Dimension reduction for hidden Markov models using the sufficiency approach

Diego Tomassi[1,2,3], Liliana Forzani[1,3], Diego Milone[2], and R. Dennis Cook[4]

[1] Instituto de Matemática Aplicada del Litoral, UNL - CONICET
[2] Centro de Investigación en Señales, Sistemas e Inteligencia Computacional FICH, Universidad Nacional del Litoral - CONICET
[3] Departamento de Matemática, FIQ, Universidad Nacional del Litoral
[4] School of Statistics, University of Minnesota
diegot@santafe-conicet.gov.ar

Abstract. Dimension reduction is often included in pattern recognizers based on hidden Markov models to lower the size of the models to estimate. Commonly used methods are heuristic in nature and do not take care of information retention after projection. In this paper we present a new method based on the approach of sufficient dimension reductions. It explicitly accounts for all the discriminative information available in the original features, while using a minimum number of linear combinations of them. We review the underlying theory and present an algorithm for practical implementation of the proposed method. In the experimental side, we use simulations to illustrate its advantages over widely-used existing alternatives. In particular, we show that it performs as good as existing techniques when data is optimal according to the assumptions of those techniques, but significantly better for heteroscedastic data with no special structure on the covariance matrix.

1 Introduction

Hidden Markov models (HMM) are used frequently to model sequential data [1,2]. When the data come from different populations and the task is to classify the sequences into one of them, a different HMM can be used to model the data from each class. In this approach, if $Y = 1, 2, \ldots, h$ indicates the class and $\mathbb{X} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n\}$, with $\mathbf{X} \in \mathbb{R}^p$, are the sequences of features, the learning task is to estimate $\mathrm{p}(\mathbb{X}|Y)$ for each class within a parametric family of HMM[1]. Typically, estimation is done using maximum likelihood and class assignment for new observations is carried out using the Bayes classification rule [2,3].

Pattern recognizers based on HMM frequently include a dimension reduction stage to lower the size of statistical models. Smaller models lead to less parameters to estimate, which for a given training sample usually improves the performance of the classifier due to the smaller variance of the estimates [4,5]. A frequent choice with HMM-based classifiers is to use *linear* dimension reduction. In this type of transformations, a matrix $\boldsymbol{\rho} \in \mathbb{R}^{p \times d}, d \leq p$, is used to

[1] We use $\mathrm{p}(\cdot)$ to refer to a probability density function.

project the original features $\mathbf{X}$ onto a lower-dimensional subspace with coordinates $\boldsymbol{\rho}^T\mathbf{X} \in \mathbb{R}^d$. These d linear combinations should not lose any information carried by $\mathbf{X}$ that is relevant for discrimination. If successful, we could estimate models for $\mathrm{p}(\boldsymbol{\rho}^T\mathbf{X}|Y)$, instead of full-sized models for $\mathrm{p}(\mathbf{X}|Y)$.

Dimension reduction for Gaussian-mixture-HMM have been explored mainly in speech recognition applications [6,7,8,9,10]. All of these methods are built in the context of reduction methods for Gaussian data. The best known of these techniques are likelihood-based extensions of linear discriminant analysis for heteroscedastic data [7,9]. As the methods are stated in a maximum likelihood framework, they can be consistently embedded into the learning process of HMM. Nevertheless, it is worth noting that these methods have been derived mainly from heuristics, without taking care of information retention.

Sufficient dimension reduction (SDR) is a relatively new approach that deals explicitly with loss of information for a particular objective [11,12]. SDR developments have been more tailored to regression problems, where the essential task is to estimate the smallest subspace of $\mathbf{X}$ that does not lose any information about Y. Theory of SDR for normal data and maximum likelihood estimators of this subspace were first developed in [13] and further extended in [14,15]. In particular, the minimal linear sufficient reduction for heteroscedastic normal data is achieved by the estimator proposed in [15], named as *likelihood acquired directions* (LAD).

In this work we introduce a sufficient dimension reduction method for HMM with Gaussian observation densities based on the LAD estimator and give an algorithm for its implementation. We use simulations to show that the proposed method inherits the properties of LAD, outperforming existing dimension reduction techniques for this type of HMM.

The paper is organized as follows. In Section 2 we start by reviewing the best-known dimension reduction method for HMM, as used widely in speech recognition applications. Then, we summarize the basic concepts and results about sufficient dimension reduction and its application to normal data. In Section 3 we rely upon these results to propose a new dimension reduction method for HMM under the sufficiency approach and give an algorithm for its implementation. The advantages of the proposed method over existing ones are illustrated in Section 4, using simulated data. The paper ends with the obtained conclusions and prospective work.

2 Background

2.1 Existing dimension reduction methods for HMM

Common dimension reduction methods included in HMM-based classifiers are linear, supervised methods. Also, they are based on maximum likelihood estimation so that they can be embedded in standard training procedures for HMM. For HMM with Gaussian or Gaussian-mixture observation densities, these methods are built from projection methods for normal data [7,8,9]. In this work we

concentrate on heteroscedastic linear discriminant analysis (HLDA) as proposed in [7], which is the state-of-the-art method available in software for managing HMM [16].

HLDA is derived as follows. Assume $\mathbf{X}|(Y=y) \sim \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Delta}_y)$ and consider a full-rank linear transformation of $\mathbf{X}$ with a matrix $\boldsymbol{\Theta} = (\boldsymbol{\rho}_{\text{HLDA}}, \boldsymbol{\rho}_0)$ so that $\boldsymbol{\Theta}^T\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}_y^*, \boldsymbol{\Delta}_y^*)$, with[2]

$$\boldsymbol{\mu}_y^* = \begin{pmatrix} \boldsymbol{\rho}^T\boldsymbol{\mu}_y \\ \boldsymbol{\rho}_0^T\boldsymbol{\mu} \end{pmatrix} \qquad \boldsymbol{\Delta}_y^* = \begin{pmatrix} \boldsymbol{\Omega}_y & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Omega}_0 \end{pmatrix},$$

where $\boldsymbol{\mu} = \mathrm{E}(\boldsymbol{\mu}_y)$ and $\boldsymbol{\Omega}_0$ is shared between all the classes. In this way, $\boldsymbol{\rho}_{\text{HLDA}}^T\mathbf{X}$ is independent of $\boldsymbol{\rho}_0^T\mathbf{X}$ and the latter is constant for all classes y. Thus, $\boldsymbol{\rho}_0^T\mathbf{X}$ does not carry any discriminative information and can be ignored for classification. Without loss of generality, assume that $\boldsymbol{\Theta}$ is an orthogonal matrix and that $\boldsymbol{\rho}_{\text{HLDA}}$ is semi-orthogonal. From [7], the optimum matrix $\boldsymbol{\Theta}$ maximizes the log-likelihood function

$$\mathcal{L}_{\text{HLDA}}(\boldsymbol{\Theta}) = -\frac{N}{2}\log|\boldsymbol{\rho}_0^T\tilde{\boldsymbol{\Sigma}}\boldsymbol{\rho}_0| - \frac{1}{2}\sum_{y=1}^{h} N_y \log|\boldsymbol{\rho}_{\text{HLDA}}^T\tilde{\boldsymbol{\Delta}}_y\boldsymbol{\rho}_{\text{HLDA}}|, \tag{1}$$

where $\tilde{\boldsymbol{\Sigma}}$ is the sample marginal covariance matrix,N_y is the sample size for population y and $N = \sum_y N_y$. The optimum does not have a closed-form solution, so numerical techniques must be employed [7,8]. Notice that in this derivation, beginning with normality for $\mathbf{X}|Y$, restrictions are imposed in the transformed feature space, not in the original space of $\mathbf{X}$. Also, the models assumed in the transformed space are strongly structured to allow statistical independence between $\boldsymbol{\rho}_{\text{HLDA}}^T\mathbf{X}$ and $\boldsymbol{\rho}_0^T\mathbf{X}$. Note that a weaker condition to reject the linear combinations $\boldsymbol{\rho}_0^T\mathbf{X}$ for classification is to have $\mathrm{p}(\boldsymbol{\rho}_0^T\mathbf{X}|\boldsymbol{\rho}^T\mathbf{X}, Y=y)$ independent of y. This is exploited in the sufficiency approach we discuss next.

2.2 Sufficient dimension reduction

Sufficient dimension reduction is a methodology that deals explicitly with information retention for a particular objective. Here we review the main facts that will be used in following sections. Formally, a linear reduction $\boldsymbol{\rho}^T\mathbf{X} \in \mathbb{R}^d$, with $d \leq p$ is sufficient if [13]

$$\mathbf{X}|(Y, \boldsymbol{\rho}^T\mathbf{X}) \sim \mathbf{X}|\boldsymbol{\rho}^T\mathbf{X}. \tag{2}$$

This definition implies that $\boldsymbol{\rho}^T\mathbf{X}$ carries all the information about Y that is contained in $\mathbf{X}$. Note that $\mathbf{X}$ is always a sufficient reduction. Thus, the essential tasks in SDR are to characterize and estimate the *smallest* sufficient reduction.

[2] Notation $S|V$ accounts for random variable S conditional on the value of a random variable V. $\mathrm{E}(\cdot)$ denotes mathematical expectation.

In addition, $\boldsymbol{\rho}$ is not unique, but we can identify the subspace spanned by the columns of $\boldsymbol{\rho}$. This subspace $\mathcal{S}_{\boldsymbol{\rho}} = \text{span}(\boldsymbol{\rho})$ is called a *dimension reduction subspace*. The smallest dimension reduction subspace is called the *central subspace* [11,17] and it is the inferential target in SDR.

Notice that the central subspace contains all the information to describe the classes Y, not only to discriminate between them. In a classification setting, the goal is actually to find a subspace of the features so that class assignment is conditionally independent of the remaining information in $\mathbf{X}$ [18,19]. The sufficient subspace for discrimination may be smaller, but it is always contained in the central subspace [18]. Nevertheless, when using the Bayes rule for classification, it was shown in [18] that this smallest discriminant subspace is the same as the central subspace when the data from each class is normally distributed. Thus, for this type of data, theory developed for regression tasks can be applied straightforwardly to discrimination problems, achieving reductions that are sufficient and also optimal from a minimality point of view. That is, we cannot find a dimension reduction subspace for discrimination that is smaller than the central subspace for this kind of data.

2.3 SDR for normal data

In this section we review the main results on SDR for normally distributed data. They will be used in Section 3 to build a reduction method for HMM that overcomes the limitations of state-of-the-art methods when the covariance matrix of conditional normal densities is not constrained to a particular structure.

Assume that $\mathbf{X}|(Y = y) \sim \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Delta}_y)$, for classes $y = 1, 2, \ldots, h$, and let $\boldsymbol{\Delta} = \text{E}(\boldsymbol{\Delta}_y)$. It is shown in [15] that $\mathcal{S}_{\boldsymbol{\rho}} = \text{span}(\boldsymbol{\rho}) \in \mathbb{R}^p$ is a sufficient dimension reduction subspace if and only if the subspace spanned by $\boldsymbol{\Delta}\boldsymbol{\rho}$ is an invariant subspace of $\boldsymbol{\Delta}_y - \boldsymbol{\Delta}$ and the translated means $\boldsymbol{\mu}_y - \boldsymbol{\mu}$ fall also in that subspace[3]. Under these conditions, the means and covariance matrices of the class models are [15]

$$\begin{aligned} \boldsymbol{\mu}_y &= \boldsymbol{\mu} + \boldsymbol{\Delta}\boldsymbol{\rho}\boldsymbol{\nu}_y, \\ \boldsymbol{\Delta}_y &= \boldsymbol{\Delta} + \boldsymbol{\Delta}\boldsymbol{\rho}\mathbf{T}_y\boldsymbol{\rho}^T\boldsymbol{\Delta}, \end{aligned} \tag{3}$$

for some $\boldsymbol{\nu}_y \in \mathbb{R}^d$, $\mathbf{T}_y \in \mathbb{R}^{d\times d}$ and $d = \dim(\mathcal{S}_{\boldsymbol{\rho}})$, with $\bar{\boldsymbol{\nu}} = \text{E}(\boldsymbol{\nu}_y) = 0$, and $\text{E}(\mathbf{T}_y) = \mathbf{0}$ to agree with the definition of $\boldsymbol{\mu}$ and $\boldsymbol{\Delta}$. It is important to emphasize that (3) are necessary and sufficient conditions derived from theory; they are not assumptions set *a priori* to derive the subspace projection method.

The model stated in (3) can be used to derive the distributions of $\boldsymbol{\rho}^T\mathbf{X}|(Y = y)$ and $\boldsymbol{\rho}_0^T\mathbf{X}|(\boldsymbol{\rho}^T\mathbf{X}, Y = y)$. With them, we can find an estimator of $\mathcal{S}_{\boldsymbol{\rho}}$ using maximum likelihood estimation. Let $\boldsymbol{\rho}$ be a semi-orthogonal basis matrix for $\mathcal{S}_{\boldsymbol{\rho}} \subseteq \mathbb{R}^p$ and let $(\boldsymbol{\rho}, \boldsymbol{\rho}_0) \in \mathbb{R}^{p\times p}$ be an orthogonal matrix. The likelihood of the

[3] $\mathcal{S} \in \mathbb{R}^p$ is an invariant subspace of $\mathbf{A} \in \mathbb{R}^{p\times p}$ if $\mathbf{A}\mathcal{S} \subseteq \mathcal{S}$.

training sample reads

$$\ell(\boldsymbol{\rho}; \mathbf{X}) = \prod_{y=1}^{h} \prod_{n=1}^{N_y} \mathrm{p}(\boldsymbol{\rho}^T \mathbf{X} | Y = y) \mathrm{p}(\boldsymbol{\rho}_0^T \mathbf{X} | \boldsymbol{\rho}^T \mathbf{X}, Y = y). \tag{4}$$

From here, it is shown in [15] that the minimal linear sufficient reduction is $\boldsymbol{\rho}_{\mathrm{LAD}}^T \mathbf{X}$, where $\boldsymbol{\rho}_{\mathrm{LAD}}$ maximizes over the Grassmann manifold of dimension d in $\mathbb{R}^p$ the log-likelihood function

$$\mathcal{L}_{\mathrm{LAD}}(\boldsymbol{\rho}) = \mathrm{const} + \frac{N}{2} \log |\boldsymbol{\rho}^T \tilde{\boldsymbol{\Sigma}} \boldsymbol{\rho}| - \frac{1}{2} \sum_y N_y \log |\boldsymbol{\rho}^T \tilde{\boldsymbol{\Delta}}_y \boldsymbol{\rho}|. \tag{5}$$

This is the LAD estimator, as named in the Statistics literature. Numerical optimization is needed to find a solution to the optimization problem.

It is important to see also that HLDA as proposed in [7] is a especial case of LAD, when $\boldsymbol{\Delta}$ in (3) has the very special structure [20]

$$\boldsymbol{\Delta} = \boldsymbol{\rho} \boldsymbol{\Omega} \boldsymbol{\rho}^T + \boldsymbol{\rho}_0 \boldsymbol{\Omega}_0 \boldsymbol{\rho}_0^T,$$

with $\boldsymbol{\Omega} = \mathrm{E}(\boldsymbol{\Omega}_y) = \mathrm{E}(\boldsymbol{\rho}^T \boldsymbol{\Delta}_y \boldsymbol{\rho})$. Due to this covariance constraint, if $\mathrm{span}(\boldsymbol{\rho}_{\mathrm{HLDA}})$ and $\mathrm{span}(\boldsymbol{\rho}_{\mathrm{LAD}})$ are both dimension reduction subspaces, it can be shown that $\mathrm{span}(\boldsymbol{\rho}_{\mathrm{LAD}}) \subseteq \mathrm{span}(\boldsymbol{\rho}_{\mathrm{HLDA}})$. This means that HLDA often needs to retain more directions than LAD in order to conserve all the information available in the original features.

3 SDR for HMM

3.1 Derivation

When an homogeneous, Gaussian HMM ϑ_i is used to model a sequence from the i-th class, it is assumed that each random vector of features $\mathbf{X}_t$ comes from a normal population, conditional on the state q_t of the underlying Markov chain at time t; that is $\mathrm{p}(\mathbf{X}_t | q_t, \vartheta_i) = \mathcal{N}(\boldsymbol{\mu}_{q_t}, \boldsymbol{\Delta}_{q_t})$. In this scenario, $\boldsymbol{\rho}$ is a basis matrix for a dimension reduction subspace if

$$\mathbf{X}_t | (\boldsymbol{\rho}^T \mathbf{X}_t, q_t, \vartheta_i) \sim \mathbf{X}_t | \boldsymbol{\rho}^T \mathbf{X}_t. \tag{6}$$

In this way, $\boldsymbol{\rho}^T \mathbf{X}_t$ and $\mathbf{X}_t$ have the same information on the state q_t of the model for class i, for $i = 1, 2, \ldots, h$. Thus, if we map (q_t, ϑ_i) onto a single index j, we recover the condition for SDR of normal data as discussed in Section 2. It is important to stress that the different populations for the dimension reduction task are all the conditional observation models in each HMM. Thus, if we have h classes, each linked to a HMM with a state space of dimension N_q, the dimension reduction task will involve hN_q normal populations.

There still remain a point to take care about. The dimension reduction problem for normal models discussed previously was a fully supervised one. That is,

Algorithm

- **Initialization**
  1. For each class i, set $\vartheta_i^* = \vartheta_i$ and $\mathcal{X}_i^* = \mathcal{X}_i$.
  2. Let $\mathcal{X}^* = \bigcup_i \mathcal{X}_i^*$.
- **Main loop:** repeat until convergence
  1. For each class i, infer the most probable sequences of states $\{\mathbf{q}^*\}_i$ that originated the data $\mathcal{X}_i^*$ according to ϑ_i^*.
  2. Form the whole labelled dataset $\Upsilon = \bigcup_i (\{\mathbf{q}^*\}_i, \mathcal{X}_i)$.
  3. Estimate the semiorthogonal basis matrix $\boldsymbol{\rho}^*$ using Υ and one of the reduction methods for normal models (LAD or HLDA).
  4. Compute $\boldsymbol{\rho}_0^*$ that spans a subspace orthogonal to span($\boldsymbol{\rho}^*$).
  5. Build the orthogonal matrix $\boldsymbol{\Theta} = (\boldsymbol{\rho}^* \boldsymbol{\rho}_0^*)$.
  6. Lineary transform the original dataset using $\boldsymbol{\Theta}$ to obtain a new $\mathcal{X}^*$.
  7. For each class i, update the observation model corresponding to each estate y of ϑ_i, doing $\boldsymbol{\mu}_y^* = \boldsymbol{\Theta}^T \boldsymbol{\mu}_y$ and $\boldsymbol{\Delta}_y^* = \boldsymbol{\Theta}^T \boldsymbol{\Delta}_y \boldsymbol{\Theta}$.
- **Finalization**
  1. Set $\widehat{\boldsymbol{\rho}} = \boldsymbol{\rho}^*$.
  2. For each class i, take the first d coordinates of the model parameters in ϑ_i^* to build the final estimates of the models, $\widehat{\vartheta}_i$.

Fig. 1: Proposed algorithm for embedding sufficient dimension reduction into the training process of a HMM-based classifier.

for each observation in the training set we knew the population from where it came. This is not the case in the current setting, as the states q_t are hidden to the observer. Thus, data must be labelled in q in some way to apply the SDR methods for normal data to HMM.

A first approach to get the labels is to train the HMM with the standard Baum-Welch algorithm using the original features and, in a second step, to use these trained models to make inference about the most probable sequences of states that describe the training sequences. In this manner, the inferred states can be used as labels for the observed vectors of random features, and SDR would be applied over this labelled dataset to obtain a basis matrix $\widehat{\boldsymbol{\rho}}$ for the dimension reduction subspace. Despite this procedure being appealing to generate a labelled dataset to apply the SDR methodology, it does not considers the central subspace properly as another parameter to estimate. Furthermore, it has been shown in previous work that embedding the estimation of the projection matrix within the iterative estimation of the parameters of the models achieves better results in classifcation [7,9]. In the next section we propose a simplified training algorithm to account for this.

3.2 Algorithm

Dimension reduction methods using likelihood-based estimators can be embedded within the Expectation-Maximization (EM) algorithm, as exploited previously in [6,7,8]. For simplicity, we show here a Viterbi-like algorithm instead of the full Baum-Welch approach (see [3] for details about these algorithms).

Assume we have pre-initialized HMM models ϑ_i, one for each class. Let $\mathcal{X}_i$ be the training set for class i and let $\mathcal{X} = \bigcup_i \mathcal{X}_i$ refer to the whole training set. The proposed algorithm is shown in Figure 1. It is interesting to note that inference about the most probable sequences of states is carried out in a (transformed) feature space without rejection of any coordinate. On the other hand, classification is carried out in the reduced feature subspace with dimension d, after projecting the new data with the estimated matrix $\widehat{\boldsymbol{\rho}}$. This alternative was found more stable than using just the d-dimensional projected features during the estimation process. The reason behind this is that information loss after rejecting coordinates in the first iteration cannot be recovered later. Thus, as the first reduction is driven by a labelling process carried out using roughly trained models, it would be common to loss important information in the rejected coordinates at the beginning of the algorithm. It would compromomise the evolution of the estimation unless very good initial estimates of model parameters are provided. Using all the transformed features during the iterative estimation, the algorithm was found stable and convergence was reached typically after a few iterations.

4 Simulation

4.1 Set up

We ran a simulation study for a two-class discrimination problem. For this experiment, data for each class was generated using a corresponding HMM with Gaussian observation densities. The number of hidden states was set to three ($N_q = 3$) for both models. Conditional on the state of the Markov chain, observed data was generated from a normal population with parameters

$$\begin{aligned}\boldsymbol{\mu}_j &= \boldsymbol{\rho}(\boldsymbol{\nu}_j - \bar{\boldsymbol{\nu}}_j),\\ \boldsymbol{\Delta}_j &= \boldsymbol{\Delta} + \boldsymbol{\Delta}\boldsymbol{\rho}(\boldsymbol{\Omega}_j - \boldsymbol{\Omega})\boldsymbol{\rho}^T\boldsymbol{\Delta},\end{aligned}$$

with $j = 1, 2, \ldots, 6$, $\bar{\boldsymbol{\nu}}_j = \sum_j \boldsymbol{\nu}_j/6$, $\boldsymbol{\Omega} = \sum_j \boldsymbol{\Omega}_j/6$ and $\boldsymbol{\Delta} = \boldsymbol{\rho}\boldsymbol{\Omega}\boldsymbol{\rho}^T + \boldsymbol{\rho}_0\boldsymbol{\Omega}_0\boldsymbol{\rho}_0^T$. Note that this normal model fulfills the conditions to make HLDA an optimal method for dimension reduction. Our objective in choosing this is two-fold: on the one hand, we want to show that when the data is exactly as assumed by HLDA, the reduction obtained with LAD is as good as the one obtained with HLDA. On the other hand, if this original data is lineary transformed with a nonsingular matrix $\boldsymbol{\eta}$, the covariance structure gets broken and HLDA is no longer optimal. We want to show that in this case, which also accounts for a general covariance matrix of the populations, LAD is significantly better than HLDA. Furthermore, this condition should illustrate that error rates achieved

HMM for class 1		
$\mathbf{A}_1 = \begin{pmatrix} 0.60 & 0.35 & 0.05 \\ 0 & 0.75 & 0.25 \\ 0 & 0 & 1.00 \end{pmatrix}$		
state 1	*state 2*	*state 3*
$\boldsymbol{\nu}_{1,1} = (1,-3)^T$	$\boldsymbol{\nu}_{1,2} = (4,2)^T$	$\boldsymbol{\nu}_{1,3} = (3,-1)^T$
$\boldsymbol{\Omega}_{2,1} = \begin{pmatrix} 1.00 & -0.25 \\ -0.25 & 3.00 \end{pmatrix}$	$\boldsymbol{\Omega}_{2,2} = \begin{pmatrix} 2.00 & 1.50 \\ 1.50 & 5.00 \end{pmatrix}$	$\boldsymbol{\Omega}_{1,3} = \begin{pmatrix} 1.00 & -0.25 \\ -0.25 & 1.00 \end{pmatrix}$

HMM for class 2		
$\mathbf{A}_2 = \begin{pmatrix} 0.75 & 0.15 & 0.10 \\ 0 & 0.75 & 0.25 \\ 0 & 0 & 1.00 \end{pmatrix}$		
state 1	*state 2*	*state 3*
$\boldsymbol{\nu}_{2,1} = (-1,0)^T$	$\boldsymbol{\nu}_{2,2} = (2,2)^T$	$\boldsymbol{\nu}_{2,3} = (2,-3)^T$
$\boldsymbol{\Omega}_{1,1} = \begin{pmatrix} 3.00 & 0.25 \\ 0.25 & 1.00 \end{pmatrix}$	$\boldsymbol{\Omega}_{2,2} = \begin{pmatrix} 2.00 & 1.50 \\ 1.50 & 5.00 \end{pmatrix}$	$\boldsymbol{\Omega}_{2,3} = \begin{pmatrix} 1.00 & -0.45 \\ -0.45 & 1.00 \end{pmatrix}$

Table 1: HMM parameters used in the simulation.

using LAD-derived estimators remains fairly the same after transforming the features, due to the equivariance property of the estimator [15].

We set $p = 10$ and a central subspace of dimension $d = 2$. Table 1 shows the values set for HMM parameters in the sufficient subspace. Note that there is not anything special in the chosen values. They could have been set at random, but specific values have been prefered to make the experiment easily reproducible. A training set and an independent testing set were randomly generated for each class using the model parameters stated above. Each generated sequence $\mathbb{X}^n = \{\mathbf{X}_1, \ldots, \mathbf{X}_{T_n}\}$ had a number T_n of feature vectors which varied randomly as $2N_q \leq T_n \leq 5N_q$. Each feature vector $\mathbf{X}_t$ was drawn from a Gaussian density conditional on the state q_t of the related hidden Markov chain at that time. For the dimension reduction stage, computations were carried out using the software available from [21].

4.2 Results

We compared the performance of the following classifiers: i) $\text{HMM}_{\text{NORED}}$, in which each HMM was trained with the Baum-Welch algorithm using the original 10-dimensional feature space; ii) $\text{HMM}_{\text{EXT-LAD}}$, which includes dimension reduction using LAD but is not embedded in HMM training; iii) HMM_{LAD}, in which LAD is embedded in the iterative training process of the HMM, using the algorithm described in Section 3.2; and iv) HMM_{HLDA}, in which the embedded method is HLDA.

Sample size	$\text{HMM}_{\text{NORED}}$	$\text{HMM}_{\text{EXT-LAD}}$	HMM_{HLDA}	HMM_{LAD}
2×100	0.1465	0.0805	0.0425	0.0220
2×1000	0.1080	0.0928	0.0424	0.0229
2×5000	0.1949	0.1014	0.0571	0.0234

Table 2: Error rate obtained with each classifier for different sizes of the training set. Reported values are means over ten runs of the experiment.

Sample size	$\text{HMM}_{\text{NORED}}$	$\text{HMM}_{\text{EXT-LAD}}$	HMM_{HLDA}	HMM_{LAD}
2×100	0.1445	0.1045	0.2045	0.0235
2×1000	0.1153	0.0954	0.1698	0.0222
2×5000	0.1549	0.1043	0.1925	0.0237

Table 3: Error rate obtained with each classifier for different sizes of the training set, after transformation of the original data with a nonsingular matrix $\boldsymbol{\eta}$. Reported values are means over ten runs of the experiment. In these experiments, data was obtained linearly transforming the datasets used in Table 2.

We ran the experiment for different sizes of the training set. In all the cases, classification was carried out over independent test sets with the same size as the training set used in the given experiment. The same datasets and the same initial estimates of the models were used for all the classifiers, so that random initialization has no effect on the relative performance of the tested schemes.

Table 2 shows the obtained results. Reported error rates are mean values over ten runs of the experiment. It can be seen that embedding the estimation of the reduction into HMM training provides better results that reducing dimensionality externally or not reducing at all, as suggested from the superior performance of both HMM_{HLDA} and HMM_{LAD} over $\text{HMM}_{\text{EXT-LAD}}$ and $\text{HMM}_{\text{NORED}}$. It is clearly seen also that HMM_{LAD} outperforms the other alternatives; results are significant at the 5% level for each size of the training set. It is interesting to note that the superiority of HMM_{LAD} is significant even at the 1% level for the smallest training sample. Thus, despite the fact that HLDA provides *a priori* a more parsimonious description of the data, estimation of such structured model is actually harder in practice. Boxplots of the achieved error rates are shown in Figure 2.a) to c). It can be seen also that achieved error rates show significant less variability using the embedded LAD method than with the other methods.

Finally, let us consider the effect of transforming the dataset with a nonsingular matrix $\boldsymbol{\eta} \in \mathbb{R}^{p \times p}$ generated randomly. Obtained results are given in Table 3, and corresponding boxplots are shown in Figure 2.d) to f). If we concentrate on the difference in error rate achieved with each reduction method after transforming the features compared to its performance with the original features, it

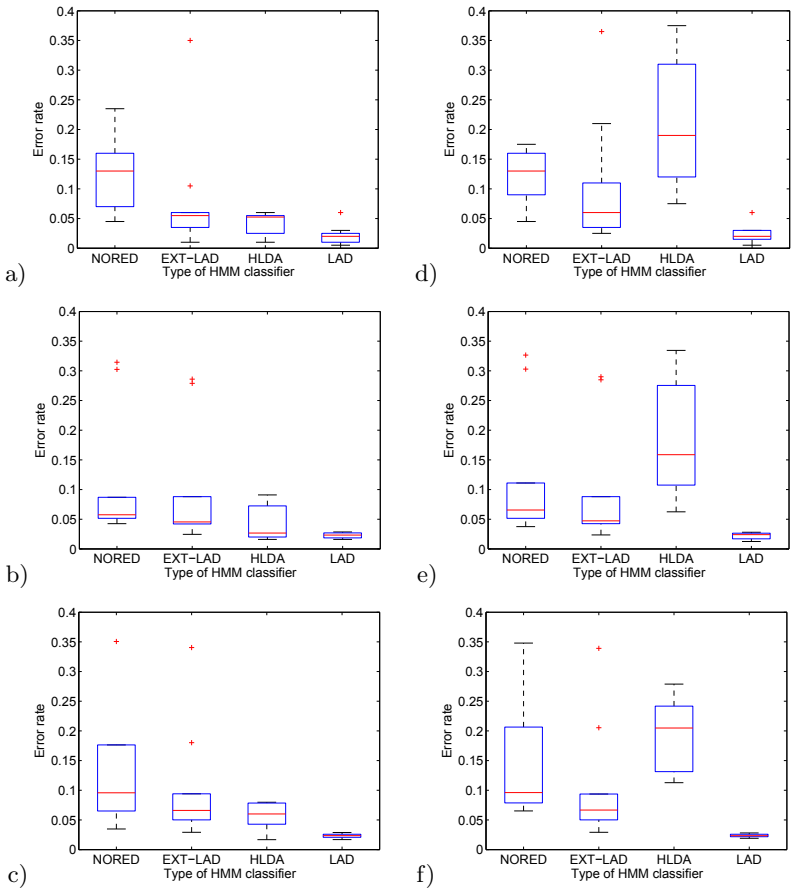

Fig. 2: Error rates achieved with classifiers $\mathrm{HMM}_{\mathrm{NORED}}$, $\mathrm{HMM}_{\mathrm{EXT\text{-}LAD}}$, $\mathrm{HMM}_{\mathrm{HLDA}}$ and $\mathrm{HMM}_{\mathrm{LAD}}$ for different sizes of the training and testing samples. a) 100 sequences per class; b) 1000 sequences per class; c) 5000 sequences per class; d)-f) same as a)-c), respectively, but after linear transformation of the data with a matrix $\boldsymbol{\eta}$.

is found that increments are not significant for $\mathrm{HMM}_{\mathrm{NORED}}$, $\mathrm{HMM}_{\mathrm{EXT\text{-}LAD}}$ and $\mathrm{HMM}_{\mathrm{LAD}}$, but $\mathrm{HMM}_{\mathrm{HLDA}}$ is strongly affected by the transformation. This clearly illustrates the fact that for a fixed dimension of the dimension reduction subspace, HLDA is an optimal reduction only for a very particular covariance structure. Thus, more directions should be retained for conservation of the original information. In addition, it is important to note that boxplots remain fairly the same after transforming the features for the reduction methods involving LAD, albeit some increase in variability for the smallest training sample.

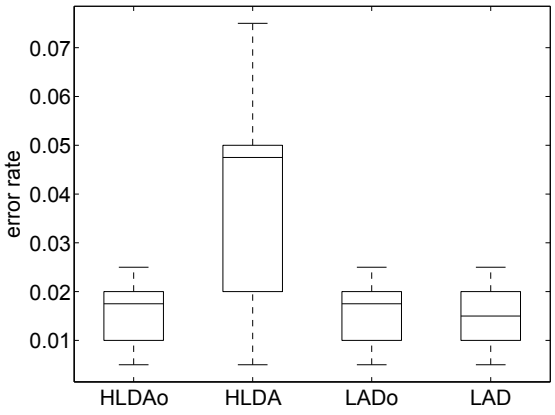

Fig. 3: Comparison of error rates achieved when using inference on the sequences of states that generated the observations (HMM_{HLDA} and HMM_{LAD}), against using the true paths for labelling the data for the dimension reduction stage ($\text{HMM}_{\text{HLDAo}}$ and HMM_{LADo}).

All of these results show the main advantages of using a reduction method based upon the LAD estimator for normal populations. It is worth noting that for HMM the LAD method seems to achieve significant better results than HLDA even for data generated from conditional normal models with the covariance structure assumed by HLDA. The reason is that the data transformations induced by LAD during the iterative learning process allow for better inference of the sequences of hidden states that most likely generated the observed data, which in turn helps to get a better labelling of the data as needed for the reduction. To show this, we compared the error rates achieved embedding HLDA and LAD as proposed in Section 3.2, against the error rates achieved when the true sequences of states that generated the data are used to label the data before the dimension reduction stage. Results are shown in Figure 3. It can be seen that, unlike LAD, inference on the optimal sequence of states degrades the performance of HLDA significantly.

5 Conclusions

In this paper we presented a subspace projection method for Gaussian-hidden Markov models that achieves a minimal sufficient reduction of the feature space. Simulations shown that the proposed method clearly outperforms them when no further structure in the covariance matrices is assumed, and that it is as good as these techniques even for their most favorable conditions. Experiments with real data are needed to further validate the method in real-world scenarios. In addition, future work should also address projection schemes onto multiple subspaces.

References

[1] Bishop, C.: Pattern Recognition and Machine Learning. Springer (2007)
[2] Rabiner, L.R., Juang, B.H.: An introduction to hidden Markov models. IEEE Acoustics Speech and Signal Processing Magazine **3**(1) (January 1986) 4–16
[3] Rabiner, L., Juang, B.: Fundamentals of Speech Recognition. Prentice-Hall (1993)
[4] Fukunaga, K.: Introduction to Statistical Pattern Recognition. Academic Press, New York (1990)
[5] Jain, A., Duin, R., Mao, J.: Statistical pattern recognition: a review. IEEE Transactions on Pattern Analysis and Machine Intelligence **22** (2000) 4–37
[6] Kumar, N.: Investigation of silicon auditory models and generalization of linear discriminant analysis for improved speech recognition. PhD thesis, John Hopkins University, Baltimore (1997)
[7] Kumar, N., Andreou, A.: Heteroscedastic discriminant analysis and reduced-rank HMMs for improved speech recognition. Speech Communication **26** (1998) 283–297
[8] Gales, M.: Maximum likelihood multiple subspace projections for hidden markov models. IEEE Transactions on Speech and Audio Processing **10** (2002) 37–47
[9] Saon, G., Padmanabhan, M., Gopinath, R., Chen, S.: Maximum likelihood discriminant feature spaces. Acoustics, Speech, and Signal Processing, IEEE International Conference on **2** (2000) 1129–1132
[10] Zhou, H., Karakos, D., Khudanpur, S., Andreou, A., Priebe, C.: On projections of gaussian distributions using maximum likelihood criteria. (2009) 431–438
[11] Cook, R.: Using dimension reduction subspaces to identify important inputs in models of physical systems. In: Proceedings of the Section on Physical and Engineering Sciences, American Statistical Association. (1994) 18–25
[12] Li, K.: Sliced inverse regression for dimension reduction (with discussion). Journal of the American Statistical Association **86** (1991) 316–342
[13] Cook, R.: Fisher lecture: Dimension reduction in regression (with discussion). Statistical Science **22** (2007) 1–26
[14] Cook, R., Forzani, L.: Principal fitted components in regression. Statistical Science **23** (2008) 485–501
[15] Cook, R., Forzani, L.: Likelihood-Based sufficient dimension reduction. Journal of the American Statistical Association **104**(485) (2008) 197–208
[16] Young, S., Young, S.: The HTK hidden markov model toolkit: Design and philosophy. Technical Report, Entropic Cambridge Research Laboratory, Ltd **2** (1994) 2–44
[17] Cook, R.: Regression Graphics. Wiley, New York (1998)
[18] Cook, R., Yin, X.: Dimension reduction and visualization in discriminant analysis (with discussion). Australia New Zeland Journal of Statistics (1994) 18–25
[19] Zhang, J., Liu, Y.: SVM decision boundary based discriminative subspace induction. Pattern Recognition **38**(10) (2005) 1746–1758
[20] Tomassi, D.: Información discriminativa en clasificadores basados en modelos ocultos de Markov. PhD thesis, Universidad Nacional del Litoral, Santa Fe, Argentina (2011)
[21] Cook, R., Forzani, L., Tomassi, D.: LDR: a package for likelihood-based dimension reduction. Journal of Statistical Software **39** (2011)