on the space $(\mathcal{X}_j,\mathcal{F}_j)$, the induced measure is equivalent to $\hat{P}_i$. Consequently, the probabilistic description of the sequence of embedded spaces is univocally characterized by the more informative probability space, $(\mathcal{X}_n,\mathcal{F}_n,\hat{P}_n)$, and the family of measurable mappings $\{\pi_{j,i}(\cdot): j>i\}$ of the embedded structure presented in Definition 1.

Theorem 2. *Let (X,Y) be the joint observation-class random variables with distribution $P_{X,Y}$ on $(\mathcal{X}\times\mathcal{Y},\sigma(\mathcal{F}_{\mathcal{X}}\times\mathcal{F}_{\mathcal{Y}}))$, where $\mathcal{X}=\mathbb{R}^K$ for some $K>0$. Let $\{\mathbb{F}_i(\cdot): i=1,\dots,n\}$ be a sequence of representation functions, with $\mathbb{F}_i(\cdot):(\mathcal{X},\mathcal{F}_{\mathcal{X}})\to(\mathcal{X}_i,\mathcal{F}_i)$, measurable $\forall\, i\in\{1,\dots,n\}$. In addition, let us assume that, $\{\mathbb{F}_i(\cdot): i=1,\dots,n\}$ is a family of dimensionally embedded transformations, satisfying $\mathbb{F}_i(\cdot)=\pi_{j,i}(\mathbb{F}_j(\cdot))$ for all $j>i$ in $\{1,\dots,n\}$. Then, considering the family of observations random variables $\{X_i=\mathbb{F}_i(X): i=1,\dots,n\}$, the Bayes error satisfies the following relationship*:

$$L_{\mathcal{X}_{i+1}}\le L_{\mathcal{X}_i},\quad \forall\, i\in\{1,\dots,n-1\}. \tag{7}$$

If in addition we have a family of empirical probability measures $\{\hat{P}_{X_i,Y}: i=1,\dots,n\}$, with $\hat{P}_{X_i,Y}$ on $(\mathcal{X}_i\times\mathcal{Y},\sigma(\mathcal{F}_i\times\mathcal{F}_{\mathcal{Y}}))$ and conditional class distribution families $\{\hat{P}_{X_i|Y}(\cdot|y): i=1,\dots,n\}$ consistent with respect to $\{\mathcal{X}_i : i=1,\dots,n\}$ $\forall\, y\in\mathcal{Y}$, then the following relationship for the estimation error applies:

$$\Delta g_{MAP}(\hat{P}_{X_i,Y})\le\Delta g_{MAP}(\hat{P}_{X_{i+1},Y}),\quad \forall\, i\in\{1,\dots,n-1\}. \tag{8}$$

This result presents a formal tradeoff between the Bayes and estimation errors by considering a family of representations of monotonically increasing complexity. In other words, by increasing complexity we improve the theoretical performance bound—-Bayes error—that we could achieve, but as a consequence of increasing the estimation error, which upper bounds the maximum deviation from the Bayes error bound, per Theorem 1. The proof of this result is presented in Appendix A.

The following corollary of Theorem 2 shows the important case when the embedded sequence of spaces is induced by coordinate projections (equivalent to a feature selection approach). In this scenario, the consistency condition of the empirical distributions can be considered natural and consequently implicitly assumed in the statement. A version of this result for coordinate projections was originally presented in [18, Theorem 5].

Corollary 2. *Let $\mathcal{X}=\mathbb{R}^K$ and the family of coordinate projections $\pi_m^K(\cdot):\mathbb{R}^K\to\mathbb{R}^m$, $m\le K$, be given by: $\pi_m^K(x_1,\dots,x_m,\dots x_K)=(x_1,\dots,x_m)$, $\forall\,(x_1,\dots,x_K)\in\mathbb{R}^K$. Let $P_{X,Y}$ and $\hat{P}_{X,Y}$ be the joint probability measure and its empirical counterpart, respectively, defined on $(\mathcal{X}\times\mathcal{Y},\sigma(\mathcal{F}_{\mathcal{X}}\times\mathcal{F}_{\mathcal{Y}}))$. Given that the coordinate projections are measurable, it is possible to induce those distributions on the sequence of embedded subspaces $\{\mathcal{X}_1,\dots,\mathcal{X}_K\}$ characterized by: $\mathcal{X}_i=\pi_i^K(\mathcal{X})$, $\forall\, i\in\{1,\dots,K\}$. Then, the Bayes estimation error tradeoff is satisfied, i.e., $L_{\mathcal{X}_{i+1}}\le L_{\mathcal{X}_i}$ and $\Delta g_{MAP}(\hat{P}_{X_{i+1},Y})\ge\Delta g_{MAP}(\hat{P}_{X_i,Y})$, $\forall\, i\in\{1,\dots,K-1\}$.*

From this corollary a natural approach to ensure that the family of empirical class conditional distributions $\{\hat{P}_{X_i|Y}(\cdot|y): i=1,\dots,n\}$ is consistent across a dimensionally embedded space sequence $\{\mathcal{X}_i : i=1,\dots,n\}$, is to constructively induce $\hat{P}_{X_i|Y}(\cdot|y)$ using the empirical distribution on the most informative representation space, $\hat{P}_{X_n|Y}(\cdot|y)$ on $(\mathcal{X}_n,\mathcal{F}_n)$, and the measurable mappings $\pi_{n,i}(\cdot),\forall\, i<n$ associated with the embedded space sequence, per Definition 1. For instance, this construction is appealing when assuming parametric class conditional distributions, like Gaussian mixture models (GMMs), and standard family of transformations, like linear operators, where inducing those distributions implies simpler operation on the parameters of $\hat{P}_{X_n|Y}(\cdot|y)$. This type of construction was considered in [18] and will be illustrated in Section 5.2.

3.2. Bayes-estimation error tradeoff: finite alphabet case (quantization)

As for Theorem 1, we also extend Theorem 2 for the case when the family of representation functions $\{\mathbb{F}_i(\cdot): i=1,\dots,n\}$ takes values in finite alphabet sets, and consequently induces quantizations of $\mathcal{X}$. In this scenario, the concept of embedded representation is better characterized by properties of the induced family of partitions. The following definition formalizes this idea.

Definition 3. Let us consider the space $(\mathcal{X}\times\mathcal{Y},\sigma(\mathcal{F}_{\mathcal{X}}\times\mathcal{F}_{\mathcal{Y}}))$ and a family of measurable functions $\{\mathbb{F}_i(\cdot): i=1,\dots,n\}$, taking values in finite alphabet sets $\{\mathcal{A}_i : i=1,\dots,n\}$, i.e., $\mathbb{F}_i(\cdot):(\mathcal{X},\mathcal{F}_{\mathcal{X}})\to(\mathcal{A}_i,2^{\mathcal{A}_i})$, with $|A_i|<\infty$. The family of representations $\{\mathbb{F}_i(\cdot): i=1,\dots,n\}$ is embedded if: $|\mathcal{A}_i|<|\mathcal{A}_{i+1}|,\forall\, i\in\{1,\dots,n-1\}$ and $\forall\, j,i\in\{1,\dots,n\}$, $j>i$, there exists a function $\pi_{j,i}(\cdot):\mathcal{A}_j\to\mathcal{A}_i$ such that

$$\mathbb{F}_i(x)=\pi_{j,i}(\mathbb{F}_j(x)),\quad \forall\, x\in\mathcal{X}.$$

Remark 2. Every representation function $\mathbb{F}_i(\cdot)$ produces a quantization of $\mathcal{X}$ by $Q_{Fi}\equiv\{\mathbb{F}^{-1}(\{a\}): a\in\mathcal{A}_i\}\subset\mathcal{F}_{\mathcal{X}}$, where the embedded condition implies that: $\forall\, i,j,\ 1\le i<j\le n$, Q_{Fj} is a refinement of Q_{Fi},[5] (notation, $Q_{Fi}\ll Q_{Fj}$), and then, $Q_{F1}\ll Q_{F2}\ll\cdots\ll Q_{Fn}$.

For the next result we also make use of the assumption of consistency for the empirical distributions across a sequence of embedded representations, which extends naturally from the continuous case presented in Definition 2.

Theorem 3. *Let (X,Y) be the joint-observation random vector and $\{\mathbb{F}_i(\cdot): i=1,\dots,n\}$ be a family of embedded representation taking values in finite alphabet sets $\{\mathcal{A}_i : i=1,\dots,n\}$. Considering the quantized observation random variables $\{X_i\equiv\mathbb{F}_i(X): i=1,\dots,n\}$ then the Bayes error satisfies: $L_{\mathcal{A}_{i+1}}\le L_{\mathcal{A}_i},\forall\, i\in\{1,\dots,n-1\}$. If in addition we have empirical probabilities $\hat{P}_{X_i,Y}$ on the family of representation spaces $\mathcal{A}_i\times\mathcal{Y}$*[6] *with conditional class probabilities, $\{\hat{P}_{X_i|Y}(\cdot|y): i=1,\dots,n\}$, consistent with respect to $\{\mathbb{F}_i(\cdot): i=1,\dots,n\},\forall\, y\in\mathcal{Y}$, then the estimation error satisfies: $\Delta g_{MAP}(\hat{P}_{X_{i+1},Y})\ge\ \Delta g_{MAP}(\hat{P}_{X_i,Y}),\forall\, i\in\{1,\dots,n-1\}$. (The proof is presented in* Appendix B*).*

Remark 3. In this context, the tradeoff is obtained as a function of the cardinality of these spaces. In particular, the cardinality is the natural choice for characterizing feature complexity, because

[5] $\overline{Q}$ is a refinement of Q if, $\forall\, A\in Q$, $\exists\overline{Q}_A\subset\overline{Q}$ such that $A=\bigcup_{B\in\overline{Q}_A}B$.

[6] In this case we consider the power set of $\mathcal{A}_i\times\mathcal{Y}$ as the sigma field, and consequently we omit it.

it is proportional to the estimation error across the embedded sequence of spaces.

The following proposition states the validity of the consistence condition for the important scenario when the empirical distribution is obtained using the maximum likelihood (ML) criterion or the well-known frequency counts [6].

Proposition 1. *For a given amount of training data, i.i.d. realizations of* (X,Y), *the ML estimator of* $P_{X_i|Y}(\cdot|y), \forall\, y \in \mathcal{Y}$ *obtained in the range of a family of finite alphabet embedded representations* $\{\mathbb{F}_i(\cdot) : i = 1,\ldots,n\}$, *per Definition3, satisfies the consistence condition stated in Definition2. (The proof is presented in Appendix* D*).*

3.3. Remarks

From the results presented in this section, having a family of embedded representations, continuous or finite alphabet version, is not enough to show the result about the evolution of the estimation error across this embedded sequence of increasing complexity, Theorems 2 and 3. The additional necessary element is to have a consistent family of empirical distributions (see proofs of the theorems for details). This last condition is clearly a function of the learning methodology used for estimating the conditional class distributions. In the original version of this result presented [18], this condition was implicit because the author considered a family of coordinate-projections for representing the embedded space sequence, where most of the learning techniques provide empirical distributions that satisfy the required consistence condition. In summary, Theorems 2 and 3 stipulate concrete conditions between a sequence of embedded spaces and a learning scheme that justify a formal tradeoff between estimation and approximation error. Complementing this observation, Appendix C shows a concrete scenario where the mentioned tradeoff is clearly illustrated, and where there are closed-form expressions for the two error terms in (2) and (6).

As explained in [18], this tradeoff formally justifies the fact that in the process of doing dimensionality-cardinality reduction, better estimation of the underlying observation-class distribution is obtained, in the KLD sense, at the expense of increasing the underlying Bayes error. In particular these results show that by constraining to a sequence of embedded representations there is one that minimizes the probability of error, and from the results presented here, the one that achieves the optimal Bayes-estimation error tradeoff (Appendix C illustrates this optimal feature solution for a concrete dimensionally embedded space sequence and learning scheme). Hence, it is natural to think that having a rich collection of feature transformations for $\mathcal{X}$, not necessarily embedded, there is one for which this tradeoff between "representation quality" and "complexity" achieves an optimal solution, a solution that is connected with the MPE-SR problem. This is the topic addressed in the next section.

4. Minimum probability of error signal representation (MPE-SR)

Let us consider again $\{(x_i,y_i) : i = 1,\ldots,N\}$ i.i.d. realizations of the observation-class (X,Y) random variables with distribution $P_{X,Y}$ on $(\mathcal{X}\times\mathcal{Y}, \sigma(\mathcal{F}_{\mathcal{X}}\times\mathcal{F}_{\mathcal{Y}}))$. In addition, let us consider a family of measurable functions $\mathbb{D}$, where any $\mathbf{f}(\cdot)\in\mathbb{D}$ is defined in $\mathcal{X}$ and takes values in a transform space $\mathcal{X}_f$. Every representation function $\mathbf{f}(\cdot)$ induces an empirical distribution $\hat{P}_{X_f,Y}$ on $(\mathcal{X}_f\times\mathcal{Y}, \sigma(\mathcal{F}_f\times\mathcal{F}_{\mathcal{Y}}))$, based on the training data and the adopted learning approach, and hence the empirical Bayes rule by

$$\hat{g}_f(x) = \arg\max_{y\in\mathcal{Y}} \hat{P}_{X_f,Y}(x,y), \quad \forall\, x \in \mathcal{X}_f. \tag{9}$$

Then, the oracle MPE-SR problem reduces to:

$$\mathbf{f}^* = \arg\max_{\mathbf{f}\in\mathbb{D}} \mathbb{E}_{X,Y}(\mathbb{1}_{\{(x,y)\in\mathcal{X}\times\mathcal{Y}:\hat{g}_f(\mathbf{f}(x))\neq y\}}(X,Y)), \tag{10}$$

where the expected value is taken with respect to the true joint distribution $P_{X,Y}$ on $(\mathcal{X}\times\mathcal{Y}, \sigma(\mathcal{F}_{\mathcal{X}}\times\mathcal{F}_{\mathcal{Y}}))$. Note that from Lemma 1, $\forall\, \mathbf{f}(\cdot)\in\mathbb{D}$, $\mathbb{E}_{X,Y}(\mathbb{1}_{\{(x,y)\in\mathcal{X}\times\mathcal{Y}:\hat{g}_f(\mathbf{f}(x))\neq y\}}(X,Y)) \geq L_{\mathcal{X}_f} \geq L_{\mathcal{X}}$, where $L_{\mathcal{X}_f}$ denotes the Bayes error associated with $\mathcal{X}_f$. Then the MPE criterion tries to find the representation framework whose performance is the closest to $L_{\mathcal{X}}$, the fundamental error bound of the problem. Using the upper bound for the risk of the empirical Bayes rule in Theorem 1, i.e.,

$$\mathbb{E}_{X,Y}(\mathbb{1}_{\{(x,y)\in\mathcal{X}\times\mathcal{Y}:\hat{g}_f(\mathbf{f}(x))\neq y\}}(X,Y)) \leq \Delta g_{MAP}(\hat{P}_{X_f,Y}) + L_{\mathcal{X}_f}, \quad \forall\, \mathbf{f}\in\mathbb{D},$$

we take the direction proposed for the structural risk minimization (SRM) principle [25], to approximate (10) with the following decision,

$$\tilde{\mathbf{f}}^* = \arg\max_{\mathbf{f}\in\mathbb{D}} \Delta g_{MAP}(\hat{P}_{X_f,Y}) + [L_{\mathcal{X}_f} - L_{\mathcal{X}}]. \tag{11}$$

Here we have introduced the normalization factor $L_{\mathcal{X}}$ to make explicit that this regularization problem implies finding the optimal tradeoff between *approximation quality*, $L_{\mathcal{X}_f} - L_{\mathcal{X}}$, and *estimation error*, $\Delta g_{MAP}(\hat{P}_{X_f,Y})$. Then, the MPE-SR is naturally formulated as a complexity regularized optimization problem whose objective function consists of a weighted combination of a fidelity criterion, reflecting the Bayes error, and a cost term, penalizing the complexity of the representation scheme. To contextualize the idea, Appendix C shows the complexity-regularized objective function in (11) in a controlled simulated scenario, where the oracle MPE-SR problem in (11) can be addressed. Note that solution to (11) is an oracle type of result, because neither the fidelity nor the cost term in (11) are available in practice — both require the knowledge of the true distribution. Appropriate approximations for the fidelity and cost terms are needed in the Bayes setting to address this problem in practice.

4.1. Approximating the MPE-SR: the cost-fidelity formulation

For approximating $\Delta g_{MAP}(\hat{P}_{X_f,Y})$, from Theorems 2 and 3 we have that this complexity indicator is proportional to the dimensionality or cardinality of the representation, respectively, and consequently a function proportional to those terms can be adopted. On the other hand for the fidelity $L_{\mathcal{X}_f}$, the first natural candidate to consider is the *empirical risk* (ER) [25,6] associated with the family of empirical Bayes rules in (9). In favor of this choice is the existence of distribution free bounds that control the uniform deviation of the ER with respect to the risk (the celebrated *Vapnik–Chervonenkis inequality* [6,25]). However, this choice of fidelity indicator raises the problem of addressing the resulting complexity regularized ER minimization, a problem that has an algorithmic solution only in very restrictive settings. In this regard, Section 5.1 presents an emblematic scenario where this problem can be efficiently solved for a family of tree-structured vector quantizations.

However, in numerous important cases, the ER is impractical because the solution to the resulting complexity regularization requires an exhaustive search. An alternative to approximating the Bayes risk $L_{\mathcal{X}_f}$ is to use some of the information theoretic quantities like the family of *Ali–Silvey* distances [26,27], formally justified for the binary hypothesis testing problem, or the widely adopted *mutual information* (MI) [11,28–30]. It is well known that MI and probability of error are connected by *Fano's inequality* and tightness has been shown asymptotically by the second *Shannon coding theorem* [9,24]. Importantly in our problem scenario, MI satisfies the same monotonic behavior under a sequence of embedded transformations as the Bayes risk; in the practical side the empirical MI, denoted by $\hat{I}(X_f,Y)$,[7] under some problem settings can offer algorithmic solutions for the resulting complexity regularized problem. One example of this is presented in following sections and another was recently shown in [31] for addressing a discriminative filter bank selection problem.

Returning to the problem, generally denoting by $\hat{I}(\mathbf{f})$ and $R(\mathbf{f})$ the approximated fidelity and cost terms for $\mathbf{f} \in \mathbb{D}$, respectively, (11) can be approximated by:

$$\mathbf{f}^*(\lambda) = \arg\max_{\mathbf{f} \in \mathbb{D}} \Psi(\hat{I}(\mathbf{f})) + \lambda \cdot \Phi(R(\mathbf{f})), \tag{12}$$

where considering the tendency of the new fidelity-cost indicators, $\Psi(\cdot)$ should be a strictly decreasing real function, $\Phi(\cdot)$, a strictly increasing function from $\mathbb{N}$ to $\mathbb{R}$ and $\lambda > 0$. Noting that the real dependency between Bayes and estimation errors in terms of our new fidelity complexity values, $\hat{I}(\mathbf{f})$ and $R(\mathbf{f})$, is hidden and, furthermore, problem dependent, then Ψ, Φ and $\lambda \in \mathbb{R}^+$ provide degrees of freedom for approximating the oracle MPE-SR in (11). Remarkable, it is important to note that independent of those degrees of freedom, (12) can be expressed by:

$$\mathbf{f}^*(\lambda) = \arg\max_{\mathbf{f} \in \{\mathbf{f}_k^* : k \in K(\mathbb{D}) t\}} \Psi(\hat{I}(\mathbf{f})) + \lambda \cdot \Phi(R(\mathbf{f})), \tag{13}$$

with $\{\mathbf{f}_k^* : k \in K(\mathbb{D})\} \subset \mathbb{D}$ the solutions of the following *cost-fidelity problem*:

$$\mathbf{f}_k^* = \arg\max_{\substack{\mathbf{f} \in \mathbb{D} \\ R(\mathbf{f}) \le k}} \hat{I}(\mathbf{f}), \tag{14}$$

$\forall\, k \in K(\mathbb{D})$, where $K(\mathbb{D}) \equiv \left\{R(\mathbf{f}) : \mathbf{f} \in \mathbb{D}\right\} \subset \mathbb{N}$. Then, the approximated MPE-SR solution in (12) can be restricted, without any loss, to what we call the *optimal achievable cost-fidelity family* $\{\mathbf{f}_k^* : k \in K(\mathbb{D})\}$. Note that the cardinality of $K(\mathbb{D})$ could be significantly smaller than $|\mathbb{D}|$ and as a result, the domain of solutions of the original problem. Finally, the empirical risk minimization criterion among $\{\mathbf{f}_k^* : k \in K(\mathbb{D})\} \subset \mathbb{D}$, for instance using cross validation [6,5], can be the final decision step for solving (13), as it has been widely used for addressing a similar complexity regularization problem in the context of regression and classification trees [5,32].

5. Applications

By stipulating a learning technique and a family of representation functions, we can obtain instances of the MPE-SR complexity regularized formulation. Two important cases where this particularization offers algorithmic solutions are presented in this section, while another interesting case was presented by the authors for the problem of discriminative Wavelet Packet filter bank selection in [31].

5.1. Classification trees: CART and minimum cost tree pruning algorithms

Let $\mathcal{X} = \mathbb{R}^K$ be a finite dimension Euclidean space, and let $\mathbb{D}$ be a family of vector quantizers (VQs) with a binary tree structure [33–35]. Then the MPE-SR problem reduces to finding an optimal binary classification tree topology. Interestingly the pruning tree algorithms proposed by Breiman, Friedman, Olshen and Stone (BFOS) [5][8] can be shown to address an instance of the complexity regularized problem presented in (12)

Let us introduce some basic terminology.[9] Using Breiman et al. conventions [5], a tree T is represented by a collection of nodes in a graph, with implicit left and right mappings reflecting the parent–child relationship among them. T is a *rooted binary tree* if every nonterminal node has two descendants, where we denote by $\mathcal{L}(T) \subset T$ the sub-collection of leaves or terminal nodes (nodes that do not have a descendent). In addition, $|T|$ denotes the norm of the tree, which is the cardinality of $\mathcal{L}(T)$. Let S and T be two binary tress, then if $S \subset T$ and both have the same root node, we say that S is a pruned version of T, and we denote this relationship by $S \ll T$. Finally, we denote by $\mathbf{T}_{full}$ the tree formed for all the nodes in the graph and by t_{root} its root.

The tree structure is used to index a family of vector quantizations for $\mathcal{X}$. In order to formalize this idea, we can consider that every node $t \in \mathbf{T}_{full}$ has associated a measurable subset $\mathcal{X}_t \subset \mathcal{X}$, such that $\mathcal{X}_{t_{root}} = \mathcal{X}$ and if t_1 and t_2 are the direct descendants of a nonterminal node t, we then have that: $\mathcal{X}_t = \mathcal{X}_{t_1} \cup \mathcal{X}_{t_2}$ and $\mathcal{X}_{t_1} \cap \mathcal{X}_{t_2} = \phi$. Therefore, any rooted binary tree $T \ll \mathbf{T}_{full}$ induces a measurable partition of the observation space given by $\mathcal{V}_T = \{\mathcal{X}_t : t \in \mathcal{L}(T)\}$, where importantly, if $T_1 \ll T_2$ then $\mathcal{V}_{T_2}$ is a refinement of $\mathcal{V}_{T_1}$. With this concept in mind, we can define a pair of tree indexed representations by $[T, \mathbf{f}_T(\cdot)]$ for all $T \ll \mathbf{T}_{full}$, with $\mathbf{f}_T(\cdot)$ being a measurable function from $(\mathcal{X}, \mathcal{F}_{\mathcal{X}})$ to $\mathcal{L}(T)$, such that $\mathcal{X}_t = \mathbf{f}_T^{-1}(\{t\}), \forall\, t \in \mathcal{L}(T)$. Hence, $\mathbf{f}_T(\cdot)$ induces the previously defined measurable partition $\mathcal{V}_T$ on $(\mathcal{X}, \mathcal{F}_{\mathcal{X}})$. Then formally, the family of tree indexed representations is given by $\mathbb{D} = \{\mathbf{f}_T(\cdot) : T \ll \mathbf{T}_{full}\}$.

Following the convention in [32], a classification tree is a triple $[T, \mathbf{f}_T(\cdot), g_T(\cdot)]$, where $g_T(\cdot)$ from $\mathcal{L}(T)$ to $\mathcal{Y}$ is the classifier that infers Y based on the quantized random variable $X_{\mathbf{f}_T} \equiv \mathbf{f}_T(X)$. In particular the Bayes rule is

$$g_T(t) = \arg\max_{y \in \mathcal{Y}} P_{X_{\mathbf{f}_T},Y}(t,y), \quad \forall\, t \in \mathcal{L}(T),$$

with Bayes error $R(T) = \mathbb{P}(\{u \in \Omega : g_T(X_{\mathbf{f}_T}(u))) \neq Y(u)\})$. Breiman et al. [5, Chapter 9] show that $R(T)$ can be written as an additive non-negative function of the terminal nodes of T,

$$R(T) = \sum_{t \in \mathcal{L}(T)} R(t), \tag{15}$$

where for the 0–1 cost function $R(t) = \mathbb{P}(X_{\mathbf{f}_T}(u) = t) \cdot (1 - \max_{y \in \mathcal{Y}} P_{Y|X_{\mathbf{f}_T}}(y|t))$. As we know if $P_{X,Y}$ is available, the best performance is obtained for the finest representation, i.e., $[\mathbf{T}_{full}, \mathbf{f}_{\mathbf{T}_{full}}(\cdot), g_{\mathbf{T}_{full}}(\cdot)]$; however, our case of interest is when under the constraint of finite i.i.d. samples of (X,Y), $D_N = \{(x_i,y_i) : i = 1, \ldots, N\}$, we want to address the MPE-SR problem formulated in Section 4. In this case, the maximum likelihood (ML) empirical distribution $\hat{P}_{X_{\mathbf{f}_T},Y}$ is considered for all $T \ll \mathbf{T}_{full}$, which reduces the problem to a family of classification trees $[T, \mathbf{f}_T(\cdot), \hat{g}_T(\cdot)]$ with the empirical Bayes decision $\hat{g}_T(\cdot)$ corresponding to the *majority vote decision rule* [5]. The next result shows that the Bayes-estimation error tradeoff holds for a sequence of embedded representations in $\mathbb{D}$.

Proposition 2. *Let us take a sequence of embedded trees $T_1 \ll T_2 \ll T_3, \ldots, \ll T_k$, subsets of $\mathbf{T}_{full}$. Then, $R(T_{i+1}) \le R(T_i)$, for all $i \in \{1, \ldots, n-1\}$. On the other hand, considering $[T_i, \mathbf{f}_{T_i}(\cdot), \hat{g}_{T_i}(\cdot)]$, for all $i \in \{1, \ldots, k\}$ and the estimation error of these empirical Bayes decisions, denoted by $\Delta g(\hat{P}_{X_{\mathbf{f}_{T_i}},Y})$ for all $i \in \{1, \ldots, k\}$ (Theorem 1), it follows that $\Delta g(\hat{P}_{X_{\mathbf{f}_{T_i}},Y}) \le \Delta g(\hat{P}_{X_{\mathbf{f}_{T_{i+1}}},Y})$, for all $i \in \{1, \ldots, n-1\}$.*

[7] $\forall\, f \in \mathbb{D}$ the empirical MI $\hat{I}(X_f,Y)$ can be obtained from the available empirical distribution $\hat{P}_{X_f,Y}$.

[8] The seminal work of Breiman et al. [5] addresses the more general case of classification and regression trees (CART), where for the context of this work we just highlight results concerning the classification part.

[9] The interested reader is referred to [5,36] for a more systematic exposition.

Proof. We have developed all the machinery to prove this result. We know that the family of representations $\{\mathbf{f}_{T_1}(\cdot),\ldots,\mathbf{f}_{T_n}(\cdot)\}$ is embedded, where by Proposition 1 the induced empirical distributions—conditional class probabilities—are consistent with respect to the embedded representation family. Consequently, the result extends directly from Theorem 3. □

This result provides a strong justification to use the complexity regularized approximation in (12) for the MPE-SR problem. For that we can use the additive structure of the Bayes error in (15), to consider the *empirical Bayes error* and cardinality as the fidelity and complexity indicators for (12), which in fact are additive tree functionals [33,37]. In fact under an additive cost assumption for the term ϕ in (12), the solution to this problem is the well known CART pruning algorithm [5], which finds an algorithmic solution for,[10]

$$T_n^*(\alpha) = \arg\max_{T \ll \mathbf{T}_{full}} \hat{R}(T) + \alpha \cdot |T|, \tag{16}$$

with $\hat{R}(T) = (1/N)\sum_{i=1}^{N} \mathbb{I}_{\{(t,y) \in \mathcal{L}_T \times \mathcal{Y}: \hat{g}_T(t) \neq y\}}(\mathbf{f}_T(x_i), y_i)$ the empirical risk. More precisely, Breiman et al. [5, Chapter 10] use the additivity of the objective function in (16) to formulate a dynamic programming solution for $T_n^*(\alpha)$ in $O(|\mathbf{T}_{full}|)$. Moreover, they proved that there exists a sequence of optimal embedded representations, denoted by $\mathbf{T}_{full} = T_1^* \gg T_2^* \gg, \ldots, \gg T_m^* = \{t_{root}\}$, which are the solutions of (16) for all possible values of the complexity weight $\alpha \in \mathbb{R}^+$. More precisely, $\exists \alpha_0 = 0 < \alpha_1 <, \ldots, < \alpha_m = \infty$, and for all $i \in \{1,\ldots,m\}$, such that,

$$T_n^*(\alpha) = T_i^*, \quad \forall\, \alpha \in [\alpha_{i-1}, \alpha_i). \tag{17}$$

Note that this result connects the MPE-SR tree pruning problem with the solutions for our cost-fidelity problem, as Scott [37] had recently pointed out. The reason is that this family of optimal embedded sequences is the solution to the cost-fidelity problem in (14), which is expressed in this context by:

$$T_j^* = \arg\max_{\substack{T \ll \mathbf{T}_{full} \\ |T| \leq m-j-1}} \hat{R}(T), \quad \forall\, j \in \{1,\ldots,m\}. \tag{18}$$

Scott coined the solution of (18) as the *minimum cost trees*, and has presented a general algorithm to solve it in $O(|\mathbf{T}_{full}|^2)$ [37]. Also the connection of this cost-fidelity problem with a more general complexity regularized objective criterion was presented, where the cost term $\alpha \cdot |T|$ is substituted for a general sized-based penalty $\alpha \cdot \Phi(|T|)$, where $\Phi(\cdot)$ is a non-decreasing function. Moreover an algorithm based on the characterization of the operational cost-fidelity boundary, was presented for finding explicitly $\alpha_0 < \alpha_1 <, \ldots, < \alpha_m$ as in (17). Scott's work is the first one that formally presented the connections between the general CART complexity regularized pruning problem and the solution of a cost-fidelity problem. Here we provide the context to show that the algorithms used to solve the cost-fidelity problem are implicitly addressing the ultimate MPE-SR problem.

5.2. Linear discriminant analysis

Let us consider again $\mathcal{X} = \mathbb{R}^K$ and the family of linear transformations as the dictionary:

$$\mathbb{D} = \{\mathbf{f} : \mathbb{R}^K \to \mathbb{R}^m : \mathbf{f}\ \textit{linear}, m \leq K\}.$$

An element $\mathbf{f} \in \mathbb{D}$ can be univocally represented by a matrix $\mathbf{A} \in \mathbb{R}(m,K)$.[11] In particular, we restrict $\mathbb{D}$ to the family of full-rank matrices. If we consider that the conditional class probability follows a multivariate Gaussian distribution, then $p_{X|Y}(\cdot|y) = \mathcal{N}(\cdot,\mu_y,\Sigma_y)$ and $p_X(\cdot) = \sum_{y \in \mathcal{Y}} \mathbb{P}(Y(u)=y) \cdot \mathcal{N}(\cdot,\mu_y,\Sigma_y)$, where $\mathcal{N}(\cdot,\mu,\Sigma)$ is a Gaussian pdf with mean μ and covariance matrix Σ.

Consider a finite amount of training data $\{(x_i,y_i) : i = 1,\ldots,N\}$ and maximum likelihood (ML) estimation techniques [4], and that the empirical distributions $\{\hat{p}_{X|Y}(\cdot|y) : y \in \mathcal{Y}\}$ and $\hat{p}_X(\cdot)$ are Gaussian and Gaussian mixtures, respectively, characterized by the empirical mean and covariance matrices:

$$\hat{\mu}_y = \frac{1}{N_y}\sum_{i=1}^{N} \mathbb{I}_{\{y\}}(y_i) \cdot x_i \quad \text{and} \quad \hat{\Sigma}_y = \frac{1}{N_y}\sum_{i=1}^{N} \mathbb{I}_{\{y\}}(y_i) \cdot (x_i - \hat{\mu}_y)(x_i - \hat{\mu}_y)^\dagger,$$

with $N_y = |\{1 \leq i \leq N : y_i = y\}|, \forall\, y \in \mathcal{Y}$.

Proposition 3. *Let* $\mathbf{A}_1,\ldots,\mathbf{A}_n$ *be a family of full-rank linear transformations taking values in* $\{\mathbb{R}^{k1},\ldots,\mathbb{R}^{kn}\}$ *with* $0 < k1 < k2 < \cdots < kn \leq K$. *In addition, let us assume that the sequence of transformations is dimensionally embedded, per Definition1, i.e.,* $\forall\, j,i,j > i$ *there exists* $B_{j,i} \in \mathbb{R}(ki,kj)$, *such that* $\mathbf{A}_i = B_{j,i} \cdot \mathbf{A}_j$. *Under the Gaussian parametric assumption the empirical sequence of class conditional pdfs* $\{\hat{p}_{\mathbf{A}_i X|Y}(\cdot|y) : i = 1,\ldots,n\}$, *estimated across* $\{\mathbb{R}^{k1},\ldots,\mathbb{R}^{kn}\}$ *(ML criterion), characterize a sequence of consistent probability measures with respect to* $\{\mathbb{R}^{k1},\ldots,\mathbb{R}^{kn}\}$, *in the sense presented in Definition2.*

Proof. Proof provided in Appendix D.

This last result formally extends Theorem 2, for the case of embedded sequences of full-rank linear transformations $\mathbf{A}_1,\ldots,\mathbf{A}_n$. and provides justification for addressing the MPE-SR problem using the cost-fidelity approach. In this context, as considered by Padmanabhan et al. [11] the empirical mutual information is used as the objective indicator. Then, the solution of the MPE-SR problem resides in the solution of:

$$\mathbf{A}_k^* = \arg\max_{\mathbf{A} \in \mathbb{R}(k,K)} \hat{I}(\mathbf{A}), \quad \forall\, k \in \{1,\ldots,K\}, \tag{19}$$

where $\hat{I}(\mathbf{A})$ denotes the empirical mutual information between $\mathbf{A}X$ and Y. Let us write $I(\mathbf{A}) = H(\mathbf{A}X) - H(\mathbf{A}X|Y)$ [24,9]. Then under the Gaussian assumption and considering $\mathbf{A} \in \mathbb{R}(k,K)$, it follows that, $H(\mathbf{A}X|Y=y) = (k/2)\log(2\pi) + \frac{1}{2}\log(|\mathbf{A}\Sigma_y\mathbf{A}^\dagger|) + \frac{1}{2}$ [9]. Given that $\mathbf{A}X$ has a Gaussian mixture distribution, a closed-form expression is not available for the differential entropy. Padmanabhan et al. [11] proposed to use an upper bound based on the well known fact that the Gaussian law maximizes the differential entropy under second moment constraints [9]. Then, denoting $\Sigma \equiv \mathbb{E}(XX^\dagger) - \mathbb{E}(X)\mathbb{E}(X)^\dagger$, we have that $H(\mathbf{A}X) \leq (k/2)\log(2\pi) + \frac{1}{2}\log\,(|\mathbf{A}\Sigma\mathbf{A}^\dagger|) + \frac{1}{2}$ and then

$$I(\mathbf{A}) \leq \frac{1}{2}\log\left[\frac{|\mathbf{A}\Sigma\mathbf{A}^\dagger|}{\prod_{y \in \mathcal{Y}} |\mathbf{A}\Sigma_y\mathbf{A}^\dagger|^{\mathbb{P}(Y(u)=y)}}\right].$$

Using this bound the cost-fidelity problem in (19) reduces to

$$\mathbf{A}_k^* = \arg\max_{\mathbf{A} \in \mathbb{R}(k,K)} \log\left[\frac{|\mathbf{A}\hat{\Sigma}\mathbf{A}^\dagger|}{\prod_{y \in \mathcal{Y}} |\mathbf{A}\hat{\Sigma}_y\mathbf{A}^\dagger|^{\mathbb{P}(Y=y)}}\right], \tag{20}$$

where $\hat{\Sigma}_y$ and $\hat{\Sigma}$ are the empirical class conditional covariance matrices and the unconditional covariance matrices, respectively. $\hat{\Sigma}$ can be written as $\hat{\Sigma}_w + \hat{\Sigma}_b$ [11], with $\hat{\Sigma}_w = \sum_{y \in \mathcal{Y}} \hat{P}_Y(\{y\}) \cdot \hat{\Sigma}_y$ and $\hat{\Sigma}_b = \sum_{y \in \mathcal{Y}} \hat{P}_Y(\{y\}) \cdot (\hat{\mu} - \hat{\mu}_y)(\hat{\mu} - \hat{\mu}_y)^\dagger$ the between-class and within-class scatter matrices used in linear discriminant analysis [4]. As pointed out in [11], under the additional assumption that class conditional covariance matrices are equivalent, the problem reduces to

$$\mathbf{A}_k^* = \arg\max_{\mathbf{A} \in \mathbb{R}(k,K)} \log\left[\frac{|\mathbf{A}\hat{\Sigma}\mathbf{A}^\dagger|}{|\mathbf{A}\hat{\Sigma}_y\mathbf{A}^\dagger|}\right],$$

[10] From this point we consider the tree index T for referring to the representation function $\mathbf{f}_T(\cdot)$ and the empirical Bayes classifier $[T, \mathbf{f}_T(\cdot), \hat{g}_T(\cdot)]$, depending on the context.

[11] $\mathbb{R}(m,n)$ represents the collection of $m \times n$ matrices with entries in $\mathbb{R}$.

which is exactly the objective function used for finding the optimal linear transformation used in multiple discriminant analysis (MDA), the case $k=1$ being the Fisher linear discriminant analysis problem [4].

6. Final discussion and connections with structural risk minimization

The MPE-SR formulation presents some interesting conceptual connection with other complexity regularization formulation developed in statistics and pattern recognition like the *minimum description length* (MDL) [38], *Bayesian information criterion* (BIC), *Akaike information criterion*, and *structural risk minimization* (SRM). All these learning principles also reduce to a complexity regularization problem, although they focus on model selection or rule selection, rather than on the signal representation, which was the focus of this work. To the best of our knowledge these techniques do not have a counterpart in the problem addressed in this work, however, they are interesting connections between them. For completeness, here we provide some analogies with the well-known SRM principle (a comprehensive treatment can be found in [39,25,6] and a good survey emphasizing results in pattern classification can be found in [40]). Similar connections can be stipulated with the other mentioned methods (an excellent exposition can be found in [38, Chapters 17.3 and 17.10]).

The *empirical risk minimization* (ERM) principle considers a class $\mathbb{C}$ of classifiers—a subset of the set of measurable functions from $\mathcal{X}$ to $\mathcal{Y}$—and naturally formalizes the learning problem as finding the decision rule $g(\cdot)$ in $\mathbb{C}$ that minimizes the empirical risk $\hat{R}(g)$ in (21), based on a finite amount of training data, $\{(x_i,y_i) : i=1,\ldots,N\}$: i.i.d. realization of the observation and class random phenomenon $(X(u),Y(u))$ with values in $\mathcal{X}\times\mathcal{Y}$:

$$\hat{R}(g) \equiv \frac{1}{N}\cdot\sum_{i=1}^{N} \mathbb{I}_{\{g(x_i)\neq y_i\}}. \tag{21}$$

In this formulation the observation feature space $\mathcal{X}$ is fixed and the learning problem is given by:

$$\hat{g}_N^*(\cdot) \equiv \arg\max_{g\in\mathbb{C}} \hat{R}(g). \tag{22}$$

Formal results have been derived to show consistency of the learning principle [6] as the number of samples tends to infinity. Furthermore, uniform bounds in $\mathbb{C}$ for the rate of convergence of empirical risk $\hat{R}(g)$ to the expected risk $R(g)\equiv\mathbb{P}(\{g(X(u))\neq Y(u)\})$ have been derived as a function of a combinatorial notion of the complexity for $\mathbb{C}$, the *Vapnik–Chervonenkis* (*VC*) *dimension* of the class [41,42]. The notion of VC dimension is particularly crucial in the development of this theory because it allows controlling the generalization ability of the learning principle, i.e., how far $R(\hat{g}_N^*)$ can be from the actual minimal risk decision in $\mathbb{C}$, $\inf_{g\in\mathbb{C}} R(g)$, independent of the joint distribution of $(X(u),Y(u))$. This is particularly crucial when dealing with sample sizes which are small relative to the VC dimension of the class $\mathbb{C}$, and consequently the gap between empirical and average risk turns out to be significant. This last scenario presents the formal justification to address the learning problem as a complexity regularized optimization, where in one hand, we have a fidelity function (empirical risk) and on the other, some notion of complexity associated with the estimation error. This complexity notion is explicitly a function of the VC dimension [39,25]. Then, the *structural risk minimization* (SRM) principle was aimed at formalizing this regularization problem. This was proposed to find the optimal tradeoff between *fidelity* and *complexity* for a given amount of training data in a sequence of classifier families, $\mathbb{C}_1\subset\mathbb{C}_2\subset\cdots$, with a structure of increasing VC dimensions.

At this point it is interesting to discuss some natural analogies and differences with the formulation of the MPE-SR problem. First, the MPE-SR makes uses of the Bayes decision approach as a way to define the optimal decision rule based on estimated empirical distributions and the plug-in Bayes rules. Second, the domain to address the MPE learning problem is with respect to a family of feature representations. On the other hand the SRM uses the ERM as learning principle in (22); and the degree of freedom is on the family of classifiers, i.e., $\mathbb{C}$. Concerning similarities, as in the SRM the MPE-SR approach provides an upper bound for measuring the deviation of the empirical Bayes rule with respect to the Bayes rule. More precisely, the deviation of the risk of the empirical Bayes rule with respect to the Bayes error bound—estimation error—is a function of the average KLD between the involved class conditional distributions, see Section 3. In this respect, it is not possible to characterize a universal closed-form expression as the one obtained in ERM theory. However, under some parametric assumption like that of multivariate Gaussian distribution, generalized Gaussian distribution, or Gaussian mixture models, KLD closed-form or KLD upper bound closed-form expressions are available [9,43]. This allows us to find distribution dependent bounds to analyze the rate of convergence of the estimation error as a function of the number of training points and the dimensionality of the feature space. This issue is interesting to address, in particular considering the fact that this family of parametric models has been used extensively under the Bayes decision approach [4]. On the other hand, in the MPE-SR the notion of complexity is directly associated with common engineering indicators, cardinality and dimensionality of the feature space depending on the family of representations considered in the problem. This is not the case for the ERM principle where the VC dimension is an abstract concept and potentially difficult to characterize for a given family of classifiers.

Results that present the tradeoff between Bayes error and estimation error across sequences of embedded representation presented in Section 3.1, and the motivation to address the MPE-SR as a complexity regularized criterion have equivalent counterpart in the SRM inductive principle. This explains why the two frameworks address similar tradeoffs between complexity and fidelity for finding the minimum probability of error decisions. Regarding practical implementation, the MPE-SR needs to address the complexity regularized optimization problem. Given the empirical probability of error as fidelity, in most of the cases this indicator does not have any closed-form expression. Empirical mutual information turns to be a naturally attractive candidate in particular considering some family of parametric models and potential embedded structure of feature representations, which is the motivation of the formulation presented in Section 4. In this direction, this paper presents two emblematic learning scenarios that show how this principle can be practically implemented: one under some parametric assumptions for the case of a finite dimensional feature family (Section 5.2), and the other considering a family of vector quantizations with a strong tree-embedded structure, where the induced combinatorial problem can be solved using dynamic programming (DP) techniques (Section 5.1). The SRM inductive principle, on the other hand, has the same issues for addressing the empirical risk minimization problem in (22). In this case some approximations based on discriminant cost functions have been considered which allow to address the optimization problem for particular families of classifiers. Examples of those practical learning frameworks are Boosting techniques, neural networks and support vector machines (SVM) [40,25].

7. Future work

It is important to emphasize that the access to the true estimation and approximation expression in the oracle MPE-SR problem in (11), $\Delta g_{MAP}(\cdot)$ and $L_{\mathcal{X}}$, is not possible, as they require the true distribution. Hence, practical approximations are strictly needed to reduce the oracle MPE-SR to problem that could be addressed as a concrete algorithm. In particular, the choice of mutual information to approximate the Bayes error is natural and offers interesting connections with practical schemes as presented in Section 5.2. However, we believe that much more can be done in this direction to understand and shrink the gap between the oracle MPE-SR and practical solutions based on approximated fidelity and cost measures. This motivates the direction of evaluating in much more detail the estimation of $\Delta g_{MAP}(\cdot)$ and $L_{\mathcal{X}}$, based on empirical data, and with that propose better expressions to address in practice the MPE-SR problem.

Acknowledgments

The authors are grateful to Antonio Ortega for helpful discussions and Joseph Tepperman for proof reading this material. We thanks the reviewers for their insightful comments that have helped us to improve the content and organization of this work. In particular, the simulated scenario in Appendix C was proposed by one the reviewer. The work of J. Silva was supported by funding from FONDECYT Grants 1090138 and 1110145, CONICYT-Chile. The work of S.S. Narayanan was supported by funding from the National Science Foundation (NSF).

Appendix A. Proof of Theorem 2: Bayes-estimation error tradeoff

Given that $\{\mathbb{F}_i(\cdot) : i=1,\ldots,n\}$ is a sequence of embedded transformations, i.e., $\forall\, i \in \{1,\ldots,n-1\}$, there exists a measurable mapping $\pi_{i+1,i} : \mathcal{X}_{i+1} \to \mathcal{X}_i$ such that $X_i = \pi_{i+1,i}(X_{i+1})$, then from Lemma 1 we have that $L_{\mathcal{X}_{i+1}} \le L_{\mathcal{X}_i}$. Concerning the estimation error inequality across the sequence of embedded spaces, a sufficient condition given in Theorem 1, is to prove that

$$D_{(\mathcal{X}_i,\mathcal{F}_i)}(P_{X_i|Y}(\cdot|y)\|\hat{P}_{X_i|Y}(\cdot|y)) \le D_{(\mathcal{X}_{i+1},\mathcal{F}_{i+1})}(P_{X_{i+1}|Y}(\cdot|y)\|\hat{P}_{X_{i+1}|Y}(\cdot|y)),$$

$$D_{(\mathcal{X}_i,\mathcal{F}_i)}(\hat{P}_{X_i|Y}(\cdot|y)\|P_{X_i|Y}(\cdot|y)) \le D_{(\mathcal{X}_{i+1},\mathcal{F}_{i+1})}(\hat{P}_{X_{i+1}|Y}(\cdot|y)\|P_{X_{i+1}|Y}(\cdot|y)), \tag{23}$$

$\forall\, i \in \{1,\ldots,n-1\}$ and $\forall\, y \in \mathcal{Y}$. We will focus on proving the first set of inequalities in (23), and the same argument can be applied for the other family. Here $D_{(\mathcal{X}_i,\mathcal{F}_i)}(P_{X_i|Y}(\cdot|y)\|\hat{P}_{X_i|Y}(\cdot|y))$ denotes the KLD of the conditional class probability $P_{X_i|Y}(\cdot|y)$ with respect to the empirical counterpart in $(\mathcal{X}_i,\mathcal{F}_i)$. The fact of considering the dependency with respect to the measurable space in the KLD notation, which is usually implicit, is conceptually important for the rest of the proof.

The main idea is to represent the empirical distribution as an underlying measure defined on the original measurable space $(\mathcal{X},\mathcal{F}_{\mathcal{X}})$. This is possible using the fact that the functions in $\{\mathbb{F}_i(\cdot) : i=1,\ldots,n\}$ are measurable [21]. Consequently given the empirical class conditional probability $\hat{P}_{X_i|Y}(\cdot|y)$ in the representation space $(\mathcal{X}_i,\mathcal{F}_i)$, we can induce a probability measure $\hat{P}_{X|Y}(\cdot|y)$ in the measurable space $(\mathcal{X},\sigma(\mathbb{F}_i))$, with $\sigma(\mathbb{F}_i)$ is the smallest sigma field that makes $\mathbb{F}_i(\cdot)$ a measurable transformation [20], where it is clear that $\sigma(\mathbb{F}_i) \subset \mathcal{F}_{\mathcal{X}}$ [20]. More precisely, $\sigma(\mathbb{F}_i) = \{\mathbb{F}_i^{-1}(B) : B \in \mathcal{F}_{\mathcal{X}}\}$ and $\hat{P}_{X|Y}(\cdot|y)$ is constructed by

$$\forall\, A \in \sigma(\mathbb{F}_i), \exists B \in \mathcal{F}_i, \quad st.\ A = \mathbb{F}_i^{-1}(B) \quad \text{and} \quad \hat{P}_{X|Y}(A|y) = \hat{P}_{X_i|Y}(B|y). \tag{24}$$

By the consistence property of $\left\{\hat{P}_{X_i|Y}(\cdot|y) : i=1,\ldots,n\right\}$, it is easy to show that there is a unique measure $\hat{P}_{X|Y}(\cdot|y)$ defined on $(X,\sigma(\mathbb{F}_n))$ that represents the family of empirical distributions $\{\hat{P}_{X_i|Y}(\cdot|y) : i=1,\ldots,n\}$ using the procedure presented in (24).[12] As a consequence, the empirical measure $\hat{P}_{X|Y}(\cdot|y)$ is uniquely characterized in $\mathcal{X}$ using the finest sigma field $\sigma(\mathbb{F}_n)$. On the other hand, the original probability measure $P_{X|Y}(\cdot|y)$ is originally defined on $(\mathcal{X},\mathcal{F}_{\mathcal{X}})$ and given that $\sigma(\mathbb{F}_n) \subset \mathcal{F}_{\mathcal{X}}$, it extends naturally to $(X,\sigma(\mathbb{F}_i)), \forall\, i \in \{1,\ldots,n\}$.

The next step is to represent the KLD in (23), as a KLD in the original observation space $\mathcal{X}$ relative to a particular sigma field. Using a classical result from measure theory [20], it is possible to prove that [24, Lemma 5.2.4]

$$D_{(\mathcal{X},\sigma(\mathbb{F}_i))}(P_{X|Y}(\cdot|y)\|\hat{P}_{X|Y}(\cdot|y)) = D_{(\mathcal{X}_i,\mathcal{F}_i)}(P_{X_i|Y}(\cdot|y)\|\hat{P}_{X_i|Y}(\cdot|y)). \tag{25}$$

Finally from proving (23), we make use of the following lemma.

Lemma 2 (*Gray [24, Lemma 5.2.5]*). *Let us consider two measurable spaces $(\mathcal{X},\mathcal{F})$ and $(\mathcal{X},\overline{\mathcal{F}})$, such that $\overline{\mathcal{F}}$ is a refinement of $\mathcal{F}$, in other words $\mathcal{F} \subset \overline{\mathcal{F}}$. In addition, let us consider two probability measures P_1 and P_2 defined on $(\mathcal{X},\overline{\mathcal{F}})$, then assuming that $P_1 \ll P_2$, the following inequality holds*:

$$D_{(\mathcal{X},\overline{\mathcal{F}})}(P_1\|P_2) \ge D_{(\mathcal{X},\mathcal{F})}(P_1\|P_2). \tag{26}$$

In our context we have $P_{X|Y}(\cdot|y)$ and $\hat{P}_{X|Y}(\cdot|y)$ defined on $(\mathcal{X},\sigma(\mathbb{F}_{i+1}))$ and consequently on $(\mathcal{X},\sigma(\mathbb{F}_i))$, because $\sigma(\mathbb{F}_{i+1})$ is a refinement of $\sigma(\mathbb{F}_i)$, then (23) follows directly from Lemma 2. □

Appendix B. Proof of the Bayes-estimation error tradeoff: finite alphabet case

This proof follows the same arguments as the one presented in Appendix A. Let us denote the Bayes rule for (X_i,Y) by $g_{P_{X_i,Y}}(\cdot)$ with error probability $L_{\mathcal{A}_i}$, given by $L_{\mathcal{A}_i} = P_{X_i,Y}(\{(x,y) \in \mathcal{A}_i \times \mathcal{Y} : g_{P_{X_i,Y}}(x) \ne y\}), \forall\, i \in \{1,\ldots,n\}$.

By the assumption that the family $\{\mathbb{F}_i(\cdot) : i=1,\ldots,n\}$ is embedded, we have that $\forall\, 0 \le i < j \le n$, $X_i \equiv \mathbb{F}_i(X) = \pi_{j,i}(\mathbb{F}_j(X)) = \pi_{j,i}(X_j)$. Consequently using Lemma 1, the Bayes error inequality, $L_{\mathcal{A}_{i+1}} \le L_{\mathcal{A}_i}, \forall\, i \in \{1, ldots, n-1\}$, follows directly. For the estimation error inequality, we will prove the following sufficient condition:

$$D(P_{X_i|Y}(\cdot|y)\|\hat{P}_{X_i|Y}(\cdot|y)) \le D(P_{X_{+1}i|Y}(\cdot|y)\|\hat{P}_{X_{i+1}|Y}(\cdot|y)),$$

$$D(\hat{P}_{X_i|Y}(\cdot|y)\|P_{X_i|Y}(\cdot|y)) \le D(\hat{P}_{X_{+1}i|Y}(\cdot|y)\|P_{X_{i+1}|Y}(\cdot|y)), \tag{27}$$

$\forall\, i \in \{1,\ldots,n-1\}$ and $\forall\, y \in \mathcal{Y}$. We focus on proving one of the set of inequalities in (27), the proof of the other set of inequalities is equivalent.

Without loss of generality let us consider a generic pair $(io,yo) \in \{1,\ldots,n-1\} \times \mathcal{Y}$. Let $\hat{P}_{io}$ be the empirical distribution on $(\mathcal{X},\sigma_{io})$ induced by the measurable transformation $\mathbb{F}_{io}(\cdot)$ and the probability space $(\mathcal{A}_i,\hat{P}_{X_{io}|Y}(\cdot|yo))$. In this case σ_{io} is the sigma filed induced by the partition $Q_{io} \equiv \{\mathbb{F}_{io}^{-1}(\{a\}) : a \in \mathcal{A}_{io}\}$, and consequently the measure $\hat{P}_{io}$ is univocally characterized by [20]:

$$\hat{P}_{io}(\mathbb{F}_i^{-1}(\{a\})) = \hat{P}_{X_{io}|Y}(\{a\}|yo), \quad \forall\, a \in \mathcal{A}_{io}. \tag{28}$$

The same process can be used to induce a measure $\hat{P}_{io+1}$ on $(\mathcal{X},\sigma_{io+1})$. Note that given that the family of representations is embedded, we have that Q_{io+1} is a refinement of Q_{io} in $\mathcal{X}$ and consequently $\sigma_{io} \subset \sigma_{io+1}$ [20]. Then we have that $\hat{P}_{io+1}$ is also well defined on $(\mathcal{X},\sigma_{io})$. Moreover, by the consistence property of the conditional class probabilities $\{\hat{P}_{X_i|Y}(\cdot|yo) : i=1,\ldots,n\}$ on the family of representation functions $\{\mathbb{F}_i(\cdot) : i=1,\ldots,n\}$, we want to

[12] It is important to note that this sequence of induced sigma fields characterizes a filtration [21], in other words $\sigma(\mathbb{F}_i) \subset \sigma(\mathbb{F}_{i+1})$, because of the existence of a measurable mapping $\pi_{i+1,i}(\cdot)$ from $(\mathcal{X}_{i+1},\mathcal{F}_{i+1})$ to $(\mathcal{X}_i,\mathcal{F}_i)$.

show that the two measures agree on σ_{io}. For that we just need to show that they agree on the set of events that generate the sigma field, i.e., in $Q_{io}=\{\mathbb{F}_{io}^{-1}(\{a\}):a\in\mathcal{A}_{io}\}$, because Q_{io} is a partition and in particular a semi-algebra [20]. Then without loss of generality let us consider the event $\mathbb{F}_{io}^{-1}(\{a\})$, then we have that:

$$\begin{aligned}\hat{P}_{io+1}(\mathbb{F}_{io}^{-1}(\{a\}))&=\hat{P}_{io+1}(\mathbb{F}_{io+1}^{-1}(\pi_{io+1,io}^{-1}(\{a\})))\\&=\hat{P}_{X_{io+1}|Y}(\pi_{io+1,io}^{-1}(\{a\})|yo)=\hat{P}_{X_{io}|Y}(\{a\}|yo)\\&=\hat{P}_{io}(\mathbb{F}_{io}^{-1}(\{a\})),\quad\forall\, a\in\mathcal{A}_{io}.\end{aligned}\tag{29}$$

The first equality is because of the fact that $\mathbb{F}_i(\cdot)=\pi_{io+1,io}(\mathbb{F}_j(\cdot))$—embedded property of the representation family, the second by (28), the third by the consistence property of the conditional class probabilities and the last again by definition of $P_{io}(\cdot)$, (28). Consequently, we can just consider $\overline{P}\equiv\overline{P}_{io+1}$ as the empirical probability measure well defined on $(\mathcal{X},\sigma_{io+1})$ and $(\mathcal{X},\sigma_{io})$. Also note that the original probability measure $P_{X|Y}(\cdot|yo)$ is well defined on $(\mathcal{X},\sigma_{io+1})$ and $(\mathcal{X},\sigma_{io})$ by the measurability of $\mathbb{F}_{io}$ and $\mathbb{F}_{io+1}$, respectively [20].

Finally, from the definition of the KLD [24, Chapter 5]:

$$D(P_{X_{io}|Y}(\cdot|yo)\|\hat{P}_{X_{io}|Y}(\cdot|yo))=D_{(\mathcal{X},\sigma_{io})}(P_{X|Y}(\cdot|yo)\|\hat{P}),\tag{30}$$

$$D(P_{X_{io+1}|Y}(\cdot|yo)\|\hat{P}_{X_{io+1}|Y}(\cdot|yo))=D_{(\mathcal{X},\sigma_{io+1})}(P_{X|Y}(\cdot|yo)\|\hat{P}),\tag{31}$$

where

$$D_{(\mathcal{X},\sigma_{io})}(P_{X|Y}(\cdot|yo)\|\hat{P})=\sum_{A\in Q_{io}}P_{X|Y}(A|yo)\log\frac{P_{X|Y}(A|yo)}{\hat{P}(A)},$$

$$D_{(\mathcal{X},\sigma_{io+1})}(P_{X|Y}(\cdot|yo)\|\hat{P})=\sum_{A\in Q_{io+1}}P_{X|Y}(A|yo)\log\frac{P_{X|Y}(A|yo)}{\hat{P}(A)},$$

and using the Lemma 2 presented in Appendix A, considering that $\sigma_{io}\subset\sigma_{io+1}$, and Eqs. (30) and (31), we prove the sufficient condition stated in (27) and consequently the result. □

Appendix C. A simulated scenario to compute the oracle estimation and approximation errors

In this section we present a controlled simulated scenario to illustrate the trade-off between the estimation and approximation error reported in Section 3, across a sequence of embedded feature spaces. Since the estimation and approximation error, i.e., $\Delta g_{MAP}(\hat{P}_{X_i,Y})$ and $L_{\mathcal{X}_i}$ considered in Theorem 2, are functions of the true joint probability measure $P_{X,Y}$, the only way to measure those quantities directly and analyze their trend is by a controlled simulated scenario. For that reason, we choose a simple two-class problem associated with the classical binary detection in the presence of *additive white Gaussian noise* (AWGN) [7,4]. In particular, we assume $\mathcal{Y}=\{0,1\}$, $P_Y(0)=P_Y(1)=1/2$, $\mathcal{X}=\mathbb{R}^d$ with $d=10$, and a multivariate Gaussian class conditional density, i.e., $P_{X|Y}(\cdot|y)\sim\mathcal{N}(\mu_y,K_y)$ for $y\in\{0,1\}$ with $\mu_0=(1,1,\ldots,1)^\dagger=-\mu_1$ and $K_0=K_1=\sigma^2\cdot I_{d\times d}$.[13] In this context, we consider the family of coordinate projections (presented in Corollary 2) to induce the family of embedded features spaces given by: $\mathcal{X}_1=\mathbb{R}$, $\mathcal{X}_2=\mathbb{R}^2,\ldots,\mathcal{X}_d=\mathbb{R}^d=\mathcal{X}$. In the Bayes context, we use the standard maximum likelihood (ML) criterion to estimate the parameters of the densities from the class-conditional empirical data. More precisely, given $X_1,X_2,\ldots,X_N$ i.i.d. realizations driven by $\mathcal{N}(\mu_y,K_y)$, we estimate the mean vector and covariance matrix by:

$$\hat{\mu}_{y,N}=\frac{1}{N}\sum_{i=1}^{N}X_i,\quad\hat{K}_{y,N}=\frac{1}{N}\sum_{i=1}^{N}X_i\cdot X_i-\hat{\mu}_{y,N}\cdot\hat{\mu}_{y,N}^\dagger.\tag{32}$$

Note that as required in Theorem 2, this learning criterion induces a family of empirical class-conditional distributions which are consistent with the sequence of coordinate projections (see Appendix E for more details on this). Since the true distribution that generates the training data is known, we have access to $P_{X_i,Y}$ and $\hat{P}_{X_i,Y}$ for all $i\in\{1,\ldots,10\}$, and consequently, we can compute the Bayes error directly by (see details in [22]):

$$L_{\mathcal{X}_i}=\mathbb{P}(W(u)>\sqrt{i}/\sigma)=\int_{\sqrt{i}/\sigma}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-(1/2)x^2}dx,\tag{33}$$

where $W(u)$ is a scalar zero mean and unit variance Gaussian random variable. On the other hand, the KLD between Gaussian densities has a closed-form expression [8,44] and consequently from (6):

$$\begin{aligned}\Delta g_{MAP}(\hat{P}_{X_i,Y})&=\sqrt{1/2\ln 2}\\&\times\sum_{y\in\{0,1\}}\sqrt{\min\{D(\mathcal{N}(\mu_y,K_y)\|\mathcal{N}(\hat{\mu}_{y,N},\hat{K}_{y,N})),D(\mathcal{N}(\hat{\mu}_{y,N},\hat{K}_{y,N})\|\mathcal{N}(\mu_y,K_y))\}},\end{aligned}\tag{34}$$

with $D(\mathcal{N}(\hat{\mu}_{y,N},\hat{K}_{y,N})\|\mathcal{N}(\mu_y,K_y))=\frac{1}{2}$ [log $(detK_y/det\hat{K}_{y,N})-i+trace$ $(K_y^{-1}\cdot\hat{K}_{y,N})+(\hat{\mu}_{y,N}-\mu_y)^\dagger K_y^{-1}(\hat{\mu}_{y,N}-\mu_y)$] [8,9].

Therefore from the oracle expressions in (33) and (34), we can measure the trade-off reported in Theorem 2 as a function of the dimension of $\mathcal{X}_i$ and the number of sample point used to estimate the empirical class-conditional densities in (32). Figs. 1–3 report the estimation and approximation errors trends across the embedded space sequence for three learning conditions ($N=100, N=10,000, N=1,000,000$) with $\sigma=1$. As theory predicts, in the three learning regimes the monotonic behavior of the estimation and Bayes error are shown. Furthermore, the estimation error dominates the approximation error in the regime of few samples (see Fig. 1 for $N=100$) and that trend is reversed as we move to the large sampling regime (in this case in the order of hundred of thousands points in Fig. 2). This is expected, because as N goes to infinity the term $D(\mathcal{N}(\hat{\mu}_{y,N},\hat{K}_{y,N})\|\mathcal{N}(\mu_y,K_y))$ in (34) tends to zero with probability one (almost-surely), as a direct consequence of the *strong law of large numbers* [21].

Finally, if we add the two error terms $\Delta g_{MAP}(\hat{P}_{X_i,Y})+L_{\mathcal{X}_i}$ for all $i\in\{1,\ldots,10\}$, we can solve the oracle MPE-SR problem in (11) and, consequently, find the feature dimension that offers the optimal balance between estimation and approximation error (see Section 4). As it can be seen in the bottom sub-figures in Figs. 1–3, the optimal dimensions (2, 6 and 10, respectively) are proportional to the number of sample points ($N=100$, $N=10{,}000$ and $N=1{,}000{,}000$) because the estimation error has a monotonic decreasing behavior as a function of N.

Appendix D. Proof that maximum likelihood estimation is consistent with respect to a sequence of finite embedded representations

Let $\{\mathbb{F}_i(\cdot):i=1,\ldots,n\}$ be the family of embedded representation functions taking values in finite alphabet sets $\{\mathcal{A}_i:i=1,\ldots,n\}$, respectively. For every representation space $\mathcal{A}_i$, the empirical distribution is obtained by the ML criterion [4,5], where the conditional class distribution is given by,

$$\hat{P}_{X_i|Y}(\{a\}|y)=\frac{\sum_{k=1}^{N}\mathbb{1}_{\{(a,y)\}}(\mathbb{F}_i(x_k),y_k)}{N_y},\tag{35}$$

$\forall\, i\in\{1,\ldots,n\},\forall\, a\in\mathcal{A}_i$ and $\forall\, y\in\mathcal{Y}$, where $N_y\equiv\sum_{k=1}^{N}\mathbb{1}_{\{y\}}(y_k)$ is assumed to be strictly greater than zero. For the proof we will

[13] $I_{d\times d}$ denotes the identity matrix.

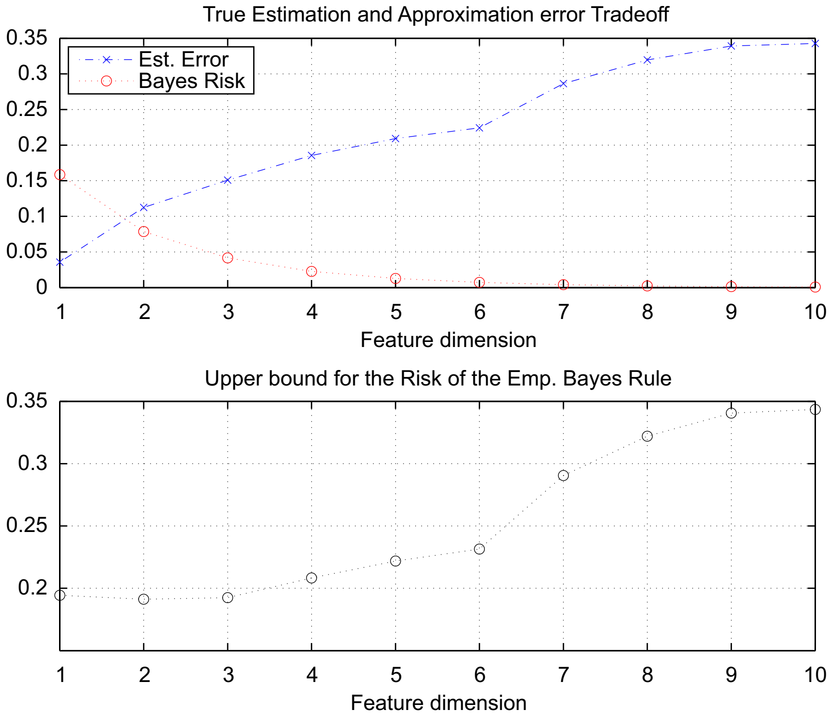

Fig. 1. The top figure reports the trend of the estimation error $\Delta g_{MAP}(\hat{P}_{X_i,Y})$ and the approximation error $L_{\mathcal{X}_i}$ as a function of the space dimensionality, for a dimensionally embedded space sequence. The bottom figure shows the sum of the two error terms, i.e., $\Delta g_{MAP}(\hat{P}_{X_i,Y})+L_{\mathcal{X}_i}$, as a function of the space dimensionality. These results were obtained for $N=100$ i.i.d. realizations of the class conditional densities.

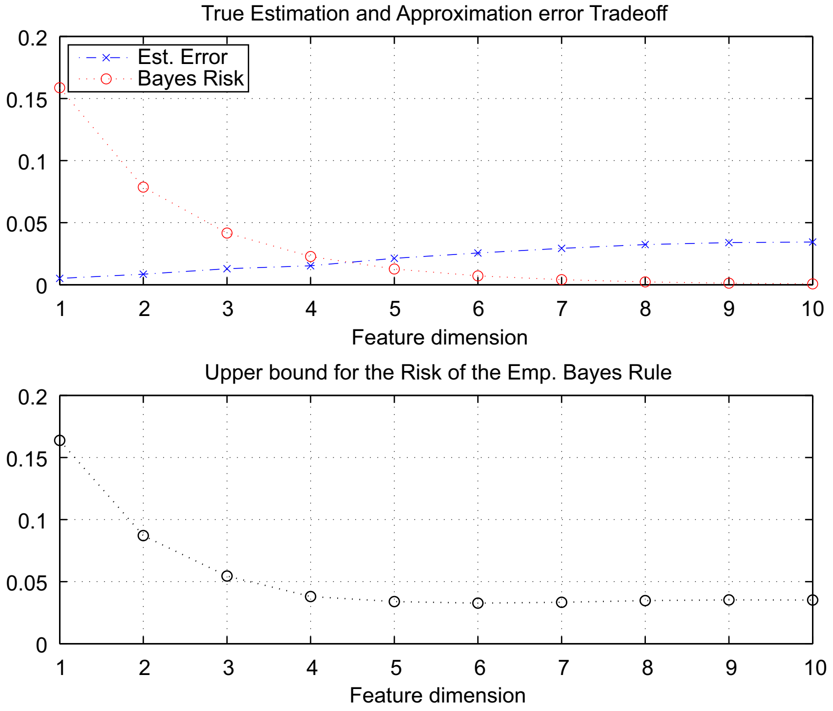

Fig. 2. Same description as Fig. 1. These results were obtained for $N=10{,}000$ i.i.d. realizations of the class conditional densities.

use the induced probability measure on the original observation space $(\mathcal{X},\mathcal{F}_{\mathcal{X}})$, that we define by

$$\hat{P}_{i|y}(\mathbb{F}_i^{-1}(\{a\})) \equiv \hat{P}_{X_i|Y}(\{a\}|y) \tag{36}$$

for all $i \in \{1,\ldots,n\}$ and $y \in \mathcal{Y}$. By (35), it is straightforward to show that for any $a \in \mathcal{A}_i$

$$\hat{P}_{i|y}(\mathbb{F}_i^{-1}(\{a\})) = \frac{\sum_{k=1}^{N} \mathbb{1}_{\{(\mathbb{F}_i^{-1}(\{a\}),y)\}}(x_k,y_k)}{N_y}. \tag{37}$$

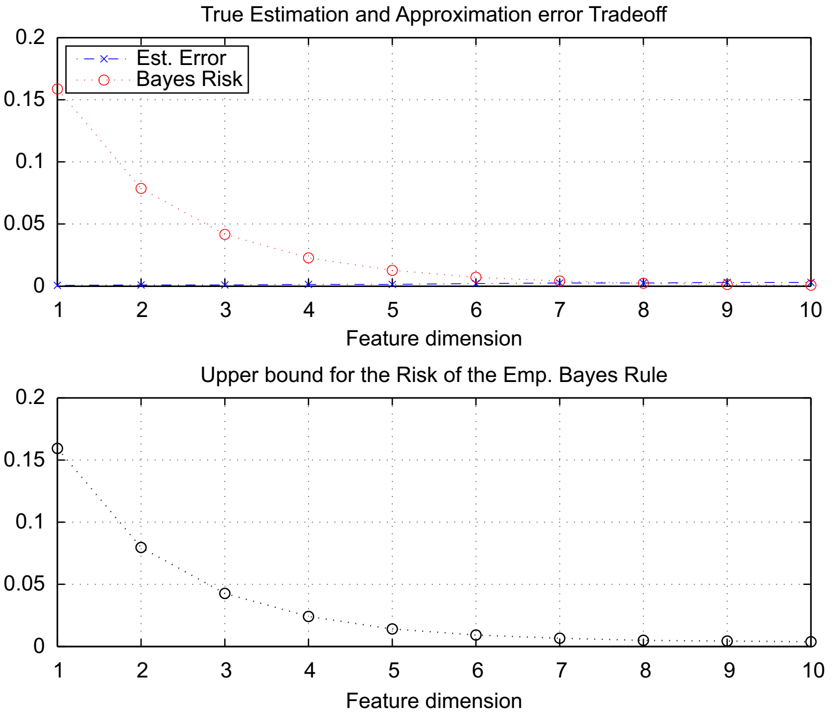

Fig. 3. Same description as Fig. 1. These results were obtained for N=1,000,000 i.i.d. realizations of the class conditional densities.

Without loss of generality, let us consider $io,jo \in \{1,\ldots,n\}$ and $yo \in \mathcal{Y}$, such that $io < jo$. For proving the consistence condition of the ML empirical distributions, we just need to show that

$$\hat{P}_{X_{io}|Y}(\{a\}|yo) = \hat{P}_{X_{jo}|Y}(\pi^{-1}_{jo,io}(\{a\})|yo), \quad \forall\ a \in \mathcal{A}_{io}. \tag{38}$$

By Remark 2, we have that the induced quantization $Q_{F_{jo}} \equiv \{\mathbb{F}^{-1}_{jo}(\{a\}) : a \in \mathcal{A}_{jo}\}$ is a refinement of $Q_{F_{io}} \equiv \{\mathbb{F}^{-1}_{io}\ (\{a\}) : a \in \mathcal{A}_{io}\}$. Then, any atom $\mathbb{F}^{-1}_{jo}(\{a\})$ indexed by $a \in \mathcal{A}_{io}$, can be expressed as disjoint unions of atoms in $Q_{F_{jo}}$; more precisely, we have that:

$$\mathbb{F}^{-1}_{io}(\{a\}) = \bigcup_{b \in \pi^{-1}_{jo,io}(\{a\})} \mathbb{F}^{-1}_{jo}(\{b\}) = \mathbb{F}^{-1}_{jo}(\pi^{-1}_{jo,io}(\{a\})) \tag{39}$$

where finally by (36) and (37), we have that:

$$\begin{aligned}\hat{P}_{X_{io}|Y}(\{a\}|yo) &= \frac{\sum_{k=1}^{N} \mathbb{1}_{\{(\mathbb{F}^{-1}_{io}(a),yo)\}}(x_k,y_k)}{N_{yo}} \\ &= \frac{\sum_{k=1}^{N} \mathbb{1}_{\{(\mathbb{F}^{-1}_{jo}(\pi^{-1}_{jo,io}(\{a\})),yo)\}}(x_k,y_k)}{N_{yo}} = \hat{P}_{X_{jo}|Y}(\pi^{-1}_{jo,io}(\{a\})|yo). \quad \square\end{aligned} \tag{40}$$

Appendix E. Proof that maximum likelihood estimation is consistent for the Gaussian parametric assumption

Without loss of generality, let us consider just $\mathbf{f}_1(x) = \mathbf{A}_1 \cdot x$ and $\mathbf{f}_2(x) = \mathbf{A}_2 \cdot x$, with $\mathbf{A}_1 \in \mathbb{R}(k1,K)$ and $\mathbf{A}_2 \in \mathbb{R}(k2,K)$ $(0 < k1 < k2 < K)$. We need to show that $\hat{P}_{\mathbf{f}_2(X)|Y}(\cdot|y)$ defined on $(\mathbb{R}^{k2},\mathcal{B}^{k2})$ is consistent with respect to $\hat{P}_{\mathbf{f}_1(X)|Y}(\cdot|y)$ defined on $(\mathbb{R}^{k1},\mathcal{B}^{k1})$, in the sense that $\hat{P}_{\mathbf{f}_2(X)|Y}(\cdot|y)$ induces $\hat{P}_{\mathbf{f}_1(X)|Y}(\cdot|y)$ by the measurable mapping $B_{2,1} : (\mathbb{R}^{k2},\mathcal{B}^{k2}) \to (\mathbb{R}^{k1},\mathcal{B}^{k1})$. However, under the Gaussian parametric assumption, this condition reduces to checking the first and second order statistics of the involved distributions [22]. Considering the training data, it is direct to show that the empirical mean and covariance matrix for $\hat{P}_{\mathbf{f}_2(X)|Y}(\cdot|y)$ is given by $\mathbf{A}_2\hat{\mu}_y$ and $\mathbf{A}_2\hat{\Sigma}_y\mathbf{A}_2^{\dagger}$, respectively, where $\hat{\mu}_y$ and $\hat{\Sigma}_y$ are the respective empirical values in the original observation space $\mathcal{X}$. Analogous results hold for the case of $\hat{P}_{\mathbf{f}_1(X)|Y}(\cdot|y)$.

Given that linear transformations preserve the multivariate Gaussian distribution, we have that $\hat{P}_{\mathbf{f}_2(X)|Y}(\cdot|y)$ induces a Gaussian distribution on $(\mathbb{R}^{k1},\mathcal{B}^{k1})$ with mean $B_{2,1}\mathbf{A}_2\hat{\mu}$ and covariance matrix $B_{2,1}\mathbf{A}_2\hat{\Sigma}_y\mathbf{A}_2^{\dagger}B_{2,1}^{\dagger}$. Finally, given that the linear transformations $\mathbf{f}_1(\cdot)$ and $\mathbf{f}_2(\cdot)$ preserve the consistence structure of $\mathbb{R}^{k1}$, $\mathbb{R}^{k2}$, we have that $B_{2,1}\mathbf{A}_2 = \mathbf{A}_1$ which is sufficient to prove the result. $\square$

References

[1] V.K. Goyal, M. Vetterli, N.T. Thao, Quantized overcomplete expansions in $\mathbb{R}^n$: Analysis, synthesis and algorithms, IEEE Transactions on Information Theory 44 (7) (1998) 16–31.
[2] K. Ramchandran, M. Vetterli, C. Herley, Wavelet, subband coding, and best bases, Proceedings of the IEEE 84 (4) (1996) 541–560.
[3] A. Cohen, I. Daubechies, O. Guleryuz, M. Orchard, On the importance of combining wavelet-based nonlinear approximation with coding strategies, IEEE Transactions on Information Theory 48 (7) (2002) 1895–1921.
[4] R.O. Duda, P.E. Hart, Pattern Classification and Scene Analysis, Wiley New York, 1983.
[5] L. Breiman, J. Friedman, R. Olshen, C. Stone, Classification and Regression Trees, Wadsworth, Belmont, CA, 1984.
[6] L. Devroye, L. Gyorfi, G. Lugosi, A Probabilistic Theory of Pattern Recognition, Springer-Verlag, New York, 1996.
[7] J. Wozencraft, I. Jacobs, Principles of Communication Engineering, Waveland Press, 1965.
[8] S. Kullback, Information Theory and Statistics, Wiley, New York, 1958.
[9] T.M. Cover, J.A. Thomas, Elements of Information Theory, Wiley Interscience, New York, 1991.
[10] Special issue: Dimensionality reduction, IEEE Signal Processing Magazine 28 (2) (2011) 1–128.
[11] M. Padmanabhan, S. Dharanipragada, Maximizing information content in feature extraction, IEEE Transactions on Speech and Audio Processing 13 (4) (2005) 512–519.
[12] K. Etemad, R. Chellapa, Separability-based multiscale basis selection and feature extraction for signal and image classification, IEEE Transactions on Image Processing 7 (10) (1998) 1453–1465.

[13] L.O. Jimenez, D.A. Landgrebe, Hyperspectral data analysis and supervised feature reduction via projection pursuit, IEEE Transactions on Geoscience and Remote Sensing 37 (6) (1999) 2653–2667.

[14] S. Kumar, J. Ghosh, M.M. Crawford, Best-bases feature extraction algorithms for classification of hyperspectral data, IEEE Transactions on Geoscience and Remote Sensing 39 (7) (2001) 1368–1379.

[15] T.F. Quatieri, Discrete-time Speech Signal Processing Principles and Practice, Prentice Hall, 2002.

[16] J. Novovicova, P. Pudil, J. Kittler, Divergence based feature selection for multimodal class densities, IEEE Transactions on Pattern Analysis and Machine Intelligence 18 (2) (1996) 218–223.

[17] A.K. Jain, R.P.W. Duin, J. Mao, Statistical pattern recognition: a review, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (2000) 4–36.

[18] N. Vasconcelos, Minimum probability of error image retrieval, IEEE Transactions on Signal Processing 52 (8) (2004) 2322–2336.

[19] J. Silva, S. Narayanan, Minimum probability of error signal representation, in: IEEE International Workshop on Machine Learning for Signal Processing, Thessaloniki, Greece, 2007.

[20] P.R. Halmos, Measure Theory, Van Nostrand, New York, 1950.

[21] L. Breiman, Probability, Addison-Wesley, 1968.

[22] R. Gray, L.D. Davisson, Introduction to Statistical Signal Processing, Cambridge University Press, 2004.

[23] N.A. Schmid, J.A. O'Sullivan, Thresholding method for dimensionality reduction in recognition system, IEEE Transactions on Information Theory 47 (7) (2001) 2903–2920.

[24] R.M. Gray, Entropy and Information Theory, Springer-Verlag, New York, 1990.

[25] V. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag New York, 1999.

[26] A. Jain, P. Moulin, M.I. Miller, K. Ramchandran, Information-theoretic bounds on target recognition performances based on degraded image data, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (9) (2002) 1153–1166.

[27] H.V. Poor, J.B. Thomas, Applications of Ali–Silvey distance measures in the design of generalized quantizers for binary decision problems, IEEE Transactions on Communications 25 (9) (1977) 893–900.

[28] J.W. Fisher III, T. Darrel, W. Freeman, P. Viola, Learning joint statistical models for audio-visual fusion and segregation, in: Advances in Neural Information Processing System, Denver, USA, Advances in Neural Information Processing Systems, 2000.

[29] M.L. Cooper, M.I. Miller, Information measures for object recognition accommodating signature variability, IEEE Transactions on Information Theory 46 (5) (2000) 1896–1907.

[30] J. Kim, J.W. Fisher III, A. Yezzi, M. Cetin, A.S. Willsky, A nonparametric statistical method for image segmentation using information theory and curve evolution, IEEE Transactions on Image Processing 14 (10) (2005) 1486–1502.

[31] J. Silva, S. Narayanan, Discriminative wavelet packet filter bank selection for pattern recognition, IEEE Transactions on Signal Processing 57 (5) (2009) 1796–1810.

[32] A.B. Nobel, Analysis of a complexity-based pruning scheme for classification tree, IEEE Transactions on Information Theory 48 (8) (2002) 2362–2368.

[33] P. Chou, T. Lookabaugh, R. Gray, Optimal pruning with applications to tree-structure source coding and modeling, IEEE Transactions on Information Theory 35 (2) (1989) 299–315.

[34] A.B. Nobel, Recursive partitioning to reduce distortion, IEEE Transactions on Information Theory 43 (4) (1997) 1122–1133.

[35] C. Scott, R.D. Nowak, Minimax-optimal classification with dyadic decision trees, IEEE Transactions on Information Theory 52 (4) (2006) 1335–1353.

[36] B.D. Ripley, Patten Recognition and Neural Networks, Cambridge University Press, 1996.

[37] C. Scott, Tree pruning with subadditive penalties, IEEE Transactions on Signal Processing 53 (12) (2005) 4518–4525.

[38] P.D. Grunwald, The Minimum Description Length Principle, The MIT Press, Cambridge, Massachusetts, 2007.

[39] V. Vapnik, Statistical Learning Theory, John Wiley, 1998.

[40] O. Bousquet, S. Boucheron, G. Lugosi, Theory of classification: a survey of recent advances, ESAIM: Probability and Statistics 9 (2005) 323–375.

[41] V. Vapnik, Estimation of Dependencies based on Empirical Data, Springer-Verlag, New York, 1979.

[42] V. Vapnik, A.J. Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities, Theory of Probability and its Application 16 (1971) 264–280.

[43] J. Silva, S. Narayanan, Upper bound Kullback–Leibler divergence for hidden Markov models with application as discrimination measure for speech recognition, in: IEEE International Symposium on Information Theory, 2006.

[44] J. Silva, S. Narayanan, Upper bound Kullback–Leibler divergence for transient hidden Markov models, IEEE Transactions on Signal Processing 56 (9) (2008) 4176–4188.

Jorge Silva is Assistant Professor at the Electrical Engineering Department, University of Chile, Santiago, Chile. He received the Master of Science (2005) and Ph.D (2008) in Electrical Engineering from the University of Southern California (USC). He is IEEE member of the Signal Processing and Information Theory Societies and he has participated as a reviewer in various IEEE journals on Signal Processing. Jorge Silva was research assistant at the Signal Analysis and Interpretation Laboratory (SAIL) at USC (2003–2008) and was also research intern at the Speech Research Group, Microsoft Corporation, Redmond (Summer 2005).

Jorge Silva is recipient of the Outstanding Thesis Award 2009 for Theoretical Research of the Viterbi School of Engineering, the Viterbi Doctoral Fellowship 2007-2008 and Simon Ramo Scholarship 2007–2008 at USC. His research interests include: non-parametric learning; optimal signal representation for pattern recognition; tree-structured vector quantization for lossy compression and statistical learning; universal quantization; sequential detection; distributive learning and sensor networks.

Shrikanth (Shri) Narayanan is the Andrew J. Viterbi Professor of Engineering at the University of Southern California (USC), and holds appointments as Professor of Electrical Engineering, Computer Science, Linguistics and Psychology. Prior to USC he was with AT&T Bell Labs and AT&T Research from 1995–2000. At USC he directs the Signal Analysis and Interpretation Laboratory. His research focuses on human-centered information processing and communication technologies.

Shri Narayanan is an Editor for the Computer Speech and Language Journal and an Associate Editor for the IEEE Transactions on Multimedia, IEEE Transactions on Affective Computing, and for the Journal of the Acoustical Society of America. He was also previously an Associate Editor of the IEEE Transactions of Speech and Audio Processing (2000–2004) and the IEEE Signal Processing Magazine (2005–2008). He served on the Speech Processing technical committee (2005–2008) and Multimedia Signal Processing technical committees (2004–2008) of the IEEE Signal Processing Society and presently serves on the Speech Communication committee of the Acoustical Society of America and the Advisory Council of the International Speech Communication Association.

Shri Narayanan is a Fellow of the Acoustical Society of America, IEEE, and the American Association for the Advancement of Science and a member of Tau–Beta–Pi, Phi Kappa Phi and Eta–Kappa–Nu. He is a recipient of a number of honors including Best Paper awards from the IEEE Signal Processing society in 2005 (with Alex Potamianos) and in 2009 (with Chul Min Lee) and appointment as a Signal Processing Society Distinguished Lecturer for 2010–2011. Papers with his students have won awards at ICSLP'02, ICASSP'05, MMSP'06, MMSP'07 and DCOSS'09 and InterSpeech2009-Emotion Challenge, Interspeech-2010 and InterSpeech2011-Speaker State Challenge. He has published over 450 papers and has twelve granted U.S. patents.