MINIMUM PROBABILITY OF ERROR SIGNAL REPRESENTATION

Jorge Silva and Shrikanth Narayanan

Signal Analysis and Interpretation Laboratory, *http://sail.usc.edu*
University of Southern California, Viterbi School of Engineering
Los Angeles, CA 90089, USA
jorgesil@usc.edu, shri@sipi.usc.edu

ABSTRACT

The problem of minimum probability of error signal representation (MPE-SR) considering issues of finite training data is revisited and extended in this paper. Results are presented that justify addressing this problem as a complexity-regularized optimization criterion, reflecting the well-known tradeoff between signal representation quality and learning complexity. A rate-distortion type of formulation is proposed to address this optimization problem by finding a sequence of signal representations achieving optimal complexity-fidelity operational points. Finally under specific assumptions, it is shown that the MPE-SR reduces to a version of Fisher linear discriminant analysis.

1. INTRODUCTION AND PRELIMINARIES

Let us consider a random vector $X(u){:}(\Omega, \mathcal{F}, \mathbb{P}) \rightarrow (\mathcal{X}, \mathcal{F}_\mathcal{X})$ taking values in a finite-dimensional Euclidian space $\mathcal{X} = \mathbb{R}^K$, and a random variable $Y(u){:}(\Omega, \mathcal{F}, \mathbb{P}) \rightarrow (\mathcal{Y}, \mathcal{F}_\mathcal{Y})$ taking values in a finite alphabet set $\mathcal{Y}$, where $(\Omega, \mathcal{F}, \mathbb{P})$ denotes the underlying probability space[1]. We refer to $X(u)$ and $Y(u)$ as the observation and the class label random phenomena, respectively. The joint observation-class random vector $(X(u), Y(u))$ induces a joint probability measure $P_{X,Y}$ in the space $(\mathcal{X} \times \mathcal{Y}, \sigma(\mathcal{F}_\mathcal{X} \times \mathcal{F}_\mathcal{Y}))$[2]. Knowing the joint distribution $P_{X,Y}$, the classification problem is to find a measurable decision function $g(\cdot)$ from $(\mathcal{X}, \mathcal{F}_\mathcal{X})$ to $(\mathcal{Y}, \mathcal{F}_\mathcal{Y})$ such that given realizations of $X(u)$, infer their discrete counterpart $Y(u)$ with the minimum risk, given by $\mathbb{E}_{X,Y}\left[l\left(g(X), Y\right)\right]$. $l(y_1, y_2)$ represents the penalization of labeling an observation with the value y_1, when its true label is $y_2, \forall y_1, y_2 \in \mathcal{Y}$. This minimum risk decision is called the *Bayes decision rule*, where for the emblematic 0-1 risk function, $l(y_1, y_2) = \delta(y_1, y_2)$, the Bayes rule minimizes the probability of error, $g_{P_{X,Y}}(\bar{x}) = \arg\max_{y \in \mathcal{Y}} P_{X,Y}(\bar{x}, y)$. In this case the optimal probability of error is given by

$$L_\mathcal{X} = 1 - \mathbb{E}_X\left[\max_{i \in \mathcal{Y}} P_{Y|X}(i|X)\right]. \qquad (1)$$

$L_\mathcal{X}$ can be seen as the optimal performance obtained as a function of the representation quality of the observation space $\mathcal{X}$, that Vasconcelos [2] denoted as the *Bayes error bound*. In real scenarios we do not have access to the true joint distribution $P_{X,Y}$, but instead we have iid realizations of $(X(u), Y(u))$, $D_N \equiv \{(x_i, y_i) : i \in \{1, .., N\}\}$, 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 its empirical Bayes classification rule that we denote as $\hat{g}_{\hat{P}_{X,Y}}(\cdot)$. The risk of the empirical Bayes rule, $\mathbb{P}\left(\left\{u \in \Omega : \hat{g}_{\hat{P}_{X,Y}}(X(u)) \neq Y(u)\right\}\right)$, differs from the Bayes error bound $L_\mathcal{X}$ as a consequence of the estimation error. It is well understood that the estimation error introduces performance degradation with respect to the Bayes error bound $L_\mathcal{X}$, and that the magnitude of this deviation is a function of some notion of complexity of the observation space [2]. This mostly explains the well known "*peaking or Hughes phenomenon*" [3] as Vasconcelos formally justifies for the case of coordinate projections [2]. At the same time, this implies a strong relationship between the number of training examples and a notion of complexity of the feature observation space, underscoring the well-known justification of dimensionality reduction as a fundamental part of feature extraction.

The focus of this work is to study the role of signal representation in pattern recognition, considering the aforementioned issues of finite training data. New results in this direction have been recently presented by Vasconcelos [2]. [2] formalizes a tradeoff between the Bayes error bound and an information-theoretic indicator of the estimation error across a sequence of feature transformations of increasing complexity — measured in terms of dimensionality — and connects this result with the concept of *minimum probability*

This material is based upon work supported by awards from National Science Foundation, ONR-MURI, DARPA and the U.S. Army.

[1] Considering $\mathcal{X} = \mathbb{R}^K$, $\mathcal{F}_\mathcal{X}$ refers to the Borel sigma field $\mathcal{B}(\mathbb{R}^K)$ [1], and $\mathcal{F}_\mathcal{Y}$ the power set of $\mathcal{Y}$.

[2] $\sigma(\mathcal{F}_\mathcal{X} \times \mathcal{F}_\mathcal{Y})$ refers to the product sigma field [1].

of error signal representation (MPE-SR). In this paper we extend those results in more generality. In particular, we extend such tradeoff for a general family of feature-embedded representations and formalize sufficient conditions for the result to hold, which not only takes into consideration the embedded structure of the feature representation family, as originally considered in [2], but also the consistent nature of the family of empirical class-obervation distributions estimated across the sequence of observation representations.

The Bayes estimation error tradeoff is used to formulate the problem of MPE-SR as a complexity-regularized optimization, with an objective function that considers a fidelity indicator, which represents the Bayes error bound, and a regularization 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 an operational cost-fidelity problem. Interestingly, a version of the *Fisher linear discriminant analysis* [3] can be considered as a particular instance of this cost-fidelity formulation.

2. RESULTS ON SIGNAL REPRESENTATION FOR CLASSIFICATION

The following theorem characterizes an information theoretic indicator for quantifying the effects of estimation error in the Bayes decision approach.

THEOREM 1 *(*Theorem 4 *in [2]) Let us consider the joint observation-class distribution $P_{X,Y}$ and its empirical version $\hat{P}_{X,Y}$, both defined on $(\mathcal{X} \times \mathcal{Y}, \sigma(\mathcal{F}_\mathcal{X} \times \mathcal{F}_\mathcal{Y}))$, which 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*

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

where $\Delta g_{MAP}(\hat{P}_{X,Y}) =$

$$\sqrt{2 \ln 2} \sum_{y \in \mathcal{Y}} P_Y(\{y\}) \cdot \sqrt{D(\hat{P}_{X|Y}(\cdot|y)||P_{X|Y}(\cdot|y))}, \quad (2)$$

and $D(\cdot||\cdot)$ denotes the Kullback-Leibler divergence *(KLD) [4] between two probability distributions on $(\mathcal{X}, \mathcal{F}_\mathcal{X})$:*

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

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 a P_Y-average of a non-decreasing function of KLDs between the conditional class probabilities and their empirical counterpart. The KLD has a well known interpretation as a statistical discrimination measure between two probabilistic models [4], however in this context it is an indicator of the performance deviation, relative to the fundamental performance bound, $L_\mathcal{X}$, as a consequence of the statistical mismatch occurring in estimating the class conditional probabilities [2]. The following result, whose first version was presented in [2], formally introduces the role of signal representation by characterizing a tradeoff between the Bayes error bound and the estimation error from *Theorem* 1. In order to present it, first we need to introduce the notion of dimensional embedded space sequences [2], 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 structure.

Definition 1 *Let us consider a family of measurable transformations $\{\mathbb{F}_i(\cdot) : i = 1, .., n\}$ from the same domain, $(\mathcal{X}, \mathcal{F}_\mathcal{X})$ and taking values in $\{\mathcal{X}_1, .., \mathcal{X}_n\}$, a sequence of finite dimensional Euclidian spaces of strictly increasing dimensionality, i.e., $dim(X_i) < dim(X_{i+1})$, $\forall i \in \{1, .., n-1\}$. The family of transformations $\{\mathbb{F}_i(\cdot) : i = 1, .., n\}$ is called dimensional embedded if, $\forall i \in \{1, .., n-1\}$, $\exists \pi_{i+1,i}(\cdot)$ measurable mapping from $(\mathcal{X}_{i+1}, \mathcal{F}_{i+1})$ to $(\mathcal{X}_i, \mathcal{F}_i)$ such that*[4]*,*

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

In this context, we say that $\{\mathcal{X}_1, .., \mathcal{X}_n\}$ is dimensional embedded with respect to $\{\mathbb{F}_i(\cdot) : i = 1, .., n\}$ and $\{\pi_{i+1,i}(\cdot) : i = 1, .., n-1\}$. Those dependencies will be implicit when not mentioned.

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

$$\hat{P}_i(B) = \hat{P}_j(\pi_{j,i}^{-1}(B)), \quad (3)$$

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

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 highest dimensional probability space, $(\mathcal{X}_n, \mathcal{F}_n, \hat{P}_n)$, and the family

[3] We assume that the involved distributions are absolutely continuous with respect to the Lebesgue measure [4]. Also it is assumed that involved KLDs are well defined, case where this upper bound is meaningful.

[4] $\mathcal{F}_i$ refers to the Borel sigma field.

of measurable mappings $\{\pi_{j,i}(\cdot) : j > i\}$ of the embedded structure.

THEOREM 2 *Consider the joint distribution $P_{X,Y}$ defined on $(\mathcal{X} \times \mathcal{Y}, \sigma(\mathcal{F}_\mathcal{X} \times \mathcal{F}_\mathcal{Y}))$, and $\{\mathbb{F}_i(\cdot) : i = 1,..,n\}$ a sequence of measurable functions, $\mathbb{F}_i{:}(\mathcal{X}, \mathcal{F}_\mathcal{X}) \rightarrow (\mathcal{X}_i, \mathcal{F}_i)$, $\forall i \in \{1,.,n\}$. In addition, let us assume that $\{\mathbb{F}_i(\cdot) : i = 1,..,n\}$ is a family of dimensional embedded transformations, satisfying $\mathbb{F}_i(\cdot) = \pi_{j,i}(\mathbb{F}_j(\cdot))$ for all $j > i$ in $\{1,..,n\}$. Then, considering the representation random variables $\{X_i(u) \equiv \mathbb{F}_i(X(u)) : i = 1,..,n\}$ their Bayes bounds satisfy:*

$$L_{\mathcal{X}_{i+1}} \leq L_{\mathcal{X}_i}, \ \ \forall i \in \{1,..,n-1\}\,. \qquad (4)$$

If in addition let $\left\{\hat{P}_{X_i,Y} : i = 1,..,n\right\}$ be the family of empirical probability measures, with $\hat{P}_{X_i,Y}$ defined on $(\mathcal{X}_i \times \mathcal{Y}, \sigma(\mathcal{F}_i \times \mathcal{F}_\mathcal{Y}))$, and with conditional class distribution families $\left\{\hat{P}_{X_i|Y}(\cdot|y) : i = 1,..,n\right\}$ consistent with respect to the embedded space sequence $\{\mathcal{X}_1, \cdots, \mathcal{X}_n\}$, $\forall y \in \mathcal{Y}$. Then $\forall i \in \{1,..,n-1\}$,

$$\Delta g_{MAP}(\hat{P}_{X_i,Y}) \leq \Delta g_{MAP}(\hat{P}_{X_{i+1},Y}). \qquad (5)$$

The proof is presented in the Appendix.

This result presents a tradeoff between the Bayes error bound and estimation error by considering a family of representations of monotonically increasing complexity. In other words, by increasing complexity we improve the theoretical performance bound that we can achieve, but as a consequence of increasing the estimation error, which upper bounds how far we can potentially be from the Bayes error bound, per *Theorem* 1. The following corollary states the original result presented in [2] (*Theorem 5*) for the tradeoff between Bayes error and estimation errors. This was presented for the important case when the embedded sequence of feature spaces is induced by coordinate projections. The consistency condition of the family of empirical distributions holds naturally and, consequently, it is assumed and not included in the statement.

COROLLARY 1 *(*Theorem 5, *[2]) Let $\mathcal{X}$=$\mathbb{R}^K$ and the family of coordinate projection $\pi^K_m(\cdot){:}\mathbb{R}^K \rightarrow \mathbb{R}^m$, $m{\leq}K$, be given by: $\pi^K_m(x_1, ..., x_m, ..x_K) = (x_1, ..., x_m)$, $\forall (x_1, ..., x_K) \in \mathbb{R}^K$. 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 $\{\mathcal{X}_1, .., \mathcal{X}_K\}$ characterized by: $\mathcal{X}_i = \pi^K_i(\mathcal{X})$, $\forall i \in \{1,..,K\}$. Then, across this sequence of dimensional embedded spaces, the Bayes bound and estimation error satisfy the following inequalities, $L_{\mathcal{X}_{i+1}} \leq L_{\mathcal{X}_i}$ and $\Delta g_{MAP}(\hat{P}_{X_{i+1},Y}) \geq \Delta g_{MAP}(\hat{P}_{X_i,Y})$, $\forall i \in \{1,..,K-1\}$.*

The proof follows directly from *Theorem* 2.

The next section formalizes the problem of minimum probability of error signal representation (MPE-SR), and shows how the main result presented in this section plays a role in addressing it as a cost-fidelity problem.

3. MINIMUM PROBABILITY OF ERROR SIGNAL REPRESENTATION (MPE-SR)

Let us consider $D_N = \{(x_i, y_i) : i = 1,..,N\}$ iid realizations of the joint observation-class phenomenon $(X(u), Y(u))$ with true probability measure $P_{X,Y}$ defined 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 representation functions $\mathbb{D}$, where any $\mathbf{f}(\cdot) \in \mathbb{D}$ is defined in $\mathcal{X}$ and takes values in $\mathcal{X}_f$. Let us assume that any 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 an implicit learning approach, where the empirical Bayes classification rule is given by: $\hat{g}_f(x) = \arg\max_{y \in \mathcal{Y}} \hat{P}_{X_f,Y}(x, y)$.

The MPE-SR problem is the solution of: $\mathbf{f}^* =$

$$\arg\min_{\mathbf{f} \in \mathbb{D}} \mathbb{E}_{X,Y}\left(\mathbb{I}_{\{(x,y) \in \mathcal{X} \times \mathcal{Y} : \hat{g}_f(\mathbf{f}(x)) \neq y\}}(X, Y)\right), \qquad (6)$$

where the expected value is with respect to the underlying true distribution $P_{X,Y}$ on $(\mathcal{X} \times \mathcal{Y}, \sigma(\mathcal{F}_\mathcal{X} \times \mathcal{F}_\mathcal{Y}))$. Note that $\forall \mathbf{f}(\cdot){\in}\mathbb{D}$, $\mathbb{E}_{X,Y}\left(\mathbb{I}_{\{(x,y) \in \mathcal{X} \times \mathcal{Y} : \hat{g}_f(\mathbf{f}(x)) \neq y\}}(X, Y)\right) \geq L_{\mathcal{X}_f} \geq L_\mathcal{X}$, then the MPE-SR criterion, Eq.(6), chooses the representation function whose performance is the closest to $L_\mathcal{X}$, the fundamental bound for the problem. Using *Theorem* 1 we have that, $\mathbb{E}_{X_f,Y}\left(\mathbb{I}_{\{(x,y) \in \mathcal{X}_f \times \mathcal{Y} : \hat{g}_f(x) \neq y\}}(X_f, Y)\right) \leq \Delta g_{MAP}(\hat{P}_{X_f,Y}) + L_{\mathcal{X}_f}$, $\forall \mathbf{f} \in \mathbb{D}$, then we can approximate Eq.(6) by

$$\mathbf{f}^* \approx \arg\min_{\mathbf{f} \in \mathbb{D}} \Delta g_{MAP}(\hat{P}_{X_f,Y}) + \left[L_{\mathcal{X}_f} - L_\mathcal{X}\right]. \qquad (7)$$

In Eq.(7), we introduce the normalization factor $L_\mathcal{X}$ to make explicit that this optimization problem tries to find the optimal tradeoff between the approximation quality, $L_{\mathcal{X}_f} - L_\mathcal{X}$, and the estimation error, $\Delta g_{MAP}(\hat{P}_{X_f,Y})$. The MPE-SR in Eq.(7) is a complexity regularized optimization whose objective function consists of a weighted combination of a fidelity criterion, reflecting the Bayes error bound, and a cost term, penalizing the complexity of the representation function.

In general the Bayes risk does not offer closed-form expressions and consequently it is not generally possible to find solutions for Eq.(7). Mutual information (MI) [4], $I(\mathbf{f}) = I(\mathbf{f}(X); Y)$, can be adopted as a fidelity indicator because of its strong connection with the probability of error[5] and because it allows us to algorithmically address Eq.(7) under particular problem scenarios — the family of representations $\mathbb{D}$ and modeling assumptions — as we shall

[5] Fano's inequality, [4], characterizes a lower bound for the probability of error.

exemplify in the following section. Regarding the estimation error — the regularization term in Eq.(7) — we use the dimensionality of $\mathcal{X}_f$, that we denote by $R(\mathbf{f})$, $\forall \mathbf{f}(\cdot) \in \mathbb{D}$. This is justified in *Theorem* 2 where the estimation error is proportional to the dimensionality of the space. Then considering our new fidelity and penalization functions, $I(\mathbf{f})$ and $R(\mathbf{f})$, Eq.(7) can be approximated by the following complexity regularized fidelity criterion:

$$\mathbf{f}^*(\lambda) = \arg \min_{\mathbf{f}\in\mathbb{D}} \Psi(I(\mathbf{f})) + \lambda \cdot \Phi(R(\mathbf{f})). \qquad (8)$$

Considering the tendency of the new fidelity-cost indicators, $\Psi(\cdot)$ and $\Phi(\cdot)$ should be a strictly decreasing and increasing real functions, respectively. Noting that the real dependency between Bayes and estimation error in terms of our new fidelity complexity values, $I(\mathbf{f})$ and $R(\mathbf{f})$, is hidden and, furthermore, problem dependent, then Ψ, Φ and λ provide degrees of freedom for approximating it and consequently approaching the solution of Eq.(7). It is interesting to note that independent of those degrees of freedom, $\mathbf{f}^*(\lambda)$ resides in the sequence of representations which are the solution to a cost-fidelity problem, Eq.(10). More precisely,

$$\mathbf{f}^*(\lambda) = \arg \min_{\mathbf{f}\in\left\{\mathbf{f}^*_k : k\in K(\mathbb{D})\right\}} \Psi(I(\mathbf{f})) + \lambda \cdot \Phi(R(\mathbf{f})), \qquad (9)$$

where $\{\mathbf{f}^*_k : k \in K(\mathbb{D})\} \subset \mathbb{D}$ is the solution to

$$\mathbf{f}^*_k = \arg \max_{\substack{\mathbf{f}\in\mathbb{D} \\ R(\mathbf{f})\leq k}} I\left(\mathbf{f}\right), \ \forall k \in K(\mathbb{D}), \qquad (10)$$

with $K(\mathbb{D}) \equiv \{R(\mathbf{f}) : \mathbf{f} \in \mathbb{D}\} \subset \mathbb{N}$.

Finally, considering the analogy with a related problem in regression and classification trees [5, 6], if we know the solutions of the cost-fidelity problem, the empirical risk minimization criterion using cross validation can be adopted as the final step for solving Eq.(9) [6]. The MPE-SR formulation presented in this section was motivated by the problem of binary classification trees, Breiman *et al.* [5], where a similar formulation for finding the minimum risk *tree-structured vector quantization* (TSVQ) was presented, [6].

3.1. Connections with Structural Risk Minimization

The MPE-SR presents some interesting connections with the *structural risk minimization* (SRM) principle [8]. It is important to briefly point out some analogies and conceptual differences. First, the MPE-SR makes uses of the Bayes decision approach as a way to choose the optimal decision rule, where SRM uses the empirical risk minimization (ERM) criterion to finding the optimal empirical rule in a collection of classifiers. Furthermore, the domain to address the MPE learning problem is with respect to a family of feature transformations, where SRM has the degree of freedom in a collection of decision rules of increasing complexity, in terms of the *VC dimension* [8]. Consequently, results from the ERM inductive principle cannot be directly extended into the Bayes decision learning framework.

On the other hand, as in the ERM the MPE-SR approach provides an upper bound for controlling the generalization ability of the empirical Bayes rule with respect to the Bayes rule. At this respect, it is not possible to characterize distribution free expressions as the one obtained in ERM theory. Finally, results that formally present the tradeoff between Bayes error and estimation error across sequences of embedded representation, in Section 2, has an equivalent counterpart in the SRM formulation. This explains why the two frameworks are formulated as complexity regularization problems, for finding their respective minimum risk decision rules constrained to a finite amount of training data.

4. REDUCING THE MPE-SR TO A LINEAR DISCRIMINANT ANALYSIS PROBLEM

Putting some restriction on the raw observation $\mathcal{X}$ and more importantly on the family of representation functions $\mathbb{D}$, we can characterize different sub-problems associated with the MPE-SR complexity regularized formulation. For instance, let us consider $\mathcal{X} = \mathbb{R}^K$ and the family of linear transformations as the dictionary, $\mathbb{D} = \left\{\mathbf{f} : \mathbb{R}^K \to \mathbb{R}^m : \mathbf{f}\ linear, m \leq K\right\}$. An element $\mathbf{f} \in \mathbb{D}$ can be univocally represented by a matrix $\mathbf{A} \in \mathbb{R}(m, K)$[7]. In particular without loss of generality, we can restrict $\mathbb{D}$ to the family of full-rank matrices. For the rest of this section, we follow the modeling assumptions proposed by Padmanabhan *et al.* [7] where the problem of dimensionality reduction is addressed under some parametric assumptions and mutual information is used as the objective indicator. 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.

Considering a finite amount of training data $\{(x_i, y_i) : i = 1, .., N\}$ and maximum likelihood (ML) estimation techniques [3], the empirical distributions $\left\{\hat{p}_{X|Y}(\cdot|y) : y \in \mathcal{Y}\right\}$ and $\hat{p}_X(\cdot)$ are Gaussian and Gaussian mixtures, respectively, characterized by the empirical mean and covariance matrices given by:

$$\hat{\mu}_y = \frac{1}{N_y} \sum_{i=1}^{N} \mathbb{I}_{\{y\}}(y_i) \cdot x_i, \qquad (11)$$

$$\hat{\Sigma}_y = \frac{1}{N_y} \sum_{i=1}^{N} \mathbb{I}_{\{y\}}(y_i) \cdot (x_i - \hat{\mu}_y)(x_i - \hat{\mu}_y)^\dagger, \qquad (12)$$

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

[6] Equivalent to finding the optimal $\lambda \in \mathbb{R}^+$ and consequently the optimal $k \in K(\mathbb{D})$ associated with $\mathbf{f}^*_k$ for solving the MPE-SR.

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

Proposition 1 *Let $\mathbf{A}_1, .., \mathbf{A}_n$ be a family of full-rank linear transformations taking values in $\left\{\mathbb{R}^{k1}, .., \mathbb{R}^{kn}\right\}$ with $0 < k1 < k2 < \cdots < kn \leq K$. In addition, let us assume that the sequence of transformations is dimensionally embedded, per* Definition *1, 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 for the class conditional distributions, the empirical sequence of class conditional pdfs $\left\{\hat{p}_{\mathbf{A}_i X|Y}(\cdot|y) : i = 1, .., n\right\}$, estimated across $\left\{\mathbb{R}^{k1}, .., \mathbb{R}^{kn}\right\}$ by the ML criterion, characterize a sequence of consistent probability measures with respect to $\left\{\mathbb{R}^{k1}, .., \mathbb{R}^{kn}\right\}$, in the sense presented in* Definition *2. Proof provided in the Appendix.*

This last result formally extends *Theorem* 2, for the case of embedded sequences of full-rank linear transformations $\mathbf{A}_1, .., \mathbf{A}_n$ as stated in *Proposition* 1. This result provides justification for addressing the MPE-SR problem under the modeling assumptions presented in this section using the cost-fidelity approach. More precisely, the solution of the MPE-SR problem resides in the solution of the following problem:

$$\mathbf{A}_k^* = \arg \max_{\mathbf{A} \in \mathbb{R}(k,K)} I\left(\mathbf{A}\right), \tag{13}$$

$\forall k \in \{1, .., K\}$, where $I\left(\mathbf{A}\right)$ denotes the mutual information between rvs. $\mathbf{A}X(u)$ and $Y(u)$. For addressing this optimization problem we follow the approximations proposed in [7]. Let us use $I(\mathbf{A}) = H(\mathbf{A}X(u)) - H(\mathbf{A}X(u)|Y(u))$. Then under the Gaussian assumption and considering $\mathbf{A} \in \mathbb{R}(k, K)$, it follows that,

$$H(\mathbf{A}X(u)|Y(u) = y) = \frac{k}{2}\log(2\pi) + \frac{1}{2}\log\left(\left|\mathbf{A}\Sigma_y\mathbf{A}^\dagger\right|\right) + \frac{1}{2}.$$

Given that $\mathbf{A}X(u)$ has a Gaussian mixture distribution, a closed-form expression is not available for the differential entropy. Padmanabhan *et al.* [7] proposed to use an upper bound based on the well known fact that the Gaussian law maximizes the differential entropy under second moment constraints. Then, denoting $\Sigma \equiv \mathbb{E}(X(u)X(u)^\dagger) - \mathbb{E}(X(u))\mathbb{E}(X(u))^\dagger$, we have that $H(\mathbf{A}X(u)) \leq \frac{k}{2}\log(2\pi) + \frac{1}{2}\log\left(\left|\mathbf{A}\Sigma\mathbf{A}^\dagger\right|\right) + \frac{1}{2}$ and then

$$I(\mathbf{A}) \leq \frac{1}{2}\log\left[\frac{\left|\mathbf{A}\Sigma\mathbf{A}^\dagger\right|}{\prod_{y\in\mathcal{Y}}\left|\mathbf{A}\Sigma_y\mathbf{A}^\dagger\right|^{\mathbb{P}(Y(u)=y)}}\right]. \tag{14}$$

Then the cost-fidelity problem reduces to $\mathbf{A}_k^* =$

$$\arg \max_{\mathbf{A} \in \mathbb{R}(k,K)} \log\left[\frac{\left|\mathbf{A}\Sigma\mathbf{A}^\dagger\right|}{\prod_{y\in\mathcal{Y}}\left|\mathbf{A}\Sigma_y\mathbf{A}^\dagger\right|^{\mathbb{P}(Y(u)=y)}}\right]. \tag{15}$$

In practice, we need to address Eq.(15) based on empirical distributions estimated with a finite amount of training data. Then, it reduces to characterizing the empirical class conditional covariance matrices $\hat{\Sigma}_y$, Eq(12), and the unconditional empirical matrix $\hat{\Sigma}$ that can be written as $\hat{\Sigma}_w + \hat{\Sigma}_b$[7],

$$\hat{\Sigma}_w = \sum_{y\in\mathcal{Y}} \hat{P}_Y(\{y\}) \cdot \hat{\Sigma}_y \tag{16}$$

$$\hat{\Sigma}_b = \sum_{y\in\mathcal{Y}} \hat{P}_Y(\{y\}) \cdot (\hat{\mu} - \hat{\mu}_y)(\hat{\mu} - \hat{\mu}_y)^\dagger \tag{17}$$

where $\hat{\Sigma}_b$ and $\hat{\Sigma}_w$ are the between-class and within-class scatter matrices used in linear discriminant analysis [3] and $\hat{\mu}$ is the unconditional empirical mean. As pointed out in [7], under the additional assumption that class conditional covariance matrixes are equivalent, the problem reduces to

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

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 [3].

5. SUMMARY AND FINAL REMARKS

The minimum probability of error signal representation (MPE-SR) problem was presented and addressed based on a complexity regularization criterion. The tradeoff between signal representation quality and learning complexity was extended for a general family of dimensional embedded transformations under certain assumptions. Formal connections were made to support the use of linear discriminant analysis as a solution of the MPE-SR problem. Similar results can be derived for finite alphabet transformations of the observation space — vector quantizations — showing concrete connections with the type of complexity regularization formulation used in classification trees [5, 6].

6. APPENDIX

6.1. Proof of Theorem 2

Proof: For proving the first inequality we invoke the well-known result, [2] (*Theorem 3*), that the Bayes error bound can not decrease under deterministic transformation of the observation space. Given that $\{\mathbb{F}_i(\cdot) : i = 1, .., n\}$ is a sequence of embedded transformations, and consequently $\forall i \in \{1, .., n-1\}$, there exists a measurable mapping $\pi_{i+1,i} : \mathcal{X}_{i+1} \to \mathcal{X}_i$ such that $X_i(u) = \pi_{i+1,i}(X_{i+1}(u))$, then it follows directly that $L_{\mathcal{X}_{i+1}} \leq L_{\mathcal{X}_i}$.

For proving the inequality regarding the estimation error, a sufficient condition, given *Theorem* 1, is to prove that[8]

$$\begin{aligned} & D_{(\mathcal{X}_i,\mathcal{F}_i)}(\hat{P}_{X_i|Y}(\cdot|y)||P_{X_i|Y}(\cdot|y)) \leq \\ & D_{(\mathcal{X}_{i+1},\mathcal{F}_{i+1})}(\hat{P}_{X_{i+1}|Y}(\cdot|y)||P_{X_{i+1}|Y}(\cdot|y)), \end{aligned} \tag{18}$$

[8] We consider $D_{(\mathcal{X}_i,\mathcal{F}_i)}(\hat{P}_{X_i|Y}(\cdot|y)||P_{X_i|Y}(\cdot|y))$ as the KLD associated with the measurable space $(\mathcal{X}_i, \mathcal{F}_i)$. The space dependency in the KLD notation, which is usually implicit, is conceptually important for the rest of the proof.

$\forall i \in \{1, .., n-1\}$ and $\forall y \in \mathcal{Y}$. Without loss of generality let us prove the result for arbitrary indices i and y.

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 functions in $\{\mathbb{F}_i(\cdot) : i = 1, .., n\}$ are measurable. 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 $(\mathcal{X}, \sigma(\mathbb{F}_i))$, where $\sigma(\mathbb{F}_i)$ is the smallest sigma field that makes $\mathbb{F}_i(\cdot)$ a measurable transformation[9]. More precisely, $\sigma(\mathbb{F}_i) = \left\{\mathbb{F}_i^{-1}(B) : B \in \mathcal{F}_i\right\}$ and $\hat{P}_{X|Y}(\cdot|y)$ is constructed by: $\forall A \in \sigma(\mathbb{F}_i)$, $\exists B \in \mathcal{F}_i$, such that $A = \mathbb{F}_i^{-1}(B)$ and $\hat{P}_{X|Y}(A|y) = \hat{P}_{X_i|Y}(B|y)$.

By the consistence property of $\left\{\hat{P}_{X_i|Y}(\cdot|y) : i = 1, .., n\right\}$, 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 $\left\{\hat{P}_{X_i|Y}(\cdot|y) : i = 1, .., n\right\}$ using the aforementioned construction. It is important to note that this sequence of induced sigma fields characterizes a filtration [1], 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 $(\mathcal{X}_i, \mathcal{F}_i)$. 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 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 $(X, \sigma(\mathbb{F}_i))$, $\forall i \in \{1, .., n\}$.

The next step is to represent the KLD in the representation space Eq.(18), as a KLD in the original observation space $\mathcal{X}$ relative to a particular sigma field. Using a classical result from measure theory, it is possible to prove that [4](*Lemma 5.2.4*),

$$\begin{aligned} D_{(\mathcal{X}, \sigma(\mathbb{F}_i))}(\hat{P}_{X|Y}(\cdot|y) || P_{X|Y}(\cdot|y)) = \\ D_{(\mathcal{X}_i, \mathcal{F}_i)}(\hat{P}_{X_i|Y}(\cdot|y) || P_{X_i|Y}(\cdot|y)). \end{aligned} \tag{19}$$

Finally for proving Eq.(18), we use the following result.

LEMMA 1 *([4],* Lemma 5.2.5*) Let us consider two measurable spaces $(\mathcal{X}, \mathcal{F})$ and $(\mathcal{X}, \bar{\mathcal{F}})$, such that $\bar{\mathcal{F}}$ is a refinement of $\mathcal{F}$, in other words $\mathcal{F} \subset \bar{\mathcal{F}}$. In addition, let us consider two probability measures P_1 and P_2 defined on $(\mathcal{X}, \bar{\mathcal{F}})$, then assuming the non-trivial case $P_1 \ll P_2$, the following inequality holds,*

$$D_{(\mathcal{X}, \bar{\mathcal{F}})}(P_1 || P_2) \geq D_{(\mathcal{X}, \mathcal{F})}(P_1 || P_2). \tag{20}$$

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 Eq.(18) follows directly from *Lemma* 1 and Eq.(19). ■

6.2. Proof of Proposition 1

Proof: Without loss of generality, let us consider $\mathbf{f_1}(x) = \mathbf{A}_1 \cdot x$ and $\mathbf{f_2}(x) = \mathbf{A}_2 \cdot x, \forall x \in \mathbb{R}^K$, 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 $B_{2,1} : (\mathbb{R}^{k2}, \mathcal{B}^{k2}) \rightarrow (\mathbb{R}^{k1}, \mathcal{B}^{k1})$. Under the Gaussian assumption, this condition reduces to checking the first and second order statistics of the involved distributions. Considering the training data, it is straightforward to show that the empirical mean and covariance matrix for $\hat{P}_{\mathbf{f}_2(X)|Y}(\cdot|y)$ are 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 $\mathcal{X}$, Eqs (11) and (12). 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 consistency 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. ■

7. REFERENCES

[1] S.R.S. Varadhan, *Probability Theory*, American Mathematical Society, 2001.

[2] Nuno Vasconcelos, “Minimum probability of error image retrieval,” *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2322–2336, August 2004.

[3] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, New York: Wiley, 1983.

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

[5] Leo Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*, Belmont, CA: Wadsworth, 1984.

[6] Andrew B. Nobel, “Analysis of a complexity-based pruning scheme for classification tree,” *IEEE Transactions on Information Theory*, vol. 48, no. 8, pp. 2362–2368, 2002.

[7] M. Padmanabhan and S. Dharanipragada, “Maximizing information content in feature extraction,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 4, pp. 512–519, July 2005.

[8] Vladimir Vapnik, *Statistical Learning Theory*, John Wiley, 1998.

[9] Given that $\mathbb{F}_i : (\mathcal{X}, \mathcal{F}_{\mathcal{X}}) \rightarrow (\mathcal{X}_i, \mathcal{F}_i)$ is measurable, we have that $\sigma(\mathbb{F}_i) \subset \mathcal{F}_{\mathcal{X}}$ [1].