Feature selection using factor analysis for Alzheimer's diagnosis using ^{18}F-FDG PET images

D. Salas-Gonzalez,[a] J. M. Górriz, J. Ramírez, I. A. Illán, M. López, F. Segovia,
R. Chaves, and P. Padilla
*Department of Signal Theory, Networking and Communications, ETSIIT, 18071, University of Granada,
Granada, Spain*

C. G. Puntonet
*Department of Computer Architecture and Computer Technology, ETSIIT, 18071, University of Granada,
Granada, Spain*

(The Alzheimer's Disease Neuroimage Initiative[b])

Purpose: This article presents a computer-aided diagnosis technique for improving the accuracy of the early diagnosis of Alzheimer's disease (AD). Two hundred and ten ^{18}F-FDG PET images from the ADNI initiative [52 normal controls (NC), 114 mild cognitive impairment (MCI), and 53 AD subjects] are studied.
Methods: The proposed methodology is based on the selection of voxels of interest using the t-test and a posterior reduction of the feature dimension using factor analysis. Factor loadings are used as features for three different classifiers: Two multivariate Gaussian mixture model, with linear and quadratic discriminant function, and a support vector machine with linear kernel.
Results: An accuracy rate up to 95% when NC and AD are considered and an accuracy rate up to 88% and 86% for NC-MCI and NC-MCI,AD, respectively, are obtained using SVM with linear kernel.
Conclusions: Results are compared to the voxel-as-features and a PCA- based approach and the proposed methodology achieves better classification performance. © *2010 American Association of Physicists in Medicine*. [DOI: 10.1118/1.3488894]

Key words: ^{18}F-FDG PET brain imaging, computer-aided diagnosis, Alzheimer's disease, image classification

I. INTRODUCTION

Alzheimer's disease (AD) is the most common cause of dementia in the elderly and affects approximately 30×10^6 individuals worldwide. With the growth of the older population in developed nations, the prevalence of AD is expected to quadruple over the next 50 yr (Refs. 1 and 2), while its early diagnosis remains being a difficult task. Therefore, Alzheimer's disease is a severe global public health problem. Nowadays, a large amount of potential treatments are under study. The early stages of Alzheimer's disease are difficult to diagnose. A definitive diagnosis is usually made once cognitive impairment compromises daily living activities. A patient will progress from mild cognitive problems, such as memory loss through increasing stages of cognitive and non-cognitive disturbances, and eventually causing death.[3]

It has been evidenced in the literature that a reduction in the cerebral metabolic rate of glucose in some regions of the brain is related to AD. AD images present lower rate of glucose metabolism in bilateral regions in the temporal and parietal lobes, posterior cingulate gyri and precunei, and also in other parts of the brain for more severe stage of the dementia. Nevertheless, brain hypometabolism is also presented in other neurodegenerative disorders. Reductions in the cerebral metabolic rate of glucose is also detected in mild cognitive impairment (MCI) subjects. Furthermore, patients labeled as MCI present a set of symptoms that might indicate the start of the disease.[4–8]

Alzheimer's diagnosis is usually based on the information provided by a careful clinical examination, a thorough interview of the patient and relatives, and a neuropsychological assessment.[9–12] An emission computed tomography (ECT) study is frequently used as a complementary diagnostic tool in addition to the clinical findings.[8,13] However, in late-onset AD, there is minimal perfusion in the mild stages of the disease and age-related changes, which are frequently seen in healthy aged people, have to be discriminated from the minimal disease-specific changes. These minimal changes in the images make visual diagnosis a difficult task that requires experienced observers.

In ECT imaging, the dimension of the feature space (number of voxels) is very large compared to the number of available training samples (usually ~ 100 images). This scenario leads to the so-called small sample size problem,[14] as the number of available samples is greater than the number of images. Therefore, a reduction in the dimension of the feature vector is desirable before performing classification. In this work, a computer-aided diagnosis (CAD) system for assisting Alzheimer's diagnosis is proposed. This methodology

is based on a preliminary automatic selection of voxels of interest using t-test. Then, selected voxels are modeled using factor analysis (FA). The observed variables are modeled as linear combinations of the factors and factor loadings are used to describe variability among selected voxels in terms of fewer unobserved variables. These factor loadings will be used as feature vectors and they allow us to reduce the dimension of the problem surmounting the small sample size problem. Classification is performed using three different approaches: Fitting a multivariate normal density to each group with a pooled estimate of covariance; fitting a multivariate normal density with covariance estimates stratified by group; and using support vector machine with linear kernels.

This work is organized as follows. In Sec. II the image acquisition, preprocessing, feature extraction, and classification methods used in this paper are presented. In Sec. III, we summarize the classification performance obtained by applying the proposed methodology and in Sec. IV, the results are discussed. Lastly, the conclusions are drawn in Sec. V.

II. MATERIAL AND METHODS

II.A. Image acquisition and labeling

^{18}F-FDG positron emission tomography (PET) images used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (www.loni.ucla.edu/ADNI). The ADNI was launched in 2003 by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, the Food and Drug Administration, private pharmaceutical companies, and nonprofit organizations, as a 60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials. The principal investigator of this initiative is Michael W. Weiner, M.D., VA Medical Center and University of California-San Francisco. ADNI is the result of efforts of many coinvestigators from a broad range of academic institutions and private corporations and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 adults, ages 55–90, to participate in the research: ∼200 cognitively normal older individuals to be followed for 3 yr, 400 people with MCI to be followed for 3 yr, and 200 people with early AD to be followed for 2 yr.

^{18}F-FDG PET scans were acquired according to a standardized protocol. A 30 min dynamic emission scan, consisting of six 5 min frames, was acquired starting 30 min after the intravenous injection of 5.0 ± 0.5 mCi of ^{18}F-FDG, as the subjects, who were instructed to fast for at least 4 h prior to the scan, lay quietly in a dimly lit room with their eyes open and minimal sensory stimulation. Data were corrected for radiation-attenuation and scatter using transmission scans from Ge-68 rotating rod sources and reconstructed using measured attenuation correction and image reconstruction algorithms specified for each scanner (www.loni.ucla.edu/ADNI/Data/ADNIData.shtml). Following the scan, each image was reviewed for possible artifacts at the University of Michigan and all raw and processed study data were archived.

Enrolled subjects were between 55 and 90 (inclusive) yr of age. Images were labeled as normal subjects, MCI subjects, and mild AD. General inclusion/exclusion criteria are as follows:

(1) Normal subjects (52 images): Mini-mental state exam[15] (MMSE) scores between 24 and 30 (inclusive), a clinical dementia rating[16] (CDR) of 0, nondepressed, non-MCI, and nondemented. The age range of normal subjects will be roughly matched to that of MCI and AD subjects. Therefore, there should be minimal enrollment of normals under the age of 70.
(2) MCI subjects (114 images): MMSE scores between 24 and 30 (inclusive), a memory complaint, have objective memory loss measured by education adjusted scores on Wechsler Memory Scale Logical Memory II,[17] a CDR of 0.5, absence of significant levels of impairment in other cognitive domains, essentially preserved activities of daily living, and an absence of dementia.
(3) Mild AD (53 images): MMSE scores between 20 and 26 (inclusive), CDR of 0.5 or 1.0, and meets NINCDS-ADRDA criteria for probable AD proposed in 1984 by the National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's Disease and Related Disorders Association.[18–20]

Let us note that ADNI patient diagnostics are not pathologically confirmed; therefore, some uncertainty on the subject's labels are introduced unavoidably.

II.B. Image registration

The complexity of brain structures and the differences among the brains of different subjects make necessary the normalization of the images with respect to a common template. This ensures that the voxels in different images refer to the same anatomical positions in the brain.

First, the images have been normalized using a general affine model, with 12 parameters.[21,22] After the affine normalization, the resulting image is registered using a more complex nonrigid spatial transformation model. The deformations are parameterized by a linear combination of the lowest-frequency components of the three-dimensional cosine transform bases[23] using the SPM5 software (http://www.fil.ion.ucl.ac.uk/spm/software/spm5). A small-deformation approach is used and regularization is achieved by the bending energy of the displacement field.

Then, we normalize the intensities of the ^{18}F-FDG PET images with respect to the maximum intensity, which is computed for each image individually by averaging over the 0.1% of the highest voxel intensities. The intensity normalization process is very important to perform a voxel-by-

voxel comparison between different images. We choose this intensity normalization procedure because it accomplishes two different goals: On the one hand, changes in intensity are linear and, on the other hand, the effect of possible outliers in the data is mitigated.

Classification results using the proposed intensity normalization procedure outperforms the results obtained when intensity values are not normalized or when each ^{18}F-FDG PET image is linearly normalized to its maximum value. Of course, we do not claim this normalization procedure is the best one, but it allows us to obtain very good classification results, as it will be shown in Sec. III. Furthermore, this intensity normalization procedure has been successfully applied in other recent works.[24–27]

II.C. Preliminary voxel selection: t-student test

Each ^{18}F-FDG PET image has 62 322 voxels ($47 \times 39 \times 34$) with intensity values ranging from 0 to 1. Some of them correspond to positions outside the brain. We discard those voxels which present an intensity value lower than 0.35. Basically, positions outside the brain and some very low-intensity regions inside the brain are discarded.

Then, voxels are ranked using the absolute value two sample t-student test with pooled variance estimate

$$I^t = \frac{\bar{I}_{\pi_1} - \bar{I}_{\pi_2}}{S_{\pi_1 \pi_2} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \tag{1}$$

where $\bar{I}_{\pi_1}$ and $\bar{I}_{\pi_2}$ denote the mean ^{18}F-FDG PET image of subjects labeled as π_1 and π_2, respectively, n_i is the number of images in population π_i, and $S_{\pi_1 \pi_2}$ is an estimator of the common standard deviation of the two samples.

$$S_{\pi_1 \pi_2} = \sqrt{\frac{(n_1 - 1)S_{\pi_1}^2 + (n_2 - 1)S_{\pi_2}^2}{n_1 + n_2 - 2}}, \tag{2}$$

where S_{π_i} is the sample standard deviation image for population π_i, $i = 1, 2$ which is defined as

$$S_{\pi_i} = \sqrt{\frac{1}{n_i - 1} \sum_{j=1}^{n_i} (I_j - \bar{I}_{\pi_i})^2}. \tag{3}$$

Figure 1 shows the brain image I^t with the value of the t-test statistic in each voxel. In this example, normals and AD images were considered in the calculation of the image I^t.

Student's t-test gives us information about voxel class separability. We select those voxels i which present a t-statistic greater than a given threshold ε

$$\text{select voxels } i : \{I^t(i) > \varepsilon\}. \tag{4}$$

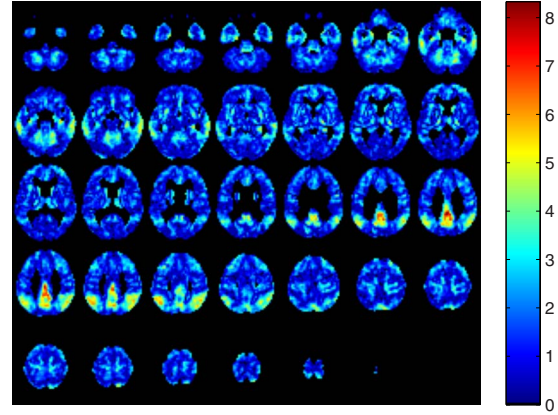Then, those selected voxels will be modeled using factor analysis.

FIG. 1. t-test value calculated in each voxel. Regions with larger value of the t-statistic are selected to perform classification.

II.D. Feature extraction: Factor analysis

Factor analysis is a statistical method used to describe variability among observed variables in terms of fewer unobserved variables called factors. The observed variables are modeled as linear combinations of the factors plus error. We use factor analysis to reduce the feature dimension. Factor analysis estimates how much of the variability in the data is due to common factors.

Suppose we have a set of N observable random variables, $x_1, \ldots, x_N$ with means $\mu_1, \ldots, \mu_N$.

Suppose for some unknown constants λ_{ij} and m unobserved random variables F_j, where $i \in 1, \ldots, N$ and $j \in 1, \ldots, m$, where $m < N$, we have

$$x_i = \lambda_{i1}F_1 + , \ldots, + \lambda_{im}F_m + \mu_i + z_i, \tag{5}$$

where z_i is independently distributed error terms with zero mean and finite variance different for each observable variable. The previous expression can be also written in matrix form

$$\mathbf{x} = \mathbf{\Lambda F} + \boldsymbol{\mu} + z, \tag{6}$$

where $\mathbf{x}$ is a vector of observed variables, $\boldsymbol{\mu}$ is a constant vector of means, $\mathbf{\Lambda}$ is a constant N-by-m matrix of factor loadings, $\mathbf{F}$ is a matrix of independent, standardized common factors, and z is a vector of independently distributed error terms.

The following assumptions are imposed to the unobserved random matrix $\mathbf{F}$: $\mathbf{F}$ and z are independent, the mean of $\mathbf{F}$ is equal to 0, and the covariance of $\mathbf{F}$ is the identity matrix. These assumptions allow computing the factor loadings $\mathbf{\Lambda}$ using a maximum likelihood approach.[28] Once the factor loadings have been estimated, we rotate them using a Varimax approach which is a change of coordinates that maximizes the sum of the variance of the squared loadings. This method attempts to maximize the variance on the new axes. Thus, we obtain a pattern of loadings on each factor that is as diverse as possible. These factor loadings will be used as features for classification purposes.

There are some known regions in AD and MCI brain PET image which present hypoperfusion [lower-intensity values
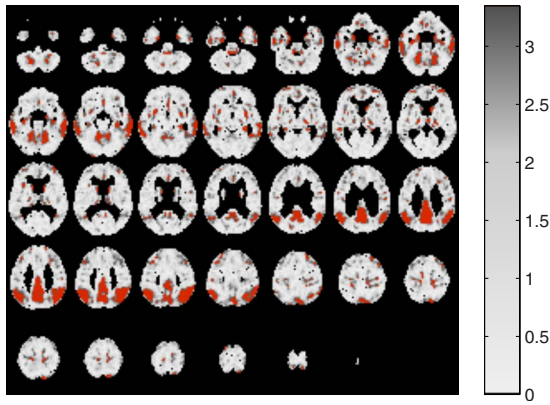
FIG. 2. Solid color: $p = 2000$ selected voxels when normal controls and AD patients are considered.

than for a given normal controls (NC) image]. This fact have been taken into account, preselecting those voxels in the images which present greater difference between NC, MCI, or AD images, measuring this difference using the t-test. By calculating the factor loadings considering only these preselected regions of the brain, we are not modeling the whole brain using factors but some specific regions of the images.

Thus, the vector of observed random variables $\mathbf{x}$ in the factor analysis model considered consists of a p-component vector of observations with the highest values of the t-test statistic measured in each voxel and intensity values greater than 0.35 for each PET brain image. In order to clarify what factors do actually model, in Fig. 2, we highlight selected voxels when $p = 2000$ and normal controls and Alzheimer's patients are considered. Therefore, in that case, $\mathbf{x}$ would consist of $p = 2000$ intensity values.

An interpretation of the meaning of the factor loadings and why they are useful to perform classification in brain image classification arises from Eq. (6). Each intensity value of the selected voxels after zero-mean normalization is expressed using a new basis using the factors $\mathbf{F}$. These factors, which are estimated via factor analysis, are common to all the images. Factor loadings $\mathbf{\Lambda}$ are the weights which are also calculated using factor analysis and they are different for each image. When images of different classes are expressed using the basis given by factors $\mathbf{F}$, they roughly present similar intraclass and different interclass values of factor loadings $\mathbf{\Lambda}$. We take advantage of this behavior and use factor loadings as features for classification purposes.

II.E. Classification

The goal of the classification task is to separate a set of binary labeled training data consisting of, in the general case, p-dimensional patterns $\mathbf{v}_i$ and class labels y_i

$$(\mathbf{v}_1, y_1), (\mathbf{v}_2, y_2), \dots, (\mathbf{v}_l, y_l) \in (R^p \times \{\pi_1, \pi_2\}), \tag{7}$$

so that a classifier is produced, which maps an object $\mathbf{v}_i$ to its classification label y_i. This classifier will be able to classify new examples $\mathbf{v}$.

There are several different procedures to build the classification rule. We utilize the following classifiers in this work.[29,30]

II.E.1. Multivariate normal model: Quadratic discriminant function

We suppose that $\mathbf{v}$ denotes a p-component random vector of observations made on any individual; $\mathbf{v}_0$ denotes a particular observed value of $\mathbf{v}$ and π_1 and π_2 denote the two populations involved in the problem. The basic assumption is that $\mathbf{v}$ has different probability distributions in π_1 and π_2. Let the probability density of $\mathbf{v}$ be $f_1(\mathbf{v})$ in π_1 and $f_2(\mathbf{v})$ in π_2. The simplest intuitive argument, termed the likelihood ratio rule, classifies $\mathbf{v}_0$ as π_1 whenever it has greater probability of coming from π_1 than from π_2. This classification rule can be written as

$$\mathbf{v}_0 \in \pi_1 \text{ if } f_1(\mathbf{v}_0)/f_2(\mathbf{v}_0) > 1, \tag{8}$$

$$\mathbf{v}_0 \in \pi_2 \text{ if } f_1(\mathbf{v}_0)/f_2(\mathbf{v}_0) \leq 1. \tag{9}$$

The most general form of the model is to assume that π_i is a multivariate normal population with mean $\boldsymbol{\mu}_i$ and dispersion matrix $\mathbf{\Sigma}_i$ for $i = 1, 2$. Thus $f_i(\mathbf{v}) = (2\pi)^{-p/2} |\mathbf{\Sigma}_i|^{-1/2} \exp\{\frac{1}{2}(\mathbf{v} - \boldsymbol{\mu}_i)' \mathbf{\Sigma}_i^{-1} (\mathbf{v} - \boldsymbol{\mu}_i)\}$, so that we obtain

$$\frac{f_1(\mathbf{v})}{f_2(\mathbf{v})} = |\mathbf{\Sigma}_2|^{1/2} |\mathbf{\Sigma}_1|^{-1/2} \exp\left[-\frac{1}{2}\{\mathbf{v}'(\mathbf{\Sigma}_1^{-1} - \mathbf{\Sigma}_2^{-1})\mathbf{v} \right.$$
$$\left. - 2\mathbf{v}'(\mathbf{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \mathbf{\Sigma}_2^{-1}\boldsymbol{\mu}_2) + \boldsymbol{\mu}_1'\mathbf{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2'\mathbf{\Sigma}_2^{-1}\boldsymbol{\mu}_2\} \right]. \tag{10}$$

Hence, on taking logarithms in Eq. (8), we find that the classification rule for this model is: Allocate $\mathbf{v}_0$ to π_1 if $Q(\mathbf{v}_0) > 0$ and otherwise to π_2, where $Q(\mathbf{v})$ is the discriminant function

$$Q(\mathbf{v}) = \frac{1}{2}\log\{|\mathbf{\Sigma}_2|/|\mathbf{\Sigma}_1|\} - \frac{1}{2}\{\mathbf{v}'(\mathbf{\Sigma}_1^{-1} - \mathbf{\Sigma}_2^{-1})\mathbf{v} - 2\mathbf{v}'(\mathbf{\Sigma}_1^{-1}\boldsymbol{\mu}_1$$
$$- \mathbf{\Sigma}_2^{-1}\boldsymbol{\mu}_2) + \boldsymbol{\mu}_1'\mathbf{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2'\mathbf{\Sigma}_2^{-1}\boldsymbol{\mu}_2\}. \tag{11}$$

Since the terms in $Q(\mathbf{v})$ include the quadratic form $\mathbf{v}'(\mathbf{\Sigma}_1^{-1} - \mathbf{\Sigma}_2^{-1})\mathbf{v}$, which will be a function of the squares of elements of $\mathbf{v}$ and cross products between pairs of them, this discriminant function is known as the quadratic discriminant.

In any practical application, the parameters $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \mathbf{\Sigma}_1$, and $\mathbf{\Sigma}_2$ are not known. Given two training sets, $\mathbf{v}_1^{(1)}, \dots, \mathbf{v}_{n_1}^{(1)}$ from π_1 and $\mathbf{v}_1^{(2)}, \dots, \mathbf{v}_{n_2}^{(2)}$ from π_2, we can estimate these parameters using the sample mean and sample standard deviation.

II.E.2. Multivariate normal model: Linear discriminant function

The presence of two different population dispersion matrices renders difficult the testing of hypothesis about the population mean vectors; therefore, the assumption $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{\Sigma}$ is a reasonable one in many practical situations. The

practical benefits of making this assumption are that the discriminant function and allocation rule simplifies. If $\Sigma_1 = \Sigma_2 = \Sigma$, then

$$f_i(\mathbf{v}) = (2\pi)^{-p/2}|\Sigma|^{-1/2}\exp\left\{-\tfrac{1}{2}(\mathbf{v}-\boldsymbol{\mu}_i)'\Sigma^{-1}(\mathbf{v}-\boldsymbol{\mu}_i)\right\} \quad (12)$$

and the classification rule reduces to allocate $\mathbf{v}$ to π_1 if $L(\mathbf{v}) > 0$ and otherwise to π_2, where $L(\mathbf{v}) = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}\{\mathbf{v} - \tfrac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\}$. No quadratic terms now exist in the discriminant function $L(\mathbf{v})$, which is therefore called the linear discriminant function.

In that case, contrary to the quadratic case, we estimate the pooled covariance matrix

$$\Sigma = \frac{1}{n_1 + n_2 - 2}\left\{\sum_{i=1}^{n_1}(\mathbf{v}_i^{(1)} - \boldsymbol{\mu}_1)(\mathbf{v}_i^{(1)} - \boldsymbol{\mu}_1)'\right. \quad (13)$$

$$\left. + \sum_{i=1}^{n_2}(\mathbf{v}_i^{(2)} - \boldsymbol{\mu}_2)(\mathbf{v}_i^{(2)} - \boldsymbol{\mu}_2)'\right\}. \quad (14)$$

II.E.3. Support vector machines with linear kernels

Support vector machines (SVMs) have recently been proposed for pattern recognition in a wide range of applications by its ability for learning from experimental data. SVMs separate a given set of binary labeled training data with a hyperplane that is maximally distant from the two classes (known as the maximal margin hyperplane). Linear discriminant functions define decision hypersurfaces or hyperplanes in a multidimensional feature space

$$g(\mathbf{v}) = \mathbf{w}^T\mathbf{v} + w_0 = 0, \quad (15)$$

where $\mathbf{w}$ is the weight vector and w_0 is the threshold. $\mathbf{w}$ is orthogonal to the decision hyperplane. The goal is to find the unknown weight vector $\mathbf{w}$ which defines the decision hyperplane.[30]

Let $\mathbf{v}_i$, $i = 1, 2, \ldots, l$ be the feature vectors of the training set. These belong to two different classes, π_1 or π_2, which for convenience in the mathematical calculations will be denoted as $+1$ and -1. If the classes are linearly separable, the objective is to design a hyperplane that classifies correctly all the training vectors. This hyperplane is not unique and it can be estimated maximizing the performance of the classifier, that is, the ability of the classifier to operate satisfactorily with new data. The maximal margin of separation between both classes is a useful design criterion. Since the distance from a point $\mathbf{v}$ to the hyperplane is given by $z = |g(\mathbf{v})|/\|\mathbf{w}\|$, the optimization problem can be reduced to the maximization of the margin $2/\|\mathbf{w}\|$, with constraints by scaling $\mathbf{w}$ and w_0 so that the value of $g(\mathbf{v})$ is $+1$ for the nearest point in π_1 and -1 for the nearest point in π_2. The constraints are the following:

$$\mathbf{w}^T\mathbf{v} + w_0 \geq 1, \ \forall \ \mathbf{v} \in \pi_1, \quad (16)$$

$$\mathbf{w}^T\mathbf{v} + w_0 \leq 1, \ \forall \ \mathbf{v} \in \pi_2, \quad (17)$$

or, equivalently, minimizing the cost function $J(\mathbf{w}) = 1/2\|\mathbf{w}\|^2$ subject to

$$y_i(\mathbf{w}^T\mathbf{v}_i + w_0) \geq 1, i = 1, 2, \ldots, l. \quad (18)$$

III. RESULTS

The performance of the classification is tested on a set of 210 ^{18}F-FDG PET images (52 normals, 114 MCI, and 53 AD) using the leave-one-out method: The classifier is trained with all but one image of the database and the remaining image, which is not used to define the classifier, is then categorized. In that way, all ^{18}F-FDG PET images are classified and the success rate is computed from the number of cor-
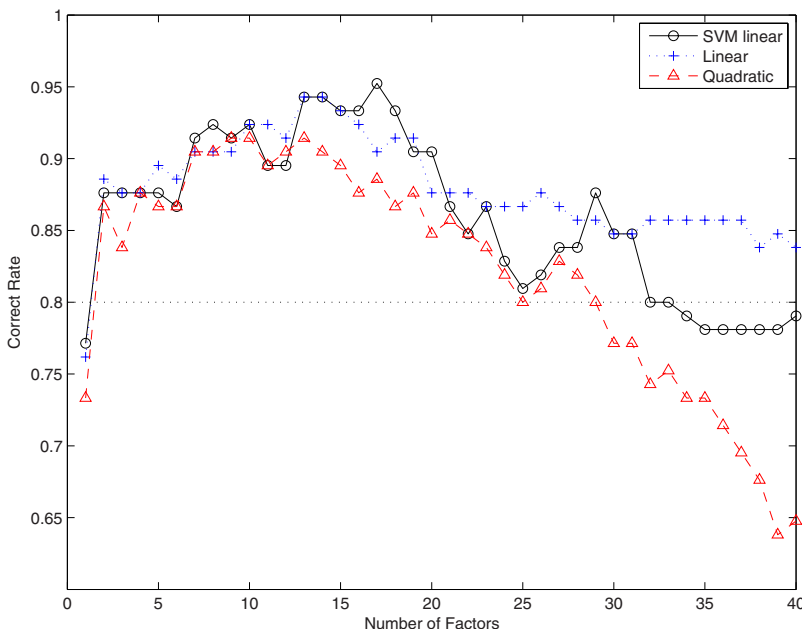

FIG. 3. Correct rate versus number of factors when NC and AD images are considered. The dotted line is the accuracy rate obtained using voxel-as-features.

rectly classified subjects. This cross-validation strategy has been previously used to assess the discriminative accuracy of different multivariate analysis methods applied to the early diagnosis of Alzheimer's disease,[26] discrimination of fronto-temporal dementia from AD,[13] and in classifying atrophy patterns based on magnetic resonance imaging images.[31]

Initially, threshold ε was set to 2000 voxels. Classification schemes used in this work are binary approaches; therefore, they are able to perform classification based on a training set of two previously labeled items. Nevertheless, we work with three classes of ^{18}F-FDG PET images, NC, MCI, and AD. Thus, we perform three different classification tasks. First, we consider only NC and AD images; second, NC versus MCI; and lastly we perform classification of NC subjects versus MCI-AD. The proposed methodology is compared to the voxel-as-features (VAF) approach with linear support vector machine classifier.[32–34]

III.A. NC versus AD

Figure 3 shows the correct rate versus the dimension *m* of factors *F* (number of unobserved random variables *F*). Linear classifiers perform better than the multivariate normal model with quadratic discriminant function. In general, the accuracy rate increases concomitantly with the dimension of factors for small values of *m*. Best correct rates are achieved when *m* is in the range 13–20 when we use SVM with linear kernel and multivariate normal model with linear discriminant function. Specifically, a correct rate up to 95% is reached when $m=17$ for SVM with linear kernels. For $m > 20$, performance of the classification task decreases with the dimension of the number of factors.

The ADNI database excludes advanced AD affected subjects. The ADNI database represents some ideal laboratory conditions, as subjects are recruited following a restricted clinical criteria to select only early AD related issues. Therefore, these favorable conditions allows proving the discrimination ability of the classifier in the discrimination of the Alzheimer's disease in an early stage of the disease.

Figure 4 shows the specificity and the sensitivity versus the number of factors for the three classifiers used in this work. The quadratic discriminant function reaches higher specificity values but very low sensitivity for a number of factors greater than 10. Specificity and sensitivity results obtained using the multivariate normal model with linear discriminant function and linear SVM are roughly the same except for number of factors greater than 30, where the linear discriminant function performs slightly better.

We have also drawn a scatter plot with samples from the loading factors when $m=2$ in Fig. 5. Figure 5 shows that even in the case in which two factor loadings are considered, most of the λ values for NC and AD images are clearly separated.

III.B. NC versus MCI

Figure 6 shows the correct rate versus number of factor loadings *F* when NC and MCI patients are considered. Our method presents, in that case, lower performance than when
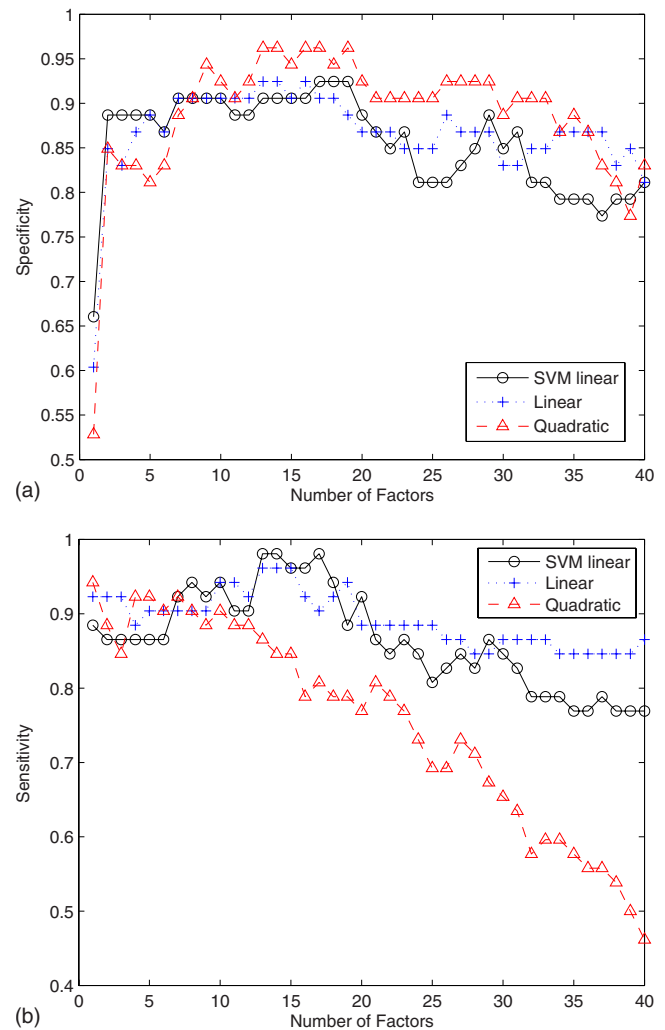
FIG. 4. (a) Specificity versus number of factors. (b) Sensitivity versus number of factors.

the NC and AD images were considered. This is due to the fact that images labeled as MCI present subjects with very different intensity patterns: Patients with memory complaint verified by a study partner, subjects presenting abnormal memory function, frontotemporal dementia, possibly MCI to AD converters, and subjects with general cognition and functional performance sufficiently preserved such that a diagnosis of Alzheimer's disease cannot be made at the time of the screening visit. This wide variability in the MCI group makes the automatic computer-aided diagnosis method presented in this paper to work slightly worse than when only NC and AD subjects were considered. Equally to the NC-AD case, the quadratic classifier performs worse than the linear ones. Furthermore, from $F=1$ to $F=14$, support vector machines with linear kernels performs better than the multivariate normal model with linear discriminant function, while for $F > 16$ they perform similarly. Our method presents better performance when the number of factor loadings is greater than 16. Specifically, the accuracy rate is greater than 85% and up to 88%.
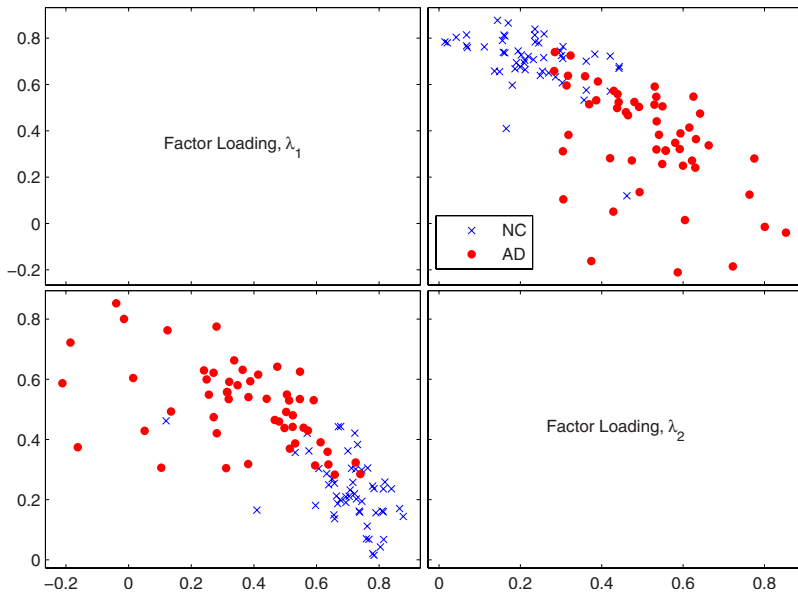
FIG. 5. Matrix of scatter plots for λ_1 and λ_2 when $m=2$.

Figure 7 shows the specificity and the sensitivity versus the number of factors. Let us note that the number of NC and MCI images are 52 and 114, respectively. In that case, the use of imbalanced training data set causes a possible poor performance of specificity or sensitivity since they are sample prevalence dependent. Quadratic discriminant function performs very bad in that case, reaching very high specificity values and very low sensitivity for a small number of factors. On the other hand, when the number of factors is greater than 35, quadratic discriminant function is not suitable for classification purposes since one obtains a sensitivity nearly 1 and a very low specificity. In that case, SVM linear and linear discriminant functions perform much better. Figure 7(a) shows that the linear discriminant function presents higher and more stable specificity values ($\sim$0.9) for a wide range of number of factors. Figure 7(b) shows a similar be-

havior for SVM linear when sensitivity is considered: High and very stable values ($\sim$0.9) for a wide range of number of factors.

III.C. NC versus MCI,AD

The number of NC and {MCI,DTA} images are 52 and 167, respectively; therefore, in that case, is inappropriate to calculate specificity or sensitivity values. A low specificity value in this largely unbalanced data set may not reflect a high false negative rate. Figure 8 depicts the accuracy rate versus the number of factors when NC subjects and a group composed of MCI and AD images are considered. In that case, SVM performs clearly better than the multivariate normal classifier with linear and quadratic discriminant function. Let note that this scenario is more difficult than classi-
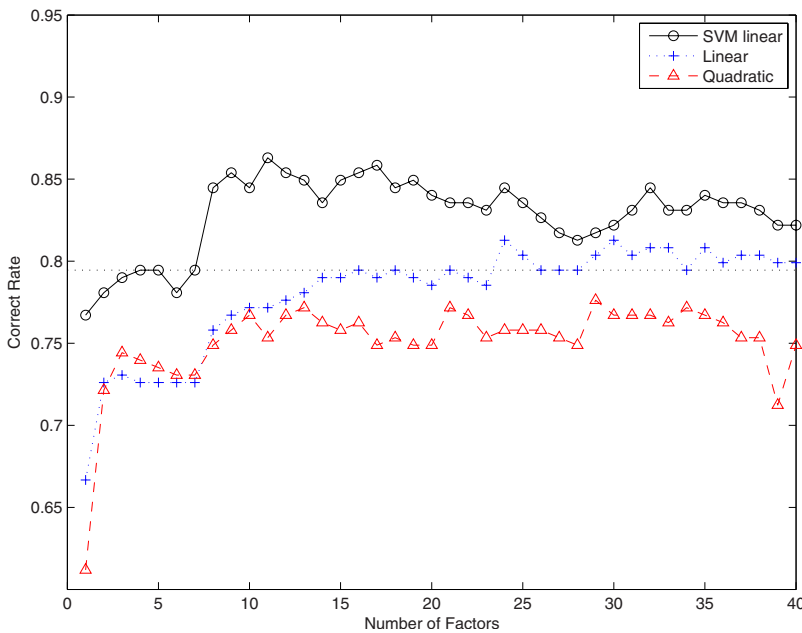

FIG. 6. Correct rate versus number of factors when NC and MCI subjects are considered. The dotted line is the accuracy rate obtained using VAF.
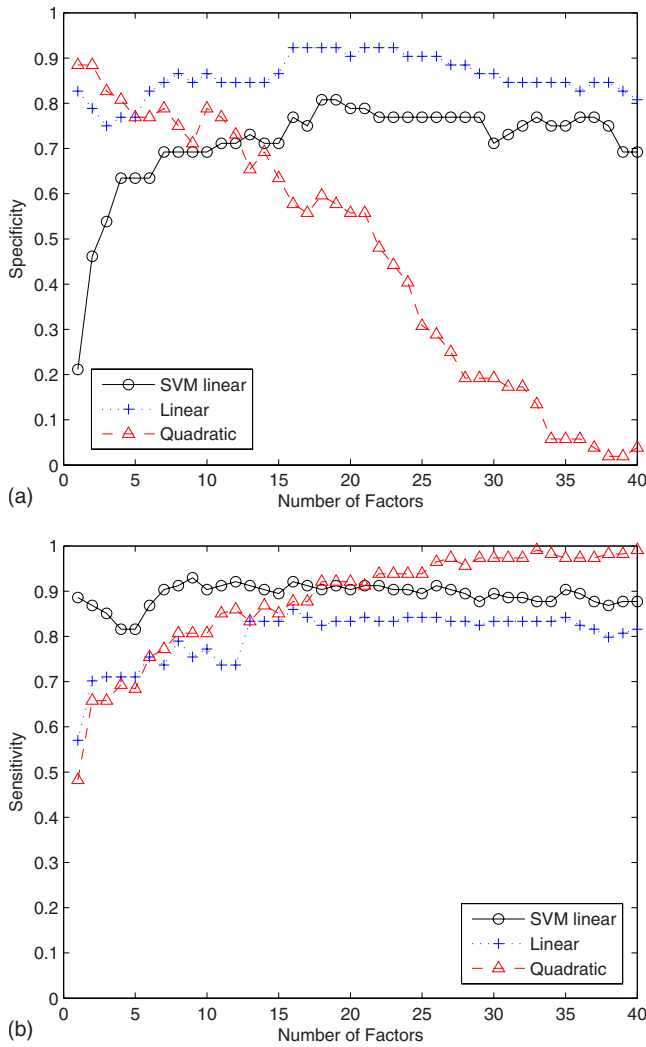
FIG. 7. (a) Specificity versus number of factors. (b) Sensitivity versus number of factors.

fying NC versus MCI or NC versus AD. In Sec. III D, the diversity of MCI subjects was pointed out. In that case, a new source of heterogeneity is included in the problem and this causes the accuracy rate to be slightly lower than when only NC and MCI were considered. Despite that, the proposed methodology reaches a correct rate greater than 85% when the number of factor loadings F ranges from 8 to 19. In that case, our method also outperforms the VAF approach when using SVM.

III.D. Performance for different threshold values ε

Up to now, threshold ε has been set to 2000 voxels. Simulations were also performed selecting 3000, 4000, and 5000 voxels (results not shown). In that case, the classification performance was slightly lower than when 2000 voxels were considered. Nevertheless, the computation time increases and the factor analysis algorithm is not able to reach convergence for large values of the number of factors. Therefore, either the relative convergence tolerance for varimax rotation (1.510^{-8}) needs to be reduced or the iteration limit (250 iterations) increased, leading to lower computational performance.

We also compare the performance of the method using different threshold values ε ($\varepsilon=2000, 1500, 1000, 500$). Table I presents the mean correct rate and standard deviation when the number of factors is in the interval $m=[10, 20]$. Best results are obtained when $\varepsilon=2000$. Nevertheless, very good performance of the proposed methodology is reached for each one of the threshold values considered.

IV. DISCUSSION

Multinormal models with linear or quadratic discriminant functions exhibit a strong dependence on outliers as they consider that the distribution of feature vector for populations π_1 and π_2 follow a Gaussian distribution. Thus, this
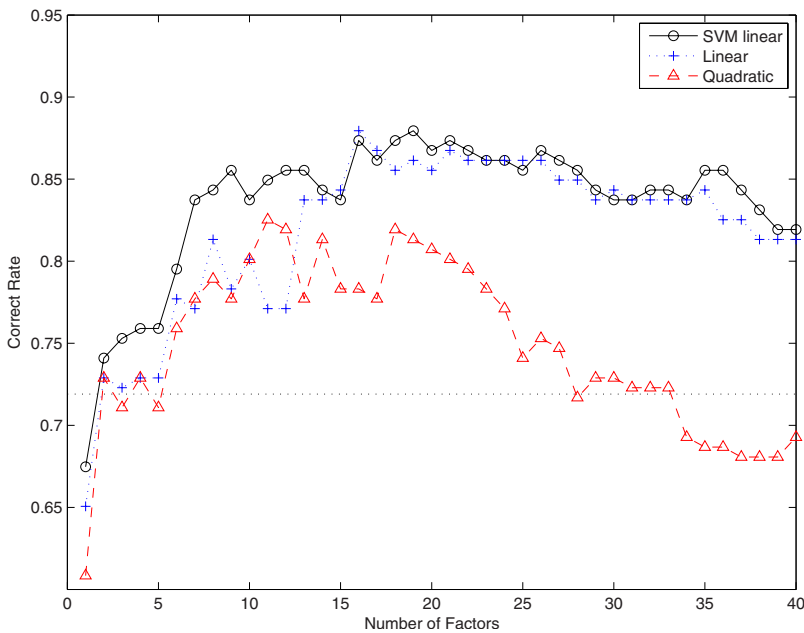

FIG. 8. Correct rate versus number of factors when NC and MCI,AD subjects are considered. The dotted line is the accuracy rate obtained using VAF.

TABLE I. Comparison of the mean accuracy rate for $m=10$ to $m=20$ number of factor loadings for different threshold values ε. Standard deviations in parentheses.

		$\varepsilon=2000$	$\varepsilon=1500$	$\varepsilon=1000$	$\varepsilon=500$
NC versus AD	SVM	0.92 (0.02)	0.90 (0.02)	0.87 (0.02)	0.86 (0.02)
	Linear	0.92 (0.01)	0.89 (0.01)	0.87 (0.01)	0.85 (0.01)
	Quadratic	0.89 (0.01)	0.85 (0.01)	0.87 (0.01)	0.80 (0.02)
NC versus MCI	SVM	0.86 (0.01)	0.86 (0.02)	0.85 (0.03)	0.82 (0.02)
	Linear	0.83 (0.03)	0.83 (0.02)	0.82 (0.03)	0.80 (0.01)
	Quadratic	0.80 (0.02)	0.78 (0.03)	0.77 (0.02)	0.80 (0.02)
NC versus MCI,AD	SVM	0.85 (0.01)	0.84 (0.01)	0.83 (0.02)	0.82 (0.01)
	Linear	0.79 (0.01)	0.77 (0.01)	0.77 (0.01)	0.75 (0.01)
	Quadratic	0.76 (0.01)	0.76 (0.01)	0.76 (0.01)	0.76 (0.01)

assumption could lead to poor results in the case in which the Gaussian assumption does not hold, as for instance, due to the presence of outliers in the populations. This is a typical problem which arises when the number of samples is small. Results in this work showed better performance when using SVM with linear kernel than in the case in which linear and quadratic discriminant functions were used. Specifically, in the case in which normal controls versus AD and MCI patients were considered, linear and quadratic discriminant functions obtained very poor performance. This could be due to the fact that including MCI and AD in the same group increases the heterogeneity of the feature vector and, therefore, the Gaussian assumption is not valid in that case. For this reason, and based on the results shown in the Sec. III, support vector machine with linear kernel is the preferred of the classifiers used in this work.

In order to discuss the behavior of multinormal models for classification and the performance of this classification methods for different number of samples, simulations have been repeated using a k-fold cross-validation scheme with $k=2$. Specifically, when $k=2$, $N/2$ images are used as training sets and the remaining $N/2$ images are considered test images. This allows checking the performance of the classification methods when a smaller number of images is considered, and therefore, it allows us to investigate the small size sample problem empirically. In Table II, the mean accuracy rate for $m=10$ to $m=20$ number of factor loadings obtained using leave-one-out and k-fold with $k=2$ are com-

TABLE II. Comparison of the mean accuracy rate for $m=10$ to $m=20$ number of factor loadings using leave-one-out and k-fold with $k=2$.

		Leave-one-out	Twofold
NC versus AD	SVM	0.92	0.87
	Linear	0.92	0.89
	Quadratic	0.89	0.85
NC versus MCI	SVM	0.86	0.83
	Linear	0.83	0.81
	Quadratic	0.80	0.77
NC versus MCI,AD	SVM	0.85	0.82
	Linear	0.79	0.77
	Quadratic	0.76	0.76

pared. The table shows that the best performance is obtained when leave-one-out is used as validation method which is an expected result. Nevertheless, let us note that good performance is also obtained when only half of the available images are used to train the classifier.

In this work, we do not propose an analytical procedure to estimate the value of the number of factor loadings. Nevertheless, some order selection technique could be possibly used to select the number of factor loadings. We estimate the performance of the classification task for a different number of λ. This procedure allows to easily estimate an "optimal" value of the threshold in the sense that it lets us select those values of the number of factor loadings which better discriminate between populations. Thus, the plot of the accuracy rate versus the threshold enables to visually select an useful value (or a range of values) of λ for classification.

Overfitting generally occurs when a model is excessively complex, such as having too many degrees of freedom, in relation to the amount of data available. In our case, this is a possible scenario when using support vector machines with nonlinear kernels or nonlinear discriminant functions to perform classification. For this reason, we only use SVM with linear kernel. Therefore, choosing linear classifiers, or in that case, SVM with linear kernel, is supported not only by the good performance results but also by the fact that this choice lets us to avoid possible overfitting.

In addition, we would like to point out that choosing a wide range of λ values and calculating the performance of the classification task is enough for our practical purposes for different reasons. On the one hand, the estimation of a very precise and specific number of factors m is not critical as the accuracy of the proposed methodology is high for a wide range of m values for the data set studied (as instance, from $m=2$ to $m=20$ when NC versus AD was considered, and $m>7$ and $m>8$ when NC versus MCI and NC versus MCI,AD). On the other hand, if we choose to estimate the threshold value using an analytical procedure, lastly, the performance of this analytical selection procedure based on a statistical criterion should be checked in terms of accuracy, specificity, and sensitivity for different values of the number of factor loadings in order to prove the validity of the method used to estimate the order model. And, actually, this is the experimental method we use in that work to estimate useful

TABLE III. Comparison between the performance of the FA proposed in this work and the PCA-based method (PCA) in Ref. 35. *m* is the number of factors or components.

		FA	PCA
NC versus AD	Acc	95.2	89.5
	Sen	98.1	84.5
	Spe	92.5	84.7
	m	17	8
NC versus MCI	Acc	88.0	81.3
	Sen	91.2	97.4
	Spe	80.8	46.1
	m	19	30
NC versus MCI,AD	Acc	86.3	82.2
	Sen	92.8	92.21
	Spe	65.4	50
	m	11	30

FIG. 10. Mean of diagonal values in matrix cov(**z**) versus number of iterations. Circles: NC,AD. Squares: NC,MCI. Triangles: NC,MCI,AD.

values of the number of factors. Furthermore, the estimation of the number of λ, equal to the training procedure, only needs to be performed once as a "batch" process.

In Ref. 35, a CAD system for an automatic evaluation of the neuroimages is presented. In that work, principal component analysis (PCA) based methods are proposed as feature extraction techniques, enhanced by other linear approaches such as linear discriminant analysis or the measure of the Fisher discriminant ratio for feature selection. These features allow surmounting the so-called small sample size problem. We have used the PCA-based feature extraction method presented there as the input vector of a SVM classifier with linear kernel in order to compare to our proposed methodology. For the data set used in this work, best values of the correct rate, specificity, and sensitivity using both approaches are shown in Table III.

Despite of the fact that an empirical estimation of *m* is enough for our purposes, we suggest a possible method to calculate an optimal *m* value using an analytical procedure studying the value of the log-likelihood. Figure 9 plots the
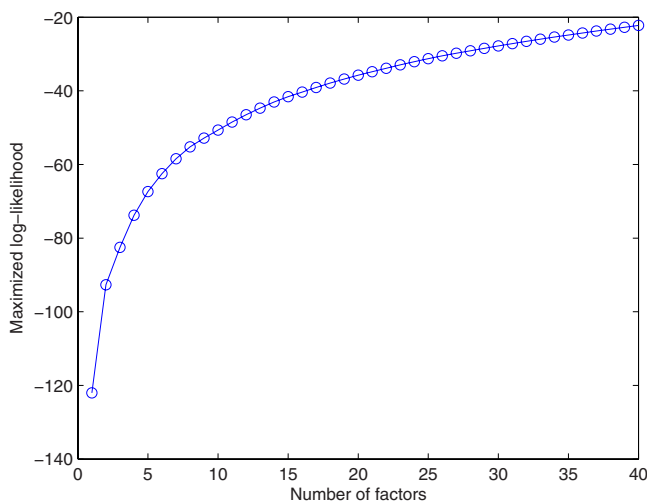

FIG. 9. Log-likelihood versus the number of factors *m* in the case in which images NC versus MCI,AD are considered.
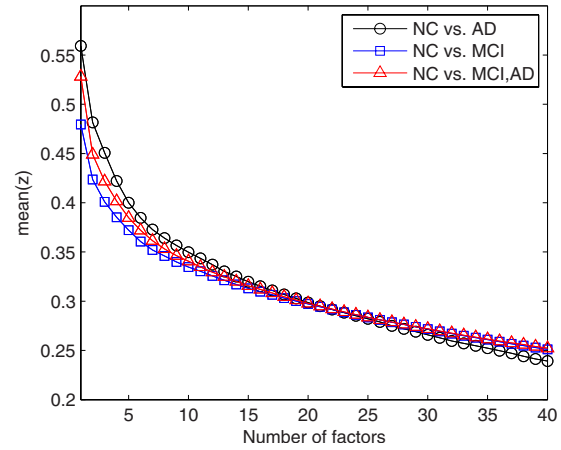
log-likelihood versus *m* (the number of factors) in the case in which NC versus MCI,AD images are considered. It can be seen how increasing *m* also increases the ability of the factor loadings to model the data more accurately (the log-likelihood increases). Log-likelihood values when *m* is small (from *m*=1 to *m*=7) present a great variability. This coincides with the same range of values in which our algorithm does not provide high correct rate, possibly due to the fact that the number of factors is not enough to model correctly the variability of the data.

The factor analysis model in Eq. (6) can also be specified as

$$\text{cov}(\mathbf{x}) = \mathbf{\Lambda}\mathbf{\Lambda}^T + \text{cov}(\mathbf{z}), \tag{19}$$

where cov(**z**) is a *p*-by-*p* diagonal matrix of specific variances and *p* is the number of observed variables. Finally, we study the values of cov(**z**) representing the mean of the diagonal values of matrix cov(**z**) for each group NC versus AD, NC versus MCI, and NC versus MCI,AD in Fig. 10. In these three cases the values are very similar and, as it was expected, they decrease as the number of factors increase, because for greater values of *m*, the factors are more capable to model the observed random variables.

V. CONCLUSIONS

In this work, an automatic procedure to assist the diagnosis of early Alzheimer's disease is presented. The proposed methodology is based on the selection of voxels of interest using the t-test and a posterior reduction of the feature dimension using factor analysis. Factor loadings were used as features of three different classifiers: Two multivariate Gaussian mixture models, with linear and quadratic discriminant function, and a support vector machine with linear kernel which was found to achieve the highest accuracy rate. The best results were obtained when normal and Alzheimer's disease subjects were considered. Specifically, an accuracy rate greater than 90% were obtained in that case for a wide range of number of factors. Furthermore, results were com-

pared to the voxel-as-features and a PCA-based approach. The proposed methodology was found to perform clearly better.

a)Electronic mail: dsalas@ugr.es; URL: http://ugr.es/~dsalas.

b)Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (www.loni.ucla.edu/ADNI). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. ADNI investigators include (complete listing available at http://www.loni.ucla.edu/ADNI/Collaboration/ADNI_Manuscript_Citations.pdf).

[1] R. Brookmeyer, S. Gray, and C. Kawas, "Projections of Alzheimer's disease in the United States and the public health impact of delaying disease onset," JAMA, J. Am. Med. Assoc. **88**, 1337–1342 (1998).

[2] R. Brookmeyer, E. Johnson, K. Ziegler-Graham, and M. Arrighi, "Forecasting the global burden of Alzheimer's disease," Alzheimer's & Dementia **3**, 186–191 (2007).

[3] E. Johnson, R. Brookmeyer, and K. Ziegler-Graham, "Modeling the effect of Alzheimer's disease on mortality," The International Journal of Biostatistics 3(1), Article 13 (2007).

[4] G. E. Alexander, P. Pietrini, S. I. Rapoport, and E. M. Reiman, "Longitudinal PET evaluation of cerebral metabolic decline in dementia: A potential outcome measure in Alzheimer's disease treatment studies," Am. J. Psychiatry **159**, 738–745 (2002).

[5] J. B. Langbaum, K. Vhen, W. Lee, C. Reschke, D. Bandy, A. S. Fleisher, G. E. Alexander, N. L. Foster, M. W. Weiner, R. A. Koeppe, W. J. Jagust, and E. M. Reiman, "Categorical and correlation analyses of baseline fluorodeoxyglucose positron emission tomography images from the Alzheimer's Disease Neuroimaging Initiative (ADNI)," Neuroimage **45**, 1107–1116 (2009).

[6] S. Minoshima, B. Goirdani, S. Berent, K. A. Frey, N. L. Foster, and D. E. Khul, "Metabolic reduction in the posterior cingulate cortex in very early Alzheimer's disease," Ann. Neurol. **42**, 85–94 (1997).

[7] L. Mosconi *et al.*, "Multicenter standardized 18F-FDG PET diagnosis of mild cognitive impairment, Alzheimer's disease and other dementias," J. Nucl. Med. **49**, 390–398 (2008).

[8] D. H. Silverman, G. W. Small, and C. Y. Chang, "Positron emission tomography in evaluation of dementia: Regional brain metabolism and long-term outcome," JAMA, J. Am. Med. Assoc. **286**, 2120–2127 (2001).

[9] H. Braak and E. Braak, "Diagnostic criteria for neuropathologic assessment of Alzheimer's disease," Neurobiol. Aging **18**, S85–S88 (1997).

[10] J. L. Cummings, H. V. Vinters, G. M. Cole, and Z. S. Khachaturian, "Alzheimer's disease: Etiologies, pathophysiology, cognitive reserve, and treatment opportunities," Neurology **51**, S2–S17 (1998).

[11] J. M. Hoffman, K. A. Welsh-Bohmer, and M. Hanson, "FDG PET imaging in patients with pathologically verified dementia," J. Nucl. Med. **41**, 1920–1928 (2000).

[12] S. Ng, V. L. Villemagne, S. Berlangieri, S.-T. Lee, M. Cherk, S. J. Gong, U. Ackermann, T. Saunder, H. Tochon-Danguy, G. Jones, C. Smith, G. O'Keefe, C. L. Masters, and C. C. Rowe, "Visual assessment versus quantitative assessment of 11C-PIB PET and 18F-FDG PET for detection of Alzheimer's disease," J. Nucl. Med. **48**, 547–552 (2007).

[13] R. Higdon, N. L. Foster, R. A. Koeppe, C. S. DeCarli, W. J. Jagust, C. M. Clark, N. R. Barbas, S. E. Arnold, R. S. Turner, J. L. Heidebrink, and S. Minoshima, "A comparison of classification methods for differentiating fronto-temporal dementia from Alzheimer's disease using FDG-PET imaging," Stat. Med. **23**, 315–326 (2004).

[14] R. P. W. Duin, "Classifiers in almost empty spaces," in *Proceedings of the 15th International Conference on Pattern Recognition* (IEEE, Barcelona, 2000), Vol. 2, pp. 1–7.

[15] M. Folstein, S. E. Folstein, and P. R. McHugh, "'Mini-mental state': A practical method for grading the cognitive state of patients for the clinician," J. Psychiatr. Res. **12**, 189–198 (1975).

[16] J. C. Morris, "Clinical dementia rating," Neurology **43**, 2412–2414 (1993).

[17] O. Spreen and E. Strauss, *A Compendium of Neuropsychological Tests: Administration, Norms, and Commentary*, 2nd ed. (Oxford University Press, New York, 1998).

[18] D. Blacker, M. S. Albert, S. S. Bassett, R. C. Go, L. E. Harrell, and M. F. Folstein, "Reliability and validity of NINCDS-ADRDA criteria for Alzheimer's disease. The National Institute of Mental Health Genetics Initiative," Arch. Neurol. **51**, 1198–1204 (1984).

[19] D. A. Loewenstein, R. M. D. Ownby, L. Schram, A. Acevedo, M. Rubert, and T. Argelles, "An evaluation of the NINCDS-ADRDA neuropsychological criteria for the assessment of Alzheimer's disease: A confirmatory factor analysis of single versus multi-factor models," J. Clin. Exp. Neuropsychol **23**, 274–284 (2001).

[20] G. McKhann, D. Drachman, M. Folstein, R. Katzman, D. Price, and E. M. Stadlan, "Clinical diagnosis of Alzheimer's disease: Report of the NINCDS-ADRDA work group under the auspices of Department of Health and Human Services Task Force on Alzheimer's disease," Neurology **34**, 939–944 (1984).

[21] D. Salas-Gonzalez, J. M. Górriz, J. Ramírez, A. Lassl, and C. G. Puntonet, "Improved Gauss-Newton optimization methods in affine registration of SPECT brain images," Electron. Lett. **44**, 1291–1292 (2008).

[22] R. P. Woods, S. T. Grafton, C. J. Holmes, S. R. Cherry, and J. C. Mazziotta, "Automated image registration: I. General methods and intrasubject, intramodality validation," J. Comput. Assist. Tomogr. **22**, 139–152 (1998).

[23] J. Ashburner and K. J. Friston, "Nonlinear spatial normalization using basis functions," Hum. Brain Mapp **7**, 254–266 (1999).

[24] R. Chaves, J. Ramírez, J. M. Górriz, M. López, D. Salas-Gonzalez, I. Álvarez, and F. Segovia, "SVM-based computer-aided diagnosis of the Alzheimer's disease using t-test NMSE feature selection with feature correlation weighting," Neurosci. Lett. **461**, 293–297 (2009).

[25] J. M. Górriz, A. Lassl, J. Ramírez, D. Salas-Gonzalez, C. G. Puntonet, and E. W. Lang, "Automatic selection of ROIs in functional imaging using Gaussian mixture models," Neurosci. Lett. **460**, 108–111 (2009).

[26] J. Ramírez, J. M. Górriz, D. Salas-Gonzalez, A. Romero, M. López, I. Alvarez, and M. Gómez-Río, "Computer aided diagnosis of Alzheimer's type dementia combining support vector machines and discriminant set of features," Information Sciences (in press).

[27]D. Salas-Gonzalez, J. M. Górriz, J. Ramírez, M. López, I. Álvarez, F. Segovia, R. Chaves, and C. G. Puntonet, "Computer aided diagnosis of Alzheimer's disease using support vector machines and classification trees," Phys. Med. Biol. **55**, 2807–2817 (2010).

[28]H. H. Harman, *Modern Factor Analysis* (University of Chicago Press, Chicago, 1976).

[29]*Principles of Multivariate Analysis: A User's Perspective*, edited by W. J. Krzanowski (Oxford University Press, New York, 1988).

[30]V. N. Vapnik, *Statistical Learning Theory* (Wiley, New York, 1998).

[31]Y. Fan, N. Batmanghelich, C. M. Clark, C. Davatzikos, and the Alzheimer's Disease Neuroimaging Initiative, "Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline," Neuroimage **39**(4), 1731–1743 (2008).

[32]G. Fung and J. Stoeckel, "SVM feature selection for classification of SPECT images of Alzheimer's disease using spatial information," Knowledge Inf. Syst. **11**, 243–258 (2007).

[33]J. Stoeckel, N. Ayache, G. Malandain, P. M. Koulibaly, K. P. Ebmeier, and J. Darcourt, "Automatic classification of SPECT images of Alzheimer's disease patients and control subjects," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2004*, edited by C. Barillot, D. R. Haynor, and P. Hellier (Springer, Berlin/Heidelberg, 2004) [Lect. Notes Comput. Sci. **3217**, 654–662 (2004).

[34]J. Stoeckel, G. Malandain, O. Migneco, P. M. Koulibaly, P. Robert, N. Ayache, and J. Darcourt, "Classification of SPECT images of normal subjects versus images of Alzheimer's disease patients," in *Medical Image Computing and Computer-Assisted Intervention Conference—MICCAI 2001*, edited by W. Niessen and M. Viergever (Springer, Berlin/Heidelberg, 2001) [Lect. Notes Comput. Sci. **2208**, 666–674 (2001).

[35]M. López, J. Ramírez, J. M. Górriz, D. Salas-Gonzalez, I. Álvarez, F. Segovia, and C. G. Puntonet, "Automatic tool for the Alzheimer's disease diagnosis using PCA and Bayesian classification rules," Electron. Lett. **45**, 389–391 (2009).